The Multidisciplinary Correlates of Chronic Stress in Canadians

by

Benjamin A. Hives

BKIN, The University of British Columbia, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Kinesiology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis/dissertation entitled:

The Multidisciplinary Correlates of Chronic Stress in Canadians

submitted by Benjamin A. Hives in partial fulfillment of the requirements for the degree of Master of Science in Kinesiology

Examining Committee:

Eli Puterman, Kinesiology
Supervisor

Mark Beauchamp, Kinesiology
Supervisory Committee Member

Yan Liu, Educational and Counselling Psychology, and Special Education
Supervisory Committee Member

Abstract

Background: Nearly one-quarter of Canadians report high levels of daily stress. This is alarming as chronic stress has been associated with several co-morbidities and premature mortality. In order to create beneficial interventions and public policy, factors associated with stress must be identified. While a wealth of research has determined a myriad of correlates of stress, the majority of this work has used approaches that focus on a very limited number of correlates per study, often from within one field of study. Currently there are no studies that analyze large-scale data sets and test multiple variables simultaneously.

Methods: This study analyzed data from the 2012 Canadian Community Health Survey - Mental Health, including 67 factors from a range of disciplines and over 23,000 participants. This study uses two approaches to test the associations between these factors and chronic stress including traditional statistics (i.e., simple linear regression and multiple linear regression) and machine learning algorithms (i.e., random forest analysis).

Results: The simple linear regression analysis showed that negative social interaction, life satisfaction, and higher levels of insomnia have the largest effect size in their association with chronic stress. Random forest analyses found that, after accounting for variance from other factors and considering complex interactions, life satisfaction, negative social interactions, and age were the most important correlates of chronic stress.

Conclusion: This study highlights that the important correlates of stress do not come from one field, but rather are a combination of psychological, social, and demographic factors. These novel findings highlight potential target pathways for devising new stress reduction interventions. However, as this study was exploratory and correlational, more research is needed regarding direction of effect and potential confounding variables.

Lay Summary

Chronic stress has proven detrimental to one's health. With almost one in four Canadians reporting high levels of stress, it is an important public health issue to identify the factors associated with stress. Using data from Statistics Canada, this study employed both traditional statistics and machine learning to determine the most important correlates of stress. This is the first study to explore stress with a large-scale dataset using machine learning. Both the traditional statistics model and machine learning highlighted the variables life satisfaction and negative social interactions as important correlates, though the results of the machine learning algorithm also noted age as a highly important correlate. More research is required to test causation; however, these results offer insight that may prove beneficial to improving stress reduction interventions.

Preface

All of the work presented in this thesis was conducted in the Fitness, Aging and Stress Laboratory at the University of British Columbia, Point Grey campus. The data presented come from the public use microdata file for the 2012 Canadian Community Health Survey - Mental Health. Access to this file was obtained through the Abacus data service, hosted by the University of British Columbia.

I was the lead investigator on this study. The Supervisory Author was Dr. Eli Puterman. With assistance from Dr. Puterman and my other committee members, Drs. Yan Liu and Mark Beauchamp, I was responsible for all major areas of concept formation, data analysis, and preparation of this manuscript.

The data presented in this manuscript are anticipated to produce one peer-reviewed publication. No publications have been submitted to date.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Term
ANOVA	Analysis of Variance
CCHS	Canadian Community Health Survey
CCHS-MH	Canadian Community Health Survey - Mental Health
CI	Confidence Interval
MSE	Mean Squared Error

Acknowledgements

Throughout the writing of this dissertation, I have received a great deal of support and assistance. I would first like to thank my supervisor, Dr. Puterman, whose expertise was invaluable in the formulation of the research topic and guiding my journey of learning over my Master's degree. I must also thank my committee (Drs. Beauchamp & Liu) whose feedback has not only strengthened this document, but helped develop my research skills.

Thank you to all of the lab mates who put up with my questions and ramblings, specifically Jean, Annie, Andrew, Luke, Adam, Renee, and Sarah.

Finally, I would like to thank my mother, Anji, for her love and support throughout all of my education.

Dedication

For Leah.

Chapter 1: Introduction

According to Statistics Canada (2014), nearly one-quarter of Canadians reported most days in their daily lives to be "quite a bit" or "extremely" stressful. This is alarming considering the serious health consequences of chronic psychological stress (Cohen, Janicki-Deverts, & Miller, 2007; Rod, Grønbaek, Schnohr, Prescott, & Kristensen, 2009; Schneiderman, Ironson, & Siegel, 2005). Chronic psychological stress is well-evidenced to contribute to poor health behaviours (Ng & Jeffery, 2003), as well as the development of psychiatric problems (Hammen, 2005) and chronic disease (Cohen et al., 2007). Seven More Years, a Canadian assessment of life expectancy conducted by Public Health Ontario, found those with high self-reported chronic stress had a lower life expectancy by approximately two years (Manuel et al., 2012). The deleterious effects of stress are recognized by several government-funded organizations that have incorporated stress reduction in their messaging. Health Canada (2008) has noted that stress is a risk factor for developing heart disease, mental illness, and bowel disease. The Heart and Stroke Foundation (2018) lists stress reduction as one of their five healthy lifestyle choices for reducing the risk of heart disease and stroke. The Canadian Cancer Society (2018) warns that stress may not just lead to physical symptoms, but also emotional symptoms and behavioral changes, such as irritability, restlessness, and poor eating and sleeping behaviours.

While the consequences of stress have been well-defined, the experience of chronic psychological stress is complex and has been linked to several factors, including genetic make-up (Kreek, Nielsen, Butelman, & LaForge, 2005), lifetime or recent exposure to serious or repeated daily stressors (Epel et al., 2018; Levine, Cole, Weir, & Crimmins, 2015), personality traits (Frost Ebstrup, Eplov, Pisinger, & Jørgensen, 2011), and engagement in a variety of health behaviours (Rod et al., 2009; Stults-Kolehmainen & Sinha, 2014). High levels of chronic

psychological stress are also more commonly reported by females compared to males (Statistics Canada, 2014), by younger compared to middle-aged adults (Statistics Canada, 2014), by ethnic minorities (Juster, McEwen, & Lupien, 2010), and by those with lower educations and incomes (Cohen & Janicki-Deverts, 2012).

Despite the multidisciplinary nature of stress research, previous studies have remained primarily intra-disciplinary and have relied on an *a priori* framework. Further, a majority of the analyses have relied exclusively on traditional statistical methods (e.g., analysis of variance (ANOVA), Pearson correlation, linear regression). This has led to two primary gaps in the literature. First, many potentially important correlates related to stress are being under-examined or not examined at all. Second, the aforementioned statistical methods are not suitable for complex interactions between correlates. Often, an individual's level of chronic psychological stress is not the product of one event, but rather the interaction between multiple behavioural, biological, psychological, and environmental factors (Epel et al., 2018). While teasing apart these factors improves a researcher's ability to quantify main effects, the many interactions between variables become impossible to measure. Machine learning, a new family of statistical modeling, has the potential to incorporate many variables and assess the complex interactions, as opposed to limiting analysis to predetermined, user-defined interactions. This not only allows for incorporating a greater number of variables from multiple fields that may have previously been ignored, but also the testing of complex interactions.

This thesis will serve to provide the reader with an overview of the literature on stress and its many correlates across disciplines. The concept of stress will then be described, including a brief history of the concept and the modern conceptualization of acute and chronic stress. The epidemiological health effects of stress will then be presented, referring to the current literature

on stress correlates and their methodological issues. Following this, a call for improved statistical methods in stress research will be presented. Finally, the findings of this study will be presented and discussed.

1.1 Psychological Stress

The most problematic element displayed in the field of stress research is the ambiguous use of the term "stress." The term has been used to describe the cause of the physiological changes (i.e., stressor or stress exposure), the immediate physiological changes (i.e., stress response), and the prolonged or repeated activations of the stress systems (i.e., chronic stress). Most commonly, psychological stress is defined as an individual's perception that demands in their environment exceed their capacity to manage these demands (Cohen, Kamarck, & Mermelstein, 1983; Lazarus & Folkman, 1984).

The general concept of stress has its roots in the field of biology with research by Hans Selye (1936). Selye (1950) demonstrated that stressors, consisting of physical threats (e.g., intense exercise, cold immersion) and emotional stimuli (e.g., rage), have reliable and consistent physiological effects, including increases in blood pressure, alterations in body temperature, and secretion of adrenocorticotropic hormones from the pituitary gland. Exposures to these stressors might perturb homeostasis (i.e., the balance between all physiological systems) in the short term and be non-damaging, or even situationally beneficial and adaptive. However, repeated or prolonged exposures can lead to a General Adaptation Syndrome, where physiological states are altered more permanently due to failure to adapt to the repeated threats (Selye, 1950). This model was further explored by Lazarus and Folkman (1984), McEwen and Stellar (1993), Smyth, Zawadzki, and Gerin (2013), and Epel and colleagues (2018) to examine the

intermediary psychological and biological steps between stressors and responses, as well as subsequent long-term effects on mental and physical health.

One of the most prominent extensions of Selye's stress model was described by Lazarus and Folkman (1984), whereby they attempted to unpack the stressor-stress response path by describing the intermediate psychological appraisals and emotional states inherent to the process. Building on the stressor-stress response model of Selye, Lazarus (1993) described the "Transactional Model of Stress and Coping," in which stress was not a stimulus nor response, but rather a relational concept. Like Selye's model of acute stress, the Lazarus model included stressor and physiological responses; however, Lazarus expanded this model to include the intermediary effects of appraisals, which take into account an individual's perceptions and offer a more nuanced analysis of the entire stress process. In the presence of a stressor, an individual would begin a subconscious appraisal process, which Lazarus and Folkman (1984, p. 31) described as "the process of categorizing an encounter, and its various facets, with respect to its significance for well-being." Appraisal can be further broken down into primary and secondary appraisals. During primary appraisals, the potential stressor is evaluated as irrelevant, benign/positive, or stressful. Stressors which are deemed stressful are further categorized as harm/loss (i.e., individual has already sustained loss or damage), threat (i.e., situation may lead to future harm or loss), or challenge (i.e., situation may lead to gain or personal growth, but requires resources to overcome). Following the primary appraisal process, the secondary appraisal is initiated. During secondary appraisal, an individual may employ coping strategies, that is, any cognitive or behaviour efforts, to manage their demands. Each appraisal outcome is associated with a unique set of emotional responses. A stressor that was found to be benign/positive during primary appraisal would elicit positive emotions, such as joy. Appraisals

of harm/loss and threat are characterized by negatively-valanced emotions (e.g., fear, anxiety, anger), whereas challenge is characterized by positively-valanced emotions (e.g., eagerness, excitement, exhilaration).

Another prominent extension of the Selye model of General Adaptation Syndrome is the Allostatic Load model by McEwen and Stellar (1993). Similar to Lazarus' model, McEwen and Stellar argued that, when faced with a psychological or physiological stressor that is appraised as a threat, there is a stress response. However, rather than the psychological response proposed by Lazarus, McEwen and Stellar noted the response primarily as a physiological one, in which an individual's autonomic nervous system (i.e., fight-or-flight system) and neuroendocrine system are activated. These systems begin an adaptive process known as allostasis. While in allostasis, the autonomic nervous system and neuroendocrine system release stress hormones in an attempt to restore homeostasis. While this adaptation is beneficial in the short-term, there is a cost. It is the repeated or prolonged activation of these systems that lead to wear-and-tear on the body, which plays a role in physiological damage and disease risk. This allostatic load model is perhaps best displayed through the paradoxical relationship between the adaptive effects of the acute stress response and harmful effects of repeated and/or prolonged activation of the stress systems (i.e., chronic stress). A review by McEwen (2004) found that during the acute stress response, the transient activation of the autonomic nervous system and neuroendocrine systems lead to situational benefits (e.g., increased blood pressure to increase blood flow to the limbs, facilitated movement of immune cells to a potential pathogen, improved situational memory), but the repeated or prolonged activation of these systems lead to damage (e.g., damaged coronary artery walls, suppressed immune function, impaired memory). This model represents one of the

most commonly suggested pathways between stress and disease (Epel et al., 2018; Smyth et al., 2013).

While McEwen investigated the pathway from acute stress to chronic stress to disease, Smyth and colleagues (2013) unraveled the transition from acute stress responses to chronic stress. Similar to Lazarus, Smyth and colleagues suggested that external or internal stimuli can be considered stressors, which, if appraised as a threat or challenge, can lead to a stress response. This process is labelled as acute stress. However, when the body undergoes a dysregulation of this response, chronic stress occurs. This dysregulation can occur through a variety of mechanisms (i.e., repeated activations, slow or no adaptations, the delay or absence of returning to homeostasis), which is similar to the conceptualization presented by McEwen. However, while McEwen considered these as physiological mechanisms, Smyth and colleagues note that these dysregulations can stem from psychological processes, such as worry and rumination. Repeated activations occur when there are a number of unrelated stressors (e.g., traffic on your way to work and missed deadlines), environmental contexts (e.g., familial caregiving, low socioeconomic status), or both. Slow or no adaptations may occur when there is a lack of habituation to a specific stressor. Delay or absence of homeostasis occurs when there is frequent or persistent stress, to which the body adapts by creating a higher activation baseline of stress systems. The authors highlighted that these various dysregulations can be influence by perseverative cognitions (e.g., rumination, perceived lack of control, imagining future stressors).

Epel and colleagues (2018) also noted that factors from many fields influenced both acute and chronic stress. They captured this idea in their Transdisciplinary Model of Stress, which incorporates physiological, psychological, social, behavioural, and cultural factors. While the acute stress process they provided was similar to that of McEwen, it expanded the chronic stress

model developed by Smyth and colleagues. Here individual context was included in the model through individual (e.g., genetic, developmental), environmental (e.g., sociocultural), and protective (e.g., sense of mastery) factors, as well as cumulative stress (e.g., history of traumatic stress, current stress states). These factors also influence an individual's habitual processes (i.e., the lens through which one sees the world). All of these play a role in allostatic load, which, in turn, leads to biological aging and early disease. Further, the acute stress response, habitual processes, and allostatic load are each influenced by health behaviors (e.g., physical activity). While this model is a comprehensive examination of the many factors that influence stress, it lacks a succinct definition of the term stress.

This thesis will use a theoretical definition of stress based on the work of Cohen, whose Stage Model of Stress and Disease incorporates stressors, stress response, and chronic stress while synthesizing the theories of McEwen, Lazarus, and Smyth (Cohen, Gianaros, & Manuck, 2016). This model begins with a stressor, which, once appraised as stressful, triggers a negative emotional response. This emotional response leads to activation of the biological systems (i.e., autonomic nervous system and neuroendocrine system), as well as poor health behaviours. It is the combination of all of these responses that leads to poor health outcomes (e.g., disease-related change in physiology, increased morbidity and mortality risk). Thus, Cohen presents stress as the degree to which one's demands exceed their available coping resources.

1.2 The Epidemiology of Stress

Four decades of research highlight the role that chronic psychological stress plays in the concurrent and future development of depression (Hammen, 2005) and cardiovascular disease (Dimsdale, 2008; Steptoe & Kivimäki, 2012, 2013). For example, in a study of 816 women

(mean age = 41), high levels of chronic stress were associated with a 16% increased likelihood in the onset of episodes of major depression (Hammen, Kim, Eberhart, & Brennan, 2009). Additionally, a systematic review analyzing six prospective studies with a total of 118,696 patients found a 27% increase in future incident of coronary heart disease in those reporting high stress (Richardson et al., 2012). Chronic stress has also been shown to predict mortality risk. A 22-year longitudinal study of 19,698 individuals found that highly-stressed men were at a 32% greater risk of all-cause mortality than less stressed men (Nielsen, Kristensen, Schnohr, & Gronbaek, 2008). While several government entities acknowledged the epidemiological health threat that chronic stress plays in the health of Canadians, no community-based or national campaigns address stress reduction as a primary target.

In the 1960s, another epidemiological health crisis existed with tobacco consumption. Similar to chronic stress, tobacco consumption had been shown to increase one's risk of cardiovascular disease and mortality (Dorn, 1959). Informed by the social and behavioural sciences, tobacco consumption interventions and policies were implemented globally at the economic (e.g., taxation on tobacco products), behavioural (e.g., cessation support, smoke-free areas), and social (e.g., mass media campaigns and legislation on tobacco-related advertisement) levels (Brown, Platt, & Amos, 2014; Mabry, Olster, Morgan, & Abrams, 2008). In countries such as the United States of America, this resulted in a drastic reduction in smoking rates, with the prevalence of adults smoking going from 42.4% in 1965 (Centers for Disease Control and Prevention, 1999) to 14% in 2017 (Centers for Disease Control and Prevention, 2019).

In order to implement similar large-scale stress reduction strategies as those seen with smoking, we need to first identify a target population - those at risk of reporting high stress. This involves a two-level approach: identifying what factors are associated with high levels of chronic

stress and then determining which of those factors are the most important. Over the past two decades, the number of articles relating to chronic stress, as indexed by Web of Science, has grown more than ten-fold (i.e., 7,796 articles in 1999 to 85,163 articles in 2019), representing an ever-expanding wealth of knowledge in the field. This body of research has found chronic stress to be associated with demographic (e.g., age, sex, education [Cohen & Janicki-Deverts, 2012]), social (e.g., pessimism [Shields, Toussaint, & Slavich, 2016], sense of belonging [Ross, 2002]), and behavioural (e.g., smoking status [Slopen et al., 2013], physical activity levels [Stults-Kolehmainen & Sinha, 2014], eating habits [Tryon, Carter, DeCant, & Laugero, 2013]) factors.

It is now known that there are correlates of stress in many fields of study, and with more research, these factors could potentially be used to identify those at risk of high stress. However, many previous studies in the stress literature employ tools to conduct explanatory or confirmatory analysis, rather than exploring overarching patterns. In order to help guide new streams of explanatory or confirmatory studies, the field of stress research requires a cross-disciplinary and expansive exploratory approach. A multidisciplinary analysis that examines chronic stress as it is related to multiple behavioural and social domains can help to develop a framework on which future investigations in the fields of epidemiology, psychology, economics, sociology, and medicine can build. This framework will allow for novel interventions and steer policy in population health towards a foundation in the behavioural and social sciences.

1.3 Machine Learning Applications

With the increase in computational power over the previous decades, advanced statistics have grown more commonplace in research, especially in fields such as epidemiology (Jean et al., 2016; Seligman, Tuljapurkar, & Rehkopf, 2018), medicine (Al'Aref et al., 2018; B. Wu et al.,

2003), and psychology (Devillers, Vidrascu, & Lamel, 2005; Gao, Calhoun, & Sui, 2018). Machine learning is an emerging field existing at the intersection between computer science and statistics (Deo, 2015).

Researchers increasingly apply machine learning approaches to research questions in the medical, biological, behavioural, and social sciences for several reasons. First, it can be used to highlight important factors as a variable selection tool. Díaz-Uriarte and Alvarez de Andrés (2006) studied variable reduction in several microarray data sets for cancer prediction. In ten data sets, each focusing on a separate type of cancer, they were able to greatly reduce the number of predictor genes (i.e., 95% or greater) required to achieve a similar predictive performance as the full set. Second, it offers a method that is less particular about the data used. With traditional statistics (e.g., correlation, ANOVA, multilevel modeling), there are usually assumptions regarding the data (e.g., normality, linearity, homogeneity of variance) that must be met (Osbourne & Waters, 2002). In cases in which a researcher collects data that do not meet these assumptions, the results of an analysis might prove inaccurate. However, in machine learning, there are no assumptions for the data (Smith, Ganesh, & Liu, 2013). This allows a greater number and variety of types of variables to be studied in any particular analysis. Finally, though the parameters (e.g., beta estimates) are often impossible to define in machine learning, the results are often seen to generate better model fit compared to traditional statistical methods, such as regression (Couronné, Probst, & Boulesteix, 2018; Seligman et al., 2018; Worachartcheewan et al., 2015).

Of the many methods of machine learning, researchers frequently use tree-based models. These models create a decision tree based on the data and sort all observations based on these decision rules. Once the data are sorted, the average value of the outcome variable for each

observation in a group is used as the predicted outcome for each new observation in the group. Hastie and colleagues (2017) note that these models produce accurate estimates (i.e., low bias), but they are highly influenced by the data sampled to create the model (i.e., high variance across models). Leo Brieman (2001) offered a new method for tree-based modeling: random forest regression. This novel ensemble tree method offered a solution to the variance-bias trade-off through two changes to traditional tree models. Firstly, rather than base the results on one tree, multiple trees are produced and the results are aggregated. Secondly, random forest regression only tests a fraction of the variables when creating decision rules, which decreases the risk of selecting two correlated variables in subsequent decision rules. This process reduces correlations between trees and leads to the overall model having fewer variables and, therefore, higher reliability. These features make random forest analysis an ideal addition to the statistician's toolbox for those in fields in which there are many interrelated factors, such as medicine or epidemiology.

1.4 Summary

If individuals are exposed to a stressor and feel that they do not have the resources to overcome it, a stress response will occur, which may induce changes in cardiovascular, immune, and neuroendocrine systems. This stress response is not necessarily harmful to the body. Rather, it is the repeated or prolonged activations of the stress systems that cause deterioration in physiological and psychological health. It is important to be able to find out who is currently experiencing high stress by examining what economic and social contexts and behavioural factors predict high stress. Such an approach can also offer insight on potential mechanisms through which stress reduction interventions can provide the greatest benefit. Numerous factors

have previously been linked to stress, creating a complex and multidisciplinary field. However, it is now time to tease apart which of the many factors associated with stress are the most important, a task for which machine learning is ideal. Further, random forest analyses, a common method of machine learning, can offer a solution that accounts for variable multicollinearity and offers lower overall model error.

1.5 Current Study

In the current investigation, random forest regression was used to examine 67 factors across demographic (n = 16), behavioural (n = 7), social (n = 7), life adversity (n = 5), physical health (n = 15) and psychological (n = 17) domains to determine their relative importance in relation to chronic stress through multiple statistical methods. First, simple linear regression was used to assess if the relationship between each variable and chronic stress was positive or negative. Second, a series of independent linear regressions were completed in which chronic stress was regressed on all the factors within each domain (e.g., health behaviours, demographic factors) to test the size of association between each domain and stress. Multiple linear regression was also employed to simultaneously examine the effect size of all variables in the dataset on chronic stress in a single analysis. Finally, a machine learning algorithm, random forest analysis, was used to test for variable importance and account for complex interactions between variables.

This study used a subsample of the 23,089 individuals from the 2012 Canadian Community Health Survey - Mental Health (CCHS-MH) who were 20 years old or older. The CCHS-MH is a nationally-representative cross-sectional study that is completed every six years to measure links between mental health and social, demographic, geographic, and health

variables. (For a full list of variables included in this analysis, see Appendix A - Variable Recoding.) As this is a data-driven analysis, there is no *a priori* hypothesis.

1.6 Research Objectives

Given that many potential correlates of stress remain unexplored, and those that are examined are usually studied in relative isolation, the objective of this study is to simultaneously analyze variables from multiple fields of study to determine the most important correlates of chronic stress.

Chapter 2: Methods

2.1 Data Source

The Canadian Community Health Survey (CCHS) was first conceived in the 1990s in response to a report from the National Task Force on Health Information. Members of the task force felt that Canadian health data were fragmented, incomplete, not well analyzed, and the results of the analyses of these data were not reaching the Canadian population (Statistics Canada, 2013). In response to these issues, the Canadian Institute for Health Information, Statistics Canada, and Health Canada (1999) created a report outlining a plan to modernize health information in Canada. Among the goals listed in the report was the development of a data source with which to track health factors on a regional, provincial, and national scale. After extensive consultation with stakeholders regarding the data that should be collected, the first CCHS was released in 2000.

In addition to the annual CCHS, a supplementary CCHS survey is conducted every other year. These surveys use a smaller sample size and rotate between different aspects of health (e.g., nutrition, healthy aging, and mental health). The current study utilized data from the 2012 Canadian Community Health Survey - Mental Health (CCHS-MH) survey. This cross-sectional survey was designed to "examine links between mental health and social, demographic, geographic, and economic variables or characteristics" (Statistics Canada, 2013, p. 3).

2.2 Participants

The sample from the 2012 CCHS-MH consisted of Canadians aged 15 or older in the 10 provinces. Excluded from the sample were institutionalized populations, individuals living on First Nation reservations and Crown land, individuals living in the three territories, and members

of the Canadian Armed Forces. Statistics Canada (2013) estimated that these exclusions represent only 3% of the population. The survey response rate was 68.9% (n = 25,113) and which, with survey weights, represents 28.3 million Canadians. This analysis employed a subsample of respondents age 20 years old or older (n = 23,089).

2.3 Outcome Measure

The CCHS-MH measured chronic psychological stress through one item. This item asks "Thinking about the amount of stress in your life, would you say that most days are...?" and includes, using a Likert-type scale, the potential responses: 'not at all stressful' (1), 'not very stressful' (2), 'a bit stressful' (3), 'quite a bit stressful' (4), and 'extremely stressful' (5). For this analysis, this variable was treated as continuous and standardized. This variable was missing for approximately 0.1% of subjects. For those subjects, multiple imputation was used to estimate values.

While this item has yet to be validated in terms of how well it measures stress, it has been included in several Statistics Canada surveys over the past two decades (e.g., Canadian Census, National Population Health Survey), allowing for comparison and generalization with other work conducted by Statistics Canada. Further, it has been used in Canadian data for nearly 20 years in the Canadian Community Health Survey (Statistics Canada, 2012). There is extensive research using the single-item stress question. For example, it has been used to show that the lives of individuals with multiple sclerosis were no more stressful than those without multiple sclerosis (Patten et al., 2012), chronic stress is associated with increased odds of depression (M. Shields, 2006), and high levels of stress are associated with worse overall health (Ross, 2002).

2.4 Predictor Variables

Sixty-seven potential correlates were selected from the CCHS-MH for the present analysis. These variables measure health behaviours (e.g., physical activity levels, sleep behaviours, drug/alcohol use), early and recent life adversity (e.g., adverse events during childhood, recent unmet healthcare needs), mental health (e.g., anxiety disorders, perceived coping abilities), physical health (e.g., chronic disease, disability), social factors (e.g., community belonging, negative social interactions), and demographic information (e.g., education, income). These variables were chosen for analysis as they represent an interdisciplinary range of factors, while minimizing redundancy (i.e., not using both a scale and the individual items on a scale). For a full list of variables analyzed, see Appendix A - Variable Recoding.

2.5 Machine Learning Approaches

There are numerous machine learning methods, each with different applications. One common feature of machine learning is validation. This is achieved by splitting the data, most commonly into two unequally-sized data sets (i.e., training data set and testing data set) or, if the sample is small, k-fold cross validation (i.e., any number, k, of equally sized data sets over which all but one train the model and the final one tests the model). These methods are how machine learning estimates prediction error (Hastie et al., 2017).

For regression analysis, tree-based learning and its derivatives are some of the most common methods. This section will begin with a description of traditional tree-based learning, followed by a brief overview of the most common ensemble tree learning methods. These are

derivatives of traditional tree-based methods that were introduced to improve upon the statistical shortcomings of tree-based models.

2.5.1 Tree-based Learning

Tree-based models are a machine learning algorithm that creates one or more decision trees based on the data provided. Although tree-based models can be used for either regression or classification, only regression will be discussed in this paper as the proposed outcome variable, chronic stress, is being treated as continuous. The following describes the methods for a regression tree as presented by James and colleagues (2015). As the sample for this study was large, the two data set validation method was used. In this method, observations are randomly divided into two subsets: two-thirds are assigned to the training set, which is the basis for building the tree model, and the remaining data are assigned to the testing set, which is used to evaluate model performance (James et al., 2015). The algorithm uses the training data to build the model using recursive binary splitting. That is, the algorithm will select a variable and, at each potential value of the variable, divide the data into two groups: those observations that meet the potential value and those that do not. For example, if income were measured as quartiles, the algorithm would first split the data by those who were in the top quartile compared to the bottom three, then the top two compared to the bottom two, and finally, the top three compared to the bottom quartile. The amount of variance in the outcome variable within each group is measured, and then the model repeats this splitting pattern for each variable in the data set. Once each variable has been tested, the algorithm uses the variable with the greatest reduction in outcome variance to split the data into two subgroups. In each of the newly created subgroups, this splitting process is repeated until variance cannot be sufficiently reduced or there are too few

observations in a subgroup. In this type of model, a clear tree diagram is shown (see Figure 1 for an example of a hypothetical tree model) and each split rule is displayed for the reader to understand how an observation is assigned its predicted value.

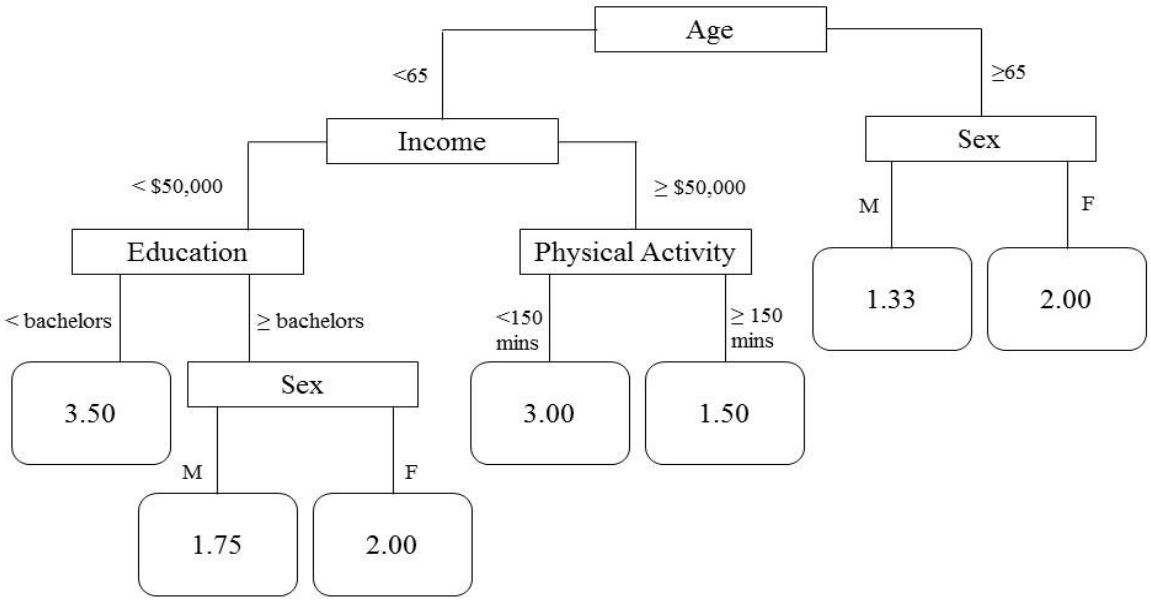

Figure 1 - An example tree model, using dummy data, showing how age, income, sex, education, and physical activity predict chronic stress scores.

Note: In this example, the greatest reduction in variance is a split on age, between those who are younger than 65, and those who are 65 or older. The leaf nodes (i.e., the bottom-most nodes), represent predicted scores; these are based on the average score of each observation that meets all of the preceding split rules. For example, a 45 year old female with a bachelor's degree who makes $45,000 per year would have a predicted stress score of 2.

While the predicted values in a regression tree model are clearly shown, the model fit is not. To test the model fit, the test data set is entered into the tree model and the squared differences between the predicted outcome of the tree and the actual outcome of the data are

averaged to produce the mean squared error (MSE). Without this testing data set to validate the model, it would be difficult to test for model fit. In machine learning, there are two common inappropriate model fits: underfitting and overfitting. (For visual examples of underfitting and overfitting, see Figure 2.) Underfitting is when a model does not account for sufficient variance; an underfit model does not perform well, though it performs equally poorly on both training and test data (James et al., 2015). Overfitting occurs when the model learns the training data too well; while these models provide minimal error in training, they often cannot account for variance in the test data (James et al., 2015). As tree models have a tendency to overfit the data, they are known to have low bias and high variance (Breiman, 1996).

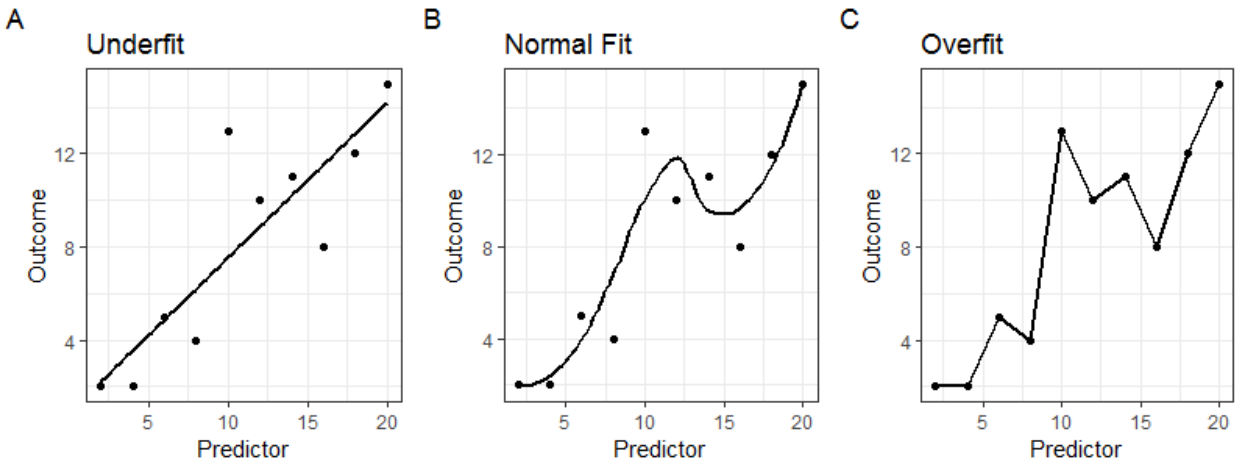

Figure 2 - A visual example of an underfit (2a), normal fit (2b), and overfit (2c) model.

Note: while the three models are based on the same ten data points, their predictions will be different. The underfit model will not be able to accurately predict outliers, or any point that is distant from the linear model. The overfit model will predict all data in the training set with minimal error; however, it will not be able to predict the testing data unless it follows the exact same pattern as the training set. The normal fit model accounts for variance equally well in both testing and training data.

2.5.2 Ensemble Tree Methods

Several algorithms have been developed in order to improve the validity of tree-based learning, including ensemble-based tree methods (Hastie et al., 2017). These algorithms aggregate multiple tree models and create one final model (see Figure 3). This allows the models to keep the low bias of decision trees, but reduce variance (James et al., 2015). Three common examples of ensemble tree methods are boosting, bootstrap aggregating (i.e., bagging), and random forests. Each of these ensemble methods improves upon traditional tree-based learning in similar ways and has merits and disadvantages. One similarity between these models is the loss of interpretability. In ensemble tree methods, the ability to follow a single tree is removed. Rather, an aggregate of predictions is used in which more than one tree is created and the results of all trees are averaged.

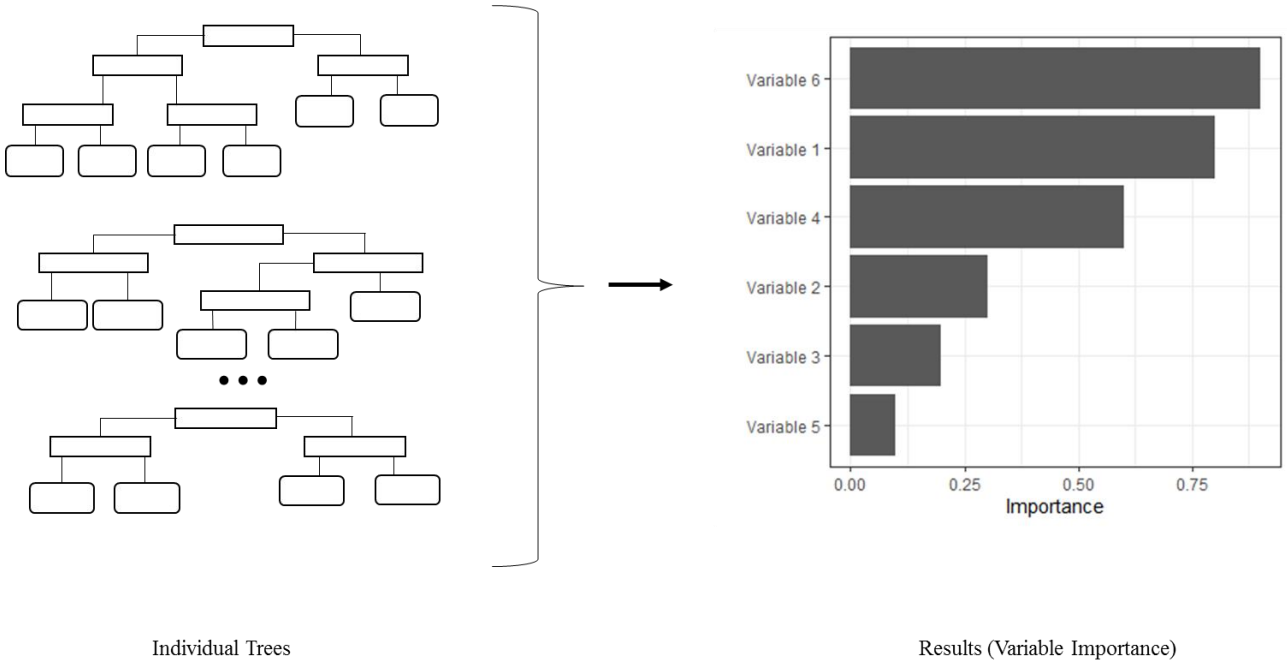

Figure 3 - A diagram of ensemble tree learning

Boosting, unlike other ensemble methods, is an additive model; that is, each tree is grown sequentially and built using the information provided by all trees grown preceding it (James et al., 2015). This slow learning process fits a model based on the previous models' residuals, allowing it to specifically improve on the areas of weakness from previous models (Berk, 2006). However, boosted models require cross-validation in order to find the optimal number of trees, as too many trees may cause overfitting (James et al., 2015).

Bagging is the simplest of the tree-based ensemble methods, both in terms of theory and computational power required. In bagging, a number of samples, each composed of a random subset of observations, are taken from the training data set. For each of these samples, a tree is created using all observations (James et al., 2015). Each tree is created as described in Section 2.4.1. Using the aggregate of many trees allows for bagging to be less influenced by any specific observation, as that observation will not be present in all trees (Grandvalet, 2004).

Random forest analysis is similar to bagging but, by employing one small change, it allows for a reduced correlation between trees. As in bagging, a number of samples, each composed of a random subset of observations, is taken from the training data set. However, when creating each tree, each potential split rule in the tree only tests a random subset of variables for amount of variance reduced (Berk, 2006). The original developer of random forests recommended that the number of features available to be tested at each split should be one-third of the total number of predictors (Breiman, 2001). For example, in an analysis of 30 variables, only 10 would be tested at each split. The lack of available variables leads to a better differentiation between trees, allowing for a more robust analysis.

Boosting and random forest models often have better performance metrics compared to bagging (Dietterich, 2000; Strobl, Malley, & Tutz, 2009). Additionally large amounts of noise

(i.e., unexplainable variance) can hinder the accuracy of boosting and cause overfitting, bagging and random forests are unaffected (Breiman, 2001; Dietterich, 2000). Furthermore, the use of variable limitation in random forest algorithms reduces the risk of multicollinearity, which is a common cause of overfitting (Breiman, 2001). In health-based survey data there are many correlated variables (e.g., alcohol consumption and depression [Graham, Massak, Demers, & Rehm, 2007], income and high blood pressure [Lemstra, Rogers, & Moraros, 2015]). Due to the potential for unexplainable variance and correlated variables in the dataset, a random forest algorithm, rather than boosting, was selected for the analysis.

An important aspect of random forest is tuning the parameters (*e.g.*, number of variables tested per split, number of trees) to obtain the optimal performance. This is achieved by using the training data in multiple models with differing parameters and testing the model fits. For number of variables tested per split, Breiman (2001) suggested one-third of the number of predictors. Nevertheless, he also recommended testing both twice and half as many variables (Breiman, n.d.). He further noted there is no upper limit to the number of trees (Breiman, 2001). It was later reported that there is only minimal improvement in model fit beyond 128 trees (Oshiro, Perez, & Baranauskas, 2012). However, it has been suggested that to get stable estimates of variable importance 1000 or 5000 trees should be used (Breiman, n.d.)

2.6 Statistical Analyses

2.6.1 Data Preprocessing

For this analysis, 67 factors from the dataset were used. Prior to analyses, variables were preprocessed, a practice sometimes described as ‘cleaning.’ All continuous and ordinal variables were converted to standardized scores. Nominal categorical variables were dummy coded and all

binary variables were coded as 0 or 1, with all responses of ‘don't know’, ‘refusal’, ‘not stated’, and ‘not applicable’ coded as NA. This cleaning led to a total of 105 variables. As an example, Marital Status was a categorical variable with three response options, ‘Single’, ‘Married’, and ‘No Longer Married’. Each option was dummy coded and became a new variable. (For a full list of these transformations, refer to Appendix A - Variable Recoding.) This dataset underwent random forest-based multiple imputation to replace all NA values from the data set with predicted values based on a respondent’s responses in all other variables.

As the majority of variables were theoretically negatively associated with stress, variables which were theoretically positively associated with stress were reverse coded. These variables included moderate-to-vigorous physical activity, life satisfaction, social provisions scale, income, and age. An (R) notes these variables in all figures and text.

2.6.2 Correlation and Regression Analysis

For each variable, Pearson correlation coefficients were calculated with respect to each other variable. Linear regression analyses were conducted in three steps. First, stress was regressed on each individual variable. Second, six linear regression models were created in which stress was separately regressed on all variables within each category of variables (i.e., health behaviours, life adversity, mental health, physical health, social factors, and demographics). Finally, a linear regression model was created in which stress was regressed on all variables. All linear regressions included the sample weighting provided by the CCHS-MH.

2.6.3 Random Forest Analysis

Two-thirds of the data were randomized into a training data set and the remaining data were used as the testing data set. The training data were analyzed with multiple tuning parameters, including the number of variables available at each split and the number of trees in the model. As this dataset includes 67 variables that were coded into 105 variables, six values for the number of variables tested parameter were assessed (see Table 1 for a more detailed justification).

Table 1 - Justification for each number of variables tried during a split in random forest analyses.

Value	Calculation	Justification
22	67 / 3	Default based on recoded count of variables (Breiman, n.d.)
45	(67 / 3) * 2	Expert opinion (Breiman, n.d.)
11	(67 / 3) / 2	Expert opinion (Breiman, n.d.)
35	105 / 3	Default based on initial count of variables (Breiman, n.d.)
70	(105 / 3) * 2	Expert opinion (Breiman, n.d.)
18	(105 / 3) / 2	Expert opinion (Breiman, n.d.)

Owing to the complex structure of the CCHS-MH data and the desired outcome of variable importance, the number of trees tested in the present analysis were 128, 1000, and 5000. Each potential combination of these parameters (i.e., number of factors and number of trees) was used to conduct individual random forest analyses using the training data. (See Table 2 for a list of parameter combinations that were tested.) The model that produced the best fit (i.e., lowest MSE, greatest R^2) was selected as the final model. Finally, the parameters employed in the final model were used to calculate the variable importance using the data of the testing data set.

Table 2 - The eighteen models used to tune model parameters.

Model Number	Number of Factors	Number of Trees
1	11	128
2	18	128
3	22	128
4	35	128
5	45	128
6	70	128
7	11	1000
8	18	1000
9	22	1000
10	35	1000
11	45	1000
12	70	1000
13	11	5000
14	18	5000
15	22	5000
16	35	5000
17	45	5000
18	70	5000

Note: each model is differentiated based on the number of variables tested at each split and the number of trees that compose the forest. The model that produced the lowest error was validated with the testing data and used to determine variable importance.

Variable importance was calculated using the permutation method created by Breiman (2001). In this method, each tree is built as previously specified, then one variable is replaced with a randomly generated set of numbers. The trees are then reconstructed using the tested variable, now replaced with this random vector of numbers, and all other original variables. The difference in error between the tree with no permutated variables and the tree using the permutated variable is calculated and averaged between all trees. This process is then repeated for each other variable in the model. These changes in error are an estimate of the effect that a single variable has on the overall model error.

2.7 Software

R statistical software was used to conduct all analyses. Functions from the *tidyverse* and *missRanger* packages were used for the data preprocessing, and functions from the *ranger* package were used for conducting the random forest analyses. There are currently several R packages available for performing random forest analyses (e.g., *party*, *randomForest*, *Rborist*, *Random Jungle*, and *ranger),* with *randomForest* being one of the most common. However, *randomForest* does not allow for survey weights nor case weights within regression analysis. *ranger*, which has been validated to produce similar results to *randomForest*, allows for case weights while also performing faster and requiring less memory than *randomForest* (Wright & Ziegler, 2017). All figures were produced by the *ggplot2* package.

Chapter 3: Results

3.1 Data Cleaning

In terms of multicollinearity, only one pair of variables (Income Relative to Province and Income Relative to Canada) had an absolute correlation coefficient greater than 0.85 ($|r| = 0.99$), and thus, Income Relative to Canada was removed from the analysis (see Appendix B - Omitted Variables). We also removed individuals who were 19 years old or younger (n = 2,024) to ensure that all participants had completed items regarding early life adversity, questions that were only asked to those 20 years of age or older. In this subset of data, several variables contained missing values. One variable, pulmonary conditions, had to be removed due to a large proportion of missing values (31%). Second, data for both coping ability and previous year employment were missing not at random, rather the missing values were a function of survey design. Sensitivity analyses were used to determine if the results were robust against these missing data. Each analysis ran the random forest procedure described above using only a subset of the data. The first sensitivity analysis included participants who were between the ages of 20 and 75. The second analysis included only those who had reported a greatest source of stress. As the results of these analyses showed similar variable importances for all analyses, missing values were input using multiple imputation. For other variables with missing data, analyses showed the patterns of missingness were not completely at random. However, the missingness for each of these variables was between 0% and 7.5% (See Appendix C - Missing Data), all of which were below the 10% threshold that has been noted to potentially bias results (Bennett, 2001). Thus, these patterns of missingness were treated as missing at random and multiple imputation was used to replace missing values. Each of the following analyses used the imputed data. Once cleaned, this imputed dataset contained 105 predictors and 23,089 observations with no missing values.

3.2 Correlations

Due to the large number of variables, correlation analysis resulted in over 5,500 unique pairwise comparisons. All correlations are shown through boxplots in Appendix D. Individual correlations ranged from -0.74 to 0.71, though the mean of all correlations was 0.02. On average, the variable with the greatest correlations with all other variables was emotional impact of health (r_{mean} = 0.10, $SD_{correlation}$ = 0.15), and the variable with the smallest correlation was living without a partner or children (r_{mean} = < 0.001, $SD_{correlation}$ = 0.15).

3.3 Simple Linear Regression

The beta estimates for factors associated with chronic stress ranged from -0.257 (Household Type - Partner and no kids) to 0.944 (General Anxiety Disorder), with 20 variables having very small beta estimates (i.e., between -0.05 and 0.05). Figure 4A shows the estimated betas of the variables with the greatest effect sizes. As each of these betas are greater than 0, each of these variables are positively associated with stress. That is, those who report a greater number of negative social interactions (Negative Social Interactions), lower levels of life satisfaction [which is reverse coded; Life Satisfaction (R)], greater sleep troubles (Level of Insomnia), being more emotionally affected by their health (Emotional Impact of Health), being younger [Age (R)], having a diagnosis of major depression (Major Depression), and those who perceive their mental health needs were not fully met (Recent Life Events - Unmet Needs) reported greater levels of stress. Additionally, those who reported less available social support when facing their greatest source of stress (Coping - Social Support), and lower abilities to cope on a day-to-day basis (Coping on a Daily Basis) reported greater stress. See Appendix E for the full list of beta estimates.

Effect sizes (i.e., R^2) represent the amount of variance accounted for by each variable and can range from 0 to 1. In the present analysis, the greatest effects sizes were Negative Social Interactions, which accounted for 12.8% of the variance in reported stress, and Life Satisfaction (R) which accounted for 9.25% of the variance in reported stress (See Figure 4B for the variables with the ten greatest effect sizes.) It is important to note that these effect sizes represent the association of a single independent variable on the outcome. It is possible, and quite likely, that some of the variance explained by any of the independent variables can also be partially explained by a different variable. Of the 105 simple linear regression models tested, 86 (82%) were statistically significant.

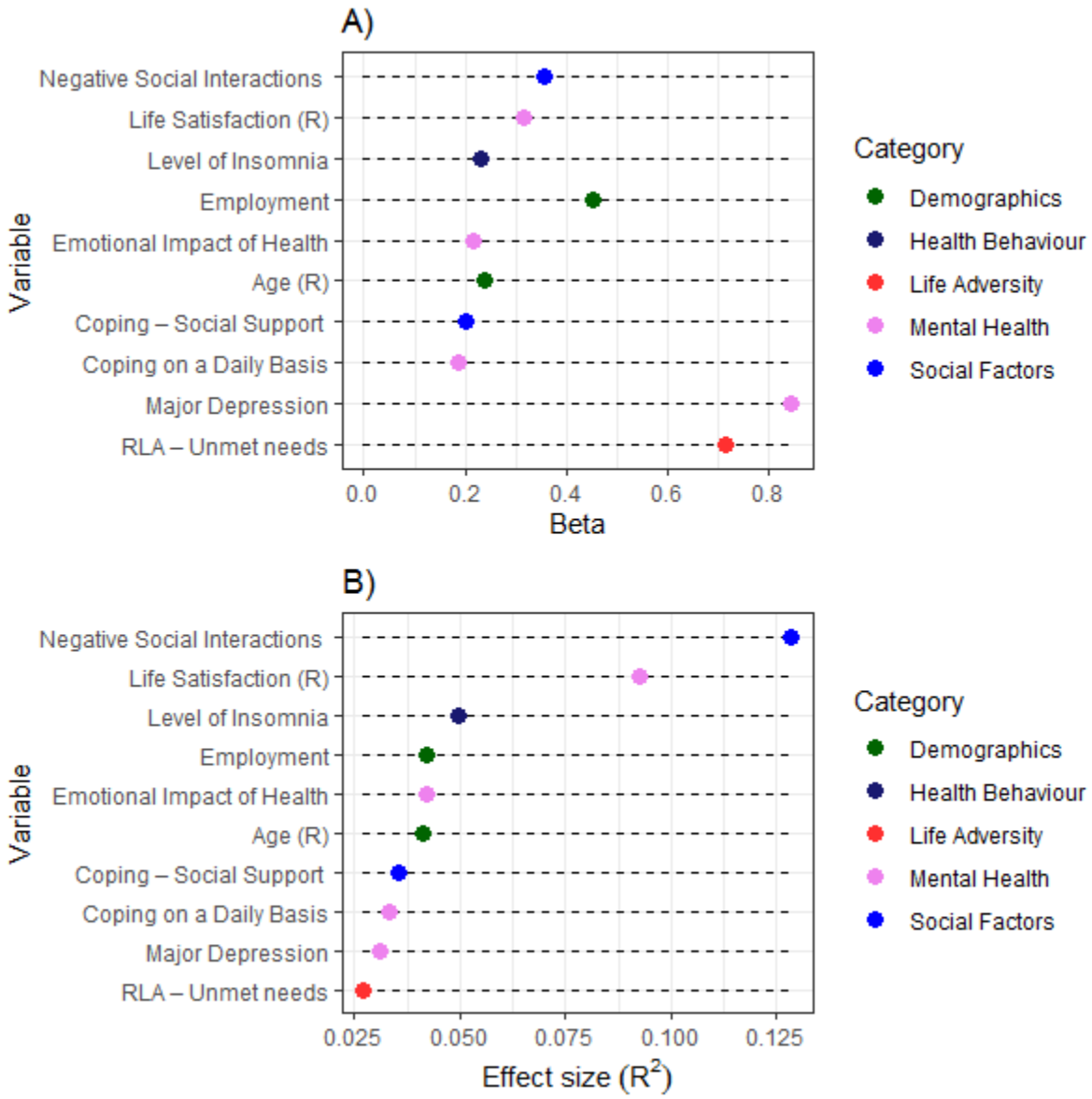

Figure 4 - A) Beta estimates and category of the variables with the greatest effect. B) Effect sizes and category of the variables with the greatest effect sizes.

Note: See Appendix E for the full list of beta estimates and effect sizes.

3.4 Multiple Linear Regression

A multiple regression models was run for each category of variables (i.e., Health Behaviours, Life Adversity, Mental Health, Physical Health, Social Factors, and Demographics). Further, an additional multiple linear regression was run using all predictors. Effect sizes (Table 3) are presented as the effect sizes (R^2) and adjusted effect sizes (Adjusted R^2), accounting for the number of predictors in the model. As all differences between unadjusted R^2 and adjusted R^2 were less than 0.01, the following results represent adjusted effect sizes.

Table 3 - Results from the multiple linear regression analyses.

Category	Number of Predictors	R^2	Adjusted R^2
Health Behaviours	11	0.060	0.059
Life Adversity	5	0.050	0.050
Mental Health	29	0.221	0.220
Physical Health	15	0.062	0.061
Social Factors	7	0.148	0.148
Demographics	33	0.140	0.139
All Variables	100	0.314	0.311

Note: for each group of variables and the combination of all predictors, the number of predictors used and effect sizes are presented. Effect sizes are represented by unadjusted R^2 and R^2 adjusted for the number of predictors. These analyses showed minimal differences between unadjusted and adjusted R^2.

In the present study, grouped effect sizes ranged from 0.05 to 0.22, and the all-variable model had an effect size of 0.31. While there were 105 simple linear regression models, each representing a unique variable, there are only 100 variables included in the grouped linear regression. This is due to the simple linear regressions including dummy coded variables that could be tested independently from other potential levels. However, in the group regression, the inclusion of all dummy variables would lead to perfect multicollinearity. For example, the marital status item was dummy coded as single, married, and no longer married. In multiple

regression, the variance accounted for by single and married will be perfectly collinear with the variance accounted for by the no longer married indicator. In practice, this leads to one level of each original categorical variable being omitted from the analysis as the baseline reference group, including marital status (no longer married), smoking (never smoked), dwelling type (other), household type (other), and province (British Columbia).

3.5 Random Forest

3.5.1 Model Tuning and Selection

Observations were randomly divided into two data sets, the training (n = 15,392) and testing (n = 7,697) data. The training data were used to determine the best model. This was achieved by using the training data to create 18 random forest models. Each model used the same data, but had slight differences in parameters, a tuning process that was done to optimize the model. The parameters tested include the number of trees and the number of variables included at each split. (For justification for each parameter that was tested, see Section 2.6.3.)

Each of the 18 random forest models tested generated similar results (Table 4). Two model diagnostic tools were used to determine the optimal model parameters. First, error was calculated using the mean squared error (MSE), which is the squared difference between the predicted and actual outcome. The second diagnostic was R^2 effect size, which represented the amount of variance accounted for by the model as a proportion of total variance. The MSE of all models ranged between 0.701 and 0.729. Model 16 had the lowest error, suggesting its predicted values were the closest to the reported outcomes. Similarly, effect sizes (R^2) ranged from 0.279 - 0.307. Model 16 also had the largest effect size, accounting for 31% of the variance in the dependent variable. According to both of these model diagnostics, Model 16, which used 5000

trees and 35 variables per split, was the best model (Table 4). Therefore, the final parameters selected were 35 variables tested per split and 5000 trees.

Table 4 - Error (MSE) and effect size (R^2) for each of the 18 models tested.

model	m.try	num.trees	MSE	R^2
1	11	128	0.722	0.287
2	18	128	0.722	0.286
3	22	128	0.721	0.288
4	35	128	0.721	0.287
5	45	128	0.723	0.285
6	70	128	0.729	0.279
7	11	1000	0.705	0.303
8	18	1000	0.706	0.302
9	22	1000	0.703	0.305
10	35	1000	0.705	0.303
11	45	1000	0.706	0.302
12	70	1000	0.712	0.296
13	11	5000	0.701	0.307
14	18	5000	0.702	0.307
15	22	5000	0.701	0.307
16	**35**	**5000**	**0.701**	**0.307**
17	45	5000	0.704	0.304
18	70	5000	0.708	0.300

Note: this table includes Model ID (model), the parameters used (m.try, number of variables tested per split; num.trees, number of trees composing the forest) and the diagnostic results, Model error (MSE) and effect size (R^2). Model 16 displayed the lowest error and highest effect size, suggesting that the parameters to use for the testing data were 35 variables tested at each split and 5000 trees.

To guard against overfitting, a random forest model was created using the testing dataset and the parameters of Model 16. This final model was shown to have similar results to the training models (MSE = 0.743, R^2 = 0.28) and was used to calculate variable importance.

3.5.2 Variable Importance

The variables with the highest importance in random forest analysis are shown in Figure 5. Importance was standardized as a percentage of the maximal variable importance (Equation 1). See Appendix G for the full list of variable importance, both raw and standardized.

$$\boldsymbol{Standardized\ Importance} = \frac{\boldsymbol{Variable\ Importance}}{\boldsymbol{Variable\ Importance_{Max}}}$$

Equation 1 - Standardized Importance

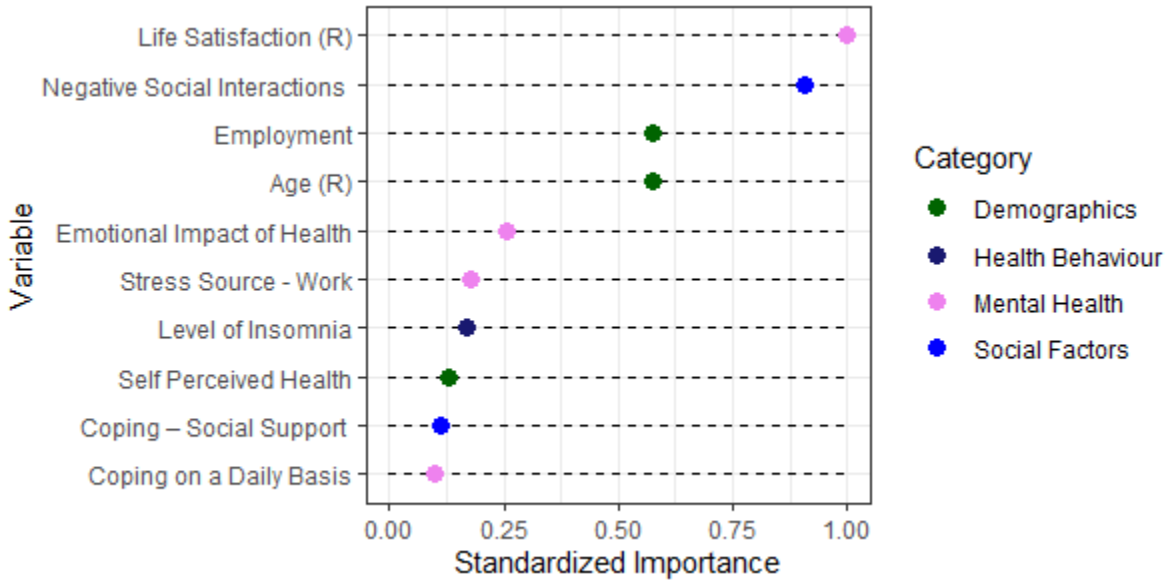

Figure 5 - Standardized variable importance and category for the most important variables.

The variables with the greatest standardized importance were Life Satisfaction (R) and Negative Social Interactions, 1.00 and 0.91 respectively. The variables with the next greatest importance were Employment (0.58) and Age (0.58), followed by Emotional Impact of Health (0.26), Stress Source - Work (0.18) and Insomnia (0.17). All other variables had a standardized importance of less than 15%. It is important to note while there are no user-defined interactions

in random forest algorithms, each of these importance values are determined after accounting for other potential sources of variance.

Surprisingly, other variables that are often linked to chronic stress were not found to have large variable importance nor effect sizes. These variables include Sex, Early Life Adversity, Caregiving Stress, and Physical Activity (Table 5).

Table 5 - Common correlates of chronic stress with low standardized variable importance and effect sizes.

Variable	Standardized Variable Importance	Effect Size (R^2)
Sex	0.05	<0.01
Early Life Adversity	0.04	0.02
Greatest Source of Stress - Caregiving	0.00	<0.01
Physical Activity	0.02	<0.01

3.6 Sensitivity Analysis

Due to survey design, there were three variables where the amount of missing responses were greater than 10%. First, 24% of observations for pulmonary disease information were missing because this item was only asked to those 35 years old or older. Due to the high level of missingness and the incomplete coverage for all respondents, this variable was removed from all analyses. Second, previous year employment was missing in 12% of observations. However, the majority of these missing values were due to this question not being asked to those 75 years old or older. Indeed, in a subset of data including only those who are between the ages of 20 and 75, less than 5% of the data were missing. Finally, the item regarding what individuals perceived as the greatest contributor to their stress included a "nothing" option that was coded as NA for these analyses. This led to 16% of the data being labeled as missing. Consequently, the coping skills item that was only asked to those who reported their greatest source of stress was also missing

16% of data. In a subsample of all those who did not report "nothing" as their greatest contributor to stress, less than 5% of data were missing. For both previous year employment and coping, missing values were imputed.

As the data missing from the two variables that were not removed (i.e., Previous Year Employment, Greatest Source of Stress) were not missing at random, two sensitivity analyses were conducted to test the potential effects of these missing values on the results. The first included a subsample of those between the ages of 20 and 75, as the employment question was not asked of those who were over 75. The second included a subsample of participants who did not respond "nothing" as the greatest contributor to their stress, as the coping question was only asked to those who reported their greatest source of stress. A sensitivity analysis was used to test if the results were driven by data from those with mental illness. In this analysis, a subsample of those without reported mental illness (i.e., depression, bipolar disorder, any anxiety disorder, mania, and hypomania) was analyzed. Each of these subsamples was tested using the random forest methods described in Section 2.

These three analyses showed similar importance results to the original data sets, highlighting the robustness of results. (See Table 6 for most important variables and Appendix F - Sensitivity Analysis Results for the full results.)

Table 6 - Standardized importance for random forest analysis and sensitivity analyses for the most important variables.

Variable	Full Analysis Importance	Age 20-75 Importance	Reported Primary Stressor Importance	No Mental Illness Importance
Life Satisfaction (R)	1.00 (1)	0.99 (2)	1.00 (1)	0.97 (3)
Negative Social Interactions	0.91 (2)	1.00 (1)	0.82 (2)	0.97 (2)

Employment	0.58 (3)	0.30 (4)	0.29 (4)	0.58 (4)
Age (R)	0.58 (4)	0.51 (3)	0.64 (3)	1.00 (1)
Emotional Impact of Health	0.26 (5)	0.11 (9)	0.16 (7)	0.15 (8)
Stress Source - Work	0.18 (6)	0.29 (5)	0.08 (13)	0.27 (5)
Level of Insomnia	0.17 (7)	0.26 (6)	0.23 (5)	0.22 (6)
Self-Perceived Health	0.13 (8)	0.11 (8)	0.18 (6)	0.10 (11)
Coping - Social Support	0.11 (9)	0.08 (10)	0.09 (11)	0.16 (7)
Coping on a Daily Basis	0.10 (10)	0.11 (7)	0.10 (10)	0.11 (10)

Note: All reported importance values use the standardized importance. Variables included are those with the greatest importance values in the random forest analysis. The bracketed number following the importance value represents the variable's ranking within each analysis in terms of greatest importance.

The results of the sensitivity analyses are quite similar to the results from the initial analysis, with only one variable having a dramatic reduction in importance. In the initial analysis, the greatest importance is seen in the following variables: Life Satisfaction, Negative Social Interactions, Age, and Employment. While Life Satisfaction, Negative Social Interactions, and Age have quite consistent importance in each of the sensitivity analyses, Employment's importance has a notable decrease in both the age 20-75 and the reported primary stressor sensitivity analyses.

Chapter 4: Discussion

4.1 Model Functions

The primary objective of this study was to quantify the relative importance of the correlates of chronic stress from various fields. To explore these relationships, the present study utilized multiple traditional and novel statistical methods to strengthen this exploratory analysis. Each of these models convey unique insights into the relationships between these factors and stress. The simple linear regression models provide the linear explanatory power of the variable (i.e., R^2 effect size) as well as whether its association is positive or negative. The grouped multiple regression models suggest which category of variables has the greatest linear predictive power. The full multiple regression model represents the effect size of all predictors without accounting for multicollinearity or interaction effects. Finally, the random forest analysis shows the correlates with the greatest importance after accounting for complex interactions.

4.2 Regression Results

The regression analyses suggest that the majority of factors are statistically significantly associated with stress. However, the majority of these associations explained less than one percent of the variance ($R^2 < .01$), suggesting these associations have low practical importance. In the simple linear regression models, the relationships with the greatest effect size were between Negative Social Interactions and Stress ($R^2 = .13$) as well as Life Satisfaction and Stress ($R^2 = .09$).

In the grouped multiple regression, mental health was shown to have the greatest effect size ($R^2 = .22$). This may be due to overall mental health being quite proximal to the experience of chronic stress. The second greatest effect size is seen in social factors ($R^2 = .15$), which had

less than one-quarter of the number of variables as mental health. A similar effect size was seen in demographic factors, though it included nearly five times the number of variables. Each of the other groups of factors (i.e., health behaviours, life adversity, and physical health) accounted for approximately 5% of the variance. The full multiple regression model had a greater effect size than any individual group, accounting for over 30% of the variance.

4.3 Random Forest Results

The random forest model had a comparable overall effect size to the multiple regression model with all predictors. However, it also showed that, when considering all variables and their interactions, Life Satisfaction, Negative Social Interactions, Age, and Employment were the most important. Sensitivity analyses were conducted to see if the results were being driven by missing data or a specific subpopulation. In these analyses, Life Satisfaction, Negative Social Interaction, and Age maintained the largest importance values. Aside from those variables, Employment, Emotional Impact Of Health, Work Stress, and Insomnia initially demonstrated high importance values, but only employment had a standardized importance of 20% or greater in the initial analysis and each sensitivity analysis. This suggests that Employment is important, though not to the extent of Life Satisfaction, Negative Social Interactions, and Age. While Age, Employment, and Insomnia each had moderate importance in the random forest models, they had very small effect sizes in the simple linear regression. However, as effect size is a measure of strength of a linear relationship, it is possible that the relationships between these factors and stress are strong but non-linear. Overall, these novel findings offer evidence that the most important correlates of stress are not from one field of study, but rather stress is associated with social, psychological, biological, behavioural, and demographic factors.

4.4 Life Satisfaction and Stress

The present study found that life satisfaction was an important correlate of stress regardless of analytic approach and had a negative linear relationship with stress demonstrating a medium effect size. This aligns with several previous studies that explored the relationship between chronic stress and life satisfaction (Lee, Kim, & Wachholtz, 2016; Matheny, Roque-Tovar, & Curlette, 2008; Smyth, Zawadzki, Juth, & Sciamanna, 2017; Strine, Chapman, Balluz, Moriarty, & Mokdad, 2008). In a study of more than 340,000 people, Strine and colleagues (2008) found that, compared to those who were very satisfied, those who reported being satisfied or dissatisfied/very dissatisfied with life were 3.8 and 13.7 times more likely to report being distressed, respectively. Smyth et al. (2017) noted that this relationship also holds true over multiple time points. Using ecological momentary assessments, they showed that those with high levels of life satisfaction consistently reported lower levels of perceived stress. Overall, there appears to be a negative relationship between life satisfaction and chronic stress (Smyth et al., 2017; Strine et al., 2008), but the causal directionality remains unclear. It is unknown if life satisfaction has a protective effect against stress, or if chronic stress may be affecting one's overall satisfaction with life.

The current literature offers correlational data linking stress and life satisfaction; therefore, the potential for a third variable problem cannot be ruled out. One potential variable that could confound this relationship is optimism. Previous correlational studies have noted that life satisfaction and optimism were positively related, while optimism and stress were negatively correlated (Chang, 2002; Extremera, Durán, & Rey, 2009). A previous systematic review of 50 studies showed that optimism was positively associated with approach coping strategies (i.e.,

managing or eliminating stressors), and negatively associated with avoidance coping strategies (i.e., ignoring or avoiding stressors; Nes & Segerstrom, 2006). It is possible that, in the relationship between life satisfaction and stress, life satisfaction is acting as a proxy for optimism. Another potential explanatory variable is self-efficacy. Jerusalem and Schawzer (2014) noted that those with low self-efficacy are more likely to assess an event as threatening and perceive deficits in their coping abilities. Indeed, Lee and colleagues (2016) found that the relationship between stress and life satisfaction was partially mediated by self-efficacy. The current study used data from a national study that did not include personality traits nor judgements, and thus these potential third variables could not be included in the current study. Future studies should attempt to prospectively investigate the relationship between life satisfaction and chronic stress, and whether it may be accounted for by dispositions, such as optimism, or judgements, such as self-efficacy.

4.5 Negative Social Interactions and Stress

Negative social interactions were also seen to be an important correlate of chronic stress. Further, there was a positive relationship between negative social interactions and chronic stress with a medium effect size. As noted by Cohen (2004), negative social interactions are themselves a source of stress (i.e., stressor). The greater importance of negative social interactions, compared to other stress sources such as early life adversity, may be a function of their proximal nature. Almeida (2005), using data from a large national study, noted that of all daily stressors reported, half were classified as interpersonal tensions. While previous models of stress have noted the potential relationship between negative social interactions and psychological stress, data from previous empirical studies have primarily focused on the link between negative social

interactions and distress, a related but different concept than psychological stress, (Newsom, Rook, Nishishiba, Sorkin, & Mahan, 2005) and positive social interactions and lower levels of chronic stress (Bernstein, Zawadzki, Juth, Benfield, & Smyth, 2018). These studies do not account for the fact that negative interactions may be more harmful than positive social interactions are helpful (Lincoln, 2008), nor do they address that distress and stress are unique but related constructs. This study is one of the first to detect the direct relationship between chronic stress and negative social interactions.

As with life satisfaction, it is also possible that a third variable problem exists for negative social interactions. One such possible variable is neuroticism, a personality trait associated with proneness to anxiety, depression, and emotional instability (Friedman & Kern, 2010). Those with higher levels of neuroticism have been shown to have greater levels of chronic stress (Park et al., 2013). This could be due to individuals with high levels of neuroticism having greater levels of relational aggression (Reardon, Tackett, & Lynam, 2018). Additionally, high levels of neuroticism have been associated with changes in the acute stress response, including smaller heart rate increases and an attenuated cortisol response (Xin et al., 2017). This suggests that neuroticism may impact chronic stress through both increasing the number of stressors and moderating the physiological responses to stressors. As mentioned earlier, personality was not assessed in the CCHS-MH, thus it is not possible to rule out that the relationship between negative social interactions and stress is not confounded by personality traits such as neuroticism.

Another potential confounding variable between negative social interactions and chronic stress are the coping strategies employed by an individual. Research in persons infected with HIV has shown that social conflicts are associated with maladaptive coping styles, including

isolation, wishful thinking, and cognitive avoidance (Fleishman et al., 2000). Further, in a study of 219 women, Manne and colleagues (2005) found that unsupportive partner behaviours led to greater use of avoidance coping and lower overall coping efficacy, but only in those receiving lower social support from friends and family. As this survey tested belief in coping ability and not types of coping strategies employed in response to stressors, it is possible that the relationship between negative social interactions and stress is partially driven by coping strategies.

4.6 Age and Stress

Age was also shown to be one of the three most important variables, with simple linear regressions showing that those who are younger report greater stress and the relationship had a small effect size. Previous studies have noted a negative relationship between age and chronic stress (Cohen & Janicki-Deverts, 2012; Stone, Schneider, & Broderick, 2017; Stone, Schwartz, Broderick, & Deaton, 2010). Again, the mechanisms of effect between age and chronic stress are unclear. Previous research has shown that older adults report fewer exposures to stressors than younger adults (Birditt, Fingerman, & Almeida, 2005). Further, older adults have also been shown to have lower emotional and physical reactivity to a stressor (Birditt et al., 2005; Scott, Sliwinski, & Blanchard-Fields, 2013).

It is conceivable that the high importance of age is also partially due to its association with multiple variables, which are, in turn, related to stress. Previously, age has been shown to be associated with other important correlates including negative social interactions, with those who are old reporting fewer negative social interactions (Birditt et al., 2005; Rauktis, Koeske, & Tereshko, 1995); life satisfaction, with those who are older reporting greater life satisfaction

(Kongarchapatara, Moschis, & Sim Ong, 2014); and employment, with those 25 to 54 years old having greater employment rates than either those younger than 25 or older than 54 (Statistics Canada, 2009). In the present study, age was correlated above 0.20 with over one-fifth of the variables analyzed, suggesting that the high importance of age may be due to interaction effects with many of the variables as opposed to one large direct effect.

The difference between low effect size and high importance for age may be due to the shape of the relationship between age and chronic stress. While some researchers describe this as a linear relationship, recent work by Stone and colleagues (2017) has shown that the relationship between reported daily stressors and age has an inverted-U trend in which daily stress slightly increases during one's 20s and 30s followed by a slow decline until age fifty, at which point a dramatic decline occurs. As simple regression only measures the linear relationship, machine learning is better suited to analyze this pattern.

4.7 Highlighting Non-Significant Findings

Several stress correlates that are commonly reported in the fields of health psychology and behaviour were found to have low explanatory power and variable importance. These variables included biological sex, early life adversity, caregiving stress, and physical activity. Each of these variables showed a low standardized importance (< 0.06) and effect size (< 0.02). While it is possible that they are truly not important to the experience of chronic stress, alternative potential explanations are outlined below.

Previous work has suggested that there is an association between biological sex and chronic stress. Cohen and Janicki-Deverts (2012) found that there were significant differences in reported stress levels. Similarly, Statistics Canada (2014) reported that females were more likely

than males to report greater chronic stress. However, these sources lack any reporting of effect sizes. Therefore, these significant findings may be a function of the sample size, rather than magnitude of effect. While the current study found a significant association between biological sex and chronic stress, there was a very small effect size.

Early life adversity had a very small effect size, but it was larger than the majority of other variables. In terms of variable importance, early life adversity was not in the top twenty most important variables. This discrepancy between effect size and variable importance may be because early life adversity is too distal; a recent review found that the relationship between early life stress and chronic stress in adulthood may be mediated by biological, cognitive, and behavioural changes stemming from the early life adversity (Epel et al., 2018). Previous studies have linked greater early life adversity to elevated markers of inflammation (Baumeister, Akhtar, Ciufolini, Pariante, & Mondelli, 2016), deficits in cognitive (e.g., executive functioning) and affective (e.g., emotional regulation) functioning (Pechtel & Pizzagalli, 2011), and unhealthy behaviours (e.g., greater intake of unhealthy foods, smoking; Duffy, McLaughlin, & Green, 2018). It is likely that the maladaptive outcomes that result from early life adversity influence chronic stress more than the adversity itself when considered in the same model, which may present more proximal and more important factors to measure in future studies. Early life adversity may not affect individuals from all nations equally. Indeed, previous research has shown that early life health and socioeconomic status play a larger role in later life health for Americans compared to the English (Banks, Oldfield, & Smith, 2012). One explanation for these differences is the availability of social services (e.g., universal healthcare); a point which would align Canada more closely to the United Kingdom. An abundance of previous early life adversity

research regarding chronic stress has been conducted using American data. Future research is required to test international differences of the effects of early life adversity on chronic stress.

This study found the effect size and importance for caregiving as a primary stress source to be small. Caregivers have been seen as a highly stressed population. However, previous research has highlighted that caregivers of individuals with dementia are significantly more stressed than caregivers of individuals with other diseases (Ory, Hoffman, Yee, Tennstedt, & Schulz, 1999; Pinquart & Sörensen, 2003). It is important to note that this item assessed whether caregiving was the greatest source of stress in one's life, rather than if the individual was a caregiver, which leads to a few limitations in this item. First, it does not identify those who are caregivers. Some individuals may be caregivers but have greater sources of stress than caregiving; these individuals would not be identified as caregivers in the data. Second, of those who identify with caregiving being their greatest source of stress, it is impossible to determine if they are caring for an individual with dementia or a different illness.

The current study showed a very small association between physical activity and chronic stress, which is contrary to previous literature. A review by Stults-Kolehmainen and Sinha (2014) noted that a majority of previous studies found an inverse relationship between stress and physical activity behaviours. Though they noted that some studies showed a positive association, they postulated such results could be driven by those who use physical activity as a coping mechanism. It is possible that the association between physical activity and chronic stress found in the current study is due to the survey item or its coding. The physical activity item used in the present analysis only measured the behaviours over the past seven days. However, a recent study suggested that to capture an individual's habitual moderate-to-vigorous physical activity requires approximately 35 days of data (Bergman, 2018). It is therefore possible that these activity levels

are not truly representative of the amount of physical activity an individual completes on a long-term basis. Second, the physical activity item only measured moderate-to-vigorous physical activity. However, in the review by Stults-Kolehmainen and Sinha (2014), the inverse relationship between physical activity and stress was synthesized using studies that included light, moderate, and vigorous physical activity. Therefore, light physical activity may also play a role, but such data was not available in the current study. Third, this variable had to be recoded due to its response options. Potential response options representing hours of moderate-to-vigorous physical activity were numeric between 0 and 13.5; however, the last option was the categorical text "14 hours or more." Due to these response options, the item was recoded into groups (i.e., less than 0.1, 0.1 - 2.5, 2.6 - 5, 5.1 - 7.5, 7.6 - 10, 10 or more hours). This recoding may have removed some of the nuance of this item. Finally, due to the self-reported nature of the data, there may be response bias. Indeed, previous work has shown dramatic over-reporting of physical activity levels (Brenner & DeLamater, 2014). For these reasons, caution is suggested when interpreting the results of the importance of physical activity on chronic stress.

4.8 Strengths of the Current Study

Using a large national data set (i.e., CCHS-MH) provided many benefits to this study. This data set was chosen for two reasons: the scope of variables and sample size. Having a data set that originally included over 500 variables allowed for the selection of variables that spanned multiple fields. The CCHS - Annual Component was briefly considered for this project due to its wide selection of variables and larger sample size. However, it had a notable dearth of mental health variables. The CCHS-MH was chosen because it better represented the social and psychological factors, while still including biological, demographic, and behavioural items.

Using the CCHS-MH allowed for increased generalizability and power due to a large sample of more than 20,000 participants.

Incorporating machine learning into an analysis provides several advantages over traditional statistics alone, especially in regards to exploratory data analyses. There is the potential to incorporate a large number of variables and their interactions in machine learning. This is not as easy in traditional statistics. For example, in a regression, two participants per variable are required to estimate regression coefficients (Austin & Steyerberg, 2015). While this means that only 200 subjects would be required to test the main effects of 100 variables, interactions would be substantially more difficult. Testing all main effects and two-way interactions would require just over 10,000 subjects and testing all interactions would require more subjects than there are people on Earth. This can result in researchers vastly limiting the number of interactions that they test due to power restrictions. In using machine learning methods, there is no maximum number of variables that can be tested in a single analysis, allowing for the incorporation of multiple interactions between variables.

Machine learning offers a less restrictive statistical approach than traditional statistics. The vast majority of stress research has been conducted using parametric methods (e.g., correlation, ANOVA, multilevel modeling). However, parametric methods have assumptions regarding the data and, if these assumptions are not met, the analysis may produce inaccurate results (Osbourne & Waters, 2002). In cases in which a researcher collects data that does not meet these assumptions, they must either 1) run an analysis which might prove inaccurate, or 2) omit a percentage of the variables. In machine learning, there are no assumptions for the data, which allows for interactions and non-linear trends to be assessed. Further, the results of these

machine learning methods are often seen with lower error compared to traditional statistical methods, such as regression (Seligman et al., 2018; Worachartcheewan et al., 2015).

Machine learning methods do not conflate meaningful and trivial results. As *p*-values are dependent on both magnitude of the effect and sample size, large epidemiological studies often find many associations statistically significant, even if they have a negligible effect in a clinical context. Researchers employing traditional statistics have, accordingly, been moving away from *p*-value based inferences toward an increased focus on effect size. The use of variable importance in random forests follows a similar idea, as it is more similar to an effect size than significance.

Beyond the overarching advantages of machine learning, there are several benefits to random forests. Tree models inherently produce accurate estimates (i.e., low bias), but they are highly influenced by the data used to create the model (i.e., high variance). However, in random forests, hundreds of trees are produced and the results are aggregated. As variance is a function of sample size, more trees allow for decreased overall model variance. Finally, random forest regression only tests a fraction of the variables when creating split rules, which reduces the correlation between the variables. It is the low bias and variance, in addition to the reduced likelihood of multi-collinearity, that make random forests ideal for large-scale epidemiological studies.

4.9 Limitations

Several limitations stem from the data used in these analyses, primary among which is the use of self-reported data. When participants complete self-reported surveys, their data may be biased by social desirability (Krumpal, 2011). That is, individuals may represent themselves

closer to what they believe society expects of them, rather than what is accurate. For example, people are more likely to over-report their habits for healthy activities, such as physical activity (Adams et al., 2005; Brenner & DeLamater, 2014). While it would be impossible to conduct such a large survey with only objective measures, in terms of both scope and number of participants, it is important to consider this source of bias. Second, as these data present a snapshot of a population at one specified time point, it cannot be considered a causal analysis, only correlational. Future studies should include a longitudinal component to add causal evidence towards these findings. Finally, the one item stress question has not been validated. This item should be validated against a previously validated measure, such as the Perceived Stress Scale, to ensure it is measuring levels of stress and not a different construct, such as distress.

Another source of limitations is the result of the methodology employed for this analysis. First, the stress variable was treated as interval, even though it is a Likert-type item. The choice was made to use regression random forests over classification random forests, which are also employed in analyses of Likert-type scales, due to how error is calculated. For example, in classification random forests, a prediction is either correct or incorrect. If an individual responds that their daily life is ‘not very stressful’ (2), the model will not distinguish between ‘a bit stressful’ (3) and ‘extremely stressful’ (5), rather either would be seen as incorrect. Using a regression tree, the error is defined as the squared difference between the predicted and actual values. Thus, a model predicting ‘a bit stressful’ (3) would have a lower error than a model predicting ‘extremely stressful’ (5). While treating Likert-type items as numeric is still controversial (H. Wu & Leung, 2017), there is previous evidence that Likert-type scales can be used as numeric in statistical analyses and achieve very accurate results (Norman, 2010). Second,

due to computational restraints and variance overlap, we only used 67 of the 584 variables included in the dataset. While it is possible that variable selection resulted in the omission of important factors, the variables used in the present analysis adequately represent each field of stress research and its multidisciplinary nature. To partially reduce the subjectivity of variable selection, we removed variables only if there was high homogeneity (e.g., less than 1% of the sample reported being pregnant), high overlap with other variables (e.g., the social provision scale total score as a single variable rather than each individual item) or variables with no theoretical connection with stress (e.g., height of respondent). Third, as is common for surveys, the CCHS-MH included a number of missing data points. While some of the data was not missing at random (i.e., coping and previous year employment), the majority of the variables with missing data seemed to be missing at random, though not missing completely at random. For the data not missing at random, sensitivity analyses showed similar results to the overall results. For the variables missing at random, only a small percent of the data was missing (< 7.5%; see Appendix C). Therefore, multiple imputation was used to replace all missing values. Although multiple imputation can lead to potentially inaccurate estimations due to unobserved variables, this dataset is likely robust to these errors due to its size, both in terms of variables and observations used to estimate the missing values.

The final limitation is in regards to weighting. Sample weights were included in the CCHS-MH data set, but are not compatible with the *ranger* and *missRanger* packages. According to the CCHS-MH documentation (Statistics Canada, 2013, p. 66), "the sampling weight can be interpreted as the number of people the respondent represents in the Canadian population." *ranger* and *missRanger* use case weights, values that are applied in the bootstrapping process and allow observations with a greater weight to be picked more frequently

when sampling for individual trees. While these types of weighting are not equivalent, the sampling weights were treated as case weights. This was done so that those who represent larger sections of the population would have a more prominent effect than smaller sections. Machine learning methods are still new to epidemiological studies and there is currently no consensus on the best way to incorporate weights. If a new method is created which employs weights during the splitting (i.e., determining which variable provides the greatest reduction in error) or prediction (i.e., calculating the predicted value of each observation) phases of random forests analyses, then these results should be validated against those methods.

4.10 Future Directions

This analysis was conducted with chronic stress, the outcome variable, operationally defined as a response to a single-item measure of chronic stress. However, as this stress measure has yet to be validated, there remains some uncertainty whether chronic stress is truly being measured. As mentioned, future research should validate this stress measure against a gold standard, such as Cohen's Perceived Stress Scale (Cohen et al., 1983). If this measure does not prove to be a valid measure of chronic stress, Statistics Canada should consider replacing this item with the Perceived Stress Scale.

As shown in the work of Epel and colleagues (2018), stress is a vast and complex construct. However, the current data sources are often lacking in important features which could be addressed through a variety of methodological changes. First, common stress correlates (e.g., personality traits, self-efficacy) should be included in future surveys. Second, researchers should consider including the components of chronic stress mentioned in Epel and colleagues' comprehensive review, as there are still several factors that remain unexplored. Stress factors

such as allostatic load, response to an acute stress, and genetic make-up would be impossible to evaluate using survey methodology. Perhaps future studies could employ a method similar to that used in the Midlife in the United States study. In that study, a random subsample of participants were asked to visit one of the testing sites for a two-day intense data collection protocol. This included a complete physical exam, collection of multiple biomarkers following blood draws, a sleep assessment, and the examination of participants' responses to a laboratory-induced challenge (Love, Seeman, Weinstein, & Ryff, 2010). By pairing a similar subsampling method with the traditional survey method, future research can better assess the correlates of stress by collecting more of the factors highlighted by Epel and colleagues. Variables often not included in large-scale chronic stress research could be investigated, including biological markers of allostatic load (e.g., resting cortisol, telomere length) and responses to a stressor (e.g., cortisol and heart rate reactivity).

Finally, an analysis similar to the one used in the present study should be conducted across multiple large national studies. Tara Gruenewald and colleagues (2018) noted that there were similar chronic stress items in national studies of the United States (Health and Retirement Study), England (English Longitudinal Study of Ageing), Ireland (Irish Longitudinal Study of Ageing), Costa Rica (Costa Rican Longevity and Healthy Aging Study), and India (Longitudinal Aging Study in India). The benefits to using data from each of the studies is two-fold. First, it would allow for comparison of international differences in stress correlates. Previous studies have noted cross-national differences in certain aspects of stress, including coping ability (Kirkcaldy & Cooper, 1992), participants' responses to a stressor (Souza-Talarico, Plusquellec, Lupien, Fiocco, & Suchecki, 2014), and number of stressful life events (Vázquez, Panadero, & Martín, 2015). However, no studies using advanced statistical methods have examined the

differences in stress correlates between nations. Identifying these differences may aid in the adaptation of interventions between nations. Second, it would present an opportunity for longitudinal data analyses. To date, the majority of studies in the field of chronic stress and its correlates have been cross-sectional. Owing to this, it is impossible to determine which correlates of stress are antecedents and which are outcomes. Using the above studies, researchers could not only test the directionality of these relationships, but also whether they are unidirectional or reciprocal.

Chapter 5: Conclusion

Stress has been shown to have numerous deleterious effects on health and well-being, including increased risk of chronic disease (Cohen et al., 2007), poor health behaviours (Ng & Jeffery, 2003), and premature mortality (Nielsen et al., 2008). In order to help combat the high rate of stress in Canadians, the most important predictors of stress must be identified. However, there currently is a paucity of multidisciplinary research that highlights the variables that are the most important. Machine learning is a logical tool for exploring the field of stress research due to its ability to analyze large volumes of data. The present study used random forests, a type of machine learning, on the Canadian Community Health Survey - Mental Health. It was found that while many factors are associated with chronic stress, life satisfaction, negative social interactions, and age were the most important. These variables need to be further explored to better understand their relationship with stress, but this study presents the next steps for future studies. If the results of the present study are consistent with the results in repeated measures research, then it can help steer new interventions and public policy.

References

Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Cunningham, J. E., Fulton, J., & Hebert, J. R. (2005). The Effect of Social Desirability and Social Approval on Self-Reports of Physical Activity. *American Journal of Epidemiology*, *161*(4), 389-398. https://doi.org/10.1093/aje/kwi054

Al'Aref, S. J., Anchouche, K., Singh, G., Slomka, P. J., Kolli, K. K., Kumar, A., … Min, J. K. (2018). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European Heart Journal*. https://doi.org/10.1093/eurheartj/ehy404

Almeida, D. M. (2005). Resilience and Vulnerability to Daily Stressors Assessed via Diary Methods. *Current Directions in Psychological Science*, *14*(2), 64-68. https://doi.org/10.1111/j.0963-7214.2005.00336.x

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*(6), 627-636. https://doi.org/10.1016/J.JCLINEPI.2014.12.014

Banks, J., Oldfield, Z., & Smith, J. P. (2012). Childhood Health and Differences in Late-Life Health Outcomes between England and the United States. In D. A. Wise (Ed.), *Investigations in the Economics of Aging* (pp. 321-339). University of Chicago Press. Retrieved from https://www.nber.org/chapters/c12445.pdf

Baumeister, D., Akhtar, R., Ciufolini, S., Pariante, C. M., & Mondelli, V. (2016). Childhood trauma and adulthood inflammation: a meta-analysis of peripheral C-reactive protein, interleukin-6 and tumour necrosis factor-α. *Molecular Psychiatry*, *21*(5), 642-649. https://doi.org/10.1038/mp.2015.67

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*. Public Health Association of Australia Inc. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

Bergman, P. (2018). The number of repeated observations needed to estimate the habitual physical activity of an individual to a given level of precision. *PLOS ONE*, *13*(2), e0192117. https://doi.org/10.1371/journal.pone.0192117

Berk, R. A. (2006). An Introduction to Ensemble Methods for Data Analysis. *Sociological Methods & Research*, *34*(3), 263-295. https://doi.org/10.1177/0049124105283119

Bernstein, M. J., Zawadzki, M. J., Juth, V., Benfield, J. A., & Smyth, J. M. (2018). Social interactions in daily life: Within-person associations between momentary social experiences and psychological and physical health indicators. *Journal of Social and Personal Relationships*. https://doi.org/10.1177/0265407517691366

Birditt, K. S., Fingerman, K. L., & Almeida, D. M. (2005). Age Differences in Exposure and Reactions to Interpersonal Tensions: A Daily Diary Study. https://doi.org/10.1037/0882-7974.20.2.330

Breiman, L. (n.d.). *Manual - Setting Up, Using, And Understanding Random Forests V4.0*. Retrieved from https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf

Breiman, L. (1996). *Bias, Variance, and Arcing Classifiers*. Berkeley, CA. Retrieved from https://www.stat.berkeley.edu/users/breiman/arcall96.pdf

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brenner, P. S., & DeLamater, J. D. (2014). Social Desirability Bias in Self-reports of Physical Activity: Is an Exercise Identity the Culprit? *Social Indicators Research*, *117*(2), 489-504. Retrieved from https://www.jstor.org/stable/24720836?seq=10#metadata_info_tab_contents

Brown, T., Platt, S., & Amos, A. (2014). Equity impact of population-level interventions and policies to reduce smoking in adults: A systematic review. *Drug and Alcohol Dependence*, *138*, 7-16. https://doi.org/10.1016/j.drugalcdep.2014.03.001

Canadian Cancer Society. (2018). Stress - Canadian Cancer Society. Retrieved August 13, 2018, from http://www.cancer.ca/en/cancer-information/cancer-journey/living-with-cancer/stress/?region=on

Canadian Institute for Health Information, Health Canada, & Statistics Canada. (1999). *Health Information Roadmap: Beginning the Journey*. Ottawa. Retrieved from http://www.cihi.ca

Centers for Disease Control and Prevention. (1999). *Achievements in Public Health, 1900-1999: Tobacco Use -- United States, 1900-1999*. Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/mm4843a2.htm

Centers for Disease Control and Prevention. (2019). Current Cigarette Smoking Among Adults in the United States | CDC. Retrieved February 7, 2019, from https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm

Chang, E. C. (2002). *Optimism-Pessimism and Stress Appraisal: Testing a Cognitive Interactive Model of Psychological Adjustment in Adults*. *Cognitive Therapy and Research* (Vol. 26). Retrieved from https://link.springer.com/content/pdf/10.1023%2FA%3A1020313427884.pdf

Cohen, S. (2004). Social relationships and health. *The American Psychologist*, *59*(8), 676-684. https://doi.org/10.1037/0003-066X.59.8.676

Cohen, S., Gianaros, P. J., & Manuck, S. B. (2016). A Stage Model of Stress and Disease. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(4), 456-463. https://doi.org/10.1177/1745691616646305

Cohen, S., & Janicki-Deverts, D. (2012). Who's Stressed? Distributions of Psychological Stress in the United States in Probability Samples from 1983, 2006, and 2009. *Journal of Applied Social Psychology*, *42*(6), 1320-1334. https://doi.org/10.1111/j.1559-1816.2012.00900.x

Cohen, S., Janicki-Deverts, D., & Miller, G. E. (2007). Psychological Stress and Disease. *JAMA*, *298*(14), 1685. https://doi.org/10.1001/jama.298.14.1685

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, *24*(4), 385-396. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6668417

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1), 270. https://doi.org/10.1186/s12859-018-2264-5

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, *132*(20), 1920-1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, *18*(4), 407-422. https://doi.org/10.1016/J.NEUNET.2005.03.007

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3. https://doi.org/10.1186/1471-2105-7-3

Dietterich, T. G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, *40*, 139-157. Retrieved from https://link.springer.com/content/pdf/10.1023%2FA%3A1007607513941.pdf

Dimsdale, J. E. (2008). Psychological stress and cardiovascular disease. *Journal of the American College of Cardiology*, *51*(13), 1237-1246. https://doi.org/10.1016/j.jacc.2007.12.024

Dorn, H. F. (1959). Tobacco consumption and mortality from cancer and other diseases. *Public Health Reports (Washington, D.C. : 1896)*, *74*(7), 581-593. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/13668008

Duffy, K. A., McLaughlin, K. A., & Green, P. A. (2018). Early life adversity and health-risk behaviors: proposed psychological and neural mechanisms. *Annals of the New York Academy of Sciences*, *1428*(1), 151-169. https://doi.org/10.1111/nyas.13928

Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., & Mendes, W. B. (2018). More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology*, *49*, 146-169. https://doi.org/10.1016/J.YFRNE.2018.03.001

Extremera, N., Durán, A., & Rey, L. (2009). The moderating effect of trait meta-mood and perceived stress on life satisfaction. *Personality and Individual Differences*, *47*(2), 116-121. https://doi.org/10.1016/J.PAID.2009.02.007

Fleishman, J. A., Sherbourne, C. D., Crystal, S., Collins, R. L., Marshall, G. N., Kelly, M., … Hays, R. D. (2000). Coping, Conflictual Social Interactions, Social Support, and Mood Among HIV-Infected Persons. *American Journal of Community Psychology*, *28*(4), 421-453. https://doi.org/10.1023/A:1005132430171

Friedman, H. S., & Kern, M. S. (2010). Contributions of Personality to Health Psychology. In J. M. Suls, K. W. Davidson, & R. M. Kaplan (Eds.), *Handbook of Health Psychology and Behavioral Medicine*. New York, NY: The Guildford Press.

Frost Ebstrup, J., Eplov, L. F., Pisinger, C., & Jørgensen, T. (2011). Association between the Five Factor personality traits and perceived stress: is the effect mediated by general self-efficacy? *Anxiety, Stress & Coping*, *24*(4), 407-419. https://doi.org/10.1080/10615806.2010.540012

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics*, *24*(11), 1037-1052. https://doi.org/10.1111/cns.13048

Graham, K., Massak, A., Demers, A., & Rehm, J. (2007). Does the Association Between Alcohol Consumption and Depression Depend on How They Are Measured? *Alcoholism: Clinical and Experimental Research*, *31*(1), 78-88. https://doi.org/10.1111/j.1530-0277.2006.00274.x

Grandvalet, Y. (2004). Bagging Equalizes Influence. *Machine Learning*, *55*(3), 251-270. https://doi.org/10.1023/B:MACH.0000027783.34431.42

Gruenewald, T. L., Crosswell, A. D., Mayer, S., & Lee, J. (2018). *Measures of Stress in the Health and Retirement Study (HRS) and the HRS Family of Studies USER GUIDE*.

Hammen, C. (2005). Stress and Depression. *Annual Review of Clinical Psychology*, *1*(1), 293-319. https://doi.org/10.1146/annurev.clinpsy.1.102803.143938

Hammen, C., Kim, E. Y., Eberhart, N. K., & Brennan, P. A. (2009). Chronic and acute stress and the prediction of major depression in women. *Depression and Anxiety*, *26*(8), 718-723. https://doi.org/10.1002/da.20571

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning*.

Health Canada. (2008). *Mental Health - Coping With Stress The Issue*. Retrieved from www.cpa-apc.org/

Heart and Stroke Foundation. (2018). Get healthy. Retrieved from http://www.heartandstroke.ca/get-healthy

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An Introduction to Statistical Learning: with Applications in R*.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science (New York, N.Y.)*, *353*(6301), 790-794. https://doi.org/10.1126/science.aaf7894

Jerusalem, M., & Schwarzer, R. (2014). *Self-Efficacy: Thought Control of Action*. (R. Schwarzer, Ed.). Taylor & Francis. https://doi.org/10.4324/9781315800820

Juster, R.-P., McEwen, B. S., & Lupien, S. J. (2010). Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neuroscience & Biobehavioral Reviews*, *35*(1), 2-16. https://doi.org/10.1016/j.neubiorev.2009.10.002

Kirkcaldy, B. D., & Cooper, C. L. (1992). Cross-cultural differences in occupational stress among british and german managers. *Work and Stress*. https://doi.org/10.1080/02678379208260352

Kongarchapatara, B., Moschis, G. P., & Sim Ong, F. (2014). Understanding the relationships between age, gender, and life satisfaction: the mediating role of stress and religiosity. *Journal of Beliefs & Values*, *35*(3), 340-358. https://doi.org/10.1080/13617672.2014.980120

Kreek, M. J., Nielsen, D. A., Butelman, E. R., & LaForge, K. S. (2005). Genetic influences on impulsivity, risk taking, stress responsivity and vulnerability to drug abuse and addiction. *Nature Neuroscience*, *8*(11), 1450-1457. https://doi.org/10.1038/nn1583

Krumpal, I. (2011). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity: International Journal of Methodology*, *47*(4), 2025-2047. https://doi.org/10.1007/s11135-011-9640-9

Lazarus, R. S. (1993). From Psychological Stress to the Emotions: A History of Changing Outlooks. *Annual Review of Psychology*, *44*(1), 1-22. https://doi.org/10.1146/annurev.ps.44.020193.000245

Lazarus, R. S., & Folkman, S. (1984). *Stress, Appraisal, and Coping*. Springer Publishing Co.

Lee, J., Kim, E., & Wachholtz, A. (2016). The effect of perceived stress on life satisfaction. *Korean Journal of Youth Studies*, *23*(10), 29. https://doi.org/10.21509/KJYS.2016.10.23.10.29

Lemstra, M., Rogers, M., & Moraros, J. (2015). Income and heart disease: Neglected risk factor. *Canadian Family Physician Medecin de Famille Canadien*, *61*(8), 698-704. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26836056

Levine, M. E., Cole, S. W., Weir, D. R., & Crimmins, E. M. (2015). Childhood and later life stressors and increased inflammatory gene expression at older ages. *Social Science &*

Medicine, *130*, 16-22. https://doi.org/10.1016/j.socscimed.2015.01.030

Lincoln, K. D. (2008). Social Support, Negative Social Interactions, and Psychological Well-Being. *Social Service Review*. https://doi.org/10.1086/514478

Love, G. D., Seeman, T. E., Weinstein, M., & Ryff, C. D. (2010). Bioindicators in the MIDUS National Study: Protocol, Measures, Sample, and Comparative Context. *Journal of Aging and Health*, *22*(8), 1059-1080. https://doi.org/10.1177/0898264310374355

Mabry, P. L., Olster, D. H., Morgan, G. D., & Abrams, D. B. (2008). Interdisciplinarity and Systems Science to Improve Population Health. *American Journal of Preventive Medicine*, *35*(2), S211-S224. https://doi.org/10.1016/j.amepre.2008.05.018

Manne, S. L., Ostroff, J., Winkel, G., Grana, G., & Fox, K. (2005). Partner unsupportive responses, avoidant coping, and distress among women with early stage breast cancer: Patient and partner perspectives. *Health Psychology*, *24*(6), 635-641. https://doi.org/10.1037/0278-6133.24.6.635

Manuel, D., Perez, R., Bennerr, C., Rosella, L., Taljaard, M., Roberts, M., … Manson, H. (2012). *Seven more years: The impact of smoking, alcohol, diet, physical activity and stress on health and life expectancy in Ontario. An ICES/PHO Report.* Toronto.

Matheny, K. B., Roque-Tovar, B. E., & Curlette, W. L. (2008). Perceived stress, coping resources, and life satisfaction among U. S. and Mexican college students: A cross-cultural study. *Anales de Psicología*, *24*(1), 49-57. Retrieved from https://psycnet.apa.org/record/2008-07540-007

McEwen, B. S. (2004). Protection and damage from acute and chronic stress: Allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders. *Annals of the New York Academy of Sciences*, *1032*, 1-7. https://doi.org/10.1196/annals.1314.001

McEwen, B. S., & Stellar, E. (1993). Stress and the Individual. *Archives of Internal Medicine*, *153*(18), 2093. https://doi.org/10.1001/archinte.1993.00410180039004

Nes, L. S., & Segerstrom, S. C. (2006). Dispositional Optimism and Coping: A Meta-Analytic Review. *Personality and Social Psychology Review*, *10*(3), 235-251. https://doi.org/10.1207/s15327957pspr1003_3

Newsom, J. T., Rook, K. S., Nishishiba, M., Sorkin, D. H., & Mahan, T. L. (2005). Understanding the relative importance of positive and negative social exchanges: examining specific domains and appraisals. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *60*(6), P304-P312. https://doi.org/10.1093/geronb/60.6.p304

Ng, D. M., & Jeffery, R. W. (2003). Relationships Between Perceived Stress and Health Behaviors in a Sample of Working Adults. *Health Psychology*, *22*(6), 638-642. https://doi.org/10.1037/0278-6133.22.6.638

Nielsen, N. R., Kristensen, T. S., Schnohr, P., & Gronbaek, M. (2008). Perceived Stress and Cause-specific Mortality among Men and Women: Results from a Prospective Cohort Study. *American Journal of Epidemiology*, *168*(5), 481-491. https://doi.org/10.1093/aje/kwn157

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625-632. https://doi.org/10.1007/s10459-010-9222-y

Ory, M. G., Hoffman, R. R., Yee, J. L., Tennstedt, S., & Schulz, R. (1999). Prevalence and Impact of Caregiving: A Detailed Comparison Between Dementia and Nondementia Caregivers. *The Gerontologist*. https://doi.org/10.1093/geront/39.2.177

Osbourne, J. W., & Waters, E. (2002). Four Assumptions of Multiple Regression That Researchers Should Always Test. *Practical Assessment, Research & Evaluation*, *8*(2), 1-5. Retrieved from https://doaj.org/article/d8f7dc53343b43b0a340f5c2169756b9

Park, J., Kitayama, S., Karasawa, M., Curhan, K., Markus, H. R., Kawakami, N., … Ryff, C. D. (2013). Clarifying the links between social support and health: Culture, stress, and neuroticism matter. *Journal of Health Psychology*, *18*(2), 226-235. https://doi.org/10.1177/1359105312439731

Patten, S. B., Williams, J. V. A., Lavorato, D. H., Berzins, S., Metz, L. M., & Bulloch, A. G. M. (2012). Health Status, Stress and Life Satisfaction in a Community Population with MS. *Canadian Journal of Neurological Science*, (39), : 206-212. https://doi.org/10.1017/S031716710001324X

Pechtel, P., & Pizzagalli, D. A. (2011). Effects of early life stress on cognitive and affective function: an integrated review of human literature. *Psychopharmacology*, *214*(1), 55-70. https://doi.org/10.1007/s00213-010-2009-2

Pinquart, M., & Sörensen, S. (2003). Differences between caregivers and noncaregivers in psychological health and physical health: A meta-analysis. *Psychology and Aging*, *18*(2), 250-267. https://doi.org/10.1037/0882-7974.18.2.250

Rauktis, M. E., Koeske, G. E., & Tereshko, O. (1995). Negative Social Interactions, Distress, and Depression Among Those Caring for a Seriously and Persistently Mentally Ill Relative. *American Journal of Community Psychology*, *23*(2).

Reardon, K. W., Tackett, J. L., & Lynam, D. (2018). The personality context of relational aggression: A Five-Factor Model profile analysis. *Personality Disorders*, *9*(3), 228-238. https://doi.org/10.1037/per0000231

Richardson, S., Shaffer, J. A., Falzon, L., Krupka, D., Davidson, K. W., & Edmondson, D. (2012). Meta-Analysis of Perceived Stress and Its Association With Incident Coronary Heart Disease. *The American Journal of Cardiology*, *110*(12), 1711-1716. https://doi.org/10.1016/j.amjcard.2012.08.004

Rod, N. H., Grønbaek, M., Schnohr, P., Prescott, E., & Kristensen, T. S. (2009). Perceived stress as a risk factor for changes in health behaviour and cardiac risk profile: a longitudinal study. *Journal of Internal Medicine*, *266*(5), 467-475. https://doi.org/10.1111/j.1365-2796.2009.02124.x

Ross, N. (2002). Community belonging and health. *Health Reports*, *13*(3), 82-85.

Schneiderman, N., Ironson, G., & Siegel, S. D. (2005). Stress and health: psychological, behavioral, and biological determinants. *Annual Review of Clinical Psychology*, *1*, 607-628. https://doi.org/10.1146/annurev.clinpsy.1.102803.144141

Scott, S. B., Sliwinski, M. J., & Blanchard-Fields, F. (2013). Age differences in emotional responses to daily stress: the role of timing, severity, and global perceived stress. *Psychology and Aging*, *28*(4), 1076-1087. https://doi.org/10.1037/a0034000

Seligman, B., Tuljapurkar, S., & Rehkopf, D. (2018). Machine learning approaches to the social determinants of health in the health and retirement study. *SSM - Population Health*, *4*, 95-99. https://doi.org/10.1016/j.ssmph.2017.11.008

Selye, H. (1936). A Syndrome produced by Diverse Nocuous Agents. *Nature*, *138*(3479), 32-32. https://doi.org/10.1038/138032a0

Selye, H. (1950). Stress and the general adaptation syndrome. *British Medical Journal*, *1*(4667), 1383-1392. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15426759

Shields, G. S., Toussaint, L. L., & Slavich, G. M. (2016). Stress-related changes in personality: A longitudinal study of perceived stress and trait pessimism. *Journal of Research in Personality*, *64*, 61-68. https://doi.org/10.1016/J.JRP.2016.07.008

Shields, M. (2006). Stress and depression in the employed population. *Health Reports*, *17*(4).

Slopen, N., Kontos, E. Z., Ryff, C. D., Ayanian, J. Z., Albert, M. A., & Williams, D. R. (2013). Psychosocial stress and cigarette smoking persistence, cessation, and relapse over 9-10 years: a prospective study of middle-aged adults in the United States. *Cancer Causes & Control : CCC*, *24*(10), 1849-1863. https://doi.org/10.1007/s10552-013-0262-5

Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods*, *220*(1), 85-91. https://doi.org/10.1016/J.JNEUMETH.2013.08.024

Smyth, J., Zawadzki, M., & Gerin, W. (2013). Stress and Disease: A Structural and Functional Analysis. *Social and Personality Psychology Compass*, *7*(4), 217-227. https://doi.org/10.1111/spc3.12020

Smyth, J., Zawadzki, M., Juth, V., & Sciamanna, C. (2017). Global life satisfaction predicts ambulatory affect, stress, and cortisol in daily life in working adults. *Journal of Behavioral Medicine*, *40*(2), 320-331. https://doi.org/10.1007/s10865-016-9790-2

Souza-Talarico, J. N., Plusquellec, P., Lupien, S. J., Fiocco, A., & Suchecki, D. (2014). Cross-country differences in basal and stress-induced cortisol secretion in older adults. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0105968

Statistics Canada. (2009). *The Canadian Labour Market at a Glance - 2007*. Retrieved from https://www150.statcan.gc.ca/n1/en/pub/71-222-x/71-222-x2008001-eng.pdf?st=Cpdb_4VV

Statistics Canada. (2012). Questionnaire(s) and Reporting guide(s) - Canadian Community Health Survey (CCHS) - Questionnaire for Cycle 1.1 - September, 2000 - November, 2001. Retrieved from http://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=getInstrumentList&Item_Id=33183&UL=1V&

Statistics Canada. (2013). *Canadian Community Health Survey (CCHS) - Mental Health User Guide*.

Statistics Canada. (2014). *Perceived life stress*. Retrieved from http://www.statcan.gc.ca/pub/82-625-x/2015001/article/14188-eng.htm

Steptoe, A., & Kivimäki, M. (2012). Stress and cardiovascular disease. *Nature Reviews Cardiology*, *9*(6), 360-370. https://doi.org/10.1038/nrcardio.2012.45

Steptoe, A., & Kivimäki, M. (2013). Stress and Cardiovascular Disease: An Update on Current Knowledge. *Annual Review of Public Health*, *34*(1), 337-354. https://doi.org/10.1146/annurev-publhealth-031912-114452

Stone, A. A., Schneider, S., & Broderick, J. E. (2017). Psychological stress declines rapidly from age 50 in the United States: Yet another well-being paradox. *Journal of Psychosomatic Research*. https://doi.org/10.1016/j.jpsychores.2017.09.016

Stone, A. A., Schwartz, J. E., Broderick, J. E., & Deaton, A. (2010). A snapshot of the age distribution of psychological well-being in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(22), 9985-9990. https://doi.org/10.1073/pnas.1003744107

Strine, T. W., Chapman, D. P., Balluz, L. S., Moriarty, D. G., & Mokdad, A. H. (2008). The Associations Between Life Satisfaction and Health-related Quality of Life, Chronic Illness, and Health Behaviors among U.S. Community-dwelling Adults. *Journal of Community Health*, *33*(1), 40-50. https://doi.org/10.1007/s10900-007-9066-4

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323-348. https://doi.org/10.1037/a0016973

Stults-Kolehmainen, M. A., & Sinha, R. (2014). The effects of stress on physical activity and exercise. *Sports Medicine*, *44*(1), 81-121. https://doi.org/10.1007/s40279-013-0090-5

Tryon, M. S., Carter, C. S., DeCant, R., & Laugero, K. D. (2013). Chronic stress exposure may affect the brain's response to high calorie food cues and predispose to obesogenic eating habits. *Physiology & Behavior*, *120*, 233-242. https://doi.org/10.1016/J.PHYSBEH.2013.08.010

Vázquez, J. J., Panadero, S., & Martín, R. M. (2015). Regional and national differences in stressful life events: The role of cultural factors, economic development, and gender. *American Journal of Orthopsychiatry*. https://doi.org/10.1037/ort0000029

Worachartcheewan, A., Shoombuatong, W., Pidetcha, P., Nopnithipat, W., Prachayasittikul, V., & Nantasenamat, C. (2015). Predicting Metabolic Syndrome Using the Random Forest Method. *The Scientific World Journal*, *2015*, 581501. https://doi.org/10.1155/2015/581501

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1-17. https://doi.org/10.18637/jss.v077.i01

Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., … Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, *19*(13), 1636-1643. https://doi.org/10.1093/bioinformatics/btg210

Wu, H., & Leung, S.-O. (2017). Can Likert Scales be Treated as Interval Scales? - A Simulation Study. *Journal of Social Service Research*, *43*(4), 527-532. https://doi.org/10.1080/01488376.2017.1329775

Xin, Y., Wu, J., Yao, Z., Guan, Q., Aleman, A., & Luo, Y. (2017). The relationship between personality and the response to acute psychological stress. *Scientific Reports*, *7*(1), 16906. https://doi.org/10.1038/s41598-017-17053-2

Appendices

Appendix A - Variable Recoding

Table 7 - The original and recoded values for the outcome variable, chronic stress, used in the analyses.

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Chronic Stress	GEN_07 / Stress	Thinking about the amount of stress in your life, would you say that most days are...?	NOT AT ALL STRESSFUL	1	1-5 = standardized	-1.65	3,555
			NOT VERY STRESSFUL	2		-0.66	5,675
			A BIT STRESSFUL	3		0.34	9,394
			QUITE A BIT STRESSFUL	4		1.33	3,809
			EXTREMELY STRESSFUL	5		2.33	632
			DON'T KNOW	7	7-9 = NA	NA	24
			REFUSAL	8			
			NOT STATED	9			

Note: Lay name is how the variable is discussed in the paper. CCHS Coding / R Coding represent the variable signifier in the CCHS-MH Dataset and how it is coded in the R script, respectively. Item text denotes how the item was worded in the CCHS-MH, with the original CCHS Response options and their CCHS-MH coding also listed. Recoding and Recoded Responses represent how the items were recoded and the potential options, respectively. Finally, sample represents the number of participants in each recoded response option.

Table 8 - The health behaviour variables used in the analyses.

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Frequency of Drinking	AUD_ 02 / Amt_drink	During the past 12 months, how often did you drink alcoholic beverages?	LESS THAN ONCE A MONTH ONCE A MONTH 2 TO 3 TIMES A MONTH ONCE A WEEK 2 TO 3 TIMES A WEEK 4 TO 6 TIMES A WEEK EVERY DAY NOT APPLICABLE DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 6 7 96 97 98 99	96 = 0 0-7 = standardized 97-99 = NA	-1.18 -0.72 -0.26 0.19 0.65 1.11 1.57 2.02 NA	5,124 4,556 1,707 2,906 2,862 3,307 1,284 1,313 30
WHO alcohol abuse or dependence	AUDDY / abuse_dep_ drink	This grouped variable identifies whether the respondent meets the 12 month CCHS - Mental Health/WHO-CIDI criteria for alcohol abuse or alcohol dependence.	YES NO NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	0 1 NA	22,103 635 351
Level of Insomnia	GEN_04 / sleep_trouble	How often do you have trouble going to sleep or staying asleep?	NONE OF THE TIME A LITTLE OF THE TIME SOME OF THE TIME MOST OF THE TIME ALL OF THE TIME DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 7 8 9	1-5 = standardized 7-9 = NA	-1.05 -0.23 0.59 1.41 2.23 NA	8,208 5,324 5,587 2,495 1,454 21

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Weekly hours of moderate or vigorous physical activity (R)	PHSGAPA / PA_levels_R	This variable indicates the average amount of time in hours that respondents reported doing moderate or vigorous physical activity in the past seven days.	HOURS 14+ HOURS NOT STATED	0 - 13.5 14 99	0 → 0, 0.1-2.5 → 1, 2.6-5.0 → 2, 5.1-7.5 → 3, 7.6-10.0 → 4, 10.1-14 → 5 99 = NA 0-5 = reverse coded and standardized	-2.31 -1.64 -.97 -.31 0.35 1.03 NA	1,515 1,116 2,107 3,937 6,135 6,786 1,493
Smoke Status • Daily Smoker • Occasional Smoker 1 • Occasional Smoker 2 • Former Smoker 1 • Former Smoker 2 • Never Smoked	SMKDSTY / Daily_smk Occ_smk_1 Occ_smk_2 form_smk_1 form_smk_2 no_smk	This variable indicates the type of smoker the respondent is, based on his/her smoking habits.	DAILY SMOKER OCCASIONAL SMOKER (FORMER DAILY SMOKER) ALWAYS AN OCCASIONAL SMOKER FORMER DAILY SMOKER FORMER OCCASIONAL SMOKER NEVER SMOKED NOT STATED	1 2 3 4 5 6 99	1→Daily_smk =1 2→Occ_smk_1 =1 3→ Occ_smk_2=1 4→form_smk_1=1 5→form_smk_2=1 6→no_smk=1	Daily_smk =1 Occ_smk_1 =1 Occ_smk_2=1 form_smk_1=1 form_smk_2=1 no_smk=1	4,172 698 378 6,414 3,623 7,773
WHO drug abuse or dependence	SUDDY / abuse_dep_drug	This grouped variable identifies whether the respondent meets the 12 month CCHS Mental Health/WHO-CIDI criteria for any drug abuse or drug dependence.	YES NO NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	0 1 NA	22,386 353 350

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Illicit Drug Use	SUDDYID / drug_use	This grouped variable identifies whether the respondent meets the 12 month CCHS Mental Health/WHO-CIDI criteria for any drug abuse or drug dependence.	YES NO NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	0 1 NA	21,243 1,611 235

Note: Lay name is how the variable is discussed in the paper. CCHS Coding / R Coding represent the variable signifier in the CCHS-MH Dataset and how it is coded in the R script, respectively. Item text denotes how the item was worded in the CCHS-MH, with the original CCHS Response options and their CCHS-MH coding also listed. Recoding and Recoded Responses represent how the items were recoded and the potential options, respectively. Finally, sample represents the number of participants in each recoded response option.

Table 9 - The mental health variables used in the analyses.

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Bipolar Disorder	BIPDL / bipolar_ disorder	This variable identifies whether respondents meet or fail to meet the CCHS - Mental Health criteria for Bipolar Disorder in the 12 months prior to the interview. The criteria for bipolar disorder 12-month episode are met when the respondent has met the criteria for lifetime bipolar disorder and has had a 12-month episode of bipolar I, bipolar II, or hypomania.	YES NO NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	1 0 NA	22,350 603 136
Anxiety Disorder	CCC_290 / anx_disorder	Do you have an anxiety disorder such as a phobia, obsessive-compulsive disorder or a panic disorder?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	21,690 1,389 10
PTSD	CCC_311 / ptsd_disorder	Do you have post-traumatic stress disorder?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	22,641 425 23
Learning Disability	CCC_331 / learning_ disability_gen	Do you have a learning disability?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	22,301 770 18
Attention Deficit Disorder	CCC_332 / learning_ disability_add	Do you have Attention Deficit Disorder?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	22,545 527 17

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Emotional Impact of Health	DAS_05 / health_ emotions	In the last 30 days, how much have you been emotionally affected by your health problems?	NONE MILD MODERATE SEVERE EXTREME / CANNOT DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 7 8 9	1-5 = standardized 7-9 = NA	-0.48 0.82 2.12 3.43 4.73 NA	17,442 3,219 1,660 508 187 73
Difficulty Concentrating	DAS_06 / diff_ concentration	In the last 30 days, how much difficulty did you have in: ...concentrating on doing something for 10 minutes?	NONE MILD MODERATE SEVERE EXTREME / CANNOT DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 7 8 9	1-5 = standardized 7-9 = NA	-0.33 1.47 3.28 5.09 6.90 NA	20,342 1,692 710 216 56 73
Major Depression	DEPDDY / MDE_occ	This is the final variable that identifies whether respondents meet or fail to meet the CCHS - Mental Health/WHO-CIDI criteria for major depressive episode in the 12 months prior to the interview. Respondents who meet the criteria reported: (1) meeting the criteria for lifetime major depressive episode; (2) having a major depressive episode in the 12 months prior to the interview; and (3) clinically significant distress or impairment in social, occupational or other important areas of functioning.	YES NO NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	0 1 NA	21,754 1,182 153

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Suicidal Thoughts	DEPFSLT / sui_thoughts	This variable classifies the respondent based on whether he/she ever thought about committing suicide or taking his/her own life.	YES NO NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	0 1 NA	19,870 3,152 67
Generalized Anxiety Disorder	GADDDY / GAD_occ	This is the final variable that identifies whether respondents meet or fail to meet the CCHS - Mental Health/WHO-CIDI criteria for Generalized Anxiety Disorder in the 12 months prior to the interview. Respondents who meet the criteria reported: (1) meeting the CCHS - Mental Health/WHO-CIDI criteria for lifetime Generalized Anxiety Disorder; (2) having an episode of generalized anxiety lasting at least six months in the 12 months prior to the interview; and (3) clinically significant distress or impairment in social, occupational or other important areas of functioning.	YES NO NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	0 1 NA	22,169 700 220
Life Satisfaction (R)	GEN_02A2 / life_ satisfaction	Using a scale of 0 to 10 where 0 means "Very dissatisfied" and 10 means "Very satisfied", how do you feel about your life as a whole right now?	Life Satisfaction Score DON'T KNOW REFUSAL NOT STATED	0-10 97 98 99	0-10 = reverse coded and standardized 97-99 = NA	-1.23 - 4.63 NA	22,985 104

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Hypomanic	HYPDEY / hypomanic_ occ	This is the final variable that identifies whether the respondent meets or fails to meet the CCHS - Mental Health/WHO-CIDI criteria for Hypomanic episode in the 12 months prior to the interview	YES NO NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	0 1 NA	22,767 197 125
Mania	MIADEY / manic_occ	This is the final variable that identifies whether respondents meet or fail to meet the CCHS Mental Health/WHO-CIDI 2002 criteria for manic episode in the 12 months prior to the interview. Respondents who meet the criteria report: (1) meeting the criteria for lifetime manic episode; (2) having a manic episode in the 12 months prior to the interview; and (3) clinically significant distress or impairment in social, occupational or other important areas of functioning.	YES NO NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	0 1 NA	22,742 231 116
Coping with Crisis	STS_1 / coping_crisis	In general, how would you rate your ability to handle unexpected and difficult problems, for example, a family or personal crisis? Would you say your ability is...?	EXCELLENT VERY GOOD GOOD FAIR POOR DON'T KNOW REFUSAL	1 2 3 4 5 7 8	1-5 = standardized 7, 8 = NA	-1.43 -0.38 0.66 1.71 2.76 NA	4,280 9,575 6,616 2,018 533 67

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Coping on a Daily Basis	STS_2 / coping_daily	In general, how would you rate your ability to handle the day-to-day demands in your life, for example, handling work, family and volunteer responsibilities? Would you say your ability is…?	EXCELLENT VERY GOOD GOOD FAIR POOR DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 7 8 9	1-5 = standardized 7, 8 = NA	-1.38 -0.23 0.92 2.07 3.22 NA	4,915 10,540 6,021 1,256 263 94
Stress source: • Time • Physical Health • Emotional Health • Money • Work • School • Caregiving • Family • Personal Relationship • Discrimination • Safety • Health of Family • Other • None • Loss	STS_3 / Stress_source_ time Stress_source_ phys_ health Stress_source_ emo_health Stress_source_ money Stress_source_ work Stress_source_ school Stress_source_ caregiving Stress_source_ family	Thinking about stress in your day-to-day life, what would you say is the most important thing contributing to feelings of stress you may have?	TIME PRESSURES / NOT ENOUGH TIME OWN PHYSICAL HEALTH PROBLEM OR CONDITION OWN EMOTIONAL/ MENTAL HEALTH PROBLEM FINANCIAL SITUATION OWN WORK SITUATION SCHOOL EMPLOYMENT STATUS CARING FOR - OWN CHILDREN CARING FOR - OTHERS	1 2 3 4 5 6 7 8 9	1 → Stress_ source_ time = 1 2 → Stress_source_ phys_health = 1 3→ Stress_source_ emo_health = 1 4 → Stress_source_ money = 1 5 or 7 → Stress_source_ work = 1 6 → Stress_source_ school = 1 8 or 9 → Stress_source_ caregiving = 1 10 → Stress_source_ family = 1 11→ Stress_source_ personal = 1	Stress_source_ time = 1 Stress_source_ phys_health = 1 Stress_source_ emo_health = 1 Stress_source_ money = 1 Stress_source_ work = 1 Stress_source_ school = 1 Stress_source_ caregiving = 1 Stress_source_ family = 1 Stress_source_ personal = 1	1,692 1,816 619 3,170 4,805 514 898 1,336 1,214

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
	Stress_source_ personal Stress_source_ discrim Stress_source_ safety Stress_source_ family_health Stress_source_ loss		OTHER PERSONAL OR FAMILY RESPONSIBILITY	10	12 → Stress_source_ discrim = 1 13 → Stress_source_ safety = 1 14 → Stress_source_ family_health = 1 17 → Stress_source_ loss = 1	Stress_source_ discrim = 1 Stress_source_ safety = 1 Stress_source_ family_health = 1 Stress_source_ loss = 1	51 501 1,680 82
			PERSONAL RELATIONSHIPS	11			
			DISCRIMINATION	12			
			PERSONAL AND FAMILY'S SAFETY	13			
			HEALTH OF FAMILY MEMBERS	14			
			OTHER	15			
			NOTHING	16			
			LOSS OF LOVED ONE	17			
			DON'T KNOW	97			
			REFUSAL	98			
			NOT STATED	99			

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Coping Skill	STS_5 / coping_ source	When faced with this source of stress, you have the personal ability to deal with the situation. Do you...?	STRONGLY AGREE AGREE NEITHER AGREE NOR DISAGREE DISAGREE STRONGLY DISAGREE NOT APPLICABLE DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 6 7 8 9	1-5 = standardized 7, 8, 9 = NA	-1.12 0.20 1.51 2.82 4.14 NA	6,231 10,998 1,344 694 156 3666

Note: Lay name is how the variable is discussed in the paper. CCHS Coding / R Coding represent the variable signifier in the CCHS-MH Dataset and how it is coded in the R script, respectively. Item text denotes how the item was worded in the CCHS-MH, with the original CCHS Response options and their CCHS-MH coding also listed. Recoding and Recoded Responses represent how the items were recoded and the potential options, respectively. Finally, sample represents the number of participants in each recoded response option.

Table 10 - The physical health variables used in the analyses.

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Asthma	CCC_031 / conditions_ asthma	Do you have asthma?	YES NO DON'T KNOW REFUSAL	1 2 7 8	1 = 1 2 = 0 7, 8 = NA	0 1 NA	21,138 1,941 10
Arthritis	CCC_051 / conditions_ arth	Do you have arthritis, excluding fibromyalgia?	YES NO DON'T KNOW REFUSAL	1 2 7 8	1 = 1 2 = 0 7, 8 = NA	0 1 NA	17,626 5,443 20
Back Problems	CCC_061 / conditions_ back	Do you have back problems, excluding fibromyalgia and arthritis?	YES NO DON'T KNOW REFUSAL	1 2 7 8	1 = 1 2 = 0 7, 8 = NA	0 1 NA	18,050 5,029 10
High Blood Pressure	CCC_071 / conditions_ high_bp	Do you have high blood pressure?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	17,581 5,474 34
Migraines	CCC_081 / conditions_ migraine	Do you have migraine headaches?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	20,687 2,391 11
Diabetes	CCC_101 / conditions_ diabetes	Do you have diabetes?	YES NO DON'T KNOW REFUSAL	1 2 7 8	1 = 1 2 = 0 7, 8 = NA	0 1 NA	21,044 2,038 7

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Heart Disease	CCC_121 / conditions_ heart_disease	Do you have heart disease?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	21,462 1,600 27
Current Cancer	CCC_131 / conditions_ cancer_current	Do you have cancer	YES NO DON'T KNOW REFUSAL	1 2 7 8	1 = 1 2 = 0 7, 8 = NA	0 1 NA	22,514 562 13
Previous Cancer	CCC_132 / conditions_ cancer_past	Have you ever been diagnosed with cancer?	YES NO NA DON'T KNOW REFUSAL NOT STATED	1 2 6 7 8 9	1 = 1 2 = 0 6-9 = NA	0 1 NA	20,962 1,557 570
Stroke	CCC_151 / conditions_ stroke	Do you suffer from the effects of a stroke?	YES NO DON'T KNOW NOT STATED	1 2 7 9	1 = 1 2 = 0 7, 9 = NA	0 1 NA	22,692 383 14
Bowel Disorders	CCC_171 / conditions_ bowel	Has bowel disorder/Crohn's Disease/ulcerative colitis	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	21,684 1,388 17
Chronic Fatigue	CCC_251 / conditions_ fatigue	Do you have chronic fatigue syndrome?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	22,694 377 18

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Difficulty Walking	DAS_07 / diff_walking	In the last 30 days, how much difficulty did you have in walking a long distance such as a kilometre (or 0.6 miles)?	NONE MILD MODERATE SEVERE EXTREME / CANNOT DO DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 7 8 9	1-5 = standardized 7, 8, 9 = NA	-0.45 0.42 1.28 2.15 3.01 NA	17,983 1,374 1,183 741 1,664 144
Difficulty Standing	DASG01 / diff_standing	In the last 30 days, how much difficulty did you have in standing for long periods such as 30 minutes?	NONE MILD MODERATE SEVERE / EXTREME / CANNOT DO NOT STATED	1 2 3 4 9	1-4 = standardized 9 = NA	-0.47 0.69 1.84 2.99 NA	18,010 1,872 1,604 1,556 47
Difficulty Household Responsibilities	DASG02 / diff_ household	In the last 30 days, how much difficulty did you have in taking care of your household responsibilities?	NONE MILD MODERATE SEVERE / EXTREME / CANNOT DO NOT STATED	1 2 3 4 9	1-4 = standardized 9 = NA	-0.42 0.98 2.39 3.80 NA	18,777 2,093 1,453 704 62

Note: Lay name is how the variable is discussed in the paper. CCHS Coding / R Coding represent the variable signifier in the CCHS-MH Dataset and how it is coded in the R script, respectively. Item text denotes how the item was worded in the CCHS-MH, with the original CCHS Response options and their CCHS-MH coding also listed. Recoding and Recoded Responses represent how the items were recoded and the potential options, respectively. Finally, sample represents the number of participants in each recoded response option.

Table 11 - The social factor variables used in the analyses.

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Difficulty In Community Activities	DAS_04 / diff_comm_ activities	In the last 30 days, how much of a problem did you have joining in community activities (for example, festivities, religious or other activities) in the same way as anyone else can?	NONE MILD MODERATE SEVERE EXTREME/ CANNOT DO NOT APPLICABLE DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 6 97 98 99	1-5 = standardized 6, 97-99 = NA	-0.31 1.12 2.55 3.98 5.42 NA	18,990 970 834 334 283 1,678
Difficulty with New People	DAS_10 / diff_new_ people	In the last 30 days, how much difficulty did you have in dealing with people you do not know?	NONE MILD MODERATE SEVERE EXTREME/ CANNOT DO DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 97 98 99	1-5 = standardized 97-99 = NA	-0.27 1.87 4.01 6.14 8.29 NA	21,136 1,205 494 138 46 70
Difficulty Maintaining Friendship	DAS_11 / diff_ maintain_ friend	In the last 30 days, how much difficulty did you have in maintaining a friendship?	NONE MILD MODERATE SEVERE EXTREME/ CANNOT DO DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 97 98 99	1-5 = standardized 97-99 = NA	-0.21 2.32 4.86 7.39 9.93 NA	21,834 690 345 104 40 76

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Community Belonging	GEN_10 / community_ belonging	How would you describe your sense of belonging to your local community? Would you say it is...?	VERY STRONG SOMEWHAT STRONG SOMEWHAT WEAK VERY WEAK DON'T KNOW REFUSAL NOT STATED	1 2 3 4 7 8 9	1-4 = standardized 7-9 = NA	-1.46 -0.31 0.84 1.97 NA	4,203 10,481 5,917 2,315 173
SPS - Total Score (R)	SPSDCON / sps_total	This variable is used to measure the overall score for the Social Provisions Scale. The range is 10-40, where a higher score reflects a higher level of perceived social support.	SOCIAL PROVISIONS SCALE NOT STATED	10-40 99	10-40 = reverse coded and standardized 99 = NA	-0.95-5.84 NA	22,250 839
Coping - Social Support*	STS_4 / coping_ social_ support	When faced with this source of stress, you can count on people that you know to help you deal with the situation. Do you...?	STRONGLY AGREE AGREE NEITHER AGREE NOR DISAGREE DISAGREE STRONGLY DISAGREE NOT APPLICABLE DON'T KNOW REFUSAL NOT STATED	1 2 3 4 5 6 7 8 9	1-5 = standardized 6-9 = NA	-0.97 0.04 1.06 2.07 3.08 NA	6,930 8,660 1,835 1,456 531 3,677
Negative Social Interactions	NSIDSC / neg_social_ interaction	This variable creates a scale from 0 - 12 in which higher scores indicate a greater amount of negative social interactions.	NEGATIVE SOCIAL INTERACTIONS SCALE NOT STATED	0-12 99	0-12 = standardized 99 = NA	-1.13-4.09 0.04	22,806 283

Note: Lay name is how the variable is discussed in the paper. CCHS Coding / R Coding represent the variable signifier in the CCHS-MH Dataset and how it is coded in the R script, respectively. Item text denotes how the item was worded in the CCHS-MH, with the original CCHS Response options and their CCHS-MH coding also listed. Recoding and Recoded Responses represent how the items were recoded and the potential options, respectively. Finally, sample represents the number of participants in each recoded response option.

*** Not asked to those who report no major source of stress**

Table 12 - The life adversity variables used in the analyses.

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
ELA - Sum Score	CEXDNUM / ela_total	This variable indicates how many of the six assessed types of child maltreatment were experienced at least once.	NUMBER OF TYPES NOT APPLICABLE NOT STATED	0-6 96 99	0-6 = standardized 96, 99 = NA	-0.71 - 3.45 NA	22,562 527
Recent Life Events - Victim of a Crime	CWP_03 / rla_police_ victim	During the past 12 months, did you come into contact with the police as a victim of a crime?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	22,153 892 44
Recent Life Events - Witness a Crime	CWP_04 / rla_police_ witness	During the past 12 months, did you come into contact with the police as a witness to a crime?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	22,266 780 43
Recent Life Events - Family Problems	CWP_07 / rla_police_ family	During the past 12 months, did you come into contact with the police for reasons related to a family member's problems with their emotions, mental health or use of alcohol or drugs?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	22,567 476 46
Recent Life Events - Unmet needs	PNCDNEED / rla_needed_ help_grouped	This variable is a summary classification of the respondent's perceived need for mental health care in the past 12 months. Respondents are grouped into one of four categories based on whether a need was reported (information, medication, counselling, other), and if so, whether their needs were met, partially met, or unmet.	NO PERCEIVED NEED ALL PERCEIVED NEEDS MET PERCEIVED NEEDS PARTIALLY MET PERCEIVED NEEDS NOT MET NOT STATED	1 2 3 4 9	1-2 = 0 3-4 = 1 9 = NA	0 1 NA	21,506 1,459 124

Note: Lay name is how the variable is discussed in the paper. CCHS Coding / R Coding represent the variable signifier in the CCHS-MH Dataset and how it is coded in the R script, respectively. Item text denotes how the item was worded in the CCHS-MH, with the original CCHS Response options

and their CCHS-MH coding also listed. Recoding and Recoded Responses represent how the items were recoded and the potential options, respectively. Finally, sample represents the number of participants in each recoded response option.

Table 13 - The demographic variables used in the analyses.

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Sex (Female = 1)	DHH_SEX / female	Enter the respondent's sex. If necessary, ask: Is respondent male or female?	MALE FEMALE	1 2	1 = 0 2 = 1	0 1	10,373 12,716
Age (R)	DHHGAGE / age	This variable indicates the age of the selected respondent.	15 TO 19 YEARS 20 TO 24 YEARS 25 TO 29 YEARS 30 TO 34 YEARS 35 TO 39 YEARS 40 TO 44 YEARS 45 TO 49 YEARS 50 TO 54 YEARS 55 TO 59 YEARS 60 TO 64 YEARS 65 TO 69 YEARS 70 TO 74 YEARS 75 TO 79 YEARS 80 YEARS OR OLDER	1 2 3 4 5 6 7 8 9 10 11 12 13 14	1-14 = reverse coded and standardized	-1.71 -1.45 -1.20 -0.95 -0.69 -0.44 -0.19 0.07 0.32 0.58 0.83 1.08 1.34	1,583 1,184 1,432 1,918 2,206 2,245 1,956 1,670 1,691 1,729 1,869 1,617 1,989
Household Size	DHHGHSZ / household_size	This variable is derived by sorting the household roster dataset by SAMPLEID and PERSONID and by counting the number of PERSONIDs within each SAMPLEID. DHHGHSZ is a grouping of DHHDHSZ.	1 PERSON 2 PERSONS 3 PERSONS 4 PERSONS 5 OR MORE PERSONS NOT STATED	1 2 3 4 5 9	1-5 = standardized 9 = NA	-1.08 -0.24 0.60 1.44 2.28 NA	7,413 9,197 2,918 2,468 1,089 4

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Dwelling Type • House • Apartment • Other	DHHGDWE / dwelling_ type_house dwelling_ type_apt dwelling_type_ other	DHHDDWE indicates the type of dwelling the respondent lives in, according to the answer given either on the phone or face-to-face.	SINGLE DETACHED APARTMENT OTHER NOT STATED	1 2 3 9	1 → dwelling_ type_house = 1 2 → dwelling_ type_apt = 1 3 → dwelling_ type_other = 1	dwelling_type house = 1 dwelling_type_ apt = 1 dwelling_type_ other = 1	14,304 5,185 3,596
Household Type • No Partner (no kids) • Partner (no kids) • No partner (kids) • Partner (kids) • Other	DHHGLVG / household_ type_ unattached household_ type_partner_ no_kids household_ type_no_ partner_kids household_ type_partner_ kids household_ type_other	The necessary data is collected using a set of relationship codes that define a link between each person in a household. All relationships with the selected respondent within each sample (relationship of selected respondent to each other person within the household) are used in creating this variable.	UNATTACHED INDIVIDUAL LIVING ALONE UNATTACHED INDIVIDUAL LIVING WITH OTHERS LIVING WITH SPOUSE / PARTNER PARENT LIVING W/SPOUSE/ PARTN., CHILDREN SINGLE PARENT LIVING WITH CHILDREN CHILD LIVING W/ONE PARENT W/WO SIBLINGS CHILD LIVING W/TWO PARENTS W/WO SIBLINGS OTHER NOT STATED	1 2 3 4 5 6 7 8 99	DHHGLVG = 1, 2 → household_ type_ unattached = 1 DHHGLVG = 3 → household_ type_ partner_no_kids = 1 DHHGLVG = 5,6 → household_ type_ no_partner_kids = 1 DHHGLVG = 4,7 → household_ type_ partner_kids = 1 DHHGLVG = 8 → household_ type_ other = 1	household_ type_unattached = 1 household_ type_partner_ no_kids = 1 household_ type_no_partner_ kids = 1 household_ type_partner_kids = 1 household_ type_ other = 1	8397 7,509 1,507 4,771 831

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Marital Status • Married • Single • No Longer Married	DHHGMS / marital_ married marital_single marital_no_ longer_married	This variable indicates the marital status for the selected respondent	MARRIED COMMON-LAW WIDOWED DIVORCED/ SEPARATED SINGLE NOT STATED	1 2 3 4 5 9	1, 2 → marital_married = 1 5 → marital_ single = 1 3,4 → marital_no_ longer_married = 1	marital_married = 1 marital_single = 1 marital_no_ longer_married = 1	12,534 5,354 5,143
No Post-Secondary Education	EDUDR04 / no_post_ secondary_ eduction	This variable indicates the highest level of education acquired by the respondent.	> SECONDARY SCHOOL GRADUATION SECONDARY SCHOOL GRADUATION SOME POST-SECONDARY POST-SECONDARY GRADUATION NOT STATED	1 2 3 4 9	1-3 = 1 4 = 0 9 = NA	0 1 NA	13,882 9,100 107
Self-Perceived Health	GEN_01 / self_health	In general, would you say your health is...?	EXCELLENT VERY GOOD GOOD FAIR POOR DON'T KNOW REFUSAL	1 2 3 4 5 7 8	1-5 = standardized 7,8 = NA	-1.38 -0.39 0.60 1.59 2.58 NA	4430 8564 6867 2450 771 7
Employment*	GEN_08 / prev_yr_ employment	Have you worked at a job or business at any time in the past 12 months?	YES NO NOT APPLICABLE REFUSAL NOT STATED	1 2 6 8 9	1 = 1 2 = 0 6, 8, 9 = NA	0 1 NA	5,741 14,574 2,774

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Province of Residence • Newfoundland and Labrador • PEI • Nova Scotia • New Brunswick • Quebec • Ontario • Manitoba • Saskatchewan • Alberta • BC	GEO_PRV / prov_nfl prov_pei prov_ns prov_nb prov_qc prov_on prov_man prov_sask prov_ab prov_bc	Province of residence of respondent	NEWFOUNDLAND AND LABRADOR PRINCE EDWARD ISLAND NOVA SCOTIA NEW BRUNSWICK QUEBEC ONTARIO MANITOBA SASKATCHEWAN ALBERTA BRITISH COLUMBIA	10 11 12 13 24 35 46 47 48 59	10 → prov_nfl = 1 11 → prov_pei = 1 12 → prov_ns = 1 13 → prov_nb = 1 24 → prov_qc = 1 35 → prov_on = 1 46 → prov_man = 1 47 → prov_sask = 1 48 → prov_ab = 1 59 → prov_bc = 1	prov_nfl = 1 prov_pei = 1 prov_ns = 1 prov_nb = 1 prov_qc = 1 prov_on = 1 prov_man = 1 prov_sask = 1 prov_ab = 1 prov_bc = 1	1,291 1,016 1,587 1,540 4,037 5,023 1,662 1,557 2,542 2,834
Income (Relative to Province)	INCDRPR / income_prov	This derived variable is a distribution of residents of each province in deciles (ten categories including approximately the same percentage of residents for each province) based on their value for INCDADR, ie. the adjusted ratio of their total household income to the low income cut-off corresponding to their household and community size. It provides, for each respondent, a relative measure of their household income to the household incomes of all other respondents in the same province.	DECILE 1 DECILE 2 DECILE 3 DECILE 4 DECILE 5 DECILE 6 DECILE 7 DECILE 8 DECILE 9 DECILE 10 NOT STATED	1 2 3 4 5 6 7 8 9 10 99	1-10 = reverse coded and standardized 99 = NA	-1.65 -1.31 -0.96 -0.61 -0.27 0.08 0.43 0.78 1.12 1.47 NA	2,092 2,023 2,267 2,155 2,194 2,337 2,212 2,550 2,685 2,565 9

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Occupation Category • Management • Finance • Sales • Trades • Manufacture	LBSGSOC / occupation_ management occupation_ finance occupation_ sales_services occupation_ trades occupation_ manufacturing	This variable groups the occupation classification of the respondent.	Occupations relating to Management, Natural and Applied Sciences , Health, Social Sciences, Education, Religion, Art, Culture and Recreation Occupations relating to Business, Finance, Administration Occupations relating to Sales and Service Occupations relating to Trades, Transport and Equipment Operator Occupations Unique to Primary Industry, Processing, Manufacturing and Utilities Respondent did not work at a job or business in the past year or age was out of range. Not Stated	1 2 3 4 5 6 9	LBSGSOC = 1 → occupation_ management = 1 LBSGSOC = 2 → occupation_ finance = 1 LBSGSOC = 3 → occupation_ sales_services = 1 LBSGSOC = 4 → occupation_ trades = 1 LBSGSOC = 5 → occupation_ manufacturing = 1	occupation_ management = 1 occupation_ finance = 1 occupation_ sales_services = 1 occupation_ trades = 1 occupation_ manufacturing = 1	5,213 2,164 2,888 1,934 860
Student Status	SDC_8 / Student	Are you currently attending a school, college, cegep or university?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	1 = 1 2 = 0 7-9 = NA	0 1 NA	21,434 1,582 73

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Recoding	Recoded Responses	Sample
Body Mass Index	HWTGBMI / bmi	The body mass index (BMI) is calculated for persons 20 to 64 years old, excluding pregnant women. BMI values have been regrouped to a minimum of 14 and a maximum of 58. Since BMI classification for respondents less than 18 is different than adults, BMI data of 15-19 year olds have been suppressed.	BMI - SELF-REPORT NA NOT STATED	14 - 55 999.96 999.99	14-55 = standardized >999 = NA	-2.32-5.21 NA	22,308 781
Immigrant Status	SDCFIMM / imm_status	This variable indicates if the respondent is an immigrant.	YES NO NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	0 1 NA	18,909 4,046 134
Minority Status	SDCGCGT / minority_status	This variable indicates the cultural or racial background of the respondent. Since the middle of cycle 3.1, this variable excludes all respondents who identify as aboriginal in SDC_41. (The exclusion of aboriginals from this variable was introduced in the middle of cycle 3.1 to align with Census 2006 procedures).	WHITE NON-WHITE (ABORIGIN. OR OTHER VIS. MIN.) NOT STATED	1 2 9	1 = 1 2 = 0 9 = NA	0 1 NA	19,372 3623 94

Note: Lay name is how the variable is discussed in the paper. CCHS Coding / R Coding represent the variable signifier in the CCHS-MH Dataset and how it is coded in the R script, respectively. Item text denotes how the item was worded in the CCHS-MH, with the original CCHS Response options

and their CCHS-MH coding also listed. Recoding and Recoded Responses represent how the items were recoded and the potential options, respectively. Finally, sample represents the number of participants in each recoded response option.

***Not asked to those older than 75 years**

Appendix B - Omitted Variables

Table 14 - The variables which were originally planned to be a part of the analysis and the reason they were omitted.

Lay name	CCHS Coding / R Coding	Item text	CCHS Response options	Code	Number of Responses	Reason for exclusion
Income (Relative to Canada)	INCDRCA/ income_can	This derived variable is a distribution of respondents in deciles (ten categories including approximately the same percentage of residents for each province) based on their value for INCDADR, ie. the adjusted ratio of their total household income to the low income cut-off corresponding to their household and community size. It provides, for each respondent, a relative measure of their household income to the household incomes of all other respondents.	DECILE 1 DECILE 2 DECILE 3 DECILE 4 DECILE 5 DECILE 6 DECILE 7 DECILE 8 DECILE 9 DECILE 10 NOT STATED	1 2 3 4 5 6 7 8 9 10 99	2,614 2,970 2,653 2,565 2,508 2,524 2,234 2,405 2,354 2,272 14	High correlation with provincial income
Pulmonary Disease	CCC_091 / conditions_ pulmonary	Do you have chronic bronchitis, emphysema or chronic obstructive pulmonary disease or COPD?	YES NO DON'T KNOW REFUSAL NOT STATED	1 2 7 8 9	954 16,648 7,499 10 2	Only asked of those who were over 35 years old

Lay name is how the variable is discussed in the paper. CCHS Coding / R Coding represent the variable signifier in the CCHS-MH Dataset and how it is coded in the R script, respectively. Item text denotes how the item was worded in the CCHS-MH, with the original CCHS Response options, CCHS-MH coding, and the number of responses for each option also listed. Finally, there is the reason the item was omitted.

Appendix C - Missing Data

Table 15 - Missingness of data for each variable in the analysis

Lay Name	Missing (%)	Category
Coping - Social Support	15.93	Social Factors
Coping Skill	15.88	Mental Health
Employment	12.01	Demographics
Difficulty In Community Activities	7.27	Social Factors
Weekly hours of MVPA	6.47	Health Behaviour
SPS - Total Score (R)	3.63	Social Factors
Body Mass Index	3.38	Demographics
Previous Cancer	2.47	Physical Health
ELA - Sum Score	2.28	Life Adversity
WHO alcohol abuse or dependence	1.52	Health Behaviour
WHO drug abuse or dependence	1.52	Health Behaviour
Negative Social Interactions	1.23	Social Factors
Illicit Drug Use	1.02	Health Behaviour
Generalized Anxiety Disorder	0.95	Mental Health
Community Belonging	0.75	Social Factors
Major Depression	0.66	Mental Health
Difficulty Walking	0.62	Physical Health
Bipolar Disorder	0.59	Mental Health
Immigrant Status	0.58	Demographics
Hypomanic	0.54	Mental Health
RLA - Unmet needs	0.54	Life Adversity
Mania	0.50	Mental Health
No Post-Secondary Education	0.46	Demographics
Life Satisfaction (R)	0.45	Mental Health
Coping on a Daily Basis	0.41	Mental Health
Minority Status	0.41	Demographics
Difficulty Maintaining Friendship	0.33	Social Factors
Household Type - No partner (kids)	0.32	Demographics
Household Type - Other	0.32	Demographics
Household Type - Partner (kids)	0.32	Demographics
Household Type - Partner (no kids)	0.32	Demographics
Household Type - No Partner (no kids)	0.32	Demographics
Difficulty Concentrating	0.32	Mental Health
Emotional Impact of Health	0.32	Mental Health
Student Status	0.32	Demographics
Difficulty with New People	0.30	Social Factors
Coping with Crisis	0.29	Mental Health

Suicidal Thoughts	0.29	Mental Health
Difficulty Household Responsibilities	0.27	Physical Health
Marital Status - Married	0.25	Demographics
Marital Status - No Longer Married	0.25	Demographics
Marital Status - Single	0.25	Demographics
Difficulty Standing	0.20	Physical Health
RLA - Family Problems	0.20	Life Adversity
RLA - Victim of a Crime	0.19	Life Adversity
RLA - Witness a Crime	0.19	Life Adversity
High Blood Pressure	0.15	Physical Health
Daily Smoker	0.13	Health Behaviour
Former Smoker 1	0.13	Health Behaviour
Former Smoker 2	0.13	Health Behaviour
Never Smoked	0.13	Health Behaviour
Occasional Smoker 1	0.13	Health Behaviour
Occasional Smoker 2	0.13	Health Behaviour
Frequency of Drinking	0.13	Health Behaviour
Heart Disease	0.12	Physical Health
Stress	0.10	Outcome
PSTD	0.10	Mental Health
Level of Insomnia	0.09	Health Behaviour
Arthritis	0.09	Physical Health
Chronic Fatigue	0.08	Physical Health
Learning Disability	0.08	Mental Health
Bowel Disorders	0.07	Physical Health
Attention Deficit Disorder	0.07	Mental Health
Stroke	0.06	Physical Health
Current Cancer	0.06	Physical Health
Migraines	0.05	Physical Health
Anxiety Disorder	0.04	Mental Health
Asthma	0.04	Physical Health
Back Problems	0.04	Physical Health
Income (Provincial) (R)	0.04	Demographics
Diabetes	0.03	Physical Health
Self-Perceived Health	0.03	Demographics
Dwelling Type - Apartment	0.02	Demographics
Dwelling Type - House	0.02	Demographics
Dwelling Type - Other	0.02	Demographics
Household Size	0.02	Demographics
Age (R)	0.00	Demographics
Sex (Female = 1)	0.00	Demographics

Occupation Category - Finance	0.00	Demographics
Occupation Category - Management	0.00	Demographics
Occupation Category - Manufacture	0.00	Demographics
Occupation Category - Sales	0.00	Demographics
Occupation Category - Trades	0.00	Demographics
Province - Alberta	0.00	Demographics
Province - BC	0.00	Demographics
Province - Manitoba	0.00	Demographics
Province - New Brunswick	0.00	Demographics
Province - Newfoundland and Labrador	0.00	Demographics
Province - Nova Scotia	0.00	Demographics
Province - Ontario	0.00	Demographics
Province - PEI	0.00	Demographics
Province - Quebec	0.00	Demographics
Province - Saskatchewan	0.00	Demographics
Stress Source - Caregiving	0.00	Mental Health
Stress Source - Discrimination	0.00	Mental Health
Stress Source - Emotional Health	0.00	Mental Health
Stress Source - Family	0.00	Mental Health
Stress Source - Health of Family	0.00	Mental Health
Stress Source - Loss	0.00	Mental Health
Stress Source - Money	0.00	Mental Health
Stress Source - Personal Relationship	0.00	Mental Health
Stress Source - Physical Health	0.00	Mental Health
Stress Source - Safety	0.00	Mental Health
Stress Source - School	0.00	Mental Health
Stress Source - Time	0.00	Mental Health
Stress Source - Work	0.00	Mental Health

Appendix D - Variable Correlations

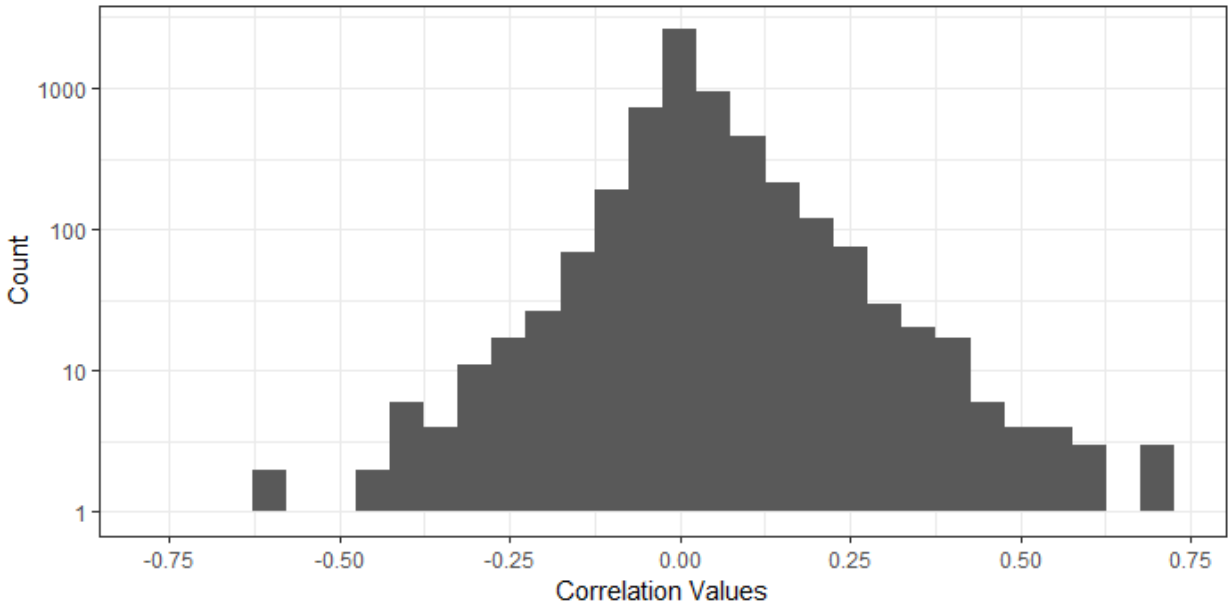

Figure 6 - Distribution of correlations for all variables.

Note: While the majority of correlation coefficients fall in the very small or small effects (N = 10,808), there are hundreds that are greater than 0.25 and less than -0.25 (N = 322). This suggests that linear regression may not be an appropriate method of analysis due to multicollinearity.

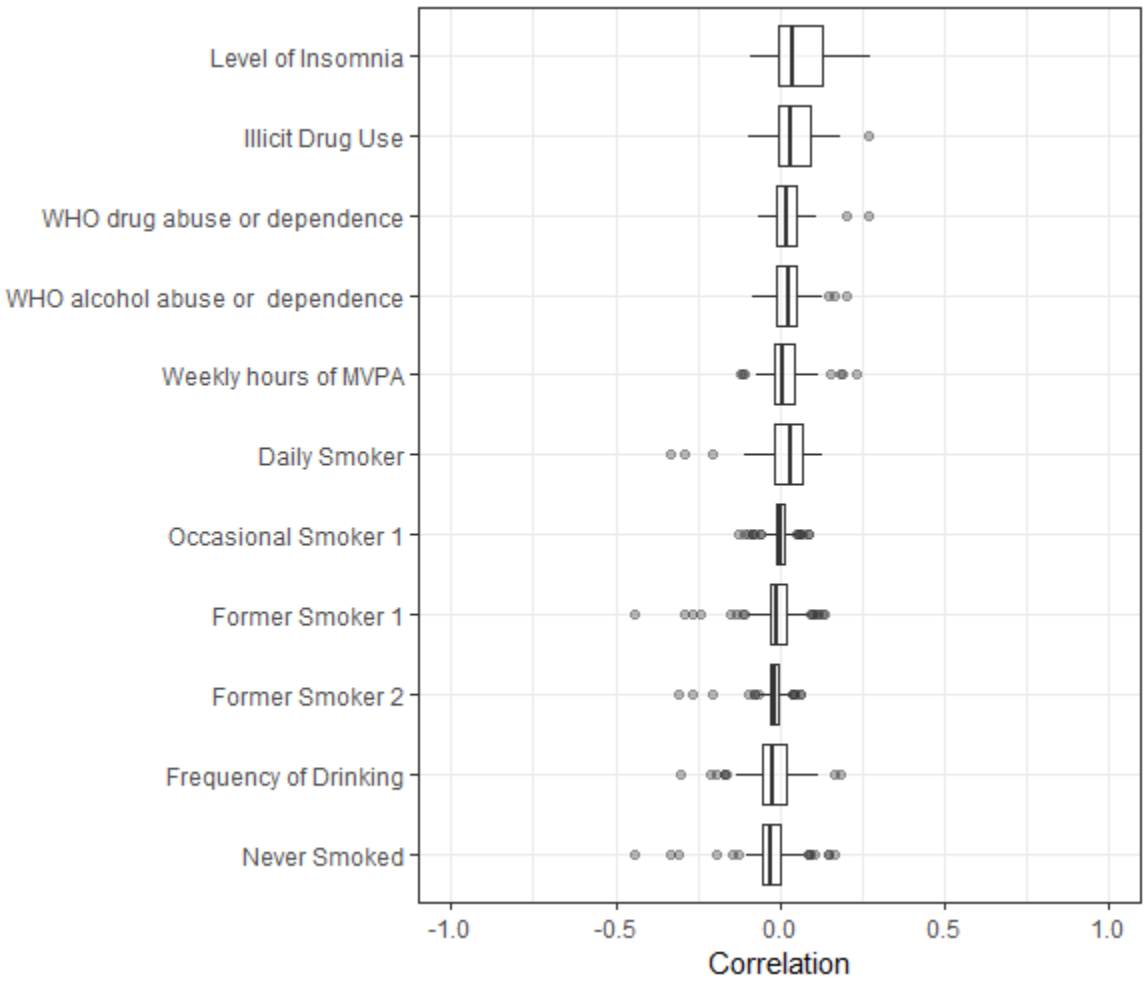

Figure 7 - Box and whisker plot showing the correlation coefficients between each health behavior and all other variables.

Note: The line in the center of the box represents the median correlation coefficient for the given variable. The box extends from the first quartile to the third quartile, and the whiskers extend to the largest value within 1.5 * interquartile range from the end of the box. Points beyond the whisker represent outliers and level of opacity on these points denote a greater concentration of outliers.

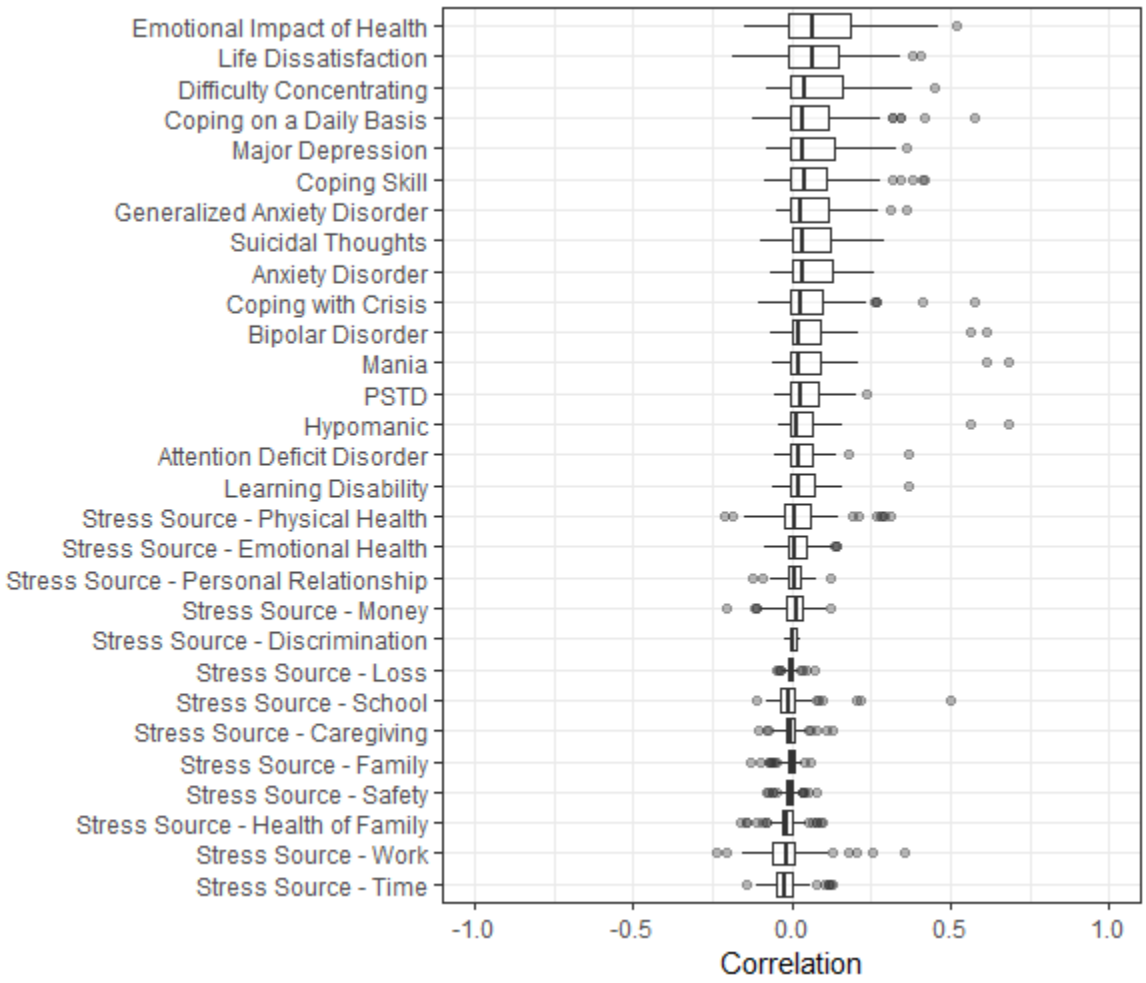

Figure 8 - Box and whisker plot showing the correlation coefficients between each mental health variable and all other variables.

Note: The line in the center of the box represents the median correlation coefficient for the given variable. The box extends from the first quartile to the third quartile, and the whiskers extend to the largest value within 1.5 * interquartile range from the end of the box. Points beyond the whisker represent outliers and level of opacity on these points denote a greater concentration of outliers.

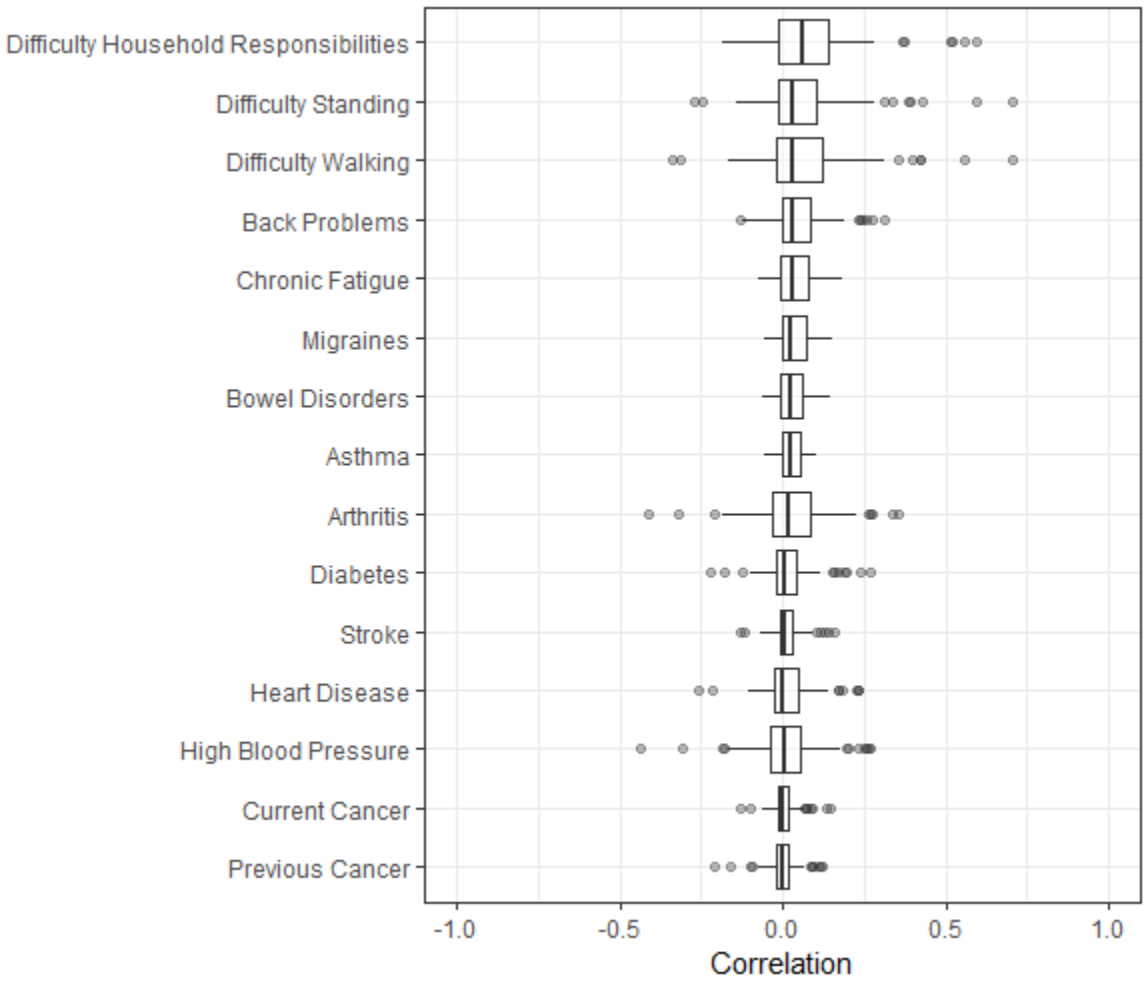

Figure 9 - Box and whisker plot showing the correlation coefficients between each physical health variable and all other variables.

Note: the line in the center of the box represents the median correlation coefficient for the given variable. The box extends from the first quartile to the third quartile, and the whiskers extend to the largest value within 1.5 * interquartile range from the end of the box. Points beyond the whisker represent outliers and level of opacity on these points denote a greater concentration of outliers.

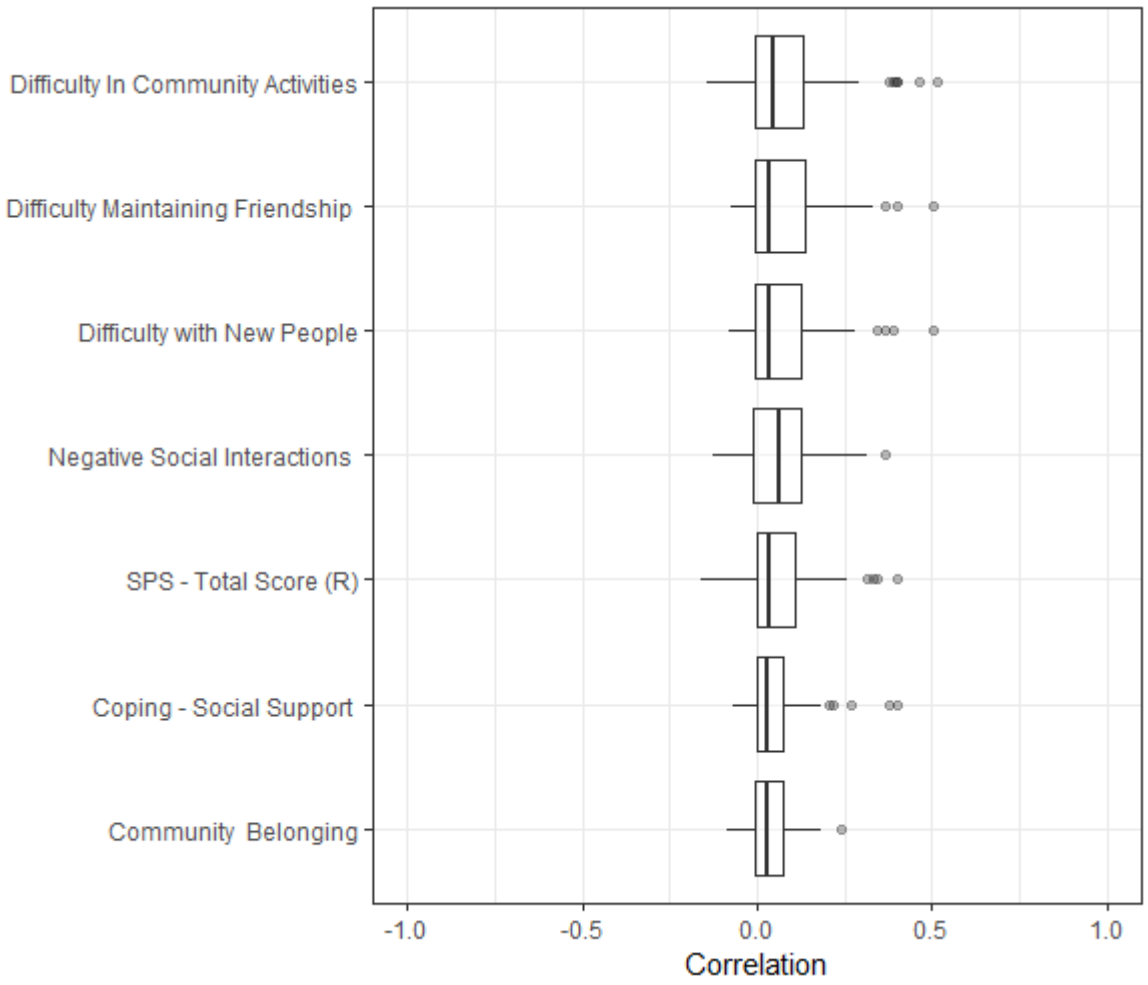

Figure 10 - Box and whisker plot showing the correlation coefficients between each social factor and all other variables.

Note: The line in the center of the box represents the median correlation coefficient for the given variable. The box extends from the first quartile to the third quartile, and the whiskers extend to the largest value within 1.5 * interquartile range from the end of the box. Points beyond the whisker represent outliers and level of opacity on these points denote a greater concentration of outliers.

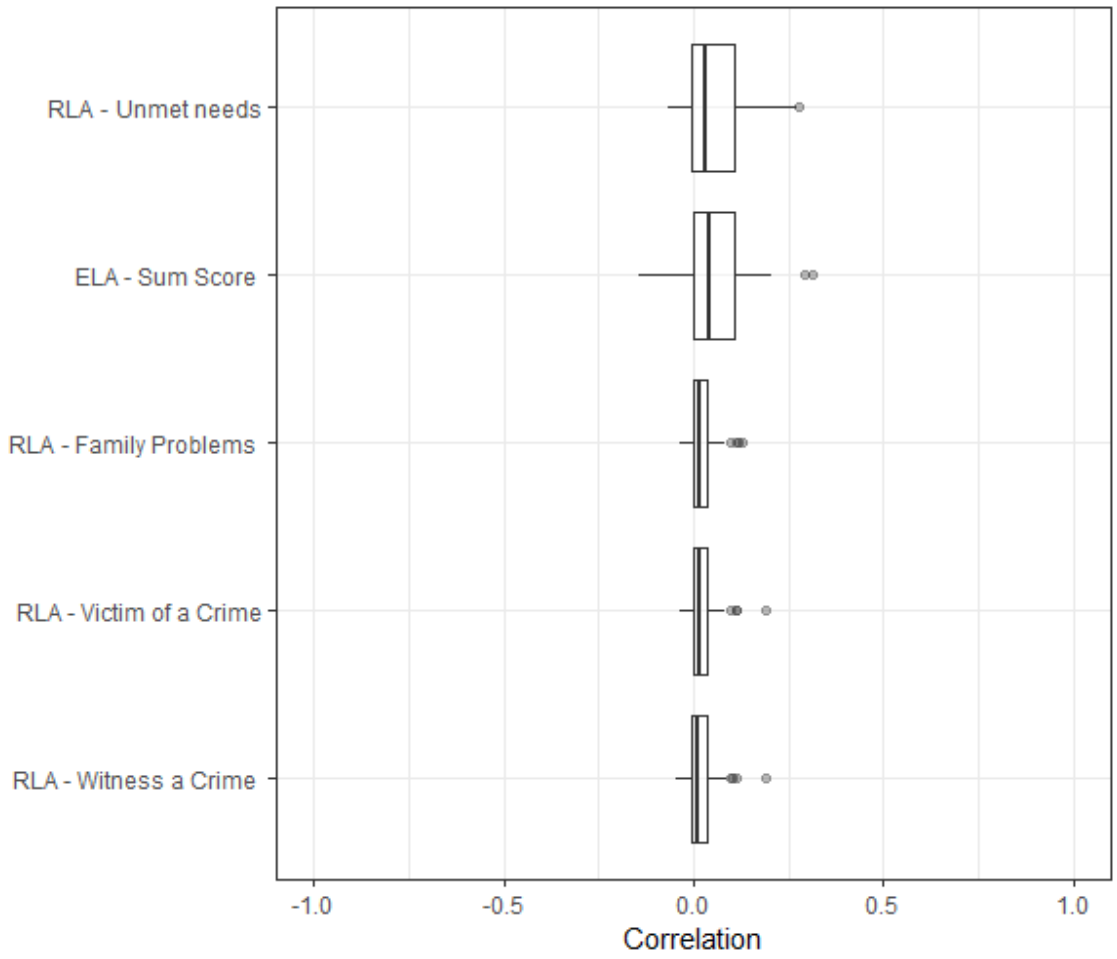

Figure 11 - Box and whisker plot showing the correlation coefficients between each life adversity variable and all other variables.

Note: The line in the center of the box represents the median correlation coefficient for the given variable. The box extends from the first quartile to the third quartile, and the whiskers extend to the largest value within 1.5 * interquartile range from the end of the box. Points beyond the whisker represent outliers and level of opacity on these points denote a greater concentration of outliers.

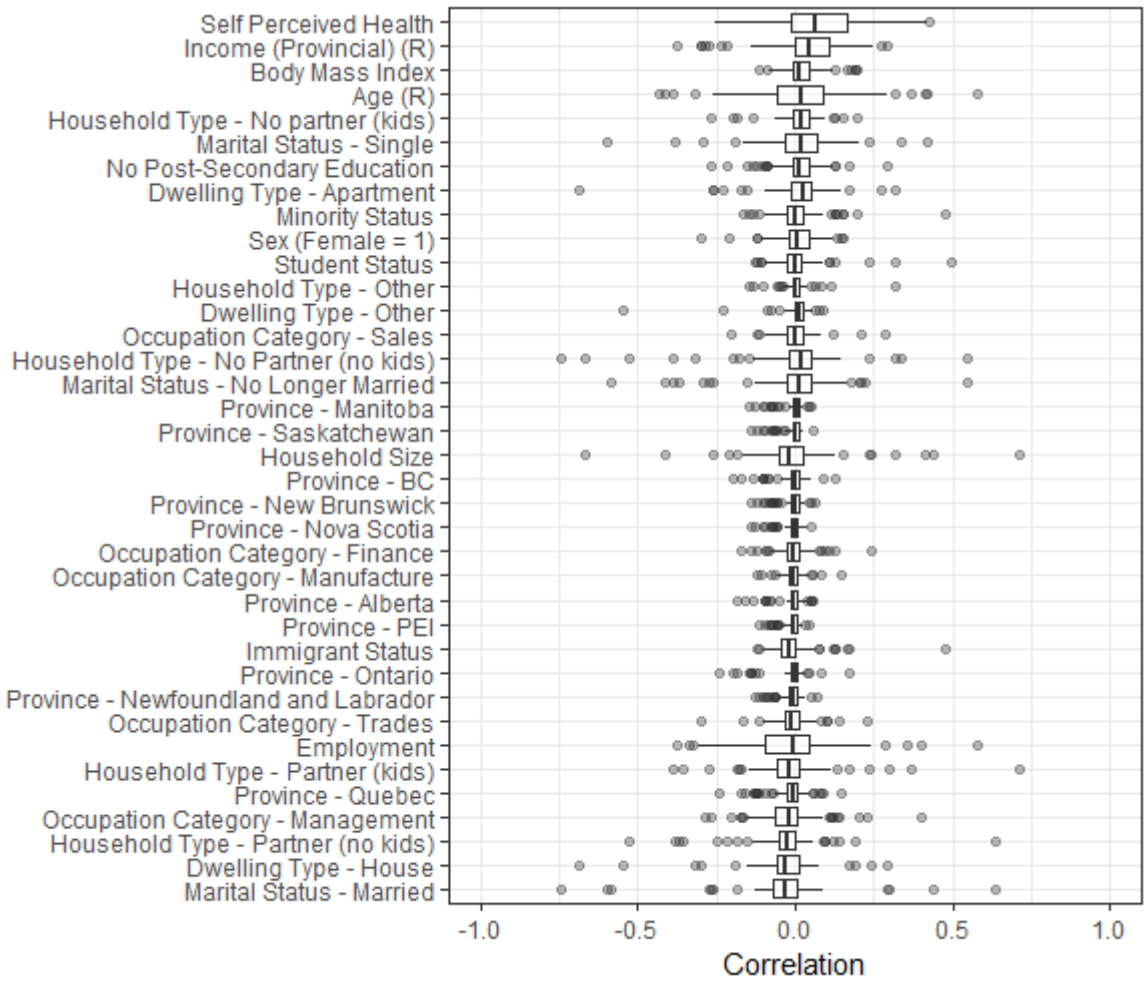

Figure 12 - Box and whisker plot showing the correlation coefficients between each demographic and all other variables.

Note: The line in the center of the box represents the median correlation coefficient for the given variable. The box extends from the first quartile to the third quartile, and the whiskers extend to the largest value within 1.5 * interquartile range from the end of the box. Points beyond the whisker represent outliers and level of opacity on these points denote a greater concentration of outliers.

Appendix E - Simple Regression Results

Table 16 - Simple linear regression results (i.e., Estimated Beta, 95% CI, and R^2)

Variable	Est. β	95% CI	R^2
Negative Social Interactions	0.356***	[0.344, 0.368]	0.13
Life Satisfaction (R)	0.318***	[0.305, 0.330]	0.09
Level of Insomnia	0.230***	[0.217, 0.243]	0.05
Emotional Impact of Health	0.216***	[0.202, 0.229]	0.04
Employment	0.452***	[0.424, 0.480]	0.04
Age (R)	0.238***	[0.223, 0.253]	0.04
Coping - Social Support	0.201***	[0.188, 0.215]	0.04
Coping on a Daily Basis	0.185***	[0.172, 0.198]	0.03
Major Depression	0.845***	[0.784, 0.905]	0.03
Recent Life Events - Unmet needs	0.715***	[0.660, 0.770]	0.03
Difficulty Concentrating	0.174***	[0.161, 0.188]	0.03
Stress Source - Work	0.359***	[0.329, 0.388]	0.02
Coping Skill	0.164***	[0.151, 0.178]	0.02
ELA - Sum Score	0.162***	[0.148, 0.175]	0.02
Generalized Anxiety Disorder	0.944***	[0.864, 1.020]	0.02
Coping with Crisis	0.150***	[0.137, 0.163]	0.02
Migraines	0.483***	[0.441, 0.525]	0.02
Self-Perceived Health	0.153***	[0.140, 0.166]	0.02
Suicidal Thoughts	0.430***	[0.391, 0.470]	0.02
Difficulty Household Responsibilities	0.149***	[0.136, 0.163]	0.02
Anxiety Disorder	0.594***	[0.535, 0.654]	0.02
Occupation Category - Management	0.290***	[0.261, 0.319]	0.02
Community Belonging	0.120***	[0.108, 0.133]	0.01
Household Type - Partner (no kids)	-0.257***	[-0.285, -0.229]	0.01
Difficulty Maintaining Friendship	0.118***	[0.104, 0.131]	0.01
Household Type - Partner (kids)	0.226***	[0.199, 0.253]	0.01
Difficulty with New People	0.110***	[0.097, 0.124]	0.01
Difficulty In Community Activities	0.109***	[0.096, 0.123]	0.01
Household Size	0.099***	[0.086, 0.111]	0.01
Occupation Category - Finance	0.314***	[0.273, 0.355]	0.01
Bipolar Disorder	0.603***	[0.521, 0.684]	<0.01
PSTD	0.717***	[0.618, 0.815]	<0.01
Chronic Fatigue	0.826***	[0.717, 0.934]	<0.01
SPS - Total Score (R)	0.094***	[0.081, 0.106]	<0.01
Back Problems	0.231***	[0.199, 0.264]	<0.01
Mania	0.851***	[0.719, 0.983]	<0.01
Household Type - No partner (kids)	0.325***	[0.276, 0.375]	<0.01
Daily Smoker	0.215***	[0.181, 0.250]	<0.01
Stress Source - Money	0.215***	[0.178, 0.251]	<0.01

Bowel Disorders	0.354***	[0.293, 0.415]	<0.01
Recent Life Events - Victim of a Crime	0.075***	[0.062, 0.088]	<0.01
Illicit Drug Use	0.281***	[0.228, 0.334]	<0.01
Hypomanic	0.81***	[0.665, 0.954]	<0.01
High Blood Pressure	-0.183***	[-0.216, -0.150]	<0.01
Recent Life Events - Family Problems	0.070***	[0.057, 0.083]	<0.01
Sex (Female = 1)	0.140***	[0.114, 0.166]	<0.01
No Post-Secondary Education	-0.149***	[-0.176, -0.123]	<0.01
Recent Life Events - Witness a Crime	0.064***	[0.051, 0.078]	<0.01
Learning Disability	0.302***	[0.229, 0.376]	<0.01
Asthma	0.217***	[0.169, 0.265]	<0.01
Heart Disease	-0.234***	[-0.292, -0.177]	<0.01
Difficulty Standing	0.058***	[0.044, 0.072]	<0.01
Immigrant Status	-0.131***	[-0.160, -0.102]	<0.01
Former Smoker 1	-0.105***	[-0.135, -0.076]	<0.01
Never Smoked	-0.088***	[-0.115, -0.062]	<0.01
Attention Deficit Disorder	0.319***	[0.234, 0.405]	<0.01
Stress Source - Time	0.159***	[0.114, 0.205]	<0.01
Stress Source - Emotional Health	0.324***	[0.239, 0.408]	<0.01
Stress Source - Personal Relationship	0.213***	[0.155, 0.271]	<0.01
Student Status	0.164***	[0.116, 0.211]	<0.01
Frequency of Drinking	0.031***	[0.018, 0.043]	<0.01
WHO alcohol abuse or dependence	0.223***	[0.144, 0.301]	<0.01
Occasional Smoker 1	0.197***	[0.124, 0.271]	<0.01
WHO drug abuse or dependence	0.290***	[0.180, 0.400]	<0.01
Stress Source - School	0.188***	[0.105, 0.272]	<0.01
Diabetes	-0.107***	[-0.157, -0.057]	<0.01
Previous Cancer	-0.108***	[-0.165, -0.050]	<0.01
Stroke	-0.234***	[-0.355, -0.112]	<0.01
Difficulty Walking	0.031***	[0.016, 0.046]	<0.01
Household Type - No Partner (no kids)	-0.074***	[-0.106, -0.043]	<0.01
Household Type - Other	-0.101***	[-0.151, -0.051]	<0.01
Marital Status - Single	0.092***	[0.06, 0.124]	<0.01
Province - Newfoundland and Labrador	-0.193***	[-0.298, -0.088]	<0.01
Province - Ontario	0.065***	[0.039, 0.092]	<0.01
Occupation Category - Sales	0.083***	[0.045, 0.120]	<0.01
Income (Provincial) (R)	-0.038***	[-0.05, -0.025]	<0.01
Body Mass Index	0.030***	[0.017, 0.044]	<0.01
Minority Status	-0.070***	[-0.101, -0.039]	<0.01
Occasional Smoker 2	0.068	[-0.019, 0.156]	<0.01
Former Smoker 2	0.027	[-0.008, 0.062]	<0.01
Stress Source - Physical Health	-0.029	[-0.082, 0.024]	<0.01
Stress Source - Caregiving	0.109**	[0.045, 0.174]	<0.01
Stress Source - Family	0.032	[-0.024, 0.088]	<0.01

Stress Source - Discrimination	-0.115	[-0.410, 0.180]	<0.01
Stress Source - Safety	-0.136**	[-0.229, -0.044]	<0.01
Stress Source - Health of Family	-0.061*	[-0.113, -0.009]	<0.01
Stress Source - Loss	-0.063	[-0.332, 0.206]	<0.01
Arthritis	-0.041*	[-0.074, -0.007]	<0.01
Current Cancer	-0.018	[-0.110, 0.074]	<0.01
Dwelling Type - House	0.021	[-0.005, 0.048]	<0.01
Dwelling Type - Apartment	-0.024	[-0.056, 0.008]	<0.01
Dwelling Type - Other	-0.008	[-0.043, 0.027]	<0.01
Marital Status - Married	-0.034*	[-0.061, -0.007]	<0.01
Marital Status - No Longer Married	-0.061**	[-0.098, -0.024]	<0.01
Province - PEI	-0.064	[-0.262, 0.134]	<0.01
Province - Nova Scotia	-0.075	[-0.154, 0.004]	<0.01
Province - New Brunswick	-0.015	[-0.103, 0.073]	<0.01
Province - Quebec	-0.023	[-0.053, 0.008]	<0.01
Province - Manitoba	0.085*	[0.014, 0.156]	<0.01
Province - Saskatchewan	-0.007	[-0.084, 0.070]	<0.01
Province - Alberta	-0.035	[-0.076, 0.007]	<0.01
Province - BC	-0.045*	[-0.083, -0.007]	<0.01
Occupation Category - Trades	0.001	[-0.041, 0.044]	<0.01
Occupation Category - Manufacture	-0.056	[-0.12, 0.009]	<0.01
Weekly hours of Moderate to Vigorous Physical Activity (R)	0.012	[-0.002, 0.025]	<0.01

Appendix F - Sensitivity Analysis Results

Table 17 - Results of the three sensitivity analyses.

Variable	Age (20-75)		Reported Primary Stressor		No Mental Illness	
	Raw Imp.	Std. Imp. (Rank)	Raw Imp.	Std. Imp. (Rank)	Raw Imp.	Std. Imp. (Rank)
Life Satisfaction (R)	0.105	0.99 (2)	0.093	1.00 (1)	0.087	0.97 (3)
Negative Social Interactions	0.106	1.00 (1)	0.077	0.82 (2)	0.087	0.97 (2)
Employment	0.032	0.30 (4)	0.027	0.29 (4)	0.052	0.58 (4)
Age (R)	0.054	0.51 (3)	0.060	0.64 (3)	0.090	1.00 (1)
Emotional Impact of Health	0.012	0.11 (9)	0.015	0.16 (7)	0.014	0.15 (8)
Stress Source - Work	0.031	0.29 (5)	0.007	0.08 (13)	0.025	0.27 (5)
Level of Insomnia	0.028	0.26 (6)	0.021	0.23 (5)	0.020	0.22 (6)
Self-Perceived Health	0.012	0.11 (8)	0.017	0.18 (6)	0.009	0.10 (11)
Coping - Social Support	0.009	0.08 (10)	0.008	0.09 (11)	0.014	0.16 (7)
Coping on a Daily Basis	0.012	0.11 (7)	0.009	0.10 (10)	0.010	0.11 (10)
SPS - Total Score (R)	0.006	0.06 (13)	0.010	0.11 (9)	0.008	0.09 (12)
Coping with Crisis	0.006	0.05 (14)	0.008	0.08 (12)	0.005	0.06 (17)
Income (Provincial) (R)	0.008	0.08 (11)	0.007	0.07 (14)	0.006	0.07 (14)
Household Size	0.004	0.03 (25)	0.007	0.07 (15)	0.011	0.12 (9)
Major Depression	0.005	0.05 (16)	0.014	0.15 (8)	<0.001	<0.01 (83)
Difficulty Household Responsibilities	0.004	0.04 (19)	0.002	0.02 (41)	0.004	0.05 (19)
Frequency of Drinking	0.005	0.04 (18)	0.003	0.03 (31)	0.004	0.04 (22)
Difficulty Concentrating	0.003	0.03 (27)	0.004	0.04 (24)	0.003	0.03 (31)
Sex (Female = 1)	0.007	0.06 (12)	0.004	0.04 (23)	0.004	0.04 (20)
Occupation Category - Management	0.004	0.04 (20)	0.005	0.05 (21)	0.008	0.09 (13)
ELA - Sum Score	0.004	0.04 (22)	0.003	0.04 (26)	0.003	0.03 (28)
Coping Skill	0.004	0.03 (26)	0.006	0.07 (17)	0.005	0.06 (16)
Marital Status - Married	0.002	0.02 (36)	0.003	0.03 (34)	0.004	0.04 (23)
Marital Status - No Longer Married	0.003	0.02 (31)	0.003	0.04 (27)	0.004	0.05 (18)
Difficulty Standing	0.004	0.04 (21)	0.003	0.04 (28)	0.002	0.03 (33)
Household Type - No Partner (no kids)	0.002	0.01 (44)	0.003	0.03 (32)	0.003	0.03 (30)
Back Problems	0.002	0.02 (40)	0.002	0.03 (40)	0.002	0.02 (38)
Difficulty Walking	0.004	0.04 (24)	0.005	0.05 (20)	0.004	0.04 (21)
High Blood Pressure	0.003	0.02 (32)	0.004	0.05 (22)	0.003	0.04 (24)
Arthritis	0.002	0.02 (34)	0.005	0.06 (19)	0.002	0.03 (34)
RLA - Unmet needs	0.004	0.04 (23)	0.003	0.03 (30)	0.003	0.04 (25)

Difficulty In Community Activities	0.001	0.01 (47)	0.002	0.02 (43)	0.001	0.01 (56)
Household Type - Partner (kids)	0.003	0.02 (30)	0.003	0.03 (36)	0.006	0.07 (15)
Generalized Anxiety Disorder	0.002	0.02 (41)	0.006	0.07 (16)	<0.001	<0.01 (84)
Migraines	0.005	0.05 (17)	0.001	0.01 (60)	0.002	0.02 (42)
Province - Quebec	0.002	0.02 (35)	0.002	0.02 (46)	0.002	0.02 (44)
Community Belonging	0.003	0.03 (28)	0.002	0.03 (38)	0.002	0.02 (37)
Marital Status - Single	0.002	0.02 (37)	0.003	0.03 (35)	0.002	0.02 (39)
Stress Source - Health of Family	0.002	0.02 (33)	0.001	0.01 (57)	0.002	0.03 (32)
Weekly hours of MVPA	0.002	0.01 (45)	0.002	0.02 (42)	0.002	0.02 (41)
Difficulty with New People	0.002	0.02 (38)	0.002	0.02 (47)	<0.001	<0.01 (76)
RLA - Witness a Crime	<0.001	<0.01 (103)	<0.001	<0.01 (81)	<0.001	<0.01 (98)
Chronic Fatigue	0.001	0.01 (53)	0.002	0.03 (39)	<0.001	<0.01 (64)
Stress Source - Money	0.001	0.01 (59)	0.001	0.01 (63)	0.002	0.02 (40)
No Post-Secondary Education	0.001	0.01 (49)	0.003	0.04 (29)	0.003	0.03 (26)
Anxiety Disorder	0.001	0.01 (48)	0.003	0.04 (25)	<0.001	<0.01 (82)
Difficulty Maintaining Friendship	<0.001	<0.01 (69)	0.001	0.01 (61)	<0.001	<0.01 (97)
Household Type - Partner (no kids)	0.003	0.02 (29)	0.003	0.03 (33)	0.003	0.03 (27)
Suicidal Thoughts	0.002	0.02 (42)	0.003	0.03 (37)	<0.001	0.01 (61)
Body Mass Index	0.005	0.05 (15)	0.005	0.06 (18)	0.002	0.03 (36)
Heart Disease	<0.001	<0.01 (67)	0.001	0.01 (58)	<0.001	0.01 (59)
Dwelling Type - House	0.002	0.02 (43)	0.001	0.01 (56)	0.001	0.01 (53)
Bowel Disorders	<0.001	<0.01 (70)	0.001	0.01 (53)	0.001	0.01 (57)
Stress Source - Time	0.001	0.01 (64)	<0.001	<0.01 (92)	0.001	0.02 (46)
Immigrant Status	0.001	0.01 (62)	<0.001	<0.01 (80)	0.001	0.01 (52)
Never Smoked	0.001	0.01 (54)	0.001	0.01 (51)	<0.001	0.01 (60)
Former Smoker 1	0.001	0.01 (65)	0.001	0.01 (55)	<0.001	<0.01 (62)
Daily Smoker	0.001	0.01 (57)	0.002	0.02 (48)	0.002	0.02 (45)
Previous Cancer	<0.001	<0.01 (72)	<0.001	<0.01 (66)	0.001	0.01 (49)
Stress Source - Physical Health	<0.001	<0.01 (79)	0.002	0.02 (45)	0.001	0.01 (48)
Asthma	0.001	0.01 (63)	<0.001	<0.01 (68)	0.001	0.01 (58)
Occupation Category - Trades	0.001	0.01 (60)	<0.001	<0.01 (72)	0.001	0.01 (55)
Student Status	0.001	0.01 (61)	0.001	0.01 (59)	0.003	0.03 (29)
Dwelling Type - Apartment	0.001	0.01 (51)	0.001	0.01 (52)	0.002	0.02 (43)
Diabetes	0.001	0.01 (55)	0.001	0.01 (64)	0.001	0.01 (51)
Minority Status	0.001	0.01 (58)	0.001	0.01 (49)	0.001	0.02 (47)
Stress Source - Family	0.001	0.01 (56)	<0.001	<0.01 (79)	0.001	0.01 (54)
Household Type - No partner (kids)	0.002	0.01 (46)	0.002	0.02 (44)	0.001	0.01 (50)
Illicit Drug Use	<0.001	<0.01 (66)	0.001	0.01 (62)	<0.001	<0.01 (63)
Stress Source - School	<0.001	<0.01 (87)	<0.001	<0.01 (76)	<0.001	<0.01 (66)

Occupation Category - Sales	0.001	0.01 (52)	0.001	0.01 (54)	0.002	0.03 (35)
RLA - Family Problems	<0.001	<0.01 (80)	<0.001	<0.01 (70)	<0.001	<0.01 (68)
RLA - Victim of a Crime	<0.001	<0.01 (74)	<0.001	0.01 (65)	<0.001	<0.01 (73)
Stress Source - Personal Relationship	<0.001	<0.01 (96)	<0.001	<0.01 (74)	<0.001	<0.01 (88)
Learning Disability	<0.001	<0.01 (68)	<0.001	<0.01 (75)	<0.001	<0.01 (95)
Province - Newfoundland and Labrador	<0.001	<0.01 (100)	<0.001	<0.01 (85)	<0.001	<0.01 (69)
Stress Source - Caregiving	0.001	0.01 (50)	<0.001	0.01 (105)	<0.001	<0.01 (72)
Stress Source - Emotional Health	<0.001	<0.01 (84)	<0.001	<0.01 (88)	<0.001	<0.01 (90)
Province - Saskatchewan	<0.001	<0.01 (92)	<0.001	<0.01 (96)	<0.001	<0.01 (89)
Stress Source - Loss	<0.001	<0.01 (89)	<0.001	<0.01 (84)	<0.001	<0.01 (67)
Bipolar Disorder	<0.001	<0.01 (71)	<0.001	<0.01 (73)	<0.001	<0.01 (81)
Dwelling Type - Other	<0.001	<0.01 (88)	<0.001	<0.01 (102)	<0.001	<0.01 (80)
Household Type - Other	<0.001	<0.01 (77)	<0.001	<0.01 (98)	<0.001	<0.01 (65)
Attention Deficit Disorder	<0.001	<0.01 (73)	<0.001	<0.01 (82)	<0.001	<0.01 (74)
PSTD	<0.001	<0.01 (76)	<0.001	<0.01 (67)	<0.001	<0.01 (77)
Province - Alberta	<0.001	<0.01 (81)	<0.001	<0.01 (99)	<0.001	<0.01 (75)
Stress Source - Safety	<0.001	<0.01 (82)	<0.001	<0.01 (69)	<0.001	<0.01 (70)
Stress Source - Discrimination	<0.001	<0.01 (86)	<0.001	<0.01 (86)	<0.001	<0.01 (87)
Current Cancer	<0.001	<0.01 (85)	<0.001	<0.01 (77)	<0.001	<0.01 (91)
WHO alcohol abuse or dependence	<0.001	<0.01 (94)	<0.001	<0.01 (71)	<0.001	<0.01 (78)
Mania	<0.001	<0.01 (83)	<0.001	<0.01 (83)	<0.001	<0.01 (86)
WHO drug abuse or dependence	<0.001	<0.01 (95)	<0.001	<0.01 (90)	<0.001	<0.01 (94)
Province - BC	<0.001	<0.01 (75)	<0.001	<0.01 (78)	<0.001	<0.01 (101)
Occasional Smoker 2	<0.001	<0.01 (98)	<0.001	<0.01 (100)	<0.001	<0.01 (93)
Hypomanic	<0.001	<0.01 (90)	<0.001	<0.01 (97)	<0.001	<0.01 (85)
Province - Manitoba	<0.001	<0.01 (105)	<0.001	<0.01 (91)	<0.001	<0.01 (99)
Province - PEI	<0.001	<0.01 (97)	<0.001	<0.01 (87)	<0.001	<0.01 (92)
Occasional Smoker 1	<0.001	<0.01 (78)	<0.001	<0.01 (101)	<0.001	<0.01 (103)
Province - Nova Scotia	<0.001	<0.01 (101)	<0.001	<0.01 (95)	<0.001	<0.01 (104)
Stroke	<0.001	<0.01 (91)	<0.001	<0.01 (93)	<0.001	<0.01 (100)
Province - Ontario	<0.001	<0.01 (102)	<0.001	<0.01 (94)	<0.001	<0.01 (96)
Occupation Category - Manufacture	<0.001	<0.01 (93)	<0.001	<0.01 (104)	<0.001	<0.01 (79)
Province - New Brunswick	<0.001	<0.01 (104)	<0.001	<0.01 (89)	<0.001	<0.01 (102)
Former Smoker 2	<0.001	<0.01 (99)	<0.001	<0.01 (103)	<0.001	0.01 (105)
Occupation Category - Finance	0.002	0.02 (39)	0.001	0.01 (50)	<0.001	<0.01 (71)

Note: For each sensitivity analysis, the raw importance and standardized importance are shown. The rank of importance denotes the position of the variable's importance, compared to each other variable in a given analysis. The Age (20-75) analysis includes only those between the ages of 20 and 75; this omits those over age

75 who were not asked the employment question. The Reported Primary Stressor includes only those who listed a greatest source of stress, which omits those who were not asked the coping item. Finally, the No Mental Illness analysis omits those who reported having mental illness including: depression, bipolar disorder, any anxiety disorder, mania, and hypomania. Negative Social Interactions, Life Satisfaction (R), and Age (R) were consistantly seen to be the three most important variables, with other variables being less stable in at least one of the sensitivity analyses.

Appendix G - Variable Importance

Table 18 - Variable importance, both raw and standardized, for each variable and their ordered rank.

Variable	Raw Imp.	Std. Imp.
Life Satisfaction (R)	0.103	1.000
Negative Social Interactions	0.094	0.908
Employment	0.060	0.576
Age (R)	0.060	0.576
Emotional Impact of Health	0.027	0.258
Stress Source - Work	0.018	0.175
Level of Insomnia	0.017	0.167
Self-Perceived Health	0.013	0.129
Coping - Social Support	0.012	0.111
Coping on a Daily Basis	0.010	0.097
SPS - Total Score (R)	0.009	0.089
Coping with Crisis	0.009	0.085
Income (Provincial) (R)	0.009	0.084
Household Size	0.008	0.080
Major Depression	0.008	0.078
Difficulty Household Responsibilities	0.005	0.051
Frequency of Drinking	0.005	0.050
Difficulty Concentrating	0.005	0.047
Sex (Female = 1)	0.005	0.047
Occupation Category - Management	0.005	0.047
ELA - Sum Score	0.004	0.043
Coping Skill	0.004	0.043
Marital Status - Married	0.004	0.043
Marital Status - No Longer Married	0.004	0.035
Difficulty Standing	0.004	0.035
Household Type - No Partner (no kids)	0.004	0.035
Back Problems	0.004	0.034
Difficulty Walking	0.003	0.033
High Blood Pressure	0.003	0.032
Arthritis	0.003	0.032
Recent Life Events - Unmet needs	0.003	0.028
Difficulty In Community Activities	0.003	0.026
Household Type - Partner (kids)	0.003	0.026

Generalized Anxiety Disorder	0.002	0.023
Migraines	0.002	0.022
Province - Quebec	0.002	0.021
Community Belonging	0.002	0.021
Marital Status - Single	0.002	0.021
Stress Source - Health of Family	0.002	0.020
Weekly hours of Moderate to Vigorous Physical Activity	0.002	0.019
Difficulty with New People	0.002	0.018
Recent Life Events - Witness a Crime	0.002	0.018
Chronic Fatigue	0.002	0.017
Stress Source - Money	0.002	0.017
No Post-Secondary Education	0.002	0.017
Anxiety Disorder	0.002	0.017
Difficulty Maintaining Friendship	0.002	0.016
Household Type - Partner (no kids)	0.002	0.015
Suicidal Thoughts	0.002	0.015
Body Mass Index	0.001	0.014
Heart Disease	0.001	0.013
Dwelling Type - House	0.001	0.013
Bowel Disorders	0.001	0.012
Stress Source - Time	0.001	0.012
Immigrant Status	0.001	0.011
Never Smoked	0.001	0.011
Former Smoker 1	0.001	0.010
Daily Smoker	0.001	0.010
Previous Cancer	0.001	0.008
Stress Source - Physical Health	0.001	0.008
Asthma	0.001	0.007
Occupation Category - Trades	0.001	0.007
Student Status	0.001	0.007
Dwelling Type - Apartment	0.001	0.007
Diabetes	0.001	0.007
Minority Status	0.001	0.006
Stress Source - Family	0.001	0.006
Household Type - No partner (kids)	0.001	0.006
Illicit Drug Use	0.001	0.005
Stress Source - School	0.001	0.005
Occupation Category - Sales	<0.001	0.004
Recent Life Events - Family Problems	<0.001	0.004

Recent Life Events - Victim of a Crime	<0.001	0.003
Stress Source - Personal Relationship	<0.001	0.003
Learning Disability	<0.001	0.003
Province - Newfoundland and Labrador	<0.001	0.002
Stress Source - Caregiving	<0.001	0.002
Stress Source - Emotional Health	<0.001	0.001
Province - Saskatchewan	<0.001	0.001
Stress Source - Loss	<0.001	0.001
Bipolar Disorder	<0.001	0.001
Dwelling Type - Other	<0.001	0.001
Household Type - Other	<0.001	0.001
Attention Deficit Disorder	<0.001	0.001
PSTD	<0.001	<0.001
Province - Alberta	<0.001	<0.001
Stress Source - Safety	<0.001	<0.001
Stress Source - Discrimination	<0.001	<0.001
Current Cancer	<0.001	<0.001
WHO alcohol abuse or dependence	<0.001	<0.001
Mania	<0.001	<0.001
WHO drug abuse or dependence	<0.001	<0.001
Province - BC	<0.001	<0.001
Occasional Smoker 2	<0.001	<0.001
Hypomanic	<0.001	<0.001
Province - Manitoba	<0.001	<0.001
Province - PEI	<0.001	<0.001
Occasional Smoker 1	<0.001	<0.001
Province - Nova Scotia	<0.001	<0.001
Stroke	<0.001	<0.001
Province - Ontario	<0.001	<0.001
Occupation Category - Manufacture	<0.001	<0.001
Province - New Brunswick	<0.001	<0.001
Former Smoker 2	<0.001	<0.001
Occupation Category - Finance	<0.001	<0.001