A Model of Health: Using Business Analytics to Identify Older Canadian Adults with Heart Disease

by

Timothy Ainge

B.Sc., Simon Fraser University, 2016

B.B.A., Simon Fraser University, 2016

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Biomedical Physiology and Kinesiology
Faculty of Science

SIMON FRASER UNIVERSITY

Fall 2018

Approval

Name:	**Timothy Ainge**
Degree:	**Master of Science**
Title:	**A Model of Health: Using Business Analytics to Identify Older Canadian Adults with Heart Disease**
Examining Committee:	**Chair:** Damon Poburko Assistant Professor
	Dawn Mackey Senior Supervisor Associate Professor
	David Whitehurst Supervisor Associate Professor
	Guy Faulkner Supervisor Professor
	Andrew Wister External Examiner Professor Gerontology Simon Fraser University
Date Defended/Approved:	December 14, 2018

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Nearly 90% of older Canadians have at least one chronic disease; 65% have two or more. The aims of my thesis were to apply business analytics techniques to predict the presence of an exemplar chronic disease, heart disease, among older Canadians, and to calculate the corresponding expected healthcare costs. I used neural networks to develop logistic regression models of heart disease using demographic, lifestyle, and health information for 15,599 older adults from the Canadian Longitudinal Study on Aging. The Economic Burden of Illness in Canada provided healthcare cost data. The best model identified 65.8% of heart disease cases from 40% of participants with the highest predicted probabilities of heart disease, accounting for $2.7 million more expected annual healthcare costs than a randomly sampled 40%. Among all older Canadians, this difference would be $1.1 billion. These methods could assist healthcare decision makers to optimize the delivery of chronic disease prevention interventions.

Keywords: Older adults; heart disease; chronic disease; primary and secondary prevention interventions; business analytics; lift charts

Dedication

I dedicate this work to Donald Valentine, a man who showed me what it meant to be genuinely kind, whose stories sparked my imagination and curiosity, and who taught me how to believe in myself by believing so fiercely in me. He was my Mom's father, my Granny's husband, and my Poppa.

He absolutely loved being a grandpa. For the first three years of my life, hardly a day went by without a visit from him. This became a little bit more challenging when my parents and I emigrated from South Africa to Canada in 1995. Being so far away meant that he and I had to resign ourselves to only seeing each other every other day – he and my Granny spent half of each year visiting us for the next decade. Like clockwork, the highlight of each and every year was running up to him in the airport and hearing him say "Hi Timo".

I'm incredibly lucky that I got to spend so much time with him. From playing tennis in the driveway to exploring the city on the SeaBus to serious conversations at the dining table, my childhood was full of opportunities to bond with and learn from him. Sometimes this learning came in the form of lessons and words of wisdom, but most of what I remember about my Poppa was how he carried himself and how he treated other people. He showed me that it was important to balance work and play, that humility and confidence weren't mutually exclusive, and that everyone deserved to be loved and respected.

He passed away in 2014, and so I unfortunately couldn't share any of this work with him. By dedicating it to him, my hope is that the pivotal role he played – and continues to play – in my journey will never be forgotten. I wouldn't be the person I am today if it weren't for him. Thanks for everything, Poppa – I love you.

Acknowledgements

I would like to thank the following people for their contributions to my thesis:

My supervisor, Dr. Dawn Mackey, for encouraging me to think deeply about issues I'm interested in, for welcoming me into her lab, and for providing immeasurable support throughout the course of my degree; my committee members, Dr. David Whitehurst and Dr. Guy Faulkner for challenging me to value quality over quantity and for always being ready to talk about soccer; and Dr. Andrew Wister, my external examiner, for providing feedback that allowed me to put the finishing touches on my thesis, and for introducing me to the Canadian Longitudinal Study on Aging.

My labmates in the Aging and Population Health Lab for their friendship and advice: Stephanie Maganja, Ashley Kwon, Kristina Collins, Nicole Whittle, Amanda Zacharuk, Rania Khelifi, Aksel Smit-Anseeuw, Chantelle Kawala, Emaan Abbasi, Valeriya Zaborska, Amandeep Gill, Jenna Chow, and Eva Habib.

My colleagues in the Department of Biomedical Physiology & Kinesiology for their camaraderie, invaluable feedback on my work as it progressed, and numerous teaching and learning opportunities.

Simon Fraser University for supporting my research goals and providing funding that allowed me to reach them; the Canadian Institutes of Health Research for allowing me to participate in the 2016 Summer Program in Aging in Montreal; the University of British Columbia and Dr. Jennifer Gardy for allowing me to take SPPH 581S to further my understanding of risk communication in public health; and the Centre for Hip Health and Mobility for providing numerous learning, presentation, and networking opportunities.

And finally, my friends, family, and wonderful girlfriend, Emily, for being there to celebrate my accomplishments, for providing support and encouragement when I struggled, and for making each day of this journey more fun that the last.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AIC	Akaike Information Criterion
BMI	Body Mass Index
CAD	Canadian Dollars
CLSA	Canadian Longitudinal Study on Aging
EBIC	Economic Burden of Illness in Canada
FHS	Framingham Heart Study
FRS	Framingham Risk Score
ICD	International Classification of Diseases
IL	Interleukin
PASE	Physical Activity Scale for the Elderly

Glossary

Cost Neutrality	A requirement that the expected costs of a given choice not exceed the expected costs of another choice.
Expected Value	The product of predicted probability and cost.
Logistic Regression	A statistical model that uses a logistic function to model a binary dependent variable.
Neural Network	A network of mathematical activation functions consisting of a single input layer, one or more hidden layers, and a single output layer. The input layer can have an unlimited number of inputs, the hidden layers can have an unlimited number of activation functions, and the activation function in the output layer can be used to model binary, categorical, or continuous outcomes.
Older Adult	A person aged 65 years or older.

Chapter 1. Background

The primary purpose of my thesis research is to predict the presence of chronic diseases among older Canadian adults based on demographic, lifestyle, and medical characteristics. A secondary purpose is to calculate the expected healthcare costs of individual older adults based on their predicted probabilities of having a chronic disease and sum these to understand how they differ across subpopulations (e.g., women from Ontario aged 65-74 versus men from Alberta aged 75-84). I use statistical methods from the field of business analytics that are novel to population health research, leveraging machine learning and data visualization tools to address these purposes in ways that support decision-making for healthcare management and public policy stakeholders. My thesis is comprised of three chapters. In Chapter 1, I review the landscape of chronic diseases among older adults in Canada, summarize the economic concerns associated with treatment of preventable chronic disease, describe a common etiology for these chronic diseases, describe analytical methods to build models that predict chronic disease outcomes among older Canadian adults, and make the case for considering heart disease as an exemplar of chronic diseases for the proposed predictive modelling. In Chapter 2, I present the methods and results of predictive modelling of heart disease, as well as a summary of expected healthcare costs for heart disease. Finally, in Chapter 3, I discuss the results of my research, including limitations, context, and recommendations for researchers and healthcare decision makers.

1.1. CHRONIC DISEASES & OLDER ADULTS

1.1.1. Chronic disease prevalence

Chronic diseases are now responsible for more deaths in first-world countries such as Canada than infectious diseases (Murray et al., 2015); currently, nearly 90% of older Canadian adults (65+ years) are living with at least one chronic disease and 65% are living with at least two chronic diseases (Canadian Institute for Health Information, 2011a; Lopez, Mathers, Ezzati, Jamison, & Murray, 2006; Public Health Agency of Canada, 2010). Chronic disease prevalence (%) among older Canadian adults is

greatest for arthritis (43.4%), heart disease (22.6%), osteoporosis (18.1%), and diabetes (17.2%) (Ramage-Morin, Shields, & Martel, 2010). Therefore, further research examining the link between risk factors and common chronic diseases is a priority for improving the health of older Canadians (Public Health Agency of Canada, 2013).

1.1.2. A common etiology for many chronic diseases

Plausible causal relationships have been described for systemic inflammation as a common etiology of many chronic diseases, with signal transduction by mitogen-activated protein kinase pathways causing insulin resistance, atherosclerosis, neurodegeneration, and tumour growth (Pedersen, 2009; Kyriakis, 2001). C-reactive protein, muscle-derived interleukin-6, and local tumor necrosis factor alpha are common biomarkers for inflammation, and have each been shown to be associated with and predictive of chronic diseases, with the largest base of support for their relationship to chronic cardiovascular diseases (Ridker et al., 1997; Ridker et al., 2000; Danesh et al., 2004; Bruunsgaard, 2005). Moreover, the reduction in C-reactive protein by anti-inflammatory agents has been shown to be effective in reducing cardiovascular disease prevalence (Ridker et al., 1997). Although damage to physiological structures associated with chronic disease can be both a cause and effect of inflammation, as with atherosclerotic lesions in the cardiovascular system, secondary inflammation produced by immune responses to cell damage is involved in further pathogenesis (Akiyama et al., 2000).

Therefore, risk factors that increase inflammation are unsurprisingly associated with the development of chronic diseases (Wellen and Hotamisligil, 2005; Bruunsgaard, 2005; Pedersen, 2009). As a consequence, the following specific diseases have been shown to be associated with risk factors that are known to increase systemic inflammation, such as physical inactivity, excess weight, smoking, and poor diet: obesity, hyperlipidemia, type 2 diabetes, cardiovascular diseases (hypertension, coronary heart disease, heart failure, cerebral apoplexy, and intermittent claudication), colon cancer, breast cancer, dementia, and depression (Krueger, Koot, Rasali, Gustin, & Pennock, 2016; Menec, 2003; Pedersen, 2009).

1.1.3. Concomitant healthcare costs of chronic disease

In epidemiology, population attributable fractions describe the burden of disease that could be eliminated if risk factors for a particular disease were eliminated (Rockhill, Newman, & Weinberg, 1998). This type of evidence provides the impetus for research related to interventions designed to modify risk factors in the hopes of reducing chronic disease risk and incidence in a population. This is important, because it helps – in part – to simplify potentially complex disease pathways for the purposes of risk prediction and cost assignment. For example, it allows us to ask the question "how much money does the healthcare system spend treating the consequences of unhealthy lifestyle choices like physical inactivity?". That answer is startling. The economic burden of treating the consequences of physical inactivity, excess weight, and smoking was estimated at $52.8 billion in Canada in 2015 (Krueger, Krueger, & Koot, 2015).

The Canadian population is aging: 25% of Canadians will be over the age of 65 within the next 20 years, compared to just 13% currently (Statistics Canada, 2010). Given that most of the costs of chronic disease are borne in later life and that activity levels decline with age among older adults, the aging of the Canadian population could increase the economic burden of chronic diseases manyfold (Aminzadeh & Dalziel, 2002; Katzmarzyk & Janssen, 2004; Thompson, Kuhle, Koepp, McCrady-Spitzer, & Levine, 2014; Wilson et al., 2009). In 2015, direct healthcare spending on chronic diseases linked to physical inactivity, excess weight, and smoking was $16.5 billion CAD (Kruger et al., 2015). According to the Canadian Institute for Health Information, healthcare costs due to increased spending on the treatment of chronic diseases are increasing at an unsustainable rate (Canadian Institute for Health Information, 2011b; Canadian Institute for Health Information, 2015). A commonly held misconception is that by the time people reach old age, it is too late for interventions aimed at reducing their risk of developing disease to be impactful; however, there is considerable evidence to suggest that many of these costs could be dramatically reduced by increasing population levels of physical activity, even among older adults (Krueger et al., 2015; Warburton, Nicol, & Bredin, 2006).

As alluded to earlier, older adults, defined as 65 years or older, are at increased risk of having one or more chronic diseases compared to middle-aged adults (Meisner, Linton, & Séguin, 2017). This is especially concerning when considered in the context of

an older adult population that has not been well represented in research studies (McMurdo et al., 2011; Verbeeten, Astles, & Prada, 2015). With population aging being an important consideration for healthcare decision makers and most chronic diseases disproportionately affecting older adults, it follows that we should expect an increase in the number of older Canadian adults at-risk of developing chronic disease in the coming years (Statistics Canada, 2017a, 2017b). Understanding who is at highest risk of developing chronic diseases within this population could contribute to the development of effective strategies to deal with the expected increase in chronic disease cases over the coming years.

It is common for cost of illness analyses of chronic diseases to report costs from a societal perspective in the form of an economic burden that takes into account not only direct hospital-, physician-, and drug-related spending, but also indirect costs associated with lost productivity due to time off from work (Hjelmgren, Berggren, & Andersson, 2001; Roux et al., 2008). However, the societal perspective does not prioritize the budgetary constraints faced by, and healthcare utilization goals of, healthcare decision makers, which reduces the extent to which results generated using this perspective are considered actionable (Russell, Fryback, & Sonnenberg, 1999). To align the costs and benefits related to primary and secondary prevention interventions with the stakeholders responsible for them, it is better to consider the healthcare system perspective (Sullivan et al., 2014). For the purposes of this thesis, I mean a healthcare system perspective to be one that includes hospital-, physician-, and drug-related spending by the healthcare system. This approach does not rely on assumptions about the productivity and value of different members of society, which might become especially tenuous as the average age of retirement continues to increase, or as younger, working adults are increasingly likely to take time off from work in order to care for aging relatives (Jacobs, Laporte, Van Houtven, & Coyte, 2014). Since a healthcare system perspective reflects the true out-of-pocket costs that healthcare decision makers within various health agencies across the country are faced with, analyses generated with this perspective in mind are likely to be both actionable and relevant for these key stakeholders (Byford, Torgerson, & Raftery, 2000; Sullivan et al., 2014).

1.1.4. Specific chronic diseases to consider

The application of the business analytics techniques used in this thesis, which are described later, are intended as a proof-of-concept rather than a final solution to the problems and gaps illustrated throughout this background chapter; therefore, I chose to focus my research on a single chronic disease – heart disease. This approach also allows much of this thesis to have a methods-oriented, rather than results-oriented, perspective, which will be useful towards the eventual implementation of this work by other researchers and healthcare decision makers. Future research could extend to other chronic diseases.

From the list of chronic diseases linked to inflammatory pathways, the following nine were measured in both the Canadian Longitudinal Study on Aging (CLSA) and the Economic Burden of Illness in Canada (EBIC), data sources that were central to this thesis: obesity, type 2 diabetes, hypertension, coronary heart disease, cerebral apoplexy (stroke), colorectal cancer, breast cancer, dementia, and depression. I chose to focus on heart disease for several reasons. Classification modelling of the type used in this thesis requires the outcome – in this case, whether a study participant has heart disease – to be sufficiently common, with 10% often used as a rule-of-thumb for the minimum prevalence rate (Putler & Krider, 2012). Among older adults in the CLSA, heart disease (17.2%), hypertension (48.2%), type 2 diabetes (20.8%), obesity (25.9%), and depression (14.0%) were quite common, whereas stroke, colorectal cancer, breast cancer, and dementia had prevalence rates less than 10%, making them less than ideal candidates (Raina, Wolfson, & Kirkland, 2018). Further, hypertension is commonly associated with heart disease, and since heart disease is a more clinically relevant outcome in terms of specificity of healthcare treatment and costs incurred, I decided not to investigate hypertension. Of the remaining four diseases, heart disease also has the largest base of evidence supporting a relationship to risk factors related to systemic inflammation, which allows me to place my work within a much broader research landscape (Katzmarzyk & Janssen, 2004; Tikkanen, Gustafsson, & Ingelsson, 2018).

1.2. BUSINESS ANALYTICS IN HEALTHCARE

1.2.1. Outcome modelling from a business analytics perspective

Compared to causal and explanatory modelling commonly seen in health research, whose goals are to describe the effect of one or more exposure variables on an outcome towards better understanding disease etiology, modelling in a business context is typified by a focus on prediction and classification (MacNally, 2000; Putler & Krider, 2012). Rather than asserting that the independent variables cause – or at least partly cause – the outcome, the implication of predictive models is that knowing the values of the independent variables will allow us to predict – or at least partly predict – what the outcome will be in a given case. This allows for a straightforward, actionable interpretation of the modelling results: in other words, did the model correctly classify an individual according to the outcome measure of choice? Against the background of chronic diseases, where people either develop a given disease or they do not, employing modelling techniques that are consistent with how diseases are diagnosed and treated in the population is a logical choice.

1.2.2. Rise of business analytics techniques in healthcare

In the age of rapid computing and big data, it is becoming increasingly possible for governments and healthcare providers to measure and track numerous metrics of health in populations over time, allowing for novel concepts such as personalized medicine (Van Poucke et al., 2016). In fact, neural networks have been used to great effect in the classification and diagnosis of chronic disease using clinically measured variables (Er, Yumusak, & Temurtas, 2010). Because neural networks ‘learn’ about the relationships between predictor and outcome variables by observing each case sequentially, they offer a good alternative to investigator-imposed variable selection and modification, especially where these relationships are complex. Given the complexity of interacting systems (e.g., lifestyle choices, built environment, genetics) at the population level (Kindig & Stoddart, 2003), it follows that a neural network-driven approach would also be well-suited to identifying important non-linear trends or interactions between demographic, lifestyle, and medical variables to enable better classification of chronic

disease. However, the use of these relatively new techniques can be fraught with errors if researchers and decision makers do not understand the contexts in which they can be used and the types of conclusions that can be drawn: “The critical importance of poor calibration is frequently underappreciated; poor calibration can lead to harmful decisions” (Shah, Steyerberg, & Kent, 2018, p. 28). Shah and colleagues identify mismatches between calibration and application populations as a common contributor to predictive models with poor validity. This underscores the need for models to be produced for and calibrated on older adults specifically, rather than simply relying on broader examples in the literature.

1.3. MODELING RISK FOR CHRONIC DISEASE

1.3.1. Non-modifiable risk factors

The body of literature related to genetic and sex-based risk of chronic disease is vast, and it suggests that there are important differences in health trajectories over the life course that are assigned at birth (Tikkanen et al., 2018). Because these attributes are intrinsic characteristics and therefore cannot be changed, they must be included in disease classification models so that the true effect of modifiable risk factors can be known. Age, sex, and ethnicity are among the most commonly researched and strongly associated non-modifiable risk factors related to chronic disease, with increasing age and belonging to the male sex often associated with increased risk; the association between ethnicity and disease is very much disease-specific (Wilson et al., 2009; Yusuf, Reddy, Ounpuu, & Anand, 2001). Gender as a social construct separate from sex is also an important consideration; however, this was not captured within the CLSA and so could not be accounted for in my thesis research (Vlassoff, 2007). From an intervention perspective, it is clearly more relevant to consider the ways in which modifiable risk factors are related to disease risk of specific populations, especially among those who were born with an increased risk.

1.3.2. Modifiable risk factors

The modifiable risk factors that contribute most to the risk and economic burden of chronic diseases in Canada are physical inactivity, smoking, and obesity (Krueger et al., 2016). While smoking rates are declining, rates of obesity and physical inactivity have increased in the older adult population in Canada (Raina et al., 2018; Tremblay, Wolfson, & Gorber, 2007). In fact, only 15% of Canadian adults get an appropriate amount of physical activity for health benefits, and 1 in 4 older Canadian adults are considered obese (Body Mass Index (BMI) ≥ 30) (Colley et al., 2011; Statistics Canada, 2017a). Among those aged 65 and older, time spent being physically active decreases by approximately 2% each year (Buchman et al., 2014; Thompson et al., 2014). This is important because numerous epidemiologic studies provide evidence that physical inactivity increases the risk of many chronic diseases (e.g., coronary heart disease, type 2 diabetes, cancer) and all-cause mortality (Bauman, Merom, Bull, Buchner, & Fiatarone Singh, 2016; Elbaz et al., 2013; Katzmarzyk & Janssen, 2004; Lee et al., 2012; Studenski, Perera, & Patel, 2011; Toots et al., 2013; Warburton et al., 2006). While there does not appear to be a similar age-dependent relationship for obesity among older adults, population levels of obesity are increasing, and obesity is also associated with the aforementioned diseases and conditions (Katzmarzyk & Janssen, 2004; Krueger et al., 2015; Pedersen & Febbraio, 2012; Statistics Canada, 2017b). Given their common association with chronic disease, it should be unsurprising that physical inactivity and obesity are often associated with one another, and that physical activity interventions, often in combination with dietary interventions, are commonly used to combat obesity (Pedersen & Saltin, 2015; Rejeski et al., 2011; Warburton et al., 2006).

There is strong evidence for mechanisms that explain why physical activity is crucial for the prevention and management of chronic diseases linked to inflammation. Skeletal muscle, the architecture responsible for physical activity, has an important role as an endocrine organ, secreting cytokines such as interleukin 6 (IL-6), IL-8, IL-15, brain-derived neurotrophic factor, and leukemia inhibitory factor when contracted (Pedersen, 2009; Pedersen & Febbraio, 2012). These cytokines contribute to lipid metabolism, neuronal health, and reductions in systemic inflammation, which makes the case for physical activity as a means to manage (i.e., secondary prevention) depression, anxiety, stress, schizophrenia, Parkinson's disease, multiple sclerosis, polycystic ovarian syndrome, type 1 diabetes, pulmonary diseases (chronic obstructive pulmonary disease,

asthma, cystic fibrosis), musculoskeletal disorders (osteoarthritis, osteoporosis, back pain, rheumatoid arthritis), and other types of cancer (Pedersen & Febbraio, 2012; Pedersen & Saltin, 2015).

Other modifiable risk factors that have been shown to be associated with chronic disease development are education level (specifically a lack of high school graduation), income, nutrition, and alcohol consumption (Broekhuizen et al., 2018; Lopez et al., 2006).

1.4. METHODOLOGY

1.4.1. Business data mining overview

New evidence about relationships between modifiable risk factors and chronic disease, as well as the means to monitor and intervene on some of these risk factors, have been enabled by evolving research and technological advances, requiring solutions at the intersection of medical science and data science. Because the fields of marketing and business analytics are replete with strategies to leverage big-data techniques to solve real-world problems, they can offer such a solution. Therefore, the methodological framework for this thesis is guided by the Rapid Model Development Framework, based on the Cross-Industry Standard Process for Data Mining model (Chapman et al., 2000; Putler & Krider, 2012). Specifically, this framework provides a set of methods to fit logistic regression models of heart disease based on a wide range of modifiable and non-modifiable risk factors.

Whereas results from randomized controlled trials are useful towards establishing causal relationships, the effect sizes they generate are not necessarily externally valid due to the controlled conditions in these trials imposing unnatural behaviour on both the usual care (control) and treatment groups. To contextualize effect sizes for a specific population – for example, older Canadian adults – and to get the most precise outcome estimates, we need to account for a wide range of predictor variables that naturally vary from one participant to another. With sufficiently large samples of observational data, the Rapid Model Development Framework can be used to map potentially complex relationships involving several variables and the chosen

outcome of heart disease (Putler & Krider, 2012). The framework proposes the following 6-step process for developing models:

Step 1. Reviewing the literature to understand modelling context.

A review of the literature is important to understand how modifiable and non-modifiable risk factors are thought to contribute to heart disease respectively. An understanding of the proposed etiology of this relationship will provide a rationale for the inclusion of variables into exploratory, minimally-adjusted models and will aid in the further selection of additional variables for more complex models. Grounding any decisions made about variable inclusion, exclusion, or modification in the heart disease literature will also help to guard against spurious correlations that commonly arise in increasingly complex models. Relying on evidence to support decision-making could also provide useful insight into how potential predictor variables could be transformed or modified if important nonlinearities are found to be present.

Step 2. Protection against overfitting by splitting the dataset.

Splitting the study population into training and validation datasets is a necessary data management step. Models should be developed with the training dataset and assessed with the validation dataset. Splitting the data will ensure that we identify the most predictive models as the ones that provide good estimates of the study outcome (heart disease) in the validation dataset without modelling any random noise in the training dataset.

Step 3. Visualization tools to identify non-linear relationships.

It is useful to employ exploratory techniques (e.g., tree models, plots of means, principal component analysis) to inform the addition or transformation of predictor variables to several candidate logistic regression models, using the training dataset. For example, if there were important thresholds of leisure-time physical activity that split participants into groups that differ significantly in terms of heart disease prevalence, as

could be visualized with a tree model, we might decide to categorize leisure-time physical activity according to these thresholds. The resulting logistic regression models will be optimized by minimizing the value of the corresponding Akaike Information Criterion (AIC) for each model. Because the AIC penalizes the inclusion of additional predictor variables, it will help to prevent overfitting the models to trends unique to the study data but not necessarily representative of the underlying relationships or generalizable to the general population.

Step 4. Training neural networks.

To approximate the best-fitted logistic regression models, neural networks of varying flexibility should be used to examine whether important nonlinear relationships are present in the relationships between modifiable and non-modifiable risk factors and heart disease, using the training dataset. Because the relationships between predictor and outcome variables in natural systems are often non-linear, achieving optimally-fitted models will likely require transformation of some of the predictor variables included in the model (e.g., ln(age)) (Putler & Krider, 2012).

Step 5. Model ranking and selection.

It is important to score each model by ranking participants in the validation dataset according to their predicted probabilities of disease under each model, and plotting these results on lift charts against neural network models. The predictive utility of the resulting models will be assessed by means of a cumulative gains lift chart, whereby the proportion of participants with a given condition is plotted against the total number of participants, ranked in order (from high to low) of their predicted likelihood of having that certain chronic condition; see Figure 1.1. Figure 1.1 also shows that better models can identify the same proportion of disease cases using fewer individuals. This type of plot allows side-by-side comparison of multiple predictive models (e.g., logistic regression models with different sets of predictor variables) to support decision making by showing the number of disease cases that could be identified per amount of effort expended in identifying certain sub-populations. The more cases that can be identified by predicting heart disease outcomes for the fewest number of participants, the higher the lift and the

better the predictive utility. This will be especially important if these models are to ultimately be useful to health care decision makers to use resources more efficiently, and to maximize the cost-effectiveness and cost-benefit of delivering disease prevention interventions.

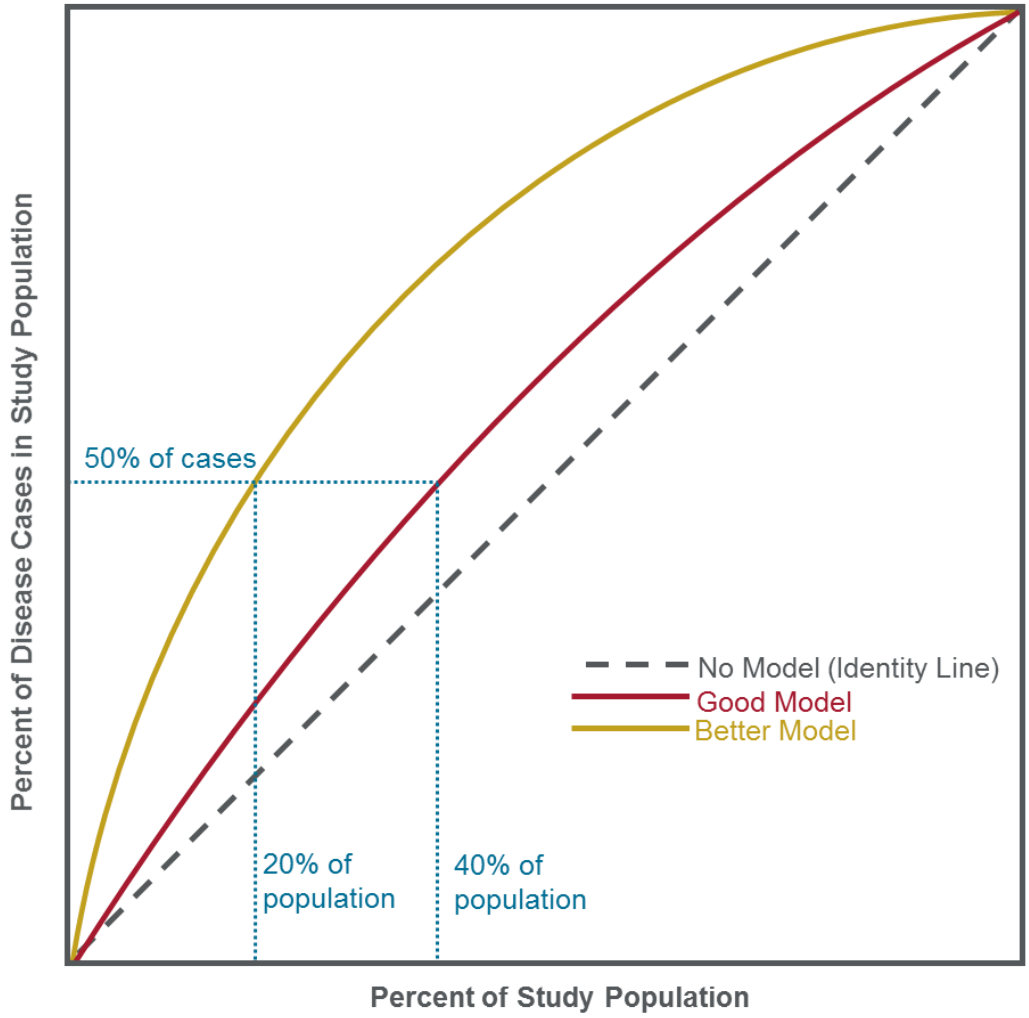

Figure 1.1. An idealized cumulative gains lift chart with models of varying predictive utility.

In the context of health policy and healthcare delivery, it is not enough just to describe the population risk of developing diseases; we need to be able to identify those individuals most at risk in order to effect change in modifiable lifestyle behaviours and to

therefore reduce their risk of developing these diseases. By utilizing cumulative gains lift charts to identify sub-populations, there exists the future potential to optimize lifestyle and behaviour change interventions for these populations to be cost neutral. That is, the expected likelihood of being meaningfully influenced by an intervention designed to prevent a certain set of chronic diseases can be set above the cost of delivering the intervention.

Step 6. Iterative model building.

After creating what are intended to be the best-fitted logistic regression models, it is important to try to improve the lift from the best logistic regression model on the validation dataset by transforming variables (e.g., ln(x), x2, 1/x, x¬1*¬x2, categorizing naturally continuous variables) and including them as inputs until the models cannot be improved further. The extent to which the predictor variables should be transformed can be determined by comparing the lift of each logistic regression model to neural network models of varying flexibility. Because neural networks can be trained to approximate infinitely complex relationships between the predictor variables and the disease outcomes, they can serve as a proxy for the true underlying relationships (Putler & Krider, 2012). When the lift of a best-fitted logistic regression model approximates that of a flexible neural network, the conclusion is that important nonlinearities in the data have been addressed and that the model-building process is complete.

1.5. RESEARCH GAPS & AIMS

1.5.1. Research Gaps

Healthcare decision makers need to know how much and who to allocate primary and secondary prevention resources to in order to have the most benefit, both in terms of health outcomes and cost-savings. However, methods that allow both disease risk estimation and cost assignment among older Canadian adults rely on categorizing or dichotomizing naturally continuous predictor variables. Being able to effectively identify people at-risk for having heart disease may require discrimination between individuals

on each predictor variable at levels finer than 2- or 3-level categories, and would benefit from the flexibility offered by classification modelling techniques.

1.5.2. Research Aims

Against the aforementioned gaps and background, I aim to address the following two specific aims:

Aim #1: *To apply modelling techniques from the field of business analytics to predict the presence of heart disease among older Canadian adults based on demographic, lifestyle, and medical characteristics.*

The ability to predict the presence of heart disease based on these characteristics depends on participants who have different values for each characteristic being more or less likely to have heart disease. For example, if the proportion of heart disease cases in a given sample of older adults decreased significantly as their scores on a measure of physical activity increased, holding other characteristics constant, then we would conclude that physical activity is associated with heart disease. This process would be repeated simultaneously for all variables using variable- and model-selection techniques that account for potentially complex relationships to get the best estimate of the cumulative association between all predictor variables and the outcome of heart disease.

Aim #2: *To calculate the expected healthcare costs of older adults based on their predicted probabilities of having heart disease.*

Using predicted probabilities of heart disease generated from the best-fitting model built in the first aim, along with sex-, age-, and province-specific treatment costs from a publicly available resource, I will assign expected treatment costs (i.e., the product of predicted probability and treatment costs) for each older adult in the CLSA.

A common approach to assigning healthcare costs in model-based economic analyses of injury and disease is by the expected value method, whereby costs are

assigned proportionally to the model-predicted probability of the outcome (Fawcett & Thorpe, 2013; Rice, Hammitt, & Evans, 2010). Compared to assigning either all or none of the costs of an outcome to an individual, this approach trades wrong individual-level predictions – because in practice, costs are either incurred or not – for better population-level predictions, because the sum of individual probabilities reflects the true population-level costs. It also allows cost assignment to be more sensitive to small, but still potentially meaningful, changes in predicted probabilities.

1.6. SIGNIFICANCE AND IMPLICATIONS

1.6.1. Identifying at-risk older adults

Ideally, the fulfillment of these aims will allow healthcare decision makers to be able to identify individuals and groups of older adults in the community who are at-risk of developing heart disease, with a view towards intervening to mitigate their risk or controlling their symptoms. While risk factor research for heart disease is quite extensive (Wilson et al., 1998; Danesh et al., 2004; Lopez et al., 2006; Rodondi et al., 2012), the majority of past studies do not classify and rank individuals according to their risk, which is a key step towards delivering something akin to personalized medicine in a primary and secondary prevention context in the community.

The novel implementation of business analytics methods towards addressing public and population health concerns is in line with directives from the Public Health Agency of Canada to “mobilize multi-sectoral action on healthy living and the prevention of chronic disease and injuries” (Public Health Agency of Canada, 2013). This focus on interdisciplinary, multi-sectoral action positions the work of this thesis to be an important proof-of-concept that there is value in considering perspectives traditionally employed by business to target customers.

Recent advancements in technology provide a method for identifying and providing interventions to at-risk individuals at scale. Mobile health interventions that utilize smartphones can be implemented effectively to a wide audience, and are generally accepted by the public (Mitchell et al., 2017). Leveraging smartphone technology would allow both collection of information about modifiable (e.g., walking-

related physical activity) and non-modifiable (e.g., age, sex, location) risk factors and model-based identification of and intervention delivery to those with a high risk of developing chronic diseases on the same device.

1.6.2. Supporting decision-making for chronic disease prevention interventions

With government agencies and national organizations advocating for increasing the proportion of healthcare resources being dedicated towards community medicine (primary and secondary prevention) (Public Health Agency of Canada, 2013; The Conference Board of Canada, 2012), it would also be useful for healthcare decision makers to have the ability to consider the cost landscape of primary and secondary prevention interventions as they relate to specific chronic diseases. By calculating the expected healthcare costs for each participant using the best classification model under Aim #2, it could be possible to explore the scale at which different interventions could be implemented. It could also allow healthcare decision makers to potentially prevent wasteful resource use on people unlikely to experience much of a risk – and expected cost – reduction even if they adhere closely to the intervention. Finally, the cost data generated by Aim #2 could assist healthcare decision makers in advocating for the wider implementation of efficacious interventions that might not otherwise have sufficient support to be scaled up.

With future access to follow-up data from the CLSA and infrastructure development by healthcare providers, a logical extension of this modelling framework could be to identify relevant subpopulations based on their calculated risk of developing heart disease, summarize the modifiable risk factors that contribute most to their disease risk, and provide interventions accordingly. By knowing the healthcare costs expected to be incurred by each targeted subpopulation, healthcare decision makers would have justification for how many people it would be likely for a given intervention to be most cost-beneficial for.

1.6.3. Approachable, relevant methods for non-technical healthcare decision makers

The choice of lift charts as both a means of model selection and visualization is unique in the healthcare sector and will aid in knowledge translation because their output is both intuitive and actionable. That is to say, by plotting actual cases against the proportion of the population that those cases were selected from by the best-fitting model, it becomes relatively straightforward for non-technical personnel to identify meaningful thresholds for targeting individuals or groups for interventions based on healthcare targets (e.g., prevent a given number of cases of heart disease), available funding (e.g., only enough money to fund an intervention for 20% of the older Canadian adult population), or – by also incorporating disease cost information – both (e.g., what percentage of the older Canadian adult population should receive a given intervention such that healthcare outlay is cost-neutral?). It is the practical application of this business-oriented approach that is perhaps its greatest benefit.

Chapter 2. A Model of Health: Using Business Analytics to Identify Older Canadian Adults with Heart Disease

2.1. INTRODUCTION

Nearly 90% of older Canadian adults (defined as ≥ 65 years of age) are living with at least one chronic disease and 65% are living with at least two chronic diseases (Canadian Institute for Health Information, 2011a; Lopez, Mathers, Ezzati, Jamison, & Murray, 2006; Public Health Agency of Canada, 2010). Concurrently, healthcare spending on the treatment of chronic diseases is increasing at a rate that the Canadian Institute for Health Information deems unsustainable (Canadian Institute for Health Information, 2011b; Canadian Institute for Health Information, 2015). A recent investigation estimated that the economic burden of chronic disease in Canada is nearly $53 billion annually (Krueger, Krueger, & Koot, 2015). Given that most of the costs of chronic disease are borne in later life, the economic burden of chronic disease is predicted to increase substantially as the Canadian population ages (Aminzadeh & Dalziel, 2002; Katzmarzyk & Janssen, 2004; Thompson, Kuhle, Koepp, McCrady-Spitzer, & Levine, 2014; Wilson et al., 2009).

Inflammatory pathways have been described as a common cause of many chronic diseases (Pedersen, 2009), and a number of chronic diseases have been associated with risk factors that are known to increase systemic inflammation, such as physical inactivity, smoking, and poor diet. These chronic diseases include obesity, hyperlipidemia, type 2 diabetes, cardiovascular diseases (hypertension, coronary heart disease, heart failure, cerebral apoplexy, and intermittent claudication), colon cancer, breast cancer, dementia, and depression (Krueger, Koot, Rasali, Gustin, & Pennock, 2016; Menec, 2003; Pedersen, 2009). There is considerable evidence to suggest that disease incidence and associated healthcare costs could be substantially reduced by targeting modifiable risk factors, such as physical inactivity, among the older adult population (Krueger et al., 2015; Warburton, Nicol, & Bredin, 2006). Therefore, leveraging knowledge of modifiable risk factors to accurately identify older adults at risk

of chronic disease is a priority for research towards improving the health of older Canadians (Public Health Agency of Canada, 2013).

Advances in data acquisition, storage, and processing technology have enabled healthcare providers to measure and track numerous metrics of health in clinical populations, creating opportunities for data-driven diagnoses and treatments of disease (Van Poucke et al., 2016). For example, methods from the field of business analytics, such as fitting classification models with neural networks, have been used to identify cases of chronic disease using large datasets of clinical records (Er, Yumusak, & Temurtas, 2010). Interactions among population health determinants (e.g., lifestyle choices, built environment, genetics) can be very complex (Kindig & Stoddart, 2003); machine learning approaches such as neural network modelling are well-suited to identify non-linear interactions between demographic, lifestyle, and medical variables towards correctly classifying participants according to their disease status. It is important that these classification models are calibrated on a sample of older Canadian adults – something that has yet to be done – to generate reliable estimates of chronic disease in the older Canadian adult population at-large (Shah et al., 2018).

It is also important to understand the financial consequences of modifiable risk factors on chronic disease in order to be able to make an economic case for primary and secondary prevention efforts. Some research has already been done to understand the relationship between physical inactivity in adulthood and a range of associated diseases and their costs at the population level, but these associations have not been examined in older adults (Krueger, Krueger, & Koot, 2015). Furthermore, these associations have simplified naturally continuous data, such as the amount of physical activity a person achieves, into broad categories. This approach overlooks potentially meaningful incremental changes in an older adult's health-related behaviour that would not result in reclassification.

Against this background, the primary aim of this study was to demonstrate the application of a methodological framework from the field of business analytics to create a classification model for a single chronic disease, heart disease, among older Canadian adults that accounts for nonlinear relationships among demographic, lifestyle, and medical variables. This work uses the relationship between risk factors, both modifiable and nonmodifiable, and heart disease as an exemplar of how risk factors are related to

chronic diseases in general. Heart disease was chosen because there is already a large evidence base suggesting that causal relationships exist between modifiable risk factors and heart disease development, and because of the relatively high prevalence (22.6%) of heart disease in the older adult population (Lee et al., 2012; Blumenthal et al., 2005; Tikkanen, Gustafsson, & Ingelsson, 2018; Ramage-Morin et al., 2010). The secondary aims of this study were to calculate the expected healthcare costs of each decile of older Canadian adults based on their predicted probabilities of having heart disease, and to compare these expected costs across deciles to identify the largest theoretically attainable cost savings versus random sampling.

2.2. METHODS

2.2.1. Study Population

The Canadian Longitudinal Study on Aging (CLSA) collected baseline data from a nationally representative sample of 51,338 Canadians aged 45-85 between 2010-2015, and will perform follow-up measurements on each participant every three years until 2033 or death (Raina et al., 2009; Kirkland & Wister, 2017). Some 30,097 participants comprise the 'comprehensive' group that undergoes direct physiological measurement (e.g., pulmonary function testing and electrocardiography) at data collection sites in Victoria, Vancouver, Surrey, Calgary, Winnipeg, Hamilton, Ottawa, Montréal, Sherbrooke, Halifax and St. John's. Another 21,241 participants comprise the 'tracking' group that completes telephone questionnaires designed to assess determinants of health, including health status, psychological wellbeing, and physical activity; the comprehensive group also completes these questionnaires. Baseline data for both groups was released at the end of 2016 and was acquired at the end of 2017 after an application to the CLSA's Data and Sample Access Committee was approved.

This work utilized data from both the comprehensive and tracking cohort, therefore only responses to the questionnaires completed by both groups was included in our analysis. There is precedent for published work that combines data from both the tracking and comprehensive cohorts (Stinchcombe, Wilson, Kortes-Miller, Chambers, & Weaver 2018).

For this study, we excluded CLSA participants who were younger than 65 years of age (n = 29847). We also excluded participants who had missing information for either the clinical outcome of interest (heart disease; n = 142) or any of the candidate predictor variables (n = 5750). The final analysis data set for this study included 15,599 participants.

2.2.2. Heart Disease

Heart disease, the outcome variable for the models developed in this study, was self-reported. Participants were asked to respond “Yes”, “No”, or “I don’t know” when asked whether a doctor had ever told them that they have heart disease (Raina et al., 2009). The CLSA study investigators ensured that all self-reported measures had been validated in previous work prior to including them in the study’s questionnaires (Raina et al., 2009).

2.2.3. Candidate Predictor Variables

Candidate predictor variables were taken directly from CLSA participants’ responses and, where relevant, were scored according to the CLSA protocol (Raina et al., 2009). The following set of candidate predictors required no scoring or modification before being included in the analysis: Age, Sex, Smoking Status, Wealth, Education, Nutritional Risk Score, Marital Status, Province, Retirement Status, Alcohol Consumption, 10-Item Center for Epidemiologic Studies Depression Scale Score, and Geographic Birth Region. Multimorbidity Classification was a derived variable, with a value of ‘Yes’ if a participant reported having two or more chronic diseases other than heart disease; the decision to exclude heart disease was made to prevent the outcome variable being included as a predictor variable in the models. Participants’ ‘Urban/Rural Classification’ was derived by coding the following responses as ‘Urban’: (i) ‘Urban Core’, (ii) ‘Urban Fringe’, (iii) ‘Urban population centre outside census metropolitan areas and census agglomerations’, (iv) ‘Secondary core’, and (v) ‘Postal code link to dissemination area’; there was only one code for ‘Rural’, which as ‘Rural’. Responses to

individual questions on the Physical Activity Scale for the Elderly (PASE) were not scored within the database, so it was necessary to convert participants' responses to a PASE score. The PASE has been shown to have acceptably high construct validity with measures of physical function, concurrent validity with accelerometry, and test-retest reliability (Washburn et al., 1999; Hagiwara, Ito, Sawai, & Kazuma, 2008). Following the PASE protocol, this involved taking the sum of the products of activity frequencies and activity-specific weights across the following categories of physical activity: household-related physical activity, work-related physical activity, walking, exercise to increase muscle strength, light-, moderate-, and vigorous- sports and recreational activities (Washburn et al., 1999). BMI was derived by dividing each participant's weight in kilograms by the square of their height in meters (Keys et al., 1972).

2.2.4. Statistical Analysis

Data were analyzed with R (version 3.0.1) using the Business and Customer Analytics package. Analysis was guided by the Rapid Model Development Framework, based on the Cross-Industry Standard Process for Data Mining model, which provides a set of methods to model binary outcomes based on complex relationships of many predictors (Putler & Krider, 2012; Chapman et al., 2000). Models were developed by the following procedures, based on the Rapid Model Development Framework:

1) We reviewed known mechanisms and risk factors for heart disease to select the first set of candidate predictors. An understanding of the proposed etiology of this relationship provided a rational for the inclusion of predictor variables into our minimally-adjusted models and contextualized the further selection and mathematical transformations of these predictors according to the methods that follow. Grounding our analysis in the heart disease literature helped to guard against spurious correlations during the model-fitting process.

2) We split the study population into training (n = 7803) and validation (n = 7796) datasets by random sampling. All models (described in steps 3 and 4) were developed on the training dataset, and selection of the fully adjusted logistic regression model was determined by performance on the validation dataset. Splitting the dataset ensured that

the model identified as the best-fitting was the one that provided the most accurate estimates of heart disease prevalence in the validation dataset without modelling any random noise in the training dataset. Risk and cost estimates reported from the best-fitting model used the entire study population, making our conclusions more generalizable to the entire older Canadian adult population.

3) We used neural networks (16 predictor variables, including 29 dummy variables to account for multilevel factor variables, interacting with a decay factor of 0.3 to produce 283 weights across 1 hidden layer with 6 nodes) to examine whether important nonlinear relationships were present in the relationships between the set of predictors and heart disease, again using the training dataset. Because the relationships between predictor and outcome variables in natural systems are often non-linear, we hypothesized that achieving optimally-fitted models would require transformation (e.g., $\ln(x)$, x^2, $1/x$, x_1*x_2, categorizing naturally continuous variables) of some of the predictor variables included in the model. Neural networks can be trained to approximate infinitely complex relationships between the predictor variables and the disease outcomes, so they were used as a proxy for the true underlying relationships (Putler & Krider, 2012). The performance of all models was compared against a neural network model to understand whether, and how much, transformation of predictor variables was necessary to account for all, or at least most, nonlinearities.

4) We used exploratory data visualization techniques to inform the removal or transformation of predictor variables to several candidate logistic regression models, using the training dataset. First, plots of means were created for each predictor variable and heart disease to identify any non-linear (e.g., exponential, U-shaped) relationships; if such relationships were evident by visual inspection, they were tested for in successive logistic regression models and retained in subsequent analysis if they resulted in a reduced Akaike Information Criterion (AIC), a measure of model fit. Because the AIC penalizes the inclusion of additional predictor variables, it helps to prevent overfitting the model (i.e., noise unique to the study data but not representative of true etiological or explanatory relationships). AIC reductions of more than two units were considered important (Burnham & Anderson, 2004). Results from each plot of means were used to identify predictor variables on whose dimensions there were significantly different groups of participants. This informed the derivation of categorical variables with the mean of each level taken from plots of means which had significantly different groups of

participants. Finally, tree models were produced to investigate whether there were any important two-way interactions. Again, each new variable was kept in the analysis if it resulted in a reduced AIC.

5) We scored each new logistic regression model against the previous best logistic regression model and against the neural network model. The comparison of each model's predictive utility utilized cumulative gains lift charts, whereby the proportion of participants with a given condition was plotted against the total number of participants, in order (from high to low) of their predicted likelihood of having heart disease. This type of plot allows side-by-side comparison of multiple predictive models (e.g., logistic regression models with different sets of predictor variables) within a decision-making framework that shows the number of disease cases identified per amount of effort expended in identifying certain sub-populations. The more cases of heart disease that can be identified for the fewest number of participants, the higher the lift and the better the predictive utility.

Predictive utility is especially important if these models are to ultimately be useful to health care decision makers to use resources more efficiently, and to maximize the cost-effectiveness of delivering disease prevention programs. For example, there is irrefutable evidence that sufficient levels of physical activity reduce the risk of heart disease development (Bauman et al., 2016; Pedersen & Saltin, 2015; Blumenthal et al., 2005; Tikkanen et al., 2018). In the context of health policy and healthcare delivery, it is not enough to describe the population risk of developing diseases based on being physically inactive; we also need to be able to target those individuals most at risk with effective physical activity promotion interventions in order to effect behavior change and therefore reduce their risk of developing heart disease.

6) We repeated steps 3-5 iteratively to improve the lift of the current best logistic regression model on the validation dataset by transforming and removing variables one-by-one. The extent to which the predictor variables were transformed was determined by comparing the lift of each logistic regression model to the neural network model, and to each previous logistic regression model. The fully adjusted logistic regression model was selected when no further improvements in lift or reductions in AIC could be made.

2.2.5. Health Care Costs

To address the study's secondary aim, we used the Economic Burden of Illness in Canada (EBIC) tool, a disease-costing tool developed by the Public Health Agency of Canada (2014). The tool monitors direct (hospital, drug, and physician) and indirect (mortality) costs at the population level. Costs were attributed to heart disease based on International Classification of Diseases codes ICD-9/ICD-10 (Wigle, Mao, Wong, & Lane, 1991; Public Health Agency of Canada, 2014). The EBIC was first published in 1991, based on data collected in 1986, and has been updated with data from 1993, 1998, and 2005-2008. EBIC permits stratification of disease costs by component (hospital, drug, physician, and mortality), province, sex, and age category, and has been validated for determining the economic burden of physical inactivity (Katzmarzyk & Janssen, 2004; Kruger et al., 2015).

We accounted for differences in treatment costs of heart disease by age, sex, and province by using heart disease prevalence data among older Canadian adults from the 2008 Canadian Community Health Survey. The ratio of 2008 cost data and 2008 prevalence data from a representative sample of older Canadian adults provided a reasonable estimate of the cost of treating one older adult heart disease patient, which allowed us to estimate the expected value of heart disease treatment costs by age (65-74, 74+), sex, and province. To do this, we multiplied predicted probability of heart disease for each participant by their estimated treatment costs. Expected costs from 2008 were inflated to 2018 values for reporting purposes using the Bank of Canada's Inflation Calculator based on historical data from Statistics Canada's Consumer Price Index (Bank of Canada, 2018; Statistics Canada, 2014).

2.3. RESULTS

Participant characteristics are provided in Table 2.1. Among participants included in our sample, 48.9% were female, the mean age was 72.9 years old, BMI was 27.5 kg/m2, and 17.2% reported having heart disease. Many participants were considered to have multimorbidity, reporting an average of 4.6 chronic diseases other than heart disease. Participants reported accumulating 505.7 minutes per week of leisure-time

physical activity, of which 275.8 minutes came from walking and 161.1 minutes came from moderate-to-vigorous physical activities. They also reported sitting for 1347.9 minutes per week. Most participants (80.5%) were born in Canada.

The unadjusted logistic regression model for heart disease (2 levels: yes, no; AIC = 6767.8) contained a set of 16 predictors (Table 2.2) taken directly from participant responses in the CLSA with as few modifications as possible (see section 2.2.3). The lift of the neural network model containing the full set of unmodified candidate predictors was greater than the lift of the logistic regression model containing the same set of predictors (Figure 2.1), indicating better performance of the neural network model at classifying CLSA participants in the training sample according to their heart disease status.

We then followed the variable selection and modification procedures outlined in the Rapid Model Development Framework (see step 4 in section 2.2.4) to remove the variables that did not contribute to predicting heart disease status and modified the variables that appeared to have nonlinear associations with heart disease. Each time a variable was removed or modified, we created a new logistic regression model; the removal or modification of a variable was only retained in subsequent analysis if it caused the AIC to decrease. Excluding the unadjusted and fully adjusted logistic regression models, there were 23 intermediate models that improved on the previous model. The fully adjusted logistic regression model (AIC = 6541.8) for heart disease included a set of 15 variables (Table 2.2), many of which were modifications or mathematical transformations of the set of 16 predictors included in the unadjusted logistic regression model. The AIC for the fully adjusted logistic regression model was 226 units less than for the unadjusted logistic regression model. The lift of the fully adjusted logistic regression model visually approximated the lift of the neural network model on the training sample (Figure 2.2), indicating that most nonlinearities were accounted for. On the validation sample, the lift of the fully adjusted logistic regression model exceeded the lift of the neural network model (Figure 2.3), which was evidence that the improved lift of the neural network model over the fully adjusted logistic regression model on the training sample was due to the neural network model being fit to noise, and that all true underlying relationships were captured in the fully adjusted logistic regression model.

In the fully adjusted logistic regression model, increasing age, male sex, more chronic conditions, wealth not meeting and barely meeting needs, living in Quebec, New Brunswick, Nova Scotia or Manitoba compared to other provinces, being retired or partly retired, increasing 10-Item Centre for Epidemiological Studies Depression Scale score, having a geographic birth region other than Africa, increasing weekly sitting time, and increasing BMI were associated with an increased probability of having heart disease. Increasing alcohol consumption frequency, an increasing number of household-related physical activities, and increasing weekly leisure-time physical activity were associated with a decreased probability of having heart disease.

Lift above the proportion of heart disease cases that would be identified by random sampling (as shown by the line of identity in Figure 2.4, representing the number of disease cases that would be expected to be identified by random sampling) was greatest among the 40% of study participants with the highest predicted probabilities of heart disease under the fully adjusted logistic regression model. At this point of maximum lift, the fully adjusted logistic regression model was able to correctly identify 65.8% of all heart disease cases in the study population, representing a 25.8% improvement over the number of heart disease cases that would be identified by random sampling. Since this is where the fully adjusted logistic regression model performs best, it will be used as a reference point in further comparisons to other models. Compared to the unadjusted logistic regression and neural network models, respectively, containing the full set of unmodified candidate predictors, this represents 5.6% and 3.5% reductions in the number of participants that would need to be targeted in order to identify 65.8% of heart disease cases (Figure 2.5).

The expected costs related to the ongoing treatment of heart disease among the 40% of CLSA participants with the highest predicted probabilities of heart disease is $6,779,138. This is $2,665,857 greater than the expected costs of treating heart disease among a randomly-sampled 40% (Figure 2.6). The expected marginal cost of identifying the next decile is $918,690, which is less than the $1,028,320 expected marginal cost of a randomly-sampled decile, indicating that the largest improvement in potential cost savings is attained by using the fully adjusted logistic regression model to identify the 40% of participants with the highest predicted probabilities of heart disease.

2.4. DISCUSSION

Analyses conducted in this chapter have demonstrated that older adults at-risk of having heart disease could be identified by using a classification model created by using business analytics techniques. The potential cost-savings associated with health promotion interventions guided by this kind of model building and selection procedure has the potential to be substantially more than interventions delivered to the population at-large. This is because the expected value of heart disease treatment costs under the model-selected scenario described in the final paragraph of the results section is less than the expected value of heart disease treatment costs under random sampling.

Within the context of chronic disease risk prediction research, the analytical methods in this study were chosen to be coherent with a decision-making framework that could eventually be used to improve health at a population level. Compared to the Framingham Risk Score models, my approach differs in several important ways (Wilson, Castelli, Kanel, 1987). First, the goal of the Framingham Risk Score models was to identify and quantify risk factors for heart disease, whereas this study selected a set of known risk factors and modified them to obtain estimates of heart disease probability. Second, in contrast to the development of the Framingham Risk Score models, the data used for this study was collected as part of an omnibus survey that did not cater to heart disease specifically. Finally, the cohort that the Framingham Risk Score models were developed on had several waves of follow-up data to allow the observation of heart disease development, rather than a cross-sectional snapshot heart disease status, as is the case in this study. While the risk estimates obtained with the Framingham Risk Score models are more accurate than those obtained in this study, they were not designed to estimate risk in the older adult population (Rodondi et al., 2012). Longitudinal cohort studies like the CLSA that focus specifically for older adults are a necessary step towards personalized care for this vulnerable subpopulation.

Neural network-driven model selection has also been used in other healthcare settings. Er and colleagues (2010) trained a neural network on a database of patient records with 38 predictors to diagnose Chronic Obstructive Lung Disease. There are a few important differences between their work and this study: their objective was to create a classification neural network, which makes their beta coefficients uninterpretable and makes it difficult to prevent including spurious correlations related to disease

development; they used clinically-measured variables (e.g., blood protein markers) which are often costlier and more difficult to collect than the self-reported responses used in this study; and they used multi-layer neural networks which are capable of capturing a wider range of nonlinearities between included predictors and the outcome of interest. Using multi-layer neural networks as a proxy for the best-fitting model comes at a much higher computational cost requires testing exponentially more predictors than for a single-layer neural network. While the single-layer neural network in this study worked well as a proof-of-concept, analyses intended to guide real decision-making scenarios would benefit from neural networks including more than one layer.

The costs assigned to healthcare treatment for CLSA participants based on their predicted probabilities of heart disease were partly inspired by the work of Krueger et al. (2015). Both this study and their work obtained cost data from the EBIC, and both used the expected value method to assign costs. However, there are a few important differences between their work and this study: Krueger et al. estimated portion of chronic disease costs attributable to obesity, smoking, and physical inactivity; their work did not focus specifically on older adults; their work included indirect costs due to losses in economic productivity in addition to direct healthcare costs; this study allowed each risk factor to take the form (i.e., continuous, categorical, mathematical transformation) that created the best classification model, compared to the binary categorical risk factors used in their work (e.g., physically inactive vs. physically active; obese vs. not obese).

The model-fitting methods are a major strength of this study, allowing models to be optimized to fit specific subpopulations, or to inform optimal intervention delivery. For example, if health care decision makers had a budget for a physical activity promotion program sufficient to enroll 20% of all older Canadian men, the predictive model that achieves the best lift on the first 20% of all older men in the validation sample would optimize its potential to be cost-neutral. The model would provide healthcare decision-makers with a ranked list of individuals most likely to benefit from a given intervention, supporting their ability to identify participants. Using an additive, rather than binary, measure of multimorbidity in the fully adjusted logistic regression model was another strength of this study and is supported by recent insights in chronic disease research (Wister, Levasseur, Griffith, & Fyffe, 2015). Another strength of this study is that it used the CLSA, a large, nationally representative sample of older Canadian adults. The large study population allowed inclusion of multilevel factor and discrete continuous variables

that have previously been dichotomised or excluded from similar analyses (Krueger et al., 2015; Epstein, 1998).

This study has several limitations. First, this work used cross-sectional data. This does not allow the assumption of a causal association between predictor variables included in the fully adjusted logistic regression model and disease development. We mitigated this limitation by relying on causal relationships between the included predictors and heart disease reported in the literature (Pedersen, 2009; Pedersen & Saltin, 2015; Tikkanen et al., 2018). Second, time-to-disease information was not available from the baseline data analyzed within this study. When longitudinal data becomes available it will be possible for future studies to calculate hazard ratios, rather than odds ratios, by conducting Cox proportional hazards regression (Basu, Manning, & Mullahy, 2004). Third, we excluded participants who had missing information for either the clinical outcome of interest (heart disease; n = 142) or any of the candidate predictor variables (n = 5750), rather than imputing missing data. This choice was made because we retained a large sample (n = 15599) even after excluding these participants, and because the additional complexity introduced by imputing values for missing data across several predictor variables for each participant would not have added much value to the proof-of-concept nature of this work. However, it will be important to impute missing data for these participants in analyses of longitudinal data because heart disease prevalence was higher among those with missing data than those without (21.3% versus 17.2%). Participants with missing data were also older (74.5 versus 72.9 years old) and were more likely to be female (53.0% versus 48.9%). Fourth, the healthcare costs obtained via the EBIC were only available at the levels of province, sex, and age category; better estimates of cost could be attained by having access to the actual treatment costs incurred by each participant in the database. Finally, all data used in this study was self-reported and therefore subject to recall bias. However, agreement between self-reported and true values for both the outcome of heart disease and the included predictors has been shown to be quite good, so despite uncertainty around magnitude and direction of self-report error, it not likely to impact this study's analyses in a meaningful way (Raina et al., 2009; Kirkland & Wister, 2017).

2.5. CONCLUSION

This project demonstrates the predictive utility of a novel modelling procedure that considers both disease incidence and cost in estimating the relationship between risk factors and heart disease among older Canadian adults. With a more rigorous modelling approach and access to forthcoming follow-up data from the CLSA, the methods presented here could help health care decision makers to advocate for, and target, the delivery of a range of health promotion interventions to at-risk populations of older Canadian adults.

Table 2.1. Participant characteristics, reported as mean (standard deviation) or n (%)

Variable	Overall	Male	Female
Sex	15599 (100%)	7964 (51.1%)	7635 (48.9%)
Age, years	72.9 (5.7)	73.0 (5.6)	72.8 (5.7)
Heart Disease, yes	2686 (17.2%)	1737 (21.8%)	949 (12.4%)
Other Chronic Diseases[1], #	4.6 (2.9)	4.0 (2.5)	5.2 (3.1)
Body Mass Index, kg/m^2	27.5 (4.9)	27.5 (4.2)	27.4 (5.5)
Physical Activity Scale for the Elderly, score	114.0 (107.9)	123.9 (97.5)	103.8 (116.9)
Walking, minutes/week	275.8 (307.3)	293.0 (319.6)	257.8 (293.0)
Light Sports and Recreational Activity, minutes/week	68.8 (201.1)	80.3 (229.9)	56.8 (165.0)
Moderate to Vigorous Physical Activity, minutes/week	161.1 (289.3)	187.1 (325.0)	133.9 (243.7)
Total Leisure-Time Physical Activity, minutes/week	505.7 (505.3)	560.4 (547.2)	448.5 (450.4)
Sitting, minutes/week	1347.9 (501.1)	1343.1 (502.6)	1352.9 (499.5)
Housework-Related Physical Activities, #	3.1 (1.4)	3.2 (1.5)	2.9 (1.3)
10-Item Centre for Epidemiological Studies Depression Scale Score	4.9 (4.2)	4.3 (3.9)	5.6 (4.5)
Smoking Status			
Current	836 (5.4%)	403 (5.7%)	433 (5.1%)
Former	8058 (51.7%)	4742 (43.4%)	3316 (59.5%)
Never	6705 (43.0%)	2819 (50.9%)	3886 (35.4%)
Alcohol Consumption			
Often (>12 times/month)	4685 (30.0%)	2943 (22.8%)	1742 (37.0%)
Sometimes (1-12 times/month)	8892 (57.0%)	4132 (62.3%)	476 (51.9%)
Never	2022 (13.0%)	889 (13.8%)	1133 (11.2%)
Province of Residence			
British Columbia	2765 (17.7%)	1388 (17.4%)	1377 (18.0%)
Alberta	1507 (9.7%)	784 (9.8%)	723 (9.5%)
Saskatchewan	407 (2.6%)	211 (2.6%)	196 (2.6%)
Manitoba	1416 (9.1%)	701 (8.8%)	715 (9.4%)
Ontario	3513 (22.5%)	1768 (22.2%)	1745 (22.9%)
Quebec	2894 (18.6%)	1463 (18.4%)	1431 (18.7%)
Nova Scotia	1349 (8.7%)	708 (8.9%)	641 (8.4%)
Prince Edward Island	361 (2.3%)	194 (2.4%)	167 (2.2%)
New Brunswick	401 (2.6%)	212 (2.7%)	189 (2.5%)
Newfoundland and Labrador	986 (6.3%)	535 (6.7%)	451 (5.9%)
Marital Status			
Single	824 (5.3%)	335 (4.2%)	489 (6.1%)
Married	10065 (64.5%)	6309 (79.2%)	3756 (47.2%)
Divorced	1646 (10.6%)	499 (6.3%)	1147 (14.4%)
Separated	257 (1.7%)	130 (1.6%)	127 (1.6%)
Widowed	2807 (18.0%)	691 (8.7%)	2116 (26.6%)

Variable	Overall	Male	Female
Retirement Status			
Retired	12239 (78.5%)	5923 (82.7%)	6316 (74.4%)
Partly Retired	2044 (13.1%)	1315 (9.5%)	729 (16.5%)
Not Retired	1316 (8.4%)	726 (7.7%)	590 (9.5%)
Wealth Meeting Basic Needs			
Insufficient	227 (1.5%)	92 (1.8%)	135 (1.2%)
Barely Sufficient	813 (5.2%)	332 (6.3%)	481 (4.2%)
Sufficient	14559 (93.3%)	7540 (91.9%)	7019 (94.7%)
Geographic Birth Region			
Canada	12557 (80.5%)	6258 (82.5%)	6299 (78.6%)
North America	381 (2.4%)	202 (2.3%)	179 (2.5%)
South America	167 (1.1%)	101 (0.9%)	66 (1.3%)
Europe	2149 (13.8%)	1177 (12.7%)	972 (14.8%)
Asia	196 (1.3%)	138 (0.8%)	58 (1.7%)
Africa	97 (0.6%)	64 (0.4%)	33 (0.8%)
Oceania	52 (0.3%)	24 (0.4%)	28 (0.3%)

[1]Other Chronic Diseases: Osteoarthritis, rheumatoid arthritis, asthma, emphysema, chronic bronchitis, chronic obstructive pulmonary disease, hypertension, diabetes, peripheral vascular disease, myocardial infarction, transient ischemic attack, dementia, Alzheimer's disease, Parkinson's disease, multiple sclerosis, epilepsy, migraine headaches, intestinal or stomach ulcers, bowel disorders, urinary disorders, cataracts, glaucoma, macular degeneration, cancer, mood disorders, allergies, osteoporosis, under-active thyroid gland, over-active thyroid gland, and kidney disease

Table 2.2 Variables included in the unadjusted and fully adjusted logistic regression models for heart disease

Variable	Unadjusted Logistic Regression Model	Fully Adjusted Logistic Regression Model
Age	Discrete	Discrete
Sex	Male Female	Male Female
Multimorbidity Classification	Yes No	*Not Included*
Count of Chronic Conditions Not Including Heart Disease	*Not Included*	Discrete
Sex × Count of Chronic Conditions Not Including Heart Disease	*Not Included*	Discrete
Smoking Status	Current Former Never	*Not Included*
Wealth Meeting Basic Needs	Totally Inadequately Not Very Well With Some Difficulty Adequately Very Well	Insufficient Barely Sufficient Sufficient
Education Level	Less Than Secondary School Graduation Secondary School Graduation Some Post-Secondary Post-Secondary Degree or Diploma	*Not Included*
Nutritional Risk Score	Discrete	*Not Included*
Marital Status	Single Married Separated Divorced Widowed	*Not Included*
Province	British Columbia Alberta Saskatchewan Manitoba Ontario Quebec Nova Scotia Prince Edward Island New Brunswick Newfoundland and Labrador	British Columbia Alberta Saskatchewan Manitoba Ontario Quebec Nova Scotia Prince Edward Island New Brunswick Newfoundland and Labrador
Urban/Rural Classification	Urban Rural	*Not Included*

Variable	Unadjusted Logistic Regression Model	Fully Adjusted Logistic Regression Model
Retirement Status	Retired Partly Retired Not Retired	Retired Partly Retired Not Retired
Alcohol Consumption Frequency	Almost Every Day 4-5 Times A Week 2-3 Times A Week Once A Week 2-3 Times A Month About Once A Month Less Than Once A Month Never	Often Sometimes Never
10-Item Centre for Epidemiological Studies Depression Scale Score	Discrete	0-4 Points 5-10 Points 11-30 Points
Geographic Birth Region	Canada North America South America Europe Asia Africa Oceania	Canada North America South America Europe Asia Africa Oceania
Physical Activity Scale for the Elderly Score	Discrete	*Not Included*
Count of Household-Related Physical Activity	*Not Included*	0-2 Activities 3-4 Activities 5-6 Activities
Natural Logarithm of Total Weekly Sitting	*Not Included*	Discrete
Total Weekly Leisure-Time Physical Activity	*Not Included*	Discrete
Square of Total Weekly Leisure-Time Physical Activity	*Not Included*	Discrete
Body Mass Index	Continuous	Continuous

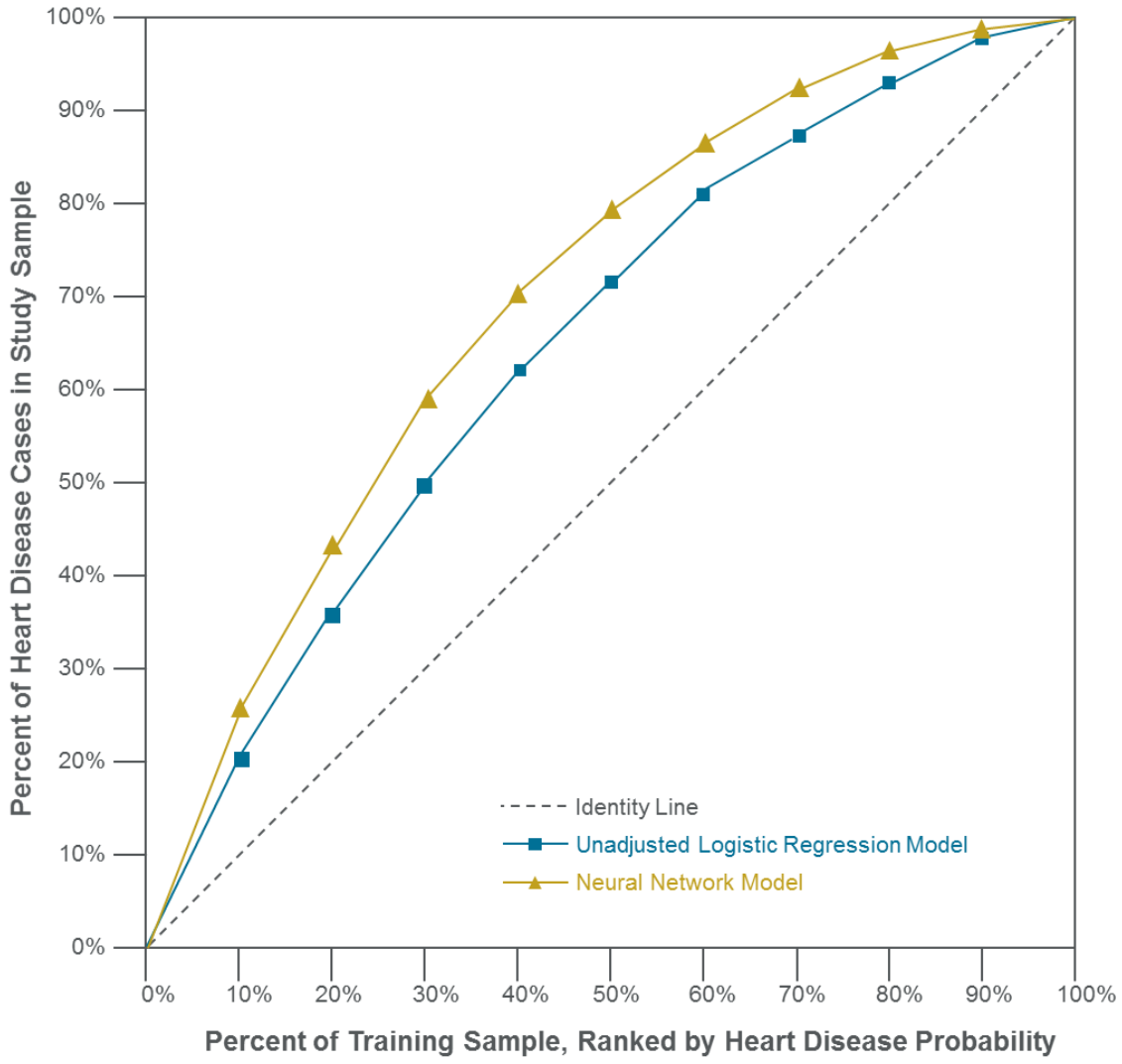

Figure 2.1. **Cumulative gains lift chart for logistic regression and neural network models containing the full set of unadjusted candidate predictors for heart disease among the training sample of CLSA participants**

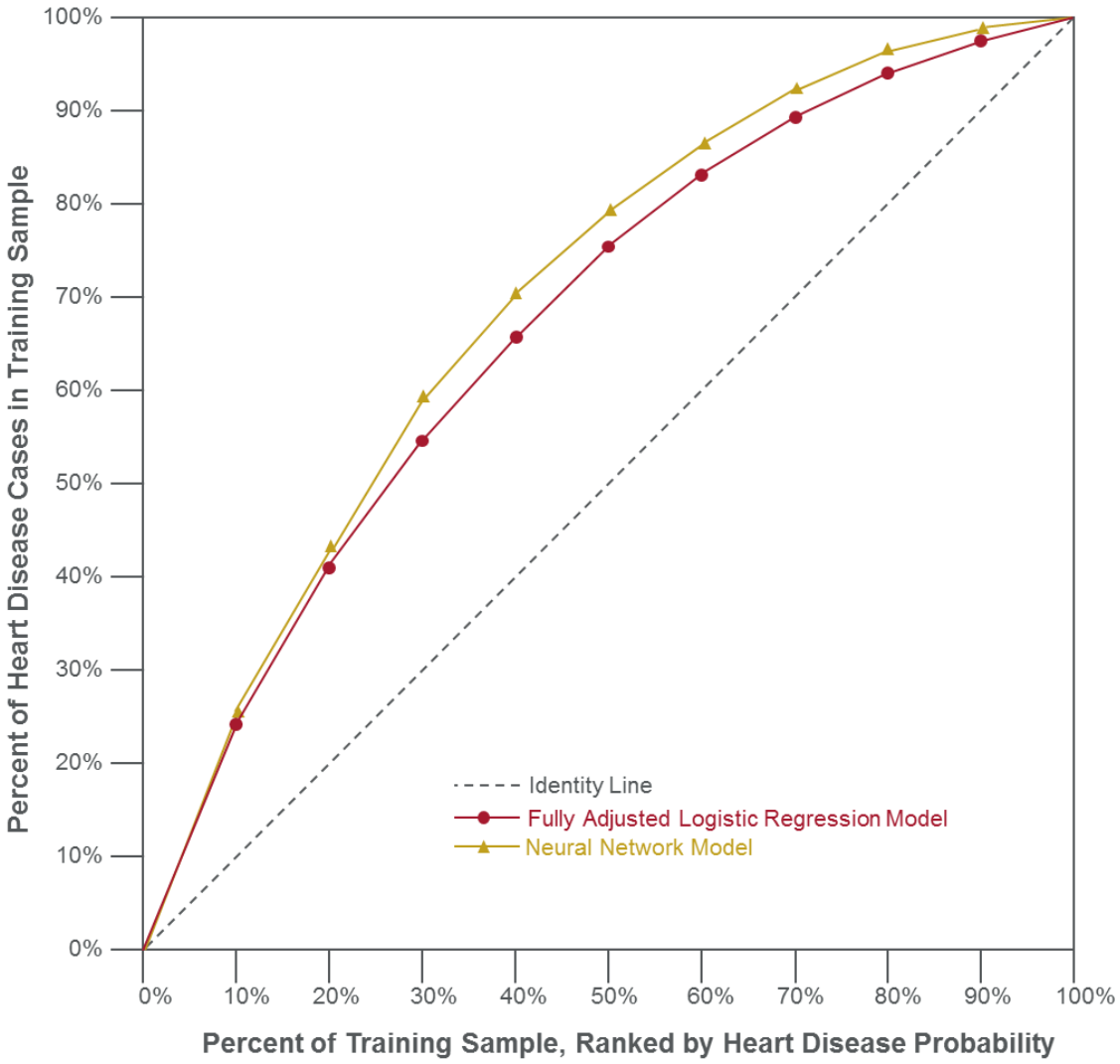

Figure 2.2. **Cumulative gains lift chart for the fully adjusted logistic regression model and original, unadjusted neural network model for heart disease among the training sample of CLSA participants**

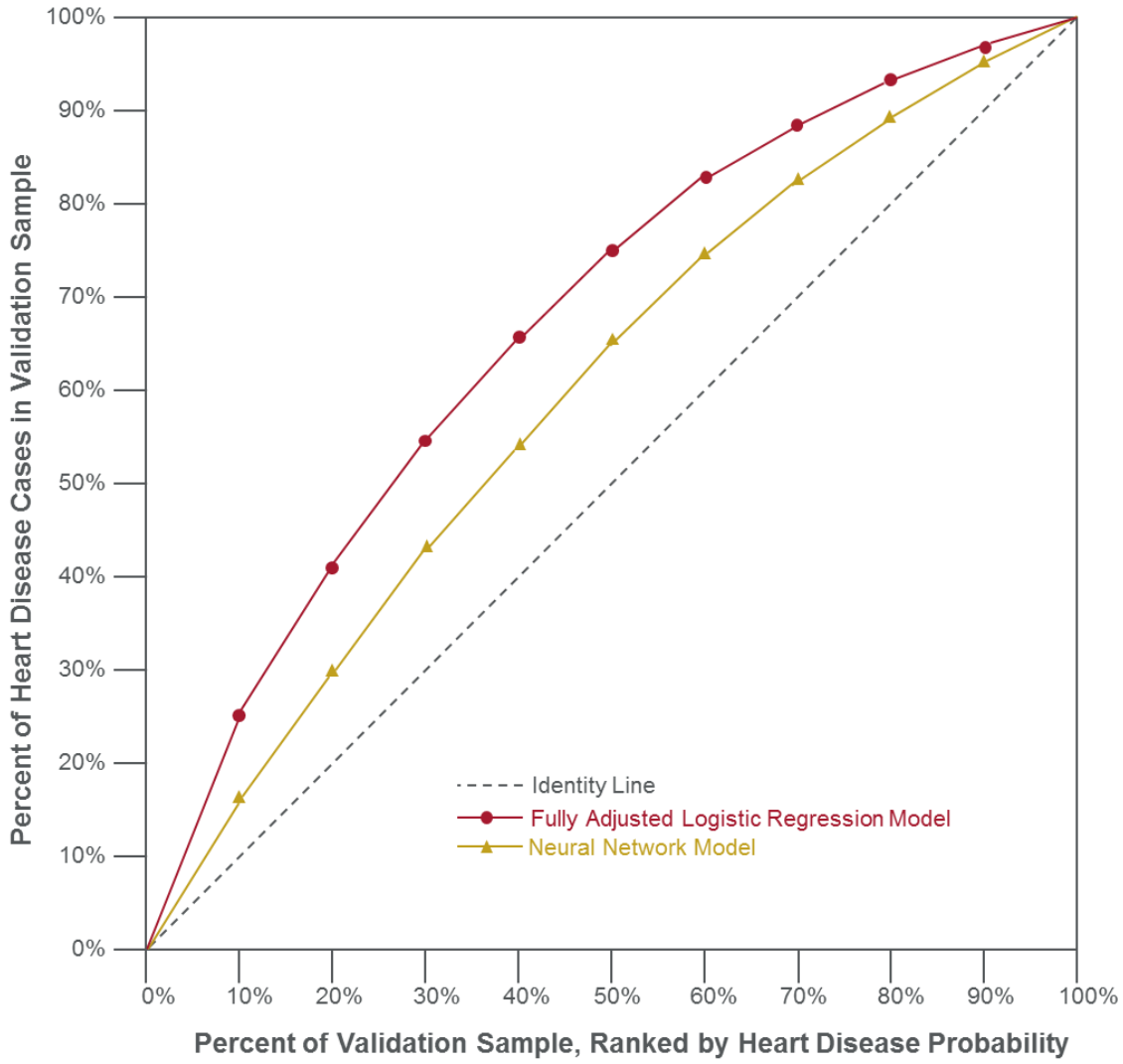

Figure 2.3. **Cumulative gains lift chart for the fully adjusted logistic regression model and original, unadjusted neural network model for heart disease among the validation sample of CLSA participants**

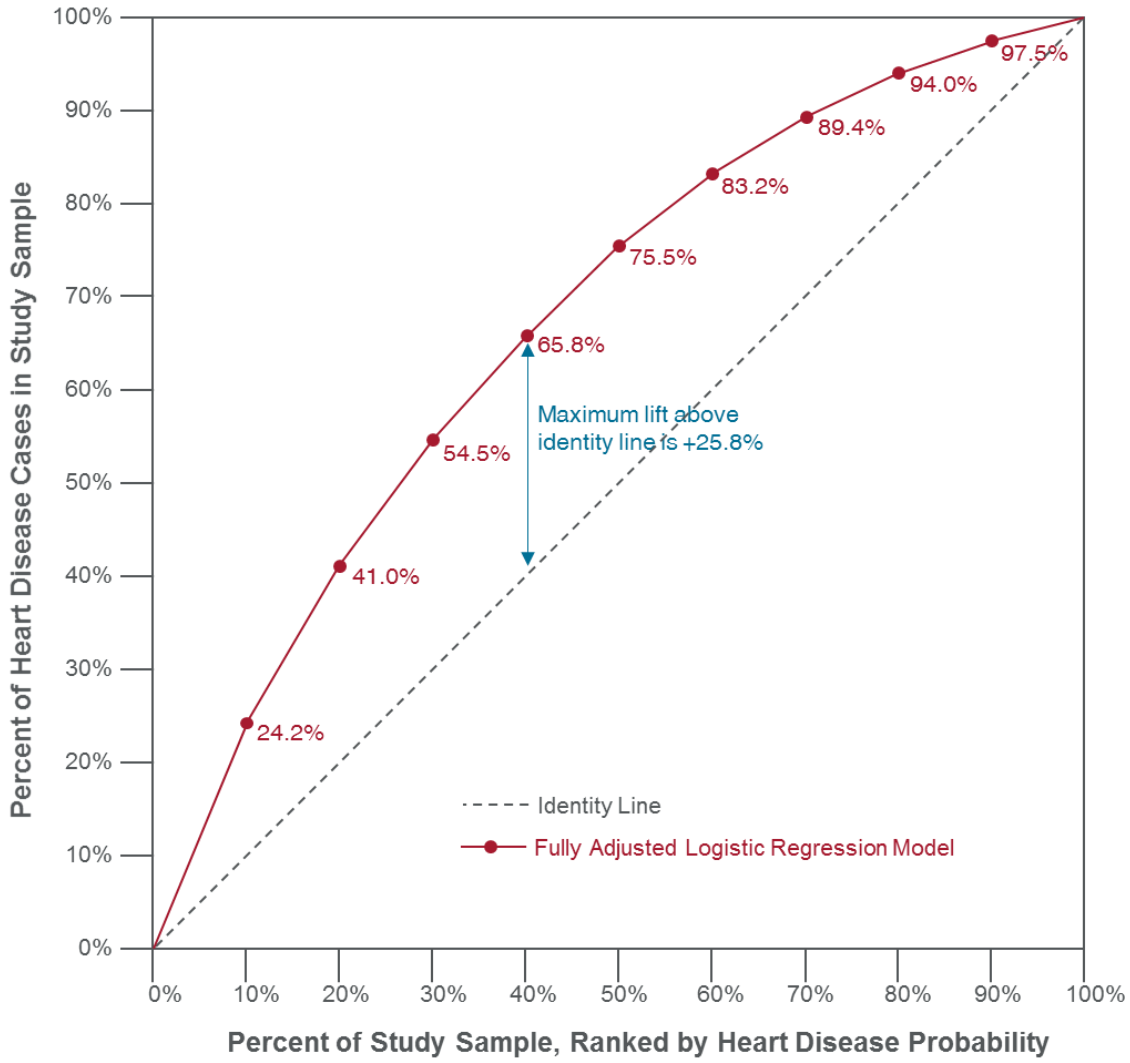

Figure 2.4. **Cumulative gains lift chart for the fully adjusted logistic regression model among all CLSA participants. The cumulative percentage of heart disease cases identified in the study population is shown for each decile, as well as the maximum lift above the cumulative percentage of heart disease cases that would be identified by random selection.**

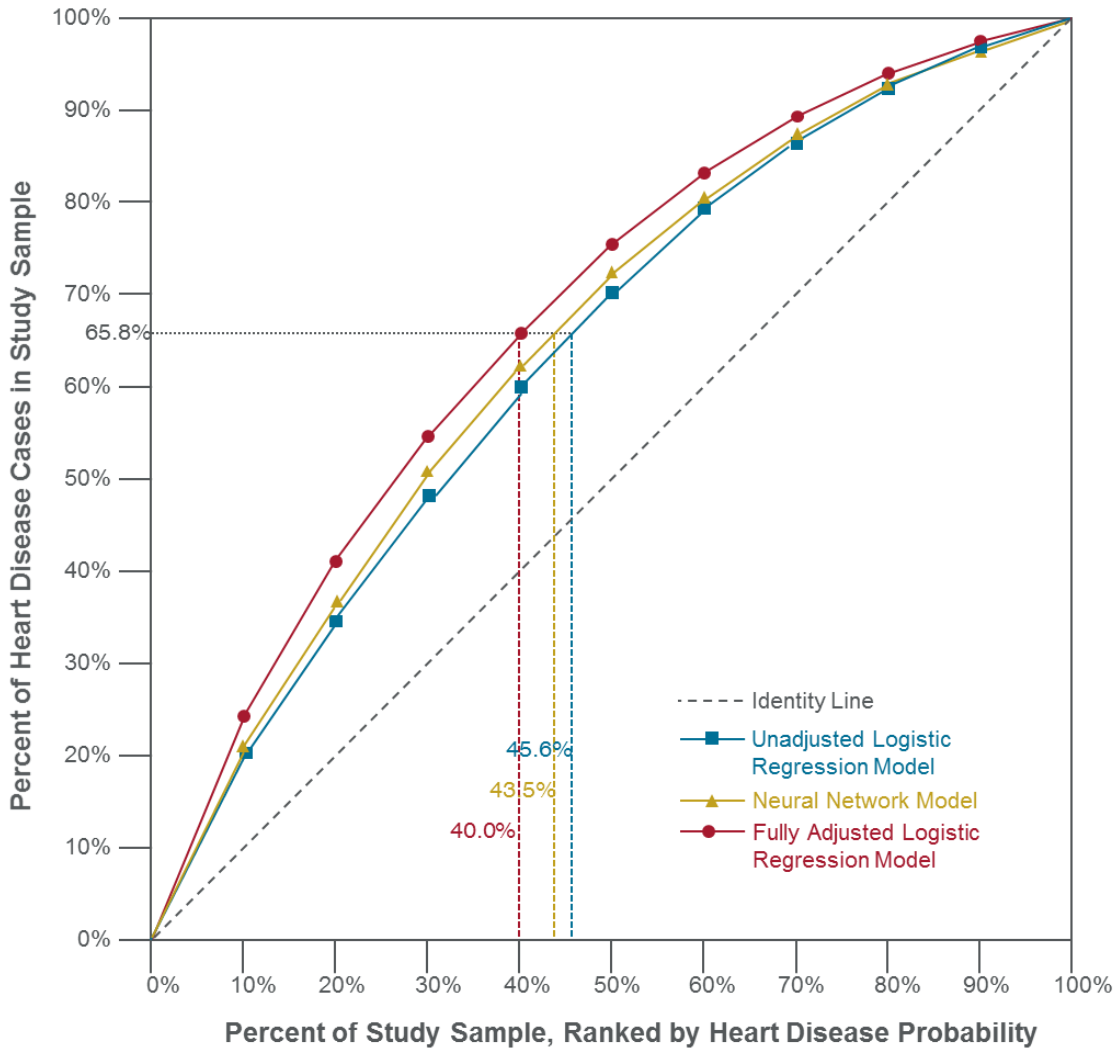

Figure 2.5. **Cumulative gains lift chart for the fully adjusted logistic regression, unadjusted logistic regression, and neural network models among all CLSA participants. The percentage of CLSA participants needed to identify 65.79% of heart disease cases is shown for each model.**

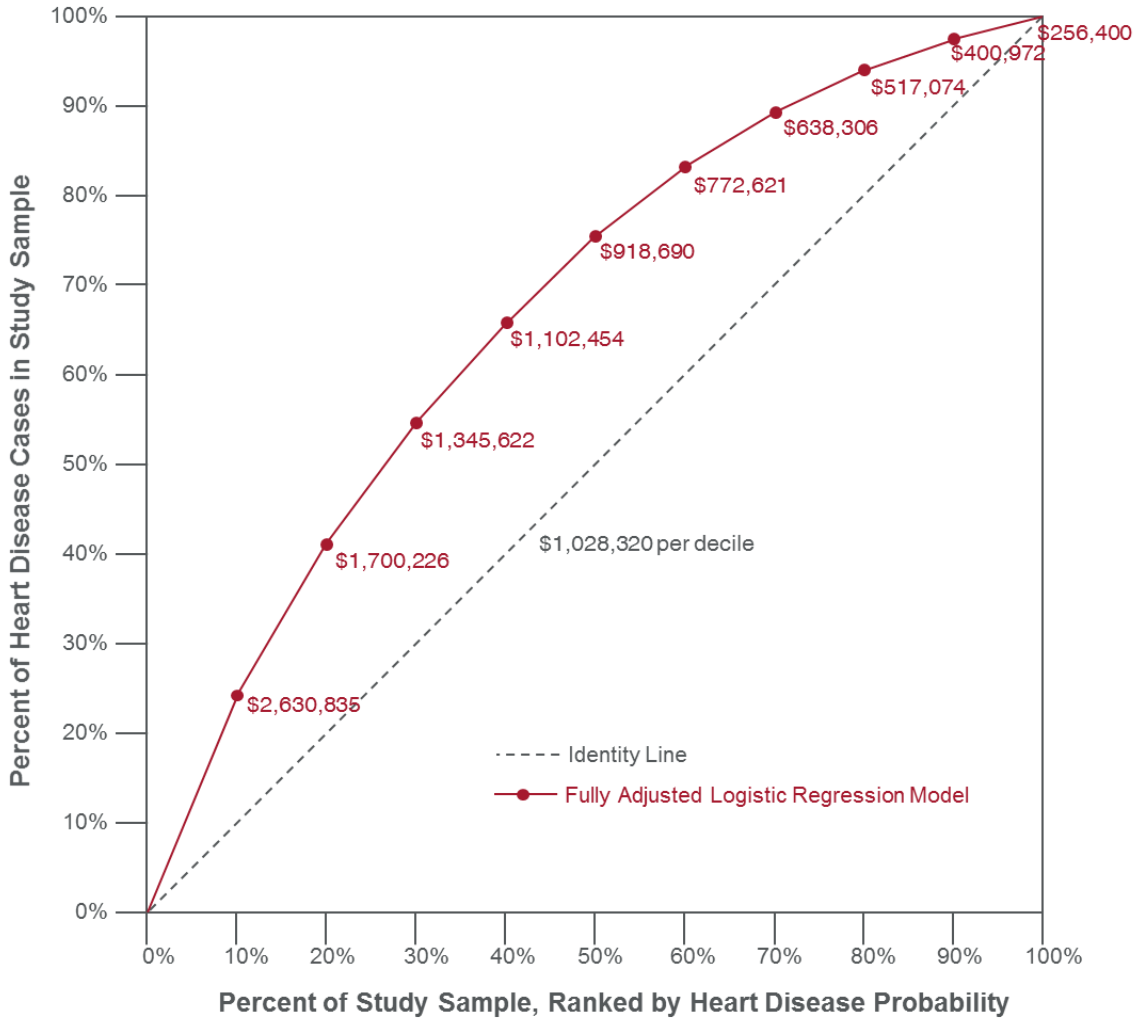

Figure 2.6. **Cumulative gains lift chart for the fully adjusted logistic regression model among all CLSA participants. The expected annual treatment costs related to heart disease for each decile are shown.**

Chapter 3. Extended Discussion

3.1. Results in context of the Canadian population

The chosen best-fitting logistic regression model was able to correctly identify 65.79% of heart disease cases from a selected subset of the study population that contained the 40% of participants with the highest predicted probabilities of having heart disease. At 40% of the study population, lift above the population-level heart disease prevalence of 17.22% was the greatest. Given that the study participants were sampled to be representative of all older Canadian adults, we can extend these conclusions to the Canadian older adult population. This means that using the model would allow us to identify 701,080 (6,195,544 x 17.2% x 65.79%) cases of heart disease if we selected the 2,478,218 older Canadian adults most likely to have heart disease, compared to only 426,253 cases by random sampling, a difference of 274,827 cases.

Likewise, the expected cost related to treating heart disease among the top 40% of study participants of $6,779,138 ($2,665,857 greater than the expected costs of a randomly-sampled 40%), translates to an expected cost for the 40% of all older Canadian adults most likely to have heart disease of $2,692,508,992 (6,779,138 x 6,195,544 / 15,599). This is $1,058,813,664 greater than the expected costs of a randomly-sampled 40%. The point at which lift above random sampling is maximized represents optimal model performance, so an improvement of one billion dollars in terms of expected healthcare cost identification is the largest potential reduction in healthcare cost versus random sampling attainable for a perfectly effective healthcare intervention.

3.2. Case Study

To understand how the methods demonstrated in Chapter 2 might be used to improve the delivery of health promotion interventions in the real world, consider the following hypothetical example. There is an intervention targeting modifiable risk factors for heart disease among an older adult population that has been shown to reduce each person's risk of heart disease by 10%, and it costs $100/person/year to deliver. From the

results of the healthcare cost assignment portion of this thesis work, the average cost of treating a heart disease patient from a population similar in demographic composition to the CLSA participants is $3804 per year.

Across the entire study population, such an intervention would require an investment of $1,559,900 and would be expected to save $1,028,320 in healthcare treatment costs, representing a loss of $531,580 to the healthcare system. However, by delivering the same intervention only to those most likely to develop heart disease, it is possible that such an intervention could be cost saving.

On the basis of individual predicted probabilities, it would be advantageous to deliver this intervention to deciles of participants who have a probability of developing heart disease that is greater than that of the i^{th} participant when participants are ranked according to their predicted probabilities of having heart disease, where ($3804*10%)(p(i))>=$100. In this context, the hypothetical i^{th} participant would have a probability of 26.3%, which would imply delivering the intervention to only the 20% of participants most likely to have heart disease (Figure 3.1), for an estimated cost saving of $113,106 (i.e., $433,106-$320,000). Future modelling efforts could involve grouping participants into smaller groups (e.g., 5% rather than deciles) to allow more precise recommendations of break-points; deciles were chosen here because they are commonly used (Putler & Krider, 2012). Linking predicted probability of disease and expected cost of treating disease allows the identification of important break-points for intervention delivery to part of the study population.

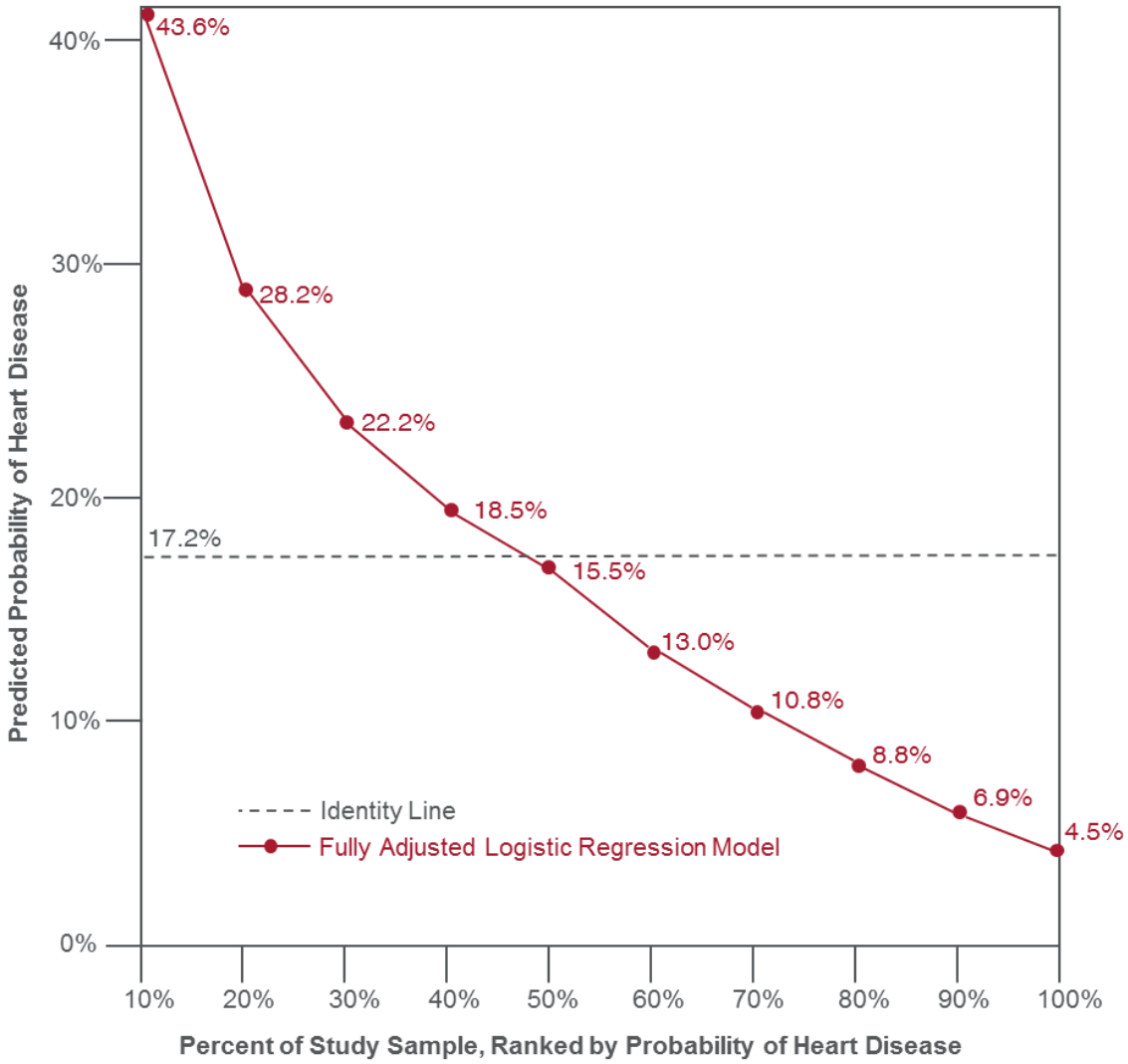

Figure 3.1. **Incremental response lift chart showing the predicted probability of each decile of the entire study population of CLSA participants ranked according to their predicted probabilities of having heart disease under the fully adjusted logistic regression model.**

3.3. Placing this work within the existing literature

Within the context of chronic disease risk prediction research, the chosen methods represent a novel application of classification modelling techniques to support decision making, and that could eventually be used to improve health at a population level. Compared to one of the most well-known risk-prediction models, the Framingham Risk Score (FRS) (Wilson, Castelli, Kanel, 1987), my approach differs in several

important ways. First, the FRS models were created from a cohort study, the Framingham Heart Study (FHS), that was specifically designed to identify and quantify risk factors for heart disease; while my work places an emphasis on obtaining the best estimates of heart disease probability given the data that I had access to. The CLSA dataset itself was collected as part of an omnibus-style investigation that did not consider heart disease specifically. The predictor variables in the FRS models are age, diabetes status, smoking status, systolic blood pressure, total cholesterol, high density lipoprotein concentration, low density lipoprotein concentration, and treatment for hypertension (Wilson et al., 1998). Of these, only age was included in this study's fully adjusted logistic regression model. Smoking was included in the unadjusted logistic regression model but was subsequently dropped, possibly due to its rarity (5.4%) or it being contraindicated for participants who have been diagnosed with heart disease. Diabetes status and hypertension were indirectly included in the fully adjusted logistic regression model as part of the multimorbidity measure. I did not have access to measurements of systolic blood pressure, cholesterol, lipoprotein concentration, or hypertension treatment within my study sample from the CLSA. So, while the types of data collected in the CLSA are not as tailored to heart disease prediction as those collected in the FHS, they also allow the business analytics-based modelling approach that I used to be easily reproduced for many other chronic disease outcomes. Second, the cohort that the Framingham Risk Score models were developed on had several waves of follow-up data to allow the observation of heart disease development, rather than a cross-sectional snapshot of who has heart disease versus who does not, as is the case in my work. This, taken with the heart disease-specific predictors that were measured, means that the risk estimates obtained with their models are much more accurate than mine. However, the FHS was not designed to allow risk estimation in the older adult population; the participants in the FHS were aged 30 to 74 with a mean age of 49 (Wilson et al., 1998). The Framingham Risk Score models have been adjusted to account for risk estimates in older adults, but they were not designed with this type of modification in mind (Rodondi et al., 2012). Clearly, longitudinal cohort studies like the CLSA that focus specifically on older adults are a necessary step towards personalized care for this vulnerable subpopulation.

The chosen method of neural network-driven model selection has also been used in other healthcare settings. In the Er et al (2010) study, researchers trained a

neural network on a database of patient records with 38 predictors to create a classification model that was able to accurately diagnose Chronic Obstructive Lung Disease. There are a few important differences between their work and mine; firstly, their objective was to create the best classification neural network, without requiring that mathematical transformations imposed on each predictor variable by the neural network model be knowable. This makes it very difficult to guard against spurious correlations that are not based on plausible mechanisms of disease development. Furthermore, they used clinically measured variables, many of which (e.g., blood protein markers) are much more invasive and more costly to collect than the self-reported responses used in my work. Finally, and most-importantly, they used multi-layer neural networks capable of capturing a wider range of nonlinearities between their included predictors and their outcome of interest. Whereas the single-hidden-layer neural network I used as a proxy for the best-fitting model was able to account for simple mathematical transformations (e.g. ln(x)) and interactions between two variables (e.g., age*sex), a multi-layer network could identify complex transformations (e.g. ln(1/x2)) and multi-variable interactions (e.g., age*sex*multimorbidity or age2*sex). Using multi-layer neural networks as a proxy for the best-fitting model would come at a much higher computational cost and would require testing exponentially more predictors than for a single-layer neural network. While my approach showed that these business analytics methods could be used to assign heart disease risks and costs, analyses intended to guide real decision-making scenarios would benefit from neural networks including more than one layer, as they would likely produce more accurate predictions.

My work bridges the gap between computationally intensive big-data approaches to creating classification models and more traditional, straightforward estimates of disease risk, demonstrating how decision making can be supported by adopting components from both schools of thought.

The costs assigned to healthcare treatment for CLSA participants based on their predicted probabilities of having heart disease under my best-fitting model bear most resemblance to the work done by Krueger et al. (2015), who determined the portion of the economic burden of several chronic diseases that is attributable to specific modifiable risk factors (obesity, smoking, and physical inactivity) among Canadian adults. In contrast, my work provided a method for estimating the expected cost of chronic diseases for groups of older adults based on knowing information about their risk

factors, but did not estimate an attributable portion of the cost of chronic diseases to any specific risk factors. Compared to their work, my study focused specifically on older adults, allowing the resulting estimates to reflect the higher treatment costs that are incurred by this under-investigated subpopulation. Furthermore, my work only considered direct healthcare spending on physician-, drug-, and hospital-related costs, whereas Krueger and colleagues also considered indirect costs due to losses in economic productivity. Finally, my work did not designate *a priori* risk factors and also allowed each risk factor to take the form (i.e., continuous, categorical, mathematical transformation) that created the best classification model, compared to the binary categorical risk factors that they used (e.g., physically inactive vs. physically active; obese vs. not obese).

3.4. Strengths

My thesis research demonstrated that many older adults at risk of having heart disease could be identified by using a classification model created by using business analytics techniques. The potential cost-savings associated with health promotion interventions guided by this kind of model building and selection procedure is also substantially more than interventions delivered to the population at-large.

The model-fitting methods described in this thesis are a major strength. By leveraging techniques from the field of business analytics, models can be optimized to fit specific subpopulations of the older adult population, or for specific types and scales of interventions. For example, if health care decision makers had a physical activity promotion budget capable of targeting 20% of all older Canadian men, the predictive model that achieves the most lift on the first 20% of all older men in the validation sample would enable the identification of those most at risk of heart disease. This approach implies that there are differences in the way that physical inactivity affects risk among different subpopulations and among people with different levels of baseline risk. This allows models to be tailored to and to reflect the heterogeneity of the older adult population.

Another strength of this thesis is that its analysis is conducted on a large, nationally representative sample of older Canadian adults, making the results generalizable to the entire population of older Canadian adults. The large study population in the CLSA also allowed the inclusion of multilevel factor and discrete continuous variables that have previously been dichotomised or excluded altogether. This is important because demographic and lifestyle factors typically explain small but meaningful components of disease outcomes (Epstein, 1998). Furthermore, because certain disease outcomes are less common (e.g., stroke, cancer), the large sample would allow sufficient numbers of observations of each disease to be able to model their prevalence in sub-populations and the older Canadian adult population at-large.

3.5. Limitations

A major limitation of this work is that the cross-sectional data used here does not allow the assumption of a causal association between predictor variables included in the fully adjusted logistic regression model and disease development. For these methods to be used to identify older adults at risk of developing chronic diseases, rather than those who already have chronic diseases, the use of follow-up data that allows observation of disease development in previously disease-free participants is necessary. Instead, a causal association between the predictor and outcome variables in our model is assumed based on the causal relationship between physical activity and heart disease reported in the literature using both physiological and longitudinal observational evidence (Pedersen, 2009; Pedersen & Saltin, 2015; Tikkanen et al., 2018). Since plausible causal relationships have been proposed and observed, it is reasonable to assume that the predictors included in the fully adjusted logistic regression model are not just correlated with, but can actually be used to predict the incidence of coronary heart disease in the study population. Time-to-disease information is not available from the baseline data analyzed within this study. When longitudinal data becomes available it will be possible to calculate hazard ratios, rather than odds ratios, by conducting Cox proportional hazards regression (Basu et al., 2004). This has the benefit of clarifying the time-component of disease onset, allowing healthcare treatment costs to be discounted accordingly, thereby improving the external validity of the resulting models.

Cross-sectional limitations notwithstanding, the extent to which the best-fitting model could be used to generate externally valid predicted probabilities of disease development for different kinds of health promotion interventions (e.g., targeting physical activity) depends on the ranges of the predictor variables (e.g., PASE responses) in the CLSA population (King & Zeng, 2006). As the CLSA cohort ages, the distribution of participants' values for each predictor variable should widen, mitigating this limitation and allowing more extreme what-if analyses over time.

EBIC data was only available at the levels of province, sex, and age category. Better estimates of treatment cost could be obtained by having access to the actual costs incurred by each participant in the database. Those data could reveal, for example, that only sex differences contribute to significantly different costs, or that there are other predisposing factors that are stronger determinants of what a person's heart disease treatment cost would be if they developed heart disease.

A further limitation is that self-reported data might not be valid representations of true values. The outcome of heart disease and all of the predictors were self-reported, introducing a potentially considerable amount of error around the generated estimates. In particular, heart disease data were self-reported by response to a question that asked if participants had ever been told by a doctor that they had coronary heart disease. Without corroboration of this outcome by angiogram, we have to accept that there will be some unknown error about this outcome measure. Both overestimation and underestimation of heart disease prevalence due to self-report error would serve to reduce the effect size of the risk factors-heart disease relationship, unless the rate and direction of error somehow depended on self-reported data, which is unlikely. However, agreement between self-reported and actual heart disease status has been shown to be acceptable (Raina et al., 2009), so despite uncertainty around magnitude and direction of this error, it is likely sufficiently small for the purposes of our analyses.

Likewise, each of the predictors was self-reported. For example, the physical activity data in the CLSA were collected by the PASE, a self-report questionnaire (Raina et al., 2009; Kirkland & Wister, 2017). While the PASE has been validated for estimating older adults' physical activity, it is subject to the same types of recall bias and floor effects as other self-reported measures of physical activity (Washburn et al., 1999, Tudor-Locke & Myers, 2001). While the direction of the bias (i.e., overestimation or

underestimation) depends on the type of activity and the length of recall, people tend to overestimate their time spent being physically active (Washburn et al., 1999, Tudor-Locke & Myers, 2001). If this was the case with the PASE questionnaire data collected in the CLSA, there will be an underestimation of the effect of physical activity – a bias toward the null – toward the abatement of the identified chronic diseases. Therefore, any conclusions drawn from our analysis are likely to be conservative estimates of the true effect. Decisions made about how to quantify physical activity data from the PASE would likely have compounded this issue. To preserve as much information as possible from participants' responses, we calculated physical activity in minutes per week by first multiplying the lower bounds of frequency (in number of times per week) and duration (in number of minutes per session) responses, then multiplying the upper bounds of frequency and duration responses, and finally taking the mean of the range. An example calculation for a hypothetical participant is given in Figure 3.2 below. Given that participants reported participating in 161.1 minutes of weekly moderate-to-vigorous physical activity, and that nearly 85% of older Canadian adults are thought to participate in less than 150 minutes of weekly moderate-to-vigorous physical activity (Colley et al., 2011), it is very likely that physical activity data reported in this thesis is overestimated.

Table 3.1 Sample leisure-time physical activity volume calculation for one participant

Mode of Physical Activity	Reported Frequency	Reported Duration	Volume, Low	Volume, High
Walking	3-4 days/week	More than 4 hours	3 days/week * 4 hours = 12 hours/week	4 days/week * 6 hours = 24 hours/week
Light Sports & Recreation	1-2 days/week	30 minutes-1 hour	1 day/week * 30 minutes = 30 minutes/week	2 days/week * 1 hour = 2 hours/week
Moderate Sports & Recreation	1-2 days/week	0-30 minutes	1 day/week * 10 minutes = 10 minutes/week	2 days/week * 30 minutes = 1 hour/week
Vigorous Sports & Recreation	Never	-	-	-
Exercising to Increase Muscle Strength	5-7 days/week	0-30 minutes	5 days/week * 10 minutes = 50 minutes	7 days/week * 30 minutes = 210 minutes
Leisure-Time Physical Activity			810 minutes/week	1830 minutes/week
Total estimate			**1117.50 minutes per week**	

The limitation of self-reported data could be partially addressed by including some measurements (e.g., short physical performance battery, systolic blood pressure) that were obtained at the data collection sites among the comprehensive cohort (Raina et al., 2009). Given that this work utilized data from both the comprehensive and tracking cohorts, only the questionnaire data collected from both groups was included. Limiting our analysis to the comprehensive cohort would provide more, and better quality, potential predictors, but would come at the expense of reducing the study sample size.

To improve cost estimation, it would be helpful to be able to differentiate between mild, moderate, and severe heart disease, for example, given that greater treatment costs tend to be borne by individuals with more severe cases of disease. The CLSA does ask about medication use, so for diseases that are treated pharmaceutically, it might be possible to obtain more accurate cost estimates if participants with higher predicted likelihood of having a disease are also more likely to be medicated for that disease.

3.6. Areas for future research

The next steps for this project towards providing a valid estimate of disease risk and associated costs among older Canadian adults are (i) to consider chronic diseases other than heart disease, (ii) to utilize follow-up data from the CLSA to model the development (rather than cross-sectional prevalence) of disease, (iii) to use multi-layer neural networks to model more complex risk factor-disease relationships, (iv) to estimate the portion of disease risk and healthcare treatment costs that can be attributed to modifiable risk factors, (v) to utilize the analytic weights from the CLSA to improve the external validity of these results, and (vi) to work with health authorities to get better estimates of per-person costs of disease treatment. Of these, only the choice of the next chronic diseases to consider and the inclusion of analytic weights have yet to be discussed in detail.

To appropriately extrapolate these results to the Canadian population at-large, it will be important to consider using sample weights, specifically the analytic weights for pooled data, to account for the probability of any given participant being included in the

study sample (Canadian Longitudinal Study on Aging, 2017). Given the numerous other limitations to this work identified above, the decision was made not to use the analytic weights in order to avoid giving the impression that the results as presently reported could be used to support healthcare decision-making. Rather, the contribution of this work is as a proof-of-concept for how these business analytics methods could be applied to chronic disease identification once follow-up data is obtained, stakeholders are involved, and the limitations – including the addition of analytic weights – are addressed.

In considering the next outcome to investigate, I observed the rapidly growing academic interest in mental health disorders, such as depression, and thought that placing my research within an emerging area would provide an insightful contrast compared to the well-established literature around heart disease. Furthermore, depression is often underdiagnosed among older adults compared to younger adults due to a reluctance to seek psychological support for negative feelings that are misattributed to the aging process, and due to the insensitivity of depression assessments for late-life depression (Mitchell, 2011; Mitchell, Rao, & Vaze, 2010). Therefore, a classification model that could identify individuals at risk for depression has the potential to improve the quality of life for people with, or at risk of developing, depressive symptoms but who have not sought or will not seek support themselves.

3.7. Significance

The Canadian healthcare system is under increasing strain from treatment costs related to preventable chronic diseases. Without corrective action, these problems are expected to increase dramatically over the coming decades. The methods demonstrated in my research allow the prediction of health and cost outcomes for older adults based on individual-level data and offer a potential method for optimizing the delivery of health promotion interventions. The flexibility of the proposed health and cost outcome models allows for demographic diversity, imputation of assumptions in regions where certain data are unavailable, and health authority-specific treatment costs. Therefore, they represent an initial step in assisting healthcare decision makers to understand how to optimize the delivery of primary and secondary prevention interventions, and to make a sound economic case for health promotion.

References

Akiyama, H., Barger, S., Barnum, S., Bradt, B., Bauer, J., Cole, G. M., ... Wyss-Coray, T. (2000). Inflammation and Alzheimer's disease. Neurobiology of Aging, 21(3), 383-421. https://doi.org/10.1016/S0197-4580(00)00124-X

Aminzadeh, F., & Dalziel, W. B. (2002). Older adults in the emergency department: A systematic review of patterns of use, adverse outcomes, and effectiveness of interventions. Annals of Emergency Medicine, 39(3), 238–247. https://doi.org/10.1067/mem.2002.121523

Bank of Canada. (2018). Inflation Calculator. Retrieved August 20, 2018, from https://www.bankofcanada.ca/rates/related/inflation-calculator/

Basu, A., Manning, W. G., & Mullahy, J. (2004). Comparing alternative models: Log vs Cox proportional hazard? Health Economics, 13(8), 749–765. https://doi.org/10.1002/hec.852

Bauman, A., Merom, D., Bull, F. C., Buchner, D. M., & Fiatarone Singh, M. A. (2016). Updating the Evidence for Physical Activity: Summative Reviews of the Epidemiological Evidence, Prevalence, and Interventions to Promote “active Aging.” Gerontologist, 56(June), S268–S280. https://doi.org/10.1093/geront/gnw031

Blumenthal, J., Sherwood, A., Babyak, M., Watkins, L., Waugh, R., Georgiades, A., . . . Hinderliter, A. (2005). Effects of Exercise and Stress Management Training on Markers of Cardiovascular Risk in Patients With Ischemic Heart Disease: A Randomized Controlled Trial. JAMA. 293(13), 1626–1634. https://doi:10.1001/jama.293.13.1626

Broekhuizen, K., Simmons, D., Devlieger, R., van Assche, A., Jans, G., Galjaard, S., … van Dongen, J. M. (2018). Cost-effectiveness of healthy eating and/or physical activity promotion in pregnant women at increased risk of gestational diabetes mellitus: Economic evaluation alongside the DALI study, a European multicenter randomized controlled trial. International Journal of Behavioral Nutrition and Physical Activity, 15(1), 1–12. https://doi.org/10.1186/s12966-018-0643-y

Bruunsgaard, H. (2005). Physical activity and modulation of systemic low-level inflammation. Journal of Leukocyte Biology, 78(4), 819–835. https://doi.org/10.1189/jlb.0505247

Buchman, A. S., Wilson, R. S., Yu, L., James, B. D., Boyle, P. A., & Bennett, D. A. (2014). Total daily activity declines more rapidly with increasing age in older adults. Archives of Gerontology and Geriatrics, 58(1), 74–79. https://doi.org/10.1016/j.archger.2013.08.001

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research. 33(2), 261–304. https://doi.org/10.1177/0049124104268644

Byford, S., Torgerson, D. J., & Raftery, J. (2000). Economic Note: Cost of illness studies. Bmj, 320(7245), 1335–1335. https://doi.org/10.1136/bmj.320.7245.1335

Canadian Institute for Health Information. (2011a). Health Care in Canada, 2011 A Focus on Seniors and Aging. CIHI, 162. https://doi.org/10.2165/00128415-200711570-00010

Canadian Institute for Health Information. (2011b). Health Care Cost Drivers: The Facts. https://doi.org/10.1016/j.healthpol.2009.04.001

Canadian Institute for Health Information. (2015). National Health Expenditure Trends, 1975 to 2015. https://doi.org/10.1007/s10916-010-9605-x

Canadian Longitudinal Study on Aging. (2017, July 11). CLSA Technical Document: Sampling and Computation of Response Rates and Sample Weights for the Tracking (Telephone Interview) Participants and Comprehensive Participants. Hamilton: Canadian Longitudinal Study on Aging. Retrieved from https://www.clsa-elcv.ca/doc/1041

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm 1.0. CRISP-DM Consortium, 76. https://doi.org/10.1109/ICETET.2008.239

Colley, R., Garriguet, D., Janssen, I., Craig, C. L., Clarke, J., & Tremblay, M. (2011). Physical activity of Canadian children and youth: Accelerometer results from the 2007 to 2009 Canadian Health Measures Survey. Health Reports, 22(1), 14–23.

The Conference Board of Canada. (2012, February). How Canada Performs: Health. Ottawa: The Conference Board of Canada. Retrieved from https://www.conferenceboard.ca/hcp/details/health.aspx

Danesh, J., Wheeler, J. G., Hirschfield, G. M., Eda, S., Eiriksdottir, G., Rumley, A., … Gudnason, V. (2004). C-Reactive Protein and Other Circulating Markers of Inflammation in the Prediction of Coronary Heart Disease. New England Journal of Medicine, 350(14), 1387–1397. https://doi.org/10.1056/NEJMoa032804

Elbaz, A., Sabia, S., Brunner, E., Shipley, M., Marmot, M., Kivimaki, M., & Singh-Manoux, A. (2013). Association of walking speed in late midlife with mortality: Results from the Whitehall II cohort study. Age, 35(3), 943–952. https://doi.org/10.1007/s11357-012-9387-9

Epstein, L. H. (1998). Integrating theoretical approaches to promote physical activity. American Journal of Preventive Medicine, 15(4), 257–265. https://doi.org/10.1016/S0749-3797(98)00083-X

Er, O., Yumusak, N., & Temurtas, F. (2010). Chest diseases diagnosis using artificial neural networks. Expert Systems with Applications, 37(12), 7648–7655. https://doi.org/10.1016/j.eswa.2010.04.078

Fawcett, L., & Thorpe, N. (2013). Mobile safety cameras: Estimating casualty reductions and the demand for secondary healthcare. Journal of Applied Statistics, 40(11), 2385–2406. https://doi.org/10.1080/02664763.2013.817547

Hagiwara, A., Ito, N., Sawai, K., & Kazuma, K. (2008). Validity and reliability of the Physical Activity Scale for the Elderly (PASE) in Japanese elderly people. Geriatrics and Gerontology International, 8(3), 143–151. https://doi.org/10.1111/j.1447-0594.2008.00463.x

Hjelmgren, J., Berggren, F., & Andersson, F. (2001). Health economic guidelines--similarities, differences and some implications. Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research, 4(3), 225–250. https://doi.org/10.1046/j.1524-4733.2001.43040.x

Jacobs, J. C., Laporte, A., Van Houtven, C. H., & Coyte, P. C. (2014). Caregiving intensity and retirement status in Canada. Social Science and Medicine. https://doi.org/10.1016/j.socscimed.2013.11.051

Katzmarzyk, P. T., & 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. https://doi.org/10.1139/h04-008

Keys, A., Fidanza, F., Karvonen, M. J., Kimura, N., & Taylor, H. L. (1972). Indices of relative weight and obesity. Journal of Chronic Diseases, 25(6–7), 329–343. https://doi.org/10.1016/0021-9681(72)90027-6

Kindig, D. A., & Stoddart, G. (2003). What is population health? American Journal of Public Health, 93(3), 380-383. https://doi.org/10.2105/AJPH.93.3.380

King, G., & Zeng, L. (2006). The Dangers of Extreme Counterfactuals. Political Analysis, 14(2), 131-159. https://doi.org/10.1093/pan/mpj004

Kirkland, S., & Wister, A. (2017). The Canadian Longitudinal Study on Aging (CLSA): A Platform For Research On Aging. Innovation in Aging, 1, 731-732. https://doi.org/10.1093/geroni/igx004.2640

Krueger, H., Koot, J. M., Rasali, D. P., Gustin, S. E., & Pennock, M. (2016). Regional variations in the economic burden attributable to excess weight, physical inactivity and tobacco smoking across British Columbia. Health promotion and chronic disease prevention in Canada: research, policy and practice, 36(4), 76-86.

Krueger, H., Krueger, J., & Koot, J. (2015). Variation across Canada in the economic burden attributable to excess weight, tobacco smoking and physical inactivity. Canadian Journal of Public Health, 106(4), e171–e177. https://doi.org/10.17269/CJPH.106.4994

Kyriakis, J. M., & Avruch, J. (2001). Mammalian Mitogen-Activated Protein Kinase Signal Transduction Pathways Activated by Stress and Inflammation. Physiological Reviews, 81(2), 807–869. https://doi.org/10.1152/physrev.2001.81.2.807

Lee, I. M., Shiroma, E. J., Lobelo, F., Puska, P., Blair, S. N., Katzmarzyk, P. T., … Wells, J. C. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. The Lancet, 380(9838), 219–229. https://doi.org/10.1016/S0140-6736(12)61031-9

Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., & Murray, C. J. L. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. The Lancet, 367(9524), 1747–57. https://doi.org/10.1016/S0140-6736(06)68770-9

MacNally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: the distinction between–and reconciliation of–'predictive'and 'explanatory'models. Biodiversity and Conservation, 9, 655–71. https://doi.org/10.1023/A:1008985925162

McMurdo, M. E. T., Roberts, H., Parker, S., Wyatt, N., May, H., Goodman, C., … Dyer, C. (2011). Improving recruitment of older people to research through good practice. Age and Ageing, 40(6), 659–665. https://doi.org/10.1093/ageing/afr115

Meisner, B. A., Linton, V., & Séguin, A. (2017). Examining Chronic Disease, Pain-Related Impairment, and Physical Activity among Middle-Aged and Older Adults in Canada: Implications for Current and Future Aging Populations. Topics in Geriatric Rehabilitation, 33(3), 182–192. https://doi.org/10.1097/TGR.0000000000000154

Menec, V. H. (2003). The Relation Between Everyday Activities and Successful Aging : A 6-Year Longitudinal Study. Social Sciences, 58(2), 74–82. https://doi.org/10.1093/geronb/58.2.S74

Mitchell, A. J. (2011). Why do physicians have difficulty diagnosing depression in the elderly? Aging Health, 7(1), 99–101. https://doi.org/10.2217/ahe.10.67

Mitchell, A. J., Rao, S., & Vaze, A. (2010). Do primary care physicians have particular difficulty identifying late-life depression? A meta-analysis stratified by age. Psychotherapy and Psychosomatics. https://doi.org/10.1159/000318295

Mitchell, M., White, L., Oh, P., Alter, D., Leahey, T., Kwan, M., & Faulkner, G. (2017). Uptake of an Incentive-Based mHealth App: Process Evaluation of the Carrot Rewards App. JMIR MHealth and UHealth, 5(5), e70. https://doi.org/10.2196/mhealth.7323

Murray, C. J. L., Barber, R. M., Foreman, K. J., Ozgoren, A. A., Abd-Allah, F., Abera, S. F., … Vos, T. (2015). Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990-2013: Quantifying the epidemiological transition. The Lancet, 386(10009), 2145–2191. https://doi.org/10.1016/S0140-6736(15)61340-X

Pedersen, B. K. (2009). The diseasome of physical inactivity - and the role of myokines in muscle-fat cross talk. Journal of Physiology, 587(23), 5559–5568. https://doi.org/10.1113/jphysiol.2009.179515

Pedersen, B. K., & Febbraio, M. A. (2012). Muscles, exercise and obesity: Skeletal muscle as a secretory organ. Nature Reviews Endocrinology. https://doi.org/10.1038/nrendo.2012.49

Pedersen, B. K., & Saltin, B. (2015). Exercise as medicine - Evidence for prescribing exercise as therapy in 26 different chronic diseases. Scandinavian Journal of Medicine and Science in Sports, 25, 1–72. https://doi.org/10.1111/sms.12581

Public Health Agency of Canada. (2010). The chief public health officer's report on the state of public health in Canada, 2010: Growing older – adding life to years. Ottawa: Government of Canada. Retrieved from http://www.phac-aspc.gc.ca/cphorsphc-respcacsp/2010/fr-rc/pdf/cpho_report_2010_e.pdf

Public Health Agency of Canada. (2013). Preventing Chronic Disease Strategic Plan 2013-2016.

Public Health Agency of Canada. (2014). Economic burden of illness in Canada, 2005-2008. Government of Canada, Ottawa, Canada. https://doi.org/HP50-1/2013E-PDF

Putler, D., & Krider, R. (2012). Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R. Boca Raton: Chapman & Hall/CRC Press.

Raina, P. S., Wolfson, C., Kirkland, S. A., Griffith, L. E., Oremus, M., Patterson, C., … Brazil, K. (2009). The Canadian Longitudinal Study on Aging (CLSA). Canadian Journal on Aging, 28(3), 221–229. https://doi.org/10.1017/S0714980809990055

Raina, P., Wolfson, C., Kirkland, S., & Griffith, L. (2018). The Canadian Longitudinal Study on Aging (CLSA) Report on Health and Aging in Canada: Findings from Baseline Data Collection 2010-2015.

Ramage-Morin P. L., Shields M., & Martel L. (2011). Health-promoting factors and good health among Canadians in mid-to late life. Statistics Canada, Catalogue no. 82-003-XPE. Health Reports. 21(3), 1-9.

Rejeski, W. J., Brubaker, P. H., Goff, D. C., Bearon, L. B., McClelland, J. W., Perri, M. G., & Ambrosius, W. T. (2011). Translating weight loss and physical activity programs into the community to preserve mobility in older, obese adults in poor cardiovascular health. Archives of Internal Medicine, 171(10), 880–6. https://doi.org/10.1001/archinternmed.2010.522

Rice, G. E., Hammitt, J. K., & Evans, J. S. (2010). A probabilistic characterization of the health benefits of reducing methyl mercury intake in the United States. Environmental Science and Technology, 44(13), 5216–5224. https://doi.org/10.1021/es903359u

Ridker P. M., Hennekens C. H., Buring J. E., & Rifai N. (2000). C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women. New Engl J Med., 342(12), 836-843.

Ridker, P. M., Cushman, M., Stampfer, M. J., Tracy, R. P., & Hennekens, C. H. (1997). Inflammation, Aspirin, and the Risk of Cardiovascular Disease in Apparently Healthy Men. New England Journal of Medicine, 336(14), 973–979. https://doi.org/10.1056/NEJM199704033361401

Rockhill, B., Newman, B., & Weinberg, C. (1998). Use and misuse of population attributable fractions. American Journal of Public Health, 88(1), 15–19. https://doi.org/10.2105/AJPH.88.1.15

Rodondi, N., Locatelli, I., Aujesky, D., Butler, J., Vittinghoff, E., Simonsick, E., … Bauer, D. (2012) Framingham Risk Score and Alternatives for Prediction of Coronary Heart Disease in Older Adults. PLoS ONE 7(3), e34287. https://doi.org/10.1371/journal.pone.0034287

Roux, L., Pratt, M., Tengs, T. O., Yore, M. M., Yanagawa, T. L., Van Den Bos, J., … Buchner, D. M. (2008). Cost Effectiveness of Community-Based Physical Activity Interventions. American Journal of Preventive Medicine, 35(6), 578–588. https://doi.org/10.1016/j.amepre.2008.06.040

Russell, L. B., Fryback, D. G., & Sonnenberg, F. A. (1999). Is the societal perspective in cost-effectiveness analysis useful for decision makers? The Joint Commission Journal on Quality Improvement. https://doi.org/10.1016/S1070-3241(16)30458-8

Shah, N. D., Steyerberg, E. W., & Kent, D. M. (2018). Big Data and Predictive Analytics: Recalibrating Expectations. JAMA. 320(1), 27–28. https://doi:10.1001/jama.2018.5602

Statistics Canada. (2010). Canada Year Book 2010. (pp. 313–326). https://doi.org/11-402-X

Statistics Canada. (2017a). Census Profile. 2016 Census. https://doi.org/98-316-X2016001

Statistics Canada. (2017b). Population size and growth in Canada: Key results from the 2016 Census. The Daily. https://doi.org/11-001-X

Statistics Canada (2014). The Canadian Consumer Price Index Reference Paper. (pp. 1–90). https://doi.org/Catalogue No. 62-553-X

Stinchcombe, A., Wilson, K., Kortes-Miller, K., Chambers, L., & Weaver, B. (2018). Physical and mental health inequalities among aging lesbian, gay, and bisexual Canadians: cross-sectional results from the Canadian Longitudinal Study on Aging (CLSA). Canadian Journal of Public Health. https://doi.org/10.17269/s41997-018-0100-3

Studenski, S., Perera, S., & Patel, K. (2011). Gait speed and survival in older adults. JAMA: The Journal of …, 305(1), 50–58. https://doi.org/10.1001/jama.2010.1923.Gait

Sullivan, S. D., Mauskopf, J. A., Augustovski, F., Jaime Caro, J., Lee, K. M., Minchin, M., … Shau, W. Y. (2014). Budget impact analysis - Principles of good practice: Report of the ISPOR 2012 budget impact analysis good practice II task force. Value in Health, 17(1), 5–14. https://doi.org/10.1016/j.jval.2013.08.2291

Thompson, W. G., Kuhle, C. L., Koepp, G. A., McCrady-Spitzer, S. K., & Levine, J. A. (2014). “Go4Life” exercise counseling, accelerometer feedback, and activity levels in older people. Archives of Gerontology and Geriatrics, 58(3), 314–319. https://doi.org/10.1016/j.archger.2014.01.004

Tikkanen, E., Gustafsson, S., & Ingelsson, E. (2018). Associations of Fitness, Physical Activity, Strength, and Genetic Risk With Cardiovascular Disease: Longitudinal Analyses in the UK Biobank Study. Circulation, 37(24), 2583-2591. https://doi.org/10.1161/CIRCULATIONAHA.117.032432

Toots, A., Rosendahl, E., Lundin-Olsson, L., Nordström, P., Gustafson, Y., & Littbrand, H. (2013). Usual Gait Speed Independently Predicts Mortality in Very Old People: A Population-Based Study. Journal of the American Medical Directors Association, 14(7), 529.e1-529.e6. https://doi.org/10.1016/j.jamda.2013.04.006

Tremblay, M., Wolfson, M., & Gorber, S. C. (2007). Canadian Health Measures Survey: rationale, background and overview. Health Reports / Statistics Canada, Canadian Centre for Health Information, 18 Suppl, 7–20. https://doi.org/10.1097/00005650-198008001-00003

Van Poucke, S., Zhang, Z., Schmitz, M., Vukicevic, M., Laenen, M. Vander, Celi, L. A., & De Deyne, C. (2016). Scalable predictive analysis in critically ill patients using a visual open data analysis platform. PLoS ONE, 11(1), 1–21. https://doi.org/10.1371/journal.pone.0145791

Verbeeten, D., Astles, P., & Prada, G. (2015). Understanding Health and Social Services for Seniors in Canada. Ottawa: The Conference Board of Canada.

Vlassoff, C. (2007). Gender differences in determinants and consequences of health and illness. Journal of Health, Population and Nutrition, 25(1), 47–61.

Warburton, D. E. R., Nicol, C. W., & Bredin, S. S. D. (2006). Health benefits of physical activity: the evidence. CMAJ: Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne, 174(6), 801–9. https://doi.org/10.1503/cmaj.051351

Washburn, R. A., Mcauley, E., Katula, J., Mihalko, S. L., & Boileau, R. A. (1999). The Physical Activity Scale for the Elderly (PASE): Evidence for Validity. Journal of Clinical Epidemiology, 52(7), 643-651.

Wellen, K. E., & Hotamisligil, G. S. (2005, May). Inflammation, stress, and diabetes. Journal of Clinical Investigation. https://doi.org/10.1172/JCI200525102

Wigle, D.T., Mao, Y., Wong, T., & Lane, R. (1991). Economic Burden of Illness in Canada, 1986. Chronic Dis Can, 12(Suppl 3). Accessed in August 2013, from http://publications.gc.ca/site/eng/448765/publication.html

Wilson, D. M., Truman, C. D., Thomas, R., Fainsinger, R., Kovacs-Burns, K., Froggatt, K., & Justice, C. (2009). The rapidly changing location of death in Canada, 1994-2004. Social Science and Medicine, 68(10), 1752–1758. https://doi.org/10.1016/j.socscimed.2009.03.006

Wilson, P. W. F., Castelli, W. P., & Kannel, W. B. (1987). Coronary risk prediction in adults (the Framingham Heart Study). American Journal of Cardiology, 59(14), 91–94. https://doi.org/10.1016/0002-9149(87)90165-2

Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837–1847. https://doi.org/10.1161/01.CIR.97.18.1837

Wister, A. V., Levasseur, M., Griffith, L. E., & Fyffe, I. (2015). Estimating multiple morbidity disease burden among older persons: a convergent construct validity study to discriminate among six chronic illness measures, CCHS 2008/09. BMC Geriatrics, 15, 12. doi:10.1186/s12877-015-0001-8

Yusuf, S., Reddy, S., Ounpuu, S., & Anand, S. (2001). Clinical Cardiology: New Frontiers Global Burden of Cardiovascular Diseases. Circulation, 104(22), 2746–2753. https://doi.org/10.1161/hc4601.099487