Predictive Modelling for Health and Health-care Utilisation

An observational study for Australians aged 45 and up

Amir Marashi

Translational Health Research Institute

University of Western Sydney

This dissertation is submitted for the degree of

Doctor of Philosophy

March 2019

I would like to dedicate this thesis to my loving parents.

ACKNOWLEDGEMENTS

I am truly indebted to both of my supervisors, Federico Girosi and Shima Ghassem Pour, for their faith in me and valuable help and guidance throughout this journey. For me, Federico was not only a great academic supervisor but also an invaluable mentor and role model and Shima became a caring friend who I am honoured to have made. I also sincerely appreciate the supportive comments and valuable feedback of my industry supervisors Vincy Li and Chris Rissel, who guided the direction of this research.

Thank you also to staff and friends at CMCRC who made these years enjoyable and whose presence and support helped in times of stress.

I am very appreciative to all administrative and executive staff of Translational Health Research Institute (THRI) who were always supportive and made my life easier with their assistance.

My heartfelt appreciation goes to my dear family, especially my incredible parents, whose unconditional love and unending support has always been with me.

Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

Amir Marashi

March 2019

DECLARATION

I would like to thank Capital Markets Cooperative Research Centres (CMCRC) for supporting this research with their generous scholarship. I also would like to thank NSW health for providing me with a valuable working experience.

I declare that I have no conflict of interest.

Amir Marashi

March 2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acronyms

Acronyms / Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
APDC	Admitted Patient Data Collection
CEM	Coarsened Exact Matching
CHeRel	Centre for Health Record Linkage
CVD	Cardiovascular Diseases
DALY	Disability-Adjusted Life-Years
DCG	Diagnostic Cost Group
DL	Deep Learning
GLM	Generalised Linear Model
ICD	International Classification of Diseases
IPF	Iterative Proportional Fitting
LDA	Latent Dirichlet Allocation
LHD	Local Health Districts
LOS	Length of Stay
LSTM	Long Short-Term Memory
LTPA	Leisure Time Physical Activity
MAE	Mean Absolute Error
MLP	Multilayer Perceptron
MVPA	Moderate to Vigorous Physical Activity
NLP	Natural Language Processing
NSW	New South Wales

OLS	Ordinary Least Squares
PA	Physical Activity
RNN	Recurrent Neural Networks
SEEF	Social, Economic and Environmental Factors
SURE	Secure Unified Research Environment
SVM	Support Vector Machines
WHO	World Health Organisation

ABSTRACT

The burden of chronic disease is growing at a fast pace, leading to poor quality of life and high healthcare expenditures in a large portion of the Australian population. Much of the burden is borne by hospitals, and therefore there is an ever-increasing interest in preventative interventions that can keep people out of hospitals and healthier for longer periods. There is a wide range of potential interventions that may be able to achieve this goal, and policy makers need to decide which one should be funded and implemented. This task is difficult for two reasons: first it is often not clear what is the short-term effectiveness of an intervention, and how it varies in specific sub-populations, and second it is also not clear what the long-term intended and unintended consequences might be.

In this thesis I make contributions to address both these difficulties. On the short-term side I focus on the use of physical activity to prevent the development of chronic disease and to reduce hospital costs. Increasing physical activity has been long heralded as a way to achieve these goals but evidence of its effectiveness has been elusive. In this thesis I provide data driven evidence to justify policies that encourage higher levels of physical activity (PA) in middle age and older Australian population. I use data from the "45 and up" and the Social, Economic and Environmental Factors (SEEF) study, linked with the Admitted Patient Data Collection (APDC), to identify and study the cost and health trajectories of individuals with different levels of physical activity. The results show a clear statistically significant association between PA and lower hospitalisation cost, as well as between PA and reduced risk of heart disease, diabetes and stroke.

On the long-term side of the analysis, I placed this thesis in the context of a larger program of work performed at Western Sydney University that aims to build a microsimulation model

for the analysis of health policy interventions. In this framework I studied predictive models that use survey and/or administrative data to predict hospital costs and resource utilisation. I placed particular emphasis on the application of methods borrowed from Natural Language Processing to understand how to use the thousands of diagnosis and procedure codes found in administrative data as input to predictive models. The methods developed in this thesis go beyond the application to hospital data and can be used in any predictive model that relies on complex coding of healthcare information.

Part I

Introduction and Background Material

CHAPTER 1

INTRODUCTION

Like many other developed countries, Australia has recently experienced a large increase in growth of health expenditures, that have reached $181 billion in 2016–17, corresponding to more than $7,400 per person and 10% of overall economic activity. Adjusting for inflation, health spending grew at a rate of 4.7% in 2016–17, which is much higher than the 3.1% growth rate observed on average over the past 5 years[1].

Continued expenditure growth at this rate is not sustainable, and therefore payers such as federal and state governments, as well as private insurers, have been looking at mechanisms that can slow the expenditure growth while still providing high quality care. Strategies of different types are being employed. Some strategies are looking at improving the efficiency of the provision of services, reducing waste and making sure that care is provided only to those who will benefit from it. Other strategies aim to prevent or delay the development of expensive and long-lasting health conditions, by identifying individuals at risk and intervening early.

A common factor in many approaches to cost containment is the increased utilisation of individual level health data, which are fed to predictive models and/or machine learning algorithms, whose results are used to guide and inform either policy or implementation. Important drivers for this phenomenon are the increased availability of health data, the increased awareness of the power of health data, and the great progress in Artificial Intelligence

[1] Australian Institute of Health and Welfare 2018. Health expenditure Australia 2016–17. Health and welfare expenditure series no. 64. Cat. no. HWE 74. Canberra: AIHW.

(AI) witnessed in the last decade, exemplified by some great successes of applications of AI in health [60, 82, 104, 134, 190, 190].

As a consequence, the field of health services research has become increasingly more inter-disciplinary, and traditional disciplines such as epidemiology, which is mainly focused on the explanation of associations, are becoming more and more intertwined with machine learning, which tends to be focused on the prediction of those associations.

This thesis is an example of such a trend. The research started as an investigation around the issue of prevention, and more specifically about whether increases in the levels of physical activity (PA) are associated with decreased hospital costs and decreased incidence of chronic conditions such as stroke, hypertension, diabetes and heart disease.

I have addressed these questions in the framework of modern epidemiology, utilising linked survey and administrative data and exploring a variety of statistical methodologies such as matching, re-weighting and instrumental variables in order to obtain an estimate of the association between PA and several primary outcomes which is not confounded. Estimates of the strength of the association are useful to payers considering the establishments of programs to increase levels of PA in the population, since they allow to answer simple "what-if" questions, especially when they are stratified over specific characteristics and allow to understand which sub-populations would benefit most from the intervention.

However, these estimates are also limited in use, since they only allow to explore short-term scenarios: while they may suggest how much is saved in hospital costs or how many fewer cases of heart disease we will observe next year, they are silent on the ramifications of improved health. Consider the case of an individual who, thanks to PA, does not prematurely die of cardiovascular disease and enjoys a longer life span. This individual could then develop another expensive chronic health condition, perhaps associated with low quality of life. In this case the savings associated to the avoided cardiovascular events may be negated by the additional expenditures for another chronic condition, and we may observe an increase in costs not compensated by a sufficiently high increase in quality adjusted life years, possibly making the intervention not cost-effective. On the other side there will also be individuals who thanks to PA go on to enjoy additional years of high quality of life, making a preventive inter-

vention highly cost-effective, even if not cost-saving. These trade-offs can only be explored in a much more sophisticated modelling context, where one simulates the population of interest over time, allowing individuals with different characteristics to be differently affected by the intervention and therefore allowing to explore the long-term consequences of increased levels of PA.

A simulation of this type is a massive enterprise that would have been out of scope for a PhD thesis, but fortunately a related simulation project was being developed by researchers at Western Sydney University for the purpose of simulating a variety of policy intervention, not only those related to PA. Therefore it was agreed that we could use the work on PA to inform the simulation project and that I could also contribute to the project by working on some of the necessary predictive algorithms. More specifically, I was assigned the task to study the best way to predict health expenditures, which are notoriously hard to predict, since they depend on many different parameters [32, 166] and are characterised by an intrinsically high level of variability [43].

As a result, this thesis is composed of two parts: one relates to the effect of PA on health and health care utilisation, and the other relates to the development of novel machine learning algorithms for the prediction of hospital costs based on individual level data. While the two parts may appear only remotely related, the deep connection between the two consists in the fact that they both are part of a bigger project, that has the objective to provide tools to policy makers that allow them to simulate the effect of a variety of policy interventions. In the following section I briefly outline the specifics of the research questions I address and my contributions to the above mentioned areas of research.

1.1 Contributions

The part of the thesis concerning the association of PA with hospital costs and the association of PA with incidence of four chronic conditions addresses and answers the following related questions:

- How much is saved in hospital cost when physically inactive individuals start to perform sufficient levels of PA?
- How do the savings vary across age groups and which age groups would benefit most from PA interventions?
- What is the association between physical activity and the incidence of diabetes, heart disease, stroke and hypertension?
- Which population sub-groups see the greatest reduction in incidence of chronic conditions when starting to perform sufficient levels of PA?

In regards to the development of predictive models for hospital costs I have delivered the following contributions:

- Evaluated the performance of several methods for predicting hospital costs, both at the level of hospital and the level of Local Health Districts (LHD).
- Developed a tool that allows to cluster hospital admissions in homogeneous and interpretable groups based on the ICD code descriptions.
- Developed a tool that applies word embedding methods and deep learning methodology to predict the LOS of a hospital admission based on the International Classification of Diseases (ICD) code descriptions.

1.2 Organisation

Chapter 2 of the thesis provides some background from the literature about the above questions. It also includes short introductions of the machine learning methods and statistical analysis that I have applied in this study. Chapter 3 introduces the data set that has been used in this thesis. Chapter 4 addresses the first and the second questions and Chapter 5 focus on the third and the forth questions. They describe the method, variables and the analytical

techniques that I have used to reduce the estimation bias and discuss the results and the challenges of the analysis. Chapter 6 and 7 present the work around predictive models for hospital costs. Chapter 6 attempts to evaluate the predictive power of the survey data for prediction of health care costs and Chapter 7 applies two different machine learning methods to analyse text data that can be found in health data. Finally, Chapter 8 provides the summary of this study and some suggestions for the future works.

CHAPTER 2

BACKGROUND AND THEORETICAL FRAMEWORK

In this thesis, I explored modelling different aspects of health and health care utilisation by investigating the association between different variables, specifically physical activity, and different health and utilisation outcomes. This line of research has several components, each of which deserved a separate literature review, that summarises and discusses methods and results of previous studies. I grouped the literature review in three distinct parts:

Epidemiology: From an epidemiological point of view in this thesis I considered one key independent variable, physical activity, and two primary outcomes: hospital costs and incidence of four chronic conditions (stroke, diabetes, heart disease and hypertension). The literature regarding the association between physical activity and hospital costs is reviewed in Section 2.1, and it is particular relevant for the material presented in Chapter 4. Literature on the association between physical activity and the incidence of chronic conditions is described in Section 2.2, and it speaks to the topic of Chapter 5.

Study Design and Statistical Methodology: In order to study the association between physical activity and the primary outcomes I have used individual level survey data. From a methodological viewpoint this poses two important questions:

- Is the data sufficiently representative of the population of interest? and if not, how can it be made more representative?

- How should the analysis be structured to minimise confounding and obtain results as close as possible as the causal effect?

The first question is investigated in the framework of reweighting methods. In particular, I had to reweight the survey data in order to make it better represent the population of New South Wales (NSW). Therefore, in Section 2.3 I reviewed the key reweighting method, the Iterative Proportional Fitting (IPF) and its implementation. We notice here that while the IPF algorithm has existed for a long time, the general and simple derivation provided in Section 2.3 is novel and it is contribution of this thesis.

The second question is of crucial importance: while a true causal effect cannot be estimated using the survey data at hand, there are different methods that can be used to ensure that our estimate of association is as close to causal as possible. A key role is played here by matching methods, which are reviewed in Section 2.4, together with basic notions of causal inference and instrumental variables.

Predictive Modelling and Machine Learning: While for the purpose of studying association between an independent variable and a primary outcome a simple statistical model may be sufficient, in this thesis I have explored more sophisticated alternatives. In particular, I have considered modern machine learning methods and how they can be applied to the problem of building a predictive model for health care utilisation or for the incidence of a chronic condition. The relevant literature is discussed in Section 2.5. Since a difficult problem in the use of health data is the choice of a representation for coded data (such as ICD-9 hospital data), in Section 2.6 I review background information on text embedding and analysis, which I will use in Chapter 6 to solve this specific problem.

2.1 Association between physical activity and health-care costs

There is strong and well-established evidence that indicates the benefits of regular physical activity on health-care expenditures.

Sari [150] has published a review of the literature, up to year 2011, on the impact of physical activity on health-care utilisation on older adults, emphasising the evidence that physical activity results in lower utilisation of health-care services but finding a lack of robust estimates for the size of the effect. A number of more recent studies support the same idea [45, 55, 87, 122, 137, 147]. There are also several studies exploring the cost effectiveness of interventions promoting physical activity. Vijay et al. [174] have done a systematic review of these studies and there are a number of examples focusing on specific interventions [2, 35, 49, 57, 126]. These studies have been performed on different populations, using different variables and with different definitions of physical activity and payments and therefore it is not surprising that they report quite different magnitudes for the effect of physical activity.

In this literature review, I consider the association between PA and health care utilisation.

The association between physical activity and health status and mortality, which in turn will cause higher health-care expenditure, has been widely discussed in the literature. However, there is less research on the direct effect of physical activity on health-care utilisation. These studies could be done on the survey data or the output of intervention programs. I divide the studies into two groups of Australian literature and international literature based on the population of the studies. The following summaries are derived directly from the studies.

2.1.1 Australian and New Zealand literature

- *Cost utility analysis of physical activity counselling in general practice [38]*
  - **Method**: Cost utility analysis using a Markov model was used to estimate the cost utility of the Green Prescription program over full life expectancy. Program effectiveness was based on published trial data (878 inactive patients presenting to NZ general practice). Costs were based on detailed costing information and were discounted at 5% per anum. The main outcome measure is cost per quality adjusted life year (QALY) gained. Extensive one-way sensitivity analyses were performed along with probabilistic (stochastic) analysis.
  - **Results**: Incremental, modelled cost utility of the Green Prescription program compared with 'usual care' was $NZ2,053 per QALY gained over full life expectancy.

  - **Conclusion**: Based on a plausible and conservative set of assumptions, if decision makers are willing to pay at least $NZ2,000 per QALY gained the Green Prescription program is likely to represent better value for money than 'usual care'.

- *Cost-Effectiveness of Interventions to Promote Physical Activity: A Modelling Study [35]*

  - **Method**: From evidence of intervention efficacy in the physical activity literature and evaluation of the health sector costs of intervention and disease treatment, this study modelled the cost impacts and health outcomes of six physical activity and interventions, over the lifetime of the Australian population and then determined cost-effectiveness of each intervention against current practice for physical activity intervention in Australia and derived the optimal pathway for implementation.

  - **Results**: Based on current evidence of intervention effectiveness, the intervention programs that encourage use of pedometers (Dominant) and mass media-based community campaigns (Dominant) are the most cost-effective strategies to implement and are very likely to be cost-saving. The internet-based intervention program, the GP physical activity prescription program, and the program to encourage more active transport, although less likely to be cost-saving, have a high probability of being under a AUS$50,000 per DALY threshold. GP referral to an exercise physiologist is the least cost-effective option if high time and travel costs for patients in screening and consulting an exercise physiologist are considered.

  - **Conclusion**: Intervention to promote physical activity is recommended as a public health measure. Despite substantial variability in the quantity and quality of evidence on intervention effectiveness, and uncertainty about the long-term sustainability of behavioural changes, it is highly likely that as a package, all six interventions could lead to substantial improvement in population health at a cost saving to the health sector.

- *Cost-effectiveness implications of GP intervention to promote physical activity: evidence from Perth, Australia [2]*

  - **Method**: The percentage of population that could potentially move from insufficiently active to sufficiently active, on GP advice was drawn from the Western Australian (WA) Premier's Physical Activity Taskforce (PATF) survey in 2006. Population impact fractions (PIF) for diseases attributable to physical inactivity together with disability adjusted life years (DALYs) and health care expenditure were used to estimate the net cost of intervention for varying subsidies. Cost-effectiveness of subsidy programs were evaluated in terms of cost per DALY saved at different compliance rates.
  - **Results**: With a 50% adherence to GP advice, an annual health care cost of AU$ 24 million could be potentially saved to the WA economy. A DALY can be saved at a cost of AU$ 11,000 with a AU$ 25 subsidy at a 50% compliance rate. Cost effectiveness of such a subsidy program decreases at higher subsidy and lower compliance rates.
  - **Conclusion**: Implementing a subsidy for GP advice could potentially reduce the burden of physical inactivity. However, the cost-effectiveness of a subsidy program for GP advice depends on the percentage of population who comply with GP advice.

- *The societal benefits of reducing six behavioural risk factors: an economic modelling study from Australia [30]*
  - **Method**: Simulation models were developed for the 2008 Australian population. A realistic reduction in current risk factor prevalence using best available evidence with expert consensus was determined. Avoidable disease, deaths, Disability Adjusted Life Years (DALYs) and health sector costs were estimated. Productivity gains included workforce (friction cost method), household production and leisure time. Multivariable uncertainty analyses and correction for the joint effects of risk factors on health status were undertaken. Consistent methods and data sources were used.

  - **Results**: Over the lifetime of the 2008 Australian adult population, total opportunity cost savings of AUD2,334 million were found if feasible reductions in the risk factors were achieved. There would be 95,000 fewer DALYs (a reduction of about 3.6% in total DALYs for Australia); 161,000 less new cases of disease; 6,000 fewer deaths; a reduction of 5 million days in workforce absenteeism; and 529,000 increased days of leisure time.
  - **Conclusion**: Reductions in common behavioural risk factors may provide substantial benefits to society.

2.1.2 International literature

- *Exercise, physical activity and health-care utilisation: A review of literature for older adults [150]*
  - **Method**: The paper reviews the literature on physical activity and its implications for health-care system, and discusses potential directions for future research by highlighting the limitations of the existing studies.
  - **Results**: Although there are significant variations in samples and methods used, both streams of reviewed literature provide evidence that physical activity leads to lower utilisation of health-care services.
  - **Conclusion**: Given differences in methods and samples in these studies, estimated effect of physical activity on health-care utilisation shows significant variation from one study to another. These results, therefore, cannot be generalised to justify population wide exercise intervention programs for older adults. Additional studies are needed to provide more robust estimates for the effects of exercise, and to examine the feasibility of population wide policies that aim to encourage participation of older adults in physical activity.
- *The Economic Costs Associated with physical inactivity and obesity in Canada: An update [89]*

  - **Method**: A prevalence-based economic burden analysis was undertaken. The relative risks of diseases associated with physical inactivity and obesity were determined from a meta-analysis of existing prospective studies and applied to the health care costs of these diseases in Ontario. The prevalence of physical inactivity and obesity were obtained from the 2009 Canadian Community Health Survey (CCHS) for the province of Ontario. Estimates of the economic burden were derived from both direct and indirect expenditure categories. Direct medical costs included hospital care expenditures, drug expenditures, physician care expenditures, expenditures for care in other institutions, and additional direct health expenditures; whereas indirect costs included the value of years of life lost due to premature death and the value of days lost due to short-term and long-term disability.
  - **Results**: The economic burden of physical inactivity was $3.4 billion ($1.02 billion in direct costs and $2.34 billion in indirect costs) while the burden associated with obesity was $4.5 billion ($1.60 billion in direct costs and $2.87 billion in indirect costs).
  - **Conclusion**: These estimates reinforce the public health importance of curbing the current epidemics of physical inactivity and obesity in Ontario

- *Is Self-Reported Physical Activity Participation Associated with Lower Health Services utilisation among Older Adults? Cross-Sectional Evidence from the Canadian Community Health Survey [56]*

  - **Method**: Cross-sectional data from 56,652 Canadian Community Health Survey respondents aged bigger or equal 50 years were stratified into three age groups and analysed using multivariate generalised linear modelling techniques. Participants were classified according to PA level based on self-reported daily energy expenditure. Non-leisure PA (NLPA) was categorised into four levels ranging from mostly sitting to mostly lifting objects.

  - **Results**: Active 50–65-year-old individuals were 27% less likely to report any GP consultations and had 8% fewer GP consultations annually than their inactive peers. Active persons aged 65–79 years were 18% less likely than inactive respondents to have been hospitalised overnight in the previous year. Higher levels of NLPA were significantly associated with lower levels of HSU, across all age groups.
  - **Conclusion**: Non-leisure PA appeared to be a stronger predictor of all types of HSU, particularly in the two oldest age groups. Considering strategies that focus on reducing time spent in sedentary activities may have a positive impact on reducing the demand for health services.

- *Is the association between physical activity and health-care utilisation affected by self-rated health and socio-economic factors? [147]*

  - **Method**: A cross-sectional public health survey was conducted in Skåne, Sweden 2012, based on a random sample with 55,000 participants (response rate 51%; 28,028 individuals included in the study) aged 18–80 years. The data was linked to individual health-care utilisation data and socio-economic data. Logistic regression analyses were conducted to study the association between LTPA and health-care utilisation.
  - **Results**: Compared to sedentary leisure time the odds ratio for health care utilisation decreased with increasing level of LPTA; physically active 0.89, for average exercise 0.74 and for vigorous exercise 0.65. The socio-economic variables attenuated this association to a small degree, but self-rated health (SRH) had a strong impact. While the mediation analysis illustrated that the indirect effects were strong (and in the expected order so that higher levels of LTPA were more negatively associated with poor health) and highly significant, the direct effects suggested that higher levels of physical activity were more positively associated with health-care utilisation than lower levels. The indirect effects were substantially stronger than the direct effects.

  - **Conclusion**: There was a significant negative association between decreased health-care utilisation and increased LPTA, and the association remained after adjustment for socio-economic variables. The mediation analysis (with SRH as the mediator between LTPA and health-care utilisation) showed that the indirect effects were strong and in the expected order, but the direct effects of LTPA on health-care utilisation was positive so that higher levels of LTPA had higher health-care utilisation. These results suggest that even though higher physical activity in total decreases the health-care utilisation, parts of the association that is not mediated through SRH actually increase health-care utilisation.

- *Are brief interventions to increase physical activity cost-effective? A systematic review [174]*

  - **Method**: Systematic review of economic evaluations.
  - **Results**: Of 1840 identified publications, 13 studies fulfilled the inclusion criteria describing 14 brief interventions. Studies varied widely in the methods used, such as the perspective of economic analysis, intervention effects and outcome measures. The incremental cost of moving an inactive person to an active state, estimated for eight studies, ranged from £96 to £986. The cost-utility was estimated in nine studies compared with usual care and varied from £57 to £14 002 per quality-adjusted life year; dominant to £6500 per disability-adjusted life year; and £15 873 per life years gained.
  - **Conclusion**: Brief interventions promoting physical activity in primary care and the community are likely to be inexpensive compared with usual care. Given the commonly accepted thresholds, they appear to be cost-effective on the whole, although there is notable variation between studies.

- *Excess Medical Care Costs Associated with Physical Inactivity among Korean Adults: Retrospective Cohort Study [122]*

  - **Method**: A total of 68,556 adults whose reported physical activity status did not change during the study period was included for this study. Propensity scores for inactive adults were used to match 23,645 inactive groups with 23,645 active

groups who had similar propensity scores. The study compared medical expenditures between the two groups using generalised linear models with a gamma distribution and a log link. Direct medical costs were based on the reimbursement records of all medical facilities from 2005 to 2010.

  - **Results**: The average total medical costs for inactive individuals were $1110.5, which was estimated to be 11.7% higher than the costs for physically active individuals. With respect to specific diseases, the medical costs of inactive people were significantly higher than those of active people, accounting for approximately 8.7% to 25.3% of the excess burden.
  - **Conclusion**: Physical inactivity is associated with considerable medical care expenditures per capita among Korean adults.

- *Physical activity and health services utilisation and costs among U.S. adults [87]*
  - **Method**: Data came from the Medical Expenditure Panel Survey-Household component from 2007 through 2011 (n=117,361). Regular physical activity was defined as spending half an hour or more in moderate or vigorous physical activity at least three times a week. The following categories of self-reported health services utilisation and costs were examined: preventive, office-based, outpatient, inpatient, emergency department, home health, and prescription medicines. The association between physical activity and health services utilisation and costs was estimated using two-part models.
  - **Results**: Adults who engaged in regular physical activity were more likely to use preventive and office-based services. Combining results from both parts of the two-part models, physically active adults incurred significantly lower utilisation of inpatient (0.09 vs 0.12 visit per person), emergency room (0.18 vs 0.19 visit per person), home health care (1.21 vs 1.92 visit per person), and prescription medicines (12.66 vs 13.75 number of prescriptions per person) and spent $27 less per capita expenditures for office-based visits, $351 less for inpatient visits, and $52 less for home health care visits.

  - **Conclusion**: Promoting regular physical activity may reduce health care costs through decreasing demand for secondary and tertiary care services.

- *The economic burden of physical inactivity: a global analysis of major non-communicable diseases [45]*

  - **Method**: Direct health-care costs, productivity losses, and disability-adjusted life-years (DALYs) attributable to physical inactivity were estimated with standardised methods and the best data available for 142 countries, representing 93·2% of the world's population. Direct health-care costs and DALYs were estimated for coronary heart disease, stroke, type 2 diabetes, breast cancer, and colon cancer attributable to physical inactivity. Productivity losses were estimated with a friction cost approach for physical inactivity related mortality. Analyses were based on national physical inactivity prevalence from available countries, and adjusted population attributable fractions (PAFs) associated with physical inactivity for each disease outcome and all-cause mortality.

  - **Results**: Conservatively estimated, physical inactivity cost health-care systems international $ (INT$) 53·8 billion worldwide in 2013, of which $31·2 billion was paid by the public sector, $12·9 billion by the private sector, and $9·7 billion by households. In addition, physical inactivity related deaths contribute to $13·7 billion in productivity losses, and physical inactivity was responsible for 13·4 million DALYs worldwide. High-income countries bear a larger proportion of economic burden (80·8% of health-care costs and 60·4% of indirect costs), whereas low-income and middle-income countries have a larger proportion of the disease burden (75·0% of DALYs). Sensitivity analyses based on less conservative assumptions led to much higher estimates.

  - **Conclusion**: In addition to morbidity and premature mortality, physical inactivity is responsible for a substantial economic burden. This paper provides further justification to prioritise promotion of regular physical activity worldwide as part of a comprehensive strategy to reduce non-communicable diseases.

2.2 Association between physical activity and chronic health conditions

The effect of physical activity on different chronic health conditions has been widely discussed in the literature. While physiologically it seems reasonable for physical activity to reduce the risk of different chronic health conditions, different epidemiological studies may report different conclusions. These contradictions may root in the amount of Physical Activity regarded in the study as sufficient, different characteristics and demography of the study population, quality of the data, the complicated interaction of PA and other risk factors or the relationships between PA and other interventions such as diet. This section focuses on the effect of PA on four chronic health conditions: Heart disease, hypertension, stroke and diabetes. These health conditions could be related together, for example, diabetes is a known risk factor of stroke [33, 85, 93]. Here I explored some of the available published research on the effect of PA on each condition separately.

2.2.1 Association between physical activity and diabetes

As mentioned before, different epidemiological studies may report mixing findings on the same topic. Some of the studies that I investigated about PA and diabetes, did not find any significant association between sedentary behaviour or PA and diabetes. One recent study on Mexican adults reported no association between occupational moderate to vigorous PA (MVPA) and diabetes [115]. Another recent study on Australian population aged ≥ 45 years reported no significant association between PA/sitting and prevalence of diabetes and suggested obesity as a risk factor for diabetes [130].

A recent systematic review over PubMed and Ovid databases explored 91 studies until March 2015. Some of these studies focus on total PA while the others consider only leisure-time PA, vigorous PA, walking, or occupational PA. Figure 2.1 and 2.2 show the relative risk for the effect of total PA and vigorous PA on type 2 diabetes for different studies investigated in this systematic review respectively.

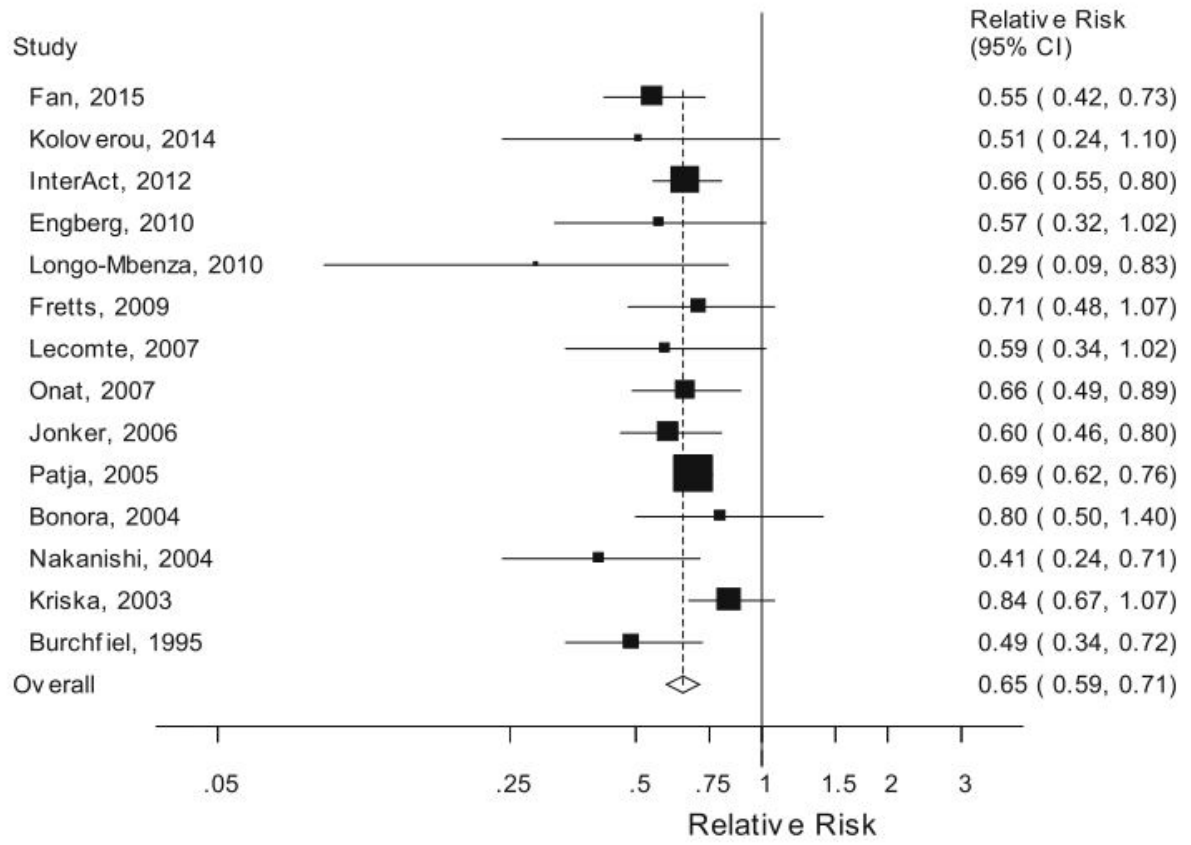

Fig. 2.1 Total physical activity and type 2 diabetes [13]

Whiles some of the studies have reported insignificant negative association and a few have reported insignificant positive association, the overall summary relative risk for total PA is 0.74 (95 % CI 0.70–0.79) and for leisure-time activity, is 0.61 (95 % CI 0.51–0.74) [13]. Another systematic review of the evidence for Canada's Physical Activity Guidelines for Adults [176] examined 20 studies in a range of 16 years. All these studies have shown that PA reduces the incidence of diabetes. Most of these studies show an incremental reductions in the risk for type 2 diabetes with increasing activity levels. The average risk reduction between the most and least active group has been reported 0.43.

The last study that I mention here is a report published by the Australian Institute of Health and Welfare in 2017 [16]. The study found that type 2 diabetes was causally associated with physical inactivity and 19% of diabetes disease burden was because of physical inactivity.

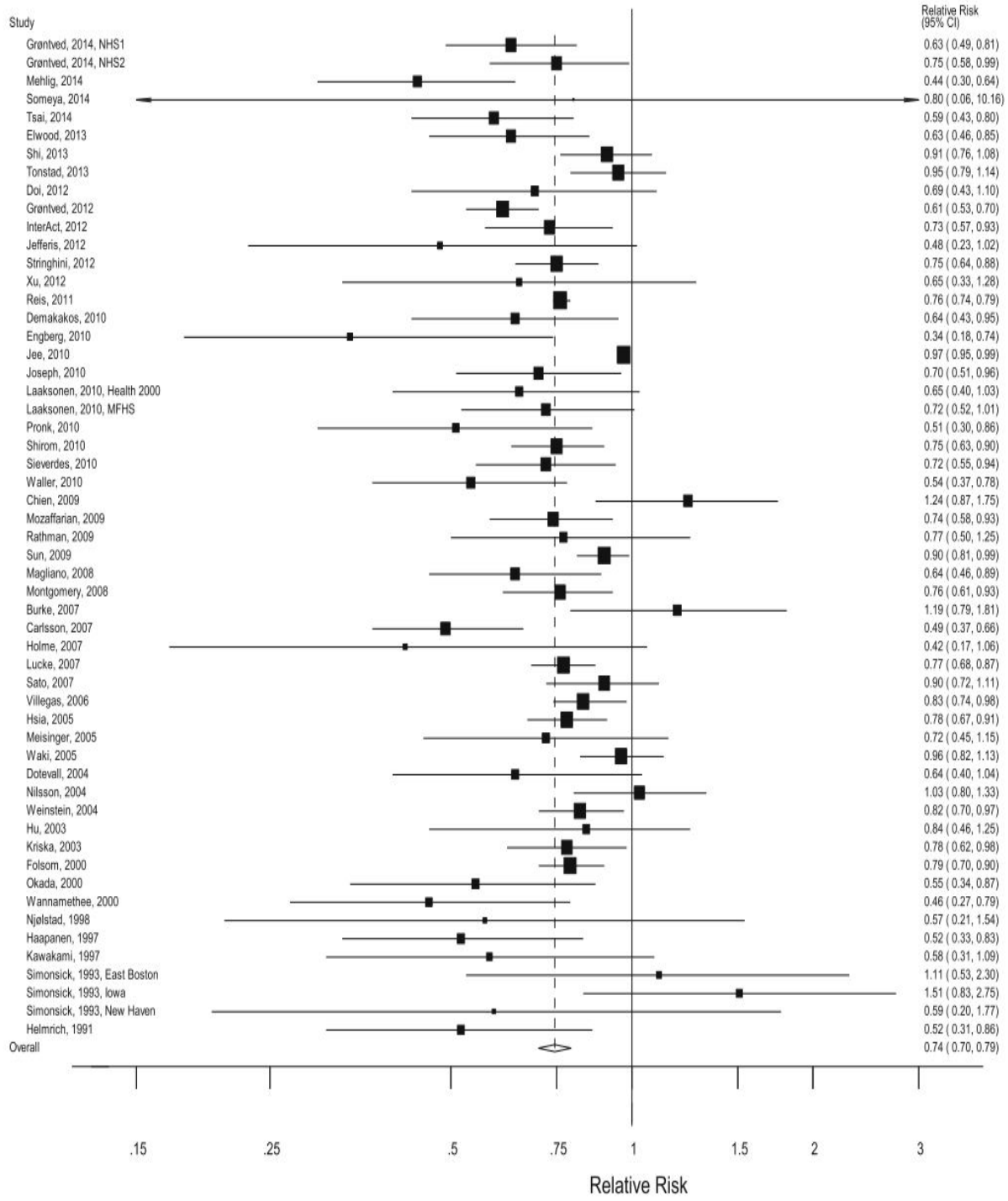

Fig. 2.2 Leisure-time physical activity and type 2 diabetes [13]

2.2.2 Association between physical activity and Hypertension

High blood pressure or hypertension could be considered both as a chronic disease and a bio-medical risk factor, meaning that it can contribute to the advancement of other chronic conditions. In 2010, hypertension was among the three leading risk factors for global disease

burden [103]. Unhealthy diet, harmful use of alcohol and physical inactivity are some of the main behavioural risk factors that can increase the prevalence of hypertension and there are numerous studies which investigate the association between these factors and hypertension. Similar to other epidemiological studies, different sample size and population characteristics, methods and study designs has caused a range of outcomes for these studies.

While A study on a sample of Korean adults reported no association between PA and hypertension regardless of age, body mass index, sleep duration, mental stress, education level, economic status, or frequency of drinking or smoking [188], some other studies limited the association to a specific gender [20, 112, 168] and many have found strong significant associations. A systematic review in 2017 investigating 24 studies on the association between PA and hypertension, reports a linear inverse association for both leisure time PA and total PA (RR 0.94, 95% CI 0.91-0.96) [105]. Figure 2.3 shows the Relative Risk statistics for these studies. Another systematic review recommended PA for the treatment of hypertension although suggesting that more randomised control trial studies are still needed to find out if physical activity can really improve the health of patients with hypertension [154].

Other factors influencing the relationship between PA and hypertension such as: dose-response relationship,temporality, high-risk population and moderating factors has been discussed in [42].

2.2.3 Association between physical activity and Heart disease and Stroke

The diseases of the heart and blood vessels are commonly known as cardiovascular diseases (CVD). Angina and heart attack are two common forms of Coronary heart disease (also known as ischaemic heart disease) and both may happen by shortage of blood supply to the heart. The interruption of the blood supply to the brain might cause a stroke. There are two main types of stroke: *Haemorrhagic* stroke happens when a vessel carrying the blood to the brain leaks or breaks and *ischaemic* stroke accrues when the vessel gets blocked by a clot or cholesterol plaque [162]. Heart failure, cardiomyopathy, congenital heart disease and peripheral vascular disease are other types of CVDs [165].

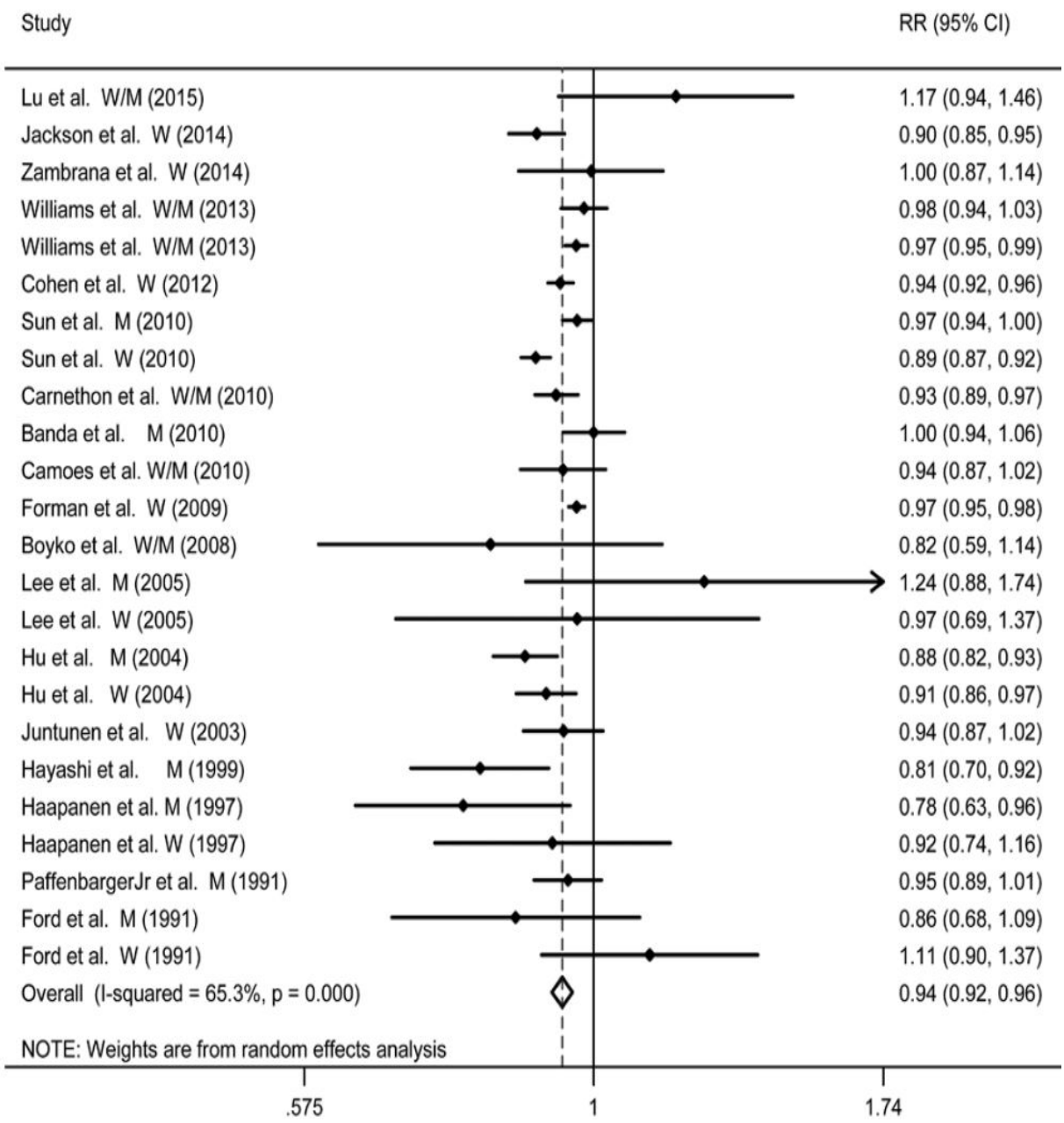

Fig. 2.3 Leisure-time physical activity and Hypertension [105]

According to the World Health Organisation (WHO), CVDs are the leading cause of death globally [182] and the association between different risk factors and CVDs has been investigated in numerous research and studies. Physical inactivity is one of the know risk factors of heart disease and stroke. Similar to other chronic conditions, the results of different studies may differ widely but overall literature reviews have found that there is a negative association between PA and both heart disease and stroke. A 2013 review on 23 prospective cohort study between 2012-2013 examines the association between low, moderate and high PA and heart

disease and stroke [101]. They suggest that moderate PA reduces the risk of PA by 20-30 percent (RR = 0.76, 95% CI 0.71–0.81) and high PA has higher reduction effect (RR = 0.66, 95% CI 0.60–0.72). They reported similar association for heart diseases and stroke. This study found a dose-response relationship between PA and both heart disease and stroke. (Figure 2.4)

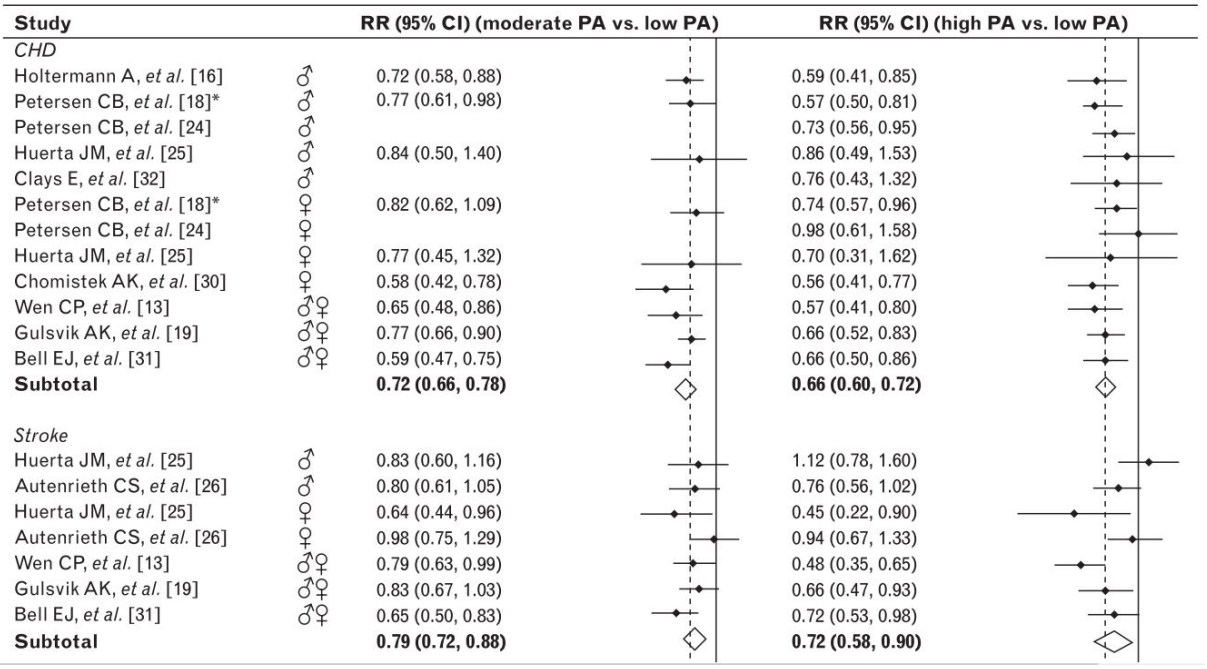

Fig. 2.4 Leisure-time physical activity and stroke and Heart disease [101]

An earlier meta analysis on the association between PA and stroke suggests that higher levels of PA may be required in women than in men to achieve similar levels of significant stroke reduction [44]. some studies discovered a u-shaped relationship between PA and incidence of stroke with the maximum effect accruing with moderate PA [146]. This finding is similar to the result of a recent study on 74,913 Japanese people age 50 to 79 years old [97]. They found that moderate PA may be optimal to reduce the chance of developing stroke.

2.3 Iterative Proportional Fitting (IPF) Re-weighting

In this section I reviewed the use of the Iterative Proportional Fitting (IPF) for updating the weights of a survey in order to match given marginals.

It is often the case that one needs to update a set of survey weights in order to match certain given marginals. Consider the case, for example, of a survey that is few years old and whose information about smoking and obesity rates is not up to date. If a newer source of information regarding smoking and obesity rates is available then one can use it to produce a set of weights which is more representative of the population. Usually the new source of information comes from the Census or from survey data, and it is in the form of a set of one-dimensional marginals.

It is the presence of one-dimensional marginals, rather than a joint distribution, that makes the problem challenging. In the example above, we might have the one-dimensional distributions of smoking and obesity status. If we had the joint distribution we could simply stratify the population in cells defined by smoking and obesity status and apply a multiplicative correction to the weights in each cell to match the desired overall weight. The lack of joint distribution implies that information about the correlation across the variables must come from somewhere else.

The approach taken in the Iterative Proportional Fitting (IPF) framework is that this information comes from the existing survey weights. In other words, the approach consists in finding a new set of weights that is as close as possible to the original set of weights, so that it preserves the correlations, but that satisfies some constraints on the marginals [40, 54, 81].

Unfortunately the IPF is usually presented just as a set of rules for the updating of the elements of a contingency table, without explaining where the algorithm actually comes from and what are the assumptions behind. In this section I derive the algorithm from its basic principles, but present it differently from what it is usually found in the literature. Usually the IPF is presented in terms of multidimensional contingency tables, making the notation difficult in dimensions larger than two. Here I write it directly in terms of the vector of survey weights, so that its implementation is straightforward in any number of dimensions and does not require the manipulation of multidimensional arrays.

2.3.1 Derivation of the algorithm

Let us denote by w_i the survey weight corresponding to individual i. Our problem consists in determining a new set of weights w_i^* which is "close" to the original weights but that matches a new set of marginals:

$$\sum_{i \in \alpha} w_i^* = N_\alpha \ , \alpha \in A \tag{2.1}$$

where α denotes a population cell (for example "male smokers age 50-54" or "overweight females age 55-59"), N_α is the desired number of people in cell α, and A is the set of cells for which we have given marginals. The only requirement on the set A and the numbers N_α is that they are internally consistent. We notice that if instead of marginals we used means of given variables the derivation that follows would be just as easy, and it is left to the reader.

Following [81] we define the distance between the weights w_i^* and w_i in terms of the Kullback-Liebler (KL) divergence. An important reason to choose the KL divergence and not, for example, the mean square error is that using KL divergence automatically ensures that the weights are positive, and therefore there is no need to introduce a positivity constraints on the weights w_i^*, which would lead to a much more complicated problem.

Therefore the problem we need to solve is the following:

$$\min_{w_i^*} \sum_i w_i^* \ln \frac{w_i^*}{w_i} \quad \text{subject to} \quad \sum_{i \in \alpha} w_i^* = N_\alpha \ , \alpha \in A \tag{2.2}$$

The Lagrangian for problem 2.2 is as follows:

$$\mathscr{L} = \sum_i w_i^* \ln \frac{w_i^*}{w_i} - \sum_{\alpha \in A} \gamma_\alpha \left[\sum_{i \in \alpha} w_i^* - N_\alpha \right]$$

where γ_α are Lagrange multipliers, to be determined. Before proceeding it is convenient to introduce the cell indicator variables $\theta_{i\alpha}$ which assume the value 1 if $i \in \alpha$ and 0 otherwise. We can then rewrite the Lagrangian as follows:

$$\mathscr{L} = \sum_i w_i^* \ln \frac{w_i^*}{w_i} - \sum_{\alpha \in A} \gamma_\alpha \left[\sum_i \theta_{i\alpha} w_i^* - N_\alpha \right]$$

where γ_α are Lagrange multipliers. The first order condition is now easily derived:

$$\ln \frac{w_i^*}{w_i} + 1 - \sum_{\alpha \in A} \gamma_\alpha \theta_{i\alpha} = 0$$

and it can be rewritten as follows:

$$w_i^* = w_i \exp\left(\sum_{\alpha \in A} \gamma_\alpha \theta_{i\alpha} - 1 \right) = w_i e^{-1} \prod_{\alpha \in A} \exp(\gamma_\alpha \theta_{i\alpha}) \tag{2.3}$$

Since the γ_α are unknown I are free to redefine them as follows:

$$\gamma_\alpha \rightsquigarrow \ln \gamma_\alpha$$

so that the first order conditions of equation 2.3 become:

$$w_i^* = w_i e^{-1} \prod_{\alpha \in A} \gamma_\alpha^{\theta_{i\alpha}} \equiv F(w_i, \theta, \gamma) \tag{2.4}$$

The equations for the Lagrange multipliers are derived by simply imposing the linear constraints $\sum_i \theta_{i\beta} w_i^* = N_\beta$, which can be written as follows:

$$\sum_i \theta_{i\beta} w_i \prod_{\alpha \in A} \gamma_\alpha^{\theta_{i\alpha}} = eN_\beta \quad \beta \in A \tag{2.5}$$

Equation 2.5 above represents a system of K multilinear equations in K unknown, where K is the total number of cells in A. The IPF is an iterative algorithm to solve this system, that takes advantage of its special multilinear form. The key observation is that since the sum in the left side of equation 2.5 runs over the individuals in cell β (as represented by the term $\theta_{i\beta}$), then $\theta_{i\beta} = 1$ in the term $\gamma_\beta^{\theta_{i\beta}}$, implying that the term γ_β can be brought outside of the sum and we can solve for it conditional on the other γ_α terms.

Formally we start by rewriting equation 2.5 as follows:

$$\sum_i \theta_{i\beta} w_i \gamma_\beta^{\theta_{i\beta}} \prod_{\alpha \in A, \alpha \neq \beta} \gamma_\alpha^{\theta_{i\alpha}} = eN_\beta \quad \beta \in A$$

Next we notice that since $\theta_{i\beta}$ is binary then $\theta_{i\beta}\gamma_\beta^{\theta_{i\beta}} = \theta_{i\beta}\gamma_\beta$ and therefore the equation above can be rewritten as:

$$\gamma_\beta \sum_i \theta_{i\beta} w_i \prod_{\alpha \in A, \alpha \neq \beta} \gamma_\alpha^{\theta_{i\alpha}} = eN_\beta \quad \beta \in A$$

This equation can then be rewritten by solving for γ_β:

$$\gamma_\beta = \frac{eN_\beta}{\sum_i \theta_{i\beta} w_i \prod_{\alpha \in A, \alpha \neq \beta} \gamma_\alpha^{\theta_{i\alpha}}} \quad \beta \in A \tag{2.6}$$

The equation above naturally leads to an iterative algorithm, since it allows to express one unknown (γ_β) in terms of all the others. The implementation of the algorithm is extremely simple, since it requires only to define the function F of equation 2.4. In fact, defining the vectors w^β and $\gamma_{-\beta}$ as follows:

$$w_i^\beta \equiv w_i \theta_{i\beta} \,, \quad \gamma_{-\beta} = (\gamma_1, \ldots, \gamma_\beta = 1, \ldots, \gamma_K)$$

we immediately see that equation 2.6 can be written as:

$$\gamma_\beta = \frac{N_\beta}{\sum_i F(w_i^\beta, \theta, \gamma_{-\beta})} \quad \beta \in A \tag{2.7}$$

The IPF algorithm start by setting all the γ_β to 1 and then iterating as follows:

$$\gamma_\beta^{(t)} = \frac{N_\beta}{\sum_i F(w_i^\beta, \theta, \gamma_{-\beta}^{(t-1)})} \quad \beta \in A \tag{2.8}$$

It is matter of simple algebra to verify that the algorithm above is exactly the IPF algorithm for contingency tables. The form of equation 2.8 is particularly simple to implement since it acts directly on the survey weights, and not surprisingly it enjoys the same property of fast convergence exhibited by the IPF algorithm in its more common form.

The algorithm was implemented by Federico Girosi in `R` in the function `IPF.Reweight`, which is reported in the Appendix.

2.4 Causality, matching methods and instrumental variables

In many epidemiological studies, researchers are interested in the effect of some exposure on some outcome. This effect could be either causation or association. To define causation and association and to illustrate the differences between them, lets limit our definition to dichotomous variables. Consider a dichotomous variable A as the "Treatment" variable or "Exposure". If a person is treated $a = 1$ and $a = 0$ if not. Y is the outcome variable. $Y_{a=1}$ is the outcome variable that would have been observed under the treatment variable $a = 1$ and

$Y_{a=0}$ is the outcome that would have been observed under the absence of the treatment, $a = 0$ for the same person. (In this thesis, Y could be the sum of the hospitalisation payments for the next day, or chance of developing chronic health condition in the future.) These two outcomes are called potential outcomes and for each person, one of these two would be the observed outcome (factual) and the other one is considered as the counterfactual outcome.

The causal effect is defined as the difference between the average outcome if "all" people in the population would have been treated and the average outcome if "no one" in the population would have been treated. The causal effect can be reported as any of the forms below [70]:

- Causal risk difference:

$$E(Y_{a=1}) - E(Y_{a=0}) \tag{2.9}$$

- Causal Risk ratio:

$$\frac{E(Y_{a=1})}{E(Y_{a=0})} \tag{2.10}$$

- Causal Odds ratio (with a binary outcome):

$$\frac{P(Y_{a=1}=1)\,/\,P(Y_{a=1}=0)}{P(Y_{a=0}=1)\,/\,P(Y_{a=0}=0)} \tag{2.11}$$

In this study, the causal effect could be the difference in the average hospitalisation payments if everyone in the study population had sufficient physical activity and if no one had sufficient physical activity. The fundamental problem of causal inference comes from the above definition and the fact that it is not possible to see both potential outcomes for the same person, however, under some assumptions, it is possible to consistently estimate these average values from the observed data.

The definition for "association" seems similar to the definition of "causation". The only difference is that the targeted groups are different. In association, we look at the average difference outcome for those who have received and those who have not received the treatment

(two separate subset of population) while for the causation, we consider the whole population for both groups. Figure 2.5 can further illustrate this difference.

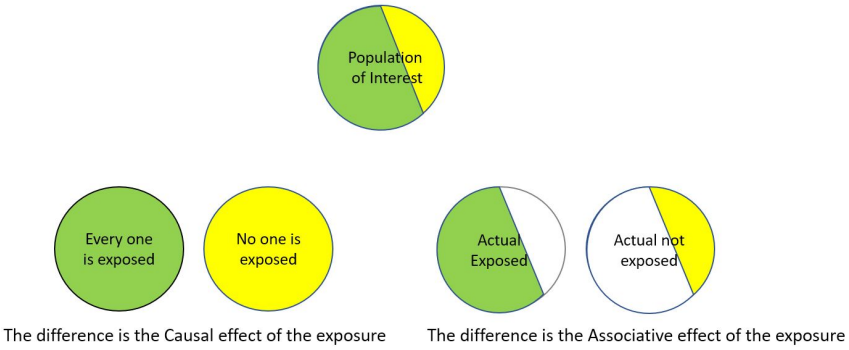

Fig. 2.5 The difference between population in Causation and Association

The formal definitions of association can be written as Equation 2.12 to 2.14

- Association risk difference:

$$E(Y|a=1) - E(Y|a=0) \tag{2.12}$$

- Association Risk ratio:

$$\frac{E(Y|a=1)}{E(Y|a=0)} \tag{2.13}$$

- Association Odds ratio (with a binary outcome):

$$\frac{P(Y=1|a=1) \,/\, P(Y=0|a=1)}{P(Y=1|a=0) \,/\, P(Y=0|a=0)} \tag{2.14}$$

2.4.1 From association to causation

If the treatment assignment is done completely randomly and the study population is large enough, it is possible to deduce that the average outcome for a sub-sample of the population with treatment $A = a$, would be the same of the average outcome for the whole population.

In this case it is possible to estimate the causal effect from the association. These required conditions could be listed as below:

- **Stable unit treatment value assumption (SUTVA)**: observations do not interfere with each other and treatment assignment of one observation does not affect the outcome of the others.
- **Consistency**: the potential outcome under treatment a, is equal to the observed outcome if the actual treatment received is $A = a$.
- **Ignorability**: given pre-treatment covariates X, treatment assignment is independent from the potential outcomes.
- **Positivity**: for every set of values for X, treatment assignment was not deterministic

If these conditions are met, it is possible to use association as the causation.

$$E(Y|A = a, X = x) = E(Y_a|A = a, X = x) = E(Y_a|X = x) \tag{2.15}$$

Randomised experiments are usually designed with these conditions in mind. Researchers try to assign the treatment or exposure randomly to the sample so both treatment and control group would be similar in all aspects but the treatment assignment. The problem with randomised experiments is that these studies are usually costly, their sample sizes are small and in some studies such as studying the effect of smoking, it is not ethical to conduct a randomised controlled trial.

Unlike data from randomised experiments, observational data are cheaper and more abundant, but in these studies, the treatment assignment is usually affected by some other pre-treatment variables (confounders) which violates the Ignorability assumption. It also would be possible to find a subset of observations with $X = x$ which just appear either in control or treatment which yields to Violation of positivity assumption. Matching and Instrument Variables Analysis are some methods that are used to randomise the treatment assignment in the observational data.

2.4.2 Matching

Matching is a technique that is used to prepare the data for the counterfactual analysis. It finds similar people from the control and the treatment groups and deletes the rest in order to get balanced groups of control and treatment. When the two groups are balanced, the distribution of the covariates (X) are similar in the control and the treatment group. In this situation, as long as there are no unobserved covariates that correlate with both treatment and outcome, the difference between the average outcomes for two groups is equal to the causal effect and controlling for the X is not needed any more. However, in practice, the two groups are only approximately balanced and we still need to control for X using a statistical tool such as a regression model.

There are different methods available to perform matching. Methods could be k-to-k (a person from treatment group matched to a person from control group) or many to many. Some methods involve using an observation more than once (non-bipartate) similar to sampling with replacement while most of the methods use each observation only once.

Propensity Score Matching (PSM)

Propensity score matching includes a popular group of matching methods. In this method, a score is calculated for each observation and the distance between two observations is calculated based on the assigned score. Propensity Score is the chance of the observation given covariates X, to belong to the treatment group. It could be computed using a logistic method or any other binary classification algorithm. Finding the matched pairs or matched groups based on their propensity score could be done through different algorithms. Greedy matching, nearest neighbour matching, genetic matching and optimal matching are some of the most common algorithms.

Some researchers believe that propensity score should not be used for matching [92].The model that is used to calculate the score could be wrong and even if the model is correct, the algorithm does not guarantee that two groups matched on propensity score are balanced.

Coarsened Exact Matching (CEM)

CEM [77] is a member of a class of matching method called Monotonic Imbalance bounding. To apply CEM, the first step is to cut the continuous and ordinal variables into intervals and combine some of the levels of categorical variables. Then the algorithm divides the population into cells defined by the categories and intervals of the variables and matches treated units with control units falling into the same cells. The intervals of the continuous variables and the categories that are selected in the first step define the limits of these cells. Cells containing observations which are all controls or all treatments are removed. One can change the level of matching by changing the predefined intervals and categories. CEM achieves a better uni-variate and multivariate balance of covariates for treatment and control groups and reduces the estimation bias more compared to other popular matching methods such as Propensity Score Matching (PSM) and Genetic Matching [144].

CEM can do a k-to-k or many-to many matches in each cell. In the first case for each treated unit in the cell, a match from the control group is selected by random or by a nearest neighbour algorithm. In the second case, all treated units in each cell are weighted one and control units get a weight such that weighted control and treatment units become balanced. Equation 2.16 shows how the weights are calculated for each matched unit.

$$w_i = \begin{cases} 1, & i \in T^s \\ \frac{m_C}{m_T}\frac{m_T^s}{m_C^s}, & i \in C^s \end{cases} \tag{2.16}$$

T^S are the treated units in cell s, C^s are the control unit in cell s, m_T^s and m_C^s are numbers of treated and control units in cell s and m_C and m_T are total number of matched control and treatment units in the data set.

The imbalance between two groups could be measured in different ways. Simplest way is to compare uni-variate absolute difference of the means in the two groups [77]:

$$I_j = |\bar{X}^w_{a=1,j} - \bar{X}^w_{a=0,j}|, \quad j = 1, \ldots, k, \tag{2.17}$$

$\bar{X}^w_{a=1,j}$ denotes weighted means of variable X_j for the treated group and weights are given by the matching algorithm. Another option is using Standardised Mean Difference (SMD) which compares the mean difference for all covariates and scales the differences. For the continuous variables it is defined as:

$$SMD_j = \frac{\bar{X}^w_{a=1,j} - \bar{X}^w_{a=0,j}}{\sqrt{\frac{s^2_{a=1,j} + s^2_{a=0,j}}{2}}} \tag{2.18}$$

Whereas $s^2_{a=1,j}$ and $s^2_{a=0,j}$ represent the sample variances of the variables in the treatment and control groups respectively.

While these numbers represent the overall average differences for the two groups, they do not show how the empirical distributions of two groups are different, so the joint distribution of the variables in two groups may be different.

Iacus et al.[76] have suggested an L_1 distance measure that measures the multivariate differences between the probability of each variables in the two groups. The outcome value is a number between 0 (two distributions overlay completely) and 1 (two distributions are completely different).

$$L_1(f, g; H) = \frac{1}{2} \sum_{\ell_1 \ldots \ell_k \in H(X)} |f_{\ell_1 \ldots \ell_k} - g_{\ell_1 \ldots \ell_k}| \tag{2.19}$$

In the above equation, $f_{\ell_1 \ldots \ell_k}$ is the relative frequency for observations belonging to the cell with coordinates $\ell_1 \ldots \ell_k$ of the multivariate cross-tabulation, and similarly, $g_{\ell_1 \ldots \ell_k}$ is the relative frequency for observations belonging to the cell with coordinates $\ell_1 \ldots \ell_k$ and $H(X)$ is the set of cells generated by the Cartesian product of all levels of different variables. This value is reported by the imbalance function in the CEM `R` package [75].

2.4.3 Instrumental Variable (IV)

In section 2.4.1 I talked about some assumptions on the data to derive casual relationship from association. I mentioned that the *Ignorability* is the assumption that is often not met in observational studies.

Even in randomised controlled trials, the participants non-compliance may result in the violation of the *Ignorability* assumption. While matching methods can adjust for the known and observed confounders between control and treatment groups, yet they can not adjust the effect of the unobserved or unknown confounders. The term "confounding bias" is used for the bias due to confounders that are not included in the analysis [34].

In a regression model that uses the treatment variable to predict the outcome, presence of unmeasured confounders yields to correlation of the treatment X and the model's error term. Such X variable is known as an *endogenous* variable.

Instrumental Variable (IV) estimation is a common method in economic studies to solve the problems caused by uncontrolled confounders and to eliminate the confounding bias. This method unlike matching or regression methods, dose not need to know or observe all the confounders that cause bias. The method is based on some variable (Z) that is associated with the treatment variable (X), it is not associated with unmeasured confounders after controlling for measured confounder and it is only associated with the outcome (Y) indirectly through X [5] (Figure 2.6). Such variable is called an Instrumental Variable.

Finding an instrument that satisfies these assumptions is always challenging. In many cases, the chosen variables is not highly associate with the treatment variable. Such instrument is considered a weak instrument which will cause some issues. When IV is weak, the violation of the mentioned assumptions may yield biased estimators and results in wide confidence intervals that may significantly reduce the power of the analysis [28, 51, 156].

Earlier examples of using IV in epidemiology can be found in the literature. Hearst et al. examine the effect of military service during the Vietnam era on subsequent mortality, and use the military draft lottery of 1970 to 1972 as the instrument variable which they call "randomised natural experiment" [69]. Physician prescribing preference [145], day of the week of hospital admission [71], the railroad division index, which measures the extent to which a

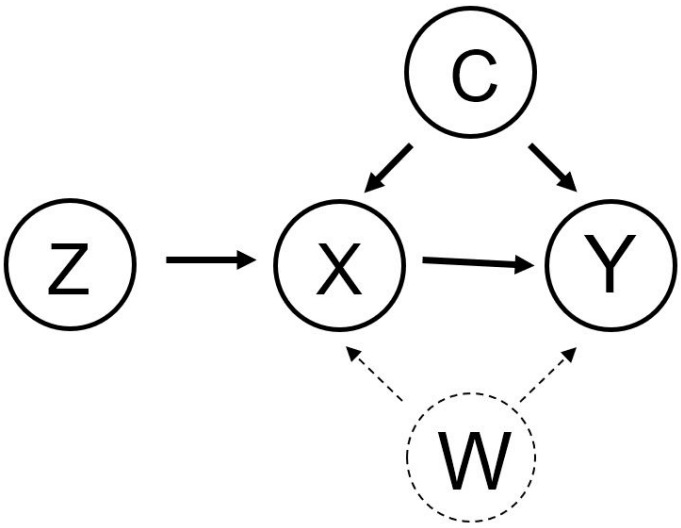

Fig. 2.6 The relationship between IV (Z), treatment (X), observed confounders (C), unobserved confounders (W), and outcome (Y) in an Instrumental Variable estimation.

metropolitan area was divided into subplots by railroad tracks [14] and distance to speciality care provider [113] are some other examples of IVs used in epidemiology.

There are many different approaches for IV analysis [95]. IV estimation methods often involve two stages. The first stage uses the IV to predict the treatment variable and the second stage finds the effect of the predicted treatments in the previous stage on the outcome.

Two-stage least squares (2SLS) method is one of the most common two-stage methods. In this methods two linear regression models are used for both stages of the analysis. Observed confounders can be used in both models. Equation 2.21 represents 2SLS method.

$$X = \alpha_0 + \alpha_z Z + \alpha_c C + \epsilon_1 \tag{2.20}$$

$$Y = \beta_0 + \beta_{IV} \hat{X} + \beta_c C + \epsilon_2 \tag{2.21}$$

In Equation 2.21, C is the measured covariates, $\hat{X}$ is the predicted value from the first regression and β_{IV} is the IV estimator. In case of binary IV and in the absence of C, β_{IV} can be calculated using Equation 2.22 which is called the Wald estimator.

$$\beta_{IV} = \frac{p(Y|Z=1) - p(Y|Z=0)}{p(X|Z=1) - p(X|Z=0)} \tag{2.22}$$

The numerator in 2.22 is known as *Intention to Treat* (ITT) estimator and shows the association between the treatment assignment by the IV and the outcome. The denominator is a measure of compliance and shows the difference between treatment rates between levels of the instrument.

2.5 Predictive Modelling of Costs and Expenses

Analysing health care utilisation data can answer many important questions about health care systems. Hospital length of stay (LOS), hospital admission costs or payments to the hospitals, and total health care expenditure for a single person or a group of people are some of the popular topics to model and predict in health care economics [43, 67]. The predicted outcome (dependent variable) can be a continuous number (for cost or LOS), a categorical response for buckets of costs or periods of LOS, a binary outcome to detect outliers in cost or to find long/short-stay hospital admissions [124, 191] or an integer count such as number of visits to general practitioner or number of hospital admissions. The data for all these outcomes have some characteristics in common [29, 111]:

- The outcomes are non-negative.
- There are significant outcomes with zero value.
- They usually have a heavily right-skewed distribution (Figure 2.7).

The positively skewed dependent variable and the pick in zero values have made the prediction of health expenditure a challenging problem. Although Ordinary Least Square (OLS) is the traditional way of modelling health expenditure by ignoring the data characteristics [25, 109], several methods have been proposed in the literature to deal with these problems.

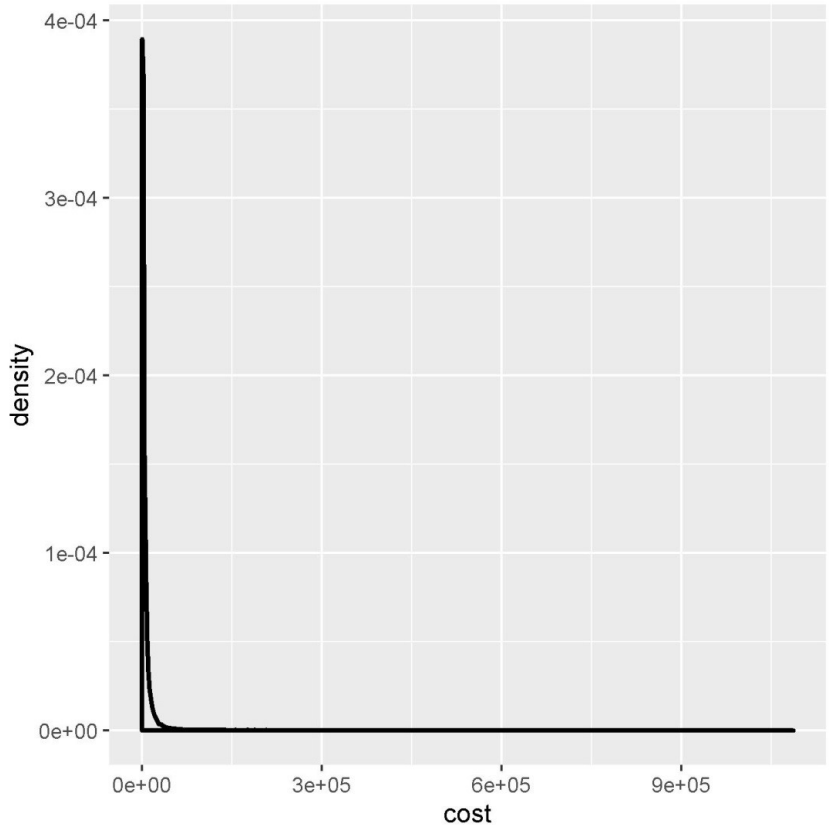

Fig. 2.7 Distribution of hospital admissions payments, An example of heavily right skewed data

Various models perform differently based on the data and the problem and there is no best model to be used in all problems. It has been shown that the choice of model can result in big difference in the predictions and results [132]. To deal with the spike of zeros, two-part models (2PM) and generalised Tobit models have been suggested [83]. A two-part model does the modelling in two stages: first, identifies no costs and positive costs and then uses a conditional regression to predict the positive costs. Generalised Tobit models are regression models which the dependent variable is constrained in some way [3]. These methods have been used vastly in the literature but still there are criticisms and debates about the usefulness of these methods [29, 48, 125]. To handle the skewed distribution, two broad classes of models are used mostly [109]:

- Transformation of the dependent variable (usually with log function).
- Using OLS on transformed data, Generalised linear models.

Literature of the health expenditure modelling is so vast and rich. Different models and machine learning algorithms have been proposed and used. In the next part of the report, the most common methods that are used in the literature will be introduced briefly and some of their cons and pros will be discussed.

2.5.1 Different approaches for modelling health expenditure data

Linear regression

As Buntin and Zaslavsky stated, "Given our research objectives, none of the estimators suggested in the health econometrics literature could be rejected out of hand" [29]. Linear Regression is one of the simplest ways to make a prediction model. Using this model, I simply ignore the characteristics of the data. The good aspect of using linear regression is that it is easy and fast to implement even when dealing with millions of observations and high dimensional data. The model directly works on the original scale of the dependent variable (e.g. dollars for prediction of costs) and needs no prior transformation [63]. In these models, the dependent variable y is a linear combination of covariates:

$$y = x\beta + \epsilon \tag{2.23}$$

In Equation 2.23, y is the original or untransformed dependent variable, x is a row vector of covariates, ϵ is an additive error term that is independent of the covariates x, and β is the vector of weights to be calculated. Ordinary Least Squares (OLS) is the most common method to estimate the weights by minimising the sum of the squares of the differences between the observed dependent variable and those predicted by the linear function. These models are sensitive to outliers and perform weak if the sample size is small to medium [118] but with a large dataset with millions of observations, it is found in the literature that OLS may perform well in relation to more sophisticated models [63].

Linear regression on transformed dependent variable

One way to deal with skewed data is to transfer it to make it more symmetric and closer to a normal distribution. Cox-Box model is a general form of such a transformation with one variable [27] (2.24).

$$\begin{aligned} \frac{y^{\lambda}-1}{\lambda} &= x\beta+\epsilon \qquad if \quad \lambda \neq 0 \\ ln(y) &= x\beta+\epsilon \qquad if \quad \lambda = 0 \end{aligned} \tag{2.24}$$

Log transformation is the special case of Cox-Box model with $\lambda = 0$ and is the most popular transformation found in the literature of health care modelling. If the log transform reduces or removes the skewness of the data, predictions based on logged models would be more proper than direct analysis of the untransformed dependent variable [108]. Figure 2.8 shows the distribution of payments in Figure 2.7 after the log transformation. As it shows, after the transformation, the distribution is closer to a normal distribution and more suitable for linear regression.

Square-root transformation is another special case of Cox-Box model with $\lambda = 0.5$

$$\sqrt{y} = x\beta+\epsilon \tag{2.25}$$

The most important issue with transforming data is that it changes the scale of the output. In the case of predicting cost, the estimated output would be log dollars while the desired output is often untransformed dollars. Re-transforming the output could lead to a wrong and biased estimation of the output.

$$\begin{aligned} ln(y) &= x\beta+\epsilon \\ y &= e^{x\beta+\epsilon} \\ E(y|x) &= E(e^{x\beta})E(e^{\epsilon}|x) \end{aligned} \tag{2.26}$$

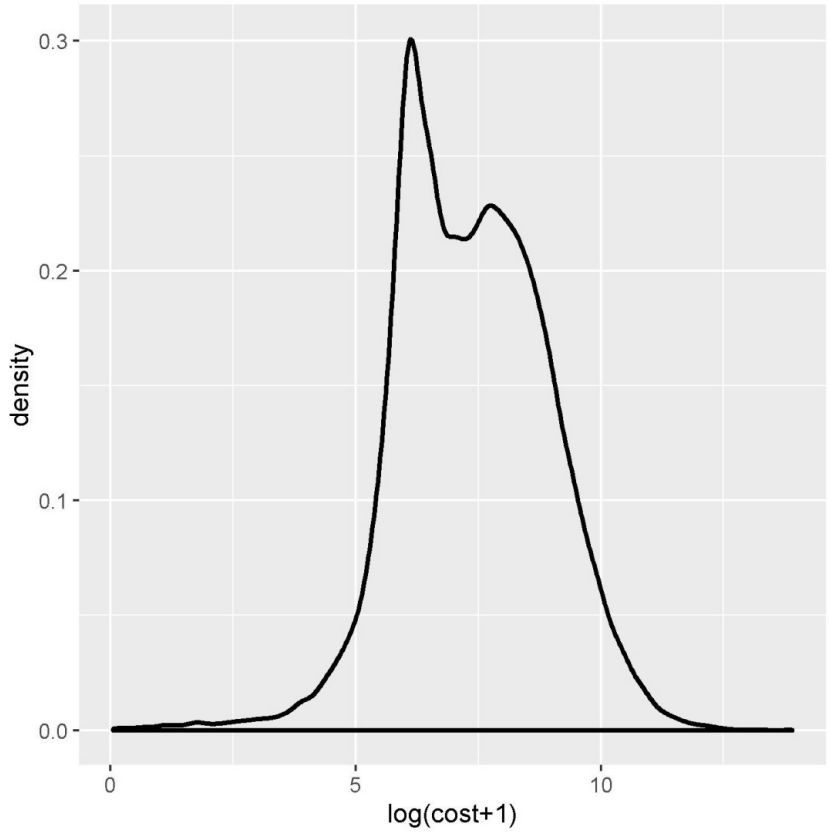

Fig. 2.8 Distribution of hospital admission payments after log transformation

In equation 2.26, the term $E(e^{\epsilon}|x)$ depends on the distribution of the error. If the error is distributed normally, the $E(e^{\epsilon}|x)$ would be equal to $e^{0.5\sigma_{\epsilon}}$. Duan [47] has shown that if the error is not normally distributed but it is homoscedastic, meaning that the variance is constant for all different independent variable, the last term in 2.26 can be replaced with *Duan's smearing factor* which is a function of sample size and number of parameters of the regression. In health-care cost estimation, the typical value for Duan's smearing factor is between 1.5 and 4 meaning that ignoring this factor could result in massive underestimation of average cost [84]. In the case of heteroscedasticity of the error term, meaning that the error variance is not homogeneous [62], more sophisticated smearing factors should be considered to avoid bias in the modelling [84, 110].

Generalised linear models (GLM)

Generalised linear model consists of 3 main parts [118]:

1. Linear part which is similar to linear regression:

$$\eta_i = x_i\beta + \epsilon_i \tag{2.27}$$

2. A link function $g(.)$ which explains how the expected values of dependent variable and linear part are related to each other:

$$g(E(y_i)) = \eta_i \tag{2.28}$$

3. Dependent variables with a probability distribution from an exponential family (Gaussian, the binomial, the Poisson, the negative binomial, the gamma and the inverse Gaussian). The choice of the distribution function defines the relation between the variance and the mean of the dependent variable.

$$var(y_i) = \Phi V(\mu_i) \tag{2.29}$$

In equation 2.29, Φ is a constant and $V(.)$ is a function defined by the selected distribution. The benefit of these models is that unlike transformed models, the estimation is happening on the original scale of the data, so back transformation is not needed. GLMs can be used both in one part and two part models. Identity link and log link are the most common links used in the literature of health-care expenditure modelling. In the case of identity link, the covariate act additively on mean and could be compared to linear regression. For log link, covariates act multiplicatively on mean [84]. Table 2.1 shows some of the popular links and the distribution that are commonly used with each of them.

Support Vector Machines (SVMs)

Support Vector Machines can be used for both classification and regression. SVM was basically developed to solve two-class classification problems. The idea is to map input vectors onto a very high dimension feature space and then separate different classes using an optimal linear decision surface [36] but with some minor changes, it can be also used for regression and

distribution	natural link function	variance function
Gaussian	μ	1
Bernoulli	$log(\frac{\mu}{1-\mu})$	$\mu(1-\mu)$
Binomial	$log(\frac{\mu}{1-\mu})$	$n\mu(1-\mu)$
Poisson	$log(\mu)$	μ
Negative Binomial	$log(\mu)$	$\mu+\mu^2/k$
Gamma	$\frac{1}{\mu}$	μ^2
Inverse Gaussian	$\frac{1}{\mu^2}$	μ^3
Quasi	$g(\mu)$	$V(\mu)$

Table 2.1 Exponential families of distributions with their links, mean and variance functions [25]

multiclass classifications. To get the best result out of SVM, some hyper parameters should be set tuned. The slow training process is another problem with SVM. The two studies in [65, 153] compare SVM with other machine learning algorithm for prediction of LOS.

Artificial Neural Networks (ANN) and Deep Learning (DL)

Neural Networks are nonlinear statistical models which are used for both regression and classification tasks. According to Trevor et al., "The central idea is to extract linear combinations of the inputs as derived features, and then model the target as a nonlinear function of these features. The result is a powerful learning method, with widespread applications in many fields" [169].

Recently with the advancements in processing powers and abundance of data, more complex models with more number of layers are developed under the name of Deep Learning (DL). ANNs and DL models can handle noisy data and in some machine learning tasks such as image processing and natural language processing perform much superior than other methods. These models require a large number of data to be trained. By increasing the availability of digital health data and advancement in transfer learning method, DL is getting more attention in health care analytic and many published articles can be found that apply ANN and DL for predicting hospital LOS [64, 65, 100, 170, 184].

Other models

Apart from statistical models and machine learning methods discussed above, there are different models and more complicated techniques used in prediction of health expenditure. A valuable summary of most of other methods used in modelling health-care expenditure are provided in [83, 84, 86, 118].

2.5.2 Risk Adjustment Systems

According to "American Academy of Actuaries", "Risk adjustment is an actuarial tool used to calibrate payments to health plans or other stakeholders based on the relative health of the at-risk populations" [141]. In a risk adjustment system, people who are probable to have higher health expenditure are expected to get a higher score based on the algorithm used in that risk adjustment system. By early detection of high-risk individuals, proper medical could be provided to them and their future medical expenditure could be reduced [32, 116].

In the previous section, some of the most popular methods of modelling health care expenditure were explored. In this section, first, I reviewed some of the "inputs" of the models which act as "regressors" of the regression and then introduce some of the available health risk adjustment systems and discuss some of the variables that are used to compare and evaluate these models.

Regressors of risk adjustment models [187]

Age and gender: Age and sex are most of the time available in the data but they have a weak prediction power.

Prior year expenditure: Prior year expenditures are correlated with next year expenditures and seem to be the best single predictor for next year total expenditures and they are often used with other regressors.

Diagnosis-based risk adjustment: It seems natural that diagnosis information for an individual is related to their future health expenditure. This information can be available from hospital records or insurance claim data. Diagnosis information is often coded using International Statistical Classification of Diseases and Related Health Problems, mostly know

by the short name International Classification of Diseases (ICD). The problem is that there are numerous ICD codes so there should be some grouping before using them as regressors of a risk adjustment model or any predictive model. Table 2.2 shows some of the famous risk adjustment systems and a short description of their algorithms. The top three items in table 2.2, "Ambulatory Care Group (ACG) system" [178], the family of "Diagnostic Cost Group (DCG)" [8] and the "Chronic Disability Payment System (CDPS)" [96] were developed primarily for US Medicaid disabled enrollees and are the 3 most famous risk adjustment systems available in the literature.

Information derived from prescription drugs: Prescribed drugs have been used to detect the presence of chronic conditions. Similar to ICD codes, prescriptions also need grouping before being used in the models. Chronic Disease Score (CDS) [175] and Pharmacy Cost Groups (PCGs) are two groupings developed for prescriptions.

Self-reported health information: Self-reported information from surveys could be another source of information for risk adjustment models. Overall summary of health status described as excellent, very good, good, fair, poor and functional health status, telling how well the individual can perform different daily works are usual examples of self-reported information.

other information: Mortality, family size, marital status, employment etc. are some of the other information that can be used in the risk adjustment models.

Metrics for evaluation of risk adjustment models

When a model is built, we need to know how accurate it works for the prediction and which model excels the others. There are different metrics reported in the literature. Here are three of the most common metrics. The first two evaluate the model in individual level and the third one measures the predictive accuracy in group level [181].

R-squared: R-squared (R^2) also known as the coefficient of determination, is the most general metric used in the evaluation of regression models. R^2 explains the proportion of the variance in the dependent variable (health-care expenditure for example) that is predictable from the independent variables. It usually gets any value between 0 and 1 and a value closer to 1

System	Developer	Input	Short description
Ambulatory Care Groups [ACGs]	Johns Hopkins	Diag[1]	Mutually exclusive groupings of diagnoses based on clinical judgement and resource implications.
Chronic Disability Payment System [CDPS]	Kronik/ UCSD	Diag	Mutually exclusive groupings of diagnoses based on clinical judgement and resource implications. Beneficiaries may be assigned to multiple categories.
DxCG DCGs	DxCG	Diag	Categories are defined on the basis of clinically coherent diagnosis groups, hierarchically combined into HCCs. Individuals may have multiple HCCs episodes.
Impact Pro	Ingenix	Med[2]+Rx[3]+ Use[4]	Episodes defined on the basis of diagnosis, procedure, and drug data; each member may have episodes falling into multiple categories.
DxCG RxGroups	DxCG	Rx	Drug therapy categories can be assigned to one or more categories.
Medicaid Rx	Glimer/UCSD	Rx	Prescription drugs mapped to medical condition categories. Cost predicted based on medical condition and age/ gender categories groupings.
Ingenix PRG	Ingenix	Rx	Groupings of prescription drugs mapped to diagnostic categories. The patient may be assigned to multiple categories.
Clinical Risk Groups	3M	Diag	Mutually exclusive categories based on diagnostic and procedural criteria.
Ingenix ERG	Ingenix	Med+Rx	All treatment information used in episode definition

Table 2.2 Risk adjustment systems [151, 181]

[1] ICD 9 diagnosis codes .
[2] ICD 9 diagnosis codes and procedure information.
[3] Pharmacy NDC codes.
[4] Measure of previous utilisation.

means a better fit in the model. A negative R^2 for a model implies that using the *mean* value generates a better prediction.

$$R^2 = 1 - \frac{\sum (y - \bar{y})^2}{\sum (y - \hat{y})^2} \tag{2.30}$$

In 2.30, y is the observed value, $\hat{y}$ is the predicted value and $\bar{y}$ is the mean of observed values. An issue with R^2 is that adding more inputs or regressors to the model will increase the value of R^2 regardless of whether the added variable improves the predictions. **Adjusted R-squared**

(R^2_{adj}) is used to address this issue. Adding new regressors to the model increases the R^2_{adj} only when the increase in R^2 is more than what expected by chance.

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{2.31}$$

Where n is the number of samples and p is the number of regressors.

Mean Absolute Prediction Error (MAPE or MAE): As the name implies this metric calculates the average value of the absolute differences between predicted value and actual value of the dependent variable and it could be expressed as the percentage of the average of the actual values.

$$\text{MAE} = \frac{\sum |y - \hat{y}|}{n} \tag{2.32}$$

Predictive Ratio: Instead of measuring the error for individuals, predictive ratio adds up the predicted values for each subgroup and compares its ratio to the sum of the actual values of the same subgroups. A value closer to 1 shows a better model.

2.6 Natural Language Processing: Topic modelling and Bi-LSTM

2.6.1 Topic Modelling with Latent Dirichlet Allocation(LDA)

A topic model is a statistical method which is used to analyse, organise and summarise large volume of text data. The topic model algorithm finds clusters of words that appear together more frequently. These clusters are considered as abstract topics.

Latent Dirichlet Allocation (LDA) is a popular example of a topic model [24]. It is a generative probabilistic model of a corpus. Intuitively, one can assume that each specific topic consists of some words that usually appear in that topic. A text document, is a combination of different words so each document is mixture of the topics that those words represent. LDA is technique that tries statistically define the topics and their distributions over the documents. LDA can be described as a generative process. It assumes that each text document is a

random mixture over the hidden topics and each topic is a specific distribution over words. This imaginary generative process for each document is as follows:

1. The model assumes that there are defined number of topics (D) over specific vocabulary (with N unique words).
2. The algorithm assigns a random distribution of different topics to each document (θ).
3. To select each word of the document, Randomly choose one of the assigned topics and randomly choose a word from that topic.

LDA assumes that the prior distribution of topics over documents (θ) and the prior distribution of words over topics(β) is *Dirichlet* distribution with parameters α and η (Equations 2.33, 2.34).

$$\theta \sim Dirichlet(\alpha) \tag{2.33}$$

$$\beta \sim Dirichlet(\eta) \tag{2.34}$$

In this algorithm, the documents are the only observed data and the final goal of the algorithm is to estimate the topic structure (distribution of words in each topic (β) and the distribution of topics over each document (θ)) by reversing the imaginary generative process. Which in plain words, it means to estimate which words are important for which topic and which topics are important in each document, respectively.

Formally, using a probabilistic modelling perspective, the LDA generative process can be described as the following equation [23]:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d)(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:k}, z_{d,n})) \tag{2.35}$$

In the above equation, β_k is a distribution over vocabulary, $\theta_{d,k}$ is the topic proportion of topic k in document d, $z_{d,n}$ represents the assigned topic to the nth word in document d and $w_{d,n}$ is the nth word in document d. The pre-defined number of topics is D and K is number of documents. Equation 2.35 formulates the described generative process as the joint distribution of observed variables (documents) and hidden variables (distribution of topics over documents and distribution of words over topics). Another way to represent Equation 2.35 is using graphical models. Each node in the graphical model in Figure 2.9 is a random variable and the node with observed variable is shaded. Each edge indicates dependence and the rectangles show replicated variables.

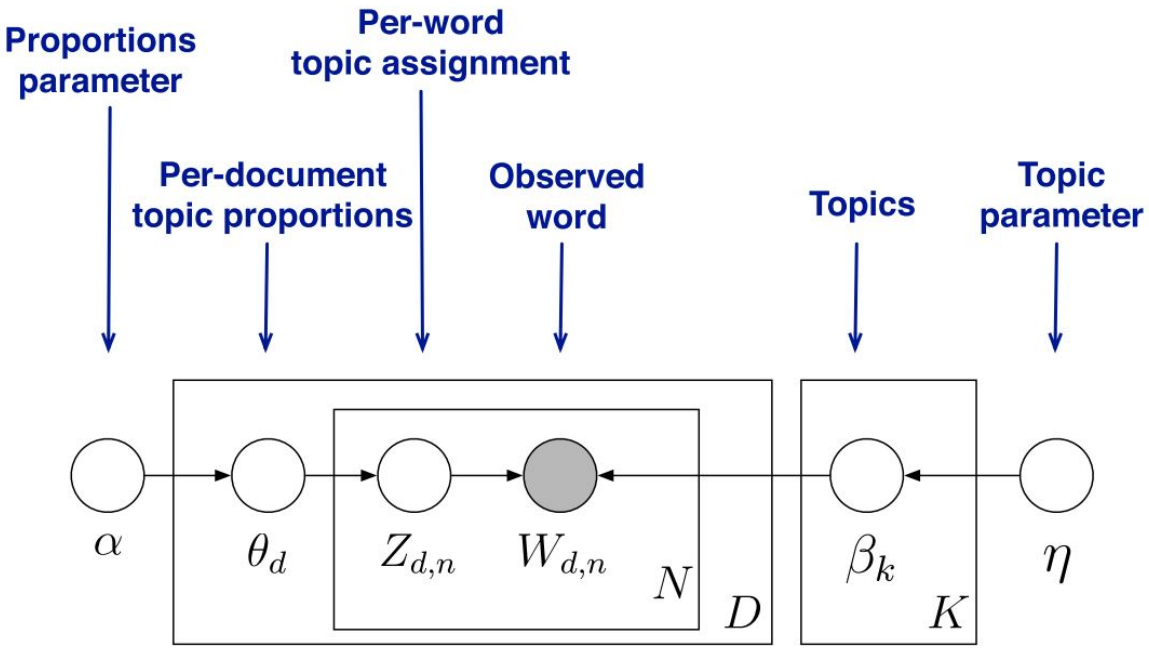

Fig. 2.9 The graphical model for Latent Dirichlet Allocation. [23]

The algorithm needs to compute the conditional distribution of the topics, given the observed documents:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{2.36}$$

Equation 2.36 could be solved using a sampling-based algorithm such as *Gibbs sampling* [160]. Gibbs sampling is from the family of Markov Chain Monte Carlo (MCMC) [59] technique. The MCMC provides a numerical approximation of the unknown posterior distribution

by sampling from conditional distributions of the variables of the posterior in an iterative algorithm.

In case of LDA, we are interested in β and θ. However, if we know $z_{d,n}$ (the assigned topic to the nth word in document d) both β and θ can be defined using 2.37 and 2.38.

$$\beta_{z,w} = \frac{n(z,w)+\eta}{\sum_{W} n(z,w)+\eta} \tag{2.37}$$

$$\theta_{d,z} = \frac{n(d,z)+\alpha}{\sum_{Z} n(d,z)+\alpha} \tag{2.38}$$

Where α and η are pre-defined distribution parameters and $n(.)$ denotes the count.

Mathematically, the desired posterior is shown in Equation 2.39 [39] where $\mathbf{z}_{-i}$ is the set of topic assignments of all words other than z_i.

$$p(z_i|\mathbf{z}_{-i},\alpha,\eta,w) \tag{2.39}$$

Intuitively, after expanding 2.39 and replacing the distributions, the desired posterior probability becomes proportional to (probability of word i given topic k)× (probability of topic k given document d). First part shows how much each topic likes a word and second part shows how much each topic is presented in a document.

So in simple words, the algorithm randomly initialises β and θ. Then goes through each word in each document and assigns a new topic to each word proportional to the posterior probability explained above. After enough iteration the algorithm converges and maximises the likelihood of the data. The final updated θ can represent each document in term of the topics and the final β defines each topic based on its vocabulary.

2.6.2 Bi-directional long-short term memory (Bi-LSTM) model

Multylayer perceptron

An artificial neural network (ANN) consists of multiple artificial neurons. Each neuron is a simple non-linear computation unit which applies a nonlinear function to the weighted sum of the inputs(Equation 2.40, Figure 2.10).

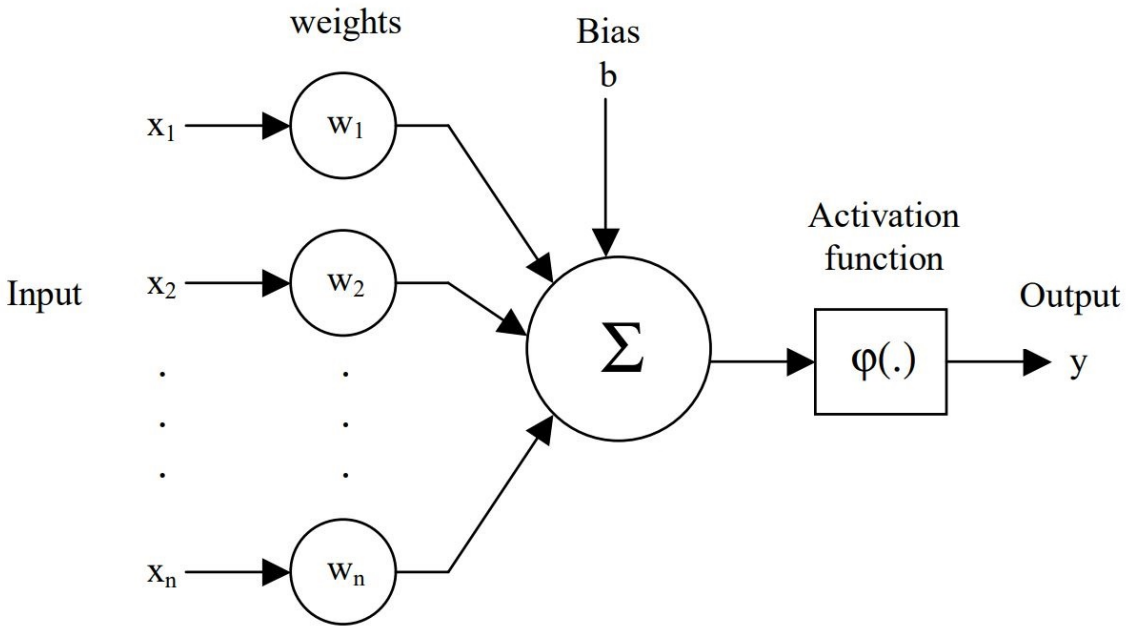

Fig. 2.10 A single artificial neuron

$$\hat{y} = \phi(\sum_{i}^{m} w_i x_i + b) \tag{2.40}$$

Arranging these neurons into a layer and adding optional number of layers between the input and the output forms an ANN commonly known as Multylayer Perceptron (MLP)(Figure 2.11).

These networks that the output signals of each layer are the inputs to the next layer and there is no cycles are referred to as feed-forward neural networks [21]. Equation 2.41 shows the mathematical representation of a MLP with one hidden layer.

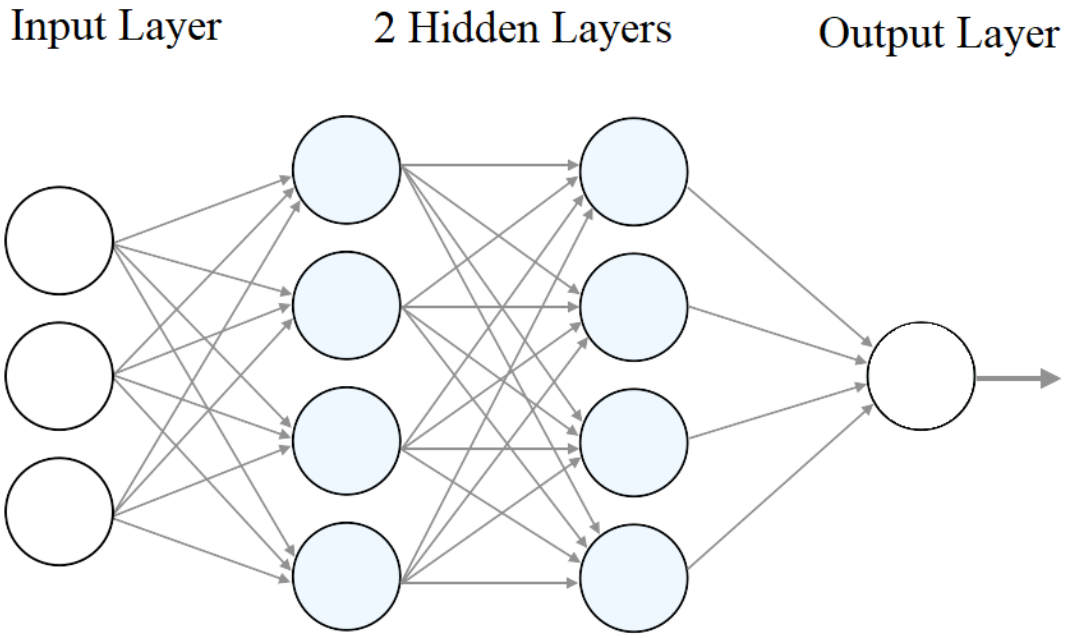

Fig. 2.11 A simple MLP with 2 hidden layers

$$\hat{y} = \phi^{[2]}(\sum_{j}^{n} w_j^{[2]} \phi^{[1]}(\sum_{i}^{m} w_{ij}^{[1]} x_i + b_j)) \tag{2.41}$$

where n is the number of neurons in hidden layer and m is the length of input x. $w_{ij}^{[1]}$ is the weight between input x_i and neuron j in the hidden layer and $w_j^{[2]}$ is the weight of the connection between the output of the jth neuron in the hidden layer and the output. $\phi^{[1]}$is the activation function for the hidden layer and $\phi^{[2]}$is the activation function for the output which could be the different than $\phi^{[1]}$.

The flow of input data through the network which yields to the calculation of the output y is called the forward pass. The learning process begins by random initiation of the w_{ij}s in all layers. Different input Xs with known output ys for each input are presented to the network and the predicted outputs $\hat{y}$s are calculated through the forward pass. An objective function is designed which represents the difference between y and $\hat{y}$ and the algorithm tries to minimise this difference by updating the the w_{ij}s. "Gradient descent" is the common

method to train the network. It calculates the derivative of the objective function with respect to each of the network weights, then adjust the weights in the direction of the negative slope. This calculation of the derivatives and updating the w_{ij} weights is formulated in a technique know as "Back propagation" algorithm and forms the backward pass of the training [149, 180].

In theory, such simple network with only one hidden layer and enough number of neurons can approximate any continuous function on a compact input domain. This is known as "universal approximation theorem" and has been proven for different non-linear activation functions [37, 74]. The short-come of feed-forward networks like MLP is that they do not have any sort of memory hence are not the best choice for dealing with sequential data such as spoken words or text. Recurrent Neural Networks (RNN)s have been proposed to deal with this problem.

Recurrent Neural Networks (RNN)

RNN can handle sequential data by making a simple change in feed-forward networks. As mentioned before, the flow of data in MLP networks in the forward pass is directly from input toward output, without any loops in between. By relaxing this assumption and allowing loops in the model, the network can keep memory of previous inputs. In an RNN, the inputs to hidden layers are from current input and the hidden layer activations one step back in time [180]. In Figure 2.12, each block has a similar structure of a MLP and the output of each network is transferred to the next block.

Back propagation through time (BPTT) is the equivalent of the back propagation for the RNNs and uses the chain rule to calculate the derivatives through different layers and through previous time steps and updates the weights [179].

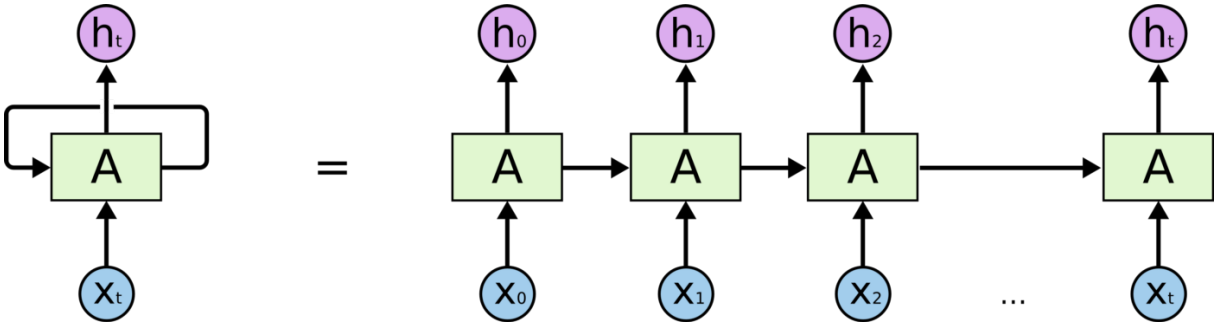

Fig. 2.12 Schematic representation of RNN

Increasing the time steps can provide more history and context, however, while training the RNN using BPTT, as the algorithm goes through the previous time step the size of gradient either explodes or vanishes. This is referred to as "vanishing error (or gradient) problem" [72]. Long-short term memory (LSTM) is type of RNNs with special units that solve this issue and allows for networks with deeper steps through time.

long-short term memory (LSTM)

An LSTM network is similar to a simple RNN, but each units in the hidden layer is a LSTM unit. Figure 2.13 shows one LSTM unit. Each unit has three gates: Input gate, forget gate and output gate. Each gate decides how much of the information can pass through the gate by scaling the gates input through multiplying it with a number between 0 and 1 so the unit can decide how much of the internal state (memory) can be transferred through the time and how much the new input x can change the state of the network. These controls solve the issue of vanishing gradient problem.

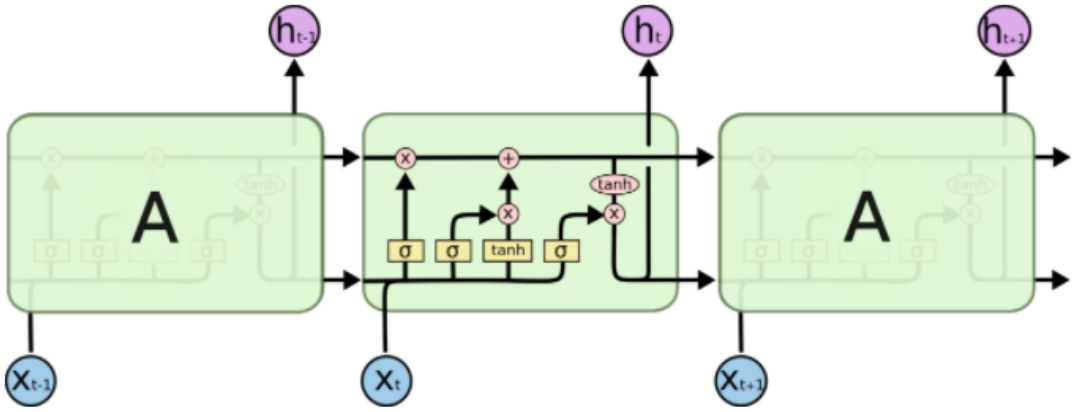

Fig. 2.13 Schematic representation of 3 LSTM units

RNNs provide a way to extract the information in sequential data and LSTM allows for deeper networks. But these networks can only use the information in the previous samples of the sequence. In many applications we have access to the full length of a sequence and future samples may contain information about the current sample. Bi-directional RNNs are models that address this limitation of normal RNNs. They train two separate networks (one for each time direction) and then merge the results [152].

Part II

Physical Activity, Hospital Costs and Chronic Conditions

CHAPTER 3

DATA

3.1 The 45 and Up Study data

In this study, I used the data from the Sax Institute's 45 and Up Study [1], a large-scale cohort study that includes more than 265,000 residents of New South Wales (NSW) Australia, aged 45 years and over, recruited between Jan 2006 and December 2009. Participants were randomly sampled from the Department of Human Services (formerly Medicare Australia) and joined the Study by completing a mailed self-administered questionnaire. People resident in non-urban areas and those aged 80 and over were over-sampled.

The overall response rate of the 45 and Up Study is 18%, accounting for approximately 11% of all individuals of age 45 years or older living in NSW. While the response rate is not high and participants tended to be of more favourable socioeconomic circumstances than average for the age group, previous work has shown that analogical findings based on internal comparisons, such as odd-ratios, are generalisable and comparable to those derived from smaller but more representative population health surveillance [114].

Centre for Health Record Linkage (CHeReL) has linked the data from the 45 and Up study to a variety of administrative data based on the participants permissions. For the purposes of this thesis I used the linked data from Admitted Patient Data Collection (APDC) and NSW Registry of Births, Marriages and Deaths in order to determine the hospital admission costs and death status of the survey respondents.

The Secure Unified Research Environment (SURE)[1] hosts these de-identified data sets and researchers can merge different linked data sets based on a unique identifier. SURE is a secure computing environment that has been purpose-built for analysis using linked health and health-related data.

The questionnaire[2] of the 45 and Up Study and full description of all the variables available in the 45 and Up study and also basic summary statistics of the data can be found on the 45 and Up Study website [3].

Data available in the baseline 45 and Up Study are rich and include age, sex, marital status, level of education, household income, smoking history, alcohol use, physical activity (Active Australia questionnaire) [17], height and weight, functional status (Medical Outcomes Study Physical Functioning scale) [7] and dietary habits among many others. There are also some self-reported chronic conditions such as (ever diagnosed) heart disease, high blood pressure, diabetes, stroke, asthma, depression and different types of cancer. Questionnaire data also include information on history of specific chronic conditions among siblings and parents of the participants.

The conduct of the 45 and Up Study was approved by the University of New South Wales Human Research Ethics Committee (HREC). Ethics approval for this study was granted by the NSW Population and Health Services Research Ethics Committee (reference: HREC/15/CIPHS/4).

3.2 SEEF data

A subset of participants of the 45 and Up Study were selected randomly and were invited through mail to participate in the Social, Economic and Environmental Factors (SEEF) study [158]. 60,404 out of the 100,000 invited people (response rate 60.4%) completed the SEEF questionnaire and provided their consent for the study.

[1] https://www.saxinstitute.org.au/our-work/sure/
[2] https://www.saxinstitute.org.au/our-work/45-up-study/questionnaires
[3] https://www.saxinstitute.org.au/our-work/45-up-study/data-book/

The SEEF data include all the original 45 and Up study data and some additional variables, aiming to provide a comprehensive view of the impact of social, economic and environmental factors on the health of Australians over age 45. The longitudinal structure of the SEEF data is used in section 5 for the estimation of the effect of sufficient physical activity on prevalence of chronic health conditions. Since individuals were recruited in the 45 and Up over a period of few years, the interval between interviews is not always the same, although it is approximately 2 and half a years on average.

3.3 Admitted Patient Data Collection (APDC)

NSW Admitted Patient Data Collection (APDC) includes records for all separations (discharges, transfers, and deaths) from all NSW public and private sector hospitals and day procedure centres as well as psychiatric and repatriation hospitals in NSW, public multipurpose services, private day procedure centres and public nursing homes. The information reported includes patient demographics, the source of referral to the service, service referred to on separation and diagnoses, procedures, and external causes of injury.

Patient separations from developmental disability institutions and private nursing homes are not included. While the APDC includes data relating to NSW residents hospitalised interstate, names and addresses are not included on these records.

Public hospital APDC data are recorded in terms of episodes of care (EOC). An episode of care ends with the patient ending a period of stay in hospital (e.g. by discharge, transfer or death) or by becoming a different "type" of patient within the same period of stay. The categories of types of care are listed under the variable "Episode of care type". For private hospitals, each APDC record represents a complete hospital stay. Private hospitals can be selected using the facility identifier code. APDC records are counted based on the date of separation (discharge) from hospital.

Although the information on Aboriginal and Torres Strait Islander peoples is available in this data set, the access is restricted to a limited to a number of studies which this thesis is not in their scope.

3.4 The NSW Registry of Births, Deaths and Marriages (RBDM)

The Registry of Births Deaths and Marriages is an agency of the Department of Justice NSW and administers the Births, Deaths and Marriages Registration Act, 1995 and the Commonwealth Marriage Act, 1961.

The role of the Registry is to register NSW life events accurately and securely for all time, ensuring their integrity and confidentiality. This includes the registration of births, deaths and marriages and official changes of name and sex. The Registrar of Births, Deaths and Marriages is the data custodian of birth registration data.

3.5 New South Wales Adult Population Health Survey

New South Wales Adult Population health surveys is a telephone survey of around 15,000 people from all over NSW, Australia and provide ongoing information on health behaviours, health status and other factors that influence the health of the adults of NSW. The the questionnaire, data collection plan, data dictionary, Weighting procedures and overview of the data is publicly available online[4].

In this thesis, I used this data to compare some statistics of the 45 and Up Study data with official NSW published data and applied the re-weighting method explained in 2.3 to change the marginal distribution of some of the variables. The selected variables and their statistics are reported in section 4.2.1.

[4]https://www.health.nsw.gov.au/surveys/adult/Pages/default.aspx

3.6 Health Roundtable Data

Health Roundtable (HRT)[5] is a non-profit membership organisation of health services across Australia and New Zealand and has a rich dataset of millions of de-identified inpatient hospital admission episodes from about 180 public hospitals in Australia and New Zealand.

I had access to 25 millions of HRT episodes between 2009 and 2014. The data includes gender and age, smoking status, obesity, and admission information such as emergency status, admission source, discharge status, diagnoses and procedures codes, assigned AR-DRG, Hospital Length of Stay (LOS) and a subset of 4000,000 episodes have a calculated cost for the admission.

[5]https://home.healthroundtable.org/

CHAPTER 4

PHYSICAL ACTIVITY AND HOSPITAL PAYMENTS FOR ACUTE ADMISSIONS

In the last decade in many countries economic growth has fallen behind the fast rising rate of health spending and in many OECD countries, around three-quarters of these spending come from public funds [131]. In Australia the total government health expenditure in 2015-16 was reported to be $114.6 billion, of which 40.9% was on public hospital services [18]. As a result, health service providers and policy makers are interested in implementing interventions that maintain people in good health for longer periods, reduce the number of hospital admissions and reduce overall payments.

Some studies about the cost effectiveness of interventions that encourage more physical activity have already been introduced in chapter 2. The different magnitudes of the effect size for PA reported in these studies implies that health-care policy planners cannot simply rely on published numbers in the literature, but instead need local studies tailored to their specific target population.

The focus of this chapter is to understand the association between physical activity in older adults and acute hospital admissions costs. This chapter of the thesis is part of a project, in collaboration with the NSW Office of Preventive Health, with a focus on the association between *sufficient* physical activity (PA) and acute hospital admission expenditures for the Australian population aged 45 and over. The goal is to estimate the size of the association, if

any, for different age groups and levels of household income, in order to help the NSW Office of Preventive Health to select target groups for interventions.

This study aims to formulate evidence-based policies to encourage higher levels of physical activity in middle age and older Australian population by understanding the long-term effects of physical activity on health and health-care utilisation. I use the baseline data from the Sax Institute's 45 and Up Study data set with more than 260,000 participants. As explained in section 3.1, this data set is linked by the Centre for Health Record Linkage (CHeReL) with the Hospital Admission Data Collection (APDC) data set and NSW Registry of Births, Deaths and Marriages (RBDM). I define a unique indicator of physical activity (PA) using the 45 and Up Study data and calculate hospitalisation payments over a one-year period for each participant using the linked APDC. I then use the matching technique called Coarsened Exact Matching (CEM) described in section 2.4 to find participants from physically active and inactive groups that are similar based on some characteristics. I develop a multivariate analyses model where the calculated hospital admission payments is the dependent variable and includes PA, chronic health conditions and standard socioeconomic variables as covariates. The results clearly indicate that there is a statistically significant association between PA and lower hospital payments. While the size of the association depends on the covariates used in the model, the conclusions are robust. I also performed a sub-group analysis and showed that the association grows significantly stronger with increasing age and with decreasing levels of household income.

Since the analysis is observational in nature, doubts always remain about the interpretation of the results as causal estimates. Therefore I performed an instrumental variable analysis to examine in greater depth the issue of causality, and whether it is possible to prove that the relationship between PA and costs is causal.

4.1 Data and Variables

4.1.1 Data sets

In this part of the study I use data from the Sax Institute's 45 and Up Study [1], which was introduced in section 3.1. The data includes 11% of the targeted NSW population (i.e. adults 45 years and older). Mealing et al. [114] show that exposure-outcome relationship patterns derived from the 45 and Up Study are comparable with those derived from the New South Wales Population Health Survey (PHS)[1]. However, higher response rate among more socioeconomic advantaged groups of the 45 and Up Study cohort has resulted in higher rate of physical activity and lower rate of smoking. Therefore, I use iterative proportional fitting (IPF) [22] to re-weight the 45 and Up data to match the distribution of key variables observed in the NSW Adult Population Health Survey, which is representative of the population. The IPF algorithm is explained in 2.3. The variables chosen for re-weighting were: physical activity, age, smoking, body mass index (BMI)[2] and income.

For the instrumental Variable analysis, I used the Long-term temperature record data from "Australian Climate Observations Reference Network – Surface Air Temperature (ACORN-SAT)[3]" to measure average temperature. The data is available online and I scraped the required data from the stations with complete data over the time period of the study.

To assign the IV to each individual, I calculated the distance between 586 available postcodes in the 45 and Up Study participants and 112 weather stations (Figure 4.1) in allover Australia based on latitude and longitude of the centre of the postcode and latitude and longitude of the stations and assigned the closest station to each postcode (Figure 4.2) then for each participant, calculated the average of minimum and maximum daily temperatures of the assigned station over the last 60 days from the date of their participation in the survey.

[1] http://www.health.nsw.gov.au/surveys/Pages/nsw-population-health-survey.aspx

[2] Body Mass Index (BMI) for adults older than 20 years is calculated by dividing weight in kilograms by height in metres squared. BMI less than 18.5 is considered Underweight, values between 18.5 and 24.9 are Normal, between 25 and 29.9 are Overweight and BMI 30 or more are Obese.

[3] www.bom.gov.au/climate/change/acorn-sat/

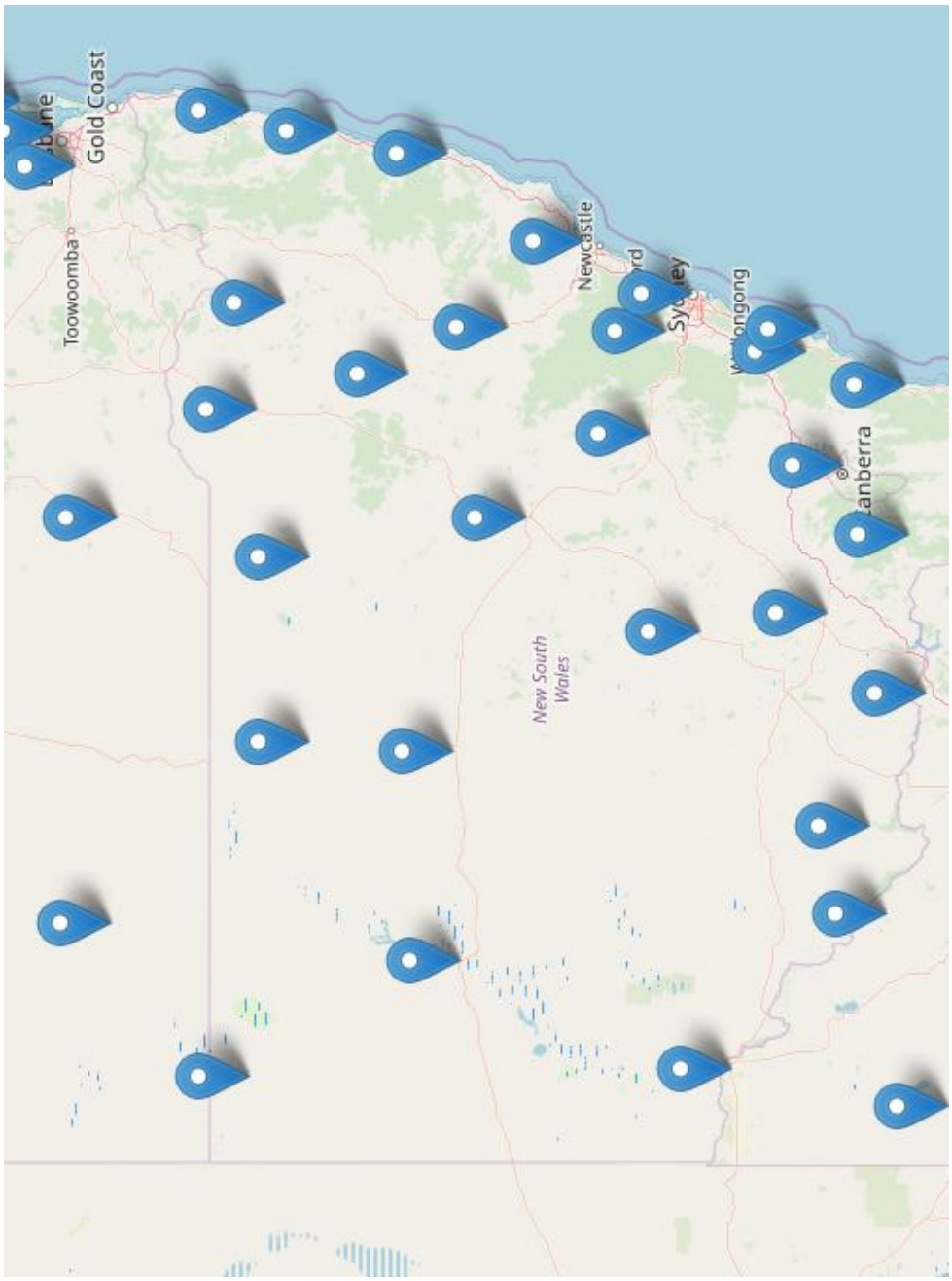

Fig. 4.1 Location of temperature stations in and around NSW.

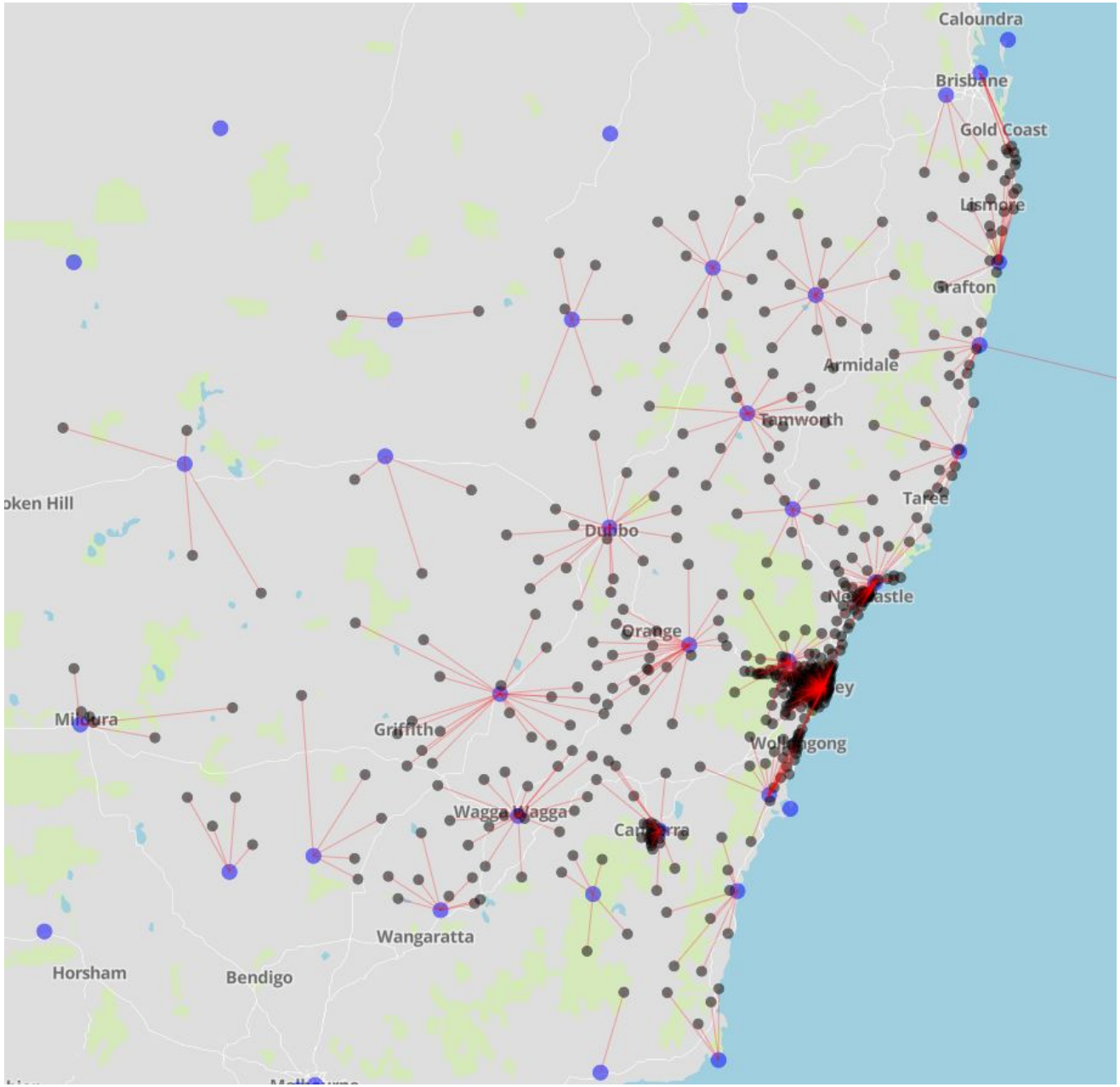

Fig. 4.2 Temperature stations and centres of postcodes in NSW. Blue points are temperature stations and black points are post codes. The red lines connect each postcode to the closest weather station.

4.1.2 Primary outcome, key predictor and covariates

I used weighted linear regression to model the association between *sufficient* physical activity, our key predictor, and hospitalisation payments, our primary outcome. Other covariates used in the model include demographic characteristics such as sex, age, income, marital status, type of private insurance (PHI) and risk factors such as history of regular smoking, body mass index (BMI), presence of four health chronic conditions (heart disease, hyper-tension, stroke, diabetes) and Physical Functioning Score. I also controlled for death in the period of study. While some of these variables are direct answers to the survey questions of the 45 and Up Study, I generated some others by combining responses. The following section describes how I create these derived variables.

Key Predictor: *sufficient* Physical Activity

I derived Physical Activity (PA) data in the 45 and Up Study from questions 16 and 17 of the questionnaire. Question 16 asks participants "How many TIMES did you do each of these activities LAST WEEK?" and question 17 asks "If you add up all the time you spent doing each activity LAST WEEK, how much time did you spend ALTOGETHER doing each type of activity?" For both questions, the questionnaire describes the activities as below:

- Walking continuously, for at least 10 minutes (for recreation or exercise or to get to or from places)
- Vigorous physical activity (that made you breathe harder or puff and pant, like jogging, cycling, aerobics, competitive tennis, but not household chores or gardening)
- Moderate physical activity (like gentle swimming, social tennis, vigorous gardening, or work around the house)

I follow the guideline of The Active Australian Survey to define a single Moderate to Vigorous Physical Activity (MVPA) variable:

> "Total time in minutes for each activity is calculated by multiplying the hours by 60 and adding the minutes. ... To avoid errors due to over-reporting, any

> times greater than 840 minutes (14 hours) for a single activity type are re-coded to 840 minutes. Missing values are not imputed. Total time in activity overall is calculated by adding the time spent in walking and moderate activity and twice the time spent in vigorous activity. The time spent in vigorous activity is doubled because vigorous activity is more intense and so confers greater health benefits than moderate activity [17]."

I defined the *sufficient* PA variable as having at least 150 minutes of MVPA and at least 5 sessions of physical activity in the past week according to the suggestion of The National Physical Activity Guidelines for Australians [155]. This guideline recommends total of at least 150 minutes of moderate activity in most days of a week.

There are some published studies that used the 45 and Up Study data and have considered the Physical Activity as one of their study's variables, but they have not all used the same definitions. Some have used the number of activities in a week [10, 12, 19, 98, 133, 157], two studies have kept walking minutes as a separate variable [11, 159], two studies only reported *sufficient* and *insufficient* PA [53, 167] and the others have used the sum of minutes spent in moderate to vigorous activity in each week and divided it into different intervals [58, 136, 140, 148, 171, 172, 189].

Table 4.1 shows the reported statistics of PA variable in these studies along with the statistic of the variable in our study and the statistics of two separate national reports [17, 142] and one report for NSW Physical Activity statistics [4] which comes from a survey on 9742 people aged over 45.

The numbers reported in Table 4.1 are derived from different subsets of the 45 and Up Study data set. Some have used all the data from the base study while the others have used the available data at the follow up and the cleaning of the PA variable would be different in each study. In general, it seems that the PA statistics in 45 and up study shows higher figures for sufficiently active people compared to the published governmental reports. I address this problem in the next section by proper re-weighting of the 45 and up physical activity variable to match the NSW data.

[4] http://www.healthstats.nsw.gov.au/Indicator/beh_phys_age/beh_phys_age_snap

	less than 150 mins/week		more than 150 mins/week	
Study	0-10 mins/week	10-150 min/week	150-300 min/week	≥300 min/week
Ding et al.	6.1	15.6	16.2	62.1
Pedisic et al.	15.9		17.3	66.8
Van der Ploeg et al. 2014	4.8	16.8	18.3	60.1
Rosenkranz et al.	3.9	15.5	15.7	64.9
George et al.	4.1	15.1	17.0	63.9
Yorston et al.	6.5	19.9	73.6	
Van der Ploeg et al. 2012	5.4	19.5	20.1	54.9
Plotnikoff et al.	6.7	18.9	18.1	56.3
This Study	7.6	14.8	15.3	62.2
ABS data (18 – 64) 2011	16	29.4	54.5	
Department of Health (18-64) 2011	60		40	
NSW (>45) 2011	67.7		32.3	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.1 Compression of PA variables for Australians older than 45 years of age in different studies

An important issue with the survey data is dealing with missing values. Since the PA variable is the sum of all minutes and hours readings against PA for all three activities, absence of any of those values will result in a missing value for the PA variable. I perform the following pre-processing four steps to reduce the number of missing values and to correct improbable values that are highly likely to be mistakes:

1. For each activity, if the number of activities in the week is zero I set to zero both the hour and the minutes variables.

2. For each activity, if the number of activities in the week is missing and both the hours and the minutes are zero, I replace the missing value with zero.

3. For each activity, if one of the hours and minutes variables has a value and the other one is missing I replace the missing value with zero.

4. For each activity, if the number of minutes is zero and the average number of hours per session of activity is greater than 10, I assume that the hour has been mistaken with the minutes, so I divide it by 60.

Table (4.2) shows the prevalence of missing values for each of the 9 physical activity variables before and after the pre-processing step.

After this pre-processing, 25.4% of items have missing values in the final *sufficient* PA variable (68,124 out of 267,897).

	walk(%)			*moderate*(%)			*vigorous*(%)		
	num.	*mins*	*hrs*	*num.*	mins	*hrs*	*num.*	mins	*hrs*
raw	6.7	34	39	10	40	36	19	43	58
preprocessed	6	8	8	8	11	11	18	18	18

Table 4.2 Prevalence of missing values for PA variables before and after pre-processing (in percent)

Primary Outcome: Hospital Payments for Acute Admissions

The financial implications of higher levels of physical activity are of interest to a variety of stakeholders, with different stakeholders interested in different financial variables. For example, policy makers, acting as representative of the people, are often interested in cost savings to society as a whole, while hospital managers may be more interested in costs to hospitals. In Australia, State Governments are responsible for managing public hospitals and for funding a large component of the care they provide. Therefore, State Health Departments have strong interest, together with private insurers, in understanding how physical activity may impact payments to hospitals, and this is the perspective I take in this study. While other perspectives are equally interesting, they would not be supported by the data at hand. For example, if one wanted to take the perspective of the hospital one would need to know how much resources are used for each admission, something which is not recorded in the administrative data set available. However, the administrative data set at our disposal has enough information to estimate the total payment to the hospital.

Therefore, I choose hospitalisation payments as the dependent variable of regression analysis. More details on how I calculated the hospitalisation payments is reported in chapter 6.

Physical Functioning Score

In the 45 and Up Study there is a survey question which asks participants whether their current health status limits them to perform some specific activities. This question is from

the RAND Medical Outcome Study, 36-Item Short Form Survey Instrument (SF-36) [22]. The activities are:

- Vigorous activity (e.g. running, strenuous sports)
- Moderate activity (e.g. pushing a vacuum cleaner, playing golf)
- Lifting or carrying shopping
- Climbing several flights of stairs
- Climbing one flight of stairs
- Walking one kilometres
- Walking half a kilometre
- Walking 100 meters
- Bending, kneeling or stooping
- Bathing or dressing yourself

Respondents can answer to these questions with three choices:

1. Yes, limited a lot (score: 0)
2. Yes, limits a little (score: 50)
3. No, not limited at all (score: 100)

The outcome variable from this question is a number between 0 and 100, which is the average score across all the items, with higher score defining a more favourable health state. This variable is highly correlated with physical activity and the effect size of PA on payments is sensitive to it. If more than 5 of the 10 items of the question are missing values, the assigned value would be missing, otherwise the available items are used to calculate the score.

Physical Functioning Score is an important variable in the model since it is highly correlated with PA, age, BMI and hospital payments. Figure 4.3 and 4.4 show the box plot of the range of the Physical Functioning Score for four different levels of PA and four BMI categories.

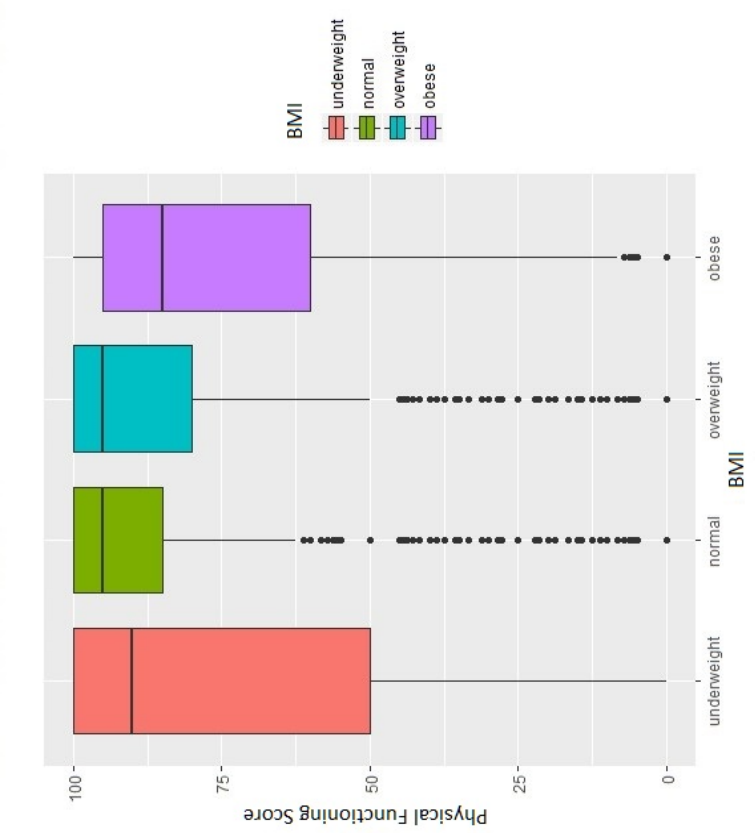

Fig. 4.4 Box plot of Physical Functioning Score for four levels of BMI

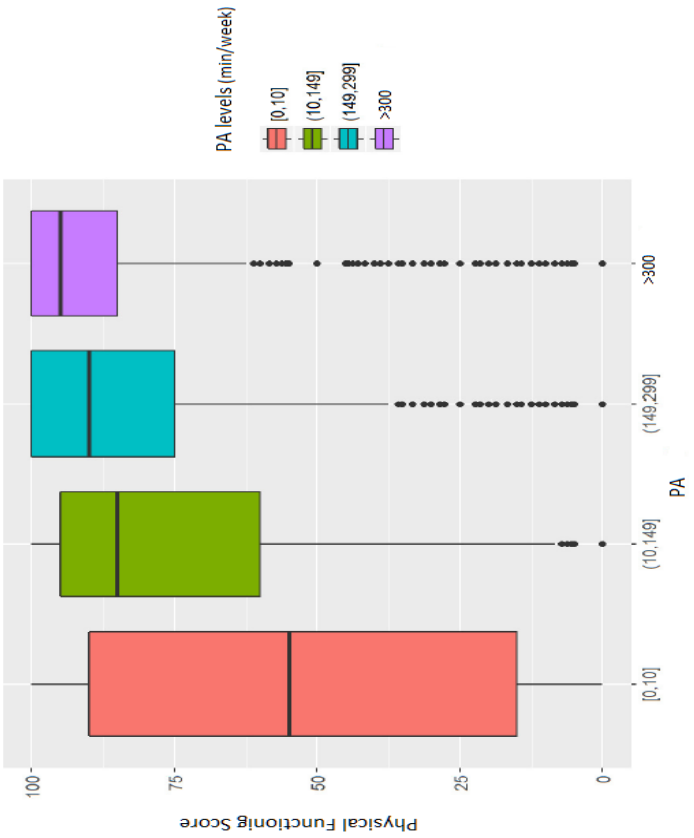

Fig. 4.3 Box plot of Physical Functioning Score for four levels of PA

The pattern in Figure 4.3 shows that the biggest difference in Physical Functioning Score is between the first category of PA (People with no PA) and the second category. Figure 4.4 implies that individuals with normal or overweight BMI have similar levels of physical functioning score and individuals who are either obese or underweight have worse physical functioning scores than those who have normal or overweight BMI.

As the figures show, lower ranges of Physical Functioning Score are associated with obese and underweight BMI groups and lower level of PA. In Figure 4.5 I break down the range of Physical Functioning Score into categories for the two populations with normal and obese BMI. The figure shows that on average categories with lower Physical Functioning Score have older population. However, for a given level of Physical Functioning Score the average age in the normal BMI group is always higher than the average age in the group with obese BMI. Since older age is highly correlated with higher payments, presence of Physical Functioning Score in the analysis may end up in a negative association between obesity and hospital payments. These complex relationships between the variables elicits the need for matching method such as the CEM, in order to balance the groups prior the analysis and make fair comparisons across groups.

Death in the next year

This variable is based on linked data from registry of death and shows whether individuals have died in the one year period after taking the survey or not. This is an important variable because expenditures in the last year of life tend to follow specific patterns which need to be controlled for.

4.2 Methods

4.2.1 Re-weighting

The 45 and Up cohort has lower rates of smoking and higher rates of physical activity when compared to the NSW population, and it represents an overall younger and healthier

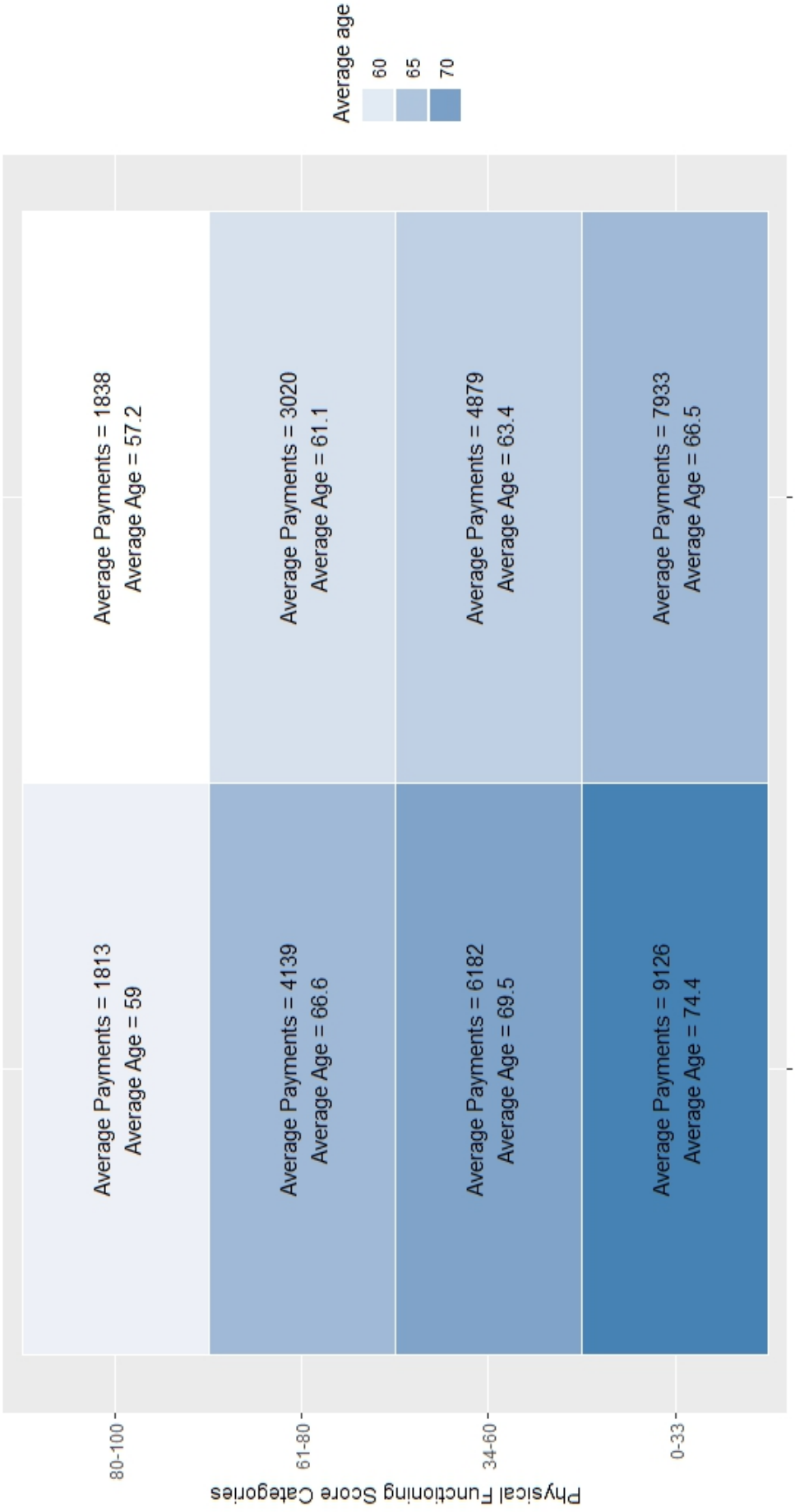

Fig. 4.5 Average age and payments for sub populations based on Physical Functioning Score and BMI.

	Before IPF	Target	After IPF
sufficient PA and Age Groups (%)			
with sufficient activity(All ages)	71.5	48.4	48.4
age group: 45-54	24.4	13.5	13.5
age group: 55-64	25.0	16.6	16.6
age group: 65-74	14.6	11.6	11.6
age group: ≥75	7.6	6.7	6.6
without sufficient activity(All ages)	28.5	51.6	51.6
age group: 45-54	8.9	10.8	10.8
age group: 55-64	8.5	14.5	14.5
age group: 65-74	5.2	12.9	12.9
age group: ≥75	5.8	13.5	13.5
BMI (%)			
normal	36.7	35.9	35.9
overweight	39.1	39.2	39.2
obese	22.8	23.2	23.2
underweight	1.3	1.7	1.7
Household income (%)			
<20K	23.8	22.0	22.0
20K-40K	22.3	23.0	23.0
≥40K	54.0	55.0	55.0
Ever smoked regularly (%)	43.6	55.4	55.4

Table 4.3 The distribution of the variables used in IPF: the 45 and Up Study (before IPF), NSW Population Health Survey (as the target), 45 and Up Study after applying IPF.

population [114]. Therefore I apply the iterative proportional fitting (IPF) [22] method to re-weight the data and make it more representative of the NSW population. The IPF algorithm, which was explained in section 2.3, assigns different weights to different individuals in the data in order to reproduce the joint or marginal distributions of some targeted variables of a reference data set, which in our case is the NSW Population Health Survey of 2008. I decided to target the joint distribution of age groups and PA and marginal distribution of smoking, body mass index (BMI), and income. Table 4.3 shows, for the variables used in the IPF, the corresponding prevalence before and after the IPF, as well as the target values from the NSW Adult Population Health Survey.

4.2.2 Matching

The CEM method, which was introduced in section 2.4, can reduce the selection bias and satisfy the "positivity" assumption by deleting unmatched samples. The "positivity" assumption requires that if a subset of the observed data only belongs to either control or treatment group it cannot be used to calculate its counterfactual.

I used the many-to-many version of the CEM algorithm and match over most of the selected covariates for the study. The variables used in the matching are: age, gender, number of chronic conditions at baseline, income, marital status, smoking history, private health insurance and the time interval between two surveys. I did not include BMI as a matching variables because we were interested in studying its role in mediating the effect of physical activity. Age and Physical Functioning Score are two continues variables of the study. I divide age into 5 year intervals and divide Physical Functioning Score into 5 categories with equal lengths. The other variables are already categorical and remained untouched, except for the 4 chronic health conditions that were replaced by a single total count variable.

4.2.3 Model selection

I used weighted linear regression model to analyse the association between PA and hospital payments. Model selection for highly skewed outcomes such as health-care expenditure has always been a much-debated topic in the literature. Violation of linear regression assumptions and heteroskedasticity [90] issues of expenditure data on one hand and re-transformation problem of logged data on the other hand has been discussed in section 2.5. Our focus here is not on individual level predictions but rather on the association between PA and payments. A previous study on the same data set, which used similar dependent variables, showed that the linear model produces the best fit compared to alternatives such as log transformed models, GLM models and two-part models [50]. Therefore, I used weighted linear regression estimated by ordinary least square (OLS) for simplicity of interpretation and in order to avoid the re-transformation issues.

4.2.4 Instrumental Variable method

I have already described in section 2.4.1 that causal estimation requires to use the observed data (factual outcome) and generalise it to situations that have not been observed (counterfactual outcome). Such generalisation requires a number of assumptions about the data in order to be valid. The main assumption is the "Ignorability assumption", which assumes that given pre-treatment covariates X, treatment assignment is independent from the potential outcomes. Since our data is collected from an observational study, the assignment of the treatment (having *sufficient* physical activity) is not randomised and the two groups with *sufficient* PA (treatment) and *insufficient* PA (control) may have systematic differences.

In section 2.4.3 I talked about the instrument variable (IV) method as a way of conducting causal estimation on observational data. In the context of this analysis an IV is a variable that affects the individual levels of PA but does not affect *directly* hospital payments, and only has an indirect effect on hospital payments through the PA variable. Finding an IV is never an easy task, which depends on the specifics of the problem and often take advantage of the environment surrounding it. Some possible options are accessibility to public transport or availability of green space, which have been proven to be associated with PA [11, 46, 127]. However, we know that these variables are associated with other socio-economic variables such as income which in turn are related to health statues and health-care costs. Another option, which has been used occasionally in the literature, is weather [6, 61, 66, 117].

In particular I hypothesised that the average temperature of the region where participants live, measured in an interval around the time the PA variable was measured, may effect people's ability to perform physical activity, introducing a random element of variation in the key independent variable. Since it is unlikely that the temperature during the week PA was measured had a direct effect on yearly hospital costs (except for intense heat waves) this variable seems to satisfy the requirements of an IV.

I used Two-stage least squares (2SLS) method as explained in section 2.4.3 for the suggested IV. In the first stage I modelled the level of PA using the IV and a number of control variables as regressors, and in the second stage I regressed the annual hospital payments

using the *predicted* PA levels and the same control variables. The coefficient of the predicted PA level is the estimated causal effect.

4.3 Results

The data set consists of 267,897 records, out of which 199,773 allow for the computation of the *sufficient* PA variable. I imputed BMI, marital status and household income using Multinomial Log-linear Models from the *nnet* R package [173] and removed items with missing value in the other relevant covariates. I also excluded 206 participants with annual hospital payments more than $100,000 in order to minimise the effect of outliers, ending up with 178,755 individuals for the analysis. Nearly 70.4 percent of males and 72.5 percent of females have reported *sufficient* physical activity, which is higher than the 47.6 percent reported statistics in the NSW Population Health Survey[5]. However after re-weighting the data using the IPF the overall *sufficient* PA is reduced to 49.7%, demonstrating the importance of the re-weighting scheme.

Table (4.4) shows the variables of the study for the groups of participants with *sufficient* and *insufficient* PA, after re-weighting for different covariates. In general, the physically active cohort are younger and in better health, and the L_1 measure of imbalance before matching is 0.363. The matching algorithm removes 11,701 (9%) participants from the active group and 6,488 (13%) participants from the group with *insufficient* PA, reducing the L_1 measure to 0.

After re-weighting, 28.4% of our selected data had at least one record of acute type hospital admission in the APDC dataset in next year. The average hospital payment in this population is about $11,111 (95% CI = 10,944 to 11,277) with median of $6,554 (95% CI = 6,439 to 6,674), while the average payment over the whole population, with and without admissions, is $3,164 (95% CI = 3,106 to 3,221).

Prior to matching, the average difference in hospital payments for people with *insufficient* PA and *sufficient* PA is $1,882 (95% CI = 1,768 to 1,996) with some of the difference due to

[5]http://www.health.nsw.gov.au/surveys/Pages/nsw-population-health-survey.aspx

the different characteristics of the two groups. After the matching, the weighted average difference reduces to $477.6 (95% CI = 393.2 to 562.0). If the matching were perfect and it had used all the possible confounders this would be our final estimate of the effect of PA on hospital payments. However matching is not always perfect and one may want to investigate the effect of including additional covariates. Therefore I applied multivariate regression models to the matched data set. In particular I was interested in understanding the effect of controlling for death of the participant, since it is an important covariate that I was not able to use effectively in the matching due to its low prevalence. I report in Table (4.5) the results of two linear regressions, with and without death covariate, and for both regressions I report the results with and without matching.

My preferred specification includes the death covariate, since it is known that hospitalisation usage is much higher in the last year of life[52, 88, 128]. The analysis shows that in this case *sufficient* physical activity on average reduces the annual hospital admissions payments by $327.7 (95% CI = 248.4 to 407.2). If death is not included, this number climbs to $426.8 (95% CI = 345.7 to 508.0). Table (4.5) also shows the coefficient for Model 1 and 2 on unmatched data. In both models, the coefficient of PA is smaller compared to the matched models.

I also performed subgroup analysis, since the results of Table (4.5) apply to the average participant. In particular I was interested in understanding how the effect of *sufficient* PA varies with key variables such as age and income. The results of the analysis for separate age groups and separate income levels are presented in figure 4.6. The figure shows that the effect of *sufficient* physical activity is statistically significant for most of the age groups and for the two lowest income groups. The effect size for the oldest group is $817.56, which is much higher than the effect for the youngest groups. The analysis based on the household income also shows that the effect size differs considerably based on income. The potential saving associated with *sufficient* PA is more than 15 times bigger for the cohort whose income is less than 20 thousand dollars per year compared to the cohort with more than 70 thousand dollars annual income.

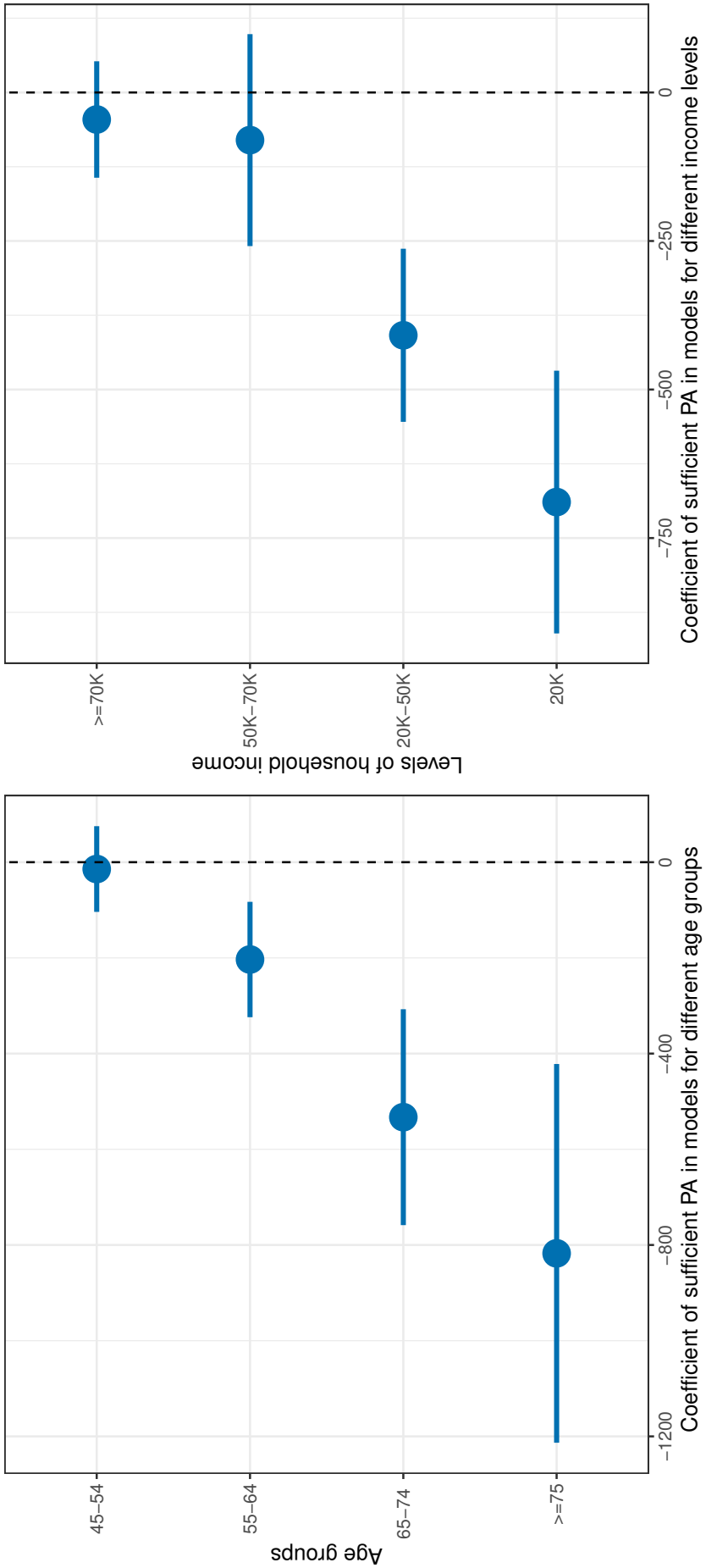

Fig. 4.6 The coefficient of *sufficient* PA in the regressions for different sub-populations

4.3.1 Instrumental Variable Analysis

For the instrumental variable analysis I used as IV the 60 day average temperature around the time that PA was measured. IV analysis is sensitive to the requirement and assumptions of the IV and in case of violation of those assumptions it may result in biased estimation [28]. Therefore, here I report results of IV analysis and result of some tests on the validation of the analysis.

The first choice I had to make is the exact definition of the IV. Analysis of the data showed that there is clearly a relationship between temperature and *sufficient* PA: as the average temperature drops (rises) the proportion of individuals with *sufficient* PA decreases (rises). The relationship is non-linear, and after some experimentation it turned that the best option, is to define the IV as a binary variables which is one when the average temperature is between 14°C and and 22.7°C and zero otherwise.

I described the 2SLS method in section 2.4.3. In the 2SLS method each stage of the analysis consists of a linear regression model. In both regression models I control for the same variables used in the previous part of the analysis. Table 4.6 presents the coefficients for both stages of the analysis. The first regression shows that the association between the suggested IV and *sufficient* PA is statistically significant and has the right sign. The coefficient of 0.03 (95% CI = 0.028 to 0.036) means that in the group with IV = 1 the rate of *sufficient* PA is higher, however the compliance rate (rate of those who have *sufficient* PA when IV = 1 and *insufficient* PA when IV = 0) is low (0.04%) which is an indicator that the IV may not be strong. The second regression shows that the calculated causal effect size is $-384.5 (95% CI = -2,467 to 1,698), with a standard error of $1,062.7. While the sign and magnitude of the effect are not only reasonable but also in line with the results of the previous section, the confidence is too large for the effect to be statistically significant. To assess the power of IV, F statistic with 1 degree of freedom for the first stage regression is often used [26]. It has been suggested that F statistic less that 10 denotes a weak instrument [161]. In this analysis the F statistic is 244.8 suggesting a strong IV. However, the size of the F statistic also depends on the sample size, and therefore I am inclined to believe that the reason for such a high value of the F statistic is the large sample size of this study [80].

4.4 Sensitivity Analysis

There are few items to be considered in the sensitivity analysis:

- In the original analysis I also included in the regression specification education level and psychological distress scale (K10). However, excluding these variables did not change the effect of PA nor affected any of the conclusions and therefore I have not included these variables in Table (4.5).

- About 50% of all APDC admission data linked with the 45 and Up is from private hospitals, which may or may not use the AR-DRG payment system. However, it is known that public and private hospitals have similar average costs [143]. Since we are interested in understanding the effect of PA on overall hospital costs, independently on whether admissions were into private or pubic hospitals, I kept both private and public admissions in the data. If I limited the analysis to public admissions the effect of *sufficient* PA would be somewhat smaller and equal to a saving of $268.2 (95% CI = 209.8 to 326.6).

- As common in this type of analysis, I have tried a variety of choices for the matching variables. For example, I excluded BMI and Physical Functioning Score from the matching and studied the effect of these variables using multivariate regression. I also tried different levels of coarseness for the age and the Physical Functioning Score variables. In all cases I have obtained estimates of the effect of *sufficient* PA on hospital payment which were consistent with the ones reported in Table (4.5). Therefore the overall assessment is that the estimate provided in this study is quite robust to the specification of both matching and regression.

- The ABF payment system takes in account the fact that individuals of Aboriginal and Torres Strait Islander origin are associated with costs that are 10% higher. In the data there was no variable allowing to identify this population and I was unable to perform this adjustment. However, given the very small size of this population this omission is unlikely to affect the results in any significant way.

- In econometrics, the Wu-Hausman test is performed to check endogeneity of the treatment [185, 186]. The null hypothesis of the test is that the IV regression is as consistent as the OLS regression without IV. Running this test using the *AER* package [94] for R software, the test fails to reject the null hypothesis (p-value = 0.91), that considering the large standard error for the IV regression is an expected result. This can suggest that there is not strong unobserved confounding in the analysis.

4.5 Discussion

In this study I have investigated the relation between *sufficient* physical activity and acute hospital payments for Australians aged 45 years and older in the state of NSW. While this relationship has been studied in the past, this is the first study that produces a quantitative estimate of this effect in Australia. The study uses data from the 45 and Up survey of the NSW population, which is not necessarily representative, but I have used the IPF [22] to re-weight the data in such a way that it becomes representative of the NSW population, and I have applied matching techniques [77] to overcome some of the limitations of observational studies and make the results less sensitive to functional form specifications.

I have found that on average, having *sufficient* physical activity reduces hospital payments of 327.8 dollars a year per person. One important finding is that the size of the effect is different for different sub-groups of the population and the potential savings in health-care payments is much higher in the oldest age group. I have also shown that the effect size of PA on hospital payments is bigger in the population with lowest household income. These results are in line with the general expectation that physical activity is associated with reductions in health-care costs and is more beneficial to the least healthy populations. The reduction in hospital costs could be due to a variety of reasons: fewer hospital admissions, shorter length of stay, or a different distribution of AR-DRGs, with smaller cost weights due to better general health status and fewer complications. The results regarding the variation of the effect size across age groups complement the findings of Khoo et al., which show that older groups

are the highest consumers of hospital resources and should be targeted for physical activity interventions [91].

At the end of 2017 the size of the NSW population was more than 7.8 million, out of which 3.17 million people (41%) were older than 45 years[6]. The reported rate of *sufficient* PA for these people in 2017 was 48.5% [7]. Applying the estimated effect size of $327.8 to the 2017 figures, one finds that the potential savings on hospital payments associated with making the entire population sufficiently active is 535 million dollars per year. Restricting this analysis to the 552,000 individuals over age 75, who had a *sufficient* PA rate of 35.5%, one obtains a potential saving of 291 million dollars per year.

The numbers reported here are conservative and are likely to underestimate the size of the effect of PA for two main reasons. The first is that I have only focused on acute care and have not considered the potential effect on other types of care such as ED admissions, medications, physicians visits, allied health, mental health and residential aged care. The second reason is that the calculated payments do not take in account all the adjustments that are usually made in the ABF payment system, which would mostly increase the costs.

I have applied a matching technique to balance the treatment and control groups in term of the known confounding variables between PA and hospital payments and make the results more robust against functional form specification. While I have experimented and controlled for a rich set of potential confounders, in every observational study there could always be some unobserved confounder, and therefore I cannot conclusively assert that reported effect sizes in the study have a causal interpretation without a proper causal estimation analysis. I offered instrumental analysis but the suggested IV did not seem to be strong enough to allow a definite conclusion. In general, though, our estimates are in line with what has been reported in the literature worldwide [4, 45, 137].

[6] http://www.healthstats.nsw.gov.au/Indicator/dem_pop_age/dem_pop_age
[7] http://www.healthstats.nsw.gov.au/Indicator/beh_phys_age/beh_phys_age_snap

Variable	Insufficient PA	Sufficient PA
n	48969	129786
Sex = M (%)	54.4	50.5
Age category (%)		
45-50	10.3	12.9
51-55	11.5	15.4
56-60	16.5	19.1
61-65	12.2	15.5
66-70	14.7	14.9
71-75	10.0	8.8
76-80	8.3	6.3
81-85	10.7	5.4
86-90	4.1	1.3
≥90	1.7	0.3
Marital status (%)		
partnered	73.7	78.5
separated	10.3	9.9
single	5.1	5.1
widowed	10.9	6.5
BMI (%)		
normal	31.4	40.3
obese	28.3	17.9
overweight	38.3	40.4
underweight	1.9	1.3
Household income (%)		
<20K	25.6	15.7
20K-50K	34.2	33.3
50K-70K	14.3	16.0
≥70K	25.9	35.1
Ever smoked regularly (%)	56.7	54.2
Hypertension (%)	42.0	34.7
Heart (%)	17.6	11.8
Stroke (%)	5.7	2.3
Diabetes (%)	13.0	7.2
Num. chronic health conditions (%)		
0	45.7	57.0
1	35.1	32.0
2	14.8	9.3
3	3.8	1.5
4	0.5	0.2
Private health insurance (%)		
none	14.9	14.6
DVA	3.3	1.9
extra	40.9	48.9
healthcare card	30.4	22.5
no extra	10.5	12.1
Physical Functioning Score (mean (sd))	69.06 (31.87)	87.18 (18.44)
Died in the next year (%)	3.2	0.7
Cost (mean (sd))	4100.35 (10234.94)	2218.39 (6826.64)

Table 4.4 Comparison of the variables of the study for two groups of people with sufficient PA and insufficient PA.

	Model 1		Model 2	
	Without Matching	After Matching	Without Matching	After Matching
(Intercept)	6113.2(122.9)***	5563.1(130.8)***	5126.1(120.7)***	4824.9(128.2)***
sex = M	697.2(42.7)***	703.4(45.4)***	585.0(41.8)***	557.64(44.1)***
age category				
51-55	104.2(79.2)	105.1(83.9)	100.1(77.5)	111.8(82.0)
56-60	236.4(75.3)**	193.1(78.9)*	244.9(73.7)***	227.5(77.1)**
61-65	528.9(81.1)***	483.4(86.3)***	545.0(79.3)***	535.0(84.4)***
66-70	1330.1(83.1)***	1330.1(87.3)***	1289.7(81.3)***	1347.3(85.3)***
71-75	1722.5(94.3)***	1699.9(99.5)***	1691.1(92.2)***	1698.8(97.3)***
76-80	2154.3(102.8)***	2572.7(108.9)***	2018.9(100.5)***	2428.3(106.5)***
81-85	2417.(103.9)***	2436.3(108.1)***	2049.8(101.7)***	2179.6(105.7)***
86-90	2644.1(149.0)***	2360.9(169.4)***	1966.0(145.8)***	1659.6(165.8)***
≥90	2352.6(223.6)***	1793.8(295.0)***	221.3(220.0)	264.7(289.0)
marital status				
separated	105.5(67.8)	141.6(69.4)*	122.9(66.3)·	148.3(67.9)*
single	−53.3(91.1)	86.0(94.5)	−11.9(89.1)	104.8(92.4)
widowed	56.3(79.3)	−7.36(82.2)	70.1(77.6)	−25.9(80.4)
household income				
20K-50K	97.0(59.2)	135.9(61.0)*	116.8(57.9)*	180.4(59.6)**
50K-70K	314.4(76.0)***	300.3(83.3)***	319.5(74.4)***	324.2(81.4)***
≥70K	220.4(73.1)**	229.7(79.5)**	217.6(71.5)**	259.3(77.7)***
stroke	543.7(104.3)***	180.3(117.7)	394.6(102.0)***	36.0(115.1)
heart	1188.8(59.7)***	1177.2(62.8)***	1104.1(58.4)***	1062.8(61.4)***
hypertension	−3.3(43.2)	60.3(44.9)	77.2(42.3)·	86.6(43.9)*
diabetes	600.2(68.4)***	521.7(71.3)***	520.7(66.9)***	371.4(69.7)***
ever smoked regularly	272.6(40.8)***	359.2(43.2)***	209.6(39.9)***	269.1(42.2)***
Private Health Insurance				
DVA	1577.5(139.7)***	1310(179.1)***	1642.2(136.6)***	1079.1(175.2)***
extra	761.8(61.9)***	733.4(66.0)***	751.1(60.5)***	710.2(64.6)***
healthcare card	453.3(69.1)***	327.6(73.8)***	440.1(67.6)***	272.3(72.1)***
no extra	360.2(79.2)***	332.6(87.3)***	395.7(77.5)***	338.2(85.3)***
BMI				
obese	−257.5(56.0)***	−7.9(57.5)	−87.1(54.8)	152.9(56.2)**
overweight	−81.2(46.6)·	76.1(50.5)	15.1(45.6)	180.5(49.4)***
underweight	233.5(160.1)	445.6(245.9)·	−203.3(156.6)	306.1(240.5)
Physical Functioning Score	−65.5(0.9)***	−60.8(0.9)***	−55.3(0.9)***	−53.3(0.9)***
died in the next year			12982.3(144.1)***	13779.3(160.7)***
sufficient PA	−325.4(42.1)***	−426.8(41.4)***	−259.8(41.2)***	−327.8(40.5)***
R^2	0.09	0.08	0.13	0.12

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 4.5 The coefficients of the least square regressions

	First stage	Second stage
	SufficientPA ~ IV + control variables	Payments ~ predicted PA + control variables
(Intercept)	0.273(0.006)***	5151.1(323.7)***
IV	0.032(0.002)***	-
sex = M	−0.039(0.002)***	419.5(55.2)***
age category		
55-64	0.041(0.002***	351.7(61.6)***
65-74	0.064(0.003)***	1253.9(87.8)***
75-79	0.009(0.004)*	1932.5(72.2)***
≥80	−0.079(0.008)***	1341.0(158.7)***
marital status		
separated	−0.004(0.003)	17.4(56.8)
single	0.008(0.004)·	26.3(75.0)
widowed	−0.005(0.004)	−0.7(73.8)
household income		
20K-50K	0.005(0.003)·	56.9(51.0)
50K-70K	−0.008(0.003)*	143.4(66.6)*
≥70K	0.004(0.003)	101.4(61.6)·
stroke	−0.048(0.006)***	347.4(116.8)**
heart	0.018(0.003)***	1071.9(59.9)***
hypertension	0.009(0.002)***	144.9(39.2)***
diabetes	−0.036(0.003)***	411.8(74.4)***
ever smoked regularly	0.010(0.002)***	214.0(36.8)***
Private Health Insurance		
DVA	0.050(0.008)***	1476.0(150.0)***
extra	0.018(0.003)***	622.9(55.8)***
healthcare card	0.036(0.003)***	261.5(71.0)***
no extra	0.017(0.004)***	349.2(69.5)***
BMI		
obese	−0.090(0.002)***	−23.19(108.5)
overweight	−0.028(0.002)***	34.1(50.2)
underweight	−0.045(0.009)***	−184.7(159.5)
Physical Functioning Score	0.005(0.000)***	−50.3(5.6)***
died in the next year	−0.114(0.009)***	13178.1(203.2)***
Predicted PA	-	−384.5(323.7)
R^2	0.10	0.11

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 4.6 The coefficients of the 2SLS regressions

CHAPTER 5

PHYSICAL ACTIVITY AND INCIDENCE OF CHRONIC HEALTH CONDITIONS

5.1 Introduction

There is agreement in the literature that physical activity may have positive effects on the reduction in the burden of noncommunicable diseases (NCDs) such as cardiovascular disease, hypertension, stroke and diabetes [101, 135, 146, 163, 177]. These effects include reduced disease incidence, increase in the quality of life and reduced health-care expenditures. Since higher levels of physical activity can be achieved through simple and low-cost life style changes, policy makers are interested in designing interventions aimed at increasing level of physical activity in different sub-groups of the population. For instance, WHO has recently published a Global Action Plan to reduce 15% of the global prevalence of physical inactivity in adults and in adolescents by 2030 [183]. Although the overall benefits of physical activity are taken for granted the effect size is likely to vary significantly across different sectors of the population and different ways of measuring physical activity.

In this chapter I investigate the effect of performing different levels of physical activity on the incidence of four chronic health conditions: hypertension, heart disease, stroke and diabetes. The aim is not only to estimate the effect size of physical activity on the incidence of these chronic conditions, but also to do this on *the same population* and using the *same*

measurements and methodology, something that has happened rarely in previous literature, that has tended to look at these conditions separately.

5.2 Data and Variables

Part of the data set, the key predictor and some of the pre-processing techniques are the same as those I have used in the previous chapter. The main outcome variables and the subsets of the data that were used are described here.

5.2.1 Data sets

I used data from the participants of the Sax Institute's 45 and Up Study survey and the Social, Economic and Environmental Factors (SEEF) study which both are introduced in section 3. I used the same re-weighting technique that was used in the previous chapter to re-weight the 45 and Up Study data to match the distribution of key variables observed in the NSW Adult Population Health Survey, which is representative of the population and removed items with missing values in the other relevant covariates.

5.2.2 Primary outcome, key predictor and covariates

I applied four separate weighted logistic regression models to model the association between physical activity, the key predictor, and incidence of heart disease, stroke, hypertension and diabetes, the primary outcomes. I estimated each of the four models on the subset of the population which does not have the targeted condition at the baseline. Other covariates used in the model include demographics such as sex and age, income and marital status, type of private insurance (PHI) and risk factors such as smoking status, body mass index (BMI), history of the condition among parents or siblings, and presence of the other three health chronic conditions at the baseline. The variables that are different from those used in the previous chapter are explained here.

Descriptive summary statistics for this data set are shown in Table 5.1. In the table, I summarise the number of observations and the distribution of the variables in the study for the four data sets associated with the four chronic conditions. Each column of the table includes the participants that have not reported the condition represented by the column name at the baseline.

Primary Outcome: Incidence of Chronic Health Conditions

I considered four categories of chronic disease: heart disease, hypertension, stroke, and diabetes. The survey questions for each of the conditions, both at the baseline and in the SEEF, asks whether a doctor ever told the participant that they have that condition. The underlying assumption, well verified in the data, is that if a chronic health condition is present at baseline it will also be present at follow up (SEEF). Therefore, for each chronic condition, I subset the data containing only the individuals who do not have that chronic condition at baseline and may develop it at follow up.

Key Predictor: *Sufficient* Physical Activity

sufficient Physical Activity (PA) is the key predictor and was already described in chapter 4.1.2.

5.3 Results

I run a separate weighted logistic regressions for each of the four chronic conditions considered, and report the results for the odds ratios (OR) in Table 5.2. For some conditions, such as heart disease and diabetes, it was useful to perform sub-group analyses and study only the set of individuals who are obese or overweight. The results for each condition are discussed in the following sections.

5.3.1 Heart disease

The data set for heart disease consists of 38,052 participants, out of which 6.4% developed heart disease at follow up. After matching, 5,244 participant were removed and the remaining were re-weighted. The first column of Table 5.2 shows that individuals who have *sufficient* PA are less likely to develop heart disease, with an odds ratio equal to 0.8 and statistically significant (95% CI = 0.73, 0.87). The ORs corresponding to the other covariates follows the expected patterns: older age and presence of other chronic conditions at baseline increase the odds of developing heart disease, as well as elevated BMI and family history of heart disease. It is interesting to notice that family history of heart disease is not only a highly significant predictor, but it also has a large effect, with an OR comparable in size to the one of other chronic conditions, and higher than the one of obesity.

For the purpose of planning and targeting interventions it may be useful to perform sub-group analyses along variables such as age, income and BMI. Most of the analyses we performed did not provide any additional insights. However, we found that the benefits of PA are mostly concentrated in the overweight and obese population. In fact, if we only consider the subgroup of individuals who are overweight or obese, the effect of *sufficient* PA is somewhat stronger (OR = 0.76 (95% CI = 0.68, 0.84)). The analysis also shows that in the normal and underweight population there is no statistically significant benefit of PA.

5.3.2 Diabetes

The incidence rate of diabetes at follow up is 2.7%. The logistic regression shows no statistically significant association between *sufficient* PA and reduced incidence of diabetes, as shown in the second column of Table 5.2. The association with other risk factors, such as age, other chronic conditions, family history and BMI follows the same pattern as heart disease. However, compared to heart disease, the association with age is much smaller and the association with obese BMI is much higher, as expected.

The lack of association between *sufficient* PA and incidence of diabetes is consistent with the finding of Nguyen et al. [130] that BMI is a more important risk factor than PA in

developing diabetes. This is confirmed by the fact that the association becomes significant if we exclude BMI from the analysis, suggesting that BMI is a mediator between PA and diabetes. I hypothesised, however, that PA could be beneficial only for the individuals who are obese or overweight. A logistic regression analysis limited to this subgroup confirms the hypothesis, showing that the OR for *sufficient* PA is 0.82 and statistically significant (95% CI = 0.72, 0.95).

5.3.3 Stroke

The incidence of stroke at follow up is relatively low and equal to 1.4%. The analysis for stroke is somewhat different from the previous ones. I did not find any potential beneficial effect of PA when comparing individuals with *sufficient* PA and *insufficient* PA. However, as shown in the third column of Table 5.2, there is a statistically significant association between the presence of *any* PA and the incidence of stroke. As expected, age is very strongly associated with stroke, even more than with heart disease, and both the presence of diabetes and heart disease at baseline are significant risk factors. Surprisingly, hypertension and BMI did not prove to be significantly associated with stroke.

5.3.4 Hypertension

The prevalence of hypertension at baseline is 34.7% and the incidence rate at follow up is 13.5%. Although the incidence rate on the matched data is slightly lower for participants with *sufficient* PA (14% compared to 15.6%), after further controlling for the additional covariates of Table 5.2, the effect of PA is not statistically significant and does not show any association between PA and hypertension. The association would become statistically significant if we relaxed the significance level, since the confidence interval barely includes one. However, the size of the OR would be quite small anyway and this would not change the interpretation of the results. Older age, BMI and diabetes at baseline all appear to be positively associated with hypertension. Subgroup analysis on a variety of subgroups failed to lead any significant insight.

5.4 Discussions

I found a positive and significant association between PA and reduced incidence of heart disease, diabetes and stroke. Interestingly, while for heart disease and diabetes the benefits may be realised if the level of PA exceeds the *sufficient* threshold, in the case of stroke it appears that any level of PA may be beneficial. The finding for stroke is consistent with evidence from [99] that even small amounts of PA can reduce stroke incidence.

In the case of diabetes the analysis shows a complex relationship between diabetes, PA and BMI. Since PA and non-normal levels of BMI are strongly correlated, and since BMI is such a strong predictor of diabetes, when both BMI and PA are included in the same regression BMI dominates and one finds no association between PA and diabetes. However, one finds a significant association when restricting the analysis to the group of individuals with elevated BMI.

In regards to hypertension there is no evidence of an association between PA and this condition. This results is robust and it is unchanged even after stratifying on BMI, age and income. This finding is not uncommon in the literature on the subject, which shows a wide variety of diverging results, from strong evidence to no evidence of an association [105].

The outcomes of this study suggest that health care experts can design different physical activity promotion programs based on their goal and the targeted population. For example, for the cohort that may be at higher risks of having stroke, a slight increase in physical activity could be beneficial. Also, if the goal is to reduce diabetes or heart disease then the highest priority target group is people who are already obese or overweight.

The findings of this analysis are in line with most of the literature, although precise comparisons are difficult to make due to differences in definitions, population profiles, size and design of the studies.

The strengths of this analysis include the prospective cohort design, the large sample data, which was re-weighted to better representative of the NSW population, and the application of the matching technique to provide more similarity between the case (without *sufficient* PA) and control (with *sufficient* PA) participants. Another strength of this analysis is the

fact this is the first epidemiological study on Australian population that reports the effect size of physical activity on four different chronic conditions over the same population with same measurements and methodology, which makes it possible to compare the effect size for different conditions.

The analysis faces several limitations. The defined PA variable is based on self reported values instead of objective measurements. Therefore, there is a possibility of misclassification of active and inactive people, although the large population number would attenuate such bias. The outcome variables are also self-reported and we cannot differentiate between different types of heart disease, stroke or diabetes. Another limitation is the short time period between the baseline and the follow up which may weaken the effect size. The last issue is that although we matched for the possible confounders, since this study is not a randomises control trial study, the possibility of unobserved confounders can not be excluded.

5.5 Conclusion

The most important finding of this chapter is that PA is significantly associated with reduced incidence of heart disease, diabetes and stroke. There appears no benefit in increased level of PA in regards to reducing the incidence of hypertension. The subgroup analyses have showed that the benefits of PA is concentrated in the population who is overweight or obese, making them a natural target to preventative interventions.

Variable	Heart	Diabetes	Stroke	Hypertension
n	38,052	39,533	41,582	28,400
Sex = M (%)	45.0	46.4	46.9	45.2
Age category (%)				
45-54	37.1	35.5	34.8	41.5
55-64	34.6	33.8	34.0	33.5
65,74	19.4	20.1	20.6	16.9
75-79	7.9	9.4	9.4	7.2
≥80	1.0	1.2	1.2	1.0
Marital status (%)				
partnered	78.5	78.5	78.4	79.4
separated	10.4	10.1	10.1	10.2
single	5.3	5.2	5.2	5.4
widowed	5.8	6.2	6.2	5.0
BMI (%)				
normal	38.7	39.5	38.2	43.8
overweight	39.7	40.3	40.0	39.2
obese	20.3	19.0	20.5	15.6
underweight	1.2	1.2	1.3	1.4
Household income (%)				
<20K	18.4	18.7	19.3	16.4
20K-50K	32.3	32.8	33.0	31.3
50K-70K	14.8	14.5	14.4	15.2
≥70K	34.5	34.0	33.3	37.2
Ever smoked regularly (%)	40.8	41.0	41.5	40.8
Chronic Health base line (%)				
Heart disease	-	9.7	10.1	7.5
Diabetes	6.2	-	6.8	4.0
Stroke	1.7	2.0	-	1.3
Hypertension	30.9	31.0	32.6	-
Chronic Health incidence at follow up (%)				
Heart disease	6.4	-	-	-
Diabetes	-	2.7	-	-
Stroke	-	-	1.4	-
Hypertension	-	-	-	13.5
Num. chronic health conditions (%)				
0	60.0	63.5	60.4	88.5
1	29.3	30.5	30.7	10.5
2	4.4	5.5	7.9	1.0
3	0.2	0.4	1.0	0.1
Private health insurance (%)				
none	14.8	14.6	14.4	15.6
DVA	1.5	1.8	1.8	1.5
extra	49.0	48.5	48.0	50.4
healthcare card	21.6	22.1	22.9	19.1
no extra	13.1	13.0	12.9	13.5
Sufficient PA (%)	76.4	76.7	76.2	77.3

Table 5.1 Variables of the study for four subsets of the data. Participants in each column are the subset of the data which do not have the condition represented by the column name at the baseline.

	Heart	Diabetes	Stroke	Hypertension
(Intercept)	0.01(0.01, 0.01)***	0.00(0.00, 0.01)***	0.00(0.00, 0.01)***	0.04(0.03, 0.05)***
sufficient PA	0.80(0.73, 0.87)***	0.98(0.86, 1.12)	-	0.93(0.87, 1.01)
any PA	-	-	0.72(0.55, 0.93)*	-
sex = M	1.52(1.38, 1.67)***	1.46(1.26, 1.68)***	1.25(1.04, 1.49)*	1.00(0.93, 1.09)
age category				
55-64	2.05(1.74, 2.42)***	1.26(1.03, 1.55)*	1.75(1.21, 2.51)**	1.45(1.31, 1.60)***
65-74	3.76(3.18, 4.45)***	1.60(1.29, 1.99)***	3.30(2.31, 4.71)***	1.80(1.60, 2.03)***
75-79	5.71(4.76, 6.84)***	1.38(1.07, 1.80)*	6.53(4.53, 9.41)***	2.15(1.86, 2.49)***
≥80	7.20(5.41, 9.57)***	1.72(1.09, 2.71)*	10.07(6.39, 15.87)***	1.46(1.06, 2.02)*
heart disease	-	1.36(1.15, 1.62)***	1.39(1.15, 1.67)***	0.99(0.87, 1.13)
diabetes	1.58(1.39, 1.80)***	-	1.37(1.09, 1.72)**	1.28(1.08, 1.52)**
stroke	1.93(1.60, 2.33)***	1.58(1.17, 2.13)**	-	1.47(1.08, 1.99)*
hypertension	1.60(1.46, 1.75)***	1.33(1.16, 1.52)***	1.00(0.85, 1.18)	-
BMI				
obese	1.31(1.16, 1.48)***	4.75(3.94, 5.72)***	1.08(0.86, 1.35)	2.32(2.08, 2.58)***
overweight	1.13(1.02, 1.25)*	1.85(1.54, 2.23)***	1.00(0.83, 1.21)	1.54(1.41, 1.69)***
underweight	0.92(0.61, 1.39)	0.59(0.21, 1.64)	1.48(0.84, 2.62)	0.95(0.68, 1.33)
family history				
heart disease	1.61(1.48, 1.76)***	-	-	-
diabetes	-	1.62(1.40, 1.87)***	-	-
stroke	-	-	1.16(0.98, 1.37)	-
hypertension	-	-	-	1.58(1.46, 1.71)***
Num. obs.	32,797	34,106	35,948	24,441

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 5.2 The odd ratios and 95% confidence intervals of the logistic regressions for four chronic health conditions.

Part III

Machine Learning for Analysis of Hospital Costs

CHAPTER 6

MODELLING HOSPITAL UTILISATION USING SURVEY DATA

Hospital admissions represent the biggest share of health care costs. Therefore it is highly important for hospital managers, policy planners and payers to gain a good understanding of what drives and predict costs, as well as of the patterns and composition of hospital admissions. From a researcher point of view, any analysis that involves the estimation of the effect of an intervention on costs or length of stay (LOS) requires to develop some statistical model of costs. The type of models and variables depend on the research questions and the data available. In this chapter I report on the cost modelling that was performed in the context of the work on the microsimulation of health policy scenarios performed in collaboration with researchers at Western Sydney University, that uses predominantly data from the 45 and Up study. Some of the simulation work also required to build models that predict hospital cost and LOS based on administrative data, and I will address this topic in chapter 7.

6.1 Measurement of cost and modelling framework

Health cost predictions can be either concurrent or prospective. In prospective models, data on year one is used to predict the health expenditure in the second year, while in

concurrent models the data of a given year is used to calculate the health expenditure in the same year. Concurrent models usually have a better predictive power than the prospective models [32]. However, for simulation purposes it is usually necessary to develop prospective models, which is the focus of this chapter.

In this thesis the cost variable represents the sum of payments for acute [1] hospital admissions for each individual in the year following participation in the 45 and Up Study survey. I only considered acute care admissions based on the acute care flag in the APDC data and I excluded the admissions related to Hemodialysis and Chemotherapy based on the assigned AR-DRG codes[2]. I also removed items with total annual payments of more than $100,000 in order to exclude outliers from the analysis. This involved 1053 people, which accounted for less than 0.4% of the data.

A key challenge in using administrative data such as the APDC collection is the definition of a valid cost variable that captures the content of the research question. The issue is complicated for two reasons:

- hospitals are usually not able to report the amount of resources used by an individual patient, which depend on a myriad of factors and are virtually impossible to track;
- the meaning of the word "cost" depends on the point of view, since it requires to specify who bears the cost;

In this work the word "cost" is used to mean the cost borne by the payer and should be seen as an expenditure for the payer. Therefore its definition depends on the Australian hospital payment system, which I summarise below.

Since 2012, the Australian Government funds public hospitals across Australia through Activity Based Funding (ABF). The idea underlying ABF is that hospitals are reimbursed based on the expected value of cost of admission. This is done by assigning an AR-DRG code to

[1] Acute care is care in which the primary clinical purpose or treatment goal is to: manage labour (obstetric), cure illness or provide definitive treatment of injury, perform surgery, relieve symptoms of illness or injury (excluding palliative care), reduce severity of an illness or injury, protect against exacerbation and/or complication of an illness and/or injury which could threaten life or normal function, perform diagnostic or therapeutic procedures. Acute care excludes care which meets the definition of mental health care.

[2] Australian Refined Diagnosis Related Groups (AR-DRGs) is an Australian admitted patient classification system which provides a clinically meaningful way of relating the number and type of patients treated in a hospital (known as hospital casemix)

each admission based on the diagnoses of the patient and the procedures that have been performed. AR-DRG codes have been designed in such a way that admissions with the same code utilise approximately the same amount of resources and have similar length of stay (LOS). Each AR-DRG is then assigned a price weight (PW), which is a relative measure of resource utilisation, and payments amounts are computed by multiplying the price weight by the National Efficient Price (NEP)[3], which represents the average cost of an acute admission. This procedure does not account for the fact that some hospitalisations can be unusually long or short and and in order to compensate the hospital appropriately some corrections based on the actual LOS are applied. Further small adjustments are also applied based on circumstances that may affect resource use and labour costs, such as time spent in the ICU and average local wages. Equation 6.1 shows the adjustment for the calculation of the price of an ABF activity:

$$\begin{aligned}\text{Price of an ABF Activity} = [A_{PPS} \times (A_{ICU} \times \text{ICU hours} + (1 + A_{Ind} + A_A) \\ \times A_{Paed} \times PW) - (A_{Acc} \times LOS)] \times NEP\end{aligned} \tag{6.1}$$

The variables of the Equation 6.1 are described in Table 6.1

symbol	description
A_A	each or any Remoteness Area Adjustment
A_{Acc}	the Private Patient Accommodation Adjustment applicable to the State of hospitalisation and length of stay
A_{ICU}	the ICU Adjustment
A_{Ind}	the Indigenous Adjustment
A_{Paed}	means the Paediatric Adjustment
A_{PPS}	the Private Patient Service Adjustment
$ICU\ hours$	the number of hours spent by a person within a Specified ICU
LOS	length of stay in hospital (in days)
NEP	National Efficient Price 2012-2013
PW	the Price Weight for an ABF Activity

Table 6.1 Adjustment variables for activity based founding.

Prior to 2012 the NSW Costs of Care Standards were a guide to estimate the costs of outputs of health services. The Standards had several applications including weighting activity in output-based funding for a range of services including acute admissions. The idea of calculation of acute care cost was similar to method explained above, using assigned

[3]The NEP is updated and published each year by the Independent Hospital Pricing Authority (IHPA) [78].

AR-DRG version 5 weights and adjusting for a range of factors. Figure 6.1 shows the flowchart of acute admitted care standard prior to introduction of ABF.

The 45 and Up Study participants were recruited before 2010, although the linked APDC data are coded with AR-DRG version 6. Since the NEP was first published in year 2012, which is very close to the data collection years, and it is consistent with the AR-DRG version 6.0 used in the data, I estimated hospital payments using the NEP of 2012 and applied the APC-CPI conversion rate [4] to obtain 2018 AUD figures. For 2012, The NEP was reported $4,808 per National Weighted Activity Unit 2012-2013.

In our current data set I was able to perform the adjustments based on LOS, but did not have all the details required to apply the additional adjustments. This implies that I might be slightly underestimating the overall payments to hospitals.

The LOS adjustment is straight forward. For each DRG group, a lower bound and an upper bound of LOS is defined. The price weight for the admissions with LOS between these two bounds are fixed. Shorter admissions and longer admission have a per diem value which adds to the base values. The lower bounds, upper bounds and per-diem values are reported for each DRG group separately. Figure 6.2 show a schematic PW over different ranges of LOS.

6.2 Cost Model Outputs

Once the cost variable has been defined it can be used as the dependant variable in an appropriate regression models. Regression models can be built for different purposes. In the context of the work on PA and prevention of chronic conditions, described in chapters 4 and 5, the rationale for developing a regression model was to obtain an unbiased estimate of the effect of PA on the hospital cost. For that purpose a linear model seemed perfectly appropriate, given its ease of interpretation.

In the more general context of microsimulation, regression models are built because one wants to predict hospital cost next year given the individual health characteristics this year.

[4] http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6401.0Dec2017?OpenDocument

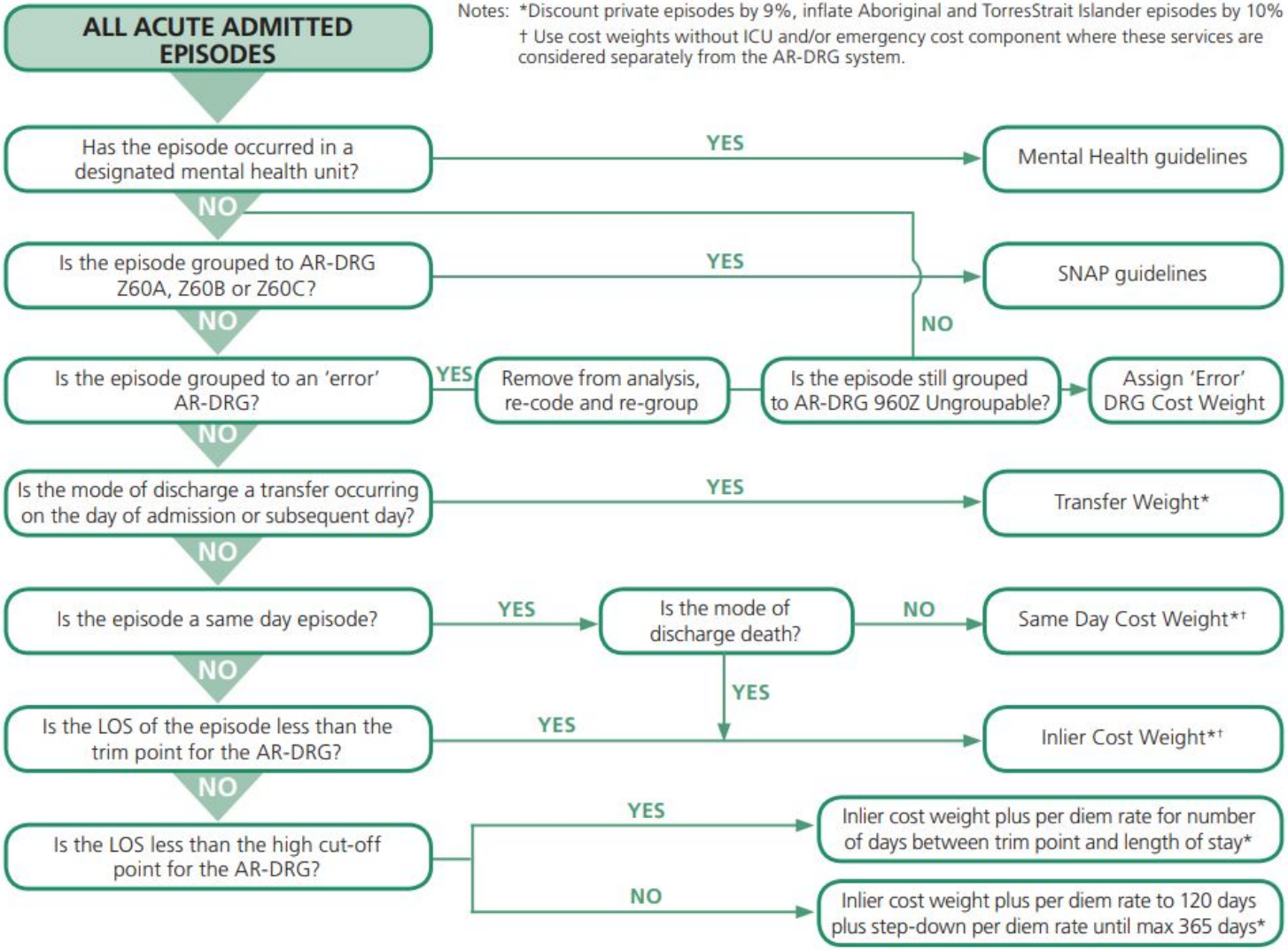

Fig. 6.1 Allocation logic for acute admitted care standards [79].

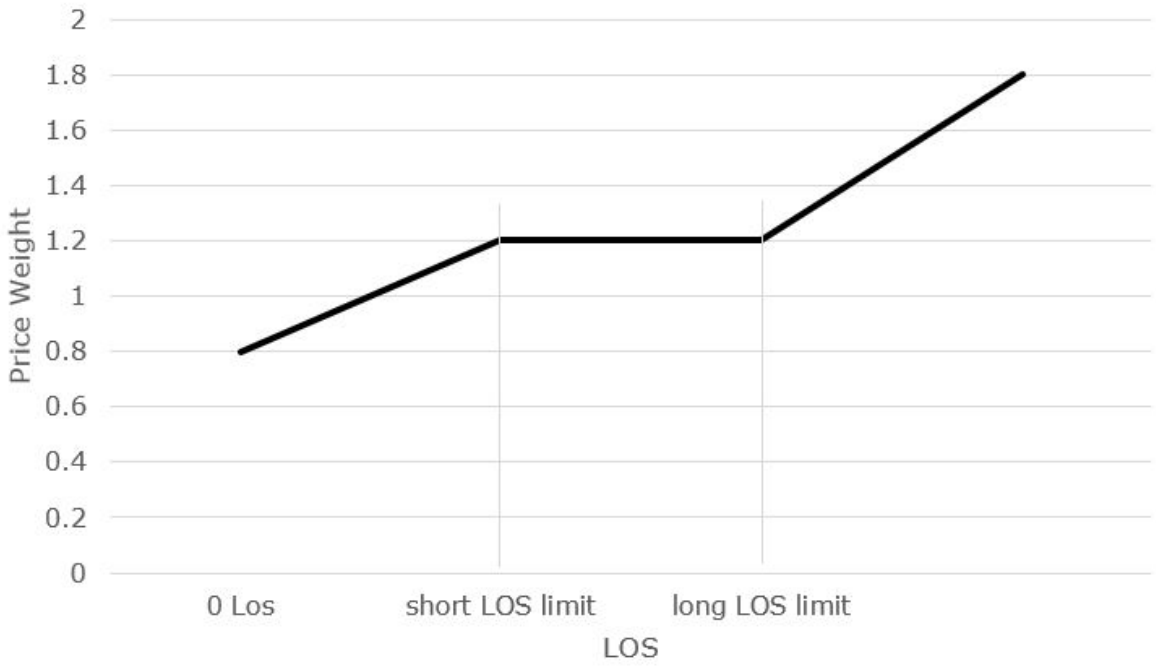

Fig. 6.2 Price weight for a sample DRG over different LOSs.

In section 2.5 I already discussed some of the general issues and modelling options for the modelling of costs. In this work I have experimented with a range of of models, including zero-inflated models, two-part models, model transformations, different members of the GLM family, neural networks, Gradient Boosting and tree methods. Somewhat surprisingly, and disappointingly, all the methods I tried failed to provide substantial, or even minimal, improvement over basic linear regression models. In this respect our conclusions is similar to what was reported in [50], where a similar analysis was performed on a comparable data set.

Regression results in the context of the effect of PA on hospital costs have been already shown in Chapter 4, on the matched data set. Here I show the results of the linear model run on the entire 45 and Up data set, for completeness. Table 6.2 shows the variables and the regression coefficient. *Sufficient* PA and Physical Functioning Score variables are described in detail in chapter 4. Other variables are directly from the 45 and Up Study questionnaire.
As table 6.2 shows, the R^2 statistics for this model is quite low. This implies that the model does not predict very well at the level of individual, and that the variables in the model do not explain most of the variance of the cost. It is possible that adding clinical level variables, such as pathology and other diagnostic results, would greatly improve the results. However

coefficient	estimate (std. error)
(Intercept)	5085.8(99.8)***
sex = M	419.2(35.4)***
age category	
55-64	336.2(41.6)***
65-74	1237.7(53.7)***
75-79	1942.8(53.7)***
≥80	1375.2(131.1)***
marital status	
separated	43.3(55.2)
single	45.1(72.6)
widowed	−11.5(71.9)
textbfBMI	
obese	−0.93(46.9)
overweight	35.53(38.9)
underweight	−199.1(148.6)
household income	
20K-50K	77.4(49.3)
50K-70K	164.9(64.3)*
≥70K	124.4(59.7)*
ever smoked regularly	209.5(34.2)***
hypertension	139.1(37.1)***
heart	1073.4(55.2)***
stroke	368.6(102.6)***
diabetes	427.9(62.0)***
Private Health Insurance	
DVA	1465.8(138.0)***
extra	613.8(50.5)***
healthcare card	260.7(57.7)***
no extra	334.7(65.1)***
Physical Functioning Score	−50.8(0.8)***
died in the next year	13219.3(159.3)***
sufficient PA	−265.9(39.2)***
R^2	0.112

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 6.2 The coefficients of the linear regression model for predicting cost

accessing clinical, survey and hospital cost data is a nearly impossible task due to the strictness of privacy legislation, and therefore the extent to which one could in theory predict hospital cost remains unknown.

Fortunately the fact that the R^2 square is so low does not necessarily mean that simulation work is bound to be inaccurate. Depending on the question that has been posed, one could still obtain reliable answers. In fact, it is often the case that one does not need the prediction to be accurate at the individual level, but requires precise estimate at a much more aggregate level. In this case it is possible that the individual level errors cancel each other out in the process of aggregation, leading to good predictions.

In order to test this hypothesis I aggregated costs at the level of local health district (LHD), something which would be quite normal to do for a state department interested in cost containment. The total hospitalisation costs for each LHD is calculated by summing the total costs of the people who live in that LHD. There are two points to notice: first, since people may use health care services in other LHDs, these numbers are not the hospital utilisation at health-care providers of LHDs. Second, each individual total cost is calculated over a period of a one year since joining the 45 and Up Study. Therefore the provided numbers do not refer to a specific year and the aim of this analysis is only to investigate the performance of this type of models at the aggregated level.

Figure 6.3 shows the calculated LHD hospital costs on the horizontal axis and the predicted LHD hospital costs on vertical axis. The size of each circle is proportional to the number of people in that LHD according to the sampled data. Ideally, we want the circles to be located on the solid black line. The accuracy of the aggregated model is measured as the coefficient of determination (R^2) of a linear fit of the data points in the plot. For the data shown in Figure 6.3 the R^2 is around 95%, showing an excellent prediction at this level of aggregation.

Another way of looking at the performance of the model is by cutting the predicted costs into buckets and comparing the sum and average of predicted and actual cost in each bucket. This approach provides a way to compare between the distribution of predicted and actual costs. Figure 6.4 shows the result of this analysis. It shows that the model underestimates in the lower and higher cost buckets but the overall aggregated performance seems acceptable.

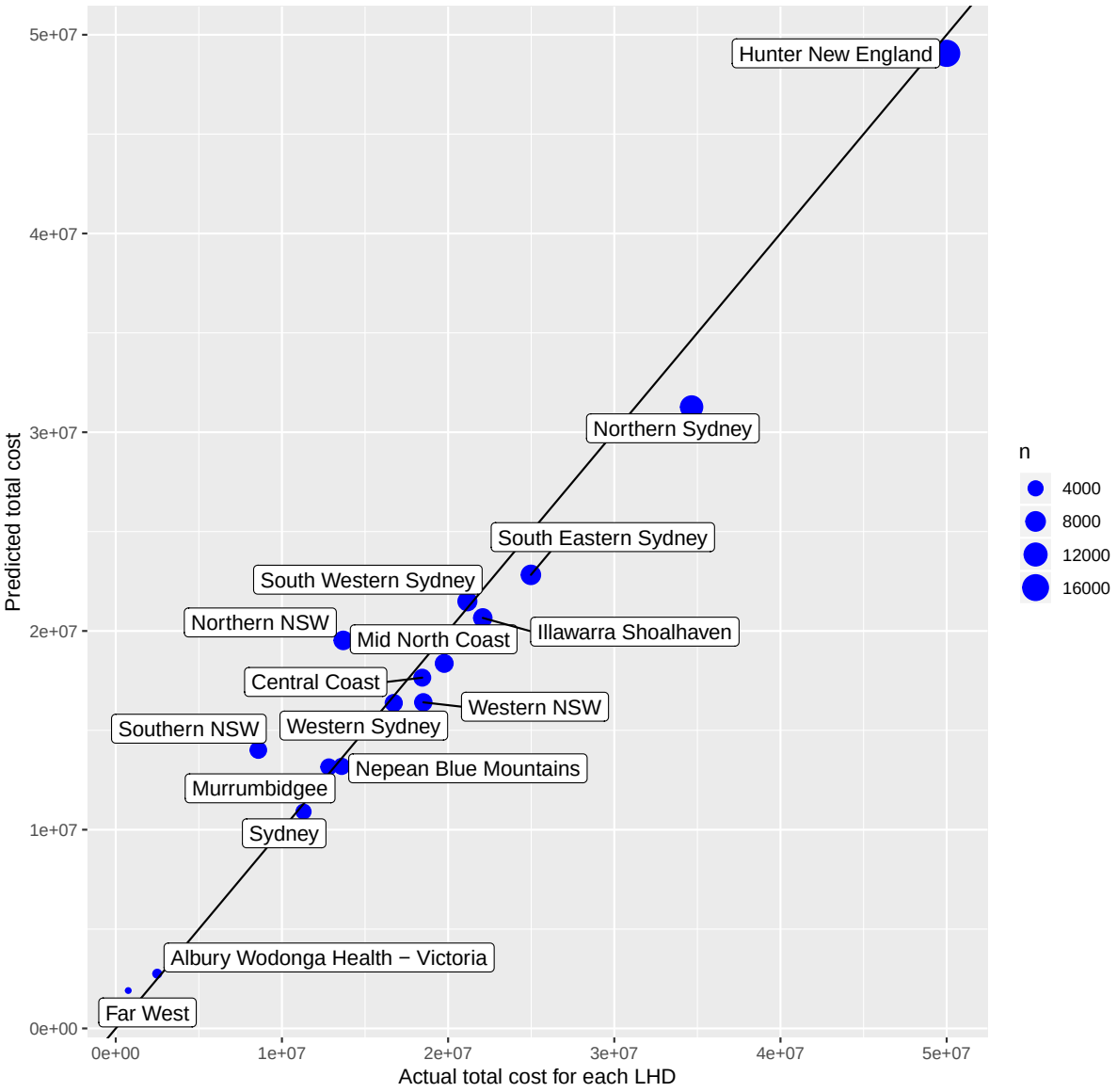

Fig. 6.3 Predicted vs. calculated costs at the level of LHD using a linear regression model.

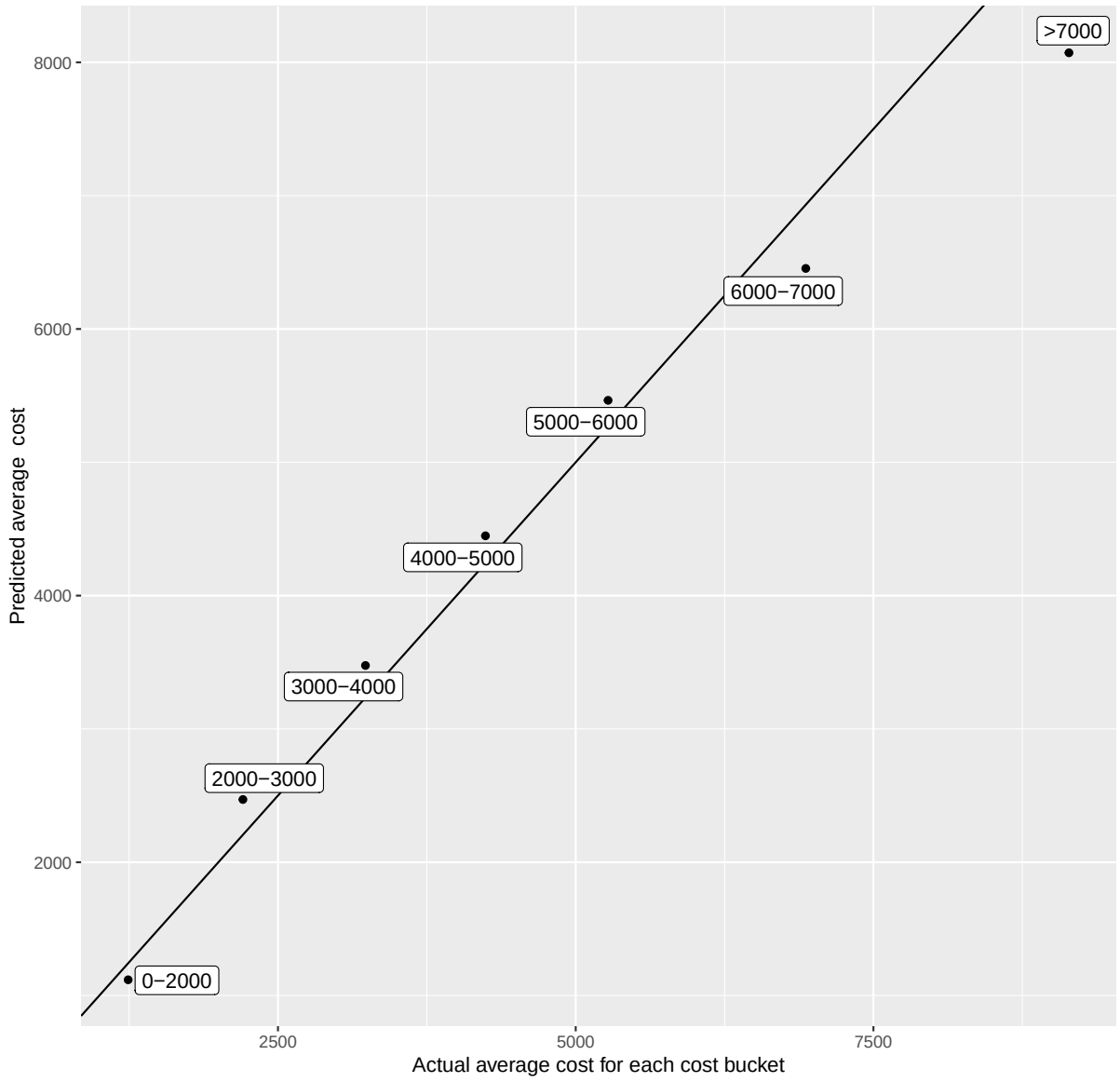

Fig. 6.4 Predicted vs. calculated costs over cost buckets

CHAPTER 7

MODELLING HOSPITAL UTILISATION USING ADMINISTRATIVE DATA AND ICD CODES

While for the work related to the effects of PA, described in chapters 4 and 5, I used individual level variables derived from self-reported survey data, it is often useful, for the purpose of simulation, to predict costs and utilisation purely on the basis of administrative hospital data. In this chapter I study the prediction of both cost and length of stay (LOS) in two data sets: the Health Roundtable (www.healthroundtable.org) data, which consist of more than 25 million de-identified inpatient hospital admission episodes from about 180 public hospitals in Australia and New Zealand and the NSW APDC data set. Both data sets were described in chapter 3.

The key variables used to predict cost and LOS in hospital administrative data are the diagnoses. Diagnosis information is usually coded using International Statistical Classification of Diseases and Related Health Problems, mostly know by the short name International Classification of Diseases (ICD). In Australia, The National Centre for Classification in Health has developed "the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification" (ICD-10-AM) which is a derived version of the World Health Organisation (WHO) ICD-10. ICD-10 uses an alphanumeric coding scheme for both diseases and external causes of injury. It is structured by body system and

aetiology, and based on the details of the code, it could consist of three, four and five character categories [15].

Each admission in administrative hospital data is usually described by several ICD codes. While these codes contain valuable information about the cost and resource utilisation of an admission, they have not been designed for such purpose. Since there is tens of thousands of ICD codes, it is not at all clear how these variable can enter a predictive models, since it is simply not possible to code them as dummy variables. Therefore some sort of grouping of these variables is in order.

A trivial way of grouping ICD codes is to classify them into ICD chapters. There are around 21 ICD chapters and each chapter is specified by a range of a letter and two digits. Such system may collect clinically related codes in a chapter, but it does not distinguish between the complexity of different episodes and therefore does not reflect the services and costs of an admission.

An alternative is offered by the Diagnosis Related Groups (DRG), that is the system designed to classify hospital admissions into smaller number of categories which each category consists of clinically similar conditions which require similar resources and services. A software known as grouper, uses ICD codes and possibly other information such as procedures, age, sex, discharge status, and the presence of complications or comorbidities to assign one single DRG to an admission. Since the software is designed to predict cost it is expected to perform well in this task. However it would not be useful if other dependent variables were considered, such as a in-hospital mortality. In addition, the grouper software is proprietary and usually not available for research purposes, posing serious challenges to researchers or developers interested in using it.

Risk adjustment models such as The Diagnostic Cost Group Hierarchical Condition Category (DCG/HCC) mentioned in section 2.5.2 are among other models that use ICD codes to summarise the health care conditions and model the health care costs of populations in the future [9]. These grouping systems are developed manually through intensive analysis of different conditions and require lots of domain knowledge. They are highly dependant on ICD scheme used and need regular manual updates.

To conclude, the issue of how ICD codes can be utilised for the purpose of modelling and predicting hospital cost and LOS is still under debate and highly important. In this chapter I propose two different, but related, methods that allow to group the ICDs in sensible ways so that they can be used in modelling exercises.

The key observation is that each ICD code is described by a short and concise text, which varies from few words to a short sentence. I suggest that it is possible to use the text descriptions of the ICD codes assigned to an episode of care to generate a synthetic clinical note. An example of how such a note may look like is shown in Table 7.1. Then I can apply Natural Language Processing (NLP) methods to extract numerical features from these documents that can then be used as a input to predictive models for cost and LOS.

I have explored this idea in two different ways:

- In the first approach I treated each admission as a short document, and applied a method known as "Topic Modelling" to group similar documents (ie. admissions) in clusters. Since the clusters contain similar documents it is then possible to label the clusters according to the main topics of those documents. So, for example, there could be a cluster of documents (admission) that all contain words related to infectious disease, and one could then label the cluster as "Infectious Disease". The grouping is probabilistic, so each document/admission has a certain probability of belonging to a cluster, and the probabilities can then be used as numerical features that can then represent the admission. The numerical features can then enter a regression model that can be used to predict costs.

- The second approach is somewhat related, but it uses sequential deep learning embedding methods to convert each document into a multi-dimensional space and uses numerical representation obtained in this way in a Deep Learning model that can be used to predicts cost or LOS.

These methodologies are described in the following two sections.

ICD code	description
I21	Acute myocardial infarction
I25.11	Atherosclerotic heart disease, of native coronary artery
T81.4	Wound infection following a procedure
I10	Essential (primary) hypertension
M19.81	Other specified arthrosis, shoulder region
text	Acute myocardial infarction, Atherosclerotic heart disease, of native coronary artery, Wound infection following a procedure, Essential (primary) hypertension, Other specified arthrosis, shoulder region

Table 7.1 An example of an episode of care with five assigned ICD codes and the generated text document by concatenating the ICD code descriptions.

7.1 Using Topic Modelling to cluster hospital admissions and predict costs

Latent Dirichlet Allocation (LDA) is a popular model to classify a corpus of unlabelled text documents. In section 2.6 I have briefly introduced the method and references for further information. I used LDA to automatically reduce the dimensionality of the space of ICD codes. Although this method is an un-supervised method and the algorithm has no prior information about clinical diseases admission costs, I believe that model can capture the semantic relationships between the words describing each episode of care and hence, develop clinically meaningful groups that could be used for predicting costs.

I sampled 1,000,000 random hospital admissions from the Health Round Table data and for each admission generated a text documents similar to the example in Table 7.1. LDA requires a pre-defined number of topics. I chose it to be 20 so I compare the predictive power of our groups with the 21 groups of ICD Chapters. After excluding ICD codes of external cause, mortality and morbidity and type of activity, and deleting the stop words and stemming the remaining text, there were 4,499 unique words in all of the documents. I used the R package *topicmodels* [73] to generate the LDA model. The output of the model includes two matrices:

The first matrix with 20 rows and 4,499 columns represents the probability of presence of each word in each topic. To explore the words in each topic and the distance of topics relative to each other, I used the R *LDAvis* package [31]. This package provides an interactive visualisation that helps to understand the topics. It maps the 20 topics into a 2d space using a Principle Component Analysis (PCA) dimension reduction (Figure 7.1, left) and for each selected topic, shows the top relevant words (Figure 7.1, right) for that topic. The relevant words are chosen relative to the frequency of the word in the topic and overall frequency of the word in all documents.

Using the tool, we can see that the relevant words in each topic are related to each other. Table 7.2 shows 5 selected topics and 5 top relevant words for each topic. Based on the top relevant words we can assume a name for each topic.

Heart	**Digestive tract cancer**	**Cellulite**	**Mental Health**	**Neoplasm**
hypertens	Haemorrhag	limb	Behaviour	Neoplasm
essenti	Intensin	skin	Mental	Malign
heart	Coliti	Cellul	Syndrom	Node
arteri	Gastroenter	staphylococcus	Alcohol	Lymph
infarct	benign	aureus	dependence	pharmacotherapi

Table 7.2 5 topics and 5 top relevant words in each topic.

The second matrix with 1,000,000 rows and 20 columns, shows the distribution of topics over each document. Each row of the second matrix could be used as feature vector with length 20 for the corresponding admission. Figure 7.2 shows a sample document and the distribution of the topics for this document. It shows that topic 10 and 19 have the most contribution in this document. By exploring the top relevant words for these two topics we can see that they are mostly about heart disease and diabetes.

While some documents such as the one in 7.2 may mostly be associated with one or two topics, others may have a mixture of several topics, as shown for example in Figure 7.3.

I used the information of the second matrix as embeddings of the ICD codes descriptions and used them in a regression model to predict costs. I control for age, gender, number of assigned diagnoses and emergency status of the admission. In order to compare the predictive

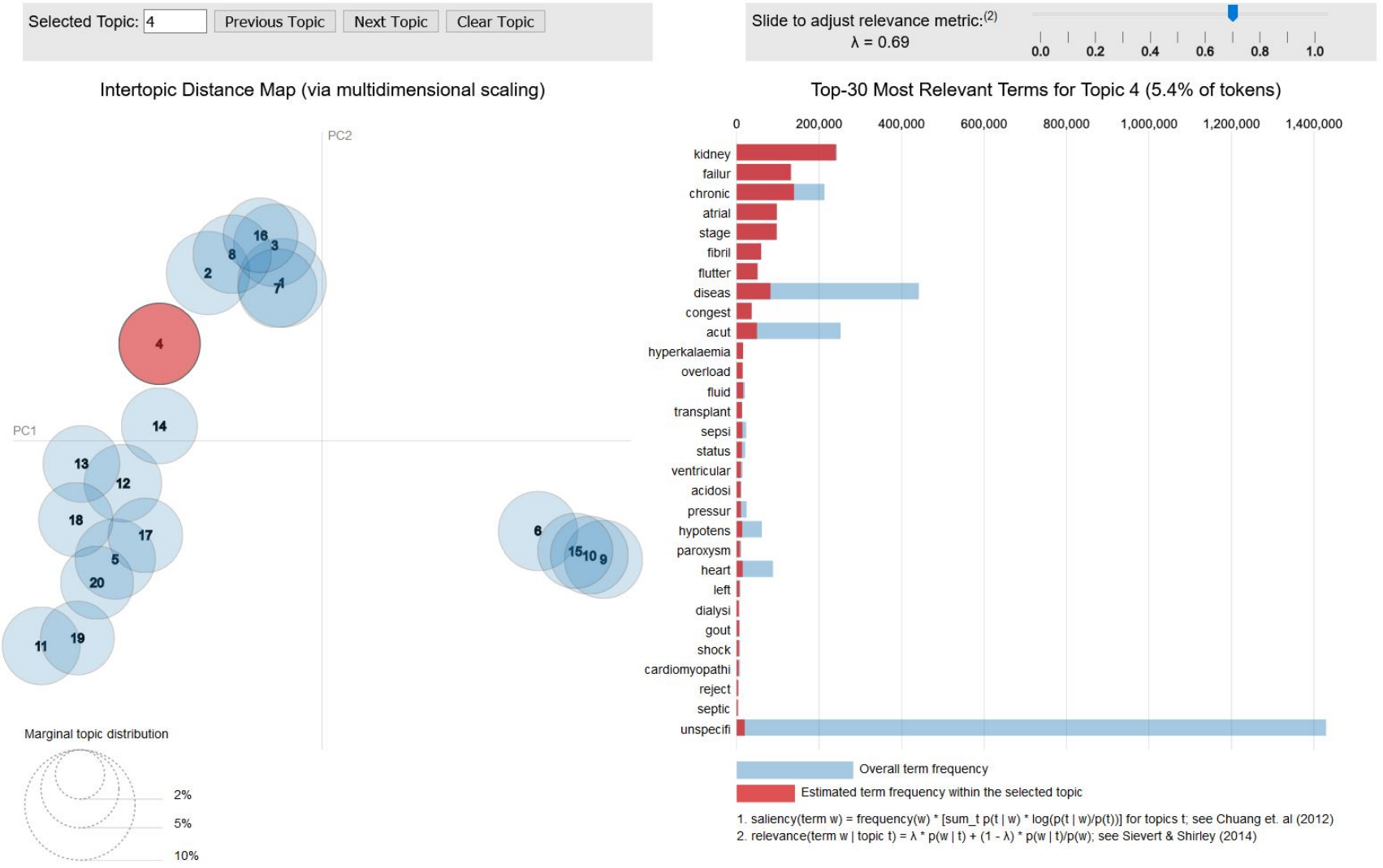

Fig. 7.1 Visualisation tool for topics: Top relevant words for topic number 4

"Type 2 diabetes mellitus with features of insulin resistance", "Essential primary hypertension", "Unstable angina", "Personal history of long-term current use of other medicaments, insulin", "Presence of aortocoronary bypass graft"

Topic 10	Topic 19
Diabetes	**Heart**
diabet	hypertens
mellitus	essenti
insulin	heart
complic	arteri
nephropathi	infarct

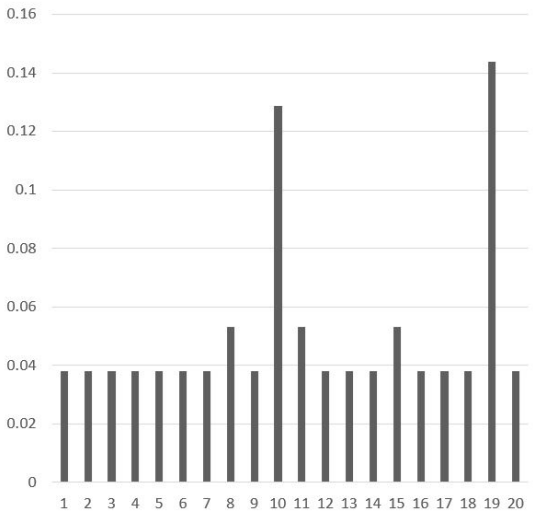

Fig. 7.2 top left: A sample document, right: The assigned distribution of topics for the sample document, bottom left: top 5 relevant words for the two dominant topics in the right.

Acute post haemorrhagic anaemia, Type 2 diabetes mellitus with established diabetic nephropathy, Type 2 diabetes mellitus with other specified kidney complication, Type 2 diabetes mellitus with peripheral angiopathy without gangrene, Type 2 diabetes mellitus with features of insulin resistance, Volume depletion, Mental and behavioural disorders due to use of tobacco, harmful use, Nerve root and plexus compressions in other dorsopathies (M45-M46+, M48.-+, M53-M54+), Angiodysplasia of colon with haemorrhage, Polyp of colon, unspecified, Melaena, Spinal stenosis, lumbosacral region, Acute kidney failure, unspecified, Chronic kidney disease, stage 3, Coma, unspecified, Syncope and collapse, Injury of stomach, Accidental puncture and laceration during a procedure, not elsewhere classified

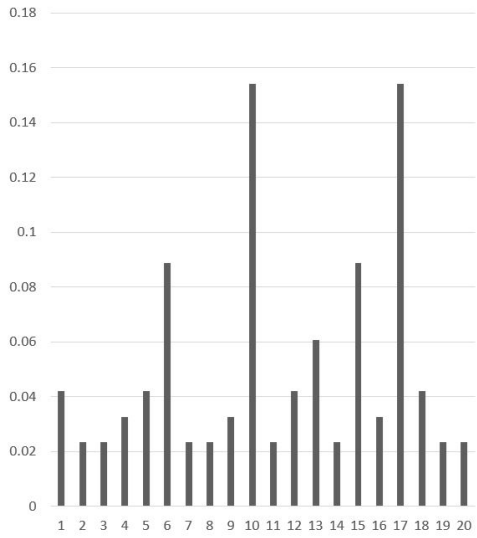

Fig. 7.3 left: A sample document with different diagnoses, right: Distribution of topics for the sample document.

power of such embedding, I developed three other models. The first is a baseline model that only uses age, gender, number of assigned diagnoses and emergency status variables. The second model uses the the same control variables together with 21 dummy variables, corresponding to the 21 ICD chapters. The third model uses the Major Diagnoses Categories (MDC) groups. MDCs are higher categories of DRG codes. Since the DRG codes are designed to illustrate the resource use of the admissions, I expected this model to perform well in term of predicting costs, although its application is limited in practice.

I evaluated the performances of the models using two metrics: the Mean Absolute Error (MAE) and the Coefficient of Determination, also known as R^2. Each regression model uses 90% of 1000,000 data as training set and 10% as validation set. I ran a 10 fold cross validation and averaged the metrics over all results. Figure 7.4 shows the result of this experiments. As expected, the model that uses the MDC categories has the best performance. Using the ICD chapters slightly improved the performance compare to the baseline and LDA features outperformed the ICD Chapter model.

In term of prediction power, the improvements in the results is not very large. However, considering some other factors may add more value to this methodology:

- This method is unsupervised. It means I do not need labelled data to train such models.
- The model is not limited to ICD codes and can be applied on other coding systems or other types of free text such as clinical notes.
- It shows that it is possible to discover the characteristics of a set of admission by analysing the text of the corresponding synthetics clinical notes and Topic modelling can tell us “what these notes are about” by extracting a number of “topic”.

An important outcome of this exercise is that it proves that the idea of using the text description of the ICD code may have merit, and this stimulated the work described in the following section.

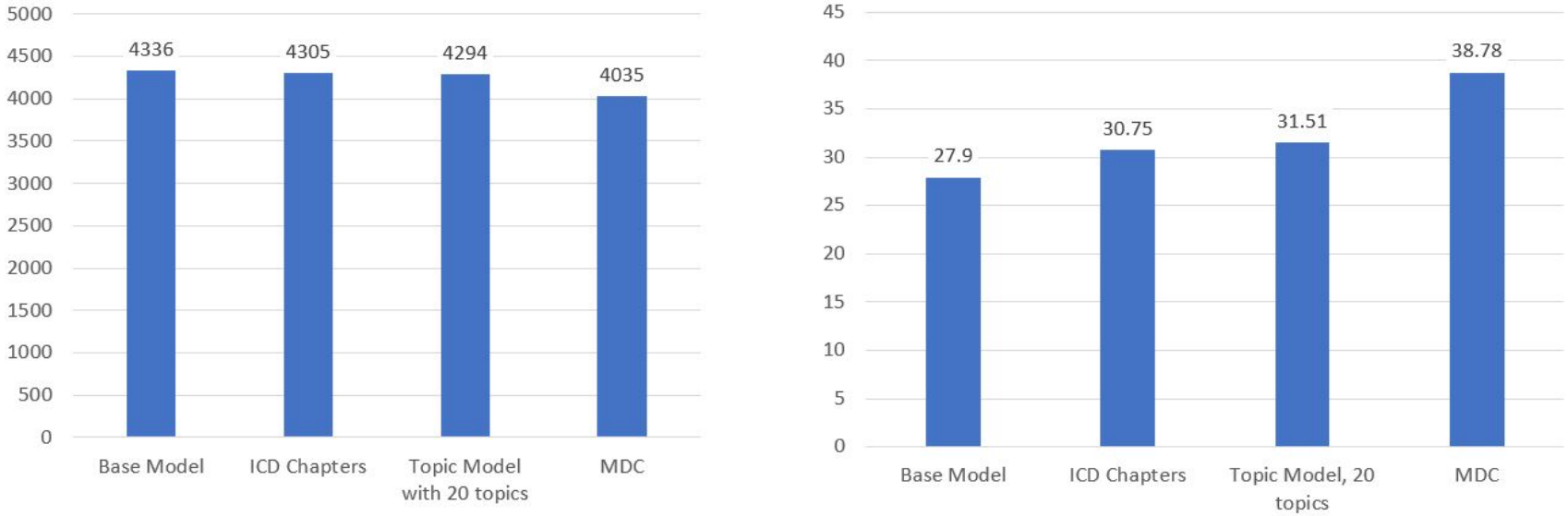

Fig. 7.4 Comparing MAE (left) and R^2 (right) for different models.

7.2 Embedding ICD code descriptions using deep learning models

At the core of Natural Language Processing (NLP) models and Text-Based Information Retrieval (IR) systems is the understanding of the semantic meaning of words in sentences. In order to use words in computer algorithms, a numerical representation is needed. Such representation is called word embedding. Traditionally, majority of statistical NLP algorithms considered words as atomic symbols. This representation is called one-hot encoding and looks like a vector of the size of the words in the dictionary with one 1 for the representing words and many zeroes. Such sparse representation neither capture the semantic meaning between two related words nor distinguishes between homographs which are the words that are spelled the same way but have different meanings such as “bat” as a baseball equipment and “bat” as an animal. To address these issues other methods have been developed to capture semantic meanings of the words and map semantically similar words to nearby points in the embedding space.

Word embedding methods can be divided into two main categories: count-based methods and predictive methods [107]. A count-based method finds the similarity of a word with any other words of the dictionary by counting the number of times that those words have appeared within a close distance of each other in a large corpus of text. The predictive methods learns the embedding vector for each word by training a model that tries to predict a word from its neighbours. These methods provide global representation of the words that can be used in different models for different tasks.

Continuous Bag of Words (CBOW), Skip Gram (SK), Word2Vec [119, 120] and Glove [138] are some of the most known word embedding in NLP. This field of machine learning is expanding so fast and new language models (Models that can predict the most probable next world given a series of words) are introduced constantly. Recent models such as ELMo [139] and BERT [41] generate word vectors from the learnt internal states of bidirectional language models that model syntax and semantics of the words across different linguistic contexts.

In this chapter, I intend to show the practical application of embedding using deep learning on real-world data. The idea, similar to the previous section, is that the words describing ICD codes which are assigned to an episode of care contain information that can be used to predict the cost or LOS of that episode.

In this work I use the APDC data collection and aim to predict LOS rather than cost, since this component is part of a larger project with these characteristics. I randomly sampled 600,000 of the episodes as training set, 60,000 samples as validation set and 120,000 samples as the test set.

Here we are interested in embedding the whole document instead of separate words. A document in this case looks like the example in Figure 7.1. Length of 99% of all documents are less than 70 words. I limited the maximum length of each document to 70 words and zero-padded the shorter documents. There were 4,538 unique words in the dictionary, out of which I only kept the most common 4,000 words.

One approach to embed the document consists of embedding each word separately and then assign the sum (or mean) of all words embedding vectors to the document. Pre-trained word embeddings can be used to embed the words. I used the pre-trained embeddings of `word2vec` algorithm on PubMed documents [123] which is available online[1]. The predictive model based on embeddings from this idea did not perform well(7.4). One probable reason is that the sum (or mean) vectors may loose the information of different embedded words. The other possible reason could be the lack of semantic information in the embeddings for the ICD descriptions. The clinical terms describing ICD codes are not commonly used in PubMed journal papers hence the embeddings learnt on those journal papers may not be a proper embedding of ICD descriptions.

Since the first approach was not successful I developed another approach that did not try to take advantage of pre-trained embeddings. In this approach the embedding is "learned" from the data within the same network that is used to predict LOS. This is achieved by using a Bi-directional LSTM model with document text as input and LOS as output. The first layer of the model embeds separate words into vectors of length 200 (initially unknown) and the last

[1]http://bio.nlplab.org/

layer before the output learns the embedding of the entire documents in vectors of length 50. Table 7.3 shows the structure of the model and dimension of each layer of the network. LSTM models have few hyper-parameters that can be tuned up to optimise performances. However, since part of this work is to demonstrate that this approach is easy to use and implement, I chose to use standard reasonable default values and not to perform any tuning.

I compared this model to a model that summarises the ICD codes using ICD chapters and to another model that uses the assigned AR-DRG codes of each admission. For each of the models I computed the corresponding R^2 and Mean Absolute Error (MAE), and reported the results in Table 7.4. The table shows that although I did not tune the hyper-parameters of the model and the structure of the network, the results from a set of reasonable default values seem to outperform the ICD chapter model and be as good as the DRG model. This is remarkable, since the DRG system is the results of years of work that was dedicated to improve the prediction of a particular type of output, while this approach is completely general and could be applied to predict any variable without modifications.

Layer	type	output shape
1-Word embedding layer	Embedding	(70,200)
2-bi-directional layer	bi-directional	(256)
3-document embedding layer	dense	(50)
4-output layer	dense	(1)

Table 7.3 Different layers of the network for embedding the documents.

Model	MAE	R^2
ICD chapters	1.71	0.36
sum of embeddings of the words	2.04	0.17
bi-lstm	1.32	0.50
AR-DRG	1.26	0.49

Table 7.4 MAE and R^2 for LOS models with different embeddings of ICDs

Since the model performs very well, the embedding must be of good quality. In order to get a better understanding of the embedding quality I plotted the embedding vectors in 2D,

using dimension reduction methods such as PCA and t-SNE [106]. These representations have shown to capture syntactic and semantic regularities in language [121]. The famous example of representation of the embedding of words "man","woman","king","queen" which captures male/female relationship and two other examples of verb-tense and country-capital are shown in Figure 7.5 [164].

I have also studied the embedding of the document as a whole. In this approach each document is embedded in a vector of length 50. I used t-SNE dimension reduction to visualise 5000 random vectors out of 600,000 training samples. Unlike the simple example of words in Figure 7.5, these documents cannot be tagged with one word and it is not easy to find similarity between them based on their text. Therefore I colour-coded the samples with LOS of each admission (Figure 7.6) and Major Diagnosis Categories (MDC) (the higher categories of the AR-DRG codes) (Figure 7.7). It can be seen in Figure 7.6 that majority of episodes with higher LOS are concentrated in the right side of the figure. Figure 7.7 also shows that some episodes with same MDC have formed separate clusters (grey and yellow dots in the middle of the plot). Grey dots belong to the MDC of "Factors influencing health status and other contacts with health services" and yellow dots belong to "Mental diseases and disorders" MDC.

Overall, the results of this approach seem to be extremely promising, especially considered that no special effort was made to optimise the deep learning network, and form the basis of future investigations.

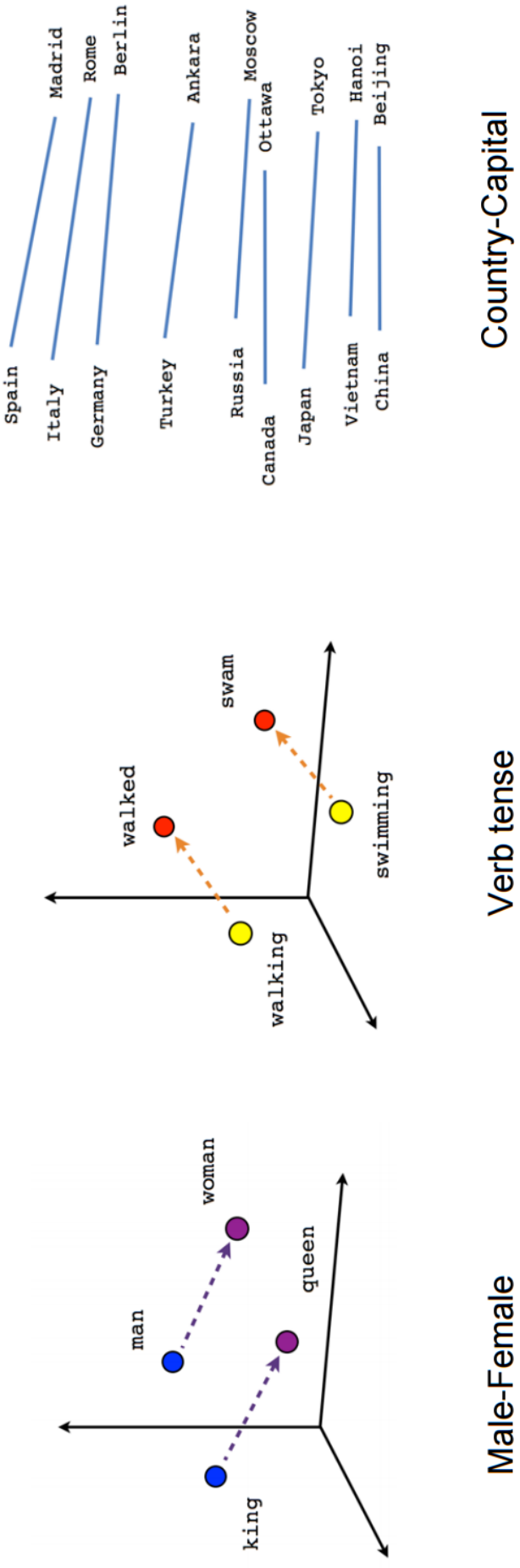

Fig. 7.5 Semantic relationships captured by word embeddings [164].

Fig. 7.6 LOS relationships captured by word embeddings [164].

Fig. 7.7 Major Diagnostic Categories (MDC) captured by word embeddings [164].

Part IV

Summary and Limitations

CHAPTER 8

SUMMARY AND LIMITATIONS

8.1 Summary

This thesis contributed to current scientific knowledge in two different areas. In the first part of the study I investigated what might be the consequences of increased levels of physical activity on hospital costs and on the incidence of chronic disease. In the second part I developed modelling methodologies for the analysis and prediction of hospital costs.

The first important results, discussed in chapter 4, is that there is a significant association between *sufficient* physical activity and acute hospital payments for Australians aged 45 years and older in the state of NSW. State departments and payers have been long interested in an estimate for this association and results exist for other countries and setting, but as far as I know this is the first study that produces a quantitative estimate of this effect in Australia.

I have found that on average, having *sufficient* physical activity reduces hospital payments of 328 dollars a year per person. The savings, however, are not distributed uniformly across the population, and tend to be much higher in the oldest age group and in the population with lowest household income, confirming the intuition that PA is likely to be more beneficial to the least healthy populations. Applying the estimated effect size of $328 to 2017 population figures, I found that the potential savings on hospital payments associated with making the entire population sufficiently active is 535 million dollars per year. If one restricted the

analysis to the much smaller population of individuals over age 75 the potential savings would still amount to 291 million dollars per year.

As noted in chapter 4 these estimates are conservative and are likely to underestimate the size of the effect of PA. While the study is observational in nature and one cannot conclusively assert that reported effect sizes in the study have a causal interpretation, I believe I have applied the best methods to minimise confounding, and it is comforting to notice that the estimates are in line with what has been reported in the literature worldwide [4, 45, 137]. These results are useful to state departments considering whether and how much to invest in programs aiming to increase the level of sufficient PA in the population, since they give them evidence that there are likely hospital cost savings associated with this goal.

However cost is the only factor considered by state departments, and even if there were no cost savings it would be still worth investing in increasing the PA levels in the population if that led to better population health. Therefore I have studied the question of whether sufficient PA is associated with decreased incidence of expensive chronic conditions such as heart disease, stroke, hypertension and diabetes.

For heart disease and diabetes the analysis strongly suggest that sufficient PA plays a significant role in reducing the probability of developing the condition in the overweight and obese population. In particular, for diabetes it seems that in the normal-weight population it is much more important to control BMI rather than PA. Interestingly, for stroke the findings were different, and the analysis suggests that what is associated with the incidence of stroke is not so much the presence of sufficient PA, but rather the presence of *any PA*, even if small, confirming some previous finding for the US population [99]. For hypertension I did not find any evidence that PA plays a role in reducing its incidence.

These findings have clearly implications on the design of how an intervention program should look like and suggests that the program would look different depending on the chronic condition whose incidence it is aiming to reduce.

These result clearly only give a short-term and limited view of what the results of an intervention may look like. In order to provide policy and planning advise to stakeholders more sophisticated tools are needed. For this reason I joined a bigger project at Western

Sydney University that aimed to simulate the effects of a range of intervention over longer time horizons, taking in account both individual level characteristics and competing risks. A simulation of this type has many moving part, and the resulted presented in the first part of this thesis will be a component of it. In the simulation project I have focused on the predictive aspects of the hospital cost modelling. When survey and self-reported health data are used to predict next-year hospital cost the problem is relatively well-posed, since the number of variables is limited and one can use a combination of variable selection methods and prior clinical knowledge to understand which variable should enter the regressions.

The main difficulty appears when using administrative data, where the health information is contained mostly in ICD coded data. Each hospitalisation in the data set is coded by few ICD codes, but there are many thousands of ICD codes. Hence treating ICD codes as dummy variables is simply not possible and a better way to encode the hospitalisation ICD codes is needed. I took advantage of the fact that ICD codes have a relatively short text description. By concatenating the text of the ICD codes I obtained a short document representing the hospitalisation. This allowed me to use state-of-the-art methods to embed documents into finite dimensional feature spaces of relatively low (and customisable) dimensionality. This is a powerful and new idea that has not been exploited yet in the literature, and I have demonstrated its usefulness in two ways.

As discussed in section 7.1, once one has a document for each hospitalisation it is natural to apply topic modelling techniques to group the hospitalisation in meaningful clusters. This process is interesting in itself, since it allows a user to browse the hospital data and understand what "types" of hospitalisations are most common just by inspecting the group of words that are used to describe them. One can also use the label of the cluster to which a hospitalisation is most likely to belong as a feature in a predictive model. While this methodology did demonstrate potential, much better performances were obtained with a more direct approach.

In order to take the most advantage of the data I trained a Long short-term memory (LSTM) Recurrent Neural Network (RNN) to perform at the same time the task of embedding the documents representing the hospitalisations into a feature vector and the task of predicting the corresponding LOS. In order to make a fair comparison I did not spend any time

optimising the architectural parameters of the RNN and have applied off-the-shelf software. The performance of the predictive algorithm was excellent, and it was at least as good as the performance of the prediction obtained using the AR-DRG codes. This is remarkable, since AR-DRG codes have been developed literally over more than three decades of continuous refinements and have involved a huge amount of manual work and expert opinion. In addition, AR-DRG have been developed exclusively for the purpose of predicting cost and LOS. Instead, the current method "learns" a feature vector representation of the documents at the same time it "learns" to predict the outcome, and therefore has two major advantages: 1) it is perfectly customisable to the prediction of any outcome of interest, and 2) it can be applied to any coding method, not just ICDs. In fact, the application of this method to administrative data coded using MBS codes is the subject of a current study that takes advantage of the publicly available 10% sample of Medicare MBS and Pharmaceutical Benefits Scheme.

8.2 Limitation of the study

Potential study limitations come from the use of the 45 and Up Study data, and apply to the first part of the thesis. The fact that people younger than 45 are missing from the study seems unlikely to be a major limitation, since the incidence of chronic conditions and hospital expenditures under that age are small. More important is the concern that the population of respondents of the 45 and Up Study may not be representative of the NSW population. It is indeed the case that the respondent of the 45 and Up tend to be of somewhat higher socio-economics status compared to the average population, and also likely to exhibit higher level of PA. However, a comprehensive study of [114], as well as some theoretical literature, has shown that just because the data set is not fully representative this does not imply that odd ratios or regression coefficients are not valid. In addition, I have made extensive use of the IPF re-weighting method to ensure that the re-weighted data represent the NSW population along many dimensions of interest. Clearly the study is sensitive to the fact that key variables, such

as the presence of chronic health conditions, are self-reported. This is also true for the PA variable, whose validity has been the subject of extensive studies in the literature [68, 102, 129].

The study is obviously sensitive to key limitations and concerns of observational studies: 1) ensuring that the observables were properly accounted for and that the results are not sensitive to functional specifications, and 2) the possibility that unobserved variables confound the results. The use of modern matching methods such as the CEM goes a long way toward making the joint distribution of the observable in the groups with and without sufficient PA. I have performed a large number of tests to ensure that the obtained results were not due to the choice of a particular specification and that the joint distribution of covariates was identical in the intervention and control group. The closeness of these joint distribution in the two groups also helps to mitigate concerns regarding the possible unobservables: to the extent that unobservables are correlated to the observables their distribution would be similar in the two groups, reducing the potential confounding effect. In addition, the component of the study that utilises instrumental variables derived from weather patterns, while not conclusive, is not inconsistent with the results.

REFERENCES

[1] 45 and Up Study Collaborators, Banks, E., Redman, S., Jorm, L., Armstrong, B., Bauman, A., and Beard, J. (2008). Cohort profile: The 45 and up study. *International Journal of Epidemiology*, 37(5):941–947.

[2] Amarasinghe, A. K. (2010). Cost-effectiveness implications of gp intervention to promote physical activity: evidence from perth, australia. *Cost Effectiveness and Resource Allocation*, 8(1):10.

[3] Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, 24(1–2):3–61.

[4] Andreyeva, T. and Sturm, R. (2006). Physical activity and changes in health care costs in late middle age. *Journal of Physical Activity and Health*, 3(s1):S6–S19.

[5] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

[6] Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4):69–85.

[7] Anita Stewart, J. (1992). *Measuring Functioning and Well-Being*. Duke University Press, Durham (NC).

[8] Ash, A. S., Ellis, R. P., Pope, G. C., Ayanian, J. Z., Bates, D. W., Burstin, H., Iezzoni, L. I., MacKay, E., and Yu, W. (2000a). Using diagnoses to describe populations and predict costs. *Health care financing review*, 21(3):7.

[9] Ash, A. S., Ellis, R. P., Pope, G. C., Ayanian, J. Z., Bates, D. W., Burstin, H., Iezzoni, L. I., MacKay, E., and Yu, W. (2000b). Using diagnoses to describe populations and predict costs. *Health care financing review*, 21(3):7.

[10] Astell-Burt, T., Feng, X., and Kolt, G. S. (2013). Mental health benefits of neighbourhood green space are stronger among physically active adults in middle-to-older age: evidence from 260,061 australians. *Preventive medicine*, 57(5):601–606.

[11] Astell-Burt, T., Feng, X., and Kolt, G. S. (2014). Green space is associated with walking and moderate-to-vigorous physical activity (mvpa) in middle-to-older-aged adults: findings from 203 883 australians in the 45 and up study. *Br J Sports Med*, 48(5):404–406.

[12] Astell-Burt, T., Feng, X., and Kolt, G. S. (2016). Large-scale investment in green space as an intervention for physical activity, mental and cardiometabolic health: study protocol for a quasi-experimental evaluation of a natural experiment. *BMJ open*, 6(4):e009803.

[13] Aune, D., Norat, T., Leitzmann, M., Tonstad, S., and Vatten, L. J. (2015). *Physical activity and the risk of type 2 diabetes: a systematic review and dose–response meta-analysis*. Springer.

[14] Austin, N., Harper, S., and Strumpf, E. (2016). Does segregation lead to lower birth weight? *Epidemiology*, 27(5):682–689.

[15] Australian Consortium for Classification Development (2018). ICD-10-AM/ACHI/ACS. Available at https://www.accd.net.au/icd10.aspx.

[16] Australian Institute of Health and Welfare (2017). Impact of physical inactivity as a risk factor for chronic conditions: Australian burden of disease study. *Australian Burden of Disease Study series*, (15).

[17] Australian Institute of Health and Welfare (AIHW) (2003). The Active Australia Survey: a guide and manual for implementation, analysis and reporting. Available at https://www.aihw.gov.au/getmedia/ff25c134-5df2-45ba-b4e1-6c214ed157e6/aas.pdf.aspx?inline=true.

[18] Australian Institute of Health and Welfare Canberra (2017). Health expenditure australia. (58). Available at https://www.aihw.gov.au/getmedia/3a34cf2c-c715-43a8-be44-0cf53349fd9d/20592.pdf.

[19] Banks, E., Jorm, L., Rogers, K., Clements, M., and Bauman, A. (2011). Screen-time, obesity, ageing and disability: findings from 91 266 participants in the 45 and up study. *Public health nutrition*, 14(1):34–43.

[20] Barros, M. V., Ritti-Dias, R. M., Honda Barros, S. S., Mota, J., and Andersen, L. B. (2013). Does self-reported physical activity associate with high blood pressure in adolescents when adiposity is adjusted for? *Journal of sports sciences*, 31(4):387–395.

[21] Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

[22] Bishop, Y., Light, R., Mosteller, F., Fienberg, S., and Holland, P. (2007). *Discrete Multivariate Analysis Theory and Practice*. Springer Science+Business Media, LLC, New York, NY.

[23] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

[24] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[25] Blough, D., Madden, C., and Hornbrook, M. (1999). Modeling risk using generalized linear models. *Journal of Health Economics*, 18(2):153–171.

[26] Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450.

[27] Box, G. and Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series A (General)*, 26(2):211–252.

[28] Brookhart, M. A., Rassen, J. A., and Schneeweiss, S. (2010). Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and drug safety*, 19(6):537–554.

[29] Buntin, M. and Zaslavsky, A. (2004). Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics*, 23(3):525–542.

[30] Cadilhac, D. A., Magnus, A., Sheppard, L., Cumming, T. B., Pearce, D. C., and Carter, R. (2011). The societal benefits of reducing six behavioural risk factors: an economic modelling study from australia. *BMC public health*, 11(1):483.

[31] Carson, S. and Kenny, S. (2015). ILDAvis: Interactive Visualization of Topic Models.

[32] Chang, H.-Y., Lee, W.-C., and Weiner, J. P. (2010). Comparison of alternative risk adjustment measures for predictive modeling: high risk patient case finding using taiwan's national health insurance claims. *BMC health services research*, 10(1):343.

[33] Chen, R., Ovbiagele, B., and Feng, W. (2016). Diabetes and stroke: epidemiology, pathophysiology, pharmaceuticals and outcomes. *The American journal of the medical sciences*, 351(4):380–386.

[34] Clarke, P. S. and Windmeijer, F. (2012). Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107(500):1638–1652.

[35] Cobiac, L. J., Vos, T., and Barendregt, J. J. (2009). Cost-effectiveness of interventions to promote physical activity: a modelling study. *PLoS medicine*, 6(7):e1000110.

[36] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[37] Cybenko, G. (1989). Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:183–192.

[38] Dalziel, K., Segal, L., and Elley, C. R. (2006). Cost utility analysis of physical activity counselling in general practice. *Australian and New Zealand journal of public health*, 30(1):57–63.

[39] Darling, W. M. (2011). A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 642–647.

[40] Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.

[41] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[42] Diaz, K. M. and Shimbo, D. (2013). Physical activity and the prevention of hypertension. *Current hypertension reports*, 15(6):659–668.

[43] Diehr, P., Yanez, D., Ash, A., Hornbrook, M., and Lin, D. (1999). Methods for analyzing health care utilization and costs. *Annual review of public health*, 20(1):125–144.

[44] Diep, L., Kwagyan, J., Kurantsin-Mills, J., Weir, R., and Jayam-Trouth, A. (2010). Association of physical activity level and stroke outcomes in men and women: a meta-analysis. *Journal of women's health*, 19(10):1815–1822.

[45] Ding, D., Lawson, K. D., Kolbe-Alexander, T. L., Finkelstein, E. A., Katzmarzyk, P. T., Van Mechelen, W., Pratt, M., Committee, L. P. A. S. . E., et al. (2016). The economic burden of physical inactivity: a global analysis of major non-communicable diseases. *The Lancet*, 388(10051):1311–1324.

[46] Djurhuus, S., Aadahl, M., Hansen, H., Gl, C., et al. (2012). Associations between accessibility of public transportation and self-reported commuting physical activity. *Journal of Science and Medicine in Sport*, 15:S71.

[47] Duan, N. (1983). Smearing estimate: A Nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383):605–610.

[48] Duan, N., Manning, W., Morris, C., and Newhouse, J. (1984). Choosing between the Sample-Selection Model and the Multi-Part Model. *Journal of Business & Economic Statistics*, 2(3):283–289.

[49] Elley, C. R., Kerse, N., Arroll, B., Swinburn, B., Ashton, T., and Robinson, E. (2004). Cost-effectiveness of physical activity counselling in general practice. *New Zealand Medical Journal*, 117(1207):8–10.

[50] Ellis, R. P., Fiebig, D. G., Johar, M., Jones, G., and Savage, E. (2013). Explaining health care expenditure variation: Large-sample evidence using linked survey and health administrative data. *Health economics*, 22(9):1093–1110.

[51] Ertefaie, A., Small, D. S., Flory, J. H., and Hennessy, S. (2017). A tutorial on the use of instrumental variables in pharmacoepidemiology. *Pharmacoepidemiology and drug safety*, 26(4):357–367.

[52] Felder, S. (2001). Health care expenditure towards the end of life. *Cardiovascular drugs and therapy*, 15(4):345–347.

[53] Feng, X., Girosi, F., and McRae, I. S. (2014). People with multiple unhealthy lifestyles are less likely to consult primary healthcare. *BMC family practice*, 15(1):126.

[54] Fienberg, S. E. et al. (1970). An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907–917.

[55] Fisher, K. L., Harrison, E. L., Reeder, B. A., Sari, N., and Chad, K. E. (2015a). Is self-reported physical activity participation associated with lower health services utilization among older adults? cross-sectional evidence from the canadian community health survey. *Journal of aging research*, 2015.

[56] Fisher, K. L., Harrison, E. L., Reeder, B. A., Sari, N., and Chad, K. E. (2015b). Is self-reported physical activity participation associated with lower health services utilization among older adults? cross-sectional evidence from the canadian community health survey. *Journal of aging research*, 2015.

[57] Frew, E. J., Bhatti, M., Win, K., Sitch, A., Lyon, A., Pallan, M., and Adab, P. (2014). Cost-effectiveness of a community-based physical activity programme for adults (Be Active) in the UK: An economic analysis within a natural experiment. *British Journal of Sports Medicine*, 48(3):207–212.

[58] George, E. S., Rosenkranz, R. R., and Kolt, G. S. (2013). Chronic disease and sitting time in middle-aged australian males: findings from the 45 and up study. *International Journal of Behavioral Nutrition and Physical Activity*, 10(1):20.

[59] Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC.

[60] Gillies, R. J., Kinahan, P. E., and Hricak, H. (2015). Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577.

[61] Graddy, K. (1995). Testing for imperfect competition at the fulton fish market. *The RAND Journal of Economics*, pages 75–92.

[62] Gregori, D., Petrinco, M., Bo, S., Desideri, A., Merletti, F., and Pagano, E. (2011). Regression models for analyzing costs and their determinants in health care: an introductory review. *International Journal for Quality in Health Care*, 23(3):331–341.

[63] Griswold, M., Parmigiani, G., Potosky, A., and Lipscomb, J. (2004). Analyzing health care costs: a comparison of statistical methods motivated by medicare colorectal cancer charges. *Biostatistics*, 1(1):1–23.

[64] Gül, M. and Güneri, A. F. (2015). Forecasting patient length of stay in an emergency department by artificial neural networks. *Journal of Aeronautics and Space Technologies*, 8(2):43–48.

[65] Hachesu, P., Ahmadi, M., Alizadeh, S., and Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare Informatics Research*, 19(2):121–129.

[66] Hanandita, W. and Tampubolon, G. (2014). Does poverty reduce mental health? an instrumental variable analysis. *Social Science & Medicine*, 113:59–67.

[67] Hanning, B. W. (2007). Length of stay benchmarking in the australian private hospital sector. *Australian Health Review*, 31(1):150–158.

[68] Haskell, W. L. and Kiernan, M. (2000). Methodologic issues in measuring physical activity and physical fitness when evaluating the role of dietary supplements for physically active people–. *The American journal of clinical nutrition*, 72(2):541S–550S.

[69] Hearst, N., Newman, T. B., and Hulley, S. B. (1986). Delayed effects of the military draft on mortality. *New England Journal of Medicine*, 314(10):620–624.

[70] Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271.

[71] Ho, V., Hamilton, B. H., and Roos, L. L. (2000). Multiple approaches to assessing the effects of delays for hip fracture patients in the united states and canada. *Health services research*, 34(7):1499.

[72] Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

[73] Hornik, K. and Grün, B. (2011). topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

[74] Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

[75] Iacus, M., S., King, Gary, Porro, and Giuseppe (2016). *cem: Coarsened Exact Matching*. R package version 1.1.17.

[76] Iacus, S. M., King, G., and Porro, G. (2011). Multivariate Matching Methods That Are Monotonic Imbalance Bounding. *Journal of the American Statistical Association*, 106(493).

[77] Iacus, S. M., King, G., Porro, G., and Katz, J. N. (2012). Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, 20:1–24.

[78] Independent Hospital Pricing Authority (2012). National Efficient Price Determination 2012-2013.

[79] Inter-Government and Funding strategies (2011). Costs of Care Standards 2009/10. Available at https://www1.health.nsw.gov.au/pds/ArchivePDSDocuments/GL2011_007.pdf.

[80] Ionescu-Ittu, R., Abrahamowicz, M., and Pilote, L. (2012). Treatment effect estimates varied depending on the definition of the provider prescribing preference-based instrumental variables. *Journal of clinical epidemiology*, 65(2):155–162.

[81] Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55(1):179–188.

[82] Johnson, K. W., Soto, J. T., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., and Dudley, J. T. (2018). Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, 71(23):2668–2679.

[83] Jones, A. M. (2000). *Health Econometrics*, pages 265–344. Elsevier.

[84] Jones, A. M. et al. (2009). *Models for health care*. University of York., Centre for Health Economics.

[85] Jørgensen, H., Nakayama, H., Raaschou, H. O., and Olsen, T. S. (1994). Stroke in patients with diabetes. the copenhagen stroke study. *Stroke*, 25(10):1977–1984.

[86] Jothi, N., Rashid, N., and Husain, W. (2015). Data Mining in Healthcare - A Review. *Procedia Computer Science*, 72:306–313.

[87] Kang, S.-w. and Xiang, X. (2017). Physical activity and health services utilization and costs among us adults. *Preventive medicine*, 96:101–105.

[88] Kardamanidis, K., Lim, K., Da Cunha, C., Taylor, L. K., and Jorm, L. R. (2007). Hospital costs of older people in new south wales in the last year of life. *Medical Journal of Australia*, 187(7):383.

[89] Katzmarzyk, P. T. and Janssen, I. (2004). The economic costs associated with physical inactivity and obesity in canada: an update. *Canadian journal of applied physiology*, 29(1):90–115.

[90] Kaufman, R. L. (2013). *Heteroskedasticity in regression: Detection and correction*, volume 172. Sage Publications.

[91] Khoo, J., Hasan, H., and Eagar, K. (2018). Examining the high users of hospital resources: implications of a profile developed from australian health insurance claims data. *Australian Health Review*, 42(5):600–606.

[92] King, G. and Nielsen, R. (2016). Why propensity scores should not be used for matching. Available at http://j.mp/1FQhySn.

[93] Kissela, B. M., Khoury, J., Kleindorfer, D., Woo, D., Schneider, A., Alwell, K., Miller, R., Ewing, I., Moomaw, C. J., Szaflarski, J. P., Gebel, J., Shukla, R., and Broderick, J. P. (2005). Epidemiology of ischemic stroke in patients with diabetes. *Diabetes Care*, 28(2):355–359.

[94] Kleiber, C. and Zeileis, A. (2008). *Applied Econometrics with R*. Springer-Verlag, New York. ISBN 978-0-387-77316-2.

[95] Klungel, O., Jamal Uddin, M., de Boer, A., Belitser, S., Groenwold, R., and Roes, K. (2015). Instrumental variable analysis in epidemiologic studies: an overview of the estimation methods. *Pharm Anal Acta*, 6(353):2.

[96] Kronick, R., Dreyfus, T., Lee, L., and Zhou, Z. (1996). Diagnostic risk adjustment for medicaid: the disability payment system. *Health care financing review*, 17(3):7.

[97] Kubota, Y., Iso, H., Yamagishi, K., Sawada, N., and Tsugane, S. (2017). Daily total physical activity and incident stroke: The japan public health center–based prospective study. *Stroke*, pages STROKEAHA–117.

[98] Lai, J. K., Lucas, R. M., Armstrong, M., and Banks, E. (2013). Prospective observational study of physical functioning, physical activity, and time outdoors and the risk of hip fracture: A population-based cohort study of 158,057 older adults in the 45 and up study. *Journal of Bone and Mineral Research*, 28(10):2222–2231.

[99] Lee, I.-M. and Paffenbarger Jr, R. S. (1998). Physical activity and stroke incidence: the harvard alumni health study. *Stroke*, 29(10):2049–2054.

[100] Lella, L., Di Giorgio, A., and Dragoni, A. F. (2015). Length of stay prediction and analysis through a growing neural gas model. In *AI-AM/NetMed@ AIME*, pages 11–21.

[101] Li, J., Loerbroks, A., and Angerer, P. (2013). Physical activity and risk of cardiovascular disease: what does the new epidemiological evidence show? *Current opinion in cardiology*, 28(5):575–583.

[102] Lim, S., Wyker, B., Bartley, K., and Eisenhower, D. (2015). Measurement error of self-reported physical activity levels in new york city: assessment and correction. *American journal of epidemiology*, 181(9):648–655.

[103] Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., AlMazroa, M. A., Amann, M., Anderson, H. R., Andrews, K. G., et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2224–2260.

[104] Lisboa, P. J. and Taktak, A. F. (2006). The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, 19(4):408–415.

[105] Liu, X., Zhang, D., Liu, Y., Sun, X., Han, C., Wang, B., Ren, Y., Zhou, J., Zhao, Y., Shi, Y., et al. (2017). Dose–response association between physical activity and incident hypertension: A systematic review and meta-analysis of cohort studies. *Hypertension*, pages HYPERTENSIONAHA–116.

[106] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[107] Mandelbaum, A. and Shalev, A. (2016). Word embeddings and their use in sentence classification tasks. *arXiv preprint arXiv:1610.08229*.

[108] Manning, W. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, 17(3):283–295.

[109] Manning, W., Basu, A., and Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*, 24(3):465–488.

[110] Manning, W., Newhouse, J., Duan, N., Keeler, E., Leibowitz, A., and Marquis, M. (1987). Health insurance and the demand for medical care: evidence from a randomized experiment. *The American Economic Review*, 77(3):251–277.

[111] Manning, W. G. and Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of health economics*, 20(4):461–494.

[112] Marti, B., Tuomilehto, J., Salonen, J. T., PUSKA, P., and Nissinen, A. (1987). Relationship between leisure-time physical activity and risk factors for coronary heart disease in middle-aged finnish women. *Acta Medica Scandinavica*, 222(3):223–230.

[113] McClellan, M., McNeil, B. J., and Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?: analysis using instrumental variables. *Jama*, 272(11):859–866.

[114] Mealing, N. M., Banks, E., Jorm, L. R., Steel, D. G., Clements, M. S., and Rogers, K. D. (2010). Investigation of relative risk estimates from studies of the same population with contrasting response rates and designs. *BMC Medical Research Methodology*, 10.

[115] Medina, C., Janssen, I., Barquera, S., Bautista-Arredondo, S., Gonzalez, M. E., and Gonzalez, C. (2018). Occupational and leisure time physical inactivity and the risk of type ii diabetes and hypertension among mexican adults: A prospective cohort study. *Scientific reports*, 8(1):5399.

[116] Meenan, R., Goodman, M., Fishman, P., Hornbrook, M., O'Keeffe-Rosetti, M., and Bachman, D. (2003). Using risk-adjustment models to identify high-cost risks. *Medical Care*, 41(11):1301–1312.

[117] Miguel, E., Satyanath, S., and Sergenti, E. (2004). Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy*, 112(4):725–753.

[118] Mihaylova, B., Briggs, A., O'Hagan, A., and Thompson, S. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, 20(8):897–916.

[119] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[120] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[121] Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

[122] Min, J.-Y. and Min, K.-B. (2016). Excess medical care costs associated with physical inactivity among korean adults: retrospective cohort study. *International journal of environmental research and public health*, 13(1):136.

[123] Moen, S. and Ananiadou, T. S. S. (2013). Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, pages 39–43.

[124] Morton, A., Marzban, E., Giannoulis, G., Patel, A., Aparasu, R., and Kakadiaris, I. A. (2014). A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. In *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*, pages 428–431. IEEE.

[125] Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, 17(3):247–281.

[126] Müller-Riemenschneider, F., Reinhold, T., and Willich, S. N. (2009). Cost-effectiveness of interventions promoting physical activity. *British journal of sports medicine*, 43(1):70–6.

[127] Mytton, O. T., Townsend, N., Rutter, H., and Foster, C. (2012). Green space and physical activity: an observational study using health survey for england data. *Health & place*, 18(5):1034–1041.

[128] Neuman, P., Cubanski, J., and Damico, A. (2015). Medicare per capita spending by age and service: new data highlights oldest beneficiaries. *Health Affairs*, 34(2):335–339.

[129] Newell, S. A., Girgis, A., Sanson-Fisher, R. W., and Savolainen, N. J. (1999). The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population: a critical review. *American journal of preventive medicine*, 17(3):211–229.

[130] Nguyen, B., Bauman, A., and Ding, D. (2017). Incident type 2 diabetes in a large australian cohort study: Does physical activity or sitting time alter the risk associated with body mass index? *Journal of Physical Activity and Health*, 14(1):13–19.

[131] OECD (2015). *Fiscal Sustainability of Health Systems*. Available at https://www.oecd-ilibrary.org/content/publication/9789264233386-en.

[132] Pagano, E., Petrelli, A., Picariello, R., Merletti, F., Gnavi, R., and Bruno, G. (2015). Is the choice of the statistical model relevant in the cost estimation of patients with chronic diseases? An empirical approach by the Piedmont Diabetes Registry. *BMC Health Services Research*, 15(1):582.

[133] Paige, E., Korda, R., Banks, E., and Rodgers, B. (2014). How weight change is modelled in population studies can affect research findings: empirical results from a large-scale cohort study. *BMJ open*, 4(6):e004860.

[134] Paschalidis, I. (2017). How machine learning is helping us predict heart disease and diabetes. *Harvard Business Review*.

[135] Pate, R. R., Pratt, M., Blair, S. N., Haskell, W. L., Macera, C. A., Bouchard, C., Buchner, D., Ettinger, W., Heath, G. W., King, A. C., et al. (1995). Physical activity and public health: a recommendation from the centers for disease control and prevention and the american college of sports medicine. *Jama*, 273(5):402–407.

[136] Pedisic, Z., Grunseit, A., Ding, D., Chau, J. Y., Banks, E., Stamatakis, E., Jalaludin, B. B., and Bauman, A. E. (2014). High sitting time or obesity: Which came first? bidirectional association in a longitudinal study of 31,787 a ustralian adults. *Obesity*, 22(10):2126–2130.

[137] Peeters, G. G., Gardiner, P. A., Dobson, A. J., and Brown, W. J. (2018). Associations between physical activity, medical costs and hospitalisations in older Australian women: Results from the Australian Longitudinal Study on Women's Health. *Journal of Science and Medicine in Sport*, 21(6):604–608.

[138] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[139] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

[140] Plotnikoff, R. C., Costigan, S. A., Short, C., Grunseit, A., James, E., Johnson, N., Bauman, A., D'Este, C., van der Ploeg, H. P., and Rhodes, R. E. (2015). Factors associated with higher sitting time in general, chronic disease, and psychologically-distressed, adult populations: findings from the 45 & up study. *PloS one*, 10(6):e0127689.

[141] Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Iezzoni, L. I., Ingber, M. J., Levy, J. M., and Robst, J. (2004). Risk adjustment of medicare capitation payments using the cms-hcc model. *Health care financing review*, 25(4):119.

[142] Population Health Division (2011). Physical Activity and Sedentary Behaviour. Available at http://www.health.gov.au/internet/main/publishing.nsf/Content/pasb.

[143] Productivity Commission (2009). Public and Private Hospitals, Research Report, Canberra. Available at https://www.pc.gov.au/inquiries/completed/hospitals/report.

[144] Qin, S. (2011). Comparing the matching properties of Coarsened Exact Matching, Propensity Score Matching, and Genetic Matching in a nationwide data and a simulation experiment. Master's thesis, University of Georgia, ATHENS, GEORGIA.

[145] Rassen, J. A., Schneeweiss, S., Glynn, R. J., Mittleman, M. A., and Brookhart, M. A. (2008). Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *American journal of epidemiology*, 169(3):273–284.

[146] Reiner, M., Niermann, C., Jekauc, D., and Woll, A. (2013). Long-term health benefits of physical activity–a systematic review of longitudinal studies. *BMC public health*, 13(1):813.

[147] Rocca, P., Beckman, A., Hansson, E. E., and Ohlsson, H. (2015). Is the association between physical activity and healthcare utilization affected by self-rated health and socio-economic factors? *BMC public health*, 15(1):737.

[148] Rosenkranz, R. R., Duncan, M. J., Rosenkranz, S. K., and Kolt, G. S. (2013). Active lifestyles related to excellent self-rated health and quality of life: cross sectional findings from 194,545 participants in the 45 and up study. *BMC Public Health*, 13(1):1071.

[149] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

[150] Sari, N. (2011). Exercise, physical activity and healthcare utilization: A review of literature for older adults. *Maturitas*, 70(3):285–289.

[151] Schone, E. and Brown, R. (2013). Risk adjustment: what is the current state of the art and how can it be improved? *POLICY*, 1:6.

[152] Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

[153] Seemab, S. and Qamar, U. (2015). Predicting patient's LOS by mining hospital data. In *3rd International Conference on Artificial Intelligence and Computer Science*.

[154] Semlitsch, T., Jeitler, K., Hemkens, L. G., Horvath, K., Nagele, E., Schuermann, C., Pignitter, N., Herrmann, K. H., Waffenschmidt, S., and Siebenhofer, A. (2013). Increasing physical activity for the treatment of hypertension: a systematic review and meta-analysis. *Sports medicine*, 43(10):1009–1023.

[155] Sims, J., Hill, K., Hunt, S., Haralambous, B., Brown, A., Engel, L., Huang, N., Kerse, N., and Ory, M. (2006). National physical activity recommendations for older Australians: Discussion document. *National Ageing Research Institute*, pages 1–164.

[156] Small, D. S. and Rosenbaum, P. R. (2008). War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association*, 103(483):924–933.

[157] Smith, D. P., Weber, M. F., Soga, K., Korda, R. J., Tikellis, G., Patel, M. I., Clements, M. S., Dwyer, T., Latz, I. K., and Banks, E. (2014). Relationship between lifestyle and health factors and severe lower urinary tract symptoms (luts) in 106,435 middle-aged and older australian men: population-based study. *PloS one*, 9(10):e109278.

[158] Stamatakis, E., Grunseit, A. C., Coombs, N., Ding, D., Chau, J. Y., Phongsavan, P., Bauman, A., et al. (2014). Associations between socio-economic position and sedentary behaviour in a large population sample of australian middle and older-aged adults: The social, economic, and environmental factor (seef) study. *Preventive medicine*, 63:72–80.

[159] Stamatakis, E., Rogers, K., Ding, D., Berrigan, D., Chau, J., Hamer, M., and Bauman, A. (2015). All-cause mortality effects of replacing sedentary time with physical activity and sleeping using an isotemporal substitution model: a prospective study of 201,129 mid-aged and older adults. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1):121.

[160] Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

[161] Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

[162] Stroke foundation (2018). Types of Stroke. Available at https://strokefoundation.org.au/About-Stroke.

[163] Taylor, D. (2014). Physical activity is medicine for older adults. *Postgraduate medical journal*, 90(1059):26–32.

[164] Tensorflow team (2018). Vector Representations of Words. Available at https://www.tensorflow.org/tutorials/representation/word2vec.

[165] The Department of Health (2018). Cardiovascular disease. Available at http://www.health.gov.au/internet/main/publishing.nsf/Content/chronic-cardio.

[166] Thomas, J. W., Grazier, K. L., and Ward, K. (2004). Comparing accuracy of risk-adjustment methodologies used in economic profiling of physicians. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 41(2):218–231.

[167] Tran, B., Falster, M. O., Douglas, K., Blyth, F., and Jorm, L. R. (2014). Health behaviours and potentially preventable hospitalisation: a prospective study of older australian adults. *PloS one*, 9(4):e93111.

[168] Treff, C., Benseñor, I., and Lotufo, P. (2017). Leisure-time and commuting physical activity and high blood pressure: the brazilian longitudinal study of adult health (elsa-brasil). *Journal of human hypertension*, 31(4):278.

[169] Trevor, H., Robert, T., and JH, F. (2009). *The elements of statistical learning: data mining, inference, and prediction.* New York, NY: Springer.

[170] Tsai, P., Chen, P., Chen, Y., Song, H., Lin, H., Lin, F., and Huang, Q. (2016). *Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network.*, volume Vol. 2016. J Heal. Eng.

[171] Van der Ploeg, H. P., Chey, T., Ding, D., Chau, J. Y., Stamatakis, E., and Bauman, A. E. (2014). Standing time and all-cause mortality in a large cohort of australian adults. *Preventive medicine*, 69:187–191.

[172] Van der Ploeg, H. P., Chey, T., Korda, R. J., Banks, E., and Bauman, A. (2012). Sitting time and all-cause mortality risk in 222 497 australian adults. *Archives of internal medicine*, 172(6):494–500.

[173] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

[174] Vijay, G., Wilson, E. C., Suhrcke, M., Hardeman, W., and Sutton, S. (2016). Are brief interventions to increase physical activity cost-effective? a systematic review. *Br J Sports Med*, 50(7):408–417.

[175] Von Korff, M., Wagner, E. H., and Saunders, K. (1992). A chronic disease score from automated pharmacy data. *Journal of clinical epidemiology*, 45(2):197–203.

[176] Warburton, D. E., Charlesworth, S., Ivey, A., Nettlefold, L., and Bredin, S. S. (2010). A systematic review of the evidence for canada's physical activity guidelines for adults. *International Journal of Behavioral Nutrition and Physical Activity*, 7(1):39.

[177] Warburton, D. E., Nicol, C. W., and Bredin, S. S. (2006). Health benefits of physical activity: the evidence. *Canadian medical association journal*, 174(6):801–809.

[178] Weiner, J., Dobson, A., Maxwell, S., Coleman, K., Starfield, B., and Anderson, G. (1996). Risk-adjusted Medicare capitation rates using ambulatory and inpatient diagnoses. *Health Care Financing Review*, 17(3):77–99.

[179] Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

[180] Williams, R. J. and Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. *Backpropagation: Theory, architectures, and applications*, 1:433–486.

[181] Winkelman, R. and Mehmud, S. (2007). A comparative analysis of claims-based tools for health risk assessment. *Society of Actuaries*, pages 1–70.

[182] World Health Organization (2018a). Cardiovascular diseases . Available at https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[183] World Health Organization (2018b). Global action plan on physical activity 2018–2030: more active people for a healthier world.

[184] Wrenn, J., Jones, I., Lanaghan, K., Congdon, C. B., and Aronsky, D. (2005). Estimating patient's length of stay in the emergency department with an artificial neural network. In *AMIA... Annual Symposium proceedings. AMIA Symposium*, volume 2005, pages 1155–1155. American Medical Informatics Association.

[185] Wu, D.-M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica: journal of the Econometric Society*, pages 733–750.

[186] Wu, D.-M. (1974). Alternative tests of independence between stochastic regressors and disturbances: Finite sample results. *Econometrica: Journal of the Econometric Society*, pages 529–546.

[187] Wynand, P., De Ven, V., and Ellis, R. P. (2000). Risk adjustment in competitive health plan markets. In *Handbook of health economics*, volume 1, pages 755–845. Elsevier.

[188] Yoon, J.-H. and So, W.-Y. (2013). Association between leisure-time physical activity and hypertension status in korean adults. *salud pública de méxico*, 55(5):492–497.

[189] Yorston, L. C., Kolt, G. S., and Rosenkranz, R. R. (2012). Physical activity and physical function in older adults: The 45 and u p s tudy. *Journal of the American Geriatrics Society*, 60(4):719–725.

[190] Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., and Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*, 7:12474.

[191] Zhao, Y., Ash, A. S., Haughton, J., and McMillan, B. (2003). Identifying future high-cost cases through predictive modeling. *Disease Management & Health Outcomes*, 11(6):389–397.

APPENDIX A

IPF.Reweight.R

```
IPF.Reweight <- function(x, weight.name, marginals, max.iter=100, tol=1e-5, n.consec=5,
                         verbose=FALSE){

### DESCRIPTION
###
###     'IPF.Reweight' performs Iterative Proportional Fitting (IPF) to adjust a
###     set of survey weights to a given number of marginals
###
### ARGUMENTS
###
### x: is the data frame that contains the weights and the variables whose
###    marginals we wish to update
###
### weight.name: is the name of the original weight variable, the one we need to
###              adjust
###
### marginals: is a named list of marginal distributions, with one element for
###            each marginal. The names of the elements must correspond to names
###            of columns of x. Each element, representing a marginal
###            distribution, is a named vector, whose values sum up to 1. The
###            names of the elements of the vector must correspond to factors
###            levels of the corresponding variable.
###            Example: marginals <- list(smokecat=c(smoker=0.2,not.smoker=0.8),
###                                       bmi=c(obese=0.2, not.obese=0.8))
###            In this example smokecat and bmi are two factors in the data x, that take
###            values in (smoker, not.smoker) and (obese, not.obese) respectively.
###
### max.iter: the maximum number of iterations of the IPF. If this number is
###           exceeded an error occurs and an error message is printed
###
### tol: a tolerance used to decide when the IPF has converged. The algorithm
###      converges when the maximum relative error between the desired and current
###      marginals is smaller than the tolerance for n.consec number of times.
###
### n.consec: see description of tol
###
### verbose: if TRUE information about convergence is printed at each step of
###          the iteration
###
```

```
### VALUE
### the function returns a list with the following elements:
###
### w.new: the new set of weights computed by the IPF
###
### res: a named list, with as many elements as marginals. Each element is matrix with 4 columns:
###           old: the old marginal, computed using the old set of weights in x
###           new: the new marginal, computed by the new set of weights w.new
###           target: the target marginal, the parameter passed to the IPF that needs to be matched
###           delta: the difference between the target and the new marginal, a
###                  very small number that should be smaller than the tolerance tol
### gamma.mat: a matrix with as many columns as Lagrange multpliers. Each row
###             correspond to an IPF iteration.
### delta: a vector with as many elements as IPF iterations, containing the
###        maximum relative error between target and current marginal for each
###        iteration. The first element is always initialized to Inf.
###
### EXAMPLE
### ### first we create a synthetic data set of size N. There are two factors, smokecat and
### ### bmi, used for the adjustment, taking values in (smoker, not.smoker) and
### ### (obese, not.obese) respectively. There is also a set of weights w.
### N <- 30
### smokecat <- sample(c("smoker","not.smoker"),size=N, replace=TRUE)
### bmi <- sample(c("obese","not.obese"),size=N, replace=TRUE)
### w <- runif(N)
### w <- w/sum(w) ### it is not necessary that the weights are normalized to 1
### dat <- data.frame(cbind(smokecat, bmi),w=w) ### this is our data set
### ### we want to change the weights w to match the following marginals
### marginals <- list(smokecat=c(smoker=0.2,not.smoker=0.8), bmi=c(obese=0.3, not.obese=0.7))
### ### here is the main call:
### ipf.result <- IPF.Reweight(x=dat, weight.name="w", marginals=marginals, verbose=TRUE)
###
### ipf.result$res
### $res$smokecat
###                   old new target          delta
### smoker      0.3949961 0.2    0.2 -8.086865e-12
### not.smoker 0.6050039 0.8    0.8  2.021716e-12
###
### $res$bmi
###                  old new target         delta
### obese      0.6034208 0.3    0.3 1.850372e-16
### not.obese 0.3965792 0.7    0.7 0.000000e+00
###
###
### ipf.result$gamma.mat
###          smoker not.smoker      obese not.obese
###  [1,] 1.376359    3.594399 0.5416057  1.569185
###  [2,] 1.739438    3.416134 0.5227507  1.593822
###  [3,] 1.757850    3.407212 0.5218175  1.595045
###  [4,] 1.758770    3.406766 0.5217709  1.595106
###  [5,] 1.758816    3.406744 0.5217686  1.595109
###  [6,] 1.758818    3.406743 0.5217685  1.595109
###  [7,] 1.758818    3.406743 0.5217685  1.595109
###  [8,] 1.758818    3.406743 0.5217685  1.595109
###  [9,] 1.758818    3.406743 0.5217685  1.595109
###
### ipf.result$delta
### [1]          Inf 1.047444e-02 5.229716e-04 2.610363e-05 1.302919e-06
### [6] 6.503299e-08 3.246010e-09 1.620190e-10 8.086865e-12
```

```
### CODE STARTS HERE

### we start with some consistency checks
    if (max.iter < n.consec + 2)
        stop("\nmax.iter is too small compared to n.consec\n")

### check that names of marginals match names of variables in x
    if (any(! names(marginals) %in% colnames(x)))
        stop("\nnames of marginals do not match any column of x\n")

### check that the marginals are possible marginals of the corresponding
### variables in x (make sure that they have values in the same sets)
    for (n in names(marginals)){
        v <- unique(x[,n])
        if (length(setdiff(v, names(marginals[[n]]))) > 0)
            stop("\nmarginal for ",n," is not valid: the variable ",n," takes different values in the data frame\n")
    }

    pop <- sum(x[, weight.name])
### N.vec stores the marginals in one single vector
    N.vec <- NULL
    for (i in 1:length(marginals)){
### make sure that marginals sum exactly to 1
        m <- marginals[[i]]
        m[length(m)] <- 1-sum(m[1:(length(m)-1)])
        marginals[[i]] <- m
        N.vec <- c(N.vec, m)
    }
    N.vec <- N.vec*pop

### gamma.vec stores the Lagrange multipliers, initialized to 1
    gamma.vec <- 0*N.vec + 1

### build the design matrix theta, and make sure its columns match N.vec
    theta <- MakeDummies(x,names(marginals)[1])
    if (length(marginals) > 1){
        nam <- names(marginals)[2:length(marginals)]
        for (n in nam)
            theta <- cbind(theta,MakeDummies(x,n))}
    theta <- theta[,names(N.vec)]

    iter <- 0
    converged <- FALSE
    w <- x[,weight.name]
    gamma.mat <- matrix(NA, nrow=max.iter,ncol=length(gamma.vec))
    colnames(gamma.mat) <- names(gamma.vec)
    delta <- rep(Inf, max.iter)

    while(!converged){
        iter <- iter + 1
        if (iter > max.iter)
            stop("\nIPF failed to converge\n")
### this for loop is the core of the IPF
        for (multiname in names(gamma.vec)){
            gamma.vec[multiname] <- IPF.EstimateMultiplier(multiname, w, theta, N.vec, gamma.vec)
        }
### we store the gammas in a matrix so we can study convergence
        gamma.mat[iter,] <- gamma.vec
```

```
###
### ESTIMATE ERROR IN MARGINALS
###

### update weights so we can compare marginals
        x$w.new <- IPF.UpdateWeights(w, gamma.vec, theta)

### compare marginals before and after IPF with target marginals
        res <- marginals
        err <- NULL
        for (n in names(marginals)){
            t1 <- my.table(x,weight.name,n, NULL, perc=TRUE)
            t2 <- my.table(x,"w.new",n, NULL, perc=TRUE)
            t1 <- t1[names(marginals[[n]])]
            t2 <- t2[names(marginals[[n]])]
            mat <- cbind(t1,t2,marginals[[n]],(t2-marginals[[n]])/marginals[[n]])
            colnames(mat) <- c("old","new","target","delta")
            res[[n]] <- mat
            err <- c(err, mat[,"delta"])
        }

### also check for the total size of population
        err <- c(err, abs((sum(x$w.new)-pop)/pop))
### delta is the maximum deviation in gamma from one iteration to the next
        if (iter > 1)
            delta[iter] <- max(abs(err))
        vcat("\nIteration:",iter,"; delta:",delta[iter],"\n",verbose=verbose)

### we say that the IPF has converged if the error delta is smaller than
### the tolerance tol more than n.consec times in a row
        if (iter >= n.consec+1){ ### we need to run at least n.consec to make this check
            if (all(delta[(iter-n.consec+1):iter] < tol))
                converged <- TRUE
        }
    } ### end of while
    w.new <- IPF.UpdateWeights(w, gamma.vec, theta)
    gamma.mat <- gamma.mat[1:iter,]
    delta <- delta[1:iter]
    vcat("\n\nIPF converged in",iter,"iterations\n",verbose=verbose)
    return(list(w.new=w.new, res=res, gamma.mat=gamma.mat,delta=delta))
}
```