Joint modeling of binary longitudinal measurement and time-to-event: An application to depression and time-to-dementia

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

in Partial Fulfillment of the Requirements

for the degree of Master of Science

in the Collaborative Program in Biostatistics of School of Public Health

University of Saskatchewan

Saskatoon, Canada

by

Md Rasel Kabir

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in wholeor part should be addressed to:

Director of School of Public Health

Health Sciences Building E-Wing, 104 Clinic Place

University of Saskatchewan

Saskatoon, Saskatchewan S7N 2Z4, Canada

OR

Dean

College of Graduate and Postdoctoral Studies

University of Saskatchewan

116 Thorvaldson Building, 110 Science Place

Saskatoon, Saskatchewan S7N 5C9, Canada

Abstract

In recent years, the methodological development of joint models of longitudinal and time-to-event data has become one of the most popular areas of studies in clinical research and its application has increased substantially over the past decades. Joint model in this area combines both the longitudinal and survival data into a single statistical model to obtain robust estimates and draw valid inference. While most of studies concentrate on continuous longitudinal measurements, little attention has been paid to joint modeling for binary longitudinal outcome and event time data. In clinical research, patients often have binary longitudinal measurement that affects the main event of interest during the follow-up time. For example, depression, a dichotomous longitudinal measurement, might have relationship with dementia. However, no study has examined this association using a joint model.

This study focuses on the joint modeling technique for binary repeated measurement and time-to-event data. This approach mainly models the longitudinal and survival processes for each individual through a shared random effect jointly, where the longitudinal part is supposed to be modeled by a generalized linear mixed model and time-to-event component is characterized by employing a parametric survival model. We applied the joint modeling technique to the Korean Health Panel Study. A generalized linear mixed model was used to model the binary repeated measurements of depression and a piecewise constant hazard model was employed for time-to-dementia. A total of 3,611 individuals aged 65 years or older were eligible for this study between 2008 and 2015. Depression and dementia were identified by the diagnosis code in medical data.

In this study, 215 (6%) were diagnosed with dementia during the 8-year follow-up period. The mean age at entry was 72.2 ($\pm$5.7) years. The overall median follow-up time was 5.8 years; 3.6 years for people living with dementia compared to 5.9 years for people without dementia. Baseline depression and sex were not significantly associated with time-to-dementia. However, time-varying depression and baseline covariates including age, economic activity, education, walking frequency/week, living with other family members and diabetes were significant in multivariable joint modeling. The risk of dementia was 2.4 times (95% CI: 1.30-4.50, p-value = 0.005) higher among depressed people compared to non-depressed people. This study also found that walking not at all or less than three days a week, being older (>70 years old), having diabetes, being less educated and living in a household with multiple generations increased the risk of dementia.

Acknowledgments

At first, I would like to express my special thanks of gratitude to my supervisor, Dr. Hyun J. "June" Lim for the excellent guidance, continuous support, strong motivation and insightful advice, she had provided throughout this whole study period. I am extremely lucky to have a supervisor who is so much careful to me, my work, and who responded to my questions and queries very promptly. Her descent approach and both financial and mental support helped me a lot to concentrate my attention fully towards the study. Without her proper monitoring, flexible schedule and unlimited time, it was entirely impossible to complete this program, particularly in this foreign country. To me, she is more than a supervisor and not less than my family members.

I am grateful to Dr. Prosanta Kumar Mondal for his great support, particularly in the context of statistical analysis and discussion. His presence, suggestion and guidelines always helped me to choose the right decision. I am thankful to my advisory committee Dr. Cindy Feng and Dr. Longhai Li for their valuable comments on each of the meetings and reports.

My sincere thanks go to the School of Public Health, University of Saskatchewan for departmental and financial support as well as the Korean Government, for providing the KHP data used for this research purpose. I would like to thank Masud Rana for spending a effective long time, mostly with statistics related discussion.

Finally, I wish to thank my family: my son Rishad, my wife, Ayesha for their sacrifices and support during my study. I am grateful to my father, Md Isahaque Ali for his

unconditional love, constant prayer and my mother, Rekha Khatun who passed away three years ago, she had always been supportive to me. This thesis is dedicated to them.

Contents

List of Tables

List of Figures

List of Abbreviations

AD	Alzheimer's disease
AIC	Akaike's Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
EM	Expectation-Maximization
GLM	General Linear Model
GLMM	Generalized Linear Mixed Effects Model
HRT	Hormone Replacement Therapy
KHP	Korean Health Panel
K-M	Kaplan–Meier
MLE	Maximum Likelihood Estimates
PH	Proportional Hazard
WHO	World Health Organization

Chapter 1

Introduction

1.1 Rationale for the study

In clinical research and health studies, it is very common to collect information about longitudinal measurements until the occurrence of an event (failure time) or censoring (censored time); interest often lies in determining the relationship between these two variables and examining the impact of longitudinal covariate(s) on survival outcome. In health sciences, longitudinal data mostly comprise the repeated measurement of a variable and binary indicators for the event of interest (live, or death). In this situation, there is always a strong possibility that the time-to-event of interest might be associated with the trajectories of longitudinal measurements and that statistical methods that ignore this repeated measurement and the course of covariates during this follow-up period might shed the etiological history of the disease (Arbeev, Akushevich, Kulminski, Ukraintseva, & Yashin, 2014; Gasparini et al., 2020). Furthermore, the dropout/ missing value of longitudinal response and measurement error complicates the survival analysis. However, joint modeling techniques in this context play a significant role in linking both models

and incorporating all information simultaneously to draw valid inferences.

The extended Cox model or time-dependent variable in the Cox hazards model is a popular method in survival analysis and can be used to examine the relationship of a covariate to survival time (Cox, 1972), particularly when incorporating an external time-dependent covariate (exogenous variable), that is, when an event that occurred at a previous time point does not affect the covariate value at a later time point (Kalbfleisch & Prentice, 2011). When the longitudinal measurements, for example CD4 (a biomarker for HIV/AIDS), are considered as endogenous, the application of the standard extended Cox model, which assumes constant values for the time-varying covariate between two occasions in this context is inappropriate and results in biased estimates and standard errors (Prentice, 1982). Alternatively, the two-stage modeling approach deals with survival as a function of a covariate that is measured longitudinally. In the first stage, the longitudinal covariate with random effects is fitted using growth curve models (Laird, Ware, et al., 1982). The estimated value from this first stage is then used as a time-dependent covariate for the Cox model in second stage and the partial likelihood is maximized (Dafni & Tsiatis, 1998; Tsiatis, Degruttola, & Wulfsohn, 1995). It is a very common scenario in most of the follow-up studies that subjects are dropping out and/ or visiting irregularly. However, this dropout due to unobserved longitudinal measurements will no longer be random; instead, it is defined as non-random and informative. Joint distribution of the longitudinal measurements and the missingness process takes this issue into account and provides valid inference (Rizopoulos, 2012).

Most of the joint modeling approaches concentrate on continuous longitudinal measure-

ment and focus on the association between a single longitudinal outcome and a single time-to-event. AIDS research is a very common example, where a longitudinal biomarker such as CD4 lymphocyte count is measured in time to determine its relationship with time-to-seroconversion or death (Faucett & Thomas, 1996; Henderson, Diggle, & Dobson, 2000; Mondal, Lim, Team, et al., 2018; Tsiatis & Davidian, 2004; Tsiatis et al., 1995; Wulfsohn & Tsiatis, 1997). There is also some available research, where a single outcome is extended to multiple longitudinal outcomes (Chi & Ibrahim, 2006; Rizopoulos & Ghosh, 2011) and multiple recurrent or competing events (Hu, Li, & Li, 2009; X. Huang, Li, Elashoff, & Pan, 2011).

Similar to HIV studies, many of the dementia disease trials collect two types of data: the time to diagnosis of dementia and longitudinal measurements of some endogenous variables. Although they are closely associated, data are analyzed separately most of the studies, which might lead to biased estimates and misleading inferences. Joint models take this association into account by combining both longitudinal and survival data into a single statistical model. In joint model, longitudinal measurements can be either continuous (e.g. Gaussian) or discrete (e.g. binary, count). This thesis focused on binary repeated measurements.

Several previous dementia studies used the joint modeling technique for determining the association between continuous longitudinal outcome and dementia risk (Jacqmin-Gadda, Commenges, & Dartigues, 2006; S. Li, Zheng, & Gao, 2017). One study conducted by Yu and Ghosh (2010) proposed a Bayesian change-point model to fit the trajectory of cognitive function for the individuals who developed dementia and determined when cognitive

decline rates started to accelerate. S. Li et al. (2017) applied this modeling approach on the basis of shared random effect to identify the association between longitudinal cholesterol measurements and the timing of the onset of dementia. Singh-Manoux et al. (2017) carried out a 28-year follow-up study of depression trajectories that failed to reveal an association between depressive symptoms and the risk of dementia.

Though joint modeling with continuous longitudinal outcome has been extensively studied over the past two decades, little work has been done for categorical longitudinal outcome. So there is room to work with binary longitudinal outcome and many interesting features of research in this area remain left. This thesis mainly focuses on joint modeling for binary longitudinal outcome and time-to-event data. Since the relationship between depression and the timing of onset of dementia is complex and unclear, we investigated it further in this study by polishing conventional statistical methods and employing a joint modeling approach that includes the mechanism to address the factors that lead to biased and misleading inferences.

1.2 Study objectives

The main objectives of this study:

1. To determine the association between binary repeated measurements of depression and time-to-dementia from the joint modeling of longitudinal submodel and time-to-event submodel;
2. To identify the potential risk factors of dementia from separate time-to-event survival model; and
3. To examine the association between covariates and repeated measurement of depression from separate longitudinal model.

Chapter 2

Literature Review

2.1 Literature review for statistical methods

In this section, we review the literature on methodology that is pertinent to our study. Joint modeling approach has received much more attention in recent decades and research activities has increased in the area of simultaneously modeling longitudinal biomarkers with a time-to-event outcome of interest. The underlying interest in this kind of research, particularly in psychology research, is always in examining the association between the long-term individual trajectories of cognitive functioning and either the diagnosis of a certain disease or death.

Before the development of the joint modeling technique, analyzing outcomes separately was well established in the literature; the mixed-effects model was frequently used for longitudinal data and the standard Cox proportional hazards model for time-to-event data (Cox, 1972; Laird et al., 1982). However, modeling the longitudinal and survival outcome separately can lead to biased estimates (Ibrahim, Chu, & Chen, 2010). Apply-

ing the extended Cox model is inappropriate for handling the longitudinal measurements (biomarker) of covariates, which are called internal covariate. Doing so results in biased estimates and standard errors (Prentice, 1982). Joint modeling, on the other hand, leads to less biased estimates and improvements in the efficiency of statistical inference (Hogan & Laird, 1998).

Tsiatis and Davidian (2004) have a comprehensive overviews of joint modeling. There are two basic approaches generally used for estimating parameters in the joint model: (i) a two-stage approach and (ii) a likelihood-based method. In two-stage approach, estimations are conducted separately in two different steps. In first stage, a linear mixed effects (LME) model is used for modeling continuous longitudinal data. The predicted longitudinal responses from first stage are used in the survival model as an independent variable in the second stage (Lin, Taylor, & Ye, 2008).

The two-stage method, however, often results in estimation biases and a loss of efficiency. There are several reasons for this. First, the estimation of parameters in the longitudinal model fitted at the first stage can be biased because it does not account for informative dropout (survival outcomes) (Albert & Shih, 2010; Faucett & Thomas, 1996; Ghisletta, McArdle, & Lindenberger, 2006; Sweeting & Thompson, 2011). That is, longitudinal trajectories for the subject experiencing the event and the subject not experiencing the event may be different. Therefore, in the first stage, the parameter estimates associated with the longitudinal model that are based only on observed data, might be biased, a bias that may depend on the strength of the association between longitudinal and event processes. Second, in all two-stage modeling approaches, uncertainty in the estimates made in first

stage may not be taken into account in the second stage (i.e., the survival model). This failure to incorporate the uncertainty of the estimation from the first stage may result in an underestimation of the standard error in parameter estimates in the survival model. Bias due to ignoring the uncertainty of estimates in Stage 1 may depend on the magnitude of measurement errors in longitudinal response (Wu, Liu, Yi, & Huang, 2012).

Several attempts have been made to improve estimation techniques and correct for biases in the two-stage approach. Joint models developed by Pawitan and Self (1993) basically fitted time-to-event parametrically. Here, a straightforward likelihood inference was facilitated to model the markers of disease as a function of time by considering the joint distribution of markers, infection time, and the time to AIDS. Other authors also adopted some modifications for improving estimates in the two-stage approach (Bycott & Taylor, 1998; Dafni & Tsiatis, 1998; Tsiatis et al., 1995). However, although various modification attempts have been made to incorporate the informative dropout and measurement error issues, the two-stage approach may still result in biased estimates (Wu et al., 2012).

In contrast, the likelihood-based approach considers joint likelihood functions from both the longitudinal and survival models to estimate parameter and draw the statistical inference. This approach mainly includes the concept of Expectation-Maximization (EM) algorithm and the Bayesian technique for maximum likelihood. There is an extensive body of literature that discusses EM-based likelihood methods (J. Choi, Cai, Zeng, & Olshan, 2015; De Gruttola & Tu, 1994; Rizopoulos, Verbeke, & Lesaffre, 2009; Tseng, Hsieh, & Wang, 2005; Wu, Liu, & Hu, 2010; Wulfsohn & Tsiatis, 1997). Other researchers have considered the Bayesian approach in joint models (Brown & Ibrahim, 2003; Chi

& Ibrahim, 2006; Faucett & Thomas, 1996; Hatfield, Boye, Hackshaw, & Carlin, 2012; Y. Huang, Dagne, & Wu, 2011; Law, Taylor, & Sandler, 2002; R. Brown & G. Ibrahim, 2003; Rizopoulos & Ghosh, 2011; Sweeting & Thompson, 2011; Wang & Taylor, 2001; Xu & Zeger, 2001). The advantage of both EM algorithm and Bayesian likelihood-based approaches is that they lead to valid and efficient inferences (Wu et al., 2012). That said, the major challenge of the EM-based approach for joint modeling is that it is computationally intensive, particularly when the dimension of random effects is not small. As well, there may also be convergence issues (Wu et al., 2012).

Most of the works mentioned above focus on continuous longitudinal outcomes. Very little attention has been paid to categorical longitudinal measurements in the context of joint modeling studies. J. Choi et al. (2015) proposed a joint model where a survival model was adopted by a stratified Cox model and the longitudinal categorical outcome was characterized by a generalized linear mixed model. In their modeling approach, the shared random effects mainly played the role of linking the survival and longitudinal process. J. Choi, Zeng, Olshan, and Cai (2018) allowed both the Gaussian process as well as distribution free assumptions for these random effects in their joint model. First, they assumed the multivariate Gaussian process for the random effects to account for the dependence between longitudinal measurement and survival time due to unobserved factors. Second, the normality assumption was relaxed by assuming that the distribution of random effects was unknown and mixture of Gaussian distribution was proposed. Both approaches adopted the EM algorithm for computing the estimates and model parameters.

A Study conducted by Rizopoulos, Verbeke, Lesaffre, and Vanrenterghem (2008) sug-

gested using a parametric survival model with random effect to accommodate unobserved heterogeneity and a mixed effects logistic regression was employed to model the binary longitudinal response. The researches assumed that the random effects from these longitudinal and survival model have a multiplicative relation. Both Garcia-Hernandez and Rizopoulos (2018) and Rizopoulos (2014) described the likelihood based approach and computational details with example for modeling jointly the categorical outcome and time-to-event data.

2.2 Literature review for dementia

2.2.1 Introduction

The estimated number of dementia people living worldwide in 2013 was 44.35 million, hitting to 75.62 million in 2030 and the figure of 135.46 million in 2050 respectively (Prince, Guerchet, & Prina, 2013). In East Asian countries, dementia has become one of the most pressing public health concerns given these countries' rapidly aging population. It has been identified as the highest burden of disease in older Korean population; Park, Eum, Bold, and Cheong (2013) project three times as many Koreans will suffer from dementia in 2050 compared to 2010. Moreover, there is currently no treatment for dementia, either to cure it or to alter its progressive course (Canada, 2018) (`https://alzheimer.ca/en/Home/About-dementia/Treatment-options`). Therefore, investigating and identifying the risk factors and determining their etiological relationship with dementia could be the potential strategy for reducing the prevalence and incidence of dementia .

2.2.2 Dementia and types

The World Health Organization, (WHO) defines dementia as a syndrome of a chronic or progressive nature- in which cognitive function deteriorates (WHO, 2019), (https://www.who.int/news-room/fact-sheets/detail/dementia). Thus, memory, thinking capacity, orientation, understanding, calculation, learning capability, language, and judgment are progressively impaired, eventually leading to the impossibility of performing regular activities. However, consciousness is not affected (PHA, 2019).

It is important to remember that dementia is set of symptoms rather than a disease. These symptoms are the result of various diseases and injuries that directly or indirectly affect the brain, such as vascular disease, Alzheimer's disease, or stroke (PHA, 2019; WHO, 2019). Dementia has become a great burden worldwide, particularly for the elderly. It is one of the main causes of both disability and dependency among this population. It also has a negative impact on mood and behavior (PHA, 2019). Finally, the physical, psychological, socio-economic impact of dementia on carers, care-givers, family, and society are enormous.

Alzheimer's disease

Alzheimer's disease irreversibly destroys brain cells. This disease reduces thinking ability and deteriorates memory capacity. Alzheimer's disease is the most common among all causes of dementia and does not occur as part of the normal aging process (PHA, 2019).

Other dementias

Like Alzheimer's disease, "Other dementias" are characterized by a progressive degener-

ation of brain cells. There are various types of dementia (see below), some of which are more common than others (PHA, 2019).

- Vascular dementia
- Lewy body dementia
- Frontotemporal dementia
- Young onset dementia
- Parkinson's disease
- Mild cognitive impairment
- Huntington's disease
- Mixed dementia
- Creutzfeldt-Jakob disease
- Down syndrome
- Rarer forms of dementia

2.2.3 Risk factors for dementia

Dementia, the very common public health problem among the older people of age over 65 has reached to the epidemic proportion and has a age related prevalence (van der Flier & Scheltens, 2005). However, research shows that this epidemic could be curbed. One study conducted by Norton, Matthews, Barnes, Yaffe, and Brayne (2014) revealed that the worldwide prevalence of Alzheimer's disease could be reduced by intervening to these

seven modifiable risk factors: diabetes, physical inactivity, midlife hypertension, midlife obesity, depression, smoking, and low educational attainment. So the disease burden of dementia in the the population could be lessened by preventing /controlling these risk factors. Some studies reported other baseline covariates as the risk factors for developing dementia. Age (Deng et al., 2018; Read, Wittenberg, Karagiannidou, Anderson, & Knapp, 2017; Yang et al., 2016); ethnicity, sex, genetic factors (Corder et al., 1993; Gatz et al., 2005); physical inactivity (Baumgart et al., 2015; Mukamal et al., 2003; Yang et al., 2016); drug use and alcohol, level of education (R. Chen et al., 2011); tobacco consumption (Deng et al., 2018); comorbidity (Hypertension (Deng et al., 2018), Type 2 diabetes), personal income (R. Chen et al., 2011); and lower income (Ren et al., 2018) were also considered in determining the risk factors of dementia. Deng et al. (2018); Starkstein and Almeida (2003) also showed that around half of the individuals with vascular cognitive impairment might develop dementia. On the other hand, protective factors for dementia include higher education levels, moderate alcohol consumption (Deng et al., 2018), use of hormone replacement therapy (HRT) for women, use of anti inflammatory drugs, and diet (J.-H. Chen, Lin, & Chen, 2009).

Studies have revealed conflicting results regarding the effect of smoking and drinking on developing dementia. Some studies (R. Chen et al., 2011; Deng et al., 2018) showed smoking as a significant risk factor, while another study (D. Choi, Choi, & Park, 2018) reported the decreasing risk of dementia among individuals who quit smoking long ago or who never smoked compared to long-term smokers. On the other hand, a study conducted among elderly people in East Boston, Massachusetts, showed no association between smoking and Alzheimer's Disease. Researchers also concluded that recent mild-to-moderate alco-

hol consumption was not substantially related to the incidence of dementia (Hebert et al., 1992). The analysis of the prospective population based cohort study found that the incidence of dementia among older adults who consume one to six drinks per week is lower compared with those who abstain (Mukamal et al., 2003).

2.3 Literature review for depression

2.3.1 Introduction

WHO states depression as a common mental disorder and identifies it as one of the leading causes of disability as well as a key contributor to the global disease burden (Organization et al., 2017). Depression, a common illness is characterized by persistent sadness, loss of interest in activities, disturbance of sleep and appetite; tiredness and poor concentration. Sometimes it can lead to suicidal thought (https://www.nia.nih.gov/health/depression-and-older-adults). It is estimated that more than 264 million people of all ages around the world suffer from depression (NIH, 2020). This is a very common problem in older population but is not a normal part of aging. Depression becomes a more serious health condition when it is long-lasting with moderate to severe intensity (https://www.nia.nih.gov/health/depression-and-older-adults).

2.3.2 Types of depression

Depression usually persists for a longer time and it can be recurrent, considerably impairing the ability of an individual to continue his work or cope with daily activities. The two

main subcategories of depressive disorder are: (i) major depressive disorder/ depressive episode (ii) dysthymia (Organization et al., 2017).

Major depressive disorder/ depressive episode

Major depressive disorder includes symptoms such as mood disorder, lack of interest and enjoyment, and less energy. A depressive episode again can be classified as mild, moderate, or severe based on the number and severity of symptoms (Organization et al., 2017).

Dysthymia

Dysthymia is a chronic form of mild depression. Although the symptoms of dysthymia are similar to the symptoms of a depressive episode, comparatively it tends to be less intense (Organization et al., 2017).

Another important distinction to make regarding depression is whether an individual has a history of manic episodes or not. Bipolar affective disorder is characterized by both manic and depressive episodes separated by periods of time where the individual's mood is normal. Individuals with bipolar disorder experience elevated or irritable moods, speech pressure, over activity, euphoria, and reduced sleep (Organization et al., 2017).

2.3.3 Risk factors for depression

The consequences and public health implications of depression among the elderly are enormous. Late life depression is associated with functional disability, suicidal tendencies, and higher rates of medical morbidity and mortality (Steffens et al., 2006).

Various factors or combinations of factors may contribute to the risk of depression. The most common risk factors for depression are sex, family history, trauma or stress, and physical illnesses (NIMH, 2020; Organization et al., 2017). For example, Anand (2015)'s study of older adults in six low and middle-income countries showed the difference in the prevalence of depression between the sexes. Female respondents were found to suffer from depression at higher rates compared to males. This finding was consistent with other studies (J.-H. Lee, Park, Park, & Jo, 2018; Meng et al., 2017; Mirkena, Reta, Haile, Nassir, & Sisay, 2018). Although depression happens at any age, it is more prevalent in adulthood, particularly older people living with serious medical condition such as diabetes, CVD, cancer etc. (NIMH, 2020).

Other risk factors for depression in older adults are stressful life events, divorce, widow, living alone, low income, low educational level (Iliffe et al., 2005; J.-H. Lee et al., 2018; N. Li et al., 2011; Rajkumar et al., 2009; Yunming et al., 2012). People living in areas with higher unemployment rates and higher proportions of visible minorities consistently had an elevated risk of experiencing a major depressive disorder (Meng et al., 2017).

Individuals in poor health are more likely to suffer from depression (Cheruvu & Chiyaka, 2019; N. Li et al., 2011). Some prevalent illnesses, such as cardiovascular disease (CVD), previous head injuries, diabetes, and cancer are associated with depression among the elderly (Alamri, Bari, & Ali, 2017; Rajkumar et al., 2009; Yunming et al., 2012). Yunming et al. (2012) also showed that the odds ratio for depression was higher among those with functional impairments, more than three chronic diseases and who had experienced

an adverse life event. Multiple medical conditions such as dual sensory loss (hearing and vision) had a significant effect on depressive symptoms (Capella-McDonnall, 2005).

Other studies have identified additional risk factors. L. Li, Wu, Gan, Qu, and Lu (2016) showed the positive relationship between insomnia and depression. Similar findings were also observed in other studies (Cui et al., 2017). Chang-Quan et al. (2009) determined a set of five risk factors independently associated with depression: poor self rated health, poor cognitive status, two or more clinic visits in the past month, and slow walking speed.

Apart from these risk factors, depression was also associated with more disability, poorer life satisfaction, alcohol intake, smoking and drug use, physical inactivity, unhealthy eating styles, living with children and medical comorbidity (Haseen & Prasartkul, 2011; He et al., 2014; J.-H. Lee et al., 2018; Meng & D'Arcy, 2013; Mirkena et al., 2018; Organization et al., 2004; Rawana, Morgan, Nguyen, & Craig, 2010; Subramaniam et al., 2016).

Chapter 3

Statistical Methods and Materials

3.1 Survival analysis

In health science and biomedical research, time to a certain event is the primary endpoint of interest. Survival analysis usually focuses on analyzing data where time-to-event is our main interest. Survival analysis deals with the time that passes from a well-defined time origin to the occurrence of specific event. This type of data is often called life time, failure time, or survival data (Collett, 2015; Lawless, 2011).

Survival analysis comprises the wide range of methods for dealing with the timing of events. Mostly, it includes the technique for positive valued random variables. For examples:

- time to diagnosis of dementia,
- time to death,
- time to failure of a machine,

- time from treatment to cure,
- time from remission to relapse of a disease, and
- time from HIV infection to AIDS

In survival analysis, interest often lies in determining the features and characteristics of the distribution of time-to-event for a given population. It also includes the statistical comparison of time-to-event among different groups (e.g., intervention vs. control group in clinical studies). Researchers are also often interested in examining the relationship of time-to-event to other potential covariates and quantifying their association through different modeling approaches (Clark, Bradburn, Love, & Altman, 2003).

3.1.1 Survival data and censoring

Standard statistical procedures are not entirely appropriate in dealing with the survival data. One reason why is that survival times are generally non-negative and are not symmetric. Usually, histogram constructed from survival time is positively skewed, that is, it has a longer tail to the right side, so it is not reasonable to consider the normality assumptions of this type of data (Collett, 2015).

Another key feature of survival data that makes conventional statistical methods inappropriate is survival times are censored. It is common in survival analyses that complete information for all observations is not available during the specified time frame; rather, some information is only partially observed, leading to what is known as "censoring" (Kleinbaum & Klein, 2011).

The followings are common reasons for censoring (Collett, 2015):

- Termination of the study before the event occurs,
- Death due to a cause not related to the event of interest,
- Failure to experience the event before the study ended, and
- Individual lost to follow-up (e.g., patient emigrates or no longer traced).

Types of censoring

There are three main types of censoring (Lawless, 2011):

- Right censoring
- Left censoring
- Interval censoring

Right censoring

Right censoring describes a situation when individuals do not experience the event before the termination of the study or are lost to follow-up (Lawless, 2011). For example, in a study of pregnancy duration, if some women are still pregnant at the end of the study or some are lost to follow-up, these observations will be right-censored. Right-censoring can further be classified as Type I, Type II, and random censoring scheme (Indrayan, 2012). Most of survival analyses deal with random censoring. This study also focused on random censoring.

Left censoring

If the actual survival time is less than what is observed by the investigator, it is called left-censoring (Lawless, 2011). For example, consider a study observing the time to the

recurrence of a particular cancer after the surgical removal of tumor. Patients are re-examined three months of their operation to see whether their cancer has reappeared. However, if some cancers have already recurred (that is, the time to the recurrence is less than three months), these observations are left-censored (Collett, 2015).

Interval censoring

If the event of interest occurs in a time interval (left, right), but we do not know exactly when is this interval, it is often call interval censoring (E. T. Lee & Wang, 2003). For instance, in a study of HIV surveillance, a subject might have two tests, where the test result is negative at first visit (say t_1) but the result is positive at second time (t_2). In this case, actual survival time lies between this two time points and this is interval censored in the interval of (t_1, t_2).

3.1.2 Terminology, notation and some important concepts

Survival function

Let T be a random variable which is continuous, non-negative and represents the true life time with probability density function (p.d.f.), $f(t)$ and cumulative distribution function (c.d.f.), $F(t) = Pr(T < t)$, determining the probability of occurring event by time t.

Survival probability, $S(t)$ is defined as the probability that a person survives longer than a specified time (say t) or is alive just before duration t, or more and it is expressed by survival function as follows (Lawless, 2011)

$$S(t) = Pr(T \geq t) = 1 - F(t) = \int_t^{\infty} f(x)dx \tag{3.1}$$

with t ranging from 0 to ∞, the survival function has the following characteristics (Lawless, 2011)

- It is non-increasing,
- At time $t = 0, S(0) = 1$, and
- At time $t = \infty, S(\infty) = 0$

Hazard function

Hazard function, $h(t)$ an alternative characteristic of the distribution of T, is defined as the probability of failure during a small interval of time given no previous events. It is also called the instantaneous failure rate (Lawless, 2011).

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{3.2}$$

The cumulative hazard, $H(t)$ represents the aggregated risk up to time t and expressed as (Lawless, 2011)

$$H(t) = \int_0^t h(u)du \tag{3.3}$$

and the survival function in-terms of cumulative hazard function can be defined as follows (Lawless, 2011)

$$S(t) = \exp(-H(t)) = \exp(-\int_0^t h(u)du) \tag{3.4}$$

3.1.3 Non-parametric estimate of survival function

The Kaplan–Meier (K-M) estimator (Kaplan & Meier, 1958), is one of the non-parametric approaches for estimating survival function and comparing survival functions between two or more groups. It is also known as the product limit estimator. Let us consider an n random sample and there are k ($k \leq n$) distinct life times $t_1 < t_2 < ... < t_k$ at which death occurs. More than one death at $t_j, j = 1, 2, ..., k$ is also allowed. Suppose d_j indicates the number of individuals who die at t_j, then the K-M estimator, $\hat{S(t)}$ takes the following form:

$$\hat{S(t)} = \prod_{j:t_j \leq t} (1 - \frac{d_j}{n_j}) \tag{3.5}$$

where n_j indicates the number of individuals at risk set (alive and not censored) just prior to time t_j.

Log-rank test

Plots of K-M curves roughly tell us which group has a better survival prognostic than that of other group. However, statistical tests are required to examine whether or not the K-M curves are statistically equivalent (Kleinbaum & Klein, 2011). A log-rank test is a commonly used non-parametric test to compare two or more survival curves. At each observed failure time, the estimated hazard functions of two groups are compared under the proportionality assumption (Bland & Altman, 2004). Under the null hypothesis of all equivalent survival curves, this log rank statistic is approximately chi-square and determines the P-value from tables of chi-square distribution (Kleinbaum & Klein, 2011). The log-rank test is more powerful if the proportionality assumption holds (Hazra & Gogtay, 2017).

Wilcoxon test

There are a few alternative techniques to the log-rank test for testing the hypothesis that two or more survival curves are equivalent. The wilcoxon test is one of them and it is more powerful when hazard functions are not proportional. Unlike log-rank test, where equal weight is given to deaths at all time points, the wilcoxon method gives more weight to deaths at the beginning of the survival curve than later failures (Kleinbaum & Klein, 2011).

3.1.4 Survival models

One of the most interesting research areas in health science is determining the association between potential covariates and time-to-event. Conventional regression approaches are not appropriate for handling the life time data due to the censoring issue and non-normal response. In this situation, Cox proportional hazard model for failure time is often recommended as it does not require assumptions about underlying survival distribution (Schober & Vetter, 2018).

Cox proportional hazards model

Cox proportional hazards (PH) model (Cox, 1972) is the most widely used statistical method to measure the impact of different covariates on failure time. The standard Cox model deals with time independent covariates. However, if the covariate changes over time (time dependent), the Cox proportional hazards model is inappropriate. The PH model describing the hazard for failure time, T with covariate vectors X, is assumed to be (Lawless, 2011; E. T. Lee & Wang, 2003).

$$h(t) = h_0(t) \exp\{\beta^T X\} \tag{3.6}$$

where, $X = (X_1, X_2, ..., X_p)^T$ is p vector of covariates and $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ denotes the corresponding coefficients of these covariates. In equation (3.6), $h_0(t)$ is the baseline hazard function that describes the risk for individuals for $x = 0$ and it usually remains unspecified. One of the interesting properties of cox model is that that it is possible to obtain estimates for β 's, even though $h_0(t)$ in the equation (3.6) is unspecified. We need to estimate β and the measure of effects, often expressed in term of hazard ratio, e^{β}, to determine the effect of exposure or explanatory variables. The PH assumption for the Cox model in the above equation requires the hazard ratio to be constant over time, that is, hazard is always proportional to each other irrespective the value of time t.

Let T be the actual event time of interest, C denote the potential censoring time (meaning that the study subject cannot be observed beyond the time of the study), and t the actual observing time. Suppose the true life times of n individuals are represented by random variables $T_1, ..., T_n$. Let us consider $t_i = min\{T_i, C_i\}$ where T_i, C_i are assumed to be independent and t_i is known as either life time or censoring time. Define another variable, $\delta = I(T_i < C_i)$ as an indicator for failure, where $I(.)$ is called an indicator function. $\delta = 1$ indicates failure time and 0 if it is censored, it tells whether $t_i, i = 1, 2, .., n$ is observed life time or censoring time. So we only observe $(t_1, \delta_1), ..., (t_n, \delta_n)$ instead of $T_1, ..., T_n$, the true life times.

Likelihood estimation for PH model

Suppose, the censored random sample (t_i, δ_i), $i = 1, 2, ..., n$, has the k distinct failure times, $t_{(1)} < t_{(2)} <, ...t_{(k)}$ and $n - k$ are censored observations. Let $R_i = R(t_{(i)})$ be the risk set at time $t_{(i)}$. This risk set indicates the group of individuals who are alive or unenclosed prior to $t_{(i)}$. Cox (1972, 1975) proposed the following partial likelihood approach for estimating the parameters of β without the involvement of $h_0(t)$

$$L(\beta) = \prod_{i=1}^{k} \left[\frac{h_i(t_i|X_{(i)})}{\sum\limits_{l \in R_i} h_l(t_i|X_l)} \right] = \prod_{i=1}^{n} \left[\frac{\exp\{\beta^T X_{(i)}\}}{\sum\limits_{l \in R_i} \exp\{\beta^T X_l\}} \right] \tag{3.7}$$

where X_i is associated covariates with individuals dying at t_i. If a death occurs at t, then the probability that it will be individual $l \in R_t$ who dies is (Lawless, 2011)

$$\frac{h(t|X_i)}{\sum\limits_{l \in R(t)} h((t)|X_l)} = \frac{\exp\{\beta^T X_{(i)}\}}{\sum\limits_{l \in R(t)} \exp\{\beta^T X_l\}}$$

However, it is also reasonable to consider that more than one death at time $t_{(j)}$ is possible where , $j = 1, 2, ..., k$ are distinct life times. For this purpose we need to define the followings

$Y_i(t) = I(t_i \geq t)$, $i = 1, ...$

Then $Y_i(t) = 1$ if $i \in R(t)$ and the equation (3.7) can be rewritten

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\{\beta^T X_i\}}{\sum_{l=1}^{n} Y_l(t_i) \exp\{\beta^T X_l\}} \right]^{\delta_i} \tag{3.8}$$

Now, the log-partial likelihood function of the form from equation (3.8) is:

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left[\beta^T X_i - \log \left\{ \sum_{l=1}^{n} Y_l(t_i) \exp\{\beta^T X_l \right\} \right] \tag{3.9}$$

Define the score vectors as $\mathbf{U(\beta)}$=$(\partial l/\partial \beta_1, ..., \partial l/\partial \beta_p)'$ and for any $t > 0$, define $p \times 1$

vector as (Lawless, 2011)

$$\bar{X}(t,\beta) = \frac{\sum_{l=1}^{n} Y_l(t) X_l \exp\{\beta^T X_l\}}{\sum_{l=1}^{n} Y_l(t) \exp\{\beta^T X_l\}}$$

which indicates the weighted mean of the covariate vectors for the individuals who are at risk at time t. The score takes the following straightforward form, defined as

$$U(\beta) = \sum_{i=1}^{n} \delta_i [X_i - \bar{X}(t_i,\beta)] \tag{3.10}$$

and the form of $p \times p$ information matrix $\boldsymbol{I(\beta)} = -\partial^2 l/\partial\beta\partial\beta'$ can be obtained as follows (Lawless, 2011)

$$I(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{l=1}^{n} Y_l(t_i) \exp\{\beta^T X_l\}[X_l - \bar{X}(t_i,\beta)][X_l - \bar{X}(t_i,\beta)]'}{\sum_{l=1}^{n} Y_l(t_i) \exp\{\beta^T X_l\}} \right\} \tag{3.11}$$

Iterative methods, such as Newton-Raphson method (Collett, 2015) is generally used to solve the score equation with a view to determining the estimates of β.

3.2 Longitudinal modeling

The key feature of longitudinal study is to measure the same outcome from same individual at multiple times or measurement from same subject taken repeatedly over the time. The primary objective of this kind of study is to characterize changes in response over time and identify the factors that influence these changes (Fitzmaurice, Laird, & Ware, 2012). The outcome variable in longitudinal study can be continuous, binary, or count. There may also be an incomplete data set due to missing/dropout.

Since responses are measured at different time points, the correlation due to repeated measurements from the same individual often violates the assumption of independence under general linear model (GLM). Therefore, in order to draw valid scientific inferences, some special statistical methods for longitudinal studies are required to take into account this correlation.

The statistical areas that address these correlated data includes (i) repeated measures analysis, (ii) linear mixed models, (iii) generalized linear mixed models, and (iv) multilevel models (Fitzmaurice et al., 2012). This thesis focuses on the random effects model, one of three generalized linear models for longitudinal data (the other models are marginal model and transitional model). Generalized linear mixed models (GLMM) can be understood as extensions of the generalized linear model to correlated data. The important feature of GLMM is its emphasis on discrete response (binary or count), although continuous response is a special case under GLMM. However, making assumption about multivariate or joint distribution is an appropriate way to deal with correlated longitudinal data. This multivariate distribution can be specified by three different modeling approaches: (i) marginal models, (ii) mixed effects models, and (iii) transitional models. Of these, transitional models are the least popular for modeling the effects of covariates because inferences made based on these models could be misleading if treatment or exposure alters the risk level during the observation period. Marginal models are used to make inference about the population averages (Fitzmaurice et al., 2012). Based on some assumptions regarding covariance structure of observations, this approach is mainly used to model the mean response conditioning on covariates but not on random effects. On the other hand, the random effects model is used to make inferences about the individual

level, rather than the general population level (Fitzmaurice et al., 2012).

3.2.1 Generalized linear mixed effects model (GLMM)

A generalized linear mixed model (GLMM) is an extension of a generalized linear model (GLM) that contains random effects along with the fixed effects. This modeling approach takes the shared random effects that results from the repeated measurements on the same individual into account (Fitzmaurice et al., 2012). The basic assumption is that there is natural heterogeneity across individuals and that a subset of regression coefficients (for example, random intercept and slope) is assumed to vary from individual to individual according to some distribution. Here we also assume that, given the random effects, the data for an individual are independent and drawn from a distribution that belongs to an exponential family (Fitzmaurice et al., 2012).

The specifications of generalized linear model with random effects are: (Fitzmaurice et al., 2012)

Suppose that, $y_i = (y_{i1}, y_{i2}, ...y_{im_i})$ is the $m_i \times 1$ vector of m_i correlated longitudinal responses for the i^{th} subject, that is, response y_{ij} for the i^{th} subject is measured at j^{th} time point, denoted by t_{ij}, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., m_i$. Here, n is total number of subjects in the longitudinal study. Let $b_i = (b_{i0}, b_{i1}, ...b_{iq})$ be the $(q + 1) \times 1$ vector of random effects and this unobservable variables, b_i is mainly responsible for the correlation among observations for a single individual. However, the response $(y_{i1}, y_{i2}, ...y_{im_i})$ are conditionally independent given the random effects $b_i = (b_{i0}, b_{i1}, ...b_{iq})$, and follows a

distribution belonging to an exponential family of density defined as (J. Choi et al., 2015; Fitzmaurice et al., 2012)

$$f(y_{ij}|b_i\beta) = \exp[\{(y_{ij}\theta_{ij} - \psi(\theta_{ij}))/\phi + c(y_{ij}, \phi)\}] \tag{3.12}$$

where $\psi(.)$ and $c(.,.)$ are known functions, ϕ is the dispersion parameter and θ is canonical parameter. Define the conditional mean as, $\mu_{ij} = E(y_{ij}|b_i) = \psi'(\theta_{ij})\phi$ and variance, $v_{ij} = Var(y_{ij}|b_i) = \psi''(\theta_{ij})$, like glm here this form has $g(\mu_{ij}) = x^T\beta + z_{ij}^T b_i$ and $v_{ij} = v(\mu_{ij}\phi)$ where $g(.)$ and v is the link and variance function receptively, $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ is vector of parameter for fixed effect, x_{ij} and z_{ij} are the vectors of covariate with length $(p+1)$ and $(q+1)$ respectively. z_{ij} may be the subset of x_{ij} (J. Choi et al., 2015).
It is also assumed that the random effect, $b_i, i = 1, 2, ..., n$, are iid with density function $f(b_i; \Sigma)$, where Σ is variance covariance matrix of random effects.

Consider the response, y_{ij} $i = 1, 2, ..., n$, $j = 1, 2, ..., m_i$ for the i^{th} subject at j^{th} time point, random effects, b_i, p covariates, x's for fixed effects and q covariates, z's for random effect coefficients. Now, using the link function $g(.)$, the relationship can be established as follows

$$g(E(y_{ij}|b_i)) = g(\mu_{ij}) = \beta_0 + \beta_1 x_{ij1} + ...\beta_1 x_{ij1p} + b_{i0} + b_{i1}z_{ij1}, ...b_{iq}z_{ijq} \tag{3.13}$$

Here, $E(.)$ is used to indicate the expected value. It is important to note that the consequences of dependence among the longitudinal measurements from the same subject are considered through the working correlation structure. However, in GLMM, the sources of

dependence among the measurements on same subject are explained by random effects instead of the working correlation structure (Fitzmaurice et al., 2012). The common choice of distributional form of random effects, b_i in equation (3.13), is that it follows normal distribution, usually expressed as $b_i \sim \mathcal{N}(0, \Sigma(\theta))$, where $\Sigma(\theta)$ has a known form with unknown parameters such as $\Sigma(\theta) = \theta\mathbf{I}$, where $\mathbf{I}$ is an identity matrix with dimension of $(q+1) \times (q+1)$. The possible covariance structure for random effects includes unstructured, Toeplitz, compound symmetry, and first order auto-regressive (AR1).

Since this thesis deals with binary longitudinal outcome, including a specific example here is useful. Consider the longitudinal data set with dichotomous response, $y_i = (y_{i1}, y_{i2}, ...y_{im_i})$, $(m_i \times 1)$ vector of m_i binary repeated measurements for the i^{th} subject, that is, response y_{ij} for the i^{th} subject is measured at j^{th} time point, t_{ij} where $i = 1, 2, ..., n$ and $j = 1, 2, ..., m_i$. Now for the given random effects b_i (say intercept only), the response y_{ij} follows a Bernouli (Binomial) distribution with probability of success (Hwang, Huang, Wang, Lin, & Tseng, 2019)

$$P(y_{ij} = 1|b_i) = \frac{e^{\beta_0+\beta_1 x_1+b_{i0}}}{1 + e^{\beta_0+\beta_1 x_1+b_{i0}}}, \tag{3.14}$$

where $b_{i0} \sim \mathcal{N}(0, \sigma_b^2)$, Here, b_{i0} in equation (3.14) is the intercept. The model defined in this equation is sometimes called a random intercept logistic regression model.

3.2.2 Likelihood based inference in GLMMs

Random effects, b_i defined in equation (3.13) is assumed as unobserved variables that is to be integrated out of the likelihood in maximum likelihood method and they are also treated as a sample of independent variables from the distribution of random effects. The

joint distribution of responses y_{ij}'s, $i = 1, 2, ..., n$, $j = 1, 2, ..., m_i$ is (Fitzmaurice et al., 2012)

$$f(\mathbf{y}|\beta, \theta) = \prod_{i=1}^{n} f(y_i|\beta, \theta) = \prod_{i=1}^{n} \int \{\prod_{j=1}^{m_i} f(y_{ij}|b_i, \beta, \theta)\} f(b_i|\theta) db_i \tag{3.15}$$

Now the likelihood function can be expressed as (Fitzmaurice et al., 2012)

$$L(\beta, \theta|\mathbf{y}) = \prod_{i=1}^{n} \int \{\prod_{j=1}^{m_i} f(y_{ij}|b_i, \beta, \theta)\} f(b_i|\theta) db_i \tag{3.16}$$

Therefore, the log-likelihood function is

$$l(\beta, \theta|\mathbf{y}) = \log f(\mathbf{y}|\beta, \theta) = \sum_{i=1}^{n} \log \left[\int \{\prod_{j=1}^{m_i} f(y_{ij}|b_i, \beta, \theta)\} f(b_i|\theta) db_i \right] \tag{3.17}$$

Now maximum likelihood estimates (MLEs) can be obtained by maximizing this log likelihood function that is:

$$(\hat{\beta}_{MLE}, \hat{\theta}_{MLE}) = argmax\ l(\beta, \theta|\mathbf{y})$$

Equation (3.16) basically reflects the kind of marginal distribution of $\mathbf{y}$ that is obtained from the integration of joint distribution of $\mathbf{y}$ and b with respect to b. For linear mixed model, there is a closed form for the integral in (3.16) and it is computationally feasible. However, numerical methods are needed to compute the integration in (3.16), particularly for non-Gaussian models in GLMMs. Here, computing is still feasible if the number of random effects, q is small but computation is really expensive for large number of random effects (Schabenberger et al., 2005; Ye & Wu, 2017). There are several approaches to deal with many random effects in GLMMs models including: Monte Carlo methods, approximation approaches, Bayes methods (McCulloch, 1997; Schabenberger et al., 2005).

However, the EM algorithm is a very common and widely used approach to find MLE of β, θ (J. Choi et al., 2015).

3.3 Joint modeling of longitudinal and time-to-event data

There are two situations where joint modeling is useful: (i) when a study interest lies in evaluating the repeated measurements over time and analyzing the longitudinal response with time to the termination of the study, adjusting for informative dropouts, and (ii) when measuring the impact of longitudinal surrogate marker on time-to-event, adjusting for the measurement error/ missing value of this longitudinal response (Garcia-Hernandez & Rizopoulos, 2018). Furthermore, this modeling approach is also employed when a survival and a longitudinal process are associated through latent variables and interest lies in determining and quantifying this association (Wu et al., 2012).

This joint modeling approach works through a group of latent variable models by which the association structure between outcomes is modeled. Normally-distributed random effects, b_i are used to capture the association between the longitudinal and the event model. Conditioning on random effects b_i, the longitudinal response y_i and the time-to-event responses T_i, are assumed to be independent and thus defined as follows (Garcia-Hernandez & Rizopoulos, 2018)

$$p(T_i, \delta_i, y_i|b_i, \theta) = p(T_i, \delta_i|b_i, \theta_T, \theta_y)p(y_i|b_i, \theta_y) \tag{3.18}$$

where T_i is the observed survival time (time-to-event) or time-to-censoring, δ_i is censoring indicator (denoting either event or censoring), $y_i = (y_{i1}, y_{i2}, ..., y_{im_i})$ the collection of m_i observations in the longitudinal response for subject $i, i = 1, 2, ...n$ where, n is the total number of subject in the study and $\theta = (\theta_T, \theta_y)$ is the vector of parameters from two parts: (i) longitudinal and (ii) survival model. Suppose the set of parameters from the model of the time-to-event response is defined by θ_T and for the longitudinal model, by θ_y.

3.3.1 Longitudinal submodel

Suppose $y_{ij} = y_i(t_{ij}),\ i = 1, 2, ..., n;\ j = 1, 2, ..., m_i$ indicates the longitudinal response for the i^{th} subject measured at j^{th} time point, defined as, t_{ij}, where n is the total number of study subjects and m_i is the number of repeated measurements for the i^{th} subject. Consider the following general conditional mixed models framework that uses latent random effects vectors b_i defined in section (3.2.1), independent across subjects, to model within-subject covariance. Given random effects b_i, the longitudinal responses, $y_i = (y_i(t_{i1}), ..., y_i(t_{im_i}))$ on the same subject are assumed to be independent (Garcia-Hernandez & Rizopoulos, 2018).

$$p(y_i|\theta_y, b_i) = \prod_j p(y_i(t_{ij})|b_i\theta_y) \tag{3.19}$$

where θ_y denotes the vectors of parameters from longitudinal model defined in equation (3.13) under section 3.2.1

More specifically, to incorporate the binary repeated measurements into the Cox model, consider the Bernoulli distribution with the probability of success defined as (Garcia-

Hernandez & Rizopoulos, 2018)

$$\pi_i(t) = \frac{e^{\beta_0+\beta_1 t+\beta_2 X_i+b_{i0}+b_{i1}t}}{1+e^{\beta_0+\beta_1 t+\beta_2 X_i+b_{i0}+b_{i1}t}} \tag{3.20}$$

where X_i is the time-independent covariate associated with i^{th} subject, β's are the regression coefficients for the fixed effects, b_i's are random effects. The model in equation (3.20) includes only two random effects (random intercept, b_{i0} and slope, b_{i1}) in addition to fixed effects t, X_i.

Therefore, the logit function can be expressed as follows (Garcia-Hernandez & Rizopoulos, 2018)

$$logit(\pi_i(t)) = \log(\frac{\pi_i(t)}{1-\pi_i(t)}) = \beta_0 + \beta_1 t + \beta_2 X_i + b_{i0} + b_{i1}t \tag{3.21}$$

More specifically, generalized linear mixed effects model can be represented as (Garcia-Hernandez & Rizopoulos, 2018):

$$g\{m_i(t)\} = g[E\{y_i(t)\}|b_i] = X_i(t)(\beta_t + b_i) + Z_i\beta_b \tag{3.22}$$

where $m_i(t)$ is the expected value for the longitudinal response, $y_i(t)$ of subject i at time t given the random effects, b_i, $g()$ denotes the link function, $X_i(t)$ is the design matrix of fixed and random effects modeling the trajectories of the longitudinal response over time. The corresponding coefficients for these fixed and random effects are β_t and b_i respectively, and Z_i, β_b are the design matrix and coefficients associated with the baseline covariates respectively.

Random intercepts and slopes

There are several options available for random-effects linear models, particularly for modeling the trajectories of longitudinal variable over time. If we consider the simple random intercept and slope model, it includes only β_0 fixed-effect intercept, β_1 fixed-effect slope, b_{i0}, intercept for random-effect; b_{i1} is slope for random-effect and β_b the vector of coefficients associated with baseline covariates, Z and can be expressed as (Garcia-Hernandez & Rizopoulos, 2018)

$$g\{m_i(t)\} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})t + Z_i\beta_b \tag{3.23}$$

The variance covariance matrix of b_{i0} and b_{i1} is denoted by Σ and defined as follows:

$$\Sigma = \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{bmatrix}$$

where σ_{00} and σ_{11} indicates the variance of random effect b_{i0} and b_{i1} receptively and σ_{01} is their covariance. However, if the random effects are assumed to be independent then the covariance term will be zero and no longer to be estimated. On the other hand, if the number of random effects increases, possible structures for covariance matrix, Σ may be unstructured, variance-components etc. Estimation procedures are discussed in section 3.2.1.

3.3.2 Time-to-event submodel

The variable of interest is the time to the event, $\tilde{T}_i$ that is, true survival time for the i^{th} subject and it might not be observed for all subjects. Two variables are observed: time-to-event or time to censoring T_i and event indicator δ_i=I($\tilde{T}_i \leq C_i$). The general expression for Cox model with a covaraite, x_i associated with i^{th} subject can be defined as follows (Cox, 1972)

$$h_i(t) = h_0(t) \exp(\lambda_1 x_i) \tag{3.24}$$

where $h_0(t)$ is the baseline hazard that is to be specified, x_i indicates the single covariate and λ_1 represents the corresponding regression coefficient. In the subsequent section, a single covariate, x_i in the Cox model is considered for simplicity, however, multiple covariates are also possible as like as the form of Cox model in equation (3.6).

Parametric survival models

The Cox model for analyzing survival data does not need to assume a specific probability distribution for survival time and there is no restriction on the functional form of the hazard function. The Cox model, therefore, has widespread application. However, if the assumptions regarding the probability distribution for the data are correct and valid, inferences will be more precise compared to the Cox model (Collett, 2015). For example, the standard error of the estimates of measures such as median survival time and the hazard ratio will be smaller than that of the model without distributional assumptions. The models with the assumption of parametric distribution for the survival time are called parametric model. This parametric version of the Cox model, where baseline hazard is specified with a parametric model, is discussed as follows:

Exponential

Assuming that the constant baseline risk is the simplest form for time-to-event of interest and that it is defined as follows (Kleinbaum & Klein, 2011; Lawless, 2011)

$$h_0(t) = \lambda \tag{3.25}$$

The full expression of the exponential model, that is, the hazard for the i^{th} subject is given as

$$h_i(t) = \lambda \exp\{\lambda_1 x_i\} \tag{3.26}$$

Here, the baseline hazard, $h_0(t)$ in cox model in equation (3.24) is replaced with λ.

Piecewise exponential

In most cases, the baseline hazard risk, $h_0(.)$ remains unspecified in the Cox model. However, this un-specification of baseline hazard risk in joint modeling may underestimate the standard errors of the parameter estimates (Hsieh, Tseng, & Wang, 2006) and it is always better idea to incorporate this function. Piecewise-constant model with following hazard function could be a standard approach (Garcia-Hernandez & Rizopoulos, 2018).

$$h_0(t) = \sum_{q=1}^{Q+1} \xi_q I(v_{q-1} < t < v_q) \tag{3.27}$$

where $v_0 = 0, v_1, ..., v_Q$ are Q knots that divide the range of times into many $Q+1$ intervals, the value of v_{Q+1} is higher than the maximum observed time and ξ_q is the hazard in the interval of (v_{q-1}, v_q). For example, the hazard has following form for different values of q

$$
\begin{aligned}
h_0(t) = & \xi_1, if, v_0 < t < v_1 \\
& . \\
& . \\
& . \\
& \xi_{Q+1}, if, v_Q < t < v_{Q+1}
\end{aligned}
\tag{3.28}
$$

The full expression of piecewise exponential model is

$$
h_{iq} = \xi_q \exp\{\lambda_1 x_i\} \tag{3.29}
$$

where h_{iq} indicates the hazard corresponding to the i^{th} individual at q^{th} interval and ξ_q is the constant baseline hazard in the q^{th} interval but it may vary across the interval. The closer form of the unspecified Cox model can be approximated with piecewise exponential model. Maintaining at least 10–20 failure per parameter is recommended to avoid the underestimation of standard errors (Harrell Jr, 2015).

Weibull

This is a very popular and commonly used distribution for baseline risk particularly for describing a scenario where the hazard risk increases or decreases over time. The hazard is given as (Kleinbaum & Klein, 2011; Lawless, 2011)

$$
h_0(t) = \frac{k}{\lambda^k} t^{k-1} \tag{3.30}
$$

where λ is the scale and k is the shape parameter, and $\lambda, k > 0$. The hazard increases when $k > 0$ and decreases when $k < 1$ with time. For $k = 1$, the hazard rate remains constant and returns to the exponential model.

The hazard for the i^{th} subject with a covariate x_i under the Weibull model is defined as

$$h_i(t) = \frac{k}{\lambda^k} t^{k-1} \exp\{\lambda_1 x_i\} \tag{3.31}$$

3.3.3 Joint model with current-value shared parameter

The most standard approach for linking the longitudinal model and the time-to-event model in the random-effects shared-parameter models framework is to assume that the risk of event at a given time t depends on the estimated value for the longitudinal response at that time. This is expressed as follows (Garcia-Hernandez & Rizopoulos, 2018)

$$h_i(t) = h_0(t) \exp\{\lambda_1 x_i + \alpha_1 \pi_i(t)\} \tag{3.32}$$

where scalar λ_1 indicates the regression coefficient corresponding to the covariate, x_i , $\pi_i(t)$ is defined in equation (3.20) in subsection (3.3.1) and α_1 is the single shared parameter that links the longitudinal process with the survival process. If the baseline hazard, $h_0(t)$ in equation (3.32) is specified with exponential model, we see that, once the shared parameter α_1 is incorporated between the subject-specific estimate of longitudinal response, $\pi_i(t)$ and event time process T_i, the resulting hazard for the i^{th} subject is no longer constant, rather it changes with the change of longitudinal response over time.

3.3.4 Joint model with slope-dependent shared parameter

Sometimes it may not be the always case that the longitudinal response is associated with the risk of event through the currentvalue shared parameter, α_1 in equation (3.32) but the rate of range (increase or decrease) of the longitudinal response may alter the risk

of event. This relationship can be expressed as follows (Garcia-Hernandez & Rizopoulos, 2018)

$$h_i(t) = h_0(t) \exp\{\lambda_1 x_i + \alpha_2 \pi_i'(t)\} \tag{3.33}$$

$h_i(t)$ specifies that the hazard for i^{th} subjects at time t is now assumed to be associated with its current rate of change (i.e., slope of longitudinal response) at time t, denoted by $\pi_i'(t)$. $\pi_i'(t)$ is a derivative of equation (3.20) with respect to t. Here, α_2 is a corresponding coefficient to $\pi_i'(t)$.

Characteristics of current-value and slope-dependent parameter

The basic features of joint model with current-value shared parameter and slope-dependent parameters are described briefly in the following table

Table 3.1: Key features of current-value and slope-dependent model

Current-value Parameter	Slope-dependent Parameter
1. This association structure assumes that the log hazard of event at time t is linearly associated with the true value of the longitudinal response evaluated at that time, t.	1. This association structure assumes that the log hazard of event at time t is linearly associated with the current rate of change (slope) of the longitudinal submodel's linear predictor.
2. The true value of the longitudinal measure at time t is predictive of the risk of experiencing the event at that same time t.	2. The true value of the slope at time t is predictive of the risk of experiencing the event at that same time t.
3. In the situation when statistical models incorporate the "true" value of the biomarker as a time-dependent covariate in the event submodel, the current-value association structure is considered in the model.	3. On the other hand, slope-dependent parameter is meaningful to use when the longitudinal scores of individuals at a specific time are same but they have different rate of change of this score (for example, one may have increasing trajectory, another may have decreasing trajectory).
4. The current-value shared parameter, α_1 in equation (3.32) is interpreted interms of hazard ratio such as the hazard ratio for one unit increase in current-value of longitudinal response is $\exp(\alpha_1)$.	4. The slope-dependent parameter, α_2 in equation (3.33) is interpreted interms of hazard ratio such as the hazard ratio for one unit increase in the current rate of change of the (true) trajectory is $\exp(\alpha_2)$.

3.3.5 Marginal log-likelihood function and inference

Let $y_i(t_{ij}), i = 1, 2, ..., n; j = 1, 2, ..., m_i$ be the longitudinal response for the i^{th} subject measured at j^{th} time point, that is, time t_{ij}, where n is total study subjects and m_i is the

number of repeated measurements for the i^{th} subject. Suppose the random effects, $b_i = (b_{i0}, b_{i1}, ...b_{iq})$ has multivariable normal distribution with mean vector, $\mu = (\mu_0, ..., \mu_q)^T$ and (q+1)$\times$ (q+1) variance covariance matrix, Σ. Here, q indicates the number random effects considered in the model. Usually random effects, b_i is assumed to follow multivariate normal distribution with mean vector, 0 and variance covariance matrix, Σ. Based on the discussion and assumptions described in the joint modeling section and subsection on the longitudinal submodel (subsection 3.3.1) and time-to-event submodel (3.3.2), the marginal (joint) likelihood of the observations on subject i, is expressed as follows (Garcia-Hernandez & Rizopoulos, 2018):

$$p(T_i, \delta_i, y_i|\theta_y, \theta_T) = \int p(T_i, \delta_i|b_i, \theta_y, \theta_T)[\prod_{j=1}^{m_i} p(y_i(t_{ij})|b_i, \theta_y)]p(b_i, \theta_b)db_i \tag{3.34}$$

where T_i is failure or censoring time, δ_i is the censoring indicator, θ_T and θ_y indicates the set of parameter from time-to-event and longitudinal model described in section 3.3.2 and section 3.3.1 respectively. θ_b denotes the unique parameters of the random effects covariance matrix, Σ.

Now the joint likelihood function for the observable data can be constructed as follows and model estimators from the likelihood function can be obtained. Thus, the joint likelihood function (Hwang et al., 2019) is

$$L(\theta) = \prod_{i=1}^{n} p(T_i, \delta_i, y_i|\theta_y, \theta_T) = \prod_{i=1}^{n} \int p(T_i, \delta_i|b_i, \theta_y, \theta_T)[\prod_{j} p(y_i(t_{ij})|b_i, \theta_y)]p(b_i, \theta_b)db_i \tag{3.35}$$

where

$p(y_i(t_{ij})|b_i, \theta_y) = \pi(t_{ij})^{y_i(t_{ij})}(1 - \pi(t_{ij}))^{(1-y_i(t_{ij}))}$,

$p(b_i|\mu, \Sigma) = \frac{\exp\{-(b_i-\mu)'\Sigma^{-1}(b_i-\mu)/2\}}{(2\pi)^{q/2}|\Sigma|^{1/2}}$ and

$p(T_i, \delta_i|b_i, \theta_y, \theta_T) = [h_0(T_i)\exp\{\lambda_1 x_i + \alpha_1\pi_i(t)\}]^{\delta_i} \times \exp[-\int_0^{T_i} h_0(s)\exp\{\lambda_1 x_i + \alpha_1\pi_i(s)\}ds]$

For the log-likelihood function, we will have log transformation both for the longitudinal and survival portion. The log transformation of probability function for longitudinal response conditional on random effects always has a closed form solution and for the time-to-event model, this log transformation of probability function conditional on random effects parameters can expressed as

$$log\{p(T_i, \delta_i|b_i, \theta_y, \theta_T)\} = \delta_i log\{h_i(t_i|b_i, \theta)\} - \int_0^{T_i} h_i(T_i|b_i, \theta)dt$$

The integration of the second part of the hazard function ranging from 0 to T_i may or may not have the closed form solution depending on the chosen model for both longitudinal and time-to-event outcome. Usually the 15-point Gauss-Kronrod rule is used for approximating this integration and is defined as follows (Garcia-Hernandez & Rizopoulos, 2018).

$$\int_0^{T_i} h_i(T_i|b_i, \theta)dt \approx \frac{T_i}{2}\sum_1^{15} w_k h_i(t_{ik}|b_i, \theta) \tag{3.36}$$

where t_{ik} are called 15 Kronrod-rule nodes for the integral re-scaling from -1 to 1 into 0 to T_i interval, and w_k are the weights for the integral from -1 to 1. However, if the $m_i(t)$, the true longitudinal response, is modeled with spline it requires to add multiple variables to work data set (Garcia-Hernandez & Rizopoulos, 2018).

The marginal loglikelihood function of the form in equation (3.35) can be expressed as

$$l(\theta) = \log \prod_{i=1}^{n} p(T_i, \delta_i, y_i | \theta_y, \theta_T) = \prod_{i=1}^{n} \log \int p(T_i, \delta_i | b_i, \theta_y, \theta_T) [\prod_j p(y_i(t_{ij}) | b_i, \theta_y)] p(b_i, \theta_b) db_i \tag{3.37}$$

This marginal log-likelihood is approximated by the approximation of integral of the conditional probability function over the random effects and adaptive or non-adaptive Gauss-Hermite quadrature rules are applied for approximating this integration (Garcia-Hernandez & Rizopoulos, 2018).

3.3.6 Expectation Maximization (EM) algorithm-based parameter estimation technique

Since the underlying random effects in our model are not directly observable, we can not find the estimates in a straightforward way. One of the solutions is to apply the EM algorithm (J. Choi et al., 2015). For current E-step, let us consider the estimates denoted as $\hat{\Omega} = \{\hat{\mu}, \hat{\Sigma}, \hat{\alpha}, \hat{\lambda}, \hat{h_0}\}$. We can compute the expected log-likelihood value in E step since we already have the estimated value, $\hat{\Omega}$. Now, given the observed data and estimated value of parameters, the conditional probability density function for random effect is expressed as (Hwang et al., 2019)

$$p(b_i | T_i, \delta_i, y_i, \hat{\Omega}) = \frac{p(T_i, \delta_i | b_i, \hat{\alpha}, \hat{\lambda}, \hat{h_0}) p(b_i | y_i, \hat{\mu}, \hat{\Sigma})}{\int_{-\infty}^{\infty} p(T_i, \delta_i | b_i, \hat{\alpha}, \hat{\lambda}, \hat{h_0}) p(b_i | y_i, \hat{\mu}, \hat{\Sigma}) db_i} \tag{3.38}$$

Let $g(.)$ be an arbitrary function. The conditional expectation of $g(.)$ given the observed

data and estimated parameter value, is defined as follows

$$E_i[g(\pi(b_i,t)|y_i,\hat{\mu},\hat{\Omega}] = E_i[g(\pi(b_i,t))] = \frac{\int_{-\infty}^{\infty} g(\pi(b_i,t))p(T_i,\delta_i|b_i,\hat{\alpha},\hat{\lambda},\hat{h_0})p(b_i|y_i,t_i,\hat{\mu},\hat{\Sigma})db_i}{\int_{-\infty}^{\infty} p(T_i,\delta_i|b_i,\hat{\alpha},\hat{\lambda},\hat{h_0})p(b_i|y_i,\hat{\mu},\hat{\Sigma})db_i} \tag{3.39}$$

However, the probability function, $p(b_i|y_i,t_i,\hat{\mu})$ in equation (3.38) is a mixture distribution and we can not generate the random sample in a straight forward way. One solution for drawing this random sample is to use the Metropolis Hasting (MH) algorithm proposed by Metropolis and Ulam (1949) and Hastings (1970). For details on how to generate a sample, particularly for this problem, see Hwang et al. (2019). Let us consider M metropolis samples denoted by $\pi(b_i^{(k)},t), k = 1,2,...,M$ and use the last B number of samples for computing this term. In the above equation in (3.39), approximation of conditional expectation is presented as follows

$$E_i[g(\pi(b_i,t))] \approx \frac{\sum_{k=B+1}^{M} g(\pi(b_i,t))p(T_i,\delta_i|b_i^{(k)},\hat{\alpha},\hat{\lambda},\hat{h_0})/(M-B)}{\sum_{k=B+1}^{M} p(T_i,\delta_i|b_i^{(k)},\hat{\alpha},\hat{\lambda},\hat{h_0})/(M-B)} \tag{3.40}$$

Now, based on the complete data obtained from E-step, the MLE of Ω can be found. . The next step is maximization of the M-step, and we will have the following estimates

$$\hat{\mu} = \frac{1}{n}\sum_{1}^{n} E_i(b_i) \tag{3.41}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{1}^{n} E_i[(b_i-\hat{\mu})(b_i-\hat{\mu})'] \tag{3.42}$$

See Viviani, Alfó, and Rizopoulos (2014) and Hwang et al. (2019) for detailed procedure.

3.3.7 Software and level of significance

Packages available in R were used to manage the entire data set and prepare the master database required for the analysis. All primary analyses were performed in R version 3.6.3 (R Core Team, 2013). The longitudinal analysis and joint modeling were carried out using SAS version 9.4 (Instiute, 2015). The level of significance was set at 0.05 for this thesis.

Chapter 4

Application to KHP Study

4.1 Background

The relationship between depression and risk of dementia is complex and still unclear. As a result, the relationship requires further investigation. In this chapter, we applied the joint modeling technique to KHP data to determine the association between depression and the risk of dementia.

Depressive symptoms in older people have been studied extensively to establish their relationship with dementia incidence, but they have been considered as a baseline covariate or have been assessed at a single point of time (Andersen, Lolk, Kragh-Sørensen, Petersen, & Green, 2005; R. Chen et al., 2008; Ganguli, Du, Dodge, Ratcliff, & Chang, 2006; Saczynski et al., 2010). These approaches neglect the longitudinal impact through the course of depression, which might provide additional insights into the complex association between depression and dementia. Some prospective studies showed the impact of depression or depressive symptoms on cognitive decline or dementia (Gatchel et al., 2019; Gracia-García

et al., 2015; Luppa et al., 2013; Mukamal et al., 2003). Another large scale longitudinal study (Helvik et al., 2019) showed the interconnection and positive association between the severity of cognitive decline and average depressive symptoms, though the strength of association became weaker with the passage of time. However, these studies' statistical analytic methodology differed greatly and led to inconsistent results. Taken together, none of these studies took into account the correlation structure due to the longitudinal measurements of depression.

Many of the dementia disease trials collect two types of data: the time to clinical diagnosis of dementia and longitudinal measurements of some endogenous variables. Although they can be strongly associated, in most of the cases, data are analyzed separately which leads to the biased estimates and misleading inference. Joint models take into account this association by combining both the longitudinal and survival data into a single statistical model. That said, most existing joint modeling approaches have concentrated on continuous longitudinal measurement (Tsiatis & Davidian, 2004) and have paid little attention to joint modeling for binary longitudinal outcome and event time data. In clinical research, patients often have binary longitudinal outcomes that affect the main event of interest during follow up. In our study, we determined the time until clinical diagnosis of dementia among a Korean cohort of subjects aged 65 or older as well as binary longitudinal information of depression diagnosed by a clinician. Since previous studies have shown that the relationship between depression and the timing of the onset of dementia is complex, we further investigate it in this study by polishing conventional statistical methods and employing a new approach that addresses the factors that commonly lead to biased and misleading inferences. Here, we mainly applied the joint modeling of binary longitudinal

outcome (depression) and the time-to-event (dementia) approach for producing less biased and more efficient inferences.

4.2 Data source and sampling design

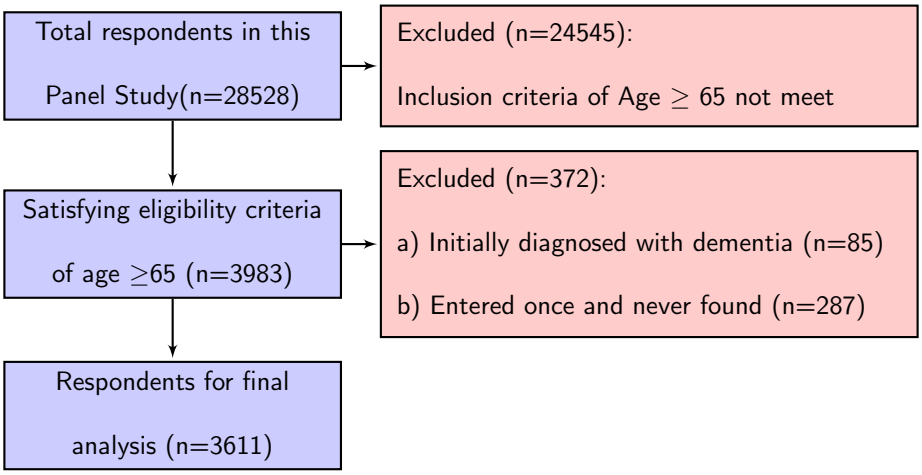
The current study utilizes the 2008 - 2015 Korea Health Panel (KHP) data as the secondary source of information. This study originally intended to focus on the use of public health care services and resulting expenditures, as well as to determine the factors affecting the use of health care services, spending on health care, and financial resources. The survey prepared a sampling frame from the list of 2005 Population and Housing Census and used 90% of the total data. The extraction method employed for determining the subject for this survey is stochastic proportional 2-step stratified cluster sampling where, region (16 metropolitan cities), East & local small town (Eup, Meon) are considered as the stratification variables. The study was approved by the University of Saskatchewan's Behavioural Research Ethics Board (See Appendix).

4.3 Study population

In 2008, information on a total of 21,283 respondents was recorded and followed up to wave 10. This sampling frame was constructed as 90% of the total data contained in the 2005 Population and Housing Census. However, due to continuous dropout from the 2010 Population and Housing Census, new household members were attracted in 2012 to secure statistical reliability using the same extraction method as in 2008. At this stage,

the sampling frame was 90% of the total data in 2010 Population and Housing Census and targeted about 2,500 new households to build a total of 8,000 panel households. Their survey data is integrated with the original annual panel data from 2014, in wave 9. Altogether the total number of respondents entered into this panel study is 28,528. Out of the total follow up subject, the number of the respondents of age 65 or older is 3,983, of them 287 respondents once enter the study were not found later and 85 of them were initially diagnosed with dementia. Finally, the data set for this analysis includes 3611 participants as our study population after excluding the number who had initially prevalent dementia and once entered the study but never found later.

Figure 4.1: Study flowchart



4.4 Variables

4.4.1 Outcome variable

Dementia is the main event of interest for this analysis and the study population were followed up to year 2015. To determine or approximate the clinically diagnosis date of dementia, we went through the inpatient, outpatient, emergency and chronic disease records for each of the patients, searched for the dementia disease codes (1501, 15011, F03) and

finally, the most backward date among them was considered as the diagnosis time of dementia. We recorded the earlier date for multiple report of dementia. Respondents once having any of these disease codes in any time during this follow up-period was identified as demented otherwise, it was non-demented or censored. Time-to-dementia was the duration from entering time to clinically diagnosis of dementia and the censoring time was the period from entrance time to the survey date when they appeared last time or date of termination of the study in 2015.

4.4.2 Potential covariates

The following socio-demographic and economic characteristics were considered as baseline covariates: sex, age, level of education, economic activity, living arrangement, living in a multi-generational household, marital status, baseline comorbidities (such as cardiovascular disease, diabetes, hypertension, and gum disease), the presence of more than three chronic diseases, and self-reported behaviors (such as alcohol intake, smoking, intense/moderate physical activity, and walking activity). Economic activity and the presence of more than three chronic disease both had two levels: yes and no. Years of schooling were categorized into three different groups: 0 = no education, 1-5 = primary education, and 5+ = above primary education. Living in a multi-generational household was recoded as first generation (i.e. single or couple), second generation (i.e. single or couple, with child), and third generation (single or couple, with child and grandchild and/or other relations). Current marital status was categorized as married (including putative marriage) versus other (separated, divorced, widowed, unmarried). Frequency of intense and medium physical activity in a week were combined to create a variable

indicating intense/medium physical activity, with three categories: never, $\leq$ 3 days and >3 days per week. Originally, the categories of alcohol consumption were never, recently non-drinking, less than once per month, once per month, 2-3 times per month, once per week, 2-3 times per week, almost daily, which were recoded as "never," "recently not drinking," "drink monthly/weekly," and "drink daily." The self-reported variable for smoking was re-categorized as "never," "current daily smoker/occasional smoker," and "ex-smoker." This data also had information regarding the subject's walking frequency during a one week period which was re-coded as "never," "three days or less," or "more than three days per week". The presence or absence of each co-morbidity (anxiety, depression, cardiovascular disease, diabetes, hypertension, and gum disease) was determined by comparing the disease codes to those contained in the data and recoding the finding as yes or no. A variable was also available that indicates the presence or absence of more than three chronic diseases.

4.5 Statistical Analysis

4.5.1 Descriptive analysis

Demographic characteristics

This study included 3,611 individuals aged 65 years or older who were eligible for the study. Of them 2,055 (56.9%) were female, the mean age was 72.2 years (s.d=$\pm$ 5.7) and 1,927 (53.4%) were older than 70. In all, 2,390 (66.2%) were married, and 2,290 (63.4%) and 781 (21.6%) lived in first-generation or second-generation household, respectively. At baseline, 17.6% reported they never walked, whereas 12.7% walked less than or equal to

three days in a week. Also, 59.1% were non-smokers, 40.1% did not drink alcohol and 61.7% did not have any economic activity (see Table 4.1).

Clinical characteristics

The majority (87.8%) had more than three chronic disease. Hypertension, cardiovascular disease, diabetes, and gum disease were present in 49.9%, 9.9%, 17.9%, and 12.5% of individuals, respectively. This analysis also found that most of the respondents (87.8%) had more than three chronic diseases. A small number of individuals had anxiety (1.9%) or depression (2.7%) at the baseline. Table 4.1 shows the details of these baseline characteristics.

Prevalence measurement for depression

This study identified 58 (2.2%) people who were initially diagnosed with depression in the beginning cohort in 2008. This percentage increased with the passage of time. Due to ongoing dropout, additional samples were added to the original population-based cohort in 2014. At that time point, the number of people with a clinical diagnosis of depression was 149 (5.35%). This figure increased to 5.48% in 2015, the end point of the study.

Table 4.1: Baseline characteristics (N=3611)

Variables	mean (±sd); n (%)	Variables	mean (±sd); n (%)
Age (mean, sd)	72.2 (±5.7)		
Gender		Hypertension	
Male	1556 (43.1)	No	1810 (50.1)
Female	2055 (56.9)	Yes	1801 (49.9)
Age category		Cardiovascular disease	
≤70 years	1687 (46.6)	No	3254 (90.1)
>70 years	1927 (53.4)	Yes	357 (9.9)
Marital status		Diabetes	
Married	2390 (66.2)	No	2964 (82.1)
Other (Separated, widowed, divorced)	1221(31.8)	Yes	647(17.9)
Living arrangement		Gum disease	
Together (not alone)	3582(99.2)	No	3161 (87.5)
Alone	29(0.8)	Yes	450 (12.5)
Economic activity		>3 chronic diseases	
No	2228 (61.7)	No	439 (12.2)
Yes	1383 (38.3)	Yes	3172 (87.8)
Generations within household		Depression	
One	2290 (63.4)	No	3514 (97.3)
Two	781 (21.6)	Yes	97 (2.7)
Three	540 (15.0)		
Education		Anxiety	
No Formal Education	700 (19.4)	No	3542 (98.1)
Primary	1570 (43.5)	Yes	69 (1.9)
Above primary	1341 (37.1)		
Walking in a week		Smoking	
Never	632 (17.6)	Never	2092 (59.1)
≤3 days	450 (12.7)	Occasionally/Current daily	488 (13.8)
>3 days	2459 (69.4)	Ex-smoker	961 (27.1)
Moderate/intense physical activity in a week		Alcohol intake	
		Never	1421 (40.1)
Never	2339(66.1)	Recently not drinking	534 (15.1)
≤3 days	1011(28.5)	Monthly/weekly	1305 (36.9)
>3 days	191(5.4)	Daily	281 (7.9)
Hospitalization		Overall follow-up time (median, IQR)	5.8 (5.7)
No	2988(82.8)		
Yes	623(17.2)		

4.5.2 Survival analysis for time-to-dementia

Variables associated with dementia risk

Kaplan-Meier (K-M) estimates

During the 8-year follow-up period, 215 (about 6%) respondents developed dementia. The overall median follow-up time was 5.8 years (3.6 years for people living with dementia versus 5.9 years for people without dementia). Figure 4.2 shows Kaplan-Meier (K-M) estimates of the survival time for diagnosis of dementia by sex, age category, educational attainment, having diabetes, depression and economic activity.

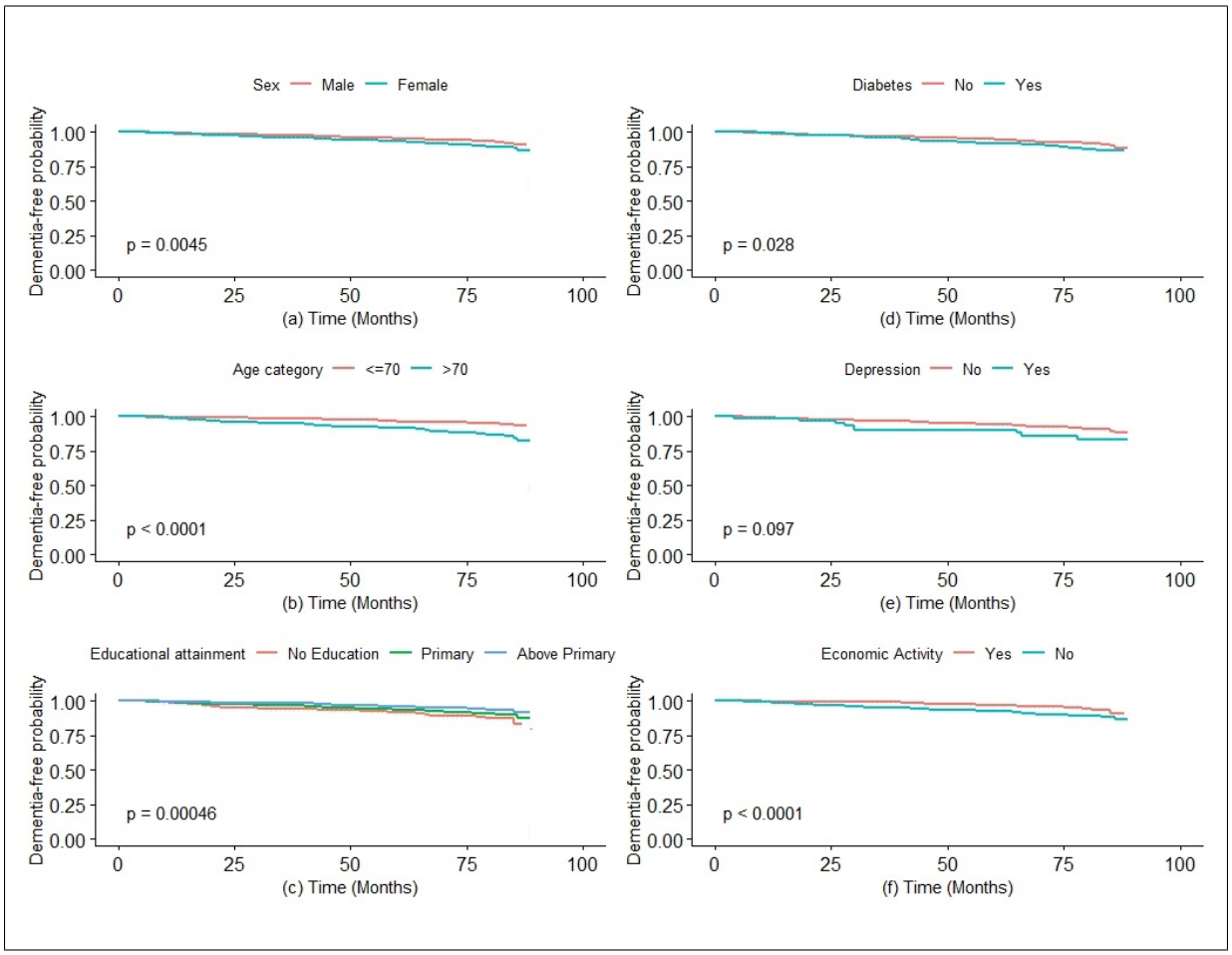


Figure 4.2: Kaplan-Meier survival estimates for sex, age category, educational attainment, diabetes, depression, and economic activity and the p-value from log-rank test

The K-M survival estimates (Figure: 4.2) showed that the survival curves for male sex, people aged 70 years or less, individual without diabetes, those who had comparatively

higher education and economic activity were significantly higher than their corresponding comparison groups. However, the survival curve for depressed people was not significantly different from the non-depressed people.

Univariate Cox PH model

Univariate analyses between potential factors and the risk of dementia were assessed in this study (Table 2). The age was highly associated with developing dementia and the risk was higher among the individuals older than 70. The dementia risk was lower among males than females (hazard ratio [HR]=0.66, 95% CI: 0.50 - 0.88, p-value=0.005). The hazard ratio was higher for those with no formal education (HR= 2.07, 95% CI: 1.43 - 3.00, p-value=0.011) and primary-educated (HR=1.53, 95% CI: 1.10 - 2.13, p-value=<0.001) participants compared to individuals who had relatively higher education. Involvement in economic activity was significantly associated with lower risk of dementia (HR=0.52, 95% CI: 0.38 - 0.71, p-value=<0.001). For individuals having more than three chronic diseases compared with the ones with fewer, the hazard ratio was 2.22 (95% CI: 1.29 - 3.82, p-value=0.004). The dementia risk among the participants who did not marry ever was 61% higher than that of ever-married participants. Individuals reporting no physical activity (HR= 2.60, 95% CI: 1.15 - 5.86, p-value=0.022) were at higher risk of dementia than the group undertaking physical activity at least three days per week. Individuals without diabetes (HR=0.70, 95% CI: 0.50 - 0.96, p-value=0.029) and hypertension (HR= 0.70, 95% CI: 0.53 - 0.91, p-value=0.009) were both at 30% lower risk of dementia than people having these conditions. Baseline walking frequency, number of generations living in the household, and hospitalization status were also significantly associated with dementia.

Table 4.2: Univariate Cox model for time-to-dementia

Variables	**Hazard Ratio (95% CI)**	**p-value**
Gender		
Female*		
Male	0.66(0.50, 0.88)	0.005
Age Category		
>70		
≤70	0.34 (0.25, 0.46)	<0.0001
Gum disease		
Yes*		
No	1.06 (0.70, 1.61)	0.778
Cardiovascular disease		
Yes*		
No	1.23 (0.75, 2.02)	0.413
Economic activity		
Yes*		
No	0.52 (0.38, 0.71)	<.0001
More than three chronic diseases		
No*		
Yes	2.22 (1.29, 3.82)	0.004
Education		
Above primary*		
No formal education	2.07(1.43, 3.00)	<0.001
Primary	1.53 (1.10, 2.13	0.011
Marital status		
Married*		
Others (Separated, widowed, divorced)	1.61 (1.23, 2.11)	<0.001
Living arrangement		
Together (not single)		
Alone	2.62 (0.84, 8.21)	0.097
Diabetes		
Yes*		
No	0.70 (0.50, 0.96)	0.029
Depression		
Yes*		
No	0.58 (0.29, 1.13)	0.107
Anxiety		
Yes*		
No	1.50 (0.48, 4.70)	0.483
Intense/medium physical activity in a week		
>3 days*		
Never	2.60 (1.15, 5.86)	0.022
≤3 days	1.33 (0.56, 3.13)	0.520
Alcohol intake		
Daily*		
Never	1.66 (0.93, 2.96)	0.087
Not recent	1.52 (0.81, 2.87)	0.197
Monthly/weekly	0.90 (0.49, 1.64)	0.721
Walking		
>3 days*		
Never	1.76 (1.25, 2.48)	0.001
≤3 days	0.96 (0.61, 1.51)	0.867

Variables	Hazard Ratio (95% CI)	p-value
Smoking		
Ex-smoker*		
Never	0.98 (0.71, 1.34)	0.888
Current daily/occasionally	0.70 (0.43, 1.14)	0.152
Hospitalization		
Yes*		
No	0.59 (0.43, 0.81)	<0.001
Hypertension		
Yes*		
No	0.70 (0.53, 0.91)	0.009
Generations within household		
Three*		
One	0.45 (0.33, 0.62)	<0.0001
Two	0.51 (0.35, 0.76)	0.001

However, this univariate analysis did not find any association of gum disease, cardiovascular disease, depression, anxiety, smoking, or alcohol intake, and living arrangement with the development of dementia.

Multivariable Cox PH model

There were fourteen variables significant at 10 % level in univariate Cox model: gender, age category, education, marital status, living arrangement, economic activity, alcohol intake, more than three diseases, hypertension, diabetes, hospitalization, physical activity, walking activity/week, and generation setting in household. Variables significant at 10 % level in univaraite Cox model were entered into the multivariable Cox model to build the final model.

Table 4.3: Multivariable Cox model for time-to-dementia

Covariates	Estimate	SE	Hazard Ratio	95% CI of HR	P-value
Age category					
≤ 70*					
>70	0.90	0.16	2.46	1.80, 3.36	<0.001
Education					
Above Primary*					
No education	0.45	0.20	1.57	1.07,2.30	0.022
Primary Level	0.38	0.17	1.46	1.05, 2.04	0.025
Economic Activity					
Yes*					
No	0.38	0.16	1.46	1.06, 2.01	0.021
Diabetic condition					
No*					
Yes	0.35	0.17	1.41	1.02, 1.97	0.04
Walking Activity/week					
>3 days*					
Never	0.35	0.18	1.42	1.01,2.01	0.046
≤ 3 days	0.18	0.20	1.20	0.81, 1.76	0.369
Generations in household					
3^{rd} generation*					
1^{st} generation	-0.56	0.17	0.57	0.41, 0.79	<0.001
2^{nd} generation	-0.45	0.20	0.64	0.43, 0.95	0.028

Along with the likelihood ratio test (LRT), Akaike information criterion (AIC) as an additional criteria were used to determine the possible and final subset of covariates in the final multivariable Cox model (Table 4.3). After multivariable adjustment, the final model retained six variables that were independently associated with the risk of dementia: age, education, economic activity, diabetes, walking activity/week, and generation setting in household.

Individuals older than 70 years were at 2.46 times higher risk for dementia than younger ones of age 70 or less (adjusted hazard ration, AHR= 2.46, 95% CI: 1.80 - 3.36, p-value=<0.001) after adjusting other covariates. Compared with people having more than primary education, individuals with no formal education or only primary education had 57% and 46% more risk, respectively (AHR=1.57, 95% CI: 1.07 - 2.30, p-value=0.022; AHR=1.46, 95% CI: 1.05 - 2.04, p-value=0.025). Never walking significantly predicted

42% higher risk of dementia than walking more than three days per week (AHR= 1.42, 95% CI: 1.01 - 2.01, p-value=0.046). People with diabetes (AHR= 1.41, 95% CI: 1.02 - 1.97, p-value=0.040) had an elevated risk of developing dementia compared to the non-diabetic people. Time-to-dementia within the non-economic activity group was 1.46 times more likely than within the economic activity group (AHR= 1.46, 95% CI: 1.06 - 2.01, p-value=0.021). We observed a lower risk of dementia among the people living in single generation (AHR= 0.57, 95% CI: 0.41 - 0.79, p-value=<0.001) or second generation (AHR= 0.64, 95% CI: 0.43 - 0.95, p-value=0.028) settings compared to those living with three generations [see Table 4.3 & Figure: 4.3].

Although sex, hypertension, marital status, and physical activity were significant at univariate analysis, they were no longer significant in the multivariable model. Variable indicating more than three chronic disease was significant for both univariate and multivariable analysis. However, since there was no specified information about which diseases were included, we decided to exclude this variable from the final analysis. Possible interactions among the variables were also tested. However, we found no significant interaction between covariates retained in the model and no violation of proportionality assumptions. The proportional hazards assumption was assessed with log-log plots and Schoenfeld residuals (Fitrianto & Jiin, 2013); the goodness of fit was assessed by the log-likelihood test.

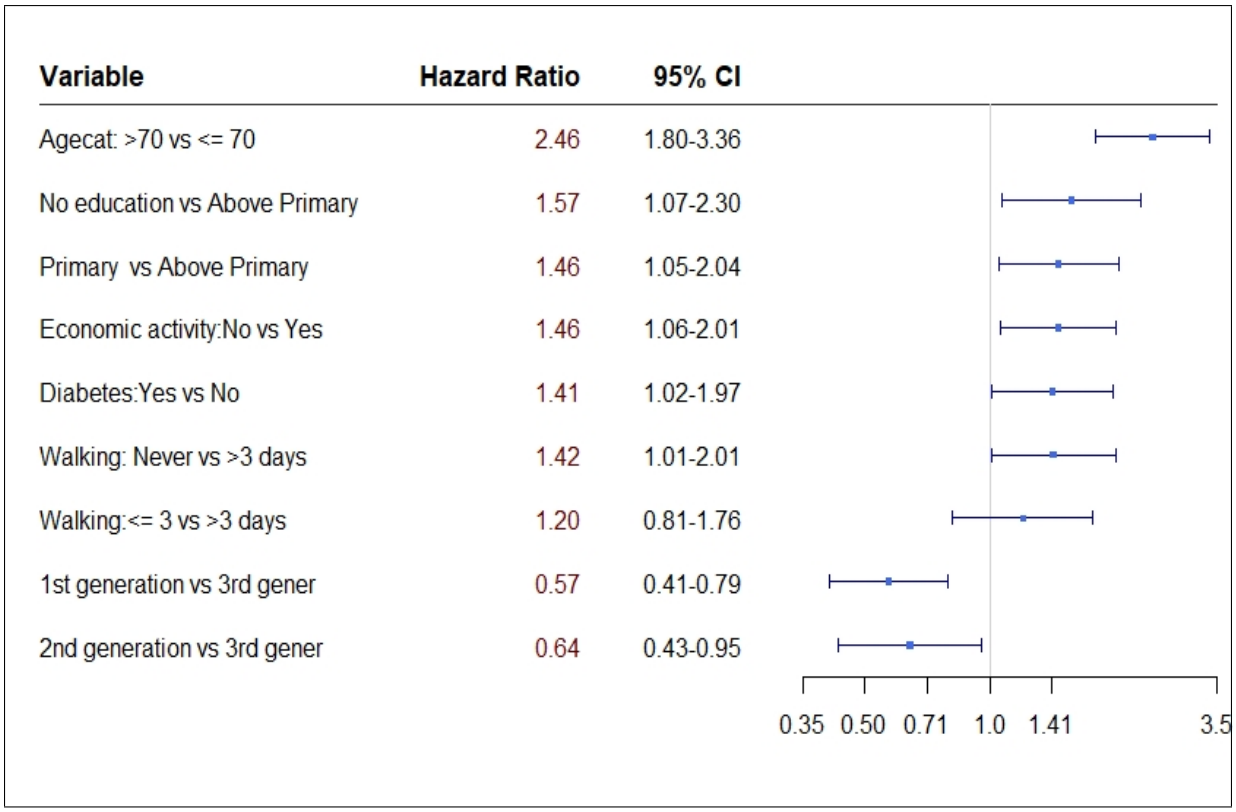


Figure 4.3: Hazard ratio plot of Cox model for time-to-dementia

4.5.3 GLMM analysis for longitudinal outcome

Depression was the longitudinal outcome in this study. Subject specific information of depression was used in joint modeling of longitudinal depression and time-to-dementia. Since the repeated measurement of depression was categorical that was dichotomous (yes or no) in this data set and the primary interest was in subject-specific effects rather than the population level, generalized linear mixed model (GLMM) was applied to fit the conditional model using G-side random effects and subject specific estimates were obtained. The covariance structure for the random effects includes unstructured, Toeplitz, compound symmetry, firs order auto-regressive (AR1). Since the GLMM for longitudinal outcome in this study included only two random effects (random intercept and random slope), choosing the covariance structure for a single covariance was not a great concern. However, during the analysis, unstructured covariance structure was specified for the random component.

Univariable analysis

Univariate generalized linear mixed effects models in which the linear predictor included random effects in addition to the fixed effects were conducted. At univariable analysis, for the binary repeated measurement of depression, single covariate was considered as a part of fixed effect and random intercept and slope was taken into account for random effects in each of the models. This univariable GLMM analysis identified four covariates including gender, CVD, more than three chronic disease, and anxiety as significant variables at 10% level. This longitudinal analysis found that depression among males, individuals who had not been diagnosed with a CVD and did not have anxiety was lower compared to their corresponding comparison group. People with more than three chronic disease had higher

odds of depression compared to the people with lower number of chronic disease. Again, since we did not have specific information about which chronic diseases were included, it was excluded and individual diseases such as hypertension, anxiety, CVD, and diabetes were used for the analysis. The rest of the covariates were insignificant in this model. The results were shown in the the Table 4.4.

Table 4.4: Univariable analysis from GLMM for repeated measurements of depression.

Covariates	Estimate (95% CI)	P-value
Gender		
Female		
Male	-0.80 (-1.62, 0.01)	0.053
Economic Activity		
Yes		
No	- 0.58 (-1.38, 0.22)	0.156
Gum disease		
Yes		
No	-0.13 (-1.21, 0.94)	0.811
CVD		
Yes		
No	-0.96 (-1.93, 0.01)	0.054
Diabetes		
Yes		
No	-0.29 (-1.20, 0.61)	0.532
Anxiety		
Yes		
No	-2.3 (-3.81, -0.79)	0.003
Age category		
>70		
≤70	-0.17 (-.90, 0.56)	0.65
Education		
Above Primary		
Illiterate	0.21(-0.80, 1.21)	0.689
Primary	0.04(-0.80, 0.87)	0.933
Living		
Together		
Not together	-0.002(-4.73, 4.73)	0.999
Marriage		
Married		
Other (Separate, widow, divorce)	0.10 (-.68, 0.87)	0.803
>3 disease persistent		
No		
Yes	1.75 (-0.33, 3.83)	0.099
Medium and intense physical activity		
Both>3		
Never	0.59 (-1.25, 2.44)	0.530
Any catrgory≤3	0.26 (-1.68, 2.19)	0.797
Drinks		
Daily		
Monthly/weekly	0.73 (-0.86, 2.32)	0.369
Never	0.16 (-1.67, 1.99)	0.868
Recently not drink	0.16 (-1.47, 1.79)	0.845

Covariates	Estimate (95% CI)	P-value
Smoking		
Smoked but not now		
Never	0.69 (-0.29, 1.67)	0.167
Current daily/ Occasionally	0.22 (-1.15, 1.59)	0.756
Walking/week		
>3		
Never	0.40 (-0.60, 1.40)	0.437
≤3	-0.38 (-1.21, 0.45)	0.368
Hypertension		
Yes		
No	-0.64 (-1.50, 0.22)	0.142
Generation		
Third		
First/couple	0.06 (-0.99, 1.12)	0.907
Second	-0.21 (-1.50, 1.09)	0.756
Time (months)	-0.001 (-0.02, 0.01)	0.877

Multivariable analysis

The covariates significant at 10% in the univariable analysis were included in the multivariable generalized linear mixed model. The four variables (that is, gender, CVD, more than three chronic diseases, and anxiety) significant in the univariable GLMM analysis were entered into the model for determining the final model. The final model retained only three variables: gender, CVD and anxiety. The findings from this multivariable GLMM model showed that female had higher depression than male. The odds of depression among those diagnosed with CVD and anxiety were also higher than people without CVD or anxiety (Table 4.5). The estimate, 0.66 corresponding to CVD from this generalized linear mixed model indicates that an individual subject's odds of having depression was $exp(0.66) = 1.93$ times higher for the subject diagnosed with CVD at baseline compared with a subject that was not diagnosed. Interaction among the covariates as well as time by other covariate were checked, however no significant interaction was found in this model.

Table 4.5: Multivariable generalized linear mixed effects model for depression

Covariates	Estimate (95% CI)	P-value
Gender		
Male		
Female	0.82 (.001, 1.62)	0.048
CVD		
No		
Yes	0.66 (0.01, 1.31)	0.046
Anxiety		
No		
Yes	2.14 (0.61, 3.67)	0.006

4.5.4 Joint analysis of longitudinal and survival data

The joint modeling approach mentioned in Chapter 3 was applied to KHP study. A likelihood ratio test was used to select the final model with a potential subset of covariates for both the longitudinal and survival submodels. The finally selected model, that is, model with current-value shared parameter also had the lowest AIC compared to the second model, that is, model with current-value plus slope-dependent parameter. Basically, we considered two joint models with the same set of covariates for both longitudinal and survival processes. The only difference between them was the addition of slope-dependent shared parameters alongside the current-value in the second model. More clearly, the first joint model considered the hazard function using only the current-value shared parameter, whereas the second model incorporated both the current-value and slope-dependent shared parameters.Finally, the best fitted model was selected using AIC (a lower AIC indicates a better fit). Since the AIC of both models were very close to each other and the slope-dependent shared parameter was not significant, the first model (the model only with the current shared parameter and that had the lowest AIC) was finally chosen as the parsimonious model.

The models are given below:

Joint model with current-value shared parameter

Longitudinal submodel

The following longitudinal submodel was considered for the i^{th} subject

$$\log(\frac{\pi_i(t_{ij})}{1-\pi_i(t_{ij})}) = \beta_0 + \beta_1 t_{ij} + \beta^T X_i + b_{i0} + b_{i1} t_{ij} \tag{4.1}$$

Time-to-event submodel

The following time-to-event model with the current-value of the longitudinal variable was used for this analysis

$$h_i(t) = h_0(t) \exp(\lambda^T W_i + \alpha_1 \pi_i(t)) \tag{4.2}$$

Joint model with current-value plus slope-dependent shared parameter

$$\log(\frac{\pi_i(t_{ij)}}{1-\pi_i(t)}) = \beta_0 + \beta_1 t_{ij} + \beta^T X_i + b_{i0} + b_{i1} t_{ij}$$

$$h_i(t) = h_0(t) \exp(\lambda^T W_i + \alpha_1 \pi_i(t) + \alpha_2 \pi_i'(t)) \tag{4.3}$$

where X_i in the longitudinal submodel is a vector of covariates and β is the corresponding vector of regression coefficients. W_i is vectors of baseline covariates in the survival model, and λ is the corresponding vector of coefficient. All or some of the covariates may or

may not be common in both X_i and W_i. Here, $h_i(t)$ is basically the hazard of dementia diagnosis at time t for the i^{th} patient and $h_0(t)$ is baseline hazard specified with piece-wise constant hazard model. The main interest is parameter α_1 in the first model defined in equation (4.2) and in second model defined in equation (4.3), that is, $\alpha = (\alpha_1, \alpha_2)$ that links the longitudinal process (the trajectory function of depression) to the survival process.

In the above model described in equation (4.1), the design matrix X_i in the longitudinal model included the baseline information of walking activity in a week, generations in household, gender. We also added some potential covariates that were expected to have a significant impact on the longitudinal process of depression. However, insignificant variables were finally dropped from the model to make it more parsimonious. We also incorporated interaction terms (covariate by time) to the longitudinal model for investigating the probable time-varying coefficients of the baseline covariates. However, no interaction terms were statistically significant so we decided to exclude those term to get our final longitudinal model. In addition, we used linear effects of time (β_1) for modeling average trajectories but more complex functions of time (for instance quadratic term or higher order polynomials or splines) are also possible.

In the survival submodel, the baseline covariate vector, W_i included age of the respondents ($\leq$70, $>$70), educational attainment (no formal education, primary education, above primary), walking activity per week (never, $\leq$ 3 days, $>$3 days), generations in household (1^{st}, 2^{nd} and other), diabetes (yes, no) and economic activity (yes, no). Similar to the longitudinal process, we incorporated the other potential covariates and possible interactions

in survival submodel. We dropped the insignificant variable and no significant interaction was found. We also incorporated the longitudinal process through shared random effects and linked it by α_1 in equation (4.2). The quantity $exp(\alpha_1)$ expressed the hazard ratio at time t for a one-unit increase in the trajectory of the longitudinal response at the same time point. This measure expresses, how much the hazard is higher among the exposed group compared to the reference group.

The above model described in equation (4.2) is a standard joint model that assumes the relationship between the current-value of longitudinal depression and the risk of dementia at the same time point. However, this naive assumption between these two processes could be more complicated in practice. The simple assumed trajectories for longitudinal response sometimes may not adequately demonstrate the longitudinal profile of the individual and sometimes only the current-value shared parameter could not be able to capture the entire picture of actual relationship. For this purpose we also compared joint models with different association structures between longitudinal and survival processes. For example, in addition to current-value $\pi_i(t)$ of depression, we also included the slope (rate change), $\pi_i'(t)$ to verify whether it contained additional information for the risk of developing dementia (see equation (4.3)). It indicates how much the longitudinal response is increasing or decreasing at a particular time point. More specifically, it can be stated that two patients with same condition of depression at the current time point do not necessarily show the same level of risk for developing dementia, for example, if the trajectory of longitudinal depression for one patient stepped up swiftly, while the trajectory of the other patient remained constant over time and always little bit higher up to a particular time point comparing to the trajectory of the first patient (the higher the trajectory, the

more they have depression). Since the trajectory of the first patient is stepping up, at some point (say current time point) it will cross the trajectory of the second patient. Then it is very natural to assume that the value of depression for this two patient are same at this current time point and they are assumed to have same risk of developing dementia. However, their trajectories were not changing in the same way and the rate of change of longitudinal response for the first patient increased, that is, the trajectory of this patient is associated with greater risk of dementia. Joint model with more complex structure such as cumulative effect, time-lagged effect, and random spline could also be considered. We avoided these models just to make our final model parsimonious.

Model selection

The model with a possible subset of covariates for both the longitudinal and survival submodel was finalized based on the likelihood ratio test. We also checked the AIC for different sets of joint model. The final model retained five variables for the longitudinal submodel and seven for the survival part. The recorded AIC for this final current-value shared parameter model was also lowest (AIC=7301.4) compared to second model, slope-dependent shared parameter (AIC=7307.0).

Common risk factors shared by depression and dementia

The link between depression and dementia could be apparent if they shared common risk factors. We found that the risk factors for depression have a little overlap with risk factors for dementia. However, preexisting vascular disease could be associated with an elevated risk of vascular dementia and may also increase depression incidence since disability of any kind often causes depression. In this study, the identified significant common risk

factors for both dementia and depression were economic activity, walking frequency in a week, and living in a multi-generation household.

In the joint modeling of longitudinal depression and time-to-dementia submodels, generation setting in household, walking in a week, and economic activity were significant in both submodel. However, diabetes, education, and age were significant only in survival submodel while gender was only significant in the longitudinal submodel (Table 4.6). The odds of depression was significantly higher among female, respondents having no economic activity, less walking frequency and the individuals living with higher generation in household. The odds of depression also increased with time. In survival submodel of joint model, association parameter corresponding to the current-value was significantly different from zero (pval=0.005), indicating a significant association between depression and the risk of dementia. The positive estimated value of the association parameters (0.88) indicated that the risk of dementia was 2.41 times higher among the depressed compared to non-depressed people (HR=2.41, 95% CI:1.30, 4.50) (see Table 4.6).

While walking frequency, time, generation setting in household and economic activity were not significant in separate analysis of longitudinal depression (Table 4.5), the longitudinal submodel of joint model identified those variables as significant predictors for depression. On the other hand, similar associations from both standard Cox model and survival submodel of joint model were observed for the potential covariates. However, the corresponding magnitudes to those covariates in separate analysis of cox model were different from the survival submodel in joint model. Comparing the results from Cox model and survival submodel, this study found that, though the depression measurement

Table 4.6: Results from joint model with current-value shared parameter

Variables	Estimate	SE	95% CI	P-value
Longitudinal depression				
Intercept	-19.77	1.63	-22.97, -16.58	<0.001
Time	0.04	0.01	0.01, 0.06	0.002
Gender				
Male*				
Female	2.64	0.57	1.52, 3.75	<0.001
Economic Activity				
Yes*				
No	1.42	0.45	0.53, 2.31	<0.001
Walking Activity/week				
>3 days*				
Never	1.29	0.46	0.38, 2.19	0.005
≤ 3 days	1.06	0.50	0.09, 2.03	0.032
Generations in household				
3^{rd} generation*				
1^{st} generation	1.37	0.57	0.25, 2.49	0.017
2^{nd} generation	0.35	0.67	-0.96, 1.67	0.601
$\log sd(b_{i0}), \sigma_0$	1.83	0.08	1.67, 1.99	<0.001
$\log sd(b_{i1}), \sigma_1$	-3.03	0.11	-3.25, -2.80	<0.001
$Cov(b_{i0}, b_{i1}), \sigma_{01}$	-0.10	0.05	-0.20, 0.001	0.054

Variables	Estimate	SE	Hazard Ratio	95% CI of HR	P-value
Time-to-dementia					
Age category					
≤ 70*					
>70	0.91	0.16	2.49	1.82, 3.39	<0.001
Education					
Above Primary*					
No education	0.41	0.20	1.50	1.02, 2.21	0.039
Primary Level	0.37	0.17	1.45	1.04, 2.02	0.028
Economic Activity					
Yes*					
No	0.36	0.16	1.43	1.06, 2.01	0.030
Diabetic condition					
No*					
Yes	0.33	0.17	1.38	1.01, 1.93	0.054
Walking Activity/week					
>3 days*					
Never	0.36	0.17	1.43	1.02, 2.02	0.038
≤ 3 days	0.15	0.20	1.16	0.79, 1.72	0.439
Generations in household					
3^{rd} generation*					
1^{st} generation	-0.59	0.17	0.56	0.40, 0.77	<0.001
2^{nd} generation	-0.42	0.20	0.66	0.44, 0.97	0.037
Associations(Current-value)	0.88	0.32	2.41	1.30, 4.50	0.005

at baseline was not predictive of dementia in standard Cox model, the time variant measurement of depression was significantly associated with dementia risk in joint model.

The results presented in Table 4.7 show additional information regarding the slope-dependent shared parameter. Two individuals might have same current-value of depression trajectories, but their trajectories' rate of change might be associated with different risks of dementia. Since this study did not find any significant association of dementia risk with the slope, we dropped the slope from the model and considered the model with a current-value shared parameter. Therefore, all of this discussion is based on the analysis of a joint model with current-value (Table 4.6).

Table 4.7: Results from joint model with current-value and slope-dependent shared parameter

Variables	Estimate	SE	95% CI	P-value
Longitudinal depression				
Intercept	-20.02	1.68	-23.32, -16.72	<0.001
Time	0.04	0.01	0.01, 0.06	0.002
Gender				
Male*				
Female	2.78	0.64	1.52, 4.04	<0.001
Economic Activity				
Yes*				
No	1.42	0.45	0.53, 2.32	<0.002
Walking Activity/week				
>3 days*				
Never	1.30	0.47	0.38, 2.21	0.005
$\leq$ 3 days	1.08	0.49	0.11, 2.04	0.028
Generations in household				
3^{rd} generation*				
1^{st} generation	1.37	0.57	0.19, 2.42	0.022
2^{nd} generation	0.30	0.66	-1.0, 1.61	0.647
$\log sd(b_{i0}), \sigma_0$	1.83	0.08	1.68, 1.20	<0.001
$\log sd(b_{i1}), \sigma_1$	-2.99	0.13	-3.25, -2.74	<0.001
$Cov(b_{i0}, b_{i1}), \sigma_{01}$	-0.10	0.05	-0.20, -0.002	0.045
Variables	Estimate	SE	95% CI	P-value
Time-to-dementia				
Age category				
$\leq$ 70*				
>70	0.91	0.16	0.60, 1.22	<0.001
Education				
Above Primary*				
No education	0.41	0.20	0.02,0.79	0.038
Primary Level	0.38	0.17	0.04, 0.71	0.027
Economic Activity				
Yes*				
No	0.34	0.16	0.02, 0.67	0.038
Diabetic condition				
No*				
Yes	0.32	0.17	-0.01, 0.65	0.056
Walking Activity/week				
>3 days*				
Never	0.36	0.17	0.02, 0.71	0.037
$\leq$ 3 days	0.14	0.20	-0.24, 0.53	0.467
Generations in household				
3^{rd} generation*				
1^{st} generation	-0.59	0.17	-0.92, -0.26	<0.001
2^{nd} generation	-0.43	0.20	-0.83, -0.04	0.033
Associations(Current-value)	0.75	0.35	0.06, 1.43	0.034
Associations(Slope-dependent)	39.05	29.21	-18.22, 96.32	0.181

Chapter 5

Discussion

In this thesis, we applied the shared random effects joint model dealing with categorical longitudinal outcomes Garcia-Hernandez and Rizopoulos (2018) to KHP study. This study mainly focused on the joint modeling approach for determining the association between endogenous time-dependent variable and time-to-dementia.

Generalized linear mixed effects model was applied to characterize binary repeated measurements of depression. Random intercept and slope terms were included in this model. A Cox model, on the other hand, was used to characterize the time-to-dementia process. Since un-specification of the baseline hazard in joint modeling approach always produces misleading inference, the baseline hazard was specified using a piece-wise constant hazard model, which is one of the most commonly used models (Garcia-Hernandez & Rizopoulos, 2018).

The dependence between time-varying depression and the risk of dementia was assumed through the current-value parameterization, which is a widely used and standard approach

to link the longitudinal and the time-to-event models in the context of the random-effects shared-parameter models framework.

Since the literature indicates that the relation between depression and the risk of dementia is unclear, we further investigated this association using the joint model. Here, the joint model combined the longitudinal model for depression and the survival model for time-to-dementia into a single statistical model to obtain more robust estimates and draw valid inferences. To the best of our knowledge, there has been no attempt to examine the impact of binary repeated measurements of depression on dementia risk using a joint model. This study found that time-varying depression, old age, lower educational attainment, diabetes, and less frequent walking were the significant factors associated with an increased risk of dementia.

The analysis of this 8-year longitudinal study showed that, although the depression measurement at the baseline was not predictive of dementia, the time variant measurement of depression was significantly associated with dementia risk. One of the important merits of this project was to properly use the longitudinal information of depression up to the time of dementia diagnosis. Ignoring this depression trajectory may conceal the association between depression and dementia . The longitudinal process in this analysis took into account the issues of random effects and the missingness of repeated depression for estimating the true value that is associated with risk of dementia.

Our finding was consistent with Holmquist, Nordström, and Nordström's (2020) recent study where depression was significantly associated with increased odds of dementia,

which was still evident more than 20 years after the diagnosis of depression. As well, the researchers discovered that the dementia risk was higher among severely depressed people compared to those with mild depression.

The Framingham Heart Study also reported a similar finding even after participants taking antidepressant medications were included (Saczynski et al., 2010). A study conducted by Almeida, Hankey, Yeap, Golledge, and Flicker (2017) showed the graded association between the severity of depressive symptoms and the risk of dementia, though this association disappeared after the five-year follow up period.

A 28-year follow-up study (Singh-Manoux et al., 2017) demonstrated that depressive symptoms in late life were associated with a higher risk for dementia and that this relation was more apparent when depressive symptoms were measured just before the decade of dementia diagnosis. Midlife depressive symptoms were not significantly associated with dementia in this study. Combining these association structures, the authors reported that depressive symptoms may be a prodromal feature of dementia.

A comprehensive review (Fountoulakis et al., 2003) suggests that geriatric depression varies from depression in younger patients in many aspects and that late-life depression is related to structural brain abnormalities and cerebrovascular changes. This type of depression is called vascular depression (Alexopoulos et al., 1997). Butters et al. (2008) proposed that depression can alter the risk of cognitive dysfunction and can also shorten the latent period between AD neuropathology development and the onset of dementia, thus accelerating incidence among older population with depression.

The finding from a study by Butters et al. (2008) could be supported by previous study of his (Butters et al., 2000), where elderly depressed people taking antidepressant may experience better conditions in specific domains but can not achieve normal performance levels. This late-life depressed group of patients were at high risk of developing dementia. Results from other studies (Johansson et al., 2019; G. Li et al., 2011; Rasmussen et al., 2018) are also consistent with our finding.

In contrast to our results, some studies (Becker et al., 2009; Ganguli et al., 2006) did not find an association between depression and the incidence of dementia. Olazaran, Trincado, and Bermejo-Pareja (2013) reported the cumulative effects of depression on both prevalence and incidence of dementia and showed that especially the combination of present and past depression was associated with dementia prevalence, but not with incidence.

One of the reasons for these conflicting findings might be the approach used for determining depressive symptoms. It is more likely for recall bias to occur in self-reported midlife depression and when discussing a past history of depression among elderly people aged 65 years or more. Thus, the retrospective or self-reported assessment of depression may substantially misclassify this risk factors and result in unreliable association. However, registry based studies and clinical diagnoses can easily avoid this recall bias by following patients routinely and recording their data longitudinally.

Another possible reason for conflicting results relates to the statistical method used to determine the association between depression and dementia. It is important to point out

that all the studies mentioned conducted separate analyses either for the survival or longitudinal portions of the studies. Most of the studies took the depression measurement at a single point, whereas a limited number of studies (G. Li et al., 2011; Singh-Manoux et al., 2017) considered the temporal issue or depression measurement at few points but separate analysis for each of the points.

Therefore, these analyses completely ignore an individual's depression trajectory which might be associated with the risk of dementia. Furthermore, the literature reviewed in this thesis shows that these separate analyses, even the time-varying nature of endogenous variables in the extended Cox model, always produce biased estimates and invalid inferences. The literature also shows that the likelihood for cognitive impairment may be increased due to recurrent depressive episodes, a finding that reinforces taking repeated measurements of depression (Donix et al., 2019; Kessing, 2012; PRESENT, 2012).

One of the main advantages of the present study over other studies is that the shared random effects joint modeling technique used for this analysis address the issues identified above by combining longitudinal process for depression and survival process for dementia into a single model. This statistical approach to data analysis increased our confidence level that our results are most likely efficient and valid. However, the causal relation can not be directly established from this study.

There is also possibility that the common risk factors for both depression and dementia may explain the observed association between two diseases. Our findings determined that a lack of economic activity and walking less frequently are common risk factors both for

depression and dementia. It is also documented that memory impairment, sleep disturbances, and impaired social functioning are common to both conditions and that common pathophysiological pathways, such as inflammation, neurodegeneration, and vascular risk factors, may very well explain the diseases' association (Singh-Manoux et al., 2017).

In addition, our study lacked information about the medications, including antidepressants, that respondents took that might have potential effect in modifying the association. Competing risk related bias is also important to take into account for true association. Older people with depression have a higher chance of dying earlier than those without. Therefore, this type of early censoring of depressed older people can lead to a biased lower risk of dementia.

Dementia is widely known to have a disproportionate impact on people over 65 (van der Flier & Scheltens, 2005) and, indeed, our study found age to be a significant risk factor for dementia risk. This is consistent with other studies (Kuo et al., 2015; Parikh et al., 2011; Qiu, De Ronchi, & Fratiglioni, 2007). Several cohort studies from the USA and Europe have also shown that the risk of dementia increases with age (Ganguli, Dodge, Chen, Belle, & DeKosky, 2000; Kukull et al., 2002; Launer et al., 1999; Ravaglia et al., 2005).

Although being male was significantly associated with a lower risk of dementia in univariate analysis, it was insignificant in the multivariable model (Andersen et al., 1999; Barnes et al., 2003; Ganguli et al., 2000; Karp et al., 2009; Ruitenberg, Ott, van Swieten, Hofman, & Breteler, 2001). This finding from our adjusted model is inconsistent with other

studies (Kuo et al., 2015; McAdams-DeMarco et al., 2018), where female has a higher risk of dementia.

In this follow-up study, respondents with no or less formal education had a greater risk of dementia compared to those who completed more than primary school (Katzman, 1993; Mortimer, Graves, et al., 1993). There are several longitudinal studies (Evans et al., 1997; Letenneur et al., 1999; Stern et al., 1994) that showed that individuals with higher education at younger age were at lower risk of developing dementia later in life. Education is therefore thought to protect against dementia-related pathology.

Though many studies (Flicker et al., 2005; Laurin, Verreault, Lindsay, MacPherson, & Rockwood, 2001; Sumic, Michael, Carlson, Howieson, & Kaye, 2007; Weuve et al., 2004) have determined that regular physical activity is a protective factor against the onset of dementia, our study did not find any form of physical activity (intense, medium, or a combination of the two) was a significant predictor of dementia. However, consistent with other works (Barendregt & Ott, 2005; Simons, Simons, McCallum, & Friedlander, 2006) walking in a week was identified as a factor in lowering the risk of dementia. A meta-analysis conducted by Quan et al. (2017) found that a low or decreased walking pace is significantly associated with an elevated risk of developing dementia in elderly. Another study also showed that regular exercise as simple as brisk walking for as little as 15 minutes a day protects brain structure and function (Sabayan & Sorond, 2017).

The univariate analysis determined the higher risk of dementia among people with co-morbidities including hypertension,and diabetes (Barendregt & Ott, 2005; Flicker et al.,

2005; Whitmer, Sidney, Selby, Johnston, & Yaffe, 2005), however, these were insignificant in multivariable model (Gilsanz et al., 2017). This finding is consistent with a cohort study of dementia in older Canadians where diabetes was not associated with mixed Alzheimer's/vascular dementia, incident Alzheimer's disease or all dementias (MacKnight, Rockwood, Awalt, & McDowell, 2002). Another study in Rochester, Minnesota, reported twice risk of Alzheimer's disease among men with diabetes vs non-diabetes and an insignificant association among women (Leibson et al., 1997). In contrast to other studies (Simard, Hudon, & van Reekum, 2009; Whitmer et al., 2005), we did not find any significant relation between the risk of dementia and other comorbidities (that is, CVD, anxiety, and gum disease) in both the univariate, and gum disease and multivariable analysis. This finding is consistent with a population-based Rotterdam study where, no association between anxiety and dementia is reported (de Bruijn et al., 2014).

Like Tyas, Manfreda, Strain, and Montgomery (2001) and Broe et al. (1990), we did not find any association between drinking alcohol and smoking and the incidence of dementia. Our findings contrast with other studies where monthly and weekly intake of alcohol (Truelsen, Thudium, & Grønbæk, 2002) or moderate alcohol use (Deng et al., 2006; Simons et al., 2006) significantly reduce the risk of dementia. Various studies found that current daily or more frequent smokers (Fillit, Nash, Rundek, & Zuckerman, 2008; Flicker et al., 2005; Ott et al., 1998; Reitz, den Heijer, van Duijn, Hofman, & Breteler, 2007; Whitmer et al., 2005) had a higher chance of developing dementia, but found no association between past smoking and the risk of dementia (Reitz et al., 2007).

Our finding is consistent with other studies (Evans et al., 1997; Fischer et al., 2009;

Scazufca, Almeida, & Menezes, 2010) that determined a significant association between no economic activity and an elevated risk dementia.

Although many studies (Helmer et al., 1999; Sundström, Westerlund, & Kotyrlo, 2016) reported a significant association between never having been married or living alone (Sundström et al., 2016) and dementia risk, our study did not find any association. This is also the finding from a case control study conducted by Beard, Kokmen, Offord, and Kurland (1992). Our study, however, observed the lowers risk of dementia among the people living with first and second generation household setting.

5.1 Strength and limitations

Among the strengths of this study is the fact that the KHPS permitted us to work with large scale, national level, prospective cohort data, where the identification of the main event of interest (dementia) and other co-morbidities was confirmed by clinical diagnosis codes. KHP is a population based study, so the results represent the general population. It also allowed for assessment of a relatively large number of potential social covariates. Moreover, this study is assumed to be free from interviewer bias as physicians and others contributing to the data did not know the hypothesis of our current study. All the co-morbidities of this data set are clinically diagnosed and there is very little chance of mis-classification of these diseases.

There are a few limitations of this study. The main limitation of this study is that we identified individuals with dementia from those who visited a medical clinic, health

complex, or hospital and were diagnosed with dementia. However, there may be additional study participants, particularly less severe cases, who were not assessed, thus potentially underestimating the true incidence of dementia. This concern also holds for all other comorbidities that were included in our analysis. Another limitation is that some relevant information (e.g. insomnia, head injury, cognitive impairment scores etc) were not available; thus, the confounding effect of these unmeasured variables cannot be controlled. Risk factors found in our study were only associated with the incidence of dementia, but causality cannot be determined with certainty due to other confounding factors which are not directly available in this data. Additionally, we have completely relied on medical record linkage without further validation. As this study was exclusively undertaken in Korea, generalizability to other countries is uncertain. Our data also did not include individuals living in long-term care, a potentially emerging demographic. From the methodological perspective, death might be the competing event of this study and this event alter the probability of occurring dementia. The total number of death in our study population was around 180, which might be not ignorable. So, there is still possibility to produce more valid and correct inference after addressing all of this information.

Another limitation of this study was that the computational part of joint analysis was very expensive. It took one and half days to get the output for current-value shared parameter model in this study. However, computing time increased as the number of complex structure and terms (number of random effects) increased. For example, when an interaction term was included in longitudinal submodel in addition to other setting of covariates, it took more than three days for getting the output of this analysis. Even the reported time by the author who developed the sas macro for this joint analysis with a

small number subjects was one and half hours (Garcia-Hernandez & Rizopoulos, 2018).

Chapter 6

Conclusion and Future Research

6.1 Concluding remarks

Determining the relation between depression and dementia is an ongoing and debated issue. We chose to investigate it once again using appropriate statistical methods to address gaps in previous research. Our research objective was to check whether endogenous time-dependent depression was associated with dementia and to measure its effects on dementia risk. Since conventional statistical methods could not properly address a time-varying variable that is endogenous in nature, we applied a joint modeling technique to deal with this problem. Our study found that single and baseline measurement of depression was not associated with the risk of developing dementia. However, repeated measurements of depression was significantly associated with dementia, meaning that people suffering from depression are at higher risk of developing dementia than those who do not suffer from depression.

We also found that walking not at all or less than three days in a week, being compar-

atively older (age>70 years old), having diabetes, having lower levels of education, and living in a household with multiple generations setting increased the risk of dementia.

Risk factors for longitudinal measurements of depression were also identified from this analysis. The univariate and even the separate multivariable analysis of these repeated measurements found that only CVD and anxiety are significant variables. However, the longitudinal submodel under joint modeling retained five variables (that is, sex, economic activity, walking in a week, and living with multiple generations in one household). The information of these variables are assumed to construct the better trajectories for individuals and this would be very useful to determine the association between longitudinal and survival outcome through current-value.

6.2 Future research

Since the longitudinal measurement of this joint model is binary and very limited research is done in this area, the joint modeling method will provide opportunity for researchers to have a more in-depth understanding of their study application. We can extend our work in several directions, some of which are listed below.

1. We know there is an extensive literature on the joint modeling technique dealing with one failure type for the time-to-event outcome. . However, in medical studies, researchers are sometimes interested in more than one possible time-to-event or where censoring is informative. These data are often called competing risks setting . To produce unbiased and efficient estimates, the informative dropout due to com-

peting risks or non-random censoring should be taken into account in joint models at the time of data analysis. Very little research considers competing risk in the joint model. Most joint modeling studies concentrate on continuous longitudinal measurement; no research focuses on joint modeling for categorical longitudinal outcomes and time-to-event data, particularly the competing risk data. For example, depression, a binary repeated measurement, might be associated with time-to-dementia and death preceding to the onset of dementia is competing event as it precludes the development of dementia. Ignoring this competing event in the joint model may lead to misleading inferences, especially in study of the elderly.

2. There might be two or more possible endogenous variables in a same study and they may be independent of each other. Since there is no available research that addresses these two independent endogenous variables at the same time in a joint model, there is an opportunity to work on this problem.

3. Most of the joint modeling approaches assume that the random effects are normally distributed. However, this assumption can be violated particularly, when dealing with medical or health data. There are several directions that this kind of research could go in, all of which are important.

Bibliography

Alamri, S. H., Bari, A. I., & Ali, A. T. (2017). Depression and associated factors in hospitalized elderly: a cross-sectional study in a saudi teaching hospital. *Annals of Saudi medicine*, *37*(2), 122–129.

Albert, P. S., & Shih, J. H. (2010). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics*, *66*(3), 983–987.

Alexopoulos, G. S., Meyers, B. S., Young, R. C., Campbell, S., Silbersweig, D., & Charlson, M. (1997). 'vascular depression'hypothesis. *Archives of general psychiatry*, *54*(10), 915–922.

Almeida, O., Hankey, G., Yeap, B., Golledge, J., & Flicker, L. (2017). Depression as a modifiable factor to decrease the risk of dementia. *Translational psychiatry*, *7*(5), e1117–e1117.

Anand, A. (2015). Understanding depression among older adults in six low-middle income countries using who-sage survey. *Behav Health*, *1*(2), 1–11.

Andersen, K., Launer, L. J., Dewey, M. E., Letenneur, L., Ott, A., Copeland, J., ... others (1999). Gender differences in the incidence of ad and vascular dementia: The eurodem studies. *Neurology*, *53*(9), 1992–1992.

Andersen, K., Lolk, A., Kragh-Sørensen, P., Petersen, N. E., & Green, A. (2005). Depression and the risk of alzheimer disease. *Epidemiology*, 233–238.

Arbeev, K. G., Akushevich, I., Kulminski, A. M., Ukraintseva, S. V., & Yashin, A. I. (2014). Joint analyses of longitudinal and time-to-event data in research on aging: Implications for predicting health and survival. *Frontiers in public health*, *2*, 228.

Barendregt, J. J., & Ott, A. (2005). Consistency of epidemiologic estimates. *European journal of epidemiology*, *20*(10), 827–832.

Barnes, L., Wilson, R., Schneider, J., Bienias, J., Evans, D., & Bennett, D. (2003). Gender, cognitive decline, and risk of ad in older persons. *Neurology*, *60*(11), 1777–1781.

Baumgart, M., Snyder, H. M., Carrillo, M. C., Fazio, S., Kim, H., & Johns, H. (2015). Summary of the evidence on modifiable risk factors for cognitive decline and dementia: a population-based perspective. *Alzheimer's & Dementia*, *11*(6), 718–726.

Beard, C. M., Kokmen, E., Offord, K. P., & Kurland, L. T. (1992). Lack of association between alzheimer's disease and education, occupation, marital status, or living arrangement. *Neurology*, *42*(11), 2063–2063.

Becker, J. T., Chang, Y.-F., Lopez, O. L., Dew, M. A., Sweet, R. A., Barnes, D., ... Reynolds III, C. F. (2009). Depressed mood is not a risk factor for incident dementia in a community-based cohort. *The American Journal of Geriatric Psychiatry*, *17*(8), 653–663.

Bland, J. M., & Altman, D. G. (2004). The logrank test. *Bmj*, *328*(7447), 1073.

Broe, G., Henderson, A., Creasey, H., McCusker, E., Korten, A., Jorm, A., ... Anthony, J. (1990). A case-control study of alzheimer's disease in australia. *Neurology*, *40*(11), 1698–1698.

Brown, E. R., & Ibrahim, J. G. (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics*, *59*(3), 686–693.

Butters, M. A., Becker, J. T., Nebes, R. D., Zmuda, M. D., Mulsant, B. H., Pollock, B. G., & Reynolds III, C. F. (2000). Changes in cognitive functioning following treatment of late-life depression. *American Journal of Psychiatry*, *157*(12), 1949–1954.

Butters, M. A., Young, J. B., Lopez, O., Aizenstein, H. J., Mulsant, B. H., Reynolds III, C. F., ... Becker, J. T. (2008). Pathways linking late-life depression to persistent cognitive impairment and dementia. *Dialogues in clinical neuroscience*, *10*(3), 345.

Bycott, P., & Taylor, J. (1998). A comparison of smoothing techniques for cd4 data measured with error in a time-dependent cox proportional hazards model. *Statistics in medicine*, *17*(18), 2061–2077.

Canada, A. S. (2018). *Treatment options.* Retrieved from https://alzheimer.ca/en/Home/About-dementia/Treatment-options

Capella-McDonnall, M. E. (2005). The effects of single and dual sensory loss on symptoms of

depression in the elderly. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, *20*(9), 855–861.

Chang-Quan, H., Xue-Mei, Z., Bi-Rong, D., Zhen-Chan, L., Ji-Rong, Y., & Qing-Xiu, L. (2009). Health status and risk for depression among the elderly: a meta-analysis of published literature. *Age and ageing*, *39*(1), 23–30.

Chen, J.-H., Lin, K.-P., & Chen, Y.-C. (2009). Risk factors for dementia. *Journal of the Formosan Medical Association*, *108*(10), 754–764.

Chen, R., Hu, Z., Wei, L., Ma, Y., Liu, Z., & Copeland, J. R. (2011). Incident dementia in a defined older chinese population. *PloS one*, *6*(9), e24817.

Chen, R., Hu, Z., Wei, L., Qin, X., McCracken, C., & Copeland, J. R. (2008). Severity of depression and risk for subsequent dementia: cohort studies in china and the uk. *The British Journal of Psychiatry*, *193*(5), 373–377.

Cheruvu, V. K., & Chiyaka, E. T. (2019). Prevalence of depressive symptoms among older adults who reported medical cost as a barrier to seeking health care: findings from a nationally representative sample. *BMC geriatrics*, *19*(1), 192.

Chi, Y.-Y., & Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, *62*(2), 432–445.

Choi, D., Choi, S., & Park, S. M. (2018). Effect of smoking cessation on the risk of dementia: a longitudinal study. *Annals of Clinical and Translational Neurology*, *5*(10), 1192–1199.

Choi, J., Cai, J., Zeng, D., & Olshan, A. F. (2015). Joint analysis of survival time and longitudinal categorical outcomes. *Statistics in biosciences*, *7*(1), 19–47.

Choi, J., Zeng, D., Olshan, A. F., & Cai, J. (2018). Joint modeling of survival time and longitudinal outcomes with flexible random effects. *Lifetime data analysis*, *24*(1), 126–152.

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, *89*(2), 232–238.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G., ... Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein e type 4 allele and the

risk of alzheimer's disease in late onset families. *Science*, *261*(5123), 921–923.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276.

Cui, S.-S., Du, J.-J., Fu, R., Lin, Y.-Q., Huang, P., He, Y.-C., ... Chen, S.-D. (2017). Prevalence and risk factors for depression and anxiety in chinese patients with parkinson disease. *BMC geriatrics*, *17*(1), 270.

Dafni, U. G., & Tsiatis, A. A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 1445–1462.

de Bruijn, R. F., Direk, N., Mirza, S. S., Hofman, A., Koudstaal, P. J., Tiemeier, H., & Ikram, M. A. (2014). Anxiety is not associated with the risk of dementia or cognitive decline: the rotterdam study. *The American Journal of Geriatric Psychiatry*, *22*(12), 1382–1390.

De Gruttola, V., & Tu, X. M. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, 1003–1014.

Deng, J., Cao, C., Jiang, Y., Peng, B., Wang, T., Yan, K., ... Wang, Z. (2018). Prevalence and effect factors of dementia among the community elderly in chongqing, china. *Psychogeriatrics*.

Deng, J., Zhou, D. H., Li, J., Wang, Y. J., Gao, C., & Chen, M. (2006). A 2-year follow-up study of alcohol consumption and risk of dementia. *Clinical neurology and neurosurgery*, *108*(4), 378–383.

Donix, M., Haussmann, R., Helling, F., Zweiniger, A., Werner, A., Linn, J., ... Buthut, M. (2019). Risk factors for dementia are not associated with cognitive dysfunction in young people with major depressive disorder. *Journal of affective disorders*, *245*, 140–144.

Evans, D. A., Hebert, L. E., Beckett, L. A., Scherr, P. A., Albert, M. S., Chown, M. J., ... Taylor, J. O. (1997). Education and other measures of socioeconomic status and risk of incident alzheimer disease in a defined population of older persons. *Archives of neurology*, *54*(11), 1399–1405.

Faucett, C. L., & Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in medicine*, *15*(15),

1663–1685.

Fillit, H., Nash, D. T., Rundek, T., & Zuckerman, A. (2008). Cardiovascular risk factors and dementia. *The American journal of geriatric pharmacotherapy*, *6*(2), 100–118.

Fischer, C., Yeung, E., Hansen, T., Gibbons, S., Fornazzari, L., Ringer, L., & Schweizer, T. (2009). Impact of socioeconomic status on the prevalence of dementia in an inner city memory disorders clinic. *International psychogeriatrics*, *21*(6), 1096–1104.

Fitrianto, A., & Jiin, R. L. T. (2013). Several types of residuals in cox regression model: an empirical study. *Int J Math Anal*, *7*, 2645–2654.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.

Flicker, L., Almeida, O. P., Acres, J., Le, M. T., Tuohy, R. J., Jamrozik, K., ... Norman, P. (2005). Predictors of impaired cognitive function in men over the age of 80 years: results from the health in men study. *Age and Ageing*, *34*(1), 77–80.

Fountoulakis, K. N., O'Hara, R., Iacovides, A., Camilleri, C. P., Kaprinis, S., Kaprinis, G., & Yesavage, J. (2003). Unipolar late-onset depression: A comprehensive review. *Annals of general hospital psychiatry*, *2*(1), 11.

Ganguli, M., Dodge, H., Chen, P., Belle, S., & DeKosky, S. (2000). Ten-year incidence of dementia in a rural elderly us community population: the movies project. *Neurology*, *54*(5), 1109–1116.

Ganguli, M., Du, Y., Dodge, H. H., Ratcliff, G. G., & Chang, C.-C. H. (2006). Depressive symptoms and cognitive decline in late life: a prospective epidemiological study. *Archives of general psychiatry*, *63*(2), 153–160.

Garcia-Hernandez, A., & Rizopoulos, D. (2018). % jm: A sas macro to fit jointly generalized mixed models for longitudinal data and time-to-event responses. *Journal of Statistical Software*, *84*(1), 1–29.

Gasparini, A., Abrams, K. R., Barrett, J. K., Major, R. W., Sweeting, M. J., Brunskill, N. J., & Crowther, M. J. (2020). Mixed-effects models for health care longitudinal data with an informative visiting process: A monte carlo simulation study. *Statistica Neerlandica*, *74*(1), 5–23.

Gatchel, J. R., Rabin, J. S., Buckley, R. F., Locascio, J. J., Quiroz, Y. T., Yang, H.-S., ... others (2019). Longitudinal association of depression symptoms with cognition and cortical amyloid among community-dwelling older adults. *JAMA Network Open*, *2*(8), e198964–e198964.

Gatz, M., Fratiglioni, L., Johansson, B., Berg, S., Mortimer, J. A., Reynolds, C. A., ... Pedersen, N. L. (2005). Complete ascertainment of dementia in the swedish twin registry: the harmony study. *Neurobiology of aging*, *26*(4), 439–447.

Ghisletta, P., McArdle, J. J., & Lindenberger, U. (2006). Longitudinal cognition-survival relations in old and very old age: 13-year data from the berlin aging study. *European Psychologist*, *11*(3), 204–223.

Gilsanz, P., Mayeda, E. R., Glymour, M. M., Quesenberry, C. P., Mungas, D. M., DeCarli, C., ... Whitmer, R. A. (2017). Female sex, early-onset hypertension, and risk of dementia. *Neurology*, *89*(18), 1886–1893.

Gracia-García, P., De-La-Cámara, C., Santabárbara, J., Lopez-Anton, R., Quintanilla, M. A., Ventura, T., ... others (2015). Depression and incident alzheimer disease: the impact of disease severity. *The American Journal of Geriatric Psychiatry*, *23*(2), 119–129.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Haseen, F., & Prasartkul, P. (2011). Predictors of depression among older people living in rural areas of thailand. *Bangladesh medical research council bulletin*, *37*(2), 51–56.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Hatfield, L. A., Boye, M. E., Hackshaw, M. D., & Carlin, B. P. (2012). Multilevel bayesian models for survival times and longitudinal patient-reported outcomes with many zeros. *Journal of the American Statistical Association*, *107*(499), 875–885.

Hazra, A., & Gogtay, N. (2017). Biostatistics series module 9: Survival analysis. *Indian journal of dermatology*, *62*(3), 251.

He, Q., Yang, L., Shi, S., Gao, J., Tao, M., Zhang, K., ... others (2014). Smoking and major depressive disorder in chinese women. *PloS one*, *9*(9), e106287.

Hebert, L. E., Scherr, P. A., Beckett, L. A., Funkenstein, H. H., Albert, M. S., Chown, M. J., & Evans, D. A. (1992). Relation of smoking and alcohol consumption to incident alzheimer's disease. *American journal of epidemiology*, *135*(4), 347–355.

Helmer, C., Damon, D., Letenneur, L., Fabrigoule, C., Barberger-Gateau, P., Lafont, S., ... others (1999). Marital status and risk of alzheimer's disease: a french population-based cohort study. *Neurology*, *53*(9), 1953–1953.

Helvik, A.-S., Barca, M. L., Bergh, S., Šaltytė-Benth, J., Kirkevold, Ø., & Borza, T. (2019). The course of depressive symptoms with decline in cognitive function-a longitudinal study of older adults receiving in-home care at baseline. *BMC Geriatrics*, *19*(1), 1–14.

Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, *1*(4), 465–480.

Hogan, J. W., & Laird, N. M. (1998). Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. *Statistical Methods in Medical Research*, *7*(1), 28–48.

Holmquist, S., Nordström, A., & Nordström, P. (2020). The association of depression with subsequent dementia diagnosis: A swedish nationwide cohort study from 1964 to 2016.

Hsieh, F., Tseng, Y.-K., & Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, *62*(4), 1037–1043.

Hu, W., Li, G., & Li, N. (2009). A bayesian approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in medicine*, *28*(11), 1601–1619.

Huang, X., Li, G., Elashoff, R. M., & Pan, J. (2011). A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime data analysis*, *17*(1), 80–100.

Huang, Y., Dagne, G., & Wu, L. (2011). Bayesian inference on joint models of hiv dynamics for time-to-event and longitudinal data with skewness and covariate measurement errors. *Statistics in Medicine*, *30*(24), 2930–2946.

Hwang, Y.-T., Huang, C.-H., Wang, C.-C., Lin, T.-Y., & Tseng, Y.-K. (2019). Joint modelling of longitudinal binary data and survival data. *Journal of Applied Statistics*, *46*(13), 2357–2371.

Ibrahim, J. G., Chu, H., & Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, *28*(16), 2796.

Iliffe, S., Kharicha, K., Harari, D., Swift, C., Gillmann, G., & Stuck, A. (2005). Self-reported visual function in healthy older people in britain: an exploratory study of associations with age, sex, depression, education and income. *Family practice*, *22*(6), 585–590.

Indrayan, A. (2012). *Medical biostatistics*. Chapman and Hall/CRC.

Instiute, S. (2015). Sas 9.4 graph template language: User's guide.

Jacqmin-Gadda, H., Commenges, D., & Dartigues, J.-F. (2006). Random changepoint model for joint modeling of cognitive decline and dementia. *Biometrics*, *62*(1), 254–260.

Johansson, L., Guerra, M., Prince, M., Hörder, H., Falk, H., Stubbs, B., & Prina, A. M. (2019). Associations between depression, depressive symptoms, and incidence of dementia in latin america: A 10/66 dementia research group study. *Journal of Alzheimer's Disease*, *69*(2), 433–441.

Kalbfleisch, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data* (Vol. 360). John Wiley & Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

Karp, A., Andel, R., Parker, M. G., Wang, H.-X., Winblad, B., & Fratiglioni, L. (2009). Mentally stimulating activities at work during midlife and dementia risk after age 75: follow-up study from the kungsholmen project. *The American Journal of Geriatric Psychiatry*, *17*(3), 227–236.

Katzman, R. (1993). Education and the prevalence of dementia and alzheimer's disease. *Neurology*.

Kessing, L. V. (2012). Depression and the risk for dementia. *Current opinion in psychiatry*, *25*(6), 457–461.

Kleinbaum, D. G., & Klein, M. (2011). Survival analysis: A self-learning text, 2005. *New York, Springer-Verlag*.

Kukull, W. A., Higdon, R., Bowen, J. D., McCormick, W. C., Teri, L., Schellenberg, G. D., ... Larson, E. B. (2002). Dementia and alzheimer disease incidence: a prospective cohort

study. *Archives of neurology*, *59*(11), 1737–1746.

Kuo, S.-C., Lai, S.-W., Hung, H.-C., Muo, C.-H., Hung, S.-C., Liu, L.-L., ... Sung, F.-C. (2015). Association between comorbidities and dementia in diabetes mellitus patients: population-based retrospective cohort study. *Journal of Diabetes and its Complications*, *29*(8), 1071–1076.

Laird, N. M., Ware, J. H., et al. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Launer, L., Andersen, K., Dewey, M., Letenneur, L., Ott, A., Amaducci, L., ... others (1999). Rates and risk factors for dementia and alzheimer's disease: results from eurodem pooled analyses. *Neurology*, *52*(1), 78–78.

Laurin, D., Verreault, R., Lindsay, J., MacPherson, K., & Rockwood, K. (2001). Physical activity and risk of cognitive impairment and dementia in elderly persons. *Archives of neurology*, *58*(3), 498–504.

Law, N. J., Taylor, J. M., & Sandler, H. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*, *3*(4), 547–563.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data* (Vol. 362). John Wiley & Sons.

Lee, E. T., & Wang, J. (2003). *Statistical methods for survival data analysis* (Vol. 476). John Wiley & Sons.

Lee, J.-H., Park, M. A., Park, M. J., & Jo, Y. S. (2018). Clinical characteristics and related risk factors of depression in patients with early copd. *International journal of chronic obstructive pulmonary disease*, *13*, 1583.

Leibson, C. L., Rocca, W. A., Hanson, V., Cha, R., Kokmen, E., O'brien, P., & Palumbo, P. (1997). Risk of dementia among persons with diabetes mellitus: a population-based cohort study. *American journal of epidemiology*, *145*(4), 301–308.

Letenneur, L., Gilleron, V., Commenges, D., Helmer, C., Orgogozo, J.-M., & Dartigues, J.-F. (1999). Are sex and educational level independent predictors of dementia and alzheimer's disease? incidence data from the paquid project. *Journal of Neurology, Neurosurgery &*

Psychiatry, *66*(2), 177–183.

Li, G., Wang, L. Y., Shofer, J. B., Thompson, M. L., Peskind, E. R., McCormick, W., ... Larson, E. B. (2011). Temporal relationship between depression and dementia: findings from a large community-based 15-year follow-up study. *Archives of general psychiatry*, *68*(9), 970–977.

Li, L., Wu, C., Gan, Y., Qu, X., & Lu, Z. (2016). Insomnia and the risk of depression: a meta-analysis of prospective cohort studies. *BMC psychiatry*, *16*(1), 375.

Li, N., Pang, L., Chen, G., Song, X., Zhang, J., & Zheng, X. (2011). Risk factors for depression in older adults in beijing. *The Canadian Journal of Psychiatry*, *56*(8), 466–473.

Li, S., Zheng, M., & Gao, S. (2017). Joint modeling of longitudinal cholesterol measurements and time to onset of dementia in an elderly african american cohort. *Biostatistics & Epidemiology*, *1*(1), 148–160.

Lin, X., Taylor, J. M., & Ye, W. (2008). A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and its Interface*, *1*(1), 33–45.

Luppa, M., Luck, T., Ritschel, F., Angermeyer, M. C., Villringer, A., & Riedel-Heller, S. G. (2013). Depression and incident dementia. an 8-year population-based prospective study. *PLoS One*, *8*(3), e59246.

MacKnight, C., Rockwood, K., Awalt, E., & McDowell, I. (2002). Diabetes mellitus and the risk of dementia, alzheimer's disease and vascular cognitive impairment in the canadian study of health and aging. *Dementia and geriatric cognitive disorders*, *14*(2), 77–83.

McAdams-DeMarco, M. A., Daubresse, M., Bae, S., Gross, A. L., Carlson, M. C., & Segev, D. L. (2018). Dementia, alzheimer's disease, and mortality after hemodialysis initiation. *Clinical Journal of the American Society of Nephrology*, *13*(9), 1339–1347.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, *92*(437), 162–170.

Meng, X., Brunet, A., Turecki, G., Liu, A., D'Arcy, C., & Caron, J. (2017). Risk factor modifications and depression incidence: a 4-year longitudinal canadian cohort of the montreal catchment area study. *BMJ open*, *7*(6), e015156.

Meng, X., & D'Arcy, C. (2013). The projected effect of increasing physical activity on reducing the prevalence of common mental disorders among canadian men and women: A national population-based community study. *Preventive medicine*, *56*(1), 59–63.

Metropolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, *44*(247), 335–341.

Mirkena, Y., Reta, M. M., Haile, K., Nassir, Z., & Sisay, M. M. (2018). Prevalence of depression and associated factors among older adults at ambo town, oromia region, ethiopia. *BMC psychiatry*, *18*(1), 338.

Mondal, P., Lim, H. J., Team, O. C. S., et al. (2018). The effect of msm and cd4+ count on the development of cancer aids (aids-defining cancer) and non-cancer aids in the haart era. *Current HIV research*, *16*(4), 288–296.

Mortimer, J. A., Graves, A. B., et al. (1993). Education and other socioeconomic determinants of dementia and alzheimer's disease. *NEUROLOGY-MINNEAPOLIS-*, *43*, 39–39.

Mukamal, K. J., Kuller, L. H., Fitzpatrick, A. L., Longstreth Jr, W. T., Mittleman, M. A., & Siscovick, D. S. (2003). Prospective study of alcohol consumption and risk of dementia in older adults. *Jama*, *289*(11), 1405–1413.

NIH, U. (2020). *Depression and older adults @ONLINE.* Retrieved from https://www.nia.nih.gov/health/depression-and-older-adults

NIMH. (2020). *Adults: Depression @ONLINE.* Retrieved from https://www.nimh.nih.gov/research/research-conducted-at-nimh/join-a-study/adults/adults-depression.shtml

Norton, S., Matthews, F. E., Barnes, D. E., Yaffe, K., & Brayne, C. (2014). Potential for primary prevention of alzheimer's disease: an analysis of population-based data. *The Lancet Neurology*, *13*(8), 788–794.

Olazaran, J., Trincado, R., & Bermejo-Pareja, F. (2013). Cumulative effect of depression on dementia risk. *International Journal of Alzheimer's Disease*, *2013*.

Organization, W. H., et al. (2004). Promoting mental health: Concepts, emerging evidence, practice: Summary report.

Organization, W. H., et al. (2017). *Depression and other common mental disorders: global health*

estimates (Tech. Rep.). World Health Organization.

Ott, A., Slooter, A., Hofman, A., van Harskamp, F., Witteman, J., Van Broeckhoven, C., ... Breteler, M. (1998). Smoking and risk of dementia and alzheimer's disease in a population-based cohort study: the rotterdam study. *The Lancet*, *351*(9119), 1840–1843.

Parikh, N. M., Morgan, R. O., Kunik, M. E., Chen, H., Aparasu, R. R., Yadav, R. K., ... Johnson, M. L. (2011). Risk factors for dementia in patients over 65 with diabetes. *International journal of geriatric psychiatry*, *26*(7), 749–757.

Park, J.-H., Eum, J.-H., Bold, B., & Cheong, H.-K. (2013). Burden of disease due to dementia in the elderly population of korea: present and future. *BMC Public Health*, *13*(1), 293.

Pawitan, Y., & Self, S. (1993). Modeling disease marker processes in aids. *Journal of the American Statistical Association*, *88*(423), 719–726.

PHA, C. (2019). *A dementia strategy for canada: Together we aspire @ONLINE.* Retrieved from https://www.canada.ca/en/public-health/services/publications/diseases-conditions/dementia-strategy.html

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, *69*(2), 331–342.

PRESENT, A. (2012). Midlife vs late-life depressive symptoms and risk of dementia. *Arch Gen Psychiatry*, *69*(5), 493–498.

Prince, M., Guerchet, M., & Prina, M. (2013). *The global impact of dementia 2013-2050.* Alzheimer's Disease International.

Qiu, C., De Ronchi, D., & Fratiglioni, L. (2007). The epidemiology of the dementias: an update. *Current opinion in psychiatry*, *20*(4), 380–385.

Quan, M., Xun, P., Chen, C., Wen, J., Wang, Y., Wang, R., ... He, K. (2017). Walking pace and the risk of cognitive decline and dementia in elderly populations: a meta-analysis of prospective cohort studies. *The Journals of Gerontology: Series A*, *72*(2), 266–270.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rajkumar, A., Thangadurai, P., Senthilkumar, P., Gayathri, K., Prince, M., & Jacob, K. (2009). Nature, prevalence and factors associated with depression among the elderly in a rural

south indian community. *International psychogeriatrics*, *21*(2), 372–378.

Rasmussen, H., Rosness, T. A., Bosnes, O., Salvesen, Ø., Knutli, M., & Stordal, E. (2018). Anxiety and depression as risk factors in frontotemporal dementia and alzheimer's disease: The hunt study. *Dementia and Geriatric Cognitive Disorders Extra*, *8*(3), 414–425.

Ravaglia, G., Forti, P., Maioli, F., Martelli, M., Servadei, L., Brunetti, N., ... Mariani, E. (2005). Incidence and etiology of dementia in a large elderly italian population. *Neurology*, *64*(9), 1525–1530.

Rawana, J. S., Morgan, A. S., Nguyen, H., & Craig, S. G. (2010). The relation between eating- and weight-related disturbances and depression in adolescence: a review. *Clinical child and family psychology review*, *13*(3), 213–230.

R. Brown, E., & G. Ibrahim, J. (2003). A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, *59*(2), 221–228.

Read, S., Wittenberg, R., Karagiannidou, M., Anderson, R., & Knapp, M. (2017). The effect of midlife risk factors on dementia in older age.

Reitz, C., den Heijer, T., van Duijn, C., Hofman, A., & Breteler, M. (2007). Relation between smoking and risk of dementia and alzheimer disease: the rotterdam study. *Neurology*, *69*(10), 998–1005.

Ren, L., Zheng, Y., Wu, L., Gu, Y., He, Y., Jiang, B., ... Li, J. (2018). Investigation of the prevalence of cognitive impairment and its risk factors within the elderly population in shanghai, china. *Scientific reports*, *8*(1), 3575.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in r*. Chapman and Hall/CRC.

Rizopoulos, D. (2014). The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *arXiv preprint arXiv:1404.7625*.

Rizopoulos, D., & Ghosh, P. (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine*, *30*(12), 1366–1380.

Rizopoulos, D., Verbeke, G., & Lesaffre, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical*

Society: Series B (Statistical Methodology), *71*(3), 637–654.

Rizopoulos, D., Verbeke, G., Lesaffre, E., & Vanrenterghem, Y. (2008). A two-part joint model for the analysis of survival and longitudinal binary data with excess zeros. *Biometrics*, *64*(2), 611–619.

Ruitenberg, A., Ott, A., van Swieten, J. C., Hofman, A., & Breteler, M. M. (2001). Incidence of dementia: does gender make a difference? *Neurobiology of aging*, *22*(4), 575–580.

Sabayan, B., & Sorond, F. (2017). Reducing risk of dementia in older age. *Jama*, *317*(19), 2028–2028.

Saczynski, J. S., Beiser, A., Seshadri, S., Auerbach, S., Wolf, P., & Au, R. (2010). Depressive symptoms and risk of dementia: the framingham heart study. *Neurology*, *75*(1), 35–41.

Scazufca, M., Almeida, O. P., & Menezes, P. R. (2010). The role of literacy, occupation and income in dementia prevention: the são paulo ageing & health study (spah). *International psychogeriatrics*, *22*(8), 1209–1215.

Schabenberger, O., et al. (2005). Introducing the glimmix procedure for generalized linear mixed models. *SUGI 30 Proceedings*, *196*.

Schober, P., & Vetter, T. R. (2018). Survival analysis and interpretation of time-to-event data: the tortoise and the hare. *Anesthesia and analgesia*, *127*(3), 792.

Simard, M., Hudon, C., & van Reekum, R. (2009). Psychological distress and risk for dementia. *Current psychiatry reports*, *11*(1), 41.

Simons, L. A., Simons, J., McCallum, J., & Friedlander, Y. (2006). Lifestyle factors and risk of dementia: Dubbo study of the elderly. *Medical Journal of Australia*, *184*(2), 68–70.

Singh-Manoux, A., Dugravot, A., Fournier, A., Abell, J., Ebmeier, K., Kivimäki, M., & Sabia, S. (2017). Trajectories of depressive symptoms before diagnosis of dementia: a 28-year follow-up study. *JAMA psychiatry*, *74*(7), 712–718.

Starkstein, S. E., & Almeida, O. P. (2003). Understanding cognitive impairment and dementia: stroke studies. *Current Opinion in Psychiatry*, *16*(6), 615–620.

Steffens, D. C., Otey, E., Alexopoulos, G. S., Butters, M. A., Cuthbert, B., Ganguli, M., ... others (2006). Perspectives on depression, mild cognitive impairment, and cognitive decline. *Archives of general psychiatry*, *63*(2), 130–138.

Stern, Y., Gurland, B., Tatemichi, T. K., Tang, M. X., Wilder, D., & Mayeux, R. (1994). Influence of education and occupation on the incidence of alzheimer's disease. *Jama*, *271*(13), 1004–1010.

Subramaniam, M., Abdin, E., Sambasivam, R., Vaingankar, J. A., Picco, L., Pang, S., ... others (2016). Prevalence of depression among older adults-results from the well-being of the singapore elderly study. *Ann Acad Med Singapore*, *45*, 123–33.

Sumic, A., Michael, Y. L., Carlson, N. E., Howieson, D. B., & Kaye, J. A. (2007). Physical activity and the risk of dementia in oldest old. *Journal of aging and health*, *19*(2), 242–259.

Sundström, A., Westerlund, O., & Kotyrlo, E. (2016). Marital status and risk of dementia: a nationwide population-based prospective study from sweden. *BMJ open*, *6*(1), e008565.

Sweeting, M. J., & Thompson, S. G. (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, *53*(5), 750–763.

Truelsen, T., Thudium, D., & Grønbæk, M. (2002). Amount and type of alcohol and risk of dementia: the copenhagen city heart study. *Neurology*, *59*(9), 1313–1319.

Tseng, Y.-K., Hsieh, F., & Wang, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika*, *92*(3), 587–603.

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 809–834.

Tsiatis, A. A., Degruttola, V., & Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, *90*(429), 27–37.

Tyas, S. L., Manfreda, J., Strain, L. A., & Montgomery, P. R. (2001). Risk factors for alzheimer's disease: a population-based, longitudinal study in manitoba, canada. *International journal of epidemiology*, *30*(3), 590–597.

van der Flier, W. M., & Scheltens, P. (2005). Epidemiology and risk factors of dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, *76*(suppl 5), v2–v7.

Viviani, S., Alfó, M., & Rizopoulos, D. (2014). Generalized linear mixed joint model for longitudinal and survival outcomes. *Statistics and Computing*, *24*(3), 417–427.

Wang, Y., & Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, *96*(455), 895–905.

Weuve, J., Kang, J. H., Manson, J. E., Breteler, M. M., Ware, J. H., & Grodstein, F. (2004). Physical activity, including walking, and cognitive function in older women. *Jama*, *292*(12), 1454–1461.

Whitmer, R. A., Sidney, S., Selby, J., Johnston, S. C., & Yaffe, K. (2005). Midlife cardiovascular risk factors and risk of dementia in late life. *Neurology*, *64*(2), 277–281.

WHO. (2019). *Dementia @ONLINE.* Retrieved from https://www.who.int/news-room/fact-sheets/detail/dementia

Wu, L., Liu, W., & Hu, X. (2010). Joint inference on hiv viral dynamics and immune suppression in presence of measurement errors. *Biometrics*, *66*(2), 327–335.

Wu, L., Liu, W., Yi, G. Y., & Huang, Y. (2012). Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, *2012*.

Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 330–339.

Xu, J., & Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *50*(3), 375–387.

Yang, L., Jin, X., Yan, J., Jin, Y., Yu, W., Wu, H., & Xu, S. (2016). Prevalence of dementia, cognitive status and associated risk factors among elderly of zhejiang province, china in 2014. *Age and ageing*, *45*(5), 708–712.

Ye, Q., & Wu, L. (2017). Two-step and likelihood methods for joint models of longitudinal and survival data. *Communications in Statistics-Simulation and Computation*, *46*(8), 6019–6033.

Yu, B., & Ghosh, P. (2010). Joint modeling for cognitive trajectory and risk of dementia in the presence of death. *Biometrics*, *66*(1), 294–300.

Yunming, L., Changsheng, C., Haibo, T., Wenjun, C., Shanhong, F., Yan, M., ... Qianzhen, H. (2012). Prevalence and risk factors for depression in older people in xi an china: a

community-based study. *International journal of geriatric psychiatry*, *27*(1), 31–39.

Appendix A

Ethical Approval Letter

UNIVERSITY OF SASKATCHEWAN

To:	Hyun Lim, Department of Community Health and Epidemiology
Sub-Investigators:	Razieh Safaripour, College of Medicine Cheng Yanzhao Cheng, School of Public Health Kabir Md Rasel Kabir, School of Public Health Kim Min Young Kim, School of Public Health
Date:	February 13, 2020
RE:	Behavioural Ethics Application ID 1759

Thank you for submitting your project entitled: "Statistical methods in epidemiology using South Korean Health Panel (KHP) Data". This project meets the requirements for exemption status as per **Article 2.2 of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2018)**, which states "Research does not require REB review when it relies exclusively on information that is:

a. publicly available through a mechanism set out by legislation or regulation and that is protected by law; or
b. in the public domain and the individuals to whom the information refers have no reasonable expectation of privacy."

It should be noted that though your project is exempt of ethics review, your project should be conducted in an ethical manner (i.e. in accordance with the information that you submitted). It should also be noted that any deviation from the original methodology and/or research question should be brought to the attention of the Behavioural Research Ethics Board for further review.

Digitally Approved by Vivian Ramsden, Vice-Chair
Behavioural Research Ethics Board
University of Saskatchewan