Analytics-assisted triage of psychological workers' compensation claims

MACQUARIE
University

Kyu Hyung Park

Supervisor: Prof. Leonie Tickle

Faculty of Business and Economics
Macquarie University, Sydney

This thesis is submitted for the degree of
Master of Research

9 October 2017

Table of contents

List of figures

List of tables

Abstract

Workers' compensation is a form of insurance for employers providing income replacement, medical benefits and rehabilitation support to eligible workers suffering a work-related injury or illness. The cost of work-related injuries and illnesses is significant in Australia, amounting to 4.1 percent of GDP in terms of total economic cost. Among all injuries, psychological injury is the most expensive form of workers' compensation claim due to its typically long duration. Despite the importance and cost of work-related psychological injury, the factors associated with prolonged claim duration are still not well understood. Using data provided by the workers' compensation agency in South Australia, this research identifies factors associated with the duration of workers' compensation claims for psychological injuries (psychological claims), and develops a practical and informative business model to aid the management of such claims by using modern analytics techniques including classification tree and association rule learning. We find that the factor most associated with duration of psychological claims is occupation, followed by bodily location of the most recent prior claim and age of injured worker. It is found that, among psychological claims, those made by claimants in high socio-economic occupations are at higher risk of prolonged claim duration. We finally develop a triage model that uses these factors to segment claims according to risk of prolonged duration. The model enables the focusing of efforts and resources on high-risk claims, thereby reducing the economic and societal burden of work-related injury.

Statement of Originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

9 October 2017

Kyu Hyung Park

Acknowledgements

Firstly, I would like to express my sincere appreciation to my supervisor Professor Leonie Tickle for the continuous guidance, insightful comments and engagement through the learning process of this thesis, without which this work would not have been possible.

My sincere thanks also goes to Associate Professor Ayse Bilgin for the valuable advice on data mining techniques and the useful comments on the thesis.

I am also grateful to ReturnToWorkSA for providing the data and to Colin Khoo for kindly and clearly answering my follow-up questions about the data.

Finally, I would also like to acknowledge with gratitude, the support and love of my family.

Chapter 1

Introduction

This research identifies factors associated with the duration of workers' compensation claims for all causes combined and for work-related psychological injuries such as depression and anxiety, and develops a practical and informative business model to aid the management of claims by using modern analytics techniques.

Workers' compensation is a form of insurance for employers providing income replacement, medical benefits and rehabilitation support to eligible workers suffering a work-related injury or illness.

The cost of work-related injuries and illnesses is significant. Of the 12.5 million persons in the Australian labour force during the 2013/2014 financial year, 4.3 percent suffered a workplace injury or illness (Australian Bureau of Statistics, 2014). The total economic cost of work-related injuries, illnesses and deaths for the 2012/2013 financial year has been estimated at $61.8 billion, or 4.1 percent of GDP.

A crucial component of managing workers' compensation costs is identifying claims with a high risk of having a long duration of income maintenance payments (claim duration). Claim durations are often used as performance measures for providers in the workers' compensation system and as estimates for the overall burden of work-related injuries or illnesses on society (Krause, Frank, Dasinger, Sullivan and Sinclair, 2001). Claim duration can be reduced by various means such as early intervention; effective claims management; workplace rehabilitation programs; cooperation, collaboration and consultation between parties involved; and incentive measures. For example, Donceel et al. (1999) found that workers are more likely to return to work within one year after injury when they are guided by medical advisers. Such plans can be more effectively applied if it can be identified in advance which claims are at high risk of becoming long duration and therefore high cost.

Factors found to be associated with claim duration can be used to develop a business model to improve outcomes. Lebedev et al. (2015) developed a triage model for workers' compensation claims. They first identified 36 characteristics of workers' compensation claims at 13 weeks after injury that indicate an increased likelihood of claim duration exceeding one year. Among the identified characteristics, most influential were type of injury, age at injury, incapacity description from the last known medical certificate, and specialty of most frequented provider. Using these characteristics, the whole set of claims are segmented into high and low risk segments with the aid of data mining techniques. This facilitates prioritising initiatives for high-risk segments.

Psychological injury such as work-related depression, anxiety and post-traumatic stress disorder is the most expensive form of workers' compensation claim due to its typically long duration, and has significant impact on injured workers and society (Safe Work Australia, 2013). In Australia, the median claim payment for such injury during the five year period to mid-2013 was $23,600, which is around 2.7 times the global median of $8,700 (Safe Work Australia, 2015*b*, 2017).

There are numerous studies involving aspects of psychological factors and return to work. For example, Carnide et al. (2016) investigated symptoms of depression following work-related musculoskeletal injuries requiring sick leave, and found that progression of the symptoms is associated with whether the claim duration goes longer than 12 months after injury. In an investigation of Korean workplaces, Lu et al. (2014) found an association between duration of leave due to work-related injury and psychosocial job factors.

However, despite the importance and cost of work-related psychological injury, the factors associated with prolonged duration of claims due to psychological injuries are still not well understood. The factors are known to be different to those affecting other injuries or illnesses. For example, Nieuwenhuijsen et al. (2004) found that while better supervisory support is associated with earlier return to work for non-depressed employees, such association could not be established for employees with a high level of depressive symptoms. In a meta-analysis, Blank et al. (2008) investigated associations between various factors and prolonged duration of claims due to psychological injuries. However, most papers studied in the analysis did not find factors directly associated with duration of work-related mental health conditions. For example, Stansfeld et al. (1999) found that lower work grade (e.g. high job demand) is associated with higher risk of psychiatric disorder, which may be different to the risk of prolonged claim duration. Therefore, generalisations of the findings are limited.

To identify the factors associated with prolonged claim duration and to develop a model to aid the management of workers' compensation claims, this study applies data mining

techniques to a large Australian dataset. The data is provided by the workers' compensation agency in South Australia, ReturnToWorkSA. It contains 209,493 accepted workers' compensation claims, filed for the eleven financial years to 2014/2015. Of the claims, 7,088 claims were primarily made for psychological injury. The data has 84 fields for which the recorded values are observable at the time a claim is made, plus the ultimate duration of claim which is the dependent variable for this study.

For a large and noisy dataset such as this, involving complex relationships between numerous fields, conventional statistical methods such as regression are unsuitable. Instead, data mining - the process of extracting new and beneficial information from big datasets by using the techniques of statistics, data analysis, and machine learning (Tufféry, 2011) - is used. The application of data mining is useful in insurance because modelling often involves a dataset containing large number of cases as well as many variables (Kolyshkina et al., 2003, p.493).

Of the data mining techniques, tree-based methods are found to be most suitable in this study. These methods are chosen because they have the ability to identify the most important predictors for prolonged claim duration, and segment claims by the risk of having a prolonged claim duration. Such segmentation of claims can be used to develop a triage model for workers' compensation claims.

There are three aspects of this research that make a new and original contribution to the existing literature. First, we identify the most important factors observable from new claims associated with duration of claims due to psychological injuries, and compare the factors with those for overall claims. Second, we develop a practical triage model to specify the claims more likely to have a prolonged duration; this can be used by claims managers and others to reduce the heavy financial and societal burden of psychological work-related injury. Third, we find that, although it is difficult to predict a duration of claims individually, robust segmentations of claims by risk of prolonged duration can be made.

The rest of this paper is organised as follows. Chapter 2 reviews the related literature. Chapter 3 provides overview of and preliminary findings from the data. Chapter 4 discusses the methodological framework. While Chapter 5 summarises the results from the analysis, they are discussed in Chapter 6. Finally, Chapter 7 concludes.

Chapter 2

Literature Review

2.1 Workers' compensation in Australia: overview and costs

Workers' compensation is a compulsory statutory form of insurance for all employers in Australia and provides income replacement, healthcare and rehabilitation support to eligible workers suffering a work-related injury or illness. An array of workers' compensation schemes in Australia has been established by Commonwealth, state and territory governments with the aim of returning ill or injured workers to the labour force while minimising the rehabilitation costs to society (Collie et al., 2016). These include the autonomous workers' compensation schemes geographically based in the six states and two territories in addition to three Commonwealth schemes (Safe Work Australia, 2016).

The worker's compensation schemes are managed differently between jurisdictions. For example, in NSW, an insurance policy is designed by the agency overseeing the scheme for NSW, State Insurance Regulatory Authority (SIRA), while a premium is determined by insurers, subject to SIRA's oversight (Safe Work Australia, 2016, p.13). However, in SA, an insurance policy and a premium are both set by the agency for SA, ReturnToWorkSA (Safe Work Australia, 2016, p.13).

The types of benefits given to an injured worker entitled to workers' compensation include income replacement payments; costs of medical and hospital treatment; permanent impairment entitlements; death entitlements; and other benefits such as settlement of future incapacity benefits (Safe Work Australia, 2016, p.89).

The cost of work-related injury or illness in Australia is significant. Of the 12.5 million persons in the Australian labour force during the 2013/2014 financial year, 4.3 percent suffered a workplace injury or illness (Australian Bureau of Statistics, 2014). The total economic cost of work-related injuries, illnesses and deaths for the 2012/2013 financial year

has been estimated at $61.8 billion, or 4.1 percent of GDP. Approximately 25 percent of the cost is estimated to be due to direct costs such as workers' compensation premiums paid by employers or payments to eligible claimant (Safe Work Australia, 2015*a*). The remainder is attributable to indirect costs such as lost productivity, loss of current and future earnings, lost potential output and the cost of providing social welfare programs for claimants. This ratio is in line with the finding of Snook and Webster (1987) for US workers' compensation systems that indirect costs are estimated to be two to four times greater than direct workers' compensation costs. Having said that, these measures are still likely to underestimate the actual magnitude of the problem due to unreported cases and unmeasurable costs of loss of quality of life (Safe Work Australia, 2013).

Attempts to reduce the cost of workers' compensation in Australia are ongoing (Safe Work Australia, 2016). Since the 1970s, schemes have put more emphasis on strengthening work health and safety, and on rehabilitation of injured workers. While initiatives have not achieved the targeted level of cost reduction, costs had fallen as much as 20 percent of total labour cost by the mid 1990s with further reform attempts focusing on cutting back benefits and making premiums more competitive. Despite the significant improvement to date, each authority continues to make efforts to reduce costs.

2.2 Duration of workers' compensation claims

Return-to-work outcomes, or durations of income maintenance payments on workers' compensation claim (claim duration) are often used as performance measures for providers in the workers' compensation system including health care providers, vocational rehabilitation services, and workers' compensation insurers (Krause, Frank, Dasinger, Sullivan and Sinclair, 2001). Combined with frequency measures, they are indicators of the social and economic cost of work-related injuries and illnesses. A claim duration measure demonstrates several aspects of the cost. First, it represents the burden of work-related injuries or illnesses on the overall society or specific segments of the society such as injured workers themselves, their families, local communities, employers, and industries (Krause, Frank, Dasinger, Sullivan and Sinclair, 2001). Second, it is a sound measure to determine the effectiveness and efficiency of provisions in the workers' compensation system including medical support, return-to-work programs, and other policies and initiatives aiming at supporting workers (Krause, Frank, Dasinger, Sullivan and Sinclair, 2001).

2.3 Reducing claim duration

Claim duration can be reduced by various means such as early intervention; effective claims management; workplace rehabilitation programs; cooperation, collaboration and consultation between parties involved; and incentive measures.

Return-to-work programs operated by employers are designed for timely, safe, and cost-efficient return to work of injured or ill workers by providing guidelines and requirements with regards to work-related injuries for an employer, a worker, an insurer and other participants through the above means. Employers are required by legislation to have such a program in most workers' compensation schemes[1] (Safe Work Australia, 2014).

An early intervention by an employer to a worker who shows signs of problems may reduce the claim duration by allowing them to act before the problem becomes severe, or may even prevent them from taking sick leave, in the case of injuries and illnesses that develop over time including mental health conditions. Many companies implement early interventions as part of employee assistance programs comprised of provisions such as fitness testing specific to job task, weight and nutrition management, cholesterol screening and stress counselling.

Studies also have found that active monitoring of claimants by insurers shortens claim duration (see Krause, Frank, Dasinger, Sullivan and Sinclair, 2001). An illustration is the finding from the study of Donceel et al. (1999) in Belgium that workers are more likely to return to work within one year after injury when they are guided by medical advisers.

Workplace rehabilitation programs promote returning workers to suitable work at the earliest possible time. These programs may aim at returning a worker to the previous job by helping the worker to keep their work skills, or assisting the worker to find another job through on-the-job training to acquire new job skills, depending on cases. Eggert (2010)'s literature review support the need for such programs by suggesting that rehabilitation programs have psychosocial influence on injured workers and thus help return to work.

Incentives are another way to shorten claim durations by influencing participant behaviours which have a significant impact on the outcomes (see Hardy et al., 2011). An example of incentive measures for workers is reducing the level of income maintenance benefit as claim duration becomes longer, while that for employers is return-to-work incentives (typically financial incentives to an employer for a timely return to work of their employees). Hardy et al. (2011) stated that the significant reduction in claim frequency rates over the last two decades in Australia is likely to be attributable to incentive measures.

[1] Two exceptions are South Australia and Commonwealth Seacare.

2.4 Factors associated with prolonged claim duration

Initiatives to reduce claim durations can be more effectively applied if it can be identified in advance which claims are at high risk of becoming long duration and therefore high cost. Long-tail claims require not only greater direct costs including income maintenance payments and medical payments, but also indirect costs including re-staffing and re-training, potential loss of productivity, and legal costs related to negotiations for issues such as level of disability and causes of injury or illness (PMIS, 2017). Studies also suggest that the longer a worker takes time off work due to work-related injury, the less likely the worker will ultimately return to work (State Insurance Regulation Authority, n.d.).

Therefore, a crucial component of managing workers' compensation costs is identifying claims with a high risk of becoming long duration. For example, if claims made by female workers have a higher risk of becoming long duration as some studies (Cheadle et al., 1994; Yelin, 1986) have suggested, early intervention to claimants can be prioritised for females.

Claim duration may depend on many factors including individual level worker characteristics, type of injury, medical and vocational rehabilitation intervention, job characteristics, disability prevention and disability management intervention, social policy, and legislative and macro-economic factors (see Cancelliere et al., 2016; Krause, Frank, Dasinger, Sullivan and Sinclair, 2001; Lane et al., 2016). Various studies have attempted to identify the factors affecting claim durations for workers' compensation claims, or to measure the impact the factors. For example, Cheadle et al. (1994), from a population-based study of Washington State workers' compensation, found that gender, age, family status (e.g. marital status), presence of dependants, type of injury, whether a claimant is hospitalised or not, firm size, and unemployment rate, have a stable and significant effect on claim duration. MacKenzie et al. (1998) found that early return to work after sickness absence due to lower extremity fracture is correlated with younger age, higher level of education and income, higher level of social support, and having a white-collar job which is not physically demanding. Dumke (2017) found that headache intensity negatively affects return to work outcomes for patients with mild traumatic brain injury. Colantonio et al. (2016) conducted interviews with injured workers in Ontario, Canada and found that among the participants who successfully returned to work, the most common factors perceived to aid the return to work are support of family and friends, support of treatment providers, and employers who provided accommodations, while the most common obstacles are difficulty thinking and concentrating, and fatigue.

Factors found to be associated with claim duration can be used to develop a business model to improve outcomes. Lebedev et al. (2015) developed a triage model for workers'

compensation claims. They first identified 36 characteristics of workers' compensation claims at 13 weeks after injury that indicate an increased likelihood of claim duration exceeding one year. Among the identified characteristics, most influential were type of injury, age at injury, incapacity description from the last known medical certificate, and specialty of most frequented provider. Using these characteristics, the whole set of claims were segmented into high and low risk segments with the aid of data mining techniques. This facilitates prioritising initiatives for high-risk segments.

2.5 Work-related psychological injuries in Australia

Work-related psychological injury has become a major concern in Australian workplaces. The Australian Work Health and Safety Strategy 2012-2022 specifies such injuries as a priority (Safe Work Australia, 2015*b*). In the superannuation and group insurance industries, there are efforts under way to understand how to better identify and support the workers at higher risk of psychological problems (Rose, 2017). Psychological injury entails a range of cognitive, emotional and behavioural symptoms that can significantly interfere with how a worker feels, thinks, behaves and works together with others in his or her work and daily life (Safe Work Australia, 2014). Typical examples include depression, anxiety and post-traumatic stress disorder that are caused by work-related mental stress. Mental stress, which accounts for around 90 percent of psychological workers' compensation claims (Safe Work Australia, 2015*b*), is associated with job factors such as high demands, low job control, unclear work role, job insecurity, bullying, and poor communication (Beyondblue, n.d.). Although the mental stress is initially a state of mind and body rather than an illness, it may cause serious psychological illnesses like depression and anxiety or other physical illnesses if a worker is exposed to the stress for a prolonged period without a resolution (Safe Work Australia, 2013).

Statistics show which workers are more likely to have psychological work injuries. Safe Work Australia (2013), by exploring accepted workers' compensation claims, found that mental stress claims are predominantly made by women, and that professional workers make more claims for mental stress than workers in any other category, with over a third of their claims made for work pressure. The study also found that the likelihood that workers make a claim increases as they get older, but the opposite relationship holds for workers aged greater than 54. Safe Work Australia (2015*b*) observed that 39 percent of claims made for mental stress are caused by harassment, bullying or exposure to violence at the workplace. The report also showed that the occupations at higher risk of having psychological injury

are "Defence force members, fire fighters and police", "Automobile, bus and rail drivers", and "Health and welfare support workers" with frequency rates[2] of 5.8, 3.7, and 2.8 claims respectively, compared to the average for all claims for mental stress of 0.5 claims.

There are various public and private policies, procedures or initiatives that address work-related mental health issues. Benefits covered by workers' compensation include medical expenses involving doctors, psychiatrists, psychologists or counsellors (Safe Work Australia, 2014). Return to work programs, which aim at returning workers to safe and suitable work, contain provisions to ensure that psychological hazards at the workplace are appropriately considered. Individual companies also usually have policies or programs to support employees' mental health. For example, ANZ's critical incident recovery planning involves a procedure to resolve a psychological problem faced by an employee through defusing, debriefing and counselling (ANZ, 2017). The Heads Up initiative, developed by Mentally Healthy Workplace Alliance and beyondblue, provides workers and employers with free tools and resources to help them manage psychological issues in the workplace (Heads Up, n.d.).

2.6 Cost of work-related psychological injuries

Psychological injury is the most expensive form of individual workers' compensation claim due to its typically long duration, and has significant impact on injured workers and society (Safe Work Australia, 2013). Total claims payments are estimated to be $481 million per year, equating to around 11 percent of the payments for all claims (Safe Work Australia, 2015*b*). Work-related psychological injury accounts for six percent of all workers' compensation claims, amounting to 7,820 claims per year on average during the five year period to mid-2013. (Safe Work Australia, 2015*b*). The median claim payment for such injury during the same period was $23,600, which is around 2.7 times the global median of $8,700 (Safe Work Australia, 2015*b*, 2017). The difference is explained by the median time lost due to psychological injury of 14.8 weeks, which is 2.8 times the global median of 5.3 weeks. In 2007, the total cost to the Australian economy of work-related mental stress was estimated to be nearly $15 billion, the direct cost of which to employers in stress-related presenteeism and absenteeism[3] alone was more than $10 billion (Medibank Private, 2012).

[2] Number of claims per million hours worked

[3] "Presenteeism is defined as the lost productivity that occurs when employees come to work but, as a consequence of illness or other conditions, are not fully functioning. In comparison, absenteeism occurs when employees do not come to work."(Medibank Private, 2012)

The estimated costs may still substantially underestimate the actual size of the problem for several reasons. First, they make no allowance for the costs of re-staffing and re-training resulting from high staff turnover stemming from mental stress. Second, workers having mental stress are less likely to lodge a workers' compensation claim compared to other injured or ill workers particularly in lower socio-economic jobs (LaMontagne et al., 2010). An Australian Bureau of Statistics (2014) analysis shows that 70 percent of workers who experienced work-related mental stress did not make a claim. Third, mental stress is also known to contribute to development of other illnesses such as musculoskeletal disorder and cardiovascular disease (Safe Work Australia, 2013).

2.7 Studies on psychological factors and return to work

Numerous studies have investigated the impact of psychological factors on work-related injuries and illnesses due to all causes (see Cantley et al., 2015; Krause, Dasinger, Deegan, Rudolph and Brand, 2001; Lesuffleur et al., 2015; Lu et al., 2014). For example, Krause, Dasinger, Deegan, Rudolph and Brand (2001), from a study in California, US, examined the impact of psychosocial job factors on claim duration for workers with low-back injuries during the acute and sub-acute disability phases. They found that, while higher psychological job demands and lower supervisory support are associated with lower return-to-work rates during all disability phases, higher job control is associated with the rates only during the sub-acute disability phase. In an investigation of Korean workplaces, Lu et al. (2014) found an association between duration of leave due to work-related injury and psychosocial job factors. Factors found to be associated with longer duration from the two studies include verbal abuse, threat of violence at work, lack of predictability at work, psychological demands, low supervisory support, low reward, workplace bullying and low job control. There are also studies on psychological factors that are not directly related to the job such as Uehli et al. (2014)'s meta-analysis suggesting that workers with sleep disorders have higher risk of workplace injury.

Other studies have considered development of psychological symptoms following work-related injuries. For example, from a study in Ontario, Canada Carnide et al. (2016) investigated symptoms of depression following work-related musculoskeletal injuries requiring sick leave, and found that progression of the symptoms is associated with whether the claim duration goes longer than 12 months after injury.

Investigation of the factors associated with prolonged duration for psychological work-related claims is necessary because the factors are known to be different to those affecting

other injuries or illnesses. For example, from a literature review, Black et al. (2016) found that increased stress when dealing with a health care provider results in lower self-efficacy of workers regarding return-to-work - that is, a belief in their ability to return to work which is by itself an important predictor of early return-to-work (Krause, Frank, Dasinger, Sullivan and Sinclair, 2001) - for workers with psychological injuries, but not for those with upper-back musculoskeletal injuries. From a study in Netherlands, Nieuwenhuijsen et al. (2004) found that while better supervisory support is associated with earlier return to work for non-depressed employees, such association could not be established for employees with a high level of depressive symptoms.

However, there are only limited studies investigating the factors associated with longer duration for psychological work-related claims. One of the most relevant works is the study of Nielsen et al. (2012) in Denmark which found that employees working in the municipal and private sector return to work slower than employees working in the governmental sector after a sickness absence due to mental health problems. From a study in Denmark, Nielsen et al. (2011) found that, among employees sick listed with mental health problems, a shorter time to return to work is associated with a positive return to work expectancy and no prior absence with mental health problems. However, unlike our analysis, the mental health problems involved in these studies were not restricted to work-related problems.

Blank et al. (2008) reviewed the studies identifying potentially[4] significant risk factors for prolonged duration of a work-related mental health condition. Factors summarised in Blank et al. (2008)'s review can be categorised as job characteristics, health risk behaviours, individual level worker characteristics, and medical factors. Job characteristic factors negatively associated with return to work outcome include job stressors, threat of unemployment, and absence of workers' insurance. Health risk behaviour factors include being underweight or overweight, smoking, and dependence on drugs, while worker characteristic factors include being widowed, divorced or single, old age, and low education. Finally, medical factors include severity of symptoms, phobia type, and presence of minor psychiatric disorder.

However, Blank et al. (2008)'s meta-analysis is different to our analysis in several ways. First, fourteen of the fifteen papers in the meta-analysis did not find factors directly associated with duration of work-related mental health conditions. For example, Stansfeld et al. (1999), from a study in London, UK, found that lower work grade (e.g. high job demand) is associated with higher risk of psychiatric disorder, which may be different to the risk of prolonged claim duration. Second, while our study focuses on the factors observable

[4]The literature review of Blank et al. (2008) identified the risk factors from the 15 papers, not all of which examined the direct impact of such factors on return to work outcome.

by an insurer when a claim is made, the meta-analysis involved other factors as well. For example, Nieuwenhuijsen et al. (2004) found that better communication with supervisor and employer leads to a higher rate of full-time return to work only in non-depressed employees. Third, unlike the meta-analysis and those papers reviewed, our analysis applies data mining techniques. Finally, the meta-analysis did not develop a triage model based on the factors found.

2.8 Data mining in insurance and healthcare

Data mining is the process of extracting new and beneficial information from big datasets by using the techniques of statistics, data analysis, and machine learning (Tufféry, 2011). It is used for a large and noisy dataset, involving complex relationships between numerous fields, for which conventional statistical methods such as regression are usually not valid.

The application of data mining is useful in insurance because modelling often involves a dataset containing large number of cases as well as many variables (Kolyshkina et al., 2003, p.493). This is analogous to applications in the retail industry where data mining was initially used to develop marketing strategies from the large amount of sales data that became collectable and storable due to the growth of bar-code technology (Agrawal and Srikant, 1994).

There are numerous examples of the use of data mining techniques in insurance. Farmers Insurance Group, by virtue of the Underwriting Profitability Analysis developed by IBM, discovered that although drivers of high-performance sports cars are known to be more likely to get into an accident, such risk is actually not much greater than that of other drivers if a driver has another vehicle in the household (Apte et al., 2002). A study of Smith et al. (2000) presents two case studies in which data mining techniques (e.g. clustering, decision trees, and neural networks) enable better understandings of retention and claim patterns of insurance policyholders. The techniques were used to identify which policyholders are at greater risks of terminating their policies or making claims from available claim characteristics. Liu et al. (2014) examined multiple data mining techniques in predicting claim durations in Australian Income Protection Insurance using available rating factors, and suggested the technique to reduce a number of variables used in data mining processes with the aim of providing a more time-efficient objective method of classifying claims into different portfolios.

To identify the factors associated with prolonged claim duration and to develop a model to aid the management of workers' compensation claims, this study applies data mining techniques to a large Australian dataset containing numerous fields.

Chapter 3

Data

3.1 Overview of data

The data is provided by ReturnToWorkSA, the agency overseeing the workers' compensation scheme in South Australia. It contains 209,493 accepted workers' compensation claims filed for the eleven financial years from 2004/2005 to 2014/2015. Of the claims, 7,088 claims were primarily made for psychological injury (psychological claims), that is, the injury description specifies the bodily location of injury as "Psychological system".

The data contains 84 fields[1] for which the recorded values are observable at the time a claim is made, plus the ultimate duration of claim (claim duration) which is the response variable to be predicted for a new claim in this study. There are five types of fields as below.

- **Demographic information**

  The dataset contains demographic information of injured workers. The information for all but one of the fields are obtained during the notification call to make a claim. They are the fields for age, gender, residential postcode, occupation, and income estimate of an injured worker, and the field to indicate if the worker requires an interpreter. By using Australian and New Zealand Standard Classification of Occupations[2] (ANZSCO), the occupation is recorded with three fields; description (e.g. baker), minor category (e.g. food trades workers) and major category (e.g. technicians and trades workers). The final field shows the number of employments experienced by an injured worker

[1] The fields are listed in Appendix A

[2] "ANZSCO is primarily a statistical classification designed to aggregate and organise data collected about jobs or individuals. The classification definitions are based on the skill level and specialisation usually necessary to perform the tasks of the specific occupation, or of most occupations in the group."(Australian Bureau Of Statistics, 2009)

in the period of five years to the time a claim is made, which is obtained from the database of ReturnToWorkSA.

- **Description of injury or disease**

  Information about injuries or diseases for which claims are made is also obtained during notification calls. These include fields for date and description of injuries and diseases, lag between date of injury and reporting date, and if an ambulance is called. The descriptions are recorded by using Type of Occurrence Classification System (TOOCS) 3.1 (Safe Work Australia, 2008). TOOCS is the coding system developed to record details of workers' compensation cases reported to workers' compensation agencies (Safe Work Australia, 2008). The system specifies injuries or diseases according to four different aspects, *Nature* (such as fractures or traumas); *Bodily location* (such as back or circulatory system); *Mechanism of incident* (such as falls from a height or being hit by an animal); and *Agency* (such as saws or boilers). Each aspect of injury or disease is recorded in the dataset with multiple fields corresponding to different levels of detail. For example, *Agency of injury* is recorded in three fields; description (e.g. manual fire tube boiler), minor category (e.g. boilers) and major category (e.g. heating, cooking, baking equipment).

- **Past claim history**

  There are the fields describing the past claim history of an injured worker if it exists both for all prior claims and for the most recent prior claim. Fields for all prior claims, include: number of claims made by the injured worker; how many of them are currently open; total lump sum payments, legal costs and income maintenance payments for all prior claims; and if any prior claim involves a payment for psychological services. Fields for the most recent prior claim include: past payments for the claim such as hospital payments, income maintenance payments, legal payments and various other payments; indications for having the injury or disease for the current claim related or connected to the most recent prior claim; descriptions using TOOCS 3.1 of the injury or disease; count of visits to medical professionals; costs of drug and potent opioids; and duration of income maintenance payments.

- **Employer-related information**

  There are three fields related to employers of injured workers: size of employer, industry classification of employer's business and premium rate of workers' compensation insurance for employers. The premium rates are calculated based on South Australian

Industry Classifications (SAIC). The rate is linked to actual claims experience of all businesses classified in a particular SAIC group (Return To Work South Australia, n.d.).

- **Duration of claim**

  The response variable, claim duration, is coded as a binary variable for whether a claim has a duration of income maintenance payments longer than two weeks (prolonged duration): this is discussed in Section 4.2.1. In the dataset for all claims, 24% have a duration greater than two weeks, or a value of 1 for the response variable.

There are some missing values for several fields in the dataset. Those fields are: descriptions of injury or disease for the current claim (around 4,400 or 2.1%) and the most recent prior claim (around 5,900 or 2.8%); demographic information including income estimate (96,178 or 45.9%), occupation (4,397 or 2.1%), residential postcode (2,153 or 1.0%) of injured worker; and premium rate for employer (882 or 0.4%)[3]. A data point with a missing value for a certain field does not always have a missing value for other fields. We have treated missing values differently for those fields. For the descriptions of injury or disease, we have taken the missing values as a special value, to investigate the effect of not having descriptions. *Income estimate for injured worker* is considered to be biased and excluded from the analysis because there is a tendency that an income estimate is not available when a claim does not involve income maintenance payments. For the other fields, missing values are handled by analysis techniques, as discussed in Section 4.2.6.

3.2 Exploration of data

3.2.1 Gender, age, occupation and mechanism of injury

We have observed the dataset by gender, age, occupation of injured workers, and mechanism of injury, to find any patterns as well as to make a comparison with the national dataset on which the findings by Safe Work Australia (2013) (see Section 2.5) are made. As shown below, they are similar overall except the finding of Safe Work Australia (2013) that professional workers made more psychological claims[4] than workers in any other category.

[3] The figures in parentheses are the numbers of missing values. Those for descriptions of injury or disease are not exact figures because the multiple fields for the descriptions have similar but different numbers of missing values.

[4] Although Safe Work Australia (2013)'s definition of psychological claim, a claim made for mental stress, is different to our definition, we find that almost all psychological claims in this study are also made for mental stress.

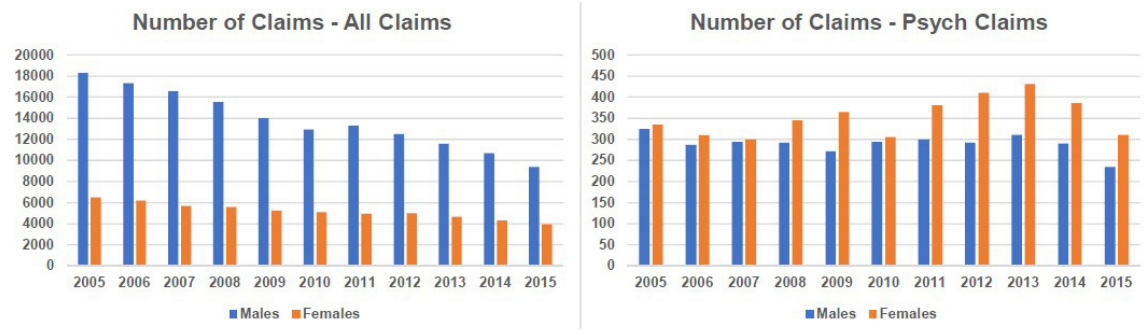

Fig. 3.1 Number of claims by gender

As seen in Figure 3.1, while workers' compensation claims for all causes combined are predominantly made by men, more psychological claims are made by women in all years. We also see that, although the number of the overall claims has decreased during the eleven-year period to 2015, such trend is not seen for the psychological claims.

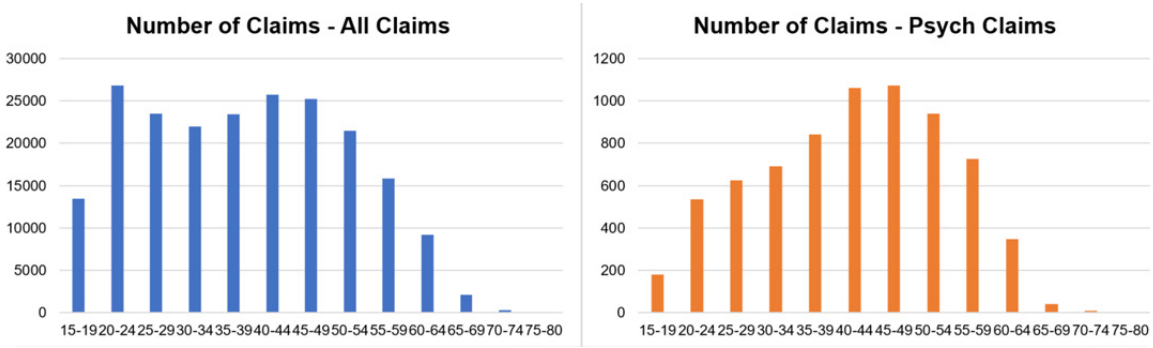

Fig. 3.2 Number of claims by age of injured Worker

Figure 3.2 shows the age patterns of the numbers of workers' compensation claims. There are peaks for the numbers of claims at the age groups, 40 to 49, followed by a decrease of the numbers for the older age groups, for both the overall and the psychological claims. However, while the numbers of claims increase consistently by age for the younger age groups for the psychological claims, there is another peak at the group, 20 to 24, for the overall claims.

From Figure 3.3 showing the numbers of workers' compensation claims made by workers in different occupation groups, we see that the patterns are quite different for overall claims and psychological claims. While Technicians and Trades Workers made the most claims, they form the group which made the least psychological claims. While the numbers of claims vary significantly by occupation group for overall claims, they are similar across occupation group for psychological claims except for the distinctly higher number of claims made by Community and Personal Service Workers.

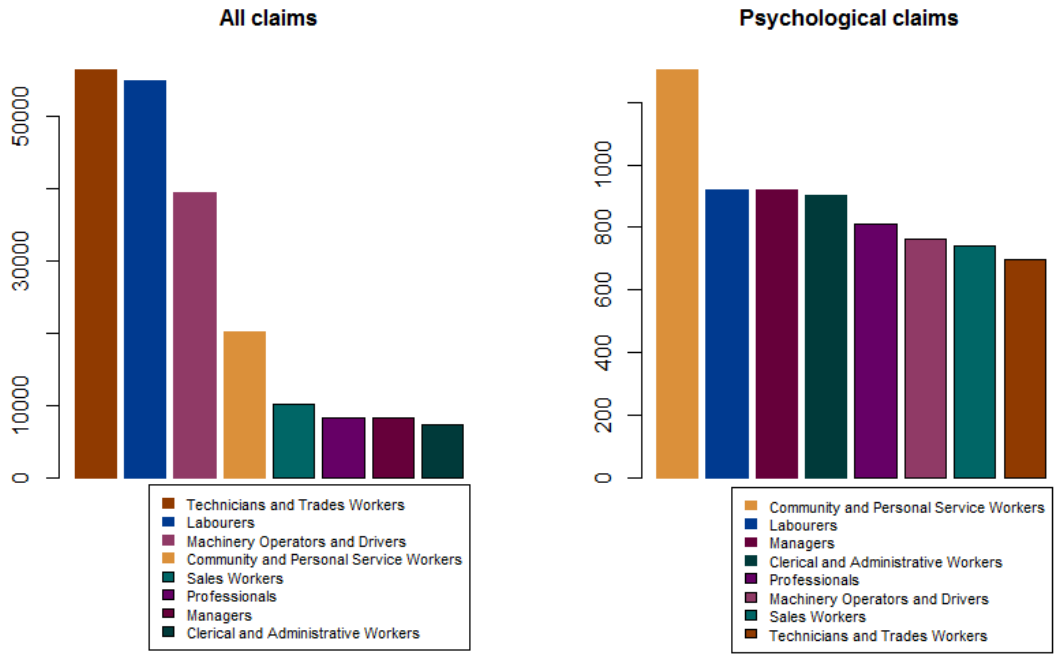

Fig. 3.3 Number of claims by occupation category

As at November 2016, the proportions of workers in the occupation groups, Managers, Professionals, Technicians and Trades Workers, Community and Personal Service Workers, Clerical and Administrative Workers, Sales Workers, Labourers are 0.129, 0.203, 0.144, 0.121, 0.127, 0.159 and 0.117, respectively (Australian Bureau Of Statistics, 2017).

Figure 3.4 shows that the mechanism of injury leading to a psychological claim is clearly different to overall claims. For psychological claims, the most claims are made due to work pressure, followed by work-related harassment and/or workplace bullying, and exposure to workplace or occupational violence.

3.2.2 Correlations between predictors

We have examined correlations between the predictor fields of the dataset. This observation is done because the extent of correlation between predictors will inform the choice of analysis techniques. We have found numerous correlations among the predictors. For example, as seen in Figure 3.5, the bodily location of injury for a claim is not independent of the bodily location of injury for the most recent prior claim made by the injured worker. The figure shows deviation of the proportion of bodily location of injury given the bodily location of injury of the most recent prior claim from the unconditional proportion of the bodily location of injury. We can see, from the diagonal line inside the table made by the colour coding, that the injury for a claim is more likely to be on a certain bodily location when there was

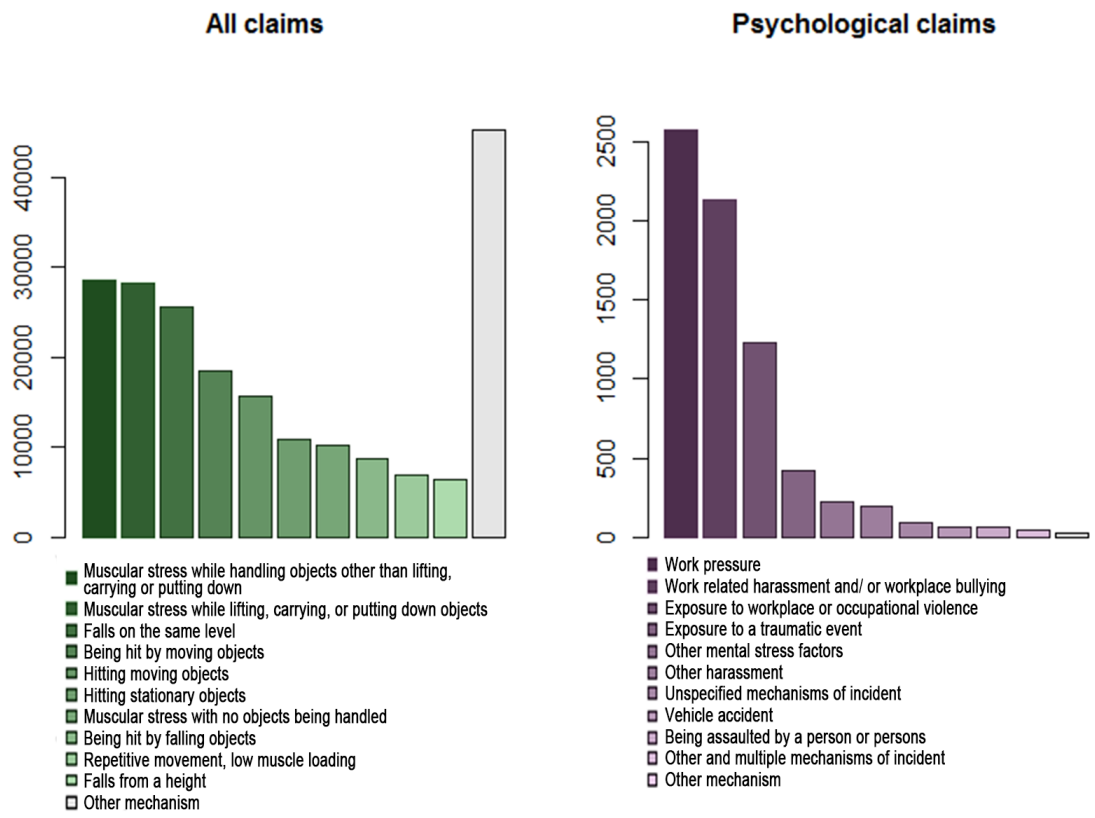

Fig. 3.4 Number of claims by mechanism of injury

a prior claim for an injury on the same bodily location, than otherwise. In Figure 3.4, we have also seen that the mechanism of injury is correlated with whether a claim is made due to psychological injury.

To further examine the relationships between predictor fields, we conduct a clustering of variables analysis which forms groups, or clusters, of correlated variables according to a homogeneity criterion based on the squared correlations for numerical variables, and sum of correlation ratios for categorical variables (Chavent et al., 2011). It generates the dendrogram in Figure 3.6 to show how we may cluster the fields closely related to one another, at some different level of grouping. For example, at the level of height of 2, by drawing a vertical line at height=2, we see that we may have four clusters of fields. We find that many relationships among fields are as expected. For example, we would expect that gender of a worker is related to industry classification and occupation, because males are more likely to have certain occupations and work in certain industries (see the top part of the dendrogram). In particular, the dendrogram shows that all payments for the most recent prior claim, such as medical payments and payment for physiotherapy treatment, are closely related. This

Bodily location of injury	Proportion	Bodily location of injury of most recent prior claim									
		HEAD	NECK	TRUNK	UPPER LIMBS	LOWER LIMBS	MULTIPLE LOCATIONS	SYSTEMIC LOCATIONS	NON-PHYSICAL LOCATIONS	UNSPECIFIED LOCATIONS	New Claim
HEAD	0.116	0.136	-0.027	-0.029	-0.005	-0.012	-0.030	-0.007	-0.042	-0.057	-0.007
NECK	0.024	-0.004	0.087	0.001	-0.003	-0.003	0.018	0.002	0.002	-0.007	-0.002
TRUNK	0.215	-0.042	-0.005	0.135	-0.023	-0.012	-0.006	-0.020	-0.029	0.046	-0.021
UPPER LIMBS	0.379	-0.035	-0.065	-0.080	0.063	-0.057	-0.055	-0.055	-0.098	0.041	0.025
LOWER LIMBS	0.173	-0.023	-0.022	-0.019	-0.013	0.096	-0.006	-0.011	-0.022	-0.030	-0.003
MULTIPLE LOCATIONS	0.048	-0.014	0.020	-0.003	-0.007	-0.005	0.063	0.000	0.018	0.011	0.003
SYSTEMIC LOCATIONS	0.011	0.000	0.002	-0.002	-0.001	-0.001	-0.002	0.074	0.003	-0.002	0.001
NON-PHYSICAL LOCATIONS	0.035	-0.018	0.011	-0.003	-0.010	-0.006	0.018	0.016	0.168	-0.009	0.005
UNSPECIFIED LOCATIONS	0.001	0.000	-0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.008	0.000

Fig. 3.5 Deviation of proportion of bodily location of injury given bodily location of injury of the most recent prior claim from unconditional proportion of bodily location of injury

For example, the table shows 0.136 for the difference between (1) the proportion of claims made due to an injury on head by those who made another prior claim due to an injury on head, 0.256, and (2) the proportion of claims made due to an injury on head among all claims, 0.116.

finding allows us to run the analysis by using the new predictor combining all those payments together instead of individually using those fields[5].

3.2.3 Predictability of claim duration

By observing the predictor fields with respect to the claim duration, we find that durations are highly unpredictable. Of those predictor fields, age of injured worker, nature of injury, premium rate for workers' compensation insurance charged to the employer of an injured worker, and size category of the employer are presented through several plots in this section.

Figure 3.7 shows scatter plots of claim duration by age of injured worker for several natures of injury. We can observe several features. First, although the lines fitted indicate that claim durations generally increase with age, the high level of variability makes this trend difficult to discern. Second, although overall claim durations seem different by nature of injury (e.g. duration of claims made for mental disorders and musculoskeletal injuries are higher than other claims), claim durations have high deviations within each nature of injury.

Figure 3.8 shows scatter plots of claim duration by the premium rate for workers' compensation insurance charged to the employer of an injured worker and the size category of the employer. There is no clear patterns for the premium rate with respect to the claim duration. We see that the claim duration is, on average, slightly higher for injured workers who work

[5]We find that such reduction in the number of predictors is useful in enhancing the robustness of the final model, with no deterioration in the performance of the model.

for small employers than those for medium or large employers. However, the durations show high deviations within each size category.

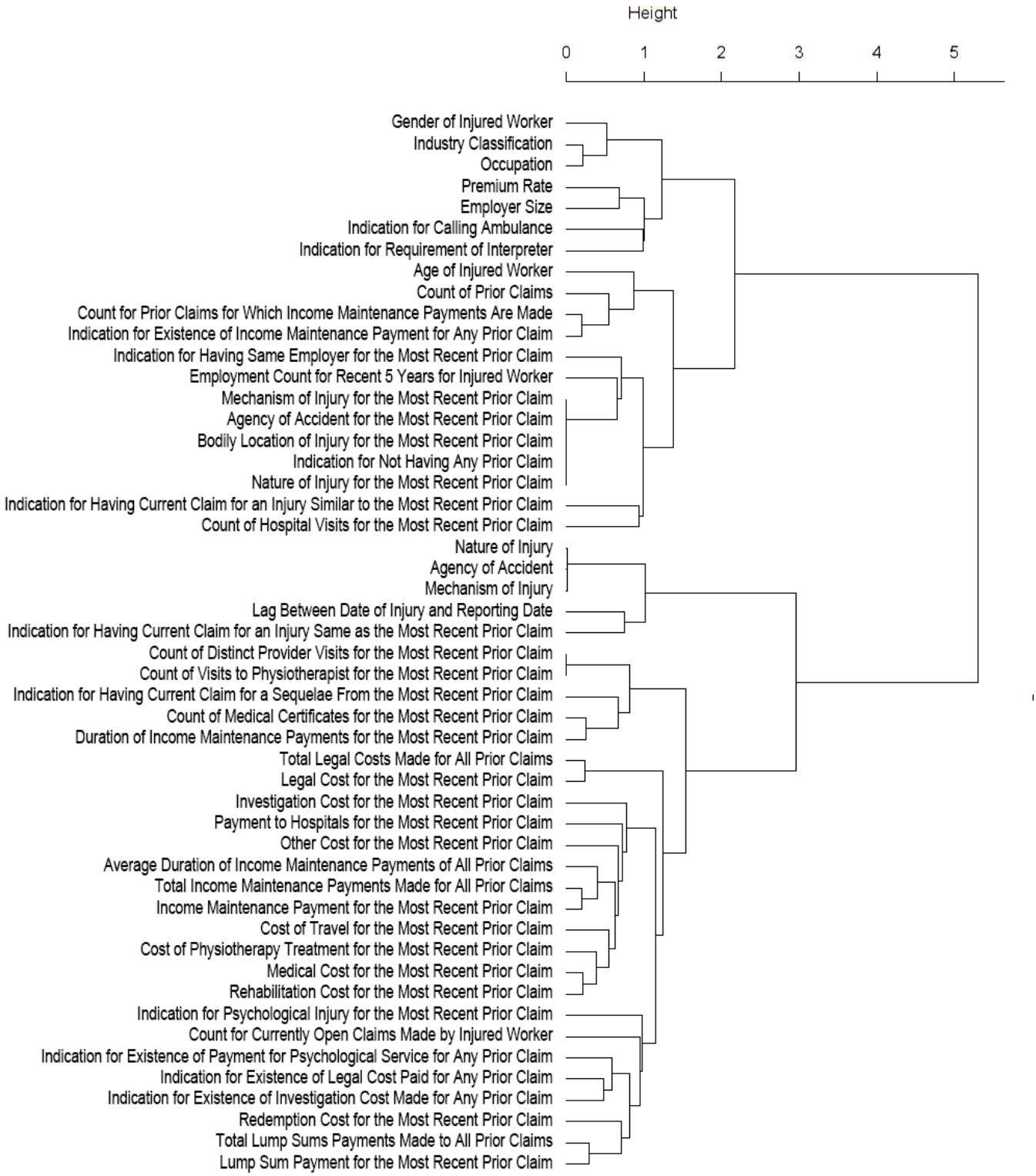

Fig. 3.6 Cluster dendrogram

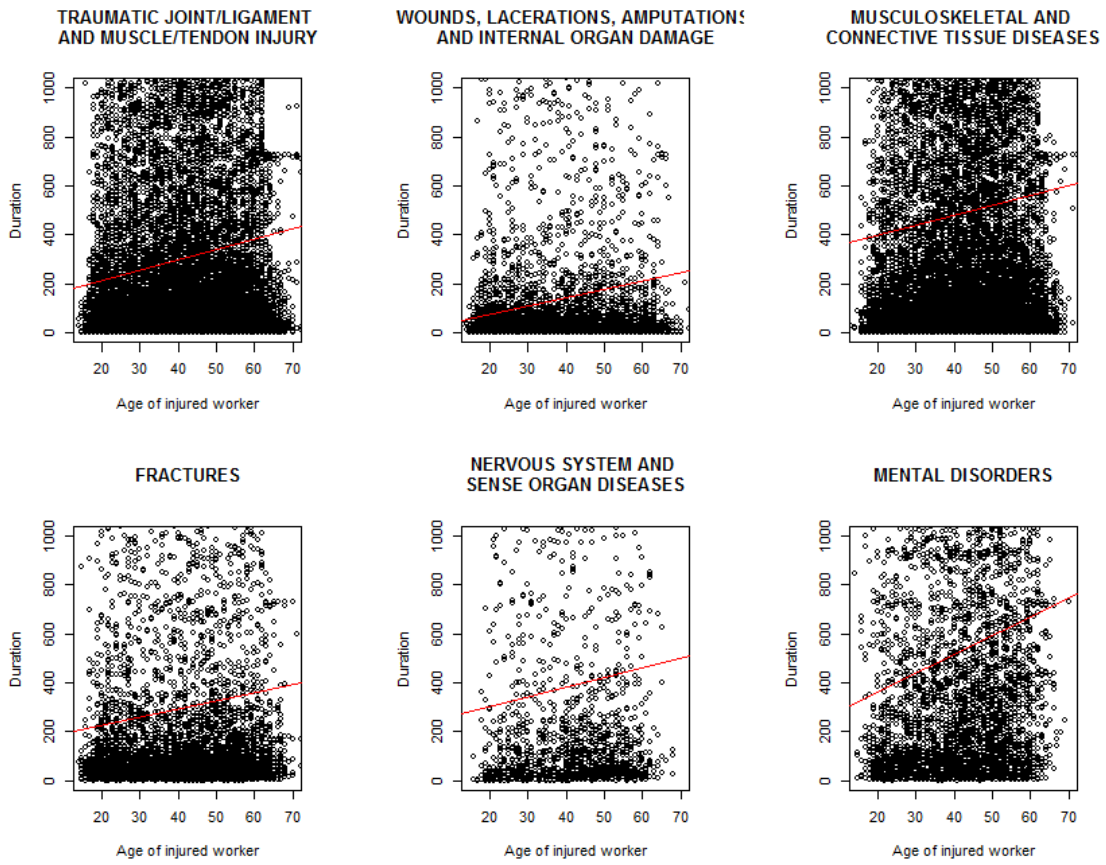

Fig. 3.7 Claim duration vs age by nature of injury

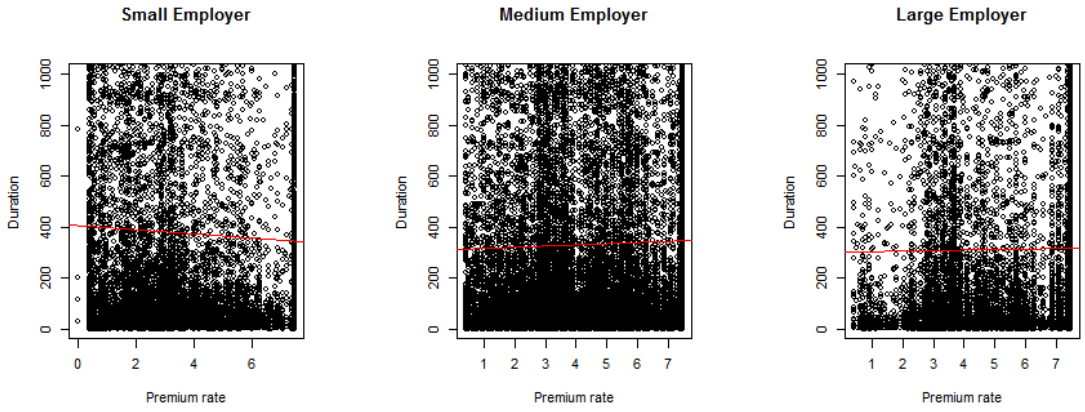

Fig. 3.8 Claim duration vs premium rate by size of employer

Chapter 4

Method

4.1 Research design

This study develops a triage model for workers' compensation claims with respect to a risk of prolonged claim duration. By first identifying predictors, or characteristics of claims, associated with claim duration, the triage model segments the whole set of claims into high and low risk segments. Two models are developed, on overall claims and on psychological claims only, and compared to each other. Therefore, all procedures stated below have been done for the two sets of claims separately.

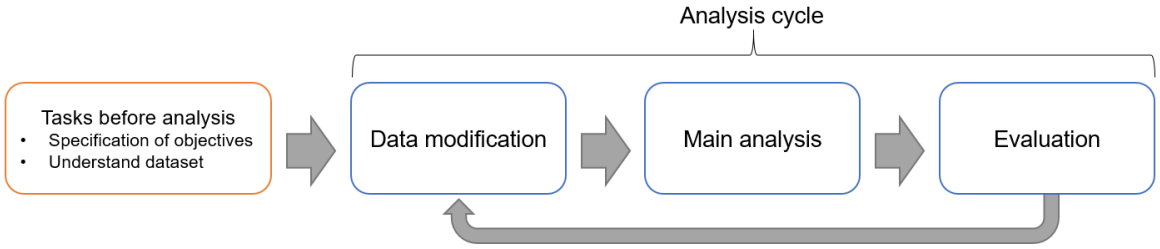

Fig. 4.1 Key stages in the research design

There are two tasks to be done before the analysis. The first task is a specification of overarching objectives. By investigating what is needed for management of workers' compensation claims, we have set the objective of developing a practical and informative model, which can be used and understood by claims managers, as follows.

> *The model should specify the claims with higher risk of having a prolonged duration based on information available at the time a claim is made, by using a set of easy rules not involving actuarial or other technical concepts.*

The second task is to understand all variables (the predictors and the response) in the data through data exploration and verification as discussed in Section 3. The analysis cycle then involves the following stages.

1. **Data modification**

   An original variable in the dataset can be transformed into another form more suitable for modelling (e.g. converting a numerical variable into a categorical variable, creating a new variable combining two or more existing variables, or reducing a number of categories of a categorical variable). A variable may be excluded when considered irrelevant due to little correlation with the response or unreliable due to an excessive number of missing values. A data point may be excluded when considered unreliable due to a missing or illogical value for key variable fields.

   This is included in the analysis cycle to optimise the final model by trying different forms and selection of predictors. We find that modelling via tree-based methods is quite sensitive to which predictors are used, and hence it may be necessary to try more than one set of predictors to improve the model performance, given that such modifications are practical in using the final triage model in practice.

2. **Main analysis**

   Main analysis is done on a training dataset which is a subset of the entire dataset. It involves developing multiple triage models by trying different approaches in terms of the tree-based algorithm and its parameters.

3. **Evaluation**

   Evaluations are done as models are applied to the test dataset.

The above three stages may be repeated to optimise the final triage model.

4.2 Triage model

4.2.1 The binary response

In this analysis, the dependent variable is the binary response, “1” for claims with a duration of income maintenance payments greater than two weeks; and “0” for other claims. Therefore, the analysis involves the prediction on whether a duration is greater than two weeks (duration

outcome). The duration is based on the income maintenance payments, to address the amount of labour time lost due to an injury or disease, which is compensated by the workers' compensation scheme.

Two weeks is chosen as the duration of interest, because a duration shorter than two weeks is considered to be a short-term claim which does not significantly affect the employer and the injured worker. Therefore, results will permit exclusion of low-risk claims. For predictions of whether a duration exceeds a time period significantly greater than two weeks, an analysis on claims already in progress, rather than new claims, involving more information obtained from the progress of such claims, would be more appropriate. For example, Lebedev et al. (2015) predicts whether a claim duration exceeds one year based on the characteristics of workers' compensation claims at 13 weeks after injury.

Several advantages exist for an analysis to predict a duration outcome for new claims. First, it enables early intervention which has been found to be helpful as shown in Section 2.3. Second, compared with an analysis requiring a dataset of claims with a duration at least for a certain length of time, the analysis can utilise a much bigger dataset including short-term claims, and thus provide more meaningful results.

A binary response allows efficient applications of some data mining techniques including tree-based methods in this study. We find that computation processes of the tree-based methods do not always converge for other types of response (continuous response or categorical response with more than two categories) on our dataset. Binary response values measure the risk for claims in each segment comprising the triage model. We can intuitively measure such risk by estimating a probability that a claim duration exceeds two weeks. Such probability would be the proportion of claims with the binary response value of 1.

4.2.2 The classification problem

The dependent variable in this study is a (binary) categorical response, so this study is referred to as a classification problem. By contrast, a numerical response would be referred to as a prediction problem.

A classification in statistics is a method to assign an observation to a category, or class, and is primarily used in predicting a categorical response. Such method often estimates the probabilities that an observation lies in each category as a basis for making a prediction.

This study is to develop a triage model for claims showing probabilities, or risks, of prolonged claim duration in each segment, rather than to predict an individual duration outcome for a claim. Therefore, we evaluate the models mainly for how accurate the

probabilities are, rather than how accurately a duration outcome is predicted for an individual claim based on such probabilities.

4.2.3 Classifiers

There are many classification techniques, or classifiers, that can be used to predict a categorical response. Three of the conventional classifiers are logistic regression, linear discriminant analysis and K-nearest neighbours (KNN) (James et al., 2014, p.127).

Logistic regression uses a logistic function to directly model a conditional distribution of a categorical response across the categories, given the predictor values (James et al., 2014, p.138).

In linear discriminant analysis, the distributions of the predictors are separately modelled in each of the response categories given the response value, and then Bayes' theorem is used to achieve the conditional distribution of the categorical response. This approach is known to generate more stable models when the categories are well-separated, or when sample size is relatively small and the distributions of the predictors are approximately normal in each of the response categories (James et al., 2014, p.138).

The K-nearest neighbours (KNN) is a non-parametric classifier that is developed by an entirely different approach from the above two classifiers. When predicting a response of an observation with a chosen positive integer K, the KNN classifier identifies the K neighbours in the dataset that are the closest to the observation, and then classifies the observation by a majority vote of its neighbours (James et al., 2014, p.151).

In this study, tree-based methods, another widely-used classifier are mainly used. The methods which involve creating a tree or trees fitted to a dataset, are non-parametric, and hence not based on any assumption on the shape of a decision boundary.

4.2.4 Classification tree and its application for the study

Classification tree is one of the tree-based methods to predict a categorical response. It segments observations according to a set of rules on predictor values (i.e. segments a multidimensional predictor space) into multiple mutually exclusive regions, so that observations classified to the same region are predicted to have the same value for a categorical response. A tree is formed by obtaining information from a training dataset, and evaluated by being applied to a test dataset. When making a prediction, a test observation is assigned to the most frequently occurring response category of the training observations in the region to which the

test observation belongs. The name is given to the method as the set of splitting rules used to segment the predictor space can be graphically displayed as a tree (James et al., 2014, p.311).

An application of a classification tree can be illustrated by an example which is based on a small subset of this research problem. Suppose we predict whether a claim has a duration longer than two weeks (a prolonged claim duration) based on the two predictors, *Bodily location of injury* and *Age of injured worker*.

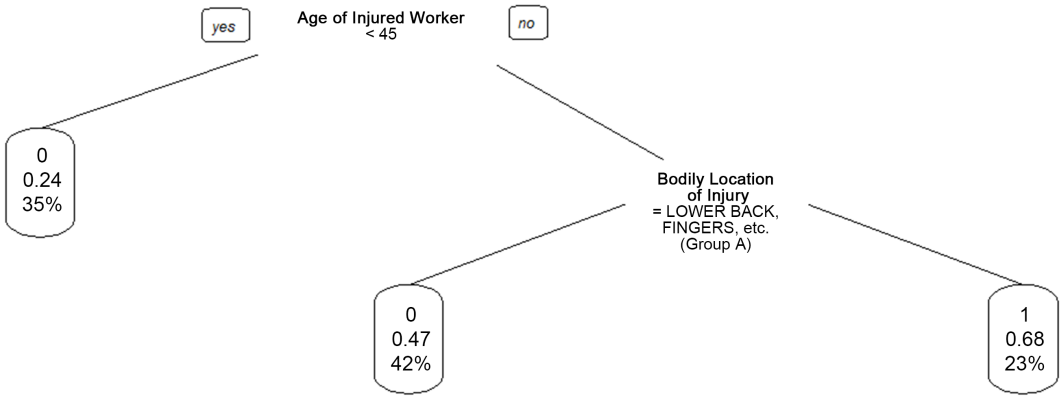

Fig. 4.2 Illustrative example of classification tree

This tree is for the illustrative purpose only. The first, second, and last figure in each region of the tree means the predicted value of response, the probability of prolonged duration, and the proportion of claims lying in the region, respectively.

Figure 4.2 shows a classification tree fitted to a training dataset (how the tree is fitted is discussed below). It is comprised of an ordered set of splitting rules, from the top of the tree.

The first split at the top of the tree assigns the claims having *Age of injured worker* less than 45 to the left branch. The probability of prolonged duration for a new claim made by an injured worker younger than 45 is given by the proportion of the claims having a prolonged duration among the claims made by the workers younger than 45 in the training dataset. As the proportion, or probability for a prolonged duration, is less than 50%, a new claim in this group is predicted not to have a prolonged duration.

Claims having *Age of injured worker* greater or equal to 45 are assigned to the right branch, and then that group is further split by *Bodily location of injury*.

As a result, the tree segments claims into three regions of predictor space: claims having *Age of injured worker* less than 45; claims having *Age of injured worker* greater or equal to 45 and *Bodily location of injury* in "Group A"; and claims having *Age of injured worker* greater or equal to 45 and *Bodily location of injury* not in "Group A". The probabilities of prolonged claim duration are 24%, 47% and 68%, and the sizes of regions are 35%, 42% and 23% of the total number of claims respectively.

In this study, regions of a tree will become each segment of a triage model.

4.2.5 The algorithm for classification trees and statistics packages used

The classification tree algorithm is "a top-down, greedy approach that is known as recursive binary splitting" (James et al., 2014, p.306). The algorithm is top-down because it starts from the top of the tree at which all training observations are in a single region, and then sequentially splits the predictor space further down the tree. Each split adds one to the total number of regions. The algorithm is "greedy because at each step of the tree-building process, the best split is made at that particular step, rather than looking ahead and picking a split that will lead to a better tree in some future step" (James et al., 2014, p.306).

To make each split, a predictor and a cut point for the predictor value are chosen. In the Figure 4.2, *Age of injured worker* and age of 45 are chosen as the predictor and the cut point respectively. In making this choice, the algorithm considers all predictors available in the dataset on which the split is to be made, and all possible values of the cut point for each of the predictors. Then it selects a predictor and a cut point such that the split results in the greatest possible reduction of the prediction error for the respective region of the tree. The prediction error after the split is the sum of the prediction errors of the resultant two regions adjusted for the size of the regions. For each region, all observations in a region are predicted to have same value of response, and the prediction error measures to what extent the training observations do not have the response predicted by the tree.

Different measures of prediction error may be used. The most general measure for any classification problem is the classification error rate which is the proportion of observations with a response value different to the predicted response value (James et al., 2014, p.312). Another widely-used measure for tree-based methods is the Gini impurity, a measure of total variance across all classes (categories) of response (defined with the formula by James et al., 2014, p.312) in each region of a tree. For each region, it sums the error rates adjusted for the proportion of the respective class, across all classes of response. For example, the classification error rate and the Gini impurity for the left-most region in Figure 4.2 are as follows.

$$\text{Classification error rate} = 1 - \max(0.24, 1 - 0.24) = 0.24$$

$$\text{Gini impurity} = 0.24 \times (1 - 0.24) + (1 - 0.24) \times (1 - (1 - 0.24)) = 0.3648$$

While the two measures both get smaller when a proportion of observations from one class gets closer to 1, the Gini impurity is more sensitive to proportions close to 1, than the classification error rate. The Gini impurity is therefore more appropriate for our analysis identifying high-risk claims with a probability of prolonged duration close to 1 as much as possible. In addition, Kolyshkina et al. (2003, p.508) finds that the model based on Gini impurity performs slightly better in terms of prediction accuracy and identification of high and low-risk claims than the model based on Twoing criterion which is another measure of prediction error. Unlike the Gini impurity used for a maximisation of class heterogeneity, the Twoing criterion selects the split that maximises the separation between classes (Kolyshkina et al., 2003, p.507).

The Gini indices and the classification error rates in several scenarios in the case of binary response are shown in Table 4.1. We can see that the Gini impurity is reduced at higher rate when a proportion of one class is getting closer to 1, than the classification error rate.

Table 4.1 Gini impurity for a region of a tree

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9
Proportion of data points with a response of 0 (a)	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Proportion of data points with a response of 1 (b)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Predicted value of response	0	0	0	0	Either 0 or 1	1	1	1	1
Classification error rate	0.1	0.2	0.3	0.4	0.5	0.4	0.3	0.2	0.1
Gini impurity	0.18	0.32	0.42	0.48	0.5	0.48	0.42	0.32	0.18

Gini impurity $= a \times (1-a) + b \times (1-b)$
Classification error rate $= 1 - \max(a,b)$

A tree is grown until it reaches the minimum possible total prediction error subject to any constraints or "stopping rules" applied such as minimum size of group or tree depth. However, the resultant tree typically overfits a training dataset. The algorithm measures how much the tree overfits by estimating the test error rate for predictions on subsets of the training dataset. It repeats the process for reduced trees with less number of splits. By using the information obtained from such a process, we can discard a part of splits made for the tree, or “prune” the tree. In this study, decisions about the level of pruning are made not only depending on the test error rate, but also on the numbers of categories in the final classification model that are usable in practice[1].

[1] A tree model that produces only a small number of regions will not specify potential smaller regions with a very high or low probability of having a prolonged duration, of which the specifications are desirable. A tree model that produces a high number of regions will not be easy to interpret.

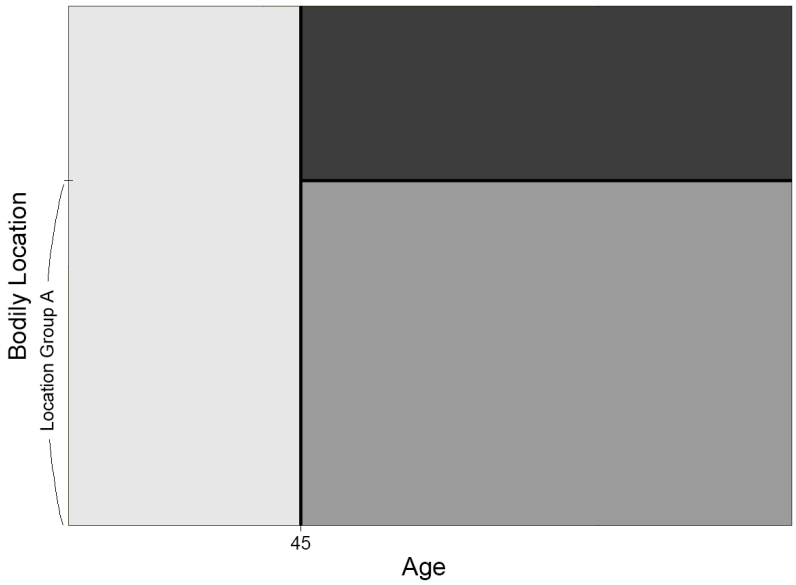

Fig. 4.3 An illustrative example of classification on a complex and non-linear decision boundary

For a robustness test, we have fitted conditional inference trees, another widely used tree-based method, and compared the results with those for the classification trees. While the conditional inference tree also segments observations to make predictions on the categorical response, it is different to the classification tree in several ways. First, to make each split, it selects a predictor that has the strongest association to the response which is measured by a p-value estimated from a test for the partial null hypothesis of a single input variable and the response (Hothorn et al., 2006). Second, it does not overfit trees and hence does not need pruning (Levshina, 2015, p.292).

We use the rpart algorithm in the R package of the same name (see Therneau et al., 2017), to implement the classification tree approach , and the ctree algorithm in R's Party package (see Hothorn et al., 2006) to implement the conditional inference tree approach.

4.2.6 Advantages and disadvantages of tree-based methods

The advantages of tree-based methods in this scenario include the following.

- A tree can be used to form rules to group observations so that each group has a distinctly different probability to fall into each category of a response. This is because the trees use such algorithms of grouping observations to classify observations. This property is used in this study to create the triage models for claims.

- When there is a complex and non-linear association between predictors and a response as indicated by model, then the tree-based methods may be fitted to a decision boundary better than the classical linear approaches such as logistic regression and linear discriminant analysis (James et al., 2014, p.315). An illustrative two-dimensional classification example is displayed in Figure 4.3. The true decision boundary indicated by the shaded regions, which is non-linear and non-continuous, can be intuitively captured by a tree-based method that involves segmenting the predictor space (the set of possible predictor values). However, linear models are unable to capture the true decision boundary without using a piece-wise function or dummy variables. In this study, we could not make any assumption on the shape of the decision boundary which involves many predictors as shown in Section 3.2.2. Therefore, this advantage is valid for the study.

- Tree-based methods involve selecting important predictors in a way specified by each algorithm. This feature is essential for our dataset which contains many predictors of which all are not necessarily important or useful together. Selecting the most important predictors from from numerous predictors is often not feasible using conventional linear methods such as Generalised Linear Models (Kolyshkina et al., 2003, p.493). This is also a comparative advantage over the KNN which does not include an output showing which predictors are important (James et al., 2014, p.151).

- Tree-based methods easily handle correlated predictors whereas classical methods applied to large datasets can have difficulty selecting important interactions between predictors (Kolyshkina et al., 2003, p.496). This property is essential for our dataset which contains numerous correlated fields.

- Tree-based methods have the ability to develop a model in the presence of missing values without discarding any useful information in the data (Kolyshkina et al., 2003, p.494). Missing predictor values in a dataset are efficiently treated by some tree-based methods. An example is the classification tree algorithm used in this study, which selects multiple predictors one at a time, to develop a tree model. It does not initially remove all data points having any missing predictor values. Instead, when a predictor is selected for a test, it omits data points with no value for the selected predictor. A developed model also specifies a means to predict for a test observation with missing values. This is a useful feature for this study as the dataset includes missing values.

- Trees created by the methods, when displayed graphically, are easy to understand even by a non-expert. (James et al., 2014, p.315).
- Qualitative predictors are easily handled by the methods without creating dummy variables (James et al., 2014, p.315).

The disadvantages of the tree-based methods in this scenario include the following.

- Trees typically have less predictive accuracy than other conventional regression and classification approaches (James et al., 2014, p.316). Although this is a clear limitation of the analysis, other approaches are not likely to be appropriate for this analysis on the dataset containing a large number of predictors. Therefore, this does not necessarily mean that the prediction accuracy can be improved otherwise. In addition, this problem is less crucial in this study because we focus on developing a triage model showing important predictors of prolonged claim duration, rather than making accurate predictions on individual claims.
- Trees may lack robustness. That is, a small change in the training dataset (or even having a new randomly sampled training dataset) can cause a significant change in the final tree (James et al., 2014, p.316). To address this, we have also used a tree ensemble method which is explained in the next section.

4.2.7 Tree ensemble methods and Random Forests

Tree ensemble methods involve building multiple trees which are then used together to make a single consensus prediction. Using plenty of trees may often help us to achieve a dramatic improvement in prediction accuracy, with some loss of interpretability (James et al., 2014, p.303).

The lower level of interpretability, or the inability to reduce multiple trees down to a single readable classification tree, means the tree ensemble method is unfortunately not appropriate to make a triage model required in this study. Hence, the improvements in prediction accuracy to be achieved by the method is not applicable to our final model.

However, we can still use the tree ensemble methods for a development of our model. First, we can take the predictors found to be important by the tree ensemble method as a reference for the predictors used in a classification tree model. In an analysis involving an excessive number of predictors, the method can also be used to measure importance of predictors and hence reduce the number of predictors to be used for the analysis (Ziegler

and König, 2014). Second, we can have an idea how close the accuracy measure of our final model is to the maximum possible accuracy approximated by the tree ensemble method.

The tree ensemble method used for the study is Random Forests, the most widely used tree ensemble method (see James et al., 2014, p.319).

Random Forests adds two layers of randomness in building a number of trees. First, it fits trees to different training datasets that are bootstrapped samples from the dataset. Second, in building these trees, when the process considers each split, it selects a predictor from a random sample of m predictors chosen as candidates from the all available p predictors. This randomises the process further because it decorrelates the trees by forcing them to select a different set of predictors while strong predictors will still be selected more often than other predictors in general. This is particularly helpful in improving the prediction accuracy as it is then possible to select a predictor that is unlikely to be chosen otherwise (because a split made with the predictor is not the best one at that particular step) but, if chosen, actually leads to a better tree in some future step. As tree algorithms are greedy, they are not able to detect such predictors.

We use the randomForest algorithm in the R package of the same name (Liaw and Wiener, 2002), to implement the Random Forests approach.

4.2.8 Method to refine the model: Association rule learning

After applying a classification tree, it can be found that there are large regions remaining after the final split is made. Such regions have a number of data points sufficient to be analysed again by using another method. To further refine those large regions, association rule learning is used for this study.

Association rule learning is a widely used data mining technique which finds all significant association rules between variables in a dataset (Agrawal et al., 1993). The concept of association rules can be best illustrated by an example in marketing strategy development where data mining was initially used. The technique may find what type of customers, specified by combinations of one or more characteristics such as age, gender or occupation, are more likely to purchase a certain product, or the association rules. In this example, the response is whether the product is purchased. Similarly, when the technique is applied to this study, it finds what type of injured workers are more likely to result in a prolonged claim duration. Therefore, the response in this study is whether a duration is greater than two weeks. It is also a non-parametric method that is not based on any assumption on a shape of a decision boundary, and hence is suitable for prediction problems involving many

predictors. In this study, the technique, which predicts duration outcomes in a different way to the tree-based methods, is used to find any patterns of data (i.e. association rules between predictors and the response) in large regions of the tree, that have not yet been captured by the tree, to provide the final triage model with additional significant rules.

Association rule learning is not used as the primary method applied to the whole set of data in this study, because it is difficult to specify a set of rules to define segments of a triage model by using the association rules provided by the method. Rather than splitting a dataset, it provides a number of combinations of predictor values indicating an increased likelihood of prolonged duration. Therefore, it is less appropriate to develop a segment-based triage model.

4.2.9 Evaluation of triage models

		Actual class	
		0	**1**
Predicted class	**0**	74,069 (True Negatives)	18,695 (False Negatives)
	1	5,353 (False Positives)	6,749 (True Positives)

Fig. 4.4 An example of confusion matrix for binary response

Many widely-used evaluation measures for classification models are calculated based on a confusion matrix which contains the number of observations correctly/ incorrectly predicted for each class of the response. For a model involving a binary response, the confusion matrix contains four components, true negatives/ positives and false negatives/ positives, as seen from the example in Figure 4.4. The example indicates that 74,069 and 6,749 observations in the test dataset are correctly predicted to have the response value of 0 and 1, respectively, while 18,495 and 5,353 observations are incorrectly predicted to have the response value of 0 and 1, respectively. Of the measures used in this study, the prediction accuracy, sensitivity, specificity and lifts are calculated by using the components of confusion matrix.

The evaluations conducted on the triage models include the following.

- **Prediction accuracy**

The prediction accuracy[2] of a model is the proportion of claims of which a duration outcome is correctly predicted in a test dataset (see Bekkar et al., 2013, p.28). A calculation example is in Appendix C.

- **Sensitivity**

  The sensitivity measures the proportion of claims for which duration outcome is correctly predicted from claims in a test dataset having a prolonged duration (see Bekkar et al., 2013, p.28). In other words, it is a measure of how many claims having a prolonged duration are correctly identified. A calculation example is in Appendix C.

- **Specificity**

  The specificity measures the proportion of claims for which duration outcome is correctly predicted from claims in a test dataset not having a prolonged duration (see Bekkar et al., 2013, p.28). In other words, it is a measure of how many claims not having a prolonged duration are correctly identified. A calculation example is in Appendix C.

- **Lift for prolonged duration**

  The lift for prolonged duration measures the proportion of claims for which duration outcome is correctly predicted from claims in a test dataset predicted to have a prolonged duration, adjusted for overall proportion of claims with a prolonged duration[3]. Such adjustment is especially important when an analysis involves a dataset which is not balanced with respect to a response, such as our dataset. For example, we can consider an unbalanced dataset in which 90% of observations actually have a response value of 1; and a model to select observations predicted to have a response value of 1 from the dataset. If 80% of the selected observations actually have the response value of 1, it seems that the model is 80% accurate. However, it is unlikely to be a meaningful model because the proportion of observations having the response value of 1 is 90%, already higher than the 80% so that even a set of randomly selected observations from the dataset will have, on average, higher chance (90%) to have the response value of 1. Therefore, we need an evaluation measure to address such an imbalance in a dataset. In other word, the lift means how much better a prediction by a model is, compared to a random prediction based on prior probabilities. For example, lift of 2 means a

[2]While this measure is often called "accuracy", we use the term "prediction accuracy" in this study to make a clearer distinction with accuracy of probability.

[3]Lift is explained here in the context of this study. For general concept of lift, see Tan et al. (2006).

prediction by the model is two times better than the random prediction. A calculation example for the lift is in Appendix C.

- **Lift for non-prolonged duration**

  Similarly, the lift for non-prolonged duration measures the proportion of claims for which duration outcome is correctly predicted from claims in a test dataset predicted not to have a prolonged duration, adjusted for overall proportion of claims not having a prolonged duration. A calculation example for the lift is in Appendix C.

- **Pseudo R^2**

  When an analysis involves a categorical response, R^2 cannot be calculated, and therefore pseudo R^2 is used instead. Of various pseudo R^2 measures, we have used Efron's R^2 (Institute for Digital Research and Education, n.d.), which is the most intuitive measure in this scenario. It is calculated on a test dataset, by multiplying 100 times the squared correlation between actual duration outcomes and predicted probabilities of having a prolonged duration. The predicted probability in this context means the best estimate for the duration outcome between 0 and 1. The pseudo R^2 and prediction accuracy both evaluate prediction ability of a model. However, unlike the prediction accuracy which evaluates predicted values that can only be 0 or 1, the pseudo R^2 looks into the best estimates, and measures to what extent those values are correct or wrong. Interpretation of this measure is similar to the R^2. For example, pseudo R^2 of 50 means 50% of the variation of the actual duration outcomes is explained by the prediction model.

- **5-fold cross validation for the model for psychological claims**

  We find that a triage model for psychological claims may not be robust. That is, evaluation outcomes vary with models developed from different training datasets. Therefore, to test our method, a 5-fold[4] cross validation is done to generate average evaluation outcomes. We randomly divide psychological claims into five subsets, develop and evaluate a triage model through the same method five times by using one of the subsets as a test dataset (and the dataset excluding the test dataset as a training dataset), and calculate averages for the evaluation outcomes.

- **Accuracy of probability**

[4] 5-fold, where "5" is relatively small, is chosen because the evaluation includes the accuracy of probability which requires quite many data points in a test dataset.

Accuracy indicates the accuracy of the probability of prolonged duration predicted for claims in each segment of the model. To quantify this, we have used the informal measure of the performance of predicted probabilities relative to the reference probability - the proportion of claims having a prolonged duration in the test dataset. The measurement is done by first calculating the deviation between predicted probabilities from the model; and actual probabilities calculated when the model is applied to a test dataset (model deviation). We then calculate another deviation between the overall proportion of claims having a prolonged duration in the test dataset and the actual probabilities (reference deviation). We then estimate how much smaller the model deviation is than the reference deviation (accuracy of probability). A calculation example is in Appendix C.

The accuracy of probability is different to the accuracy of prediction on duration of an individual claim, for which conventional evaluation methods for classification trees (e.g. prediction accuracy and lifts) are used. Given the accuracy of probability is high, the prediction accuracy and lifts would also be high only if predicted probabilities of prolonged duration are mostly close to either 0 or 1 so that not many claims are given wrong predictions. However, a model with high accuracy of probability and low prediction accuracy and lifts may still be useful. For instance, a model which has a big segment with the predicted probability of 0.49, and one small segment with the probability of 0.99 is very likely to be evaluated to have low prediction accuracy and lifts. That is because, with the predicted probability of 0.49 (close to neither 0 nor 1), nearly half of the claims in the big segment will, on average, be predicted incorrectly. However, if the probabilities are sufficiently accurate, it will be useful for a claims manager to identify high-risk claims in the small segment accurately. Hence, it is important not to dismiss a model with relatively low individual claim prediction accuracy, as such a model may still be very successful at segmenting claims into different risk levels and therefore at enabling the targeting of initiatives to reduce claims durations and costs.

4.3 Other considerations

4.3.1 Splitting training dataset and test dataset

The training dataset is created by random sampling of 50% of the claims in the whole dataset. The test dataset is comprised of the claims not in the training dataset. We find that the results

are similar for splits with different proportions of the training dataset including 25% and 75%.

4.3.2 Combining hierarchical variables into a variable of single layer

Multiple variables of hierarchical descriptions have been combined into a variable of single layer for *Nature of injury* and *Bodily location of injury*.

Nature of injury is described at two levels, detailed and group descriptions. Therefore, there are two fields, or variables describing *Nature of injury*. For example, all claims having the detailed descriptions as "Dislocation", "Trauma to joints and ligaments, not elsewhere classified" and "Trauma to joints and ligaments, unspecified" have the same group description as "Trauma to joints and ligaments". It is considered necessary to leave sufficient details for the Nature of injury which is found to be an important predictor for prolonged duration from our preliminary analysis. However, use of the detailed descriptions would make specifications of segments of the final model too complex, while the use of only the group descriptions would discard a large amount of information. Therefore, we have used new variable combining the detailed and group descriptions, by taking the group descriptions only for nature of injuries of which the detailed descriptions are used for relatively a small number of data points. For example, while "Dislocation" is used as a class for the new variable, for "Trauma to joints and ligaments, not elsewhere classified" and "Trauma to joints and ligaments, unspecified", "Trauma to joints and ligaments except dislocation" is used as a class.

We have done a similar process for the *Bodily location of injury* which is also found to be an important predictor for prolonged duration from our preliminary analysis.

4.3.3 Creating binary variables

Binary coding was used for several numerical variables for the costs (e.g. *Total legal cost made for all prior claims*) in addition to the response variable. This is to create a predictor that shows whether the cost is non-zero. Such coding is to simplify the data structure, and sometimes to focus on the existence, rather than magnitude, of an event.

4.3.4 Excluding unreliable variables

Several variables were initially excluded from the analysis due to an excessive number of missing values or other special reasons. For example, *Income estimate for injured worker* is

considered to be biased and excluded because there is a tendency that an income estimate is not available when a claim does not involve income maintenance payments.

4.3.5 Selecting predictors

Predictors are included in the analysis only when their values are to be known by an insurer of workers' compensation when claims are made. This is to ensure that a final triage model can be used for new claims.

4.3.6 Variable clustering

As discussed in Section 3.2.2, we conduct a clustering of variables analysis - finding the groups of variables that are highly related to one another and therefore provide the same information - which is known to be is useful for dimension reduction and variable selection (Chavent et al., 2011). The analysis is done by forming groups, or clusters, of correlated variables according to a homogeneity criterion based on the squared correlations for numerical variables, and sum of correlation ratios for categorical variables (Chavent et al., 2011). We have used the hclustvar algorithm in R's ClustOfVar package.

4.3.7 Exclusion of time variables

An association between time variables (year and date of injury) and the response, if found, indicates a time trend of the response. If a time variable is selected to define any segment of the triage model developed from the training dataset, the model will involve a process to divide claims by determining if a claim is made before or after a certain time point within the time range of the dataset. However, it is not practical to use such rule for the triage model, because all new claims will be made after that time point. Therefore, the final model is to be developed without the time variables, and any relevant findings are to be reported separately in Section 6.

4.3.8 Combining regions of tree output

Two (or more) regions segmented by the tree fitted to the training dataset are combined in the final triage model when probabilities, or risks, of prolonged claim duration of those regions are similar. This is done because the final model aims to segment claims solely by level of risk.

Chapter 5

Results

The main research objective is to find factors associated with durations of psychological claims, and to develop a triage model for such claims. For comparison, the tree-based methods are applied both to all claims and only to the psychological claims.

Modelling is repeated with different parameter values to specify trees and with different predictor settings, to give the best model in terms of accuracy of probability and robustness (as explained in Section 4.1).

While the classification tree almost always generates an accurate triage model with an accuracy of probability greater than 98% for all claims, it is not always so for psychological claims. This is likely because of the smaller number of data points for psychological claims. Therefore, we made a few attempts to refine the model (through association rule learning), and validate the method for the analysis on psychological claims (through 5-fold cross validation), as shown below.

Key evaluation measures are summarised in Table 5.2. An example of calculations of evaluation measures is in Appendix C.

Table 5.1 Components of confusion matrix

	True negatives	False negatives	False positives	True positives
Actual value	0	1	0	1
Model				
Random Forest - All Claims	76677	24986	143	171
Classification tree - All Claims	74069	18495	5353	6749
Conditional inference tree - All Claims	73238	18582	6184	6662
Random Forest - Psych Claims	1494	1880	46	83
Classification tree - Psych Claims	355	250	1212	1719
Conditional inference tree - Psych Claims	334	297	1245	1660

Table 5.2 Summary of evaluations

Model	Prediction accuracy	Sensitivity	Specificity	Lift for prolonged duration	Lift for non-prolonged duration	Psuedo R^2	Accuracy of probability
Random Forest - All Claims	0.7536	0.0068	0.9981	2.2075	1.0012	14.9141	-
Classification tree - All Claims	0.7722	0.2674	0.9326	2.3122	1.0545	10.2785	0.9899
Conditional inference tree - All Claims	0.7634	0.2639	0.9221	2.1502	1.0511	13.5526	0.9978
Random Forest - Psych Claims	0.4502	0.0423	0.9701	1.1482	1.0072	2.4722	-
Classification tree - Psych Claims	0.5852*	0.7568*	0.3725*	1.0843*	1.2589*	2.2466*	0.7326*
Conditional inference tree - Psych Claims	0.5639	0.8482	0.2115	1.0325	1.1854	1.8048	0.8514

The measures are calculated based on the four components of the confusion matrices in Table 5.1
* Evaluation results from 5-fold cross validation.

Table 5.3 Important predictors in order of importance

Random Forest - All Claims	Random Forest - Psych Claims
Nature of injury	Occupation
Occupation	Bodily location of injury of the most recent prior claim
Bodily location of injury	Lag between date of injury and reporting date
Bodily location of injury of the most recent prior claim	Age of injured worker
Age of injured worker	Industry classifications
Lag between date of injury and reporting date	Premium rate
Industry classifications	Nature of injury
Premium rate	Agency of accident of the most recent prior claim
Mechanism of injury	Total payments on the most recent prior claim
Agency of accident	Nature of injury of the most recent prior claim
Classification tree - All Claims	**Classification tree - Psych Claims***
Nature of injury	Occupation
Bodily location of injury	Bodily location of the most recent prior claim
Mechanism of injury	Age of injured worker
Occupation	Industry classification
Agency of accident	Nature of injury of the most recent prior claim
Industry classification	Agency of accident of the most recent prior claim
Premium rate	Total payments on the most recent prior claim
Nature of injury of the most recent prior claim	Lag between date of injury and reporting date
Agency of accident of the most recent prior claim	
Bodily location of the most recent prior claim	

* The output from R shows only eight important predictors.

5.1 Models for all claims

5.1.1 Random Forest on all claims

The Random Forest model is developed for all claims, by creating 500 trees. Each split of a tree is made by only considering six[1] randomly selected predictors from all selected predictors. When the model developed from the training dataset is applied to the test dataset, it gives prediction accuracy of 0.75, sensitivity of 0.01, specificity of 0.99, lift for prolonged duration of 2.21, lift for non-prolonged duration of 1.00 and pseudo R^2 of 14.91. We find

[1]The default value by the Random Forest algorithm is 6, which is square-root of total number of predictors rounded to the nearest integer. The results did not change dramatically with a use of different values.

that only a small number of claims are predicted to have a prolonged duration[2]. The ten most important factors in predicting a duration outcome indicated by the model in order of importance are in Table 5.3.

5.1.2 Classification tree on all claims

The classification tree is fitted for all claims, by aiming at having seven to thirteen regions. Such number of regions or segments of a triage model is considered manageable by claims managers and also specific enough to specify relatively small segments. Tree size is adjusted by two parameters, the complexity, which determines the level of pruning, and the minimum size of a region. The classification tree for the final triage model is in Figure 5.1. The predicted probabilities of prolonged duration in the eleven regions created by the tree range between 17% and 69%. The model gives prediction accuracy of 0.77, sensitivity of 0.27, specificity of 0.93, lift for prolonged duration of 2.31, lift for non-prolonged duration of 1.05, pseudo R^2 of 10.28, and accuracy of probability of 0.99. The ten most important factors in predicting a duration outcome indicated by the model are in Table 5.3. Only a part of those important factors is used to make the splitting rules for the classification tree because multiple factors may be correlated and containing similar information.

5.1.3 Conditional inference tree on all claims

The conditional inference tree tree is fitted for all claims as a robustness test for the classification tree. It also aims to have seven to thirteen regions. Tree size is adjusted by two parameters, the confidence level and the minimum size of a region. The predicted probabilities of prolonged duration in the eleven regions created by the tree range between 1.5% and 51.4%. Although the conditional inference tree model tends to expand more on low-risk claims, the evaluation measures are quite similar to the classification tree model. The conditional inference tree model gives prediction accuracy of 0.76, sensitivity of 0.26, specificity of 0.92, lift for prolonged duration of 2.15, lift for non-prolonged duration of 1.05, pseudo R^2 of 13.55, and accuracy of probability of 0.998.

[2]In Table 5.2, the second and third columns show numbers of claims predicted not to have a prolonged duration, while the fourth and fifth columns show numbers of claims predicted to have a prolonged duration.

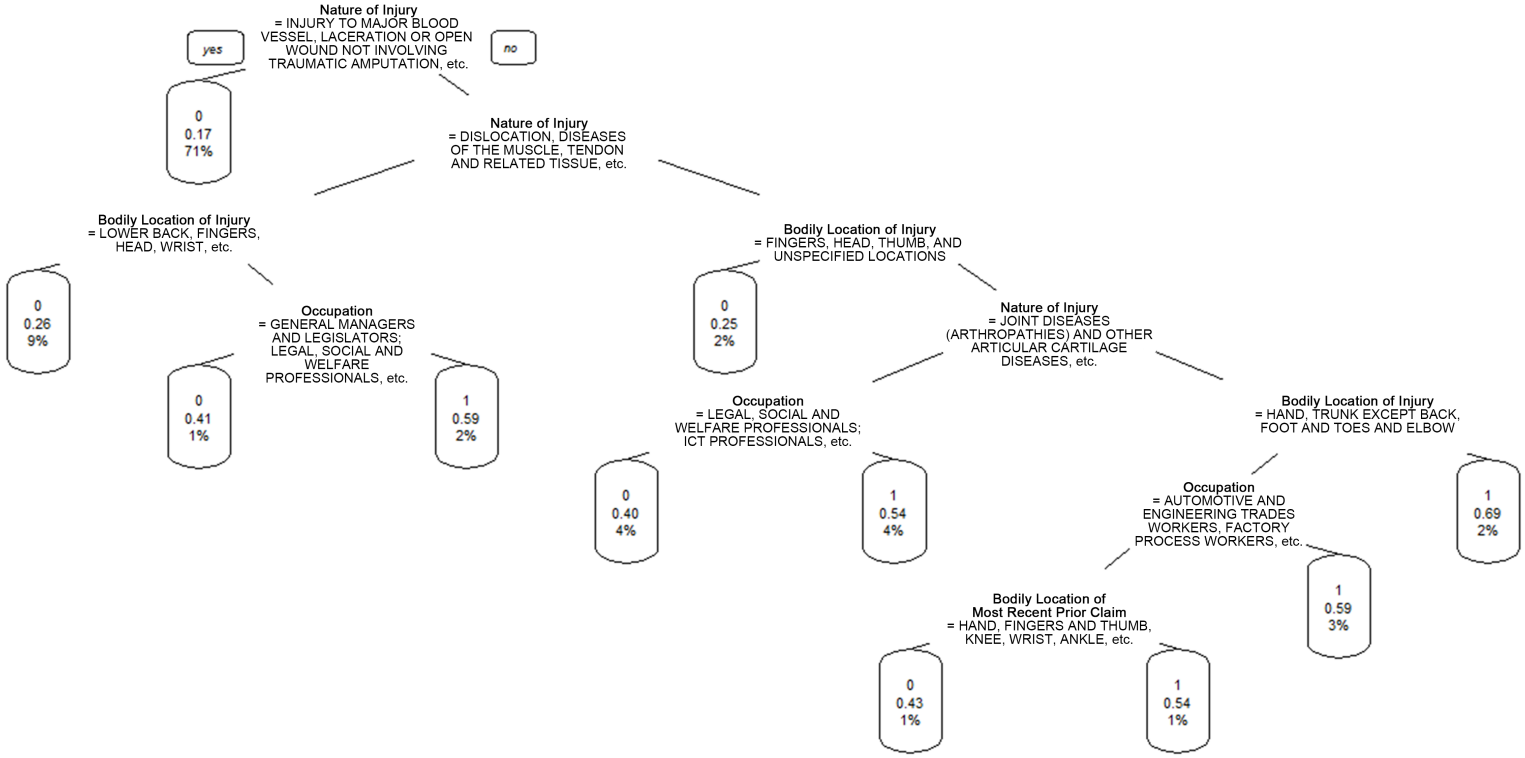

Fig. 5.1 Classification tree on all claims

Detailed specification of the tree is in Appendix B. The first, second, and last figure in each region of the tree means the predicted value of response, the probability of prolonged duration, and the proportion of claims lying in the region, respectively. For each split, the tree sends claims with higher probability of prolonged duration to the right region.

5.1.4 Final triage model for all claims

The final triage model developed based on the classification tree is summarised in Appendix D. Along with the probabilities of prolonged duration and the proportions of claims lying in each segment, average durations for each segment are also given. Segment 7 is created by combining two regions of the classification tree as they have a similar level of risk.

5.2 Models for psychological claims

5.2.1 Random Forest on psychological claims

The Random Forest model is developed for psychological claims, by creating 500 trees. Each split of a tree is made by only considering six[3] randomly selected predictors from all predictors. When the model developed from the training dataset is applied to the test dataset, it gives prediction accuracy of 0.45, sensitivity of 0.04, specificity of 0.97, lift for prolonged duration of 1.15, lift for non-prolonged duration of 1.01, and pseudo R^2 of 2.47. The low values of evaluation measures imply the model is not useful in predicting the duration outcome for a single claim, but it may - as noted previously - still be useful in segmenting claims into different risk regions. We find that only a small number of claims are predicted to have a prolonged duration. The eight most important factors in predicting a duration outcome indicated by the model are in Table 5.3.

5.2.2 Classification tree on psychological claims

The classification tree is fitted for psychological claims. We first aimed at having seven to thirteen regions for the tree, like our aim for the classification on all claims. However, such number did not result in an acceptable level of accuracy of probability likely due to the smaller number of data points for psychological claims only. Therefore, we restrict the number of regions to six. In addition, to mitigate the robustness problem[4] found, the tree is fitted by using only the ten most important factors found in the Random Forest model for psychological claims. It is known that the Random Forest can be used to measure importance of factors and hence reduce the number of factors used in an analysis (Ziegler and König, 2014). The classification tree for the final triage model is in Figure 5.2. The predicted probabilities of prolonged duration in the four regions created by the tree range between 34%

[3] Square-root of total number of predictors rounded to the nearest integer

[4] A specification of an classification tree on psychological claims is quite sensitive to settings such as predictor selection and user-defined parameter values

and 61%. The model gives low value of evaluation measures related to individual prediction for a duration outcome (i.e. prediction accuracy of 0.59, sensitivity of 0.87, specificity of 0.23, lift for prolonged duration of 1.05, lift for non-prolonged duration of 1.32 and pseudo R^2 of 2.05). However, the accuracy of probability is 0.92. The eight most important factors in predicting a duration outcome indicated by the model are in Table 5.3. Only a part of those important factors is used to make the splitting rules for the classification tree because multiple factors may be correlated and containing similar information.

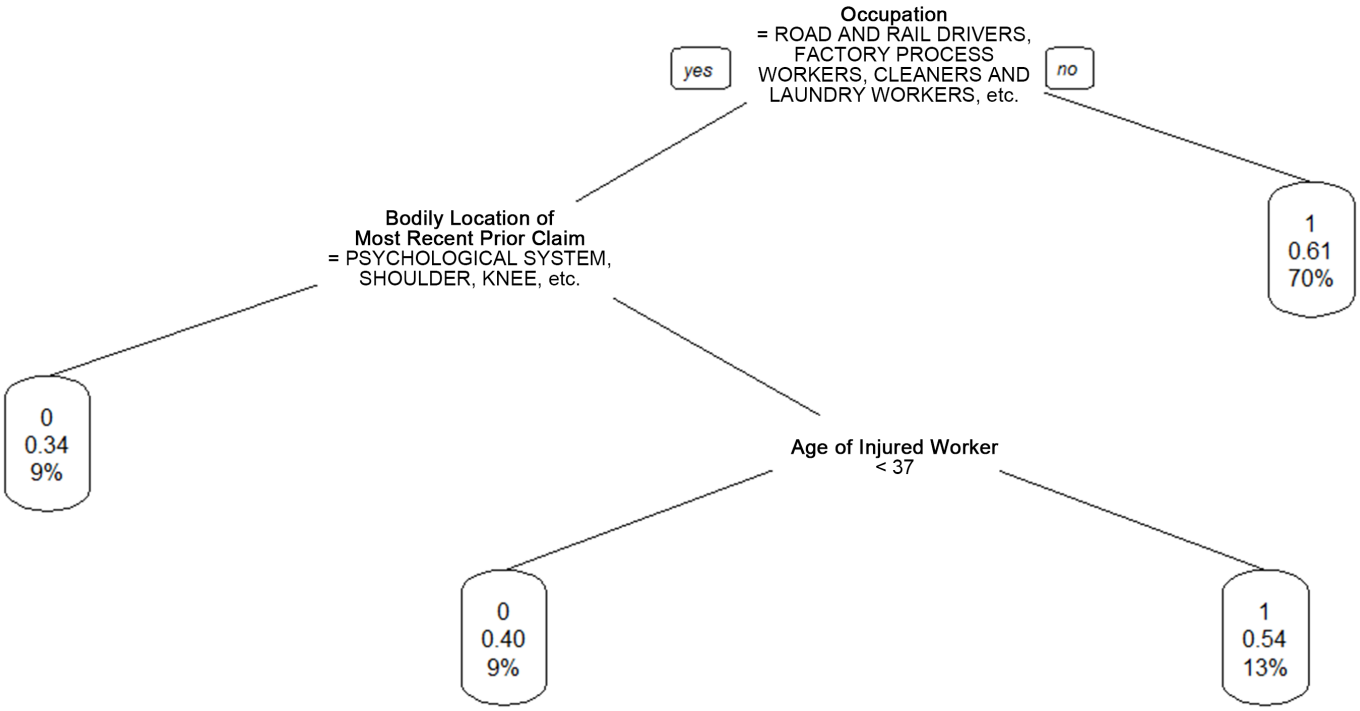

Fig. 5.2 Classification tree on psychological claims

Detailed specification of the tree is in Appendix B. The first, second, and last figure in each region of the tree means the predicted value of response, the probability of prolonged duration, and the proportion of claims lying in the region, respectively. For each split, the tree sends claims with higher probability of prolonged duration to the right region.

5.2.3 5-fold cross validation for classification tree on psychological claims

We find that classification tree is quite sensitive to settings such as parameter values to specify a tree and predictor selection. Therefore, we conduct 5-fold cross validation for the settings used for the classification tree above. We find, on average, prediction accuracy of 0.59, sensitivity of 0.76, specificity of 0.37, lift for prolonged duration of 1.08, lift for

non-prolonged duration of 1.26, pseudo R^2 of 2.25, and accuracy of probability of 0.73 as in Table 5.4.

Table 5.4 Result of Cross Validation

	Run 1	Run 2	Run 3	Run 4	Run 5	Average
Prediction Accuracy	0.6031	0.5826	0.5816	0.5783	0.5802	0.5812
Sensitivity	0.8381	0.8323	0.8342	0.663	0.6131	0.7568
Specificity	0.2974	0.2859	0.2729	0.4725	0.5339	0.3725
Lift for prolonged duration	1.0756	1.0694	1.0613	1.1045	1.1106	1.0843
Lift for non-prolonged duration	1.3469	1.2894	1.2753	1.1909	1.1918	1.2589
Pseudo R^2	2.3722	2.1863	1.5099	2.8146	2.3500	2.2466
Accuracy of Probability	0.7903	0.8750	0.6329	0.7227	0.6421	0.7326

5.2.4 Association rule learning on the big region of classification tree model

We find that 70% of claims in the training dataset fall in the region of the highest risk of prolonged duration in the classification tree (the high-risk region). To refine the region further, we use an association rule learning algorithm, apriori (Hahsler et al., 2017), to identify claims with even higher risk of prolonged duration in the high-risk region[5]. Association rule learning finds thousands of association rules from the training dataset. The top five combinations of predictor values (by lift for prolonged duration) which lead to a prolonged duration are summarised in Table 5.5.

For example, rule 1 with lift for prolonged duration of 1.22 states that when Nature of injury is Anxiety/ depression combined; Nature of injury of the most recent prior claim is not Musculoskeletal and connective tissue disease; and Industry classification is not Transport, postal and warehousing, the probability of prolonged duration is higher within the high-risk region. When the rule is tested on the test dataset, 336 out of 2,405 claims satisfy such conditions, and actually have the higher probability of prolonged duration of 0.6815.

We find that, in the high-risk region, when the rules are applied to the test dataset, the claims having the predictor values stated by the five rules have a 12-15% higher probability of prolonged duration than other claims. However, by applying only the common rules from

[5]apriori algorithm of R's arules package, with minimal number of items per item set of 1; minimal support of an item set of 0.1; and minimal confidence of rules/ association hyperedges of 0.1, has been applied to the dataset comprised of the 70% claims assigned to the biggest region of the classification tree. The categorical variables in the dataset are transformed to multiple binary variables by using dummyvar algorithm of R's caret package, as required by the apriori algorithm.

the five rules (see Table 5.5), we also find that having Nature of injury of Anxiety/ depression combined explains most of the probability increase.

Therefore, we add the rule learned, a claim with Nature of injury of Anxiety/ depression combined has a higher risk of prolonged duration, to the final triage model.

Table 5.5 Combination of predictor values leading to a prolonged duration within the high-risk region of the classification tree

	Predictor	Predictor Value	Lift for prolonged duration	Probability of prolonged duration	Number of such claims in testset
All claims				0.6017	2405
Rule 1	Nature of injury	Anxiety/depression combined	1.22	0.6815	336
	Nature of injury of most recent prior claim	Not musculoskeletal and connective tissue disease			
	Industry classification	Not transport, postal and warehousing			
Rule 2	Nature of injury	Anxiety/depression combined	1.22	0.6783	401
	Nature of injury of most recent prior claim	Not musculoskeletal and connective tissue disease			
	Agency of injury of most recent prior claim	Not non-powered handtools, appliances and equipment			
Rule 3	Nature of injury	Anxiety/depression combined	1.21	0.6938	369
	Nature of injury of most recent prior claim	Neither musculoskeletal and connective tissue disease nor traumatic joint/ligament and muscle/tendon injury			
Rule 4	Nature of injury	Anxiety/depression combined	1.21	0.6794	393
	Nature of injury of most recent prior claim	Neither musculoskeletal and connective tissue disease nor wounds, lacerations, amputations and internal organ damage			
Rule 5	Nature of injury	Anxiety/depression combined	1.21	0.6716	402
	Nature of injury of most recent prior claim	Not musculoskeletal and connective tissue disease			
	Bodily location of most recent prior claim	Not hand, fingers and thumb			
Common rule 1	Nature of injury	Anxiety/depression combined		0.6765	510
Common rule 2	Nature of injury	Anxiety/depression combined		0.6770	421
	Nature of injury of most recent prior claim	Not musculoskeletal and connective tissue disease			

5.2.5 Conditional inference tree on psych claims

The conditional inference tree is fitted for psychological claims as a robustness test for the classification tree. We also restrict the number of regions to six. The predicted probabilities of prolonged duration in the four regions created by the tree range between 46% and 71%. Although the conditional inference tree model tends to more evenly allocate the claims to each region compared to the classification tree model, the evaluation measures are quite similar to the classification tree model. The conditional inference tree model gives prediction accuracy of 0.56, lift for prolonged duration of 1.03, lift for non-prolonged duration of 1.19, pseudo R^2 of 1.80, and accuracy of probability of 0.85.

5.2.6 Final triage model for psychological claims

The final triage model for psychological claims developed based on the classification tree (and the association rule found from the big region of the tree) is summarised in Appendix E. Along with the probabilities of prolonged duration and the proportions of claims lying in each segment, average durations for each segment are also given. For the segment of the highest risk (Segment 4), specified solely by Occupation group, we also provide the ten occupation groups with higher risk of prolonged duration.

Chapter 6

Discussion

6.1 Triage model for all claims

The triage model for all claims is developed by the classification tree, to have ten segments as shown in Appendix D. The segments of the model are specified by dividing claims by using four factors *Nature of injury*; *Bodily location of injury*; *Occupation of injured worker*; and/ or *Bodily location of the most prior claim*, clearly in this order. The order shows that subsequent factors are found to be informative in predicting duration outcomes for the segments created by more prioritised factors. For example, *Occupation of injured worker* is found to be an informative factor for claims with some natures of injury and bodily locations of injury. These four factors are found to be important also by the Random Forest model. Although a number of other factors are also found to be important, they have not been used for specifications of segments, because they are not sufficiently informative to make additional specifications for the segments already specified by the four factors.

We find that the majority (71%) of claims belong to Segment 1 of the lowest-risk (probability of 17% for prolonged duration), which is solely specified by *Nature of injury*. In general, the claims in this segment are made due to injuries while those in other segments are made due to diseases. Segment 1 involves all categories of injuries except fracture and dislocation including *Traumatic joint/ ligament and muscle/ tendon injury except dislocation*; *Wounds, Lacerations, Amputations and Internal organ damage*; and *Burn*. This finding is consistent with a study of Johnson and Ondrich (1990) which finds that workers with amputations or bruises and contusions are likely to return to work earlier than those with fractures or dislocations. This is likely because the majority of work-related injuries are mild acute injuries which are cured soon after an affected area is healed. For example, the majority of *Wounds, Lacerations, Amputations and Internal organ damage* are injuries likely to be

mild such as “open wound not involving traumatic amputation”; and “contusion, bruising and superficial crushing”.

Segment 10 of the highest risk (probability of 68% for prolonged duration) is specified by *Nature of injury* and *Bodily location of injury*. The injuries in this segment are “Fractures on back, shoulder, wrist and neck”. Similar findings are seen from other studies. Cheadle et al. (1994) find that a diagnosis of back or neck sprain is associated with longer duration of work-related injuries. Baldwin et al. (1996) find that workers with back injuries are more likely to experience multiple sickness absences than workers with other type of injuries.

While types of injury are used to specify all segments, *Occupation of injured worker* is used to specify Segment 3, 4, 6, 7 and 8. We find the occupation is associated with duration outcomes differently between types of injury. This finding is discussed in Section 6.2 by comparing psychological claims and the other claims.

We find that, psychological claims are assigned to relatively high-risk segments[1]. As this finding is in line with Safe Work Australia (2015*b*)’s strategy to specify psychological injuries as a priority, we have been encouraged to develop the separate triage model for psychological claims.

6.2 Triage model for psychological claims

The triage model for psychological claims is developed by the classification tree, to have four segments, by specifying *Occupation of injured worker*, *Bodily location of injury of the most recent prior claim*, and/ or *Age of injured worker* (see Appendix E). These factors are found to be important also by the Random Forest model. All three factors for any workers would be known even before a claim is made by providers of workers’ compensation. Therefore, the triage model may be used for early intervention as well as for the other measures (see Section 2.3) to reduce claim durations.

We should note that more detailed segmentation for the triage model may be possible if there are more data points. We restricted the number of splits for the tree for psychological claims to have a stable accuracy of probability (see Section 4.2.9) for the model. The accuracy of probability is calculated based on proportions of claims with prolonged duration within each segment. Therefore, for a stable accuracy of probability, the model requires a sufficient number of data points in each segment, achievable by restricting the number of splits for the tree. This means we need to be cautious in interpreting the result. For example, although *Age*

[1]They appear in Segments 7, 8 and 9 which have probability of prolonged duration 52%, 47% and 58%, respectively

of injured worker is used as a predictor for prolonged duration only for lower-risk occupation groups, it may also be a predictor for the other occupation groups, if a tree is made by using a bigger dataset.

The factors used for the specifications are very different for the models for all claims and psychological claims. Excluding the area of the classification tree for all claims to which most of psychological claims are sent[2], we find the only factor commonly used for both trees is *Occupation of injured worker*, for which we could make a comparison.

Occupation of injured worker is found to be the most important predictor from the Random Forest model for psychological claims, on which the first split of the classification tree is also based. The segment of the highest risk (probability of 60% for prolonged duration) includes all occupations in *Managers* and *Community and personal service workers*; and the majority of *Professionals* and *Clerical and administrative workers*[3]. These categories include the majority of high socio-economic occupations such as chief executives, legal professionals and education professionals. The tree for all claims also makes a similar split based on *Occupation of injured worker* for the area of the tree where most of psychological claims are sent. However, in the region not involving psychological claims, it tends to send claims with such occupations to the lower-risk segment.

It is well known that, consistent with our findings, early return to work is correlated with higher socio-economic occupations for non-psychological injuries or diseases. For example, MacKenzie et al. (1998) finds that workers with a lost time injury of lower extremity fracture with higher socio-economic status have returned to work earlier.

It is more difficult to find studies directly examining an association between occupation and duration of mental health conditions. Blank et al. (2008)'s study on sickness absence due to psychological reasons find that obstacles for successful return to work include high job stressors and re-organisational stress. However, it does not indicate which occupations have such characteristics. Stansfeld et al. (1999)'s longitudinal, prospective cohort study find that higher job demand, which is often greater in jobs entailing greater authority, is associated with increased risk of psychiatric disorder. However, this study examines incidence rather than duration of disorders. Brenninkmeijer et al. (2008) find that low education is associated with a less favourable course of depressive symptoms. However, while education and occupation are expected to be correlated, this analysis is restricted to depression, and evaluates course of illness through surveys, rather than duration measures. Therefore, further investigation

[2] see the left region after the right-most split by using *Bodily Location of injury* in Figure 5.1

[3] Jobs in the other occupation categories are allocated across the different segments.

is needed to support and explain our finding of longer psychological claim duration for managerial and professional occupations.

Bodily location of injury for the most recent prior claim (*Prior bodily location*) is found to be a predictor for prolonged duration for the lower-risk occupations. However, we find that prior bodily locations in the lowest-risk segment (Segment 1) include all top seven bodily locations for which claims have long durations (*Neck and trunk*; *Psychological system*; *Other specified multiple locations*; *Neck*; *Neck and shoulder*; *Shoulder*; and *Trunk and limbs* in descending order of average duration). This means that the duration of a psychological claim made after another claim with a long duration is more likely to have a short duration. This finding is different to our initial expectation that workers who have made serious claims are again more likely to make a claim with longer duration. A possible but unproven explanation is that workers with an experience of serious claim know the system better, and therefore may know how to seek psychological treatments effectively and recover quickly. Another explanation may be that such better knowledge about the system makes them more likely to make a short-term claim for minor psychological issues. This finding raises interesting questions for future research.

Age of injured worker is found to be a predictor of prolonged duration (dividing Segment 2 and 3) with workers aged 37 and over assigned to the higher-risk segment (probability of 53% of prolonged duration). We previously have found that more psychological claims are made by workers aged 37 and over (see Figure 3.2), and now find here that the risk of prolonged duration is also higher for them. It is well established in the literature that duration of psychological injuries is longer for older workers (for example, see Cornelius et al., 2011; Dewa et al., 2002; Nieuwenhuijsen et al., 2006).

6.3 Factors excluded from the triage models

There are several factors that have not been used for the triage models.

The time factors, although they are found to be associated with duration outcomes, have been excluded from the main analysis. In the preliminary analysis, we find that *Date of injury* is associated with duration outcomes for psychological claims. The triage model from the analysis assigns claims made after August 2008 to segments of higher risk of prolonged duration. However, we have excluded this factor for our main analysis because, if such triage models are used in practice, all new claims made later than the claims in this study will be assigned to the same region, making such specification unhelpful. We find that the risks of prolonged duration have been significantly higher between 2010 and 2013 than the

years before as well as after that period. This is likely because of the significant legislative amendments to the WorkCover Scheme[4] in June 2008 (Safe Work Australia, 2009, p.14), and the transition following the amendments.

Lag between date of injury and reporting date (*Lag*) is found to be the fifth most important factor for overall claims, and the third most important factor for psychological claims by the Random Forest models. However, it has not been used as a factor for the triage models. This is because, while *Lag* is significant in predicting a prolonged duration if used earlier than the used factors in the classification trees, it is not sufficiently informative to be used for an additional specification for segments already specified by more prioritised factors that have been used. From the separate investigation on *Lag*[5], we find that higher *Lag* is associated with a higher risk of prolonged duration for psychological claims. This indicates, most information contained in *Lag* is captured by the two factors already used for psychological claims, occupation and age. Further investigation is needed to explain, and clarify the direction of, this association. This is important because *Lag* can be reduced by early intervention.

Of the four aspects of injury described by TOOCS 3.1 (Safe Work Australia, 2008), *Nature of injury* and *Bodily location of injury* have been used to develop the triage model for overall claims, while the other two aspects have not been used. The unused aspects, *Mechanism of injury* and *Agency of accident*, are also found to be the ninth and tenth most important factors, respectively, by the Random Forest model on overall claims. However, they are highly correlated with the used aspects, and therefore do not enhance a model already containing those aspects. For example, when they are set to be used before the used aspects, they also specify the majority of psychological claims as high risk by assigning *Agency of accident* as "non-physical agencies" or *Mechanism of injury* as "mental stress" to high-risk segments.

Total payments on the most recent prior claim is found to be associated with the duration outcome for psychological claims, but is excluded from the triage model because the information contained in the factor is already captured by *Bodily location of injury for the most recent prior claim*. It is found that claims with payments greater than $49,050 have lower risk of prolonged duration. However, this finding is different to our initial expectation that workers who have made serious claims are again more likely to make a claim with longer duration. This is similar to the finding that *Bodily location of injury for the most recent prior*

[4]The former name for Return To Work South Australia

[5]We have created several classification trees involving just one factor in addition to *Lag*

claim is associated with risk of prolonged duration for psychological claims in Section 6.2. Possible explanations are also discussed in Section 6.2.

6.4 Discussion of methods

We should note that our measure of risk, whether duration is greater than two weeks, is clearly different to the numerical value of duration. Indeed, we find that, while the two measures are positively correlated for the segments of the triage models, they are not always consistent. For example, in the triage model for all claims, the claims in Segment 1 compared to Segment 2, have a lower risk of prolonged duration, but have a higher average duration. The inconsistency is found to be due to the small number of claims with very long durations over 2,000 days in Segment 1. It suggests that claims in Segment 1 have a lower chance to have a prolonged duration, but a higher chance to have a very long duration. This finding is consistent with the fact that most injuries in Segment 1 are minor and acute, but a small number (e.g. serious trauma to muscles and tendons) are very serious and prolonged.

Using duration of claim as the (numerical) dependent variable would capture such effects, and therefore was considered as an alternative dependent variable to that chosen, the two week flag. However, for this study predicting duration outcomes based only on information available for new claims, use of duration as the dependent variable is not feasible. First, for predictions on very long durations, an analysis on claims already in progress, rather than new claims, involving more information obtained from the progress of such claims, would be more appropriate. That means, claims managers should identify any claim likely to become a non-short-term claim for a new claim, and may be able to identify a claim likely to become long-term for a claim already in progress. Second, due to the additional complexity, models based on a numerical dependent variable were found to be less robust and - in some cases - not to converge. A useful topic for future research would be development of modelling structure and techniques that treat claim duration as a continuous variable or a categorical variable with more than two levels, without sacrificing robustness.

We find that the all triage models or classification trees developed in this study do not make an accurate prediction of duration outcome for each claim (individual prediction), while they provide useful estimates for the risk for each segment (as evaluated by accuracy of probability) as shown in Table 5.2. For example, the classification tree model for all claims is evaluated to have relatively low level of individual prediction ability (prediction accuracy of 0.77, sensitivity of 0.27, specificity of 0.93, lift for prolonged duration of 2.31, lift for non-prolonged duration of 1.05 and pseudo R^2 of 10.28). The prediction accuracy in

this instance is very low given our imbalance dataset comprised of 24% of claims having prolonged claim durations. When we simply predict all claims not to have a prolonged duration, the prediction accuracy is already 76%. Although the specificity is quite high, that is mostly due to the unbalanced dataset. This becomes more obvious when we see the low level of the sensitivity. From the lifts, we see that, while the model has some ability to correctly identify high-risk claims, the identified set of claims include only a small portion of claims having a prolonged duration. However, the model is shown to produce segments that do differ in their risk levels, and to estimate the average risk-levels in these segments accurately.

The evaluations on the Random Forests models (see Table 5.2) show difficulties of individual predictions. The Random Forests algorithm aims to improve the accuracy of individual predictions by taking a consensus value from a number of randomised trees. However, the Random Forests models predict almost all claims not to have a prolonged duration, and achieve the prediction accuracies close to merely the proportion of claims not having a prolonged duration. The psuedo R^2 are also at low levels. From this finding, we conclude that these models are not useful for individual predictions.

We have used Neural Network (see Fritsch et al., 2016), as an alternative method, to further examine the possibility of individual predictions[6]. However, while the modelling for all claims did not converge likely due to the large data size, the Neural Network model for psychological claims provided similar results[7] to the tree methods, while being less interpretable.

For a triage model to make an accurate individual prediction, all segments should have probabilities of prolonged duration close to either 0 or 1. If not, although the estimated probabilities are accurate, a tree model cannot have a low error rate for individual predictions. However, a model with low level of individual prediction ability can still be useful given that the model can separate high and low-risk claims. Our triage models for overall claims and psychological claims segment claims by the probabilities of prolonged duration ranging between 17% and 68% and between 37% and 60%, respectively. This idea can be illustrated by the following example.

[6] An explanation or discussion for the method is not included in this paper because it is not used to develop the models as it provides results not better than the methods used. The neuralnet algorithm in the R's package of the same name, with three hidden layers and maximum steps of 1,000,000, has been applied to the training dataset of which categorical variables are transformed to multiple binary variables by using dummyvar algorithm of R's caret package, as required by the neuralnet algorithm. The developed model is evaluated on the test dataset.

[7] Prediction accuracy of 0.5731, sensitivity of 0.8166, specificity of 0.2647, lift for prolonged duration of 1.0459, lift for non-prolonged duration of 1.2071 and psuedo R^2 of 1.4232

> We have a tree with two segments. 90% of claims fall into the first segment of which the probability of prolonged duration is 0.49. The other 10% of claims fall into the second segment of which the probability is 0.99. We suppose the probabilities and the sizes of the segments are accurate. Then, if we make predictions for new claims by using this tree model, all claims assigned to the first segment will be predicted not to have a prolonged duration because the probability is less than 0.5. Therefore, the error rate for the segment will be around 0.49. Similarly, the error rate for the second segment will be around 0.01. Therefore, the overall error rate, considering the sizes of segments, will be around 0.49 times 90% plus 0.01 times 10%, or 0.442. We can conclude that the tree in this example with very high error rate cannot make accurate individual predictions. However, this tree model would be clearly useful for claims managers in identifying 10% of new claims at higher risk of prolonged duration.

In many cases, classification trees are used to make individual predictions by dividing observations into regions with different likelihoods to have each categorical value of the response. However, in this study, the trees are used to develop a triage model by utilising one of the processes of its algorithm for the predictions, that is dividing observations. We believe such a triage model can be useful for claims managers who need to make segments for claims rather than predict a duration outcome for every claim.

6.5 Other limitations

Although South Australia has a somewhat different distribution of industries to other states and territories of Australia, the findings from this study would be applicable throughout Australia if individual workers have similar properties with respect to work-related injuries or disease, regardless of the distribution of workers.

While the triage model for all claims tends to specify small segments for high-risk claims among all claims, that for psychological claims does so for low-risk claims among other claims. We can see this tendency from the size of the segments (e.g. 69% of psychological claims assigned to the highest-risk segment) as well as from the lift measures (e.g. the lift for non-prolonged duration greater than the lift for prolonged duration for the models for psychological claims). In other words, all psychological claims are at high risk except for a small portion of low-risk claims. This result would be less desirable than the reverse, which would enable better allocation of cost and effort by claims managers.

Chapter 7

Conclusion

We have identified the important factors associated with duration of claims, and developed a triage model for overall claims and claims due to psychological injuries, by using classification tree and association rule learning. The models segment the whole set of claims into high and low-risk segments according to easily interpretable rules. The probabilities of prolonged duration range between 17% and 68% for overall claims and between 37% and 60% for psychological claims. The models can be used by claims managers in prioritising initiatives for high-risk segments.

The most important factor associated with the duration of overall claims is found to be the type of injury or disease. In general, claims due to injuries are assigned to the lowest-risk segment of the model, while claims due to diseases are assigned across higher-risk segments, with some exceptions. A remarkable exception is that "Fractures on back, shoulder, wrist and neck" are assigned to the highest-risk segment.

Occupation of an injured worker is found to be the most important predictor for prolonged duration for claims due to psychological injuries. We find that the majority of high socio-economic occupations such as chief executives, legal professionals and education professionals are associated with increased risk of prolonged duration.

We find that, among the lower-risk psychological claims in terms of occupation, claims are shorter on average when they follow a serious previous claim as defined by bodily location. This finding is different to our initial expectation that workers who have made serious claims are again more likely to make a claim with longer duration, and raises interesting questions for further research. Of those following a serious previous claim, higher age is found to be associated with the increased risk of prolonged duration.

We suggest a more detailed segmentation for the triage model for psychological claims based on a larger dataset. In our analysis, the reduced number of segments for the model was

needed to assign a sufficient number of data points in our dataset to each segment, which is required to achieve robust probabilities of prolonged duration provided by the model. Therefore, a more detailed analysis would be possible on a larger dataset.

From a methodological perspective, we find that, although it is difficult to accurately predict an individual claim duration, robust segmentations of claims by risk of prolonged duration can be made. Duration is associated with many factors and is highly variable, so an accurate prediction on the duration for an individual claim is very difficult. However, by using classification tree and association rule learning, we have developed a relatively robust and practical triage model based on claim segmentations. Given the very significant financial and societal burden of work-related injuries, such a model can have a major positive impact through enabling faster return to work.

References

Agrawal, R., Imieliński, T. and Swami, A. (1993), 'Mining association rules between sets of items in large databases', *SIGMOD Rec.* **22**(2), 207–216.
URL: *http://doi.acm.org/10.1145/170036.170072*

Agrawal, R. and Srikant, R. (1994), Fast algorithms for mining association rules, *in* 'Proc. 20th int. conf. very large data bases, VLDB', Vol. 1215, pp. 487–499.

ANZ (2017), 'Health and safety management system overview', http://www.anz.com/resources/c/9/c9fc791a-e613-4485-855f-9a3f1962905e/2017-healthsafety-mgt-sys.pdf?MOD=AJPERES.

Apte, C., Liu, B., Pednault, E. P. D. and Smyth, P. (2002), 'Business applications of data mining', *Commun. ACM* **45**(8), 49–53.
URL: *http://doi.acm.org/10.1145/545151.545178*

Australian Bureau Of Statistics (2009), '1220.0 - anzsco - australian and new zealand standard classification of occupations, first edition, revision 1', http://www.abs.gov.au/ausstats/abs@.nsf/0/2EFCC979A2B4A78DCA2575DF002DA70B?opendocument.

Australian Bureau of Statistics (2014), 'Work-related injuries, australia, jul to jun 2014', http://www.abs.gov.au/ausstats/abs@.nsf/mf/6324.0.

Australian Bureau Of Statistics (2017), 'Eq08 - employed persons by occupation unit group of main job (anzsco), sex, state and territory, august 1986 onwards (pivot table)', http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6291.0.55.003Nov%202016?OpenDocument.

Baldwin, M. L., Johnson, W. G. and Butler, R. J. (1996), 'The error of using returns-to-work to measure the outcomes of health care', *American Journal of Industrial Medicine* **29**(6), 632–641.
URL: *http://dx.doi.org/10.1002/(SICI)1097-0274(199606)29:6<632::AID-AJIM7>3.0.CO;2-L*

Bekkar, M., Djemaa, H. K. and Alitouche, T. A. (2013), 'Evaluation measures for models assessment over imbalanced data sets', *Iournal Of Information Engineering and Applications* **3**(10).

Beyondblue (n.d.), 'Men in the workplace', https://www.beyondblue.org.au/who-does-it-affect/men/what-causes-anxiety-and-depression-in-men/men-in-the-workplace.

Black, O., Sim, M., Collie, A. and Smith, P. (2016), 'O20-4 modifiable factors associated with return-to-work self-efficacy; exploring early-claim differences between workers with a psychological or upper-body musculoskeletal injury', *Occupational and Environmental Medicine* **73**(Suppl 1), A38–A38.
URL: *http://oem.bmj.com/content/73/Suppl_1/A38.2*

Blank, L., Peters, J., Pickvance, S., Wilford, J. and MacDonald, E. (2008), 'A systematic review of the factors which predict return to work for people suffering episodes of poor mental health', *Journal of Occupational Rehabilitation* **18**(1), 27–34.
URL: *http://dx.doi.org/10.1007/s10926-008-9121-8*

Brenninkmeijer, V., Houtman, I. and Blonk, R. (2008), 'Depressed and absent from work: predicting prolonged depressive symptomatology among employees.', *Occupational Medicine (Oxford, England)* **58**(4), 295 – 301.
URL: *http://simsrad.net.ocs.mq.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=18434294&site=ehost-live*

Cancelliere, C., Donovan, J., Stochkendahl, M. J., Biscardi, M., Ammendolia, C., Myburgh, C. and Cassidy, J. D. (2016), 'Factors affecting return to work after injury or illness: best evidence synthesis of systematic reviews', *Chiropractic & Manual Therapies* **24**(1), 32.
URL: *http://dx.doi.org/10.1186/s12998-016-0113-z*

Cantley, L. F., Tessier-Sherman, B., Slade, M. D., Galusha, D. and Cullen, M. R. (2015), 'Expert ratings of job demand and job control as predictors of injury and musculoskeletal disorder risk in a manufacturing cohort', *Occup Environ Med* pp. oemed–2015.

Carnide, N., Franche, R.-L., Hogg-Johnson, S., Côté, P., Breslin, F. C., Severin, C. N., Bültmann, U. and Krause, N. (2016), 'Course of depressive symptoms following a workplace injury: A 12-month follow-up update', *Journal of Occupational Rehabilitation* **26**(2), 204–215.
URL: *http://dx.doi.org/10.1007/s10926-015-9604-3*

Chavent, M., Kuentz, V., Liquet, B. and Saracco, L. (2011), 'Clustofvar: an r package for the clustering of variables', *arXiv preprint arXiv:1112.0295* .

Cheadle, A., Franklin, G., Wolfhagen, C., Savarino, J., Liu, P., Salley, C. and Weaver, M. (1994), 'Factors influencing the duration of work-related disability: a population-based study of washington state workers' compensation.', *American Journal of Public Health* **84**(2), 190–196.

Colantonio, A., Salehi, S., Kristman, V., Cassidy, J. D., Carter, A., Vartanian, O., Bayley, M., Kirsh, B., Hébert, D. and Lewko, J. (2016), 'Return to work after work-related traumatic brain injury', *NeuroRehabilitation* **39**(3), 389–399.

Collie, A., Lane, T. J., Hassani-Mahmooei, B., Thompson, J. and McLeod, C. (2016), 'Does time off work after injury vary by jurisdiction? a comparative study of eight australian workers' compensation systems', *BMJ Open* **6**(5).
URL: *http://bmjopen.bmj.com/content/6/5/e010910*

Cornelius, L. R., van der Klink, J. J. L., Groothoff, J. W. and Brouwer, S. (2011), 'Prognostic factors of long term disability due to mental disorders: A systematic review', *Journal of Occupational Rehabilitation* **21**(2), 259–274.
URL: *https://doi.org/10.1007/s10926-010-9261-5*

Dewa, C. S., Goering, P., Lin, E. and Paterson, M. (2002), 'Depression-related short-term disability in an employed population', *Journal of Occupational and Environmental Medicine* **44**(7).
URL: *http://journals.lww.com/joem/Fulltext/2002/07000/Depression_Related_Short_Term_Disability_in_an.7.aspx*

Donceel, P., Du Bois, M. and Lahaye, D. (1999), 'Return to work after surgery for lumbar disc herniation: A rehabilitation-oriented approach in insurance medicine', *Spine* **24**(9), 872–876.

Dumke, H. A. (2017), 'Posttraumatic headache and its impact on return to work after mild traumatic brain injury', *The Journal of Head Trauma Rehabilitation* **32**(2), E55–E65.
URL: *http://journals.lww.com/headtraumarehab/Fulltext/2017/03000/Posttraumatic_Headache_and_Its_Impact_on_Return_to.16.aspx*

Eggert, S. (2010), 'Psychosocial factors affecting employees' abilities to return to work', *AAOHN Journal* **58**(2), 51–55.

Fritsch, S., Guenther, F., Suling, M. and Mueller, S. M. (2016), 'Package 'neuralnet''.
URL: *https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf*

Hahsler, M., Buchta, C., Gruen, B., Hornik, K., Johnson, I. and Borgelt, C. (2017), 'Package 'arules''.
URL: *https://cran.r-project.org/web/packages/arules/arules.pdf*

Hardy, P., Knight, B. and Edwards, B. (2011), 'The role of incentive measures in workers' compensation schemes', *Institute of Actuaries in Australia, Sydney* pp. 20–22.

Heads Up (n.d.), 'About us', https://www.headsup.org.au/general/about-us.

Hothorn, T., Hornik, K. and Zeileis, A. (2006), 'Unbiased recursive partitioning: A conditional inference framework', *Journal of Computational and Graphical Statistics* **15**(3), 651–674.
URL: *http://dx.doi.org/10.1198/106186006X133933*

Institute for Digital Research and Education (n.d.), 'Faq: What are pseudo r-squareds?'.
URL: *https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/*

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2014), *An Introduction to Statistical Learning: with Applications in R*, Springer Texts in Statistics, Springer New York.
URL: *https://books.google.com.au/books?id=at1bmAEACAAJ*

Johnson, W. G. and Ondrich, J. (1990), 'The duration of post-injury absences from work', *The Review of Economics and Statistics* **72**(4), 578–586.
URL: *http://www.jstor.org/stable/2109597*

Kolyshkina, I., Steinberg, D. and Cardell, N. S. (2003), *Intelligent And Other Computational Techniques In Insurance: Theory and Applications*, Vol. Volume 6 of *Series on Innovative Intelligence*, WORLD SCIENTIFIC, chapter Using Data Mining for Modeling Insurance Risk and Comparison of Data Mining and Linear Modeling Approaches, pp. 493–522. 0.
URL: *https://doi.org/10.1142/9789812794246_0014*

Krause, N., Dasinger, L. K., Deegan, L. J., Rudolph, L. and Brand, R. J. (2001), 'Psychosocial job factors and return-to-work after compensated low back injury: A disability phase-specific analysis', *American Journal of Industrial Medicine* **40**(4), 374–392.
URL: *http://dx.doi.org/10.1002/ajim.1112*

Krause, N., Frank, J. W., Dasinger, L. K., Sullivan, T. J. and Sinclair, S. J. (2001), 'Determinants of duration of disability and return-to-work after work-related injury and illness: Challenges for future research', *American Journal of Industrial Medicine* **40**(4), 464–484.
URL: *http://dx.doi.org/10.1002/ajim.1116*

LaMontagne, A. D., Keegel, T., Louie, A. M. and Ostry, A. (2010), 'Job stress as a preventable upstream determinant of common mental disorders: A review for practitioners and policy-makers', *Advances in Mental Health* **9**(1), 17–35.
URL: *http://dx.doi.org/10.5172/jamh.9.1.17*

Lane, T., Collie, A. and Hassani-Mahmooei, B. (2016), 'Work-related injury and illness in australia, 2004 to 2014', http://www.iscrr.com.au/__data/assets/pdf_file/0020/540830/118_Work-injury-in-Australia-Review-2004-2014.pdf.

Lebedev, I., Kolyshkina, I., Brownlow, M. and Khoo, C. (2015), 'Analytics-assisted triage of workers' compensation claims', http://www.actuaries.asn.au/Library/Events/ACS/2015/LebedevEtAlWorkersComp.pdf.

Lesuffleur, T., Chastang, J.-F., Sandret, N. and Niedhammer, I. (2015), 'Psychosocial factors at work and occupational injury: Results from the french national sumer survey', *Journal of Occupational and Environmental Medicine* **57**(3).
URL: *http://journals.lww.com/joem/Fulltext/2015/03000/Psychosocial_Factors_at_Work_and_Occupational.6.aspx*

Levshina, N. (2015), *How to Do Linguistics with R: Data Exploration and Statistical Analysis*, John Benjamins Publishing Company.
URL: *https://books.google.com.au/books?id=Zmx3rgEACAAJ*

Liaw, A. and Wiener, M. (2002), 'Classification and regression by randomforest', *R news* **2**(3), 18–22.

Liu, Q., Pitt, D. and Wu, X. (2014), 'On the prediction of claim duration for income protection insurance policyholders', *Annals of Actuarial Science* **8**(1), 42–62.

Lu, M.-L., Nakata, A., Park, J. B. and Swanson, N. G. (2014), 'Workplace psychosocial factors associated with work-related injury absence: A study from a nationally representative sample of korean workers', *International Journal of Behavioral Medicine* **21**(1), 42–52.
URL: *http://dx.doi.org/10.1007/s12529-013-9325-y*

MacKenzie, E. J., Morris, J. A., Jurkovich, G. J., Yasui, Y., Cushing, B. M., Burgess, A. R., DeLateur, B. J., McAndrew, M. P. and Swiontkowski, M. F. (1998), 'Return to work following injury: the role of economic, social, and job-related factors.', *American Journal of Public Health* **88**(11), 1630–1637.
URL: *http://dx.doi.org/10.2105/AJPH.88.11.1630*

Medibank Private (2012), 'The cost of workplace stress in australia', http://www.medibank.com.au/Client/Documents/Pdfs/The-Cost-of-Workplace-Stress.pdf.

Nielsen, M. B. D., Bültmann, U., Madsen, I. E., Martin, M., Christensen, U., Diderichsen, F. and Rugulies, R. (2012), 'Health, work, and personal-related predictors of time to return to work among employees with mental health problems', *Disability and Rehabilitation* **34**(15), 1311–1316. PMID: 22200251.
URL: *http://dx.doi.org/10.3109/09638288.2011.641664*

Nielsen, M. B. D., Madsen, I. E. H., Bültmann, U., Christensen, U., Diderichsen, F. and Rugulies, R. (2011), 'Predictors of return to work in employees sick-listed with mental health problems: findings from a longitudinal study', *European Journal of Public Health* **21**(6), 806–811.
URL: + *http://dx.doi.org/10.1093/eurpub/ckq171*

Nieuwenhuijsen, K., Verbeek, J. H. A. M., de Boer, A. G. E. M., Blonk, R. W. B. and van Dijk, F. J. H. (2004), 'Supervisory behaviour as a predictor of return to work in employees absent from work due to mental health problems', *Occupational and Environmental Medicine* **61**(10), 817–823.
URL: *http://oem.bmj.com/content/61/10/817*

Nieuwenhuijsen, K., Verbeek, J. H., de Boer, A. G., Blonk, R. W. and van Dijk, F. J. (2006), 'Predicting the duration of sickness absence for patients with common mental disorders in occupational health care', *Scandinavian Journal of Work, Environment & Health* **32**(1), 67–74.
URL: *http://www.jstor.org/stable/40967545*

PMIS (2017), 'Manufacturers can manage long-tail workers' comp claims', http://www.pmiservices.com/workers-compensation/manufacturers-manage-workers-comp-claims/.

Return To Work South Australia (n.d.), 'Industry classifications and rates', https://www.rtwsa.com/insurance/insurance-with-us/premium-calculations/industry-classifications-and-rates.

Rose, S. (2017), 'Mental health and group insurance: Early intervention works', *Investment Magazine* .
URL: *https://investmentmagazine.com.au/2017/08/mental-health-and-group-insurance-early-intervention-works/*

Safe Work Australia (2008), 'Type of occurrence classification system 3rd edition revision 1', https://www.safeworkaustralia.gov.au/doc/type-occurrence-classification-system-3rd-edition-revision-1.

Safe Work Australia (2009), 'Comparison of workers' compensation arrangements in australia and new zealand', https://www.safeworkaustralia.gov.au/system/files/documents/1702/comparisonworkerscompensationarrangementsaustnz_2008_pdf.pdf.

Safe Work Australia (2013), 'The incidence of accepted workers' compensation claims for mental stress in australia', http://www.safeworkaustralia.gov.au/sites/SWA/about/Publications/Documents/769/The-Incidence-Accepted-WC-Claims-Mental-Stress-Australia.pdf.

Safe Work Australia (2014), 'Workers' compensation legislation and psychological injury - fact sheet', http://www.safeworkaustralia.gov.au/sites/SWA/about/Publications/Documents/852/WC-psychological-injury-fact-sheet.pdf.

Safe Work Australia (2015*a*), 'The cost of work-related injury and illness for australian employers, workers and the community: 2012–13', http://www.safeworkaustralia.gov.au/sites/swa/about/publications/documents/940/cost-of-work-related-injury-and-disease-2012-13.docx.pdf.

Safe Work Australia (2015*b*), 'Work-related mental disorders profile 2015', http://www.safeworkaustralia.gov.au/sites/SWA/about/Publications/Documents/945/work-related-mental-disorders-profile.pdf.

Safe Work Australia (2016), 'Comparison of workers' compensation arrangements in australia and new zealand', http://www.safeworkaustralia.gov.au/sites/SWA/about/Publications/Documents/924/comparison-wc-2015.pdf.

Safe Work Australia (2017), 'Australian workers' compensation statistics 2014–15', http://www.safeworkaustralia.gov.au/sites/swa/statistics/workers-compensation-data/pages/compendiumofworkerscompensationstatistics.

Smith, K. A., Willis, R. J. and Brooks, M. (2000), 'An analysis of customer retention and insurance claim patterns using data mining: A case study', *The Journal of the Operational Research Society* **51**(5), 532–541.
URL: *http://www.jstor.org/stable/254184*

Snook, S. H. and Webster, B. S. (1987), 'The cost of disability.', *Clinical Orthopaedics and Related Research* **221**.
URL: *http://journals.lww.com/corr/Fulltext/1987/08000/The_Cost_of_Disability_.9.aspx*

Stansfeld, S. A., Fuhrer, R., Shipley, M. J. and Marmot, M. G. (1999), 'Work characteristics predict psychiatric disorder: prospective results from the whitehall ii study.', *Occupational and Environmental Medicine* **56**(5), 302–307.
URL: *http://oem.bmj.com/content/56/5/302*

State Insurance Regulation Authority (n.d.), 'Workers compensation guide for employers: when a worker is injured', http://www.sira.nsw.gov.au/resources-library/workers-compensation-resources/publications/workers-compensation-policies/workers-compensation-guide-for-employers.

Tan, P.-N., Steinbach, M. and Kumar, V. (2006), *Introduction to Data Mining*, Boston : Pearson Addison Wesley.

Therneau, T., Atkinson, B. and Ripley, B. (2017), *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-11.
URL: *https://CRAN.R-project.org/package=rpart*

Tufféry, S. (2011), *Data Mining and Statistics for Decision Making*, John Wiley & Sons, Hoboken, NJ.
URL: *http://cds.cern.ch/record/1486168*

Uehli, K., Mehta, A. J., Miedinger, D., Hug, K., Schindler, C., Holsboer-Trachsler, E., Leuppi, J. D. and Künzli, N. (2014), 'Sleep problems and work injuries: A systematic review and meta-analysis', *Sleep Medicine Reviews* **18**(1), 61 – 73.
URL: *http://www.sciencedirect.com/science/article/pii/S1087079213000087*

Yelin, E. (1986), 'The myth of malingering: Why individuals withdraw from work in the presence of illness', *The Milbank Quarterly* **64**(4), 622–649.
URL: *http://www.jstor.org/stable/3349928*

Ziegler, A. and König, I. R. (2014), 'Mining data with random forests: current options for real-world applications', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **4**(1), 55–63.
URL: *http://dx.doi.org/10.1002/widm.1114*

Appendix A

Fields in data

* Field used for analysis
** Modified field, which is not in the original dataset, used for analysis

Indentification and demographic information

CLAIM ID
WORKER ID
RESIDENTIAL POSTCODE FOR INJURED WORKER
AGE OF INJURED WORKER *
GENDER OF INJURED WORKER *
INDICATION FOR REQUIREMENT OF INTERPRETER *
DESCRIPTION FOR OCCUPATION OF INJURED WORKER
MINOR GROUP FOR OCCUPATION OF INJURED WORKER *
MAJOR GROUP FOR OCCUPATION OF INJURED WORKER
INCOME ESTIMATE FOR INJURED WORKER
EMPLOYMENT COUNT FOR RECENT 5 YEARS FOR INJURED WORKER *

Description for injury or disease

DATE OF INJURY
FINANCIAL YEAR OF INJURY
LAG BETWEEN DATE OF INJURY AND REPORTING DATE *
DESCRIPTION FOR AGENCY OF ACCIDENT
MINOR GROUP FOR AGENCY OF ACCIDENT

```
MAJOR GROUP FOR AGENCY OF ACCIDENT *
DESCRIPTION FOR AGENCY OF INJURY
MINOR GROUP FOR AGENCY OF INJURY
MAJOR GROUP FOR AGENCY OF INJURY
DESCRIPTION FOR NATURE OF INJURY *
DESCRIPTION FOR NATURE OF INJURY WITH CLAIMS IN A SMALL DESCRIPTION
   ↪ CATEGORY COMBINED FOR THE GROUP FOR NATURE OF INJURY **
GROUP FOR NATURE OF INJURY
DESCRIPTION FOR MECHANISM OF INJURY
GROUP FOR MECHANISM OF INJURY *
DESCRIPTION FOR BODILY LOCATION OF INJURY
DESCRIPTION FOR BODILY LOCATION OF INJURY WITH CLAIMS IN A SMALL
   ↪ DESCRIPTION CATEGORY COMBINED FOR THE MINOR OR MAJOR GROUP FOR
   ↪ BODILY LOCATION **
MINOR GROUP FOR BODILY LOCATION OF INJURY
MAJOR GROUP FOR BODILY LOCATION OF INJURY
INDICATION FOR PSYCHOLOGICAL INJURY *
INDICATION FOR CALLING AN AMBULANCE *
```

Past claim history

```
COUNT FOR PRIOR WORKERS' COMPENSATION CLAIM MADE BY INJURED WORKER *
COUNT FOR CURRENTLY OPEN CLAIMS MADE BY INJURED WORKER *
TOTAL LUMP SUMS PAYMENTS MADE TO ALL PRIOR CLAIMS *
INDICATION FOR EXISTENCE OF LEGAL COST PAID FOR ANY PRIOR CLAIM *
TOTAL LEGAL COSTS MADE FOR ALL PRIOR CLAIMS
INDICATION FOR EXISTENCE OF INCOME MAINTENANCE PAYMENTS FOR ANY PRIOR
   ↪ CLAIM *
COUNT FOR PRIOR CLAIMS FOR WHICH INCOME MAINTENANCE PAYMENTS ARE MADE *
   ↪
TOTAL INCOME MAINTENANCE PAYMENTS MADE FOR ALL PRIOR CLAIMS *
INDICATION FOR EXISTENCE OF INVESTIGATION COST MADE FOR ANY PRIOR
   ↪ CLAIM *
INDICATION FOR EXISTENCE OF PAYMENT FOR PSYCHOLOGICAL SERVICE FOR ANY
   ↪ PRIOR CLAIM *
```

TOTAL PAYMENTS ON THE MOST RECENT PRIOR CLAIM **
PAYMENT TO HOSPITALS FOR THE MOST RECENT PRIOR CLAIM
INCOME MAINTENANCE PAYMENT FOR THE MOST RECENT PRIOR CLAIM
INVESTIGATION COST FOR THE MOST RECENT PRIOR CLAIM
LEGAL COST FOR THE MOST RECENT PRIOR CLAIM
LUMP SUM PAYMENT FOR THE MOST RECENT PRIOR CLAIM
MEDICAL COST FOR THE MOST RECENT PRIOR CLAIM
COST OF PHYSIOTHERAPY TREATMENT FOR THE MOST RECENT PRIOR CLAIM
REDEMPTION COST FOR THE MOST RECENT PRIOR CLAIM
REHABILITATION COST FOR THE MOST RECENT PRIOR CLAIM
COST OF TRAVEL FOR THE MOST RECENT PRIOR CLAIM
OTHER COST FOR THE MOST RECENT PRIOR CLAIM
AVERAGE DURATION OF INCOME MAINTENANCE PAYMENTS OF ALL PRIOR CLAIMS *
INDICATION FOR HAVING CURRENT CLAIM FOR AN INJURY SAME AS THE MOST
↪ RECENT PRIOR CLAIM *
INDICATION FOR HAVING CURRENT CLAIM FOR AN INJURY SIMILAR TO THE MOST
↪ RECENT PRIOR CLAIM *
INDICATION FOR HAVING CURRENT CLAIM FOR A SEQUELAE FROM THE MOST
↪ RECENT PRIOR CLAIM *
INDICATION FOR HAVING SAME EMPLOYER FOR THE MOST RECENT PRIOR CLAIM *
DESCRIPTION FOR AGENCY OF ACCIDENT FOR THE MOST RECENT PRIOR CLAIM
MINOR GROUP FOR AGENCY OF ACCIDENT FOR THE MOST RECENT PRIOR CLAIM
MAJOR GROUP FOR AGENCY OF ACCIDENT FOR THE MOST RECENT PRIOR CLAIM *
DESCRIPTION FOR AGENCY OF INJURY FOR THE MOST RECENT PRIOR CLAIM
MINOR GROUP FOR AGENCY OF INJURY FOR THE MOST RECENT PRIOR CLAIM
MAJOR GROUP FOR AGENCY OF INJURY FOR THE MOST RECENT PRIOR CLAIM
DESCRIPTION FOR NATURE OF INJURY FOR THE MOST RECENT PRIOR CLAIM
GROUP FOR NATURE OF INJURY FOR THE MOST RECENT PRIOR CLAIM *
DESCRIPTION FOR MECHANISM OF INJURY FOR THE MOST RECENT PRIOR CLAIM
GROUP FOR MECHANISM OF INJURY FOR THE MOST RECENT PRIOR CLAIM *
DESCRIPTION FOR BODILY LOCATION OF INJURY FOR THE MOST RECENT PRIOR
↪ CLAIM
MINOR GROUP FOR BODILY LOCATION OF INJURY FOR THE MOST RECENT PRIOR
↪ CLAIM *

```
MAJOR GROUP FOR BODILY LOCATION OF INJURY FOR THE MOST RECENT PRIOR
   ↪ CLAIM
INDICATION FOR PSYCHOLOGICAL INJURY FOR THE MOST RECENT PRIOR CLAIM *
COUNT OF MEDICAL CERTIFICATES FOR THE MOST RECENT PRIOR CLAIM *
COUNT OF HOSPITAL VISITS FOR THE MOST RECENT PRIOR CLAIM *
COUNT OF DISTINCT PROVIDER VISITS FOR THE MOST RECENT PRIOR CLAIM *
COUNT OF VISITS TO PHYSIOTHERAPIST FOR THE MOST RECENT PRIOR CLAIM *
DESCRIPTION FOR TYPE OF VISITS TO PREDOMINANT SPECIALIST FOR THE MOST
   ↪ RECENT PRIOR CLAIM
COUNT OF VISITS TO CHEMIST FOR THE MOST RECENT PRIOR CLAIM
COST OF DRUG FOR THE MOST RECENT PRIOR CLAIM
COST OF POTENT OPIOID FOR THE MOST RECENT PRIOR CLAIM
INDICATION FOR USE OF MEDICATION FOR THE MOST RECENT PRIOR CLAIM
DURATION OF INCOME MAINTENANCE PAYMENTS FOR THE MOST RECENT PRIOR
   ↪ CLAIM *
POSTCODE OF PREDOMINANT CLINIC FOR THE MOST RECENT PRIOR CLAIM
INDICATION FOR NOT HAVING ANY PRIOR CLAIM **
```

Employer-related information

```
EMPLOYER SIZE CATEGORY IN THE FINANCIAL YEAR OF INJURY *
SOUTH AUSTRALIAN INDUSTRY CLASSIFICATIONS (SAIC) FOR EMPLOYER OF
   ↪ INJURED WORKER *
PREMIUM RATE FOR WORKERS' COMPENSATION * INSURANCE FOR EMPLOYER OF
   ↪ INJURED WORKER
```

Duration of claim

```
INDICATION FOR DURATION LONGER THAN TWO WEEKS *
DURATION OF INCOME MAINTENANCE PAYMENTS *
```

Appendix B

Detailed specification of trees

B.1 Specification of rpart classification tree on all claims from R

```
n= 104667

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 104667 25233 0 (0.7589211 0.2410789)
    2) Nature of Injury Category=INTRACRANIAL INJURIES; INTERNAL INJURY
       ↪  OF CHEST, ABDOMEN AND PELVIS; TRAUMATIC AMPUTATION; MEDICAL
       ↪  SHARP/NEEDLE-STICK PUNCTURE; SUPERFICIAL INJURY; INJURY TO
       ↪ NERVES AND SPINAL CORD; TRAUMATIC JOINT/LIGAMENT AND MUSCLE/
       ↪ TENDON INJURY EXCEPT DISLOCATION; INJURY TO MAJOR BLOOD
       ↪ VESSEL; LACERATION OR OPEN WOUND NOT INVOLVING TRAUMATIC
       ↪ AMPUTATION; CONTUSION; BRUISING AND SUPERFICIAL CRUSHING;
       ↪ BURN; TRAUMA TO JOINTS AND LIGAMENTS; TRAUMA TO MUSCLES AND
       ↪ TENDONS; SOFT TISSUE INJURIES DUE TO TRAUMA OR UNKNOWN
       ↪ MECHANISMS WITH INSUFFICIENT; OTHER INJURIES; DISEASES OF
       ↪ THE MUSCLE; TENDON AND RELATED TISSUE EXCEPT TENDINITIS AND
       ↪ EPICONDYLITIS; OTHER MENTAL DISEASES NOT ELSEWHERE
       ↪ CLASSIFIED; MENTAL DISEASES UNSPECIFIED; DIGESTIVE SYSTEM
       ↪ DISEASES; SKIN AND SUBCUTANEOUS TISSUE DISEASES; NERVOUS
```

```
    ↪ SYSTEM AND SENSE ORGAN DISEASES; RESPIRATORY SYSTEM DISEASES;
    ↪  OTHER DISEASES; OTHER CLAIMS; NOT KNOWN 74157 12366 0
    ↪ (0.8332457 0.1667543) *
  3) Nature of Injury Category=FRACTURES; DISLOCATION; JOINT DISEASES
    ↪  (ARTHROPATHIES) AND OTHER ARTICULAR CARTILAGE DISEASES;
    ↪ SPINAL VERTEBRAE AND INTERVERTEBRAL DISC DISEASES -
    ↪ DORSOPATHIES; DISEASES INVOLVING THE SYNOVIUM AND RELATED
    ↪ TISSUE; DISEASES OF THE MUSCLE; TENDON AND RELATED TISSUE;
    ↪ OTHER SOFT TISSUE DISEASES; OTHER MUSCULOSKELETAL AND
    ↪ CONNECTIVE TISSUE DISEASES, NOT ELSEWHERE CLASSIFIED;
    ↪ TENDINITIS AND EPICONDYLITIS; BURSITIS; OCCUPATIONAL OVERUSE
    ↪  SYNDROME; FIBROMYALGIA; FIBROSITIS AND MYALGIA; POST-
    ↪ TRAUMATIC STRESS DISORDER; ANXIETY/STRESS DISORDER;
    ↪ DEPRESSION; ANXIETY/DEPRESSION COMBINED; SHORT TERM SHOCK
    ↪ FROM EXPOSURE TO DISTURBING CIRCUMSTANCES; REACTION TO
    ↪ STRESSORS - OTHER, MULTIPLE OR NOT SPECIFIED; CIRCULATORY
    ↪ SYSTEM DISEASES; INFECTIOUS AND PARASITIC DISEASES;
    ↪ NEOPLASMS (CANCER) 30510 12867 0 (0.5782694 0.4217306)
   6) Nature of Injury Category =DISLOCATION; DISEASES INVOLVING THE
     ↪  SYNOVIUM AND RELATED TISSUE; DISEASES OF THE MUSCLE,
     ↪ TENDON AND RELATED TISSUE; OTHER SOFT TISSUE DISEASES;
     ↪ OTHER MUSCULOSKELETAL AND CONNECTIVE TISSUE DISEASES NOT
     ↪ ELSEWHERE CLASSIFIED; TENDINITIS AND EPICONDYLITIS;
     ↪ CIRCULATORY SYSTEM DISEASES; INFECTIOUS AND PARASITIC
     ↪ DISEASES; NEOPLASMS (CANCER) 12273 3975 0 (0.6761183
     ↪ 0.3238817)
    12) Bodily Location Category=HEAD EXCEPT EYE; EYE; NECK; TRUNK
       ↪ EXCEPT BACK - UPPER AND LOWER; UPPER BACK; BACK - OTHER
       ↪ AND MULTIPLE AND BACK - UNSPECIFIED; LOWER BACK; UPPER
       ↪ LIMBS EXCEPT SHOULDER, ELBOW, WRIST, HAND, FINGERS AND
       ↪ THUMB; ELBOW; WRIST; HAND; HAND, FINGERS AND THUMB - OTHER
       ↪  AND MULTIPLE; HAND, FINGERS AND THUMB - UNSPECIFIED;
       ↪ FINGERS; THUMB; LOWER LIMBS EXCEPT KNEE, ANKLE, FOOT AND
       ↪ TOES; FOOT AND TOES; MULTIPLE LOCATIONS; SYSTEMIC
```

```
    ↪ LOCATIONS; PSYCHOLOGICAL SYSTEM IN GENERAL; UNSPECIFIED
    ↪ LOCATIONS; NOT KNOWN; 9036 2332 0 (0.7419212 0.2580788) *
  13) Bodily Location Category =SHOULDER; KNEE; ANKLE 3237 1594 1
    ↪ (0.4924313 0.5075687)
    26) Occupation Category=10 MANAGERS,11 Chief Executives,
      ↪ General Managers and Legislators,12 Farmers and Farm
      ↪ Managers, 13 Specialist Managers,14 Hospitality, Retail
      ↪ and Service Managers,21 Arts and Media Professionals, 22
      ↪  Business, Human Resource and Marketing Professionals,23
      ↪  Design, Engineering, Science and Transport
      ↪ Professionals, 24 Education Professionals,25 Health
      ↪ Professionals,26 ICT Professionals,27 Legal, Social and
      ↪ Welfare Professionals, 31 Engineering, ICT and Science
      ↪ Technicians,32 Automotive and Engineering Trades Workers
      ↪ ,35 Food Trades Workers, 39 Other Technicians and Trades
      ↪  Workers,45 Sports and Personal Service Workers,51
      ↪ Office Managers and Program Administrators, 52 Personal
      ↪ Assistants and Secretaries,53 General Clerical Workers
      ↪ ,54 Inquiry Clerks and Receptionists,55 Numerical Clerks
      ↪ , 56 Clerical and Office Support Workers,59 Other
      ↪ Clerical and Administrative Workers,61 Sales
      ↪ Representatives and Agents, 62 Sales Assistants and
      ↪ Salespersons,71 Machine and Stationary Pland Operators,
      ↪ 80 Labourers, 83 Factory Process Workers, NOT KNOWN 1568
      ↪  650 0 (0.585459 0.414541)
    27) Occupation Category=30 TECHNICIANS AND TRADES WORKERS,33
      ↪ Construction Trades Workers,34 Electrotechnology and
      ↪ Telecommunications Trades Workers, 36 Skilled Animal and
      ↪  Horticultural Workers,40 COMMUNITY AND PERSONAL SERVICE
      ↪  WORKERS,41 Health and Welfare Support Workers,42 Carers
      ↪  and Aides, 43 Hospitality Workers,44 Protective Service
      ↪  Workers,63 Sales Support Workers,70 MACHINERY OPERATORS
      ↪  AND DRIVERS, 72 Mobile Plant Operators,73 Road and Rail
      ↪  Drivers,81 Cleaners and Laundry Workers,82 Construction
      ↪  and Mining Labourers, 84 Farm, Forestry and Garden
```

```
        ↪ Workers,85 Food Preparation Assistants,89 Other
        ↪ Labourers 1669 676 1 (0.4050330 0.5949670) *
    7) Nature of Injury Category=FRACTURES; JOINT DISEASES (
       ↪ ARTHROPATHIES) AND OTHER ARTICULAR CARTILAGE DISEASES;
       ↪ SPINAL VERTEBRAE AND INTERVERTEBRAL DISC DISEASES -
       ↪ DORSOPATHIES; BURSITIS; OCCUPATIONAL OVERUSE SYNDROME;
       ↪ FIBROMYALGIA; FIBROSITIS AND MYALGIA; POST-TRAUMATIC STRESS
       ↪  DISORDER; ANXIETY/STRESS DISORDER; DEPRESSION; ANXIETY/
       ↪ DEPRESSION COMBINED; SHORT TERM SHOCK FROM EXPOSURE TO
       ↪ DISTURBING CIRCUMSTANCES; REACTION TO STRESSORS ? OTHER,
       ↪ MULTIPLE OR NOT SPECIFIED 18237 8892 0 (0.5124198
       ↪ 0.4875802)
     14) Bodily Location Category=HEAD EXCEPT EYE; EYE; FINGERS;
        ↪ THUMB; UNSPECIFIED LOCATIONS 511 0 (0.7544450 0.2455550) *
     15) Bodily Location Category=NECK; TRUNK EXCEPT BACK - UPPER AND
        ↪  LOWER, UPPER BACK, BACK - OTHER AND MULTIPLE AND BACK -
        ↪ UNSPECIFIED; LOWER BACK; UPPER LIMBS EXCEPT SHOULDER,
        ↪ ELBOW, WRIST, HAND, FINGERS AND THUMB; SHOULDER; ELBOW;
        ↪ WRIST; HAND; HAND, FINGERS AND THUMB - OTHER AND MULTIPLE;
        ↪  HAND, FINGERS AND THUMB - UNSPECIFIED; LOWER LIMBS EXCEPT
        ↪  KNEE, ANKLE, FOOT AND TOES; KNEE; ANKLE; FOOT AND TOES;
        ↪ MULTIPLE LOCATIONS; PSYCHOLOGICAL SYSTEM IN GENERAL 7775 1
        ↪  (0.4812454 0.5187546)
      30) Nature of Injury Category=JOINT DISEASES (ARTHROPATHIES)
         ↪ AND OTHER ARTICULAR CARTILAGE DISEASES; SPINAL VERTEBRAE
         ↪  AND INTERVERTEBRAL DISC DISEASES - DORSOPATHIES;
         ↪ BURSITIS; OCCUPATIONAL OVERUSE SYNDROME; FIBROMYALGIA;
         ↪ FIBROSITIS AND MYALGIA 8945 4195 0 (0.5310229 0.4689771)
        60) Occupation Category=12 Farmers and Farm Managers,13
           ↪ Specialist Managers,14 Hospitality, Retail and Service
           ↪  Managers,21 Arts and Media Professionals,22 Business,
           ↪  Human Resource and Marketing Professionals,23 Design,
           ↪  Engineering, Science and Transport Professionals,24
           ↪ Education Professionals,26 ICT Professionals,27 Legal,
           ↪  Social and Welfare Professionals,31 Engineering, ICT
```

```
      ↪ and Science Technicians,32 Automotive and Engineering
      ↪ Trades Workers,34 Electrotechnology and
      ↪ Telecommunications Trades Workers,39 Other Technicians
      ↪  and Trades Workers,50 CLERICAL AND ADMINISTRATIVE
      ↪ WORKERS,51 Office Managers and Program Administrators
      ↪ ,52 Personal Assistants and Secretaries,53 General
      ↪ Clerical Workers,54 Inquiry Clerks and Receptionists
      ↪ ,55 Numerical Clerks,59 Other Clerical and
      ↪ Administrative Workers,61 Sales Representatives and
      ↪ Agents,62 Sales Assistants and Salespersons,71 Machine
      ↪  and Stationary Plant Operators,74 Storepersons,83
      ↪ Factory Process Workers,84 Farm, Forestry and Garden
      ↪ Workers 4437 1783 0 (0.598152 0.401848)
    61) Occupation Category=11 Chief Executives, General
      ↪ Managers and Legislators,25 Health Professionals,33
      ↪ Construction Trades Workers,35 Food Trades Workers,36
      ↪ Skilled Animal and Horticultural Workers,40 COMMUNITY
      ↪ AND PERSONAL SERVICE WORKERS,41 Health and Welfare
      ↪ Support Workers,42 Carers and Aides,43 Hospitality
      ↪ Workers,44 Protective Service Workers,45 Sports and
      ↪ Personal Service Workers,56 Clerical and Office
      ↪ Support Workers,63 Sales Support Workers,70 MACHINERY
      ↪ OPERATORS AND DRIVERS,72 Mobile Plant Operators,73
      ↪ Road and Rail Drivers,80 LABOURERS,81 Cleaners and
      ↪ Laundry Workers,82 Construction and Mining Labourers
      ↪ ,85 Food Preparation Assistants,89 Other Labourers,NOT
      ↪  KNOWN 4508 2096 1 (0.4649512 0.5350488) *
  31) Nature of Injury Category=FRACTURES; POST-TRAUMATIC STRESS
    ↪  DISORDER; ANXIETY/STRESS DISORDER; DEPRESSION; ANXIETY/
    ↪ DEPRESSION COMBINED; SHORT TERM SHOCK FROM EXPOSURE TO
    ↪ DISTURBING CIRCUMSTANCES; REACTION TO STRESSORS ? OTHER,
    ↪  MULTIPLE OR NOT SPECIFIED 7211 3025 1 (0.4194980
    ↪ 0.5805020)
```

```
62) Bodily Location Category=TRUNK EXCEPT BACK - UPPER AND
    ↪ LOWER; ELBOW; HAND, FOOT AND TOES; PSYCHOLOGICAL
    ↪ SYSTEM IN GENERAL 5134 2371 1 (0.4618231 0.5381769)
 124) Occupation Category=21 Arts and Media Professionals,23
    ↪ Design, Engineering, Science and Transport
    ↪ Professionals,24 Education Professionals,31
    ↪ Engineering, ICT and Science Technicians,32
    ↪ Automotive and Engineering Trades Workers,33
    ↪ Construction Trades Workers,34 Electrotechnology and
    ↪ Telecommunications Trades Workers,36 Skilled Animal
    ↪ and Horticultural Workers,39 Other Technicians and
    ↪ Trades Workers,40 COMMUNITY AND PERSONAL SERVICE
    ↪ WORKERS,56 Clerical and Office Support Workers,59
    ↪ Other Clerical and Administrative Workers,61 Sales
    ↪ Representatives and Agents,62 Sales Assistants and
    ↪ Salespersons,63 Sales Support Workers,70 MACHINERY
    ↪ OPERATORS AND DRIVERS,71 Machine and Stationary Plant
    ↪  Operators,72 Mobile Plant Operators,73 Road and Rail
    ↪  Drivers,80 LABOURERS,81 Cleaners and Laundry Workers
    ↪ ,83 Factory Process Workers 2392 1149 0 (0.5196488
    ↪ 0.4803512)
  248) Prior Claim Bodily Location Category=11 CRANIUM,12
      ↪ EYE,16 FACE, NOT ELSEWHERE SPECIFIED,21 NECK,31 BACK
      ↪  - UPPER OR LOWER,34/35 ABDOMEN AND PELVIC REGION
      ↪ ,38 TRUNK - MULTIPLE LOCATIONS,41 SHOULDER,42 UPPER
      ↪  ARM,43 ELBOW,45 WRIST,46 HAND, FINGERS AND THUMB,51
      ↪  HIP,53 KNEE,55 ANKLE,56 FOOT AND TOES,59 LOWER
      ↪ LIMB - UNSPECIFIED LOCATIONS,61 NECK AND TRUNK,68
      ↪ OTHER SPECIFIED MULTIPLE LOCATIONS,71 CIRCULATORY
      ↪ SYSTEM,80 PSYCHOLOGICAL SYSTEM,NOT KNOWN 1296 561 0
      ↪  (0.5671296 0.4328704) *
  249) Prior Claim Bodily Location Category=13 EAR,14 MOUTH
      ↪ ,15 NOSE,18 HEAD- MULTIPLE LOCATIONS,33 CHEST (
      ↪ THORAX),44 FOREARM,52 UPPER LEG,54 LOWER LEG,58
      ↪ LOWER LIMB - MULTIPLE LOCATIONS,62 HEAD AND NECK,63
```

```
      ↪  HEAD AND OTHER,64 TRUNK AND LIMBS,65 UPPER AND
      ↪ LOWER LIMBS,66 NECK AND SHOULDER,69 UNSPECIFIED
      ↪ MULTIPLE LOCATIONS,72 RESPIRATORY SYSTEM,78 OTHER
      ↪ AND MULTIPLE SYSTEMIC CONDITIONS,90 UNSPECIFIED
      ↪ LOCATIONS,NewClaim 1096 508 1 (0.4635036 0.5364964)
      ↪  *
  125) Occupation Cetegory=11 Chief Executives, General
     ↪ Managers and Legislators,12 Farmers and Farm Managers
     ↪ ,13 Specialist Managers,14 Hospitality, Retail and
     ↪ Service Managers,22 Business, Human Resource and
     ↪ Marketing Professionals,25 Health Professionals,26
     ↪ ICT Professionals,27 Legal, Social and Welfare
     ↪ Professionals,30 TECHNICIANS AND TRADES WORKERS,35
     ↪ Food Trades Workers,41 Health and Welfare Support
     ↪ Workers,42 Carers and Aides,43 Hospitality Workers,44
     ↪  Protective Service Workers,45 Sports and Personal
     ↪ Service Workers,50 CLERICAL AND ADMINISTRATIVE
     ↪ WORKERS,51 Office Managers and Program Administrators
     ↪ ,52 Personal Assistants and Secretaries,53 General
     ↪ Clerical Workers,54 Inquiry Clerks and Receptionists
     ↪ ,55 Numerical Clerks,74 Storepersons,82 Construction
     ↪ and Mining Labourers,84 Farm, Forestry and Garden
     ↪ Workers,85 Food Preparation Assistants,89 Other
     ↪ Labourers,NOT KNOWN 2742 1128 1 (0.4113786 0.5886214)
     ↪  *
   63) Bodily Location Category=NECK; UPPER BACK; BACK - OTHER
     ↪ AND MULTIPLE AND BACK - UNSPECIFIED; LOWER BACK; UPPER
     ↪  LIMBS EXCEPT SHOULDER, ELBOW, WRIST, HAND, FINGERS
     ↪ AND THUMB; SHOULDER; WRIST; HAND, FINGERS AND THUMB -
     ↪ OTHER AND MULTIPLE; HAND, FINGERS AND THUMB -
     ↪ UNSPECIFIED; LOWER LIMBS EXCEPT KNEE, ANKLE, FOOT AND
     ↪ TOES; KNEE; ANKLE; MULTIPLE LOCATIONS 654 1 (0.3148772
     ↪  0.6851228) *
```

B.2 Specification of rpart classification tree on psychological claims from R

```
n= 3537

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 3537 1562 1 (0.4416172 0.5583828)
   2) Occupation Category=10 MANAGERS,20 PROFESSIONALS,21 Arts and
      ↪ Media Professionals,23 Design, Engineering, Science and
      ↪ Transport Professionals,25 Health Professionals,32 Automotive
      ↪  and Engineering Trades Workers,33 Construction Trades
      ↪ Workers,35 Food Trades Workers,56 Clerical and Office Support
      ↪  Workers,63 Sales Support Workers,71 Machine and Stationary
      ↪ Plant Operators,72 Mobile Plant Operators,73 Road and Rail
      ↪ Drivers,81 Cleaners and Laundry Workers,82 Construction and
      ↪ Mining Labourers,83 Factory Process Workers,85 Food
      ↪ Preparation Assistants 1066 472 0 (0.5572233 0.4427767)
     4) Prior Claim Bodily Location Category=11 CRANIUM,13 EAR,14 MOUTH
        ↪ ,18 HEAD- MULTIPLE LOCATIONS,21 NECK,38 TRUNK - MULTIPLE
        ↪ LOCATIONS,41 SHOULDER,42 UPPER ARM,45 WRIST,51 HIP,52 UPPER
        ↪ LEG,53 KNEE,59 LOWER LIMB - UNSPECIFIED LOCATIONS,61 NECK
        ↪ AND TRUNK,63 HEAD AND OTHER,64 TRUNK AND LIMBS,66 NECK AND
        ↪ SHOULDER,68 OTHER SPECIFIED MULTIPLE LOCATIONS,71
        ↪ CIRCULATORY SYSTEM,80 PSYCHOLOGICAL SYSTEM 311 107 0
        ↪ (0.6559486 0.3440514) *
     5) Prior Claim Bodily Location Category =12 EYE,16 FACE, NOT
        ↪ ELSEWHERE SPECIFIED,31 BACK - UPPER OR LOWER,33 CHEST (
        ↪ THORAX),34/35 ABDOMEN AND PELVIC REGION,43 ELBOW,44 FOREARM
        ↪ ,46 HAND, FINGERS AND THUMB,54 LOWER LEG,55 ANKLE,56 FOOT
        ↪ AND TOES,58 LOWER LIMB - MULTIPLE LOCATIONS,62 HEAD AND NECK
        ↪ ,65 UPPER AND LOWER LIMBS,72 RESPIRATORY SYSTEM,78 OTHER AND
        ↪  MULTIPLE SYSTEMIC CONDITIONS,NewClaim,NOT KNOWN 755 365 0
        ↪ (0.5165563 0.4834437)
```

```
    10) Age of injured worker< 36.5 311 125 0 (0.5980707 0.4019293) *
    11) Age of injured worker >=36.5 444 204 1 (0.4594595 0.5405405)
      ↪ *
  3) Occupation Category=11 Chief Executives, General Managers and
    ↪ Legislators,12 Farmers and Farm Managers,13 Specialist
    ↪ Managers,14 Hospitality, Retail and Service Managers,22
    ↪ Business, Human Resource and Marketing Professionals,24
    ↪ Education Professionals,26 ICT Professionals,27 Legal, Social
    ↪  and Welfare Professionals,31 Engineering, ICT and Science
    ↪ Technicians,34 Electrotechnology and Telecommunications
    ↪ Trades Workers,36 Skilled Animal and Horticultural Workers,39
    ↪  Other Technicians and Trades Workers,40 COMMUNITY AND
    ↪ PERSONAL SERVICE WORKERS,41 Health and Welfare Support
    ↪ Workers,42 Carers and Aides,43 Hospitality Workers,44
    ↪ Protective Service Workers,45 Sports and Personal Service
    ↪ Workers,51 Office Managers and Program Administrators,52
    ↪ Personal Assistants and Secretaries,53 General Clerical
    ↪ Workers,54 Inquiry Clerks and Receptionists,55 Numerical
    ↪ Clerks,59 Other Clerical and Administrative Workers,61 Sales
    ↪ Representatives and Agents,62 Sales Assistants and
    ↪ Salespersons,70 Machine and Stationary Plant Operators,74
    ↪ Storepersons,80 LABOURERS,84 Farm, Forestry and Garden
    ↪ Workers, Other Labourers, NOT KNOWN 2471 968 1 (0.391744
    ↪ 0.608256)
```

Appendix C

Examples of calculations of accuracy, lift and accuracy of probability

A confusion matrix for the classification tree on all claims is obtained by applying the tree obtained from a training dataset to a test dataset, and shown in Table C.2

Table C.1 Components of confusion matrix for classification tree on all claims

Model	True negatives	False negatives	False positives	True positives
Classification tree - All Claims	74069	18495	5353	6749

Table C.2 Summary of evaluations for classification tree on all claims

Model	Prediction accuracy	Sensitivity	Specificity	Lift for prolonged duration	Lift for non-prolonged duration	Psuedo R^2	Accuracy of probability
Classification tree - All Claims	0.7722	0.2674	0.9326	2.3122	1.0545	10.2785	0.9899

The measures are calculated based on the four components of the confusion matrices in Table C.1

$$\text{Prediction accuracy} = \frac{74069+6749}{74069+18495+5353+6749} = 0.7722$$

$$\text{Sensitivity} = \frac{6749}{18495+6749} = 0.2674$$

$$\text{Specificity} = \frac{74069}{74069+5353} = 0.9326$$

$$\text{Lift for prolonged duration} = \frac{\frac{6749}{5353+6749}}{\frac{18495+6749}{74069+18495+5353+6749}} = 2.3122$$

$$\text{Lift for non-prolonged duration} = \frac{\frac{74069}{74069+18495}}{\frac{74069+5353}{74069+18495+5353+6749}} = 1.0545$$

Accuracy of probability is calculated based on Model deviation (the deviation between predicted probabilities from the model; and actual probabilities calculated when the model is applied to a test dataset) and Reference deviation (deviation between the proportion of claims having a prolonged duration in the test dataset and the actual probabilities). The Model deviation and Reference deviation are calculated by using values in Table C.3 showing actual probabilities from each region of the classification tree, when the tree is applied to the test dataset.

$$\text{Model deviation} = \sum_{i=1}^{11} (b_i - a_i)^2 \times n = 20.0223$$

$$\text{Reference deviation} = \sum_{i=1}^{11} (b_i - c)^2 \times n = 1988.042$$

$$\text{Accuracy of probability} = 1 - \frac{18.3058}{2317.0940} = 0.9899$$

Table C.3 Predicted and actual probabilities in test dataset

Region* (i)	Predicted probability (a)	Actual probability (b)	Reference probability (c)	Number of claims (n)
1	0.1668	0.1666	0.2412	74332
2	0.2456	0.2461	0.2412	2048
3	0.2581	0.2622	0.2412	8883
4	0.4018	0.4306	0.2412	4440
5	0.4145	0.4514	0.2412	1544
6	0.4329	0.5065	0.2412	1317
7	0.535	0.5051	0.2412	4512
8	0.5365	0.5229	0.2412	1157
9	0.5886	0.5741	0.2412	2761
10	0.595	0.566	0.2412	1659
11	0.6851	0.6662	0.2412	2013

The actual probability is the actual proportion of claims with a prolonged duration within the claims in each region of the test dataset by predicted probability. The reference probability is the proportion of claims with a prolonged duration in the test dataset.
* Regions created by the classification tree

Appendix D

Final triage model for all claims

Probability of prolonged duration, average duration and size of each segment are calculated by applying the rules, that are obtained from the training dataset and tested on the test dataset, to the whole dataset.

D.1 Segment 1

Probability of prolonged duration: 17%
Average duration: 59 days
Size of the segment: 71% of claims

Claims in Segment 1

A claim with one of the following Nature of injury.

```
INTRACRANIAL INJURIES
INJURY TO NERVES AND SPINAL CORD
TRAUMATIC JOINT/LIGAMENT AND MUSCLE/TENDON INJURY EXCEPT DISLOCATION
WOUNDS, LACERATIONS, AMPUTATIONS AND INTERNAL ORGAN DAMAGE
BURN
DISEASES OF THE MUSCLE, TENDON AND RELATED TISSUE EXCEPT TENDINITIS
    ↪ AND EPICONDYLITIS
OTHER MENTAL DISEASES NOT ELSEWHERE CLASSIFIED
MENTAL DISEASES UNSPECIFIED
DIGESTIVE SYSTEM DISEASES
SKIN AND SUBCUTANEOUS TISSUE DISEASES
```

NERVOUS SYSTEM AND SENSE ORGAN DISEASES
RESPIRATORY SYSTEM DISEASES
OTHER INJURIES AND DISEASES
NOT KNOWN

D.2 Segment 2

Probability of prolonged duration: 26%
Average duration: 125 days
Size of the segment: 9% of claims

Claims in Segment 2

A claim with one of the following Nature of injury (Nature of injury Group A);

FRACTURES
DISLOCATION
DISEASES INVOLVING THE SYNOVIUM AND RELATED TISSUE
DISEASES OF THE MUSCLE, TENDON AND RELATED TISSUE
OTHER SOFT TISSUE DISEASES
OTHER MUSCULOSKELETAL AND CONNECTIVE TISSUE DISEASES NOT ELSEWHERE CLASSIFIED
TENDINITIS AND EPICONDYLITIS
CIRCULATORY SYSTEM DISEASES
INFECTIOUS AND PARASITIC DISEASES
NEOPLASMS (CANCER)

AND with one of the following Bodily Locations.

HEAD
NECK
TRUNK
UPPER LIMBS EXCEPT SHOULDER
LOWER LIMBS EXCEPT KNEE AND ANKLE
MULTIPLE LOCATIONS
SYSTEMIC LOCATIONS
PSYCHOLOGICAL SYSTEM IN GENERAL
UNSPECIFIED LOCATIONS
NOT KNOWN

D.3 Segment 3

Probability of prolonged duration: 43%
Average duration: 178 days
Size of the segment: 1% of claims

Claims in Segment 3

A claim with Nature of injury in Nature of injury Group A;
AND with one of the following Bodily Locations;

SHOULDER
KNEE
ANKLE

AND with one of the following Occupation.

MANAGERS
CHIEF EXECUTIVES, GENERAL MANAGERS AND LEGISLATORS
FARMERS AND FARM MANAGERS
SPECIALIST MANAGERS
HOSPITALITY, RETAIL AND SERVICE MANAGERS
PROFESSIONALS
ARTS AND MEDIA PROFESSIONALS
BUSINESS, HUMAN RESOURCE AND MARKETING PROFESSIONALS
DESIGN, ENGINEERING, SCIENCE AND TRANSPORT PROFESSIONALS
EDUCATION PROFESSIONALS
HEALTH PROFESSIONALS
ICT PROFESSIONALS
LEGAL, SOCIAL AND WELFARE PROFESSIONALS
ENGINEERING, ICT AND SCIENCE TECHNICIANS
AUTOMOTIVE AND ENGINEERING TRADES WORKERS
FOOD TRADES WORKERS
OTHER TECHNICIANS AND TRADES WORKERS
SPORTS AND PERSONAL SERVICE WORKERS
CLERICAL AND ADMINISTRATIVE WORKERS
OFFICE MANAGERS AND PROGRAM ADMINISTRATORS
PERSONAL ASSISTANTS AND SECRETARIES

GENERAL CLERICAL WORKERS
INQUIRY CLERKS AND RECEPTIONISTS
NUMERICAL CLERKS
CLERICAL AND OFFICE SUPPORT WORKERS
OTHER CLERICAL AND ADMINISTRATIVE WORKERS
SALES WORKERS
SALES REPRESENTATIVES AND AGENTS
SALES ASSISTANTS AND SALESPERSONS
MACHINE AND STATIONARY PLANT OPERATORS
STOREPERSONS
LABOURERS
FACTORY PROCESS WORKERS
FOOD PREPARATION ASSISTANTS
NOT KNOWN

D.4 Segment 4

Probability of prolonged duration: 58%
Average duration: 259 days
Size of the segment: 2% of claims

Claims in Segment 4

A claim with Nature of injury in Nature of injury Segment 1;
AND with one of the following Bodily Locations;

SHOULDER
KNEE
ANKLE

AND with one of the following Occupation.

TECHNICIANS AND TRADES WORKERS
CONSTRUCTION TRADES WORKERS
ELECTROTECHNOLOGY AND TELECOMMUNICATIONS TRADES WORKERS
SKILLED ANIMAL AND HORTICULTURAL WORKERS
COMMUNITY AND PERSONAL SERVICE WORKERS

HEALTH AND WELFARE SUPPORT WORKERS
CARERS AND AIDES
HOSPITALITY WORKERS
PROTECTIVE SERVICE WORKERS
SALES SUPPORT WORKERS
MACHINERY OPERATORS AND DRIVERS
MOBILE PLANT OPERATORS
ROAD AND RAIL DRIVERS
CLEANERS AND LAUNDRY WORKERS
CONSTRUCTION AND MINING LABOURERS
FARM, FORESTRY AND GARDEN WORKERS
OTHER LABOURERS

D.5 Segment 5

Probability of prolonged duration: 25%
Average duration: 51 days
Size of the segment: 2% of claims

Claims in Segment 5

A claim with one of the following Nature of injury;

FRACTURES
JOINT DISEASES (ARTHROPATHIES) AND OTHER ARTICULAR CARTILAGE DISEASES
SPINAL VERTEBRAE AND INTERVERTEBRAL DISC DISEASES - DORSOPATHIES
BURSITIS
OCCUPATIONAL OVERUSE SYNDROME
FIBROMYALGIA
FIBROSITIS AND MYALGIA
POST-TRAUMATIC STRESS DISORDER
ANXIETY/STRESS DISORDER; DEPRESSION
ANXIETY/DEPRESSION COMBINED
SHORT TERM SHOCK FROM EXPOSURE TO DISTURBING CIRCUMSTANCES
REACTION TO STRESSORS - OTHER, MULTIPLE OR NOT SPECIFIED

AND with one of the following Bodily Locations.

HEAD
FINGERS
THUMB
UNSPECIFIED LOCATIONS

D.6 Segment 6

Probability of prolonged duration: 42%
Average duration: 232 days
Size of the segment: 4% of claims

Claims in Segment 6

A claim with one of the following Nature of injury;

JOINT DISEASES (ARTHROPATHIES) AND OTHER ARTICULAR CARTILAGE DISEASES
SPINAL VERTEBRAE AND INTERVERTEBRAL DISC DISEASES - DORSOPATHIES
BURSITIS
OCCUPATIONAL OVERUSE SYNDROME
FIBROMYALGIA
FIBROSITIS AND MYALGIA

AND with one of the following Bodily Locations;

NECK
TRUNK
UPPER LIMBS
LOWER LIMBS
MULTIPLE LOCATIONS
PSYCHOLOGICAL SYSTEM IN GENERAL

AND with one of the following Occupation.

FARMERS AND FARM MANAGERS
SPECIALIST MANAGERS
HOSPITALITY, RETAIL AND SERVICE MANAGERS
ARTS AND MEDIA PROFESSIONALS
BUSINESS, HUMAN RESOURCE AND MARKETING PROFESSIONALS

DESIGN, ENGINEERING, SCIENCE AND TRANSPORT PROFESSIONALS
EDUCATION PROFESSIONALS
ICT PROFESSIONALS
LEGAL, SOCIAL AND WELFARE PROFESSIONALS
ENGINEERING, ICT AND SCIENCE TECHNICIANS
AUTOMOTIVE AND ENGINEERING TRADES WORKERS
ELECTROTECHNOLOGY AND TELECOMMUNICATIONS TRADES WORKERS
OTHER TECHNICIANS AND TRADES WORKERS
CLERICAL AND ADMINISTRATIVE WORKERS
OFFICE MANAGERS AND PROGRAM ADMINISTRATORS
PERSONAL ASSISTANTS AND SECRETARIES
GENERAL CLERICAL WORKERS
INQUIRY CLERKS AND RECEPTIONISTS
NUMERICAL CLERKS
OTHER CLERICAL AND ADMINISTRATIVE WORKERS
SALES REPRESENTATIVES AND AGENTS
SALES ASSISTANTS AND SALESPERSONS
MACHINE AND STATIONARY PLANT OPERATORS
STOREPERSONS
FACTORY PROCESS WORKERS
FARM, FORESTRY AND GARDEN WORKERS
NOT KNOWN

D.7 Segment 7

Probability of prolonged duration: 52%
Average duration: 287 days
Size of the segment: 5% of claims

Claims in Segment 7

A claim with one of the following Nature of injury;

JOINT DISEASES (ARTHROPATHIES) AND OTHER ARTICULAR CARTILAGE DISEASES
SPINAL VERTEBRAE AND INTERVERTEBRAL DISC DISEASES - DORSOPATHIES
BURSITIS

OCCUPATIONAL OVERUSE SYNDROME
FIBROMYALGIA
FIBROSITIS AND MYALGIA

AND with one of the following Bodily Locations;

NECK
TRUNK
UPPER LIMBS
LOWER LIMBS
MULTIPLE LOCATIONS
PSYCHOLOGICAL SYSTEM IN GENERAL

AND with one of the following Occupation.

CHIEF EXECUTIVES, GENERAL MANAGERS AND LEGISLATORS
HEALTH PROFESSIONALS
CONSTRUCTION TRADES WORKERS
FOOD TRADES WORKERS
SKILLED ANIMAL AND HORTICULTURAL WORKERS
COMMUNITY AND PERSONAL SERVICE WORKERS
HEALTH AND WELFARE SUPPORT WORKERS
CARERS AND AIDES
HOSPITALITY WORKERS
PROTECTIVE SERVICE WORKERS
SPORTS AND PERSONAL SERVICE WORKERS
CLERICAL AND OFFICE SUPPORT WORKERS
SALES SUPPORT WORKERS
MACHINERY OPERATORS AND DRIVERS
MOBILE PLANT OPERATORS
ROAD AND RAIL DRIVERS
LABOURERS
CLEANERS AND LAUNDRY WORKERS
CONSTRUCTION AND MINING LABOURERS
FOOD PREPARATION ASSISTANTS
OTHER LABOURERS

OR, a claim with one of the following Nature of injury;

FRACTURES
POST-TRAUMATIC STRESS DISORDER
ANXIETY/STRESS DISORDER
DEPRESSION
ANXIETY/DEPRESSION COMBINED
SHORT TERM SHOCK FROM EXPOSURE TO DISTURBING CIRCUMSTANCES
REACTION TO STRESSORS - OTHER, MULTIPLE OR NOT SPECIFIED

AND with one of the following Bodily Locations;

TRUNK EXCEPT BACK - UPPER AND LOWER
ELBOW
HAND, FOOT AND TOES
PSYCHOLOGICAL SYSTEM IN GENERAL

AND with one of the following Occupation;

ARTS AND MEDIA PROFESSIONAL
DESIGN, ENGINEERING, SCIENCE AND TRANSPORT PROFESSIONALS
EDUCATION PROFESSIONALS
ENGINEERING, ICT AND SCIENCE TECHNICIANS
AUTOMOTIVE AND ENGINEERING TRADES WORKERS
CONSTRUCTION TRADES WORKERS
ELECTROTECHNOLOGY AND TELECOMMUNICATIONS TRADES WORKERS
SKILLED ANIMAL AND HORTICULTURAL WORKERS
OTHER TECHNICIANS AND TRADES WORKERS
COMMUNITY AND PERSONAL SERVICE WORKERS
CLERICAL AND OFFICE SUPPORT WORKERS
OTHER CLERICAL AND ADMINISTRATIVE WORKERS
SALES REPRESENTATIVES AND AGENTS
SALES ASSISTANTS AND SALESPERSONS
SALES SUPPORT WORKERS
MACHINERY OPERATORS AND DRIVERS
MACHINE AND STATIONARY PLANT OPERATORS
MOBILE PLANT OPERATORS
ROAD AND RAIL DRIVERS
LABOURERS

CLEANERS AND LAUNDRY WORKERS
FACTORY PROCESS WORKERS

AND with one of the following Bodily location of the most recent prior claim.

EAR
MOUTH
NOSE
HEAD - MULTIPLE LOCATIONS
CHEST (THORAX)
FOREARM
UPPER LEG
LOWER LEG
LOWER LIMB - MULTIPLE LOCATIONS
HEAD AND NECK
HEAD AND OTHER
TRUNK AND LIMBS
UPPER AND LOWER LIMBS
NECK AND SHOULDER
UNSPECIFIED MULTIPLE LOCATIONS
RESPIRATORY SYSTEM
OTHER AND MULTIPLE SYSTEMIC CONDITIONS
UNSPECIFIED LOCATIONS
NEW CLAIM

D.8 Segment 8

Probability of prolonged duration: 47%
Average duration: 202 days
Size of the segment: 1% of claims

Claims in Segment 8

A claim with one of the following Nature of injury;

FRACTURES
POST-TRAUMATIC STRESS DISORDER

ANXIETY/STRESS DISORDER
DEPRESSION
ANXIETY/DEPRESSION COMBINED
SHORT TERM SHOCK FROM EXPOSURE TO DISTURBING CIRCUMSTANCES
REACTION TO STRESSORS - OTHER, MULTIPLE OR NOT SPECIFIED

AND with one of the following Bodily Locations;

TRUNK EXCEPT BACK - UPPER AND LOWER
ELBOW
HAND, FOOT AND TOES
PSYCHOLOGICAL SYSTEM IN GENERAL

AND with one of the following Occupation;

ARTS AND MEDIA PROFESSIONAL
DESIGN, ENGINEERING, SCIENCE AND TRANSPORT PROFESSIONALS
EDUCATION PROFESSIONALS
ENGINEERING, ICT AND SCIENCE TECHNICIANS
AUTOMOTIVE AND ENGINEERING TRADES WORKERS
CONSTRUCTION TRADES WORKERS
ELECTROTECHNOLOGY AND TELECOMMUNICATIONS TRADES WORKERS
SKILLED ANIMAL AND HORTICULTURAL WORKERS
OTHER TECHNICIANS AND TRADES WORKERS
COMMUNITY AND PERSONAL SERVICE WORKERS
CLERICAL AND OFFICE SUPPORT WORKERS
OTHER CLERICAL AND ADMINISTRATIVE WORKERS
SALES REPRESENTATIVES AND AGENTS
SALES ASSISTANTS AND SALESPERSONS
SALES SUPPORT WORKERS
MACHINERY OPERATORS AND DRIVERS
MACHINE AND STATIONARY PLANT OPERATORS
MOBILE PLANT OPERATORS
ROAD AND RAIL DRIVERS
LABOURERS
CLEANERS AND LAUNDRY WORKERS
FACTORY PROCESS WORKERS

AND with one of the following Bodily location of the most recent prior claim.

CRANIUM
EYE
FACE, NOT ELSEWHERE SPECIFIED
NECK
BACK - UPPER OR LOWER
ABDOMEN AND PELVIC REGION
TRUNK - MULTIPLE LOCATIONS
SHOULDER
UPPER ARM
ELBOW
WRIST
HAND, FINGERS AND THUMB
HIP
KNEE
ANKLE
FOOT AND TOES
LOWER LIMB - UNSPECIFIED LOCATIONS
NECK AND TRUNK
OTHER SPECIFIED MULTIPLE LOCATIONS
CIRCULATORY SYSTEM
PSYCHOLOGICAL SYSTEM
NOT KNOWN

D.9 Segment 9

Probability of prolonged duration: 58%
Average duration: 281 days
Size of the segment: 3% of claims

Claims in Segment 9

A claim with one of the following Nature of injury;

FRACTURES
POST-TRAUMATIC STRESS DISORDER

ANXIETY/STRESS DISORDER
DEPRESSION
ANXIETY/DEPRESSION COMBINED
SHORT TERM SHOCK FROM EXPOSURE TO DISTURBING CIRCUMSTANCES
REACTION TO STRESSORS - OTHER, MULTIPLE OR NOT SPECIFIED

AND with one of the following Bodily Locations;

TRUNK EXCEPT BACK - UPPER AND LOWER
ELBOW
HAND, FOOT AND TOES
PSYCHOLOGICAL SYSTEM IN GENERAL

AND with one of the following Occupation.

CHIEF EXECUTIVES, GENERAL MANAGERS AND LEGISLATORS
FARMERS AND FARM MANAGERS
SPECIALIST MANAGERS
HOSPITALITY, RETAIL AND SERVICE MANAGERS
BUSINESS, HUMAN RESOURCE AND MARKETING PROFESSIONALS
HEALTH PROFESSIONALS
ICT PROFESSIONALS
LEGAL, SOCIAL AND WELFARE PROFESSIONALS
TECHNICIANS AND TRADES WORKERS
FOOD TRADES WORKERS
HEALTH AND WELFARE SUPPORT WORKERS
CARERS AND AIDES
HOSPITALITY WORKERS
PROTECTIVE SERVICE WORKERS
SPORTS AND PERSONAL SERVICE WORKERS
CLERICAL AND ADMINISTRATIVE WORKERS
OFFICE MANAGERS AND PROGRAM ADMINISTRATORS
PERSONAL ASSISTANTS AND SECRETARIES
GENERAL CLERICAL WORKERS
INQUIRY CLERKS AND RECEPTIONISTS
NUMERICAL CLERKS
STOREPERSONS

```
CONSTRUCTION AND MINING LABOURERS
FARM, FORESTRY AND GARDEN WORKERS
FOOD PREPARATION ASSISTANTS
OTHER LABOURERS
NOT KNOWN
```

D.10 Segment 10

Probability of prolonged duration: 68%
Average duration: 275 days
Size of the segment: 2% of claims

Claims in Segment 10

A claim with one of the following Nature of injury;

```
FRACTURES
POST-TRAUMATIC STRESS DISORDER
ANXIETY/STRESS DISORDER
DEPRESSION
ANXIETY/DEPRESSION COMBINED
SHORT TERM SHOCK FROM EXPOSURE TO DISTURBING CIRCUMSTANCES
REACTION TO STRESSORS - OTHER, MULTIPLE OR NOT SPECIFIED
```

AND with one of the following Bodily Locations.

```
NECK
UPPER BACK
LOWER BACK
BACK - OTHER AND MULTIPLE
BACK - UNSPECIFIED
SHOULDER
UPPER ARM
FOREARM
WRIST
HAND, FINGERS AND THUMB - OTHER AND MULTIPLE
HAND, FINGERS AND THUMB - UNSPECIFIED
```

LOWER LIMBS EXCEPT FOOT AND TOES
MULTIPLE LOCATIONS

Appendix E

Final triage model for psychological claims

Probability of prolonged duration, average duration and size of each segment are calculated by applying the rules, that are obtained from the training dataset and tested on the test dataset, to all psychological claims in the whole dataset.

E.1 Segment 1

Probability of prolonged duration: 37%
Average duration: 241 days
Size of the segment: 9% of all psychological claims

Claims in Segment 1

A claim with one of the following Occupation groups (Occupation Group 1);

```
ARTS AND MEDIA PROFESSIONALS
DESIGN, ENGINEERING, SCIENCE AND TRANSPORT PROFESSIONALS
HEALTH PROFESSIONALS
AUTOMOTIVE AND ENGINEERING TRADES WORKERS
CONSTRUCTION TRADES WORKERS
FOOD TRADES WORKERS
CLERICAL AND OFFICE SUPPORT WORKERS
SALES SUPPORT WORKERS
MACHINE AND STATIONARY PLANT OPERATORS
```

MOBILE PLANT OPERATORS
ROAD AND RAIL DRIVERS
CLEANERS AND LAUNDRY WORKERS
CONSTRUCTION AND MINING LABOURERS
FACTORY PROCESS WORKERS
FOOD PREPARATION ASSISTANTS

AND with one of the following Bodily Locations of the most recent prior claim.

CRANIUM
EAR
MOUTH
HEAD - MULTIPLE LOCATIONS
NECK
TRUNK - MULTIPLE LOCATIONS
SHOULDER
UPPER ARM
WRIST
HIP
UPPER LEG
KNEE
LOWER LIMB - UNSPECIFIED LOCATIONS
NECK AND TRUNK
HEAD AND OTHER
TRUNK AND LIMBS
NECK AND SHOULDER
OTHER SPECIFIED MULTIPLE LOCATIONS
CIRCULATORY SYSTEM
PSYCHOLOGICAL SYSTEM

E.2 Segment 2

Probability of prolonged duration: 42%
Average duration: 163 days
Size of the segment: 9% of all psychological claims

Claims in Segment 2

A claim with Occupation groups in Occupation Group 1 (defined above);
AND with one of the following Bodily Locations of the most recent prior claim (Prior Bodily Location Sector 1);

```
EYE
FACE, NOT ELSEWHERE SPECIFIED
BACK - UPPER OR LOWER
CHEST (THORAX)
ABDOMEN AND PELVIC REGION
ELBOW
FOREARM
HAND, FINGERS AND THUMB
LOWER LEG
ANKLE
FOOT AND TOES
LOWER LIMB - MULTIPLE LOCATIONS
HEAD AND NECK
UPPER AND LOWER LIMBS
RESPIRATORY SYSTEM
OTHER AND MULTIPLE SYSTEMIC CONDITIONS
NOT KNOWN
```

AND with Age of injured worker less than 36.5.

E.3 Segment 3

Probability of prolonged duration: 53%
Average duration: 291 days
Size of the segment: 14% of all psychological claims

Claims in Segment 3

A claim with Occupation groups in Occupation Sector 1 (defined above);
AND with Bodily Locations of the most recent prior claim in Prior Bodily Location Sector 1 (defined above);
AND with Age of injured worker greater than or equal to 36.5.

E.4 Segment 4

Probability of prolonged duration: 60%
Average duration: 337 days
Size of the segment: 69% of all psychological claims

Claims in Segment 4

A claim with one of the following Occupation groups.

CHIEF EXECUTIVES, GENERAL MANAGERS AND LEGISLATORS
FARMERS AND FARM MANAGERS
SPECIALIST MANAGERS
HOSPITALITY, RETAIL AND SERVICE MANAGERS
BUSINESS, HUMAN RESOURCE AND MARKETING PROFESSIONALS
EDUCATION PROFESSIONALS
ICT PROFESSIONALS
LEGAL, SOCIAL AND WELFARE PROFESSIONALS
ENGINEERING, ICT AND SCIENCE TECHNICIANS
ELECTROTECHNOLOGY AND TELECOMMUNICATIONS TRADES WORKERS
SKILLED ANIMAL AND HORTICULTURAL WORKERS
OTHER TECHNICIANS AND TRADES WORKERS
HEALTH AND WELFARE SUPPORT WORKERS
CARERS AND AIDES
HOSPITALITY WORKERS
PROTECTIVE SERVICE WORKERS
SPORTS AND PERSONAL SERVICE WORKERS
OFFICE MANAGERS AND PROGRAM ADMINISTRATORS
PERSONAL ASSISTANTS AND SECRETARIES
GENERAL CLERICAL WORKERS
INQUIRY CLERKS AND RECEPTIONISTS
NUMERICAL CLERKS
OTHER CLERICAL AND ADMINISTRATIVE WORKERS
SALES REPRESENTATIVES AND AGENTS
SALES ASSISTANTS AND SALESPERSONS
MACHINERY OPERATORS AND DRIVERS
STOREPERSONS

FARM, FORESTRY AND GARDEN WORKERS
OTHER LABOURERS

High-risk Occupation groups in Segment 4 (probability of prolonged duration, average duration)

ICT PROFESSIONALS (86%, 485 days)
PERSONAL ASSISTANTS AND SECRETARIES (76%, 396 days)
CHIEF EXECUTIVES, GENERAL MANAGERS AND LEGISLATORS (74%, 592 days)
SPECIALIST MANAGERS (70%, 456 days)
BUSINESS, HUMAN RESOURCE AND MARKETING PROFESSIONALS (69%, 398 days)
LEGAL, SOCIAL AND WELFARE PROFESSIONALS (68%, 330 days)
PROTECTIVE SERVICE WORKERS (64%, 394 days)
EDUCATION PROFESSIONALS (63%, 375 days)
OFFICE MANAGERS AND PROGRAM ADMINISTRATORS (62%, 331 days)
HOSPITALITY, RETAIL AND SERVICE MANAGERS (61%, 392 days)

High-risk Nature of injury in Segment 4 (probability of prolonged duration, average duration)

DEPRESSION/ ANXIETY COMBINED (68%, 432 days)