GROUP-BASED TRAJECTORY MODELING WITH BINARY AND ZERO-INFLATED COUNT OUTCOMES: APPLICATION TO GERIATRIC PNEUMONIA

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 Biostatistics Program
within the School of Public Health
University of Saskatchewan
Saskatoon
by

Min young Kim

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 whole or part should be addressed to:

Department Head of School of Public Health
College of Medicine
University of Saskatchewan Health Sciences Building E-Wing, 104 Clinic Place
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

A developmental trajectory is defined as an evolution of an outcome over age or time (Nagin, 2005). Several statistical approaches are available for trajectory analysis. Hierarchical modeling and latent growth curve modeling are most commonly used. However, this thesis is focused on widely used method, “Group-based trajectory modeling.”

Group-based trajectory modeling (GBTM) is an application of finite mixture modeling that the population is composed of distinct groups, each with a different underlying trajectory and every subject in the group approximately follows the same patterns of behavior of outcome over age or time (Nagin, 1999).

This thesis utilized the Korean Health Panel Study, which included 4007 individuals 65 years old or older at the baseline. Trajectory analysis was conducted with GBTM for geriatric pneumonia with binary and count outcomes. The models were compared and the binary outcome trajectory model was considered a better fit model. Both the binary outcome trajectory model and the zero-inflated count outcome trajectory model identified three trajectory groups with similar shapes: “low-flat,” “low-to-high,” and “high-to-low.” The majority of the participants belonged to the “low-flat” group. In the binary outcome trajectory model, having three household members, having a disability, and having a chronic respiratory disease were significant risk factors for the pneumonia trajectory groups. In the zero-inflated count outcome trajectory model, being male and having a chronic respiratory disease were the significant risk factors.

ACKNOWLEDGMENTS

Foremost, I would like to express my deep and sincere gratitude to my supervisor Dr. Hyun J. Lim for the excellent guidance, strong motivation, advice, and continuous support. Her guidance helped me throughout all the research and the writing of this thesis. I really appreciate her time, patience and mentoring.

My sincere thanks also go to Dr. Yanzhao Cheng for his support and advice, especially in the context of statistical analysis.

My thanks and appreciations go to my supervisory committee members Dr. Prosanta Mondal and Dr. Micheal Szafron, for taking their precious time to assess my thesis and giving me constructive comments.

My sincere thanks go to the School of Public Health, University of Saskatchewan for departmental and financial support during the program. Also, I would like to thank the Korean government, for allowing me to use the Korea Health Panel Study dataset in this thesis.

Last but not least, I could not have undertaken this journey without my family and friends. I am really grateful for their continuous and strong support. They had faith in me even when I didn't believe in myself. This thesis is dedicated to them.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

CAP Community-Acquired Pneumonia

CDC Centers for Disease Control and Prevention

CI Confidence Interval

COVID-19 Coronavirus Disease of 2019

EM Expectation-Maximization

GBTM Group-Based Trajectory Modeling

GEE Generalized Estimating Equations

HAP Hospital-Acquired Pneumonia

HCAP Health Care-Associated Pneumonia

HIV Human Immunodeficiency Virus

ICU Intensive Care Unit

IRLS Iteratively Reweighted Least Squares

KHPS Korea Health Panel Survey

MLE Maximum Likelihood Estimation

PP Posterior Probability

VAP Ventilator-Associated Pneumonia

VIF Variance Inflation Factor

WHO World Health Organization

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

1.1 Background

Most population-based study data sets have many observations that might not contain certain medical events of interest (Yang et al., 2017). In such data, the outcome is discrete and may be over-dispersed. This type of data is referred to as “zero-inflated” data since the data have a higher proportion of zero counts. In longitudinal follow-up data, modeling change over time with an outcome representing a count is challenging. Besides transforming, it is important to apply a relevant distributional form when analyzing data with zero-inflated count outcomes. For example, applying the Poisson distribution to a dataset with many zero observations will become increasingly positively skewed as the mean of the outcome decreases (Grimm & Stegmann, 2019; Yang et al., 2017). In general, zero-inflated Poisson distribution, zero-inflated negative binomial distribution, and zero-altered Poisson distribution (also known as a hurdle model) could be used to model count data with an excess of zero counts (Yang et al., 2017).

In various medical / health studies, it is common to see binary outcomes or count outcomes converted to binary outcomes (Guddattua et al., 2015). For several reasons, researchers prefer binary outcomes. Binary data can offer a simple interpretation or classification, establish eligibility criteria for future studies, and make data summarization more efficient (Williams et al., 2006). On the other hand, using binary data also has some limitations. In the binary data, some information can be ignored, such as individual differences (MacCallum et al., 2002). The loss of information could lead to a loss of power, and maintaining the power might require a larger sample size (Fedorov et al., 2009). Also, when data contains information about relatively

rare conditions, caution should be taken before utilizing the binary data (Ferraro & Wilmoth, 2000). However, the decision on the type of outcome data will depend on the researcher's study objective, and it can be based on prior information and compared through different models. Trajectory specifies evaluating one or more outcomes over age or time (Nagin, 2005), and several statistical approaches are used for analyzing developmental trajectories. Traditionally hierarchical modeling and latent growth curve modeling were most commonly used, but recently the number of studies that applied group-based trajectory modeling has increased (Nagin, 2014; Nagin & Odgers, 2010).

Group-based trajectory modeling is defined as "Finite mixture modeling application that uses trajectory groups as a statistical device for approximating unknown trajectories across population members" (Nagin & Odgers, 2010). Group-based trajectory assumes that the population is composed of distinct groups, each with a different underlying trajectory and every subject in the group approximately follows the same patterns of behavior of outcome over age or time (Nagin, 1999). Group-based trajectory can identify distinctive developmental paths in complex longitudinal data, which can be useful when handling non-monotonic trajectories (Nagin, 2005).

The top ten causes of death surveyed in 2019 accounted for 55 percent of worldwide deaths (WHO, 2020). Of them, the leading causes of death can be categorized broadly into three topics: cardiovascular, respiratory, and neonatal conditions (WHO, 2020). Since the COVID-19 (SARS-CoV-2) outbreak, respiratory diseases have received a lot of attention. Research into the causes of death is essential and can improve lives (WHO, 2020). However, very little work has been done in trajectory analysis for pneumonia, a well-known respiratory-related disease. Therefore, our thesis investigated the trajectory of pneumonia. Also, for pneumonia, age is considered as a risk factor where people at most risk are adults aged more than 65 years, young

children, and infants (Vinogradova et al., 2009). Since pneumonia is a common cause of death increasing in the elderly, we focused on geriatric pneumonia.

In conclusion, this thesis conducted a trajectory analysis for geriatric pneumonia with dichotomous and count outcomes in longitudinal data with group-based trajectory modeling. Also, both models were compared and based on the results, we discussed what type of outcome generates a better fit model.

1.2 Research objectives

This study has four study objectives:

- Objective 1: To develop trajectories with binary and zero-inflated count outcomes using group-based trajectory modeling for geriatric pneumonia.
- Objective 2: To compare the trajectory shape and membership differences in the group-based trajectory model with binary and zero-inflated count outcomes.
- Objective 3: To identify relevant risk factors from the group-based trajectory model with binary and zero-inflated count outcomes.
- Objective 4: To compare the trajectory models with binary and zero-inflated count outcomes and discuss what type of outcome fits better.

CHAPTER 2. LITERATURE REVIEW

2.1 Review of group-based trajectory modeling

Several statistical approaches are available to analyze trajectories. Standard statistical approaches include hierarchical modeling and latent curve analysis (Nagin, 2014). Group-based trajectory modeling (GBTM) is also a trajectory analysis method. It is an application of finite mixture modeling that uses trajectory groups to find sub-group trajectories within a population (Nagin & Odgers, 2010). Group-based trajectory assumes that the population is composed of distinct groups, each with a different underlying trajectory, and each subject in the group is approximately following similar trajectories of an outcome over time (Nagin, 1999). However, the final selected trajectories should not be taken as exact trajectories that individuals follow. Rather, the trajectory should be considered as a powerful tool to discover the heterogeneity in the data and more clearly visualize the change and continuity (Wojciechowski, 2017).

In a trajectory model, the parameters are estimated by maximum likelihood estimation, which means that unbiased estimates can be calculated in the presence of missing data under the assumption of missing at random (Nagin, 1999; Nagin & Odgers, 2010). To perform the maximum likelihood estimation, the Quasi-Newton procedure is used in GBTM. In GBTM, the posterior probability is assigned to each individual. Posterior probability measures an individual's probability of belonging to a particular group, and the individual is assigned to the group with the highest probability (Nagin, 2005).

Three types of longitudinal data are applicable in GBTM: continuous outcomes following the normal distribution, binary outcomes following the binary logistic distribution, and count

outcomes following the Poisson or the zero-inflated Poisson distribution (Jones et al., 2001).

There are several motivations for applying GBTM for trajectory analysis. It shows strength in testing taxonomic theories, capturing the connectedness of behavior over time, and identifying distinctive developmental paths in complex longitudinal datasets (Nagin & Odgers, 2010). In GBTM, it is assumed that there is no variation between individuals in the same group (Nagin & Odgers, 2010). Therefore, when GBTM is applied in the research, the differences between subgroups are only discussed, which can be easier to interpret since it is assumed there are differences within subgroups (Nguefack et al., 2020).

Group-based trajectory modeling was first applied in criminology studies and expanded to other fields, such as psychology, medicine, and clinical research (Nagin & Land, 1993; Nagin & Odgers, 2010). For example, GBTM was used in criminology research to investigate social disorganization and police arrest trajectories (Wong et al., 2021). In medical research, GBTM was applied to identify health status trajectories among outpatients with heart failure (Flint et al., 2017). Also, there were several other disease trajectory studies. There was a study of trajectories of depression and their predictors in a population-based study of Korean older adults (Lim et al., 2020). Trajectories of cardiovascular disease risk were identified (Koohi et al., 2021), and asthma trajectories among Danish children were also studied (Pape et al., 2021). These trajectory studies show the changes over time and the predictors related to the trajectory groups, which will enable clinicians and policy decision-makers to plan their interventions.

2.2 Review of binomial and zero-inflated Poisson distribution

2.2.1 Binomial distribution

A binomial distribution is modeled with repeated binary outcomes, and each binary outcome is referred to as a Bernoulli trial. In Bernoulli trials, there are only two possible outcomes for each trial: 'success' and 'failure'. The probability of having 'success' is p ($0 \leq p \leq 1$), and the probability for 'failure' is $q = 1 - p$. According to Attwood (2000), there are four conditions to apply the binomial distribution: (i) There must be a fixed number of trials; (ii) The trials should be independent, which means that one trial will not affect another trial's results; (iii) The trials must have only two possible outcomes; and (iv) The probabilities will remain the same from trial to trial.

If we assume that random variable Y follows the binomial distribution, the probability will be calculated as:

$$P(Y=k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0, 1, \ldots, n;\ n = 1, 2, \ldots, N \text{ and } 0 \leq p \leq 1$$

where k is the number of successes, n is the number of independent Bernoulli trials, and $\binom{n}{k}$ is the number of different possible patterns of k successes and n-k failures in n trials. For a binomial distribution, the expected number of successes (mean of binomial distribution) and variance would be:

$$\mu = np,\ \sigma^2 = npq$$

As a general rule, if n is sufficiently large, the binomial random variable has a probability distribution that can be approximated by a normal distribution (Pitman, 1993). Also, when n is sufficiently large, and p is small, it can be approximated by a Poisson distribution (Pitman, 1993).

Binomial distribution occurs in various fields such as business, social sciences, and medical research. For example, a binomial distribution was applied in academic and school health issues among children exposed to maternal intimate partner abuse (Kernic et al., 2002) and in studies with patients with aortic stenosis: cardiac complications in non-cardiac surgery (Raymer et al., 1998). Also, binomial distribution was applied in the trajectory modeling of depression among Korean older adults and asthma trajectories among Danish children (Lim et al., 2020; Pape et al., 2021).

2.2.2 Poisson distribution

The Poisson distribution is a possibility distribution that is used to model the number of events occurring in a fixed interval (Triola et al., 2006). It is often used for describing rare events. When the random variable Y follows the Poisson distribution with parameter λ (λ>0), the probability will be calculated as:

$$P(Y=k) = \frac{\lambda^k e^{-\lambda}}{k!} \text{ for } k=0,1,2,\ldots$$

where k is the number of occurrences of an event. The Poisson distribution depends only on one parameter λ, which is both the mean and the variance of the Poisson distribution. In the Poisson distribution, it is assumed that the occurrences will be random and independent of each other (Triola et al., 2006). While a binomial distribution is affected by sample size and probability, the Poisson distribution is only affected by λ. Also, the Poisson distribution has no upper limit for k, while in a binomial distribution, it is up to n (Triola et al., 2006).

However, if the counts have many extreme values, such as extra zeros, the mean would be closer to zero, and the variance would be larger than the mean. Thus, it would violate the

assumption of mean-variance equality in a Poisson distribution. In this case, we encounter an over-dispersion issue (Hu et al., 2011).

The Poisson distribution has been applied in various fields, including medical and health studies. The Poisson distribution was applied to discover the suicides rates in the recovery phase after the earthquake in Japan (Masatsugu, 2020). In Massaro's (2017) study, the Poisson distribution was utilized to classify North Carolina counties by sudden infant death syndrome (SIDS) counts. There was also a study that applied the Poisson distribution in group-based trajectory modeling, which identified the trajectory of arrests (Neil et al., 2021).

2.2.3 Zero-inflated Poisson distribution

Data containing a large number of zeros are easily seen in various population-based medical / health studies. For example, we can consider counting the number of seizure attacks among epilepsy patients. After taking anti-epileptic drugs, about three-fourths of patients become seizure-free, but the rest continue to experience seizure attacks (Tigistu et al., 2018). In this case, we can see that the data will have a higher proportion of zero counts than expected. Data like this are called "zero-inflated." Applying the traditional Poisson model to the zero-inflated data without accounting for the extra zeros, the results could be biased and incorrect (Lambert, 1992). Gupta et al. (1996) showed that more errors occur in small values of the count if we do not adjust the model for extra zeros. Different kinds of models were applied to avoid these problems, such as the zero-inflated Poisson model.

The zero-inflated Poisson model can be explained as a mixture model. The probability equation would be:

$$\Pr(\mathrm{Y}=\mathrm{k}) = \begin{cases} \mathrm{p} + (1-\mathrm{p})\exp(-\lambda), k = 0 \\ \frac{(1-\mathrm{p})\exp(-\lambda)\lambda^{\mathrm{k}}}{\mathrm{k!}}, \ k = 1,2,3 \ldots \end{cases}$$ (Lambert, 1992)

where with probability p ($0 \leq p \leq 1$), a random variable Y equals 0, and with probability (1-p), Y follows the Poisson distribution with mean λ ($\lambda > 0$) (Lambert, 1992). λ is the parameter for the Poisson distribution, and k is the number of occurrences.

Zero observations can have two different origins: structural zero and sampling zero. Structural zero is derived from the zero-component distribution, and the sampling zero is derived from the Poisson distribution, which assumes that it happened by chance (Loeys et al., 2012; Hu et al., 2011).

The zero-inflated Poisson model has been widely used in various fields. Lambert first introduced the zero-inflated Poisson model with an application to defects in manufacturing (Lambert, 1992). Including manufacturing defects from Lambert (Lambert, 1992), we can find the zero-inflated Poisson model applied in econometrics, psychology, agriculture, etc. For example, in epidemiological and medical research, the influence of gender and neighborhood deprivation on alcohol consumption (Matheson et al., 2012), life-course socioeconomic circumstances and multimorbidity among older adults (Tucker-Seeley et al., 2011), finding the number of decayed, missing or filled teeth (Bohning et al., 1999) applied zero-inflated Poisson distribution in their studies. Also, Xia et al (2012) compared different models, including the zero-inflated Poisson model, for count outcomes from HIV risk reduction interventions. The zero-inflated Poisson distribution has been applied in various trajectory studies too. For example, the trajectory of psychotropic agent use and adverse outcomes among older people (Huang et al., 2019) and trajectory patterns of dental caries experience (Broadbent et al., 2008) studies applied zero-inflated Poisson distribution in their group-based trajectory modeling.

2.3 Review of epidemiology of geriatric pneumonia

Pneumonia is a disease that affects the lungs in the forms of acute respiratory infection (WHO, 2020). The main causes of pneumonia are viruses, bacteria, and fungi (WHO, 2020). When someone has pneumonia, the alveoli are filled with pus and fluid, limiting oxygen intake, making it hard to breathe (WHO, 2020). According to the WHO (2020), in 2019, 2.6 million deaths were due to lower respiratory infections, which mostly included pneumonia. It remained the world's most deadly communicable disease, and it was ranked as the fourth leading cause of death worldwide. Pneumonia was included in the top leading causes of death in many countries. In the United States, influenza and pneumonia were ranked as the ninth leading cause of death, with 14.4 deaths per 100,000 population (CDC, 2021; Murphy et al 2021). Similarly, in Canada, influenza and pneumonia were ranked as the eighth leading cause of death, with 12.9 deaths per 100,000 population (Statistics Canada, 2022). Compared to the United States and Canada, South Korea had more deaths per 100,000 population due to pneumonia. In South Korea, pneumonia was the third leading cause of death, and the number of pneumonia deaths increased steadily every year from 14.9 per 100,000 population in 2010 to 43.3 in 2020 (Statistics Korea, 2021).

Pneumonia can be classified by several categories. First, it can be categorized by the place where it was acquired. For elderly patients, common categories of pneumonia are: community-acquired pneumonia (CAP), hospital-acquired pneumonia (HAP), ventilator-associated pneumonia (VAP), and health care-associated pneumonia (HCAP) (Lanks et al., 2019).

When someone gets pneumonia without any hospitalization or exposure to any health care system, it is defined as community-acquired pneumonia (Musher & Thorner, 2014). Many pathogens are associated with CAP, but the most common pathogen for CAP is streptococcus

pneumoniae, accounting for more than 25% of cases of CAP worldwide (Kaysin & Viera, 2016; Musher & Thorner, 2014). Hospital-acquired pneumonia (HAP) is defined as pneumonia that develops 48 hours or more after hospitalization (American Thoracic Society et al., 2005). HAP is one of the most common nosocomial infections, which increases mortality, morbidity, extended length of stay, and excessive cost in hospitalized patients (Kieninger & Lipsett, 2009; Raghavendran et al., 2007). As the hospitalized patients are already concomitant with other diseases, the pathogen for HAP is more complex than CAP. Ventilator-associated pneumonia (VAP) and healthcare-associated pneumonia (HCAP) are a subset of health-associated pneumonia. Ventilator-associated pneumonia occurs in mechanically ventilated patients more than 48-72 hours after endotracheal intubation (American Thoracic Society et al., 2005). Healthcare-associated pneumonia can be classified when pneumonia is associated with health care risk factors such as residing in a nursing home or long-term care facility, prior hospitalization, dialysis, chemotherapy etc. (Micek et al., 2020).

Pathogens can also categorize pneumonia. Broadly it can be classified into four categories: bacterial pneumonia, viral pneumonia, mycoplasma pneumonia, and fungal pneumonia. Bacteria cause bacterial pneumonia, and the most common bacterium that causes pneumonia is Streptococcus pneumoniae (Cunha, 2010). Viral pneumonia is caused by various viruses. For adults, the influenza virus is the most common virus (Cunha, 2010). Mycoplasma pneumonia is caused by mycoplasma pneumonia, because it has some different symptoms and physical signs, it is also referred to as atypical pneumonia (Smith, 2010). Fungal pneumonia is caused by various fungi. Commonly, immunocompromised patients with HIV infection or transplantation patients are subject to fungal pneumonia (Yamada et al., 2003).

Pneumonia is the single largest infectious cause of death in children accounting for 14% of all deaths of children under 5 years old worldwide (WHO, 2021). Pediatric pneumonia and

geriatric pneumonia have different presenting features. Pediatric pneumonia includes remarkable risk factors, such as breastfeeding status, low paternal education, low birth weight, and air pollution (Victora et al., 1994; Gritly et al., 2018; Sutriana et al., 2021; WHO, 2021). Even though pneumonia is a leading cause of death among children, pneumonia is less often fatal for children, which leads to a shorter stay in the hospital (Ziss et al., 2003; American Thoracic Society, 2019). On the other hand, geriatric pneumonia patients have higher mortality and longer lengths of stay (Furman et al., 2021). Furthermore, the WHO (2021) suggested that children can be prevented from pneumonia with simple interventions, and treated with low-cost, low-tech medication and care.

Common symptoms of pneumonia are cough, fever, shortness of breath, heavy sweating, fatigue, chest pain, nausea, vomiting, and diarrhea (WHO, 2020). According to Janssens and Krause (2004), these common symptoms are sometimes absent in older patients. Instead, other symptoms, such as falls or confusion occur. Because of this difference, the diagnosis of pneumonia in the elderly is often delayed, contributing to high morbidity and mortality.

To prevent pneumonia, getting vaccinated, maintaining good hygiene, arranging environmental factors, and having a healthy lifestyle would help (WHO, 2020). However, older adults are not aware of vaccination, and it is generally less accepted than vaccination for children (Janssens & Krause, 2004). Various studies suggested individuals over 65 years should get vaccinated (Furman et al., 2021; Kline et al., 2016; Kaysin & Viera, 2016). Kaysin and Viera (2016) also recommended receiving both influenza and pneumococcal vaccines for pneumonia prevention. There are two kinds of pneumococcal vaccine recommended: the 13-valent pneumococcal conjugate vaccine (PCV13; Prevnar 13) and the 23-valent pneumococcal polysaccharide vaccine (PPSV23; Pneumovax 23) (Kaysin & Viera, 2016).

As the population is aging, and with the advent of new viruses that can lead to pneumonia, morbidity and mortality of pneumonia could increase even more. There are several reasons why the elderly are at risk for pneumonia. First, as we get older, our organ systems (especially the respiratory system and immune system) get old too, and physiological changes occur: a decrease in the elastic recoil of the lung, the compliance of the chest wall, and the strength of respiratory muscles (Janssens & Krausea, 2004). Also, the elderly are more likely to have comorbidities that can increase the risk of pneumonia incidence and severity (Janssens & Krausea, 2004).

According to the UN's report on world population ageing 2020, the proportion of people aged 65 years or over is expected to increase from 9.3% in 2020 to 16.0% percent in 2050 worldwide (United Nations, 2020). We would need to prepare before the disease burden gets larger. Geriatric pneumonia shows different aspects from younger patients (Nierderman et al., 2007). For example, older adults have a lower symptom index but have higher mortality, hospitalization rate, re-hospitalization rate, and longer length of stay (Furman et al., 2021). In Welte's (2011) study, percentage of respiratory and general symptoms by age were presented and compared. Elderly patients had lower percentage in most of the symptoms. Especially the percentage difference was larger in general symptoms when we compared them to 18-44 years old patients. For example, 67% of patients between 18 and 44 years old had body aches, but the percentage was only 30% and 25% for patients in 65-74 years old and patients over 75 years old. Therefore, geriatric pneumonia would require a different approach and management.

2.3.1 Risk factors for geriatric pneumonia

In various studies, the risk factors for pneumonia are: age, smoking status, gender, alcoholism, underlying conditions (diabetes, chronic heart disease, chronic respiratory disease), history of hospitalization for pneumonia, hospitalization at a nursing home, etc.

Age is the most common risk factor for pneumonia that has been stated in various studies (Koivular et al., 1994; Jackson et al., 2004; Kline et al., 2015; Chang, 2010; Vila-Corcoles et al., 2008; Kaysin & Viera, 2016; Yoshikawa & Marrie, 2000; Loeb et al., 2009; Skull et al., 2009). Some studies, which only utilized data from seniors, stated that the risk increases with age (Koivular et al., 1994; Jackson et al., 2004; Vila-Corcoles et al., 2008; Yoshikawa & Marrie, 2000; Loeb et al., 2009; Skull et al., 2009). Another common risk factor for pneumonia is smoking status (Jackson et al., 2004; Jackson et al., 2009; Gau et al., 2010; Loeb et al., 2009). In some studies that utilized hospitalized pneumonia patients, not only current smoker, but also ex-smoker (Gau et al., 2010) and secondhand smoke (Loeb et al., 2009) were significant factors. As more men use tobacco than women (WHO, 2010), some studies stated male sex as a risk factor (Jackson et al., 2004; Vila-Corcoles et al., 2008; Yoshikawa & Marrie, 2000; Skull et al., 2009). In one longitudinal cohort study from Japan, smoking was an important risk factor. However, when they analyzed it by gender, smoking wasn't significant among females (Inoue et al., 2007). Some studies also included alcoholism as a risk factor (Koivular et al 1994; Kaysin & Viera, 2016; Yoshikawa & Marrie, 2000; Skull et al., 2009). Also, comorbidities are common in older adults, so various studies reported many other diseases as risk factors for pneumonia. Respiratory diseases (Koivular et al., 1994; Chang, 2010; Vila-Corcoles et al., 2008; Jackson et al., 2004; Jackson et al., 2009; Kaysin & Viera, 2016; Kline et al., 2015; Yoshikawa & Marrie, 2000; Gau et al., 2010; Loeb et al., 2009; Skull et al., 2009), diabetes (Jackson et al., 2004; Skull et al., 2009), and heart diseases (Koivular et

al., 1994; Vila-Corcoles et al., 2008; Jackson et al., 2009, Gau et al., 2010) were identified multiple times in previous studies. As pneumonia is often acquired in hospitals or health care facilities, history of hospitalization for pneumonia (Jackson et al., 2004; Vila-Corcoles et al., 2008), and hospitalization at a nursing home (Chang, 2010) were also considered as risk factors for geriatric pneumonia. Also, such as living in a nursing home, crowded living conditions was considered a risk factor (Coughlin, 2007; WHO, 2020).

CHAPTER 3. METHODS

3.1 Longitudinal studies

3.1.1 Introduction

Longitudinal data are widely used in health/clinical studies. In a longitudinal study, individuals are measured repeatedly over time (Diggle et al., 2002). Compared to cross-sectional studies, it allows separating between-subject variation and within-subject variation. Thus, researchers can assess multiple aspects from the data, such as the change over time of an outcome and associated risk factors, the timing of disease onset, and individual and group patterns (Garcia & Marder, 2017). However, when analyzing longitudinal data, several challenges can arise. First, as the measurements are taken repeatedly over time, missing data is common in a longitudinal dataset (Fitzmaurice et al., 2012). Also, the measurements could be collected at irregularly spaced visits, making the data unbalanced (Fitzmaurice et al., 2012). Lastly, repeated measurements from the same individual can be correlated (Fitzmaurice et al., 2012). Therefore, special statistical techniques, such as a general linear mixed model and generalized estimating equations (GEE), are required for analyzing longitudinal data (Edwards, 2000).

3.1.2 Analysis of longitudinal data

Let Y_{ij} represent a response variable for the i^{th} individual (i=1,2,…N) at j^{th} observation (j=1,2,…n_i), where N indicates the total number of individuals and n_i indicates the number of

observed responses on the i^{th} individual (Fitzmaurice et al., 2012). The model for changes in the mean response over time and for relating the changes to the covariates would be:

$$\mathrm{Y_{ij}} = \beta_1 \mathrm{X}_{ij1} + \beta_2 \mathrm{X}_{ij2} + \cdots + \beta_\mathrm{p} \mathrm{X}_{ijp} + e_{ij},\ \mathrm{j}=1,2,\ldots,n_i$$

where $\beta_1, \beta_2, \ldots, \beta_\mathrm{p}$ are unknown regression coefficients relating the mean of $\mathrm{Y_{ij}}$ to its corresponding covariates (Fitzmaurice et al., 2012). As we have n_i repeated measurements of the response variable on the same individual i, n_i x 1 response vector $\mathbf{Y}_\mathrm{i}$ is denoted as (Fitzmaurice et al., 2012):

$$\mathbf{Y}_\mathrm{i} = \begin{pmatrix} \mathrm{Y}_{i1} \\ \mathrm{Y}_{i2} \\ \ldots \\ \mathrm{Y}_{in_i} \end{pmatrix},\ i=1,2,\ldots,\mathrm{N}$$

where the N individuals are assumed to be independent.

In many longitudinal studies, the main interest is focused on the changes in the mean response over time and how these changes relate to the covariates (Fitzmaurice et al., 2012). For covariates, let $\mathbf{X}_\mathrm{ij}$ as a p x 1 vector of covariates, which are associated with the response variable $\mathrm{Y_{ij}}$ (Fitzmaurice et al., 2012).

$$\mathbf{X}_\mathrm{ij} = \begin{pmatrix} \mathrm{X}_{ij1} \\ \mathrm{X}_{ij2} \\ \ldots \\ \mathrm{X}_{ijp} \end{pmatrix},\ i=1,2,\ldots,\mathrm{N};\ j=1,2,\ldots,n_i$$

Every p row of $\mathbf{X}_\mathrm{ij}$ corresponds to different covariates, and the covariates could be time-dependent or independent (Fitzmaurice et al., 2012). The covariates' vectors could also be grouped into an n_i x p matrix of covariates (Fitzmaurice et al., 2012).

$$\mathrm{X_i} = \begin{pmatrix} \mathbf{X}'_{i1} \\ \mathbf{X}'_{i2} \\ \ldots \\ \mathbf{X}'_{in_i} \end{pmatrix} = \begin{pmatrix} \mathrm{X_{i11}} & \cdots & \mathrm{X_{i1p}} \\ \vdots & \ddots & \vdots \\ \mathrm{X_{in_i1}} & \cdots & \mathrm{X_{in_ip}} \end{pmatrix},\ i=1,2,\ldots,\mathrm{N}$$

where $\mathbf{X}'_{ij}$ is a 1 x p row vector of covariates for the i^{th} individual at the j^{th} observation (Fitzmaurice et al., 2012). Lastly, n_i x 1 vector of random errors would be:

$$\boldsymbol{\varepsilon}_{\text{ij}} = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \dots \\ e_{in_i} \end{pmatrix}, \text{i=1,2,…,N}$$

Other than creating summary measures of the response variables and covariates, there are several other approaches to analyze longitudinal data. Diggle et al. (2002) suggested three methods to analyze the longitudinal data, considering the correlation in individual responses.

The first approach is to fit a marginal model. The marginal model focuses on the mean response over the population and its dependence on the covariates rather than the random effects or previous responses (Fitzmaurice et al., 2012). The advantage of marginal analysis is that we can model the mean and covariance separately (Diggle et al., 2002). A generalized Estimating Equation (GEE) is a method to estimate the parameters in a marginal model that is commonly used for longitudinal data analysis.

The second approach is the random effects model. Since each individual has distinct regression coefficients in the random effects model, it assumes that correlation arises among repeated responses. (Diggle et al., 2002). However, not enough information is provided to estimate the regression coefficient in the random effects model (Diggle et al., 2002). Thus, the random effects are assumed to be independent realizations from some distribution (Diggle et al., 2002). Random effects models are useful when we are more focused on the individuals than the population average (Diggle et al., 2002).

The third approach is the transition model. The transition model focuses on modeling the response conditional on the past response and covariates (Diggle et al., 2002). The markov model is commonly used for this approach.

In addition to these approaches, a trajectory approach is now arising as another approach analyzing longitudinal data (Jones, 2001). This thesis will be focused on the trajectory approach, specifically on group-based trajectory modeling, which is one of the methods of the trajectory approach.

3.2 Finite mixture modeling

3.2.1 Introduction

The statistical foundation of group-based trajectory modeling is in finite mixture modeling (Salonen, 2020). A finite mixture model is often used to investigate and model the heterogeneity in the data. In finite mixture modeling, it is assumed that the observed distribution is a mixture of different distributions (Salonen, 2020). As it provides convenient representations for modeling complex distributions of data, it is widely used in various fields such as biology, agriculture, astronomy, engineering, and many other fields (McLachlan et al., 2019).

3.2.2 Definition

Let Y_n be a random variable of size n where y_i is an observed outcome of the random variable Y_n. Then the random variable Y_n follows a J-component finite mixture distribution with the density function as (McLachlan & Peel, 2000):

$$f(y_i; \boldsymbol{\Psi}) = \sum_{j=1}^{J} \pi_j \, f_j(y_i; \boldsymbol{\theta_j})$$

where π_j meets the assumption of:

$$0 \leq \pi_j \leq 1,$$

$$\sum_{j=1}^{J} \pi_j = 1$$

Here $\boldsymbol{\Psi}$ is the vector containing all the unknown parameters that are needed in the mixture model (McLachlan & Peel, 2000).

$$\boldsymbol{\Psi} = (\pi_1, \pi_2, \ldots, \pi_{J-1}, \xi^T)^T$$

where T is defined as the vector transpose and ξ is the vector that contains all the parameters in $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_J}$.

In the equation, J is the number of mixture components, π_j is the mixing probability for the j^{th} component, f(.) is the component specific density function, and $\boldsymbol{\theta_j}$ refers to the vector of unknown parameters for the j^{th} component density (McLachlan & Peel, 2000).

3.2.3 Likelihood function

Maximum likelihood estimation is the most common method of fitting the mixture distributions in finite mixture modeling (McLachlan et al., 2019). If we suppose that identically distributed y_i are coming from the mixture distribution, the likelihood function would be (McLachlan & Peel, 2000):

$$L(\boldsymbol{\Psi}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\Psi})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{J} \pi_j \, f_j(y_i; \boldsymbol{\theta_j})$$

We can also obtain the maximum likelihood estimation by maximizing the log-likelihood with respect to $\boldsymbol{\Psi}$. The log-likelihood would be as:

$$l(\boldsymbol{\Psi}) = \log\ \mathrm{L}(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{j=1}^{J} \pi_j\ \mathrm{f_j}(\mathrm{y_i}; \boldsymbol{\theta_j})$$

There is no closed-form solution for maximum likelihood estimates. One of the most common methods used to derive the maximum likelihood estimation is the Expectation-Maximization (EM) algorithm (McLachlan & Peel, 2000).

3.3 Group-based trajectory modeling

Group-based trajectory modeling is an application of finite mixture modeling. While the analysis aims to find sub-group trajectories within a population, the estimated parameters are not derived from cluster analysis but depend on maximum likelihood estimation (Nagin, 2005).

Let $\mathrm{Y_i} = \{\mathrm{y_{i1}}, y_{i2}, \ldots\ldots, y_{it}\}$ denote the longitudinal trajectory data for subject i over t's measurement, t = 1,2,…,T. $P(Y_i)$ represents the probability of Y_i, and the trajectory group is indexed by j, j = 1,2,…,J. If we assume there are J groups of trajectories from the population, the group-based trajectory modeling would be (Nagin, 2005):

$$P(Y_i) = \sum_{j=1}^{J} \pi_j\ P^j\ (Y_i)$$

where π_j represents the probability of a randomly chosen population member belonging to group j, and $P^j(Y_i)$ represents the probability of Y_i given membership in group j. π_j should meet the two assumptions of (Jones & Nagin, 2007; Nagin, 2005):

$$0 \leq \pi_j\ \leq 1,$$

$$\sum_{j=1}^{J} \pi_j\ = 1$$

For given j, measurement for the ith subject at time t is assumed to be independent over the T periods, where T is the maximum number of measures, t = 1, 2,. . . ,T (Nagin, 2005). Thus,

$$P^j(Y_i) = \prod_{t=1}^{T} p^j(y_{it})$$

where $p^j(y_{it})$ is the probability distribution function of y_{it} given the ith subject at time t for group j (Nagin, 2005).

For binary data, $P(Y_i)$ follows the binary logistic distribution. For count data, $P(Y_i)$ is specified as the Poisson distribution or the zero-inflated Poisson distribution, and for censored data, $P(Y_i)$ is specified as a censored normal distribution (Nagin, 2005). To estimate the parameters in the model, maximum likelihood is used, and the likelihood function for the entire sample of N individuals would be (Nagin, 2005):

$$L = \prod_{i=1}^{N} P(Y_i)$$

The effect of adding time-stable covariates and time-dependent covariates to the model are described in Figure 3.1. If we consider time-stable covariates into the model, which would be risk factors, it will affect the likelihood of a particular data trajectory (Jones et al., 2001). Meanwhile, including time-dependent covariates can directly influence the trajectory paths (Jones, 2001). Overall, the observed trajectory depends on group membership and the time-dependent covariates, but the risk factors and the observed trajectory are independent (Jones et al., 2001). (Figure 3.1)

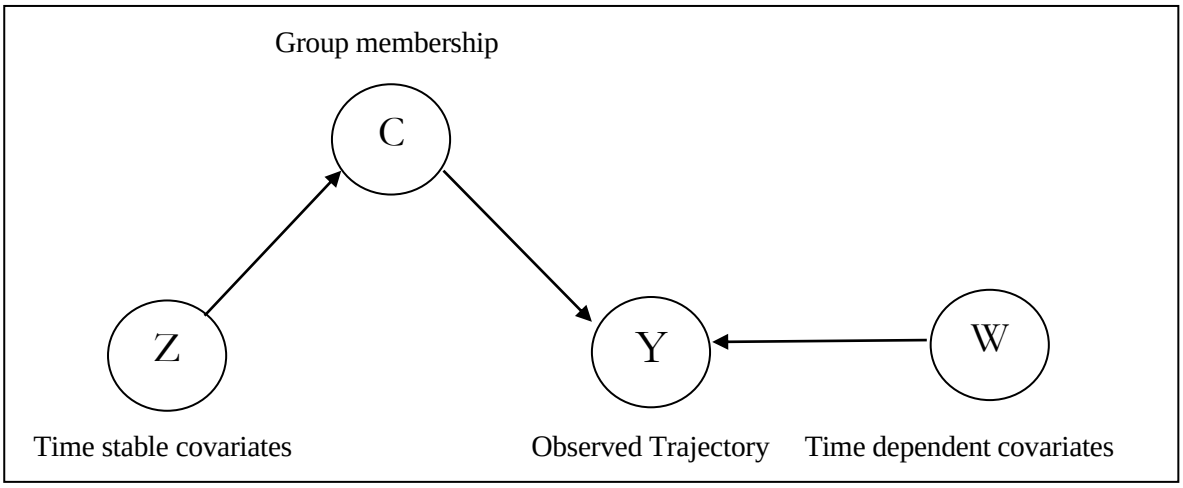


Figure 3.1 Directed Acyclic Graph representing the independence assumptions (Jones et al., 2001)

3.3.1 Group membership probabilities

The probability of membership in group j is stated as π_j. It measures the proportion of the population that belongs to group j (Nagin, 2005). The value should be between 0 and 1, and it is not estimated directly, but it is estimated by a multinomial logit function which is linked to a set of parameters θ_j, $j = 1,2,\dots,J$, and the calculation would be:

$$\pi_j = \frac{e^{\theta_j}}{\sum_1^J e^{\theta_j}}$$

where θ_j can take on any value without violating the two assumptions of π_j (Jones & Nagin, 2007; Nagin, 2005).

$$0 \leq \pi_j \leq 1,$$

$$\sum_{j=1}^{J} \pi_j = 1$$

If we consider risk factors into the model, the calculation of π_j would be different. It would be calculated as:

$$\pi_j = P_r(\mathrm{C} = \mathrm{j} \mid \boldsymbol{X} = \mathbf{x}) = \frac{\exp(\theta_j + \mathbf{w_j'}\boldsymbol{x})}{\sum_1^J \exp(\theta_j + \mathbf{w_j'}\boldsymbol{x})}$$

where x is a risk factor, and $\mathbf{X}=\{x_1, x_1, \ldots, x_r\}$ is a vector of random variables for risk factor x. $\boldsymbol{w_j}$ is a vector of parameters that represents the coefficients of the risk factors (Jones et al., 2001).

3.3.2 Posterior group membership probabilities

A posterior group membership probability measures the individual's likelihood of belonging to each trajectory group (Nagin, 2005). It is different from the group membership probability π_j, which measures the size of each trajectory group (Nagin, 2005). Thus, π_j is not focused on the individual, and it is used to aggregate the size of each trajectory group. Meanwhile, the posterior probability assigns the individual to a specific trajectory group that best matches their features (Nagin, 2005).

The posterior probability of individual *i*'s membership in group j is denoted as (Nagin, 2005):

$$\widehat{\mathrm{P}}(\mathrm{j}|\mathrm{Y_i}) = \frac{\widehat{\mathrm{P}}(\mathrm{Y_i}|\mathrm{j})\hat{\pi}_j}{\sum_1^J \widehat{\mathrm{P}}(\mathrm{Y_i}|\mathrm{j})\hat{\pi}_j}$$

where $\widehat{\mathrm{P}}(\mathrm{Y_i}|\mathrm{j})$ is the probability of $\mathrm{Y_i}$ conditional on membership in group j, and $\hat{\pi}_j$ is the estimated group membership probability in group j (Nagin, 2005).

A maximum-probability assignment rule is applied to the posterior probability, so each individual is assigned to the trajectory group in which their posterior membership probability is the largest (Nagin, 2005). Kass and Wasserman (1995) suggested selecting the model with a larger posterior probability, and Nagin (2005) suggested applying it to determine the model fit adequacy.

3.3.3 Group-based trajectory modeling for continuous outcomes

Adaption of the general model to the data requires two assumptions: (i) an appropriate distributional form for the outcome, (ii) specification of a link function (Nagin, 2005). The censored normal model is useful when conducting the group-based trajectory modeling with continuous outcomes. For example, the censored normal distribution can be applied to psychometric scale data (Nagin, 1999). The censored normal distribution allows the data to cluster at the minimum of the scale and at the scale maximum (Jones et al., 2001). Also, it is appropriate for continuous data that are approximately normally distributed, with or without censoring (Jones et al., 2001). Let S_{min} and S_{max} be the minimum and maximum possible scores on the measurement scale. Then the model assumes:

$$y_{it} = S_{min} \quad \text{if} \quad y_{it}^{*} < S_{min}$$

$$y_{it} = y_{it}^{*} \quad \text{if} \quad S_{min} < y_{it}^{*} < S_{max}$$

$$y_{it} = S_{max} \quad \text{if} \quad y_{it}^{*} > S_{min}$$

where y_{it}^{*} is the latent variable. Therefore, the equation would be composed of three parts:

$$P_r(Y = Y_i|C = j) =$$

$$\prod_{y_{it}=S_{min}} \Phi\left(\frac{S_{min} - \mu_{itj}}{\sigma}\right) \prod_{S_{min}<y_{it}<S_{max}} \frac{1}{\sigma}\phi\left(\frac{y_{it} - \mu_{itj}}{\sigma}\right) \prod_{y_{it}=S_{max}} [1 - \Phi\left(\frac{S_{max} - \mu_{itj}}{\sigma}\right)]$$

where Φ is the cumulative distribution function, and ϕ is the probability density function of a normal random variable with mean μ_{itj} and standard deviation σ (Nagin, 2005). C is defined as the unobserved discrete variable indicating the latent class of the ith subject, C = 1, 2, . . . , J (Nagin, 1999).

If we consider that the data is normally distributed, which means there is no scale minimum or maximum, or all the data lies inside the range, the probability distribution function of y_{it} given membership in group j at time t would be as:

$$P_r(Y = Y_i|\mathrm{C} = \mathrm{j}) = \prod_{y_{it}} \frac{1}{\sigma}\phi(\frac{y_{it} - \mu_{itj}}{\sigma})$$

In addition, the link between time and the models' parameters is modeled as a polynomial relationship (Jones et al., 2001). The link function for the censored normal distribution would be:

$$\mu_{itj} = \beta_{0t} + Time_{it}\beta_{1t} + Time_{it}^2\beta_{2t} + Time_{it}^3\beta_{3t} + \varepsilon_{\mathrm{it}}$$

where β's are the parameters of time and $\varepsilon_{\mathrm{it}}$ is a disturbance assumed to be normally distributed with a zero mean and a constant standard deviation σ (Nagin, 2005). Considering the application, the model allows for up to a cubic relationship (Nagin, 2005). The parameters of time would determine the shape of the polynomial function (Nagin, 2005).

3.3.4 Group-based trajectory modeling for binary outcomes

Binary outcomes are common in longitudinal data. For example, each individual either had or had not pneumonia in a certain period. Therefore, the outcome in this case would be recorded binomial as $Y_i = 0$ if no, $Y_i = 1$ if yes.

As y_{it} is assumed to be a binary outcome, $p^j(y_{it})$ will follow the binary logit distribution (Jones et al., 2001), for the ith subject in group j at time t. For the binary logistic model, the probability of observing the trajectory for subject i, given that subject belongs to group j, the equation would be:

$$P_r(Y = Y_i|\mathrm{C} = \mathrm{j}) = \prod_{y_{it}=1} \rho_{itj} \prod_{y_{it}=0} (1 - \rho_{itj})$$

where ρ_{itj} denotes the probability when y_{it}=1 for ith subject in group j at time t. The link function for the binary logistic distribution would be:

$$\rho_{itj} = \frac{\exp\left(\beta_{0j} + Time_{it}\beta_{1j} + Time_{it}^2\beta_{2j} + Time_{it}^3\beta_{3j}\right)}{1 + \exp\left(\beta_{0j} + Time_{it}\beta_{1j} + Time_{it}^2\beta_{2j} + Time_{it}^3\beta_{3j}\right)}$$

3.3.5 Group-based trajectory modeling for count outcomes

The most common probability distribution for count outcomes is the Poisson distribution. For a Poisson distribution, the probability distribution function of y_{it} given membership in group j would be as:

$$P_r(Y = Y_i|\mathrm{C} = \mathrm{j}) = \frac{\exp(-\lambda_{jit})\,\lambda_{jit}^{y_{it}}}{y_{it}!}$$

where λ_{jit} ($\lambda_{jit} > 0$) is a parameter measuring the mean rate of occurrence of the event for the ith subject in the group j at time t. C is defined as the unobserved discrete variable indicating the latent class of the ith subject, C = 1,2,…,J. When λ_{jit} increases, the Poisson distribution will resemble the normal distribution. Therefore, if the mean rate is large enough, the results based on the Poisson distribution and normal distribution will be similar (Nagin, 2005).

However, if there are too many zeros in the data, the results based on the Poisson distribution can be biased (Lambert, 1992). In this case, we can use the zero-inflated Poisson distribution in group-based trajectory modeling. The equation for the zero-inflated Poisson distribution would be divided into two parts where y_{it}=0 and y_{it}>0.

$$P_r(Y = Y_i|\mathrm{C} = \mathrm{j}) =$$

$$\prod_{y_{it}=0}[\rho_{itj}+(1-\rho_{itj})\exp(-\lambda_{jit})]\prod_{y_{it}>0}(1-\rho_{itj})\frac{\exp(-\lambda_{jit})*\lambda_{jit}^{y_{it}}}{y_{it}!}$$

In the equation, ρ_{itj} denotes the probability that the ith subject in the group j at time t will have zero counts. λ_{jit} ($\lambda_{jit}>0$) is the Poisson distribution parameter, which is the expected number of occurrences of the event of subject i at time t, given membership in group j (Nagin, 2005). Likewise, the Poisson and the zero-inflated Poisson model also require the specification of the link function that connects the parameters with time. The link function would be:

$$\log(\lambda_{jit})=\beta_{0j}+Time_{it}\beta_{1j}+Time_{it}^{2}\beta_{2j}+Time_{it}^{3}\beta_{3j}$$

3.3.6 Parameter estimation

The parameters of group-based trajectory models are estimated by maximum likelihood estimation. In most cases for finite mixture modeling, the Expectation-Maximization (EM) algorithm is used to compute the maximum likelihood estimates (McLachlan & Peel, 2000). Several studies compare the maximum likelihood estimation in group-based trajectory modeling among different methods. In Nawa's (2014) study, the parameters were estimated and compared using the EM algorithm and the Quasi-Newton method. The results were mostly similar but had some differences. The Quasi-Newton method was highly dependent on starting values, and the EM algorithm showed a slow convergence rate (Nawa, 2014). However, the EM algorithm was preferred. Chu and Koval (2014), considered the Quasi-Newton method and three EM based algorithms: standard EM, EM algorithm with the iteratively reweighted least squares (IRLS) method at maximization stage, and EM algorithm with the Quasi-Newton method at the maximization stage. Similarly, they recommended the EM-IRLS algorithm to reduce convergence problems and result in precise estimations (Chu & Koval, 2014). However,

in this thesis the Quasi-Newton method was applied, since the package running in SAS 9.4 for group-based trajectory modeling applies the Quasi-Newton method.

If we let $\beta^j Time_{it} = \beta_{0j} + Time_{it}\beta_{1j} + Time_{it}^2\beta_{2j} + Time_{it}^3\beta_{3j}$. The likelihood function that should be maximized based on Quasi-Newton method for binary outcomes would be:

$$L(\boldsymbol{\Psi}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \sum_{j=1}^{J} \pi_j \prod_{\mathrm{t}=1}^{\mathrm{T}} \left(\frac{e^{\beta^j Time_{it}}}{1 + e^{\beta^j Time_{it}}} \right)^{y_{it}} \left(\frac{1}{1 + e^{\beta^j Time_{it}}} \right)^{1-y_{it}}$$

where T is the maximum number of measures, t = 1, 2,. . . ,T.

For the continuous outcome that follows the censored normal distribution would be:

$$L(\boldsymbol{\Psi}) = \frac{1}{\sigma} \prod_{\mathrm{i}=1}^{\mathrm{n}} \sum_{j=1}^{J} \pi_j \prod_{\mathrm{t}=1}^{\mathrm{T}} \phi \left(\frac{y_{it} - \beta^j Time_{it}}{\sigma} \right)$$

For count data, the log-likelihood function would be used. First, for the count outcome which follows the Poisson distribution, the equation would be:

$$logL(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{j=1}^{J} \pi_j \prod_{t=1}^{\mathrm{T}} \frac{\mathrm{e}^{-\beta^j Time_{it}} \left(\beta^j Time_{it}\right)^{y_{it}}}{y_{it}!}$$

The count data which follows the zero-inflated Poisson distribution's log-likelihood function would be as:

$$logL(\boldsymbol{\Psi}) =$$

$$\sum_{i=1}^{n} \log \sum_{j=1}^{J} \pi_j \prod_{t=1}^{\mathrm{T}} \left\{ \rho_{itj} + (1 - \rho_{itj}) \mathrm{e}^{-\beta^j Time_{it}} \right\}^{a_{it}} \left\{ (1 - \rho_{itj}) \frac{\mathrm{e}^{-\beta^j Time_{it}} \left(\beta^j Time_{it}\right)^{y_{it}}}{y_{it}!} \right\}^{1-a_{it}}$$

where ρ_{itj} denotes the probability that the ith subject in the group j at time t will have zero counts and $a_{it} = \begin{cases} 1, & \text{if } y_{it} = 0 \\ 0, & \text{if } y_{it} > 0 \end{cases}$

3.3.7 Model selection

The model selection follows the process suggested by Nagin (Nagin, 2005). First, the numbers of trajectory groups are determined, and then the best polynomial trajectory function is chosen for the shape of the trajectories. The maximum number of groups is based on prior knowledge, for which we consider the size of the population or any existing evidence (Walsh et al., 2020). Constant, linear, quadratic, or cubic are considered to decide the polynomial trajectory function.

The model selection decision is mainly based on Bayesian Information Criteria (BIC). Also, Akaike information criterion (AIC) and probability of a correct model are considered when selecting the groups and the shape. BIC and AIC are calculated as in equation:

$$BIC = \log(L) - 0.5k \log(N)$$

$$AIC = \log(L) - 0.5k$$

where L is the value of the model's maximum likelihood, N is the sample size, and k denotes the number of parameters in the model. AIC is similar to BIC, but the difference is that AIC does not vary with sample size (Nagin, 2005). If we look at the first component of BIC and AIC, the logarithm of the likelihood will always increase as more groups are added to the model. The second part of the equation would be the counterbalance for the first component. Additional groups mean that the number of parameters will also increase. For BIC, a large sample size will decrease the value too. The model with the largest value is recommended as the final model (Nagin, 2005).

The Bayes factor can be used to compare the magnitude of change in the BIC between two models (Nest et al., 2020; Kass & Raftery, 1995). The Bayes factor is denoted as B_{ij}, and it measures the posterior odds of model *i* being the correct model compared to the model *j* being

the correct model (Nagin, 2005). Thus, if the Bayes factor is larger than 1, it implies that model *i* is more likely to be correct. Jeffrey's scale is considered for a more detailed interpretation for the Bayes factor (Wasserman, 2000). Jeffrey's scale is shown in Table 3.1.

Table 3.1 Jeffrey's scale of evidence for Bayes factors

Bayes factors	**Interpretation**
$B_{ij} < 1/10$	Strong evidence for model *j*
$1/10 < B_{ij} < 1/3$	Moderate evidence with model *j*
$1/3 < B_{ij} < 1$	Weak evidence with model *j*
$1 < B_{ij} < 3$	Weak evidence with model *i*
$3 < B_{ij} < 10$	Moderate evidence with model *i*
$B_{ij} > 10$	Strong evidence with model *i*

However, the computation of the Bayes factor is difficult. Thus, Schwarz (1978) and Kass and Wasserman (1995) suggested to compute the Bayes factor as $e^{BIC_i - BIC_j}$. Also, they suggested comparing the probability p_j, where p_j denotes the probability that a model with j groups is the correct model from a set of J different models. The probability p_j would be calculated as:

$$p_j = \frac{e^{BIC_j - BIC_{max}}}{\sum_{j=1}^{J} e^{BIC_j - BIC_{max}}}$$

where BIC_{max} is the maximum score of different J models. The model with the largest p_j will be considered the best model (Kass & Wasserman, 1995).

Similarly, Jones et al. (2001) suggested another criterion using a logged Bayes factor that measures the strength of evidence against the null model. The logged Bayes factor is calculated

as $2(BIC_{complex} - BIC_{null})$. The interpretation of the logged Bayes factor is shown in Table 3.2 (Jones et al., 2001).

Table 3.2 Interpretation of logged Bayes factor

Logged Bayes factor	**Evidence against the null model**
0 to 2	Not worth mentioning
2 to 6	Positive
6 to 10	Strong
>10	Very strong

We should consider various aspects and compare the models to determine the best trajectory model. Nagin stated the model selection as "one of the key decision points in group-based trajectory modeling" (Nagin & Odgers, 2010). However, he also said that there is no correct model. Selecting a model is not about maximizing the statistic fit of the model but choosing the model that summarizes the data's distinctive features (Nagin, 2005).

3.3.8 Software and level of significance

Statistical software packages are currently available in SAS, R, and Stata for conducting group-based trajectory modeling. All the analyses in this thesis were completed using SAS 9.4 (SAS Institute, Cary, NC). GBTM was performed with PROC TRAJ, a macro package running under SAS 9.4 (Jones, 2020). For this thesis, $\alpha = 0.05$ was selected for the significance level unless stated otherwise.

CHAPTER 4. APPLICATION

4.1 Introduction

In this chapter, we applied group-based trajectory modeling (GBTM) to a real dataset to identify groups with pneumonia. It was applied to both binary and count outcomes independently. First, we considered whether the individual had visited the hospital or not due to pneumonia as a binary outcome, and the number of hospital visits was considered as a count outcome. Also, the trajectory shape and membership differences in the models were reported with relevant risk factors that may influence the trajectory groups. The dataset of this study was approved by the Behavioural Research Ethics Board, University of Saskatchewan (ID: 1759).

4.2 Study data

4.2.1 Study design

This research utilizes a subset of ten-year longitudinal survey data from the Korea Health Panel Study (KHPS). The data were collected from 2008 to 2017 that mainly cover public health care services. The KHPS aims to establish panel data that provides information on medical use and medical expenditure and helps to analyze factors affecting medical use and medical expenditure (KHPS, 2021). The KHPS used a stratified sampling frame taken from the 2005 Korean Population and Housing Census. After the data was adjusted for unequal selection probabilities and non-responses, the sample weights for the data were calculated. They also went through the process of making a population distribution disclosure via post-stratification

corresponding to the sample distribution (Lim et al., 2020). Data were collected using computer-assisted personal interviews, and trained staff conducted the survey that was divided into households, individuals, and case-based sections by subdividing the survey areas (Cheng, 2021).

4.2.2 Study population

The data were first collected in 2008 and incorporated 28,970 individuals. However, as the dropouts increased, 5,424 additional people were recruited in 2014. The additional subjects in 2014 were included based on dropout households/members with the same sampling frame and sampling weight as in 2008 to secure statistical reliability.

The survey's questions were based on 13 essential sectors, which include household and household member information, health insurance data, chronic disease data, medical service use data, drug use and medical expenditure data, long-term care data for adult household members, and emergency medical use data, etc. (Lim et al., 2020). From 2008 to 2011, the annual data disease (diagnosis) code was used to record the medical data, and Korean standard disease sign classifications (KSCD) were used to record the medical data from 2012. For this thesis, participants with ages below 65 from the baseline were excluded from the data. A total of 4,007 individuals met our study criteria and were used for the trajectory analysis.

Figure 4.1 describes the study flow diagram of the outcome data for trajectory analysis. The dataset is composed of two parts: 2,946 participants aged 65 or older started in 2008 and continued for ten years, and 1,077 additional participants continued for four years from 2014. Participants added in 2014 were moved to the year 2008, which is considered baseline. As they were moved to the baseline, they had a total of four measurements, and the other six measurements were considered missing.

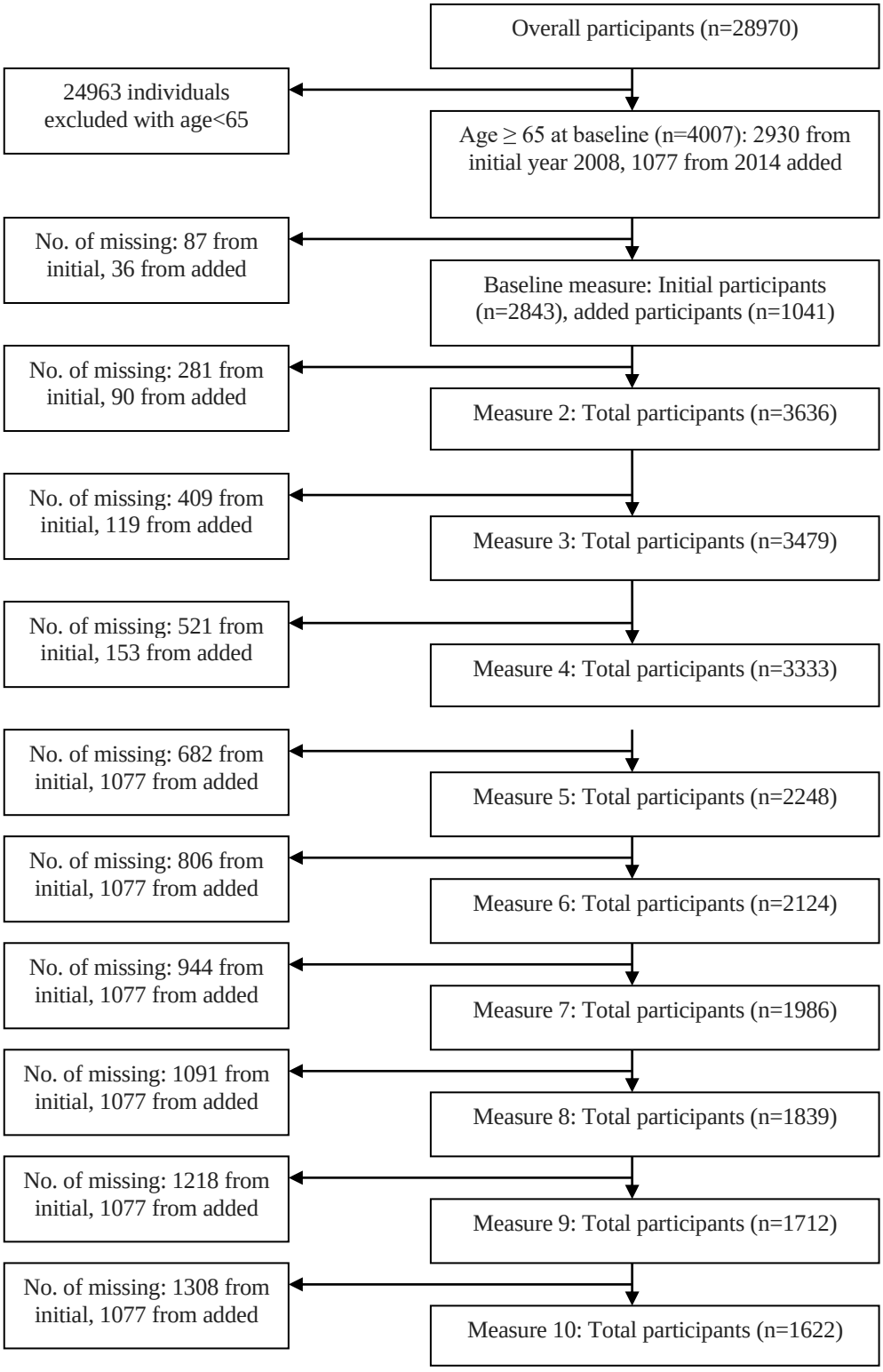


Figure 4.1 Study flow diagram

4.3 Variables

4.3.1 Outcome Variable

Pneumonia is the main outcome variable for this analysis. To determine if the patient is clinically diagnosed with pneumonia, we used inpatient, outpatient, and emergency room records for each patient from KHPS. For the binary outcome, respondents who had inpatient, outpatient, or emergency room pneumonia records in any given year were recorded as 1=yes, and people who didn't have any record of pneumonia were recorded as 0=no. The frequency of their visit each year due to pneumonia was counted for count outcome.

4.3.2 Covariates

The following characteristics were considered as baseline covariates: gender, age, level of education, number of household members, housing type, household income percentile, disability, economic activity, baseline comorbidities (such as chronic respiratory disease, chronic heart disease, diabetes), the presence of more than three chronic diseases, and self-reported behaviors (such as alcohol intake, smoking, and physical activity).

Gender was coded 0 = female and 1 = male. Age was categorized as 65–69, 70–74, 75–79, and 80 years and older. Education was coded as 0 = none, 1 = elementary school, 2 = middle / high school, and 3 = university or higher. The number of household members was grouped as 1, 2, 3, 4, and more than 5 people. Housing type was categorized as 1 = detached house, 2 = multi-unit and townhouse, 3 = apartment, and 4 = other types of houses. The household income percentile was divided into five categories by every 20%. Economic activity was coded as 0 = no, and 1 = yes. Disability was recorded as 1 = yes and 0 = no, according to the official

disability record data. Smoking was coded as 0 = never smoked, 1 = current smoker, and 2 = former smoker. The drinking variable was scored on an 8-point Likert scale that asked how often they drank over the past year. Based on their answers, it was re-categorized as 0 = never, 1 = didn't drink for past 1 year, 2 = less than twice/week, 3 = 2-3 times/week, and 4 = almost daily. Exercise and walking were also scored separately on an 8-point Likert scale that asked respondents how many days they did moderate physical activity or walked more than 10 minutes a day during the past week. Responses ranged from 0 to 7 days a week. However, in this thesis, exercise and walking variables were categorized as 0 = none, 1 = ≤ 3 days/week, and 2 = >3 days/week. The number of chronic diseases was coded as 1 = having 3 or more of these chronic diseases, and 0 = otherwise. In this thesis, chronic diseases included hypertension, heart disease, diabetes, asthma, and all kinds of diseases that can impact their daily functioning. For baseline comorbidities, diabetes, chronic heart disease, and chronic respiratory disease were selected. Chronic heart disease includes myocardial infarction, ischemic heart disease, angina, pulmonary embolism, arrhythmia, conduction disorder, heart failure, heart valve syndrome, mitral stenosis, and other heart diseases. Chronic respiratory disease includes chronic obstructive pulmonary disease, bronchitis, asthma, pulmonary edema, and any other disease classified as 'disease of respiratory system' in KHPS data. Baseline comorbidities were coded as 1 = yes if they had any one of the diseases, and 0 = no according to their chronic disease record. Participants who weren't confirmed about their chronic disease were not recorded in the chronic disease record. All the baseline covariates were measured in 2008.

4.4 Statistical Analysis

4.4.1 Descriptive analysis

Table 4.1 provides the binary outcome of pneumonia by year, and Table 4.2 provides the count outcome, which is the frequency of the hospital visit due to pneumonia by year. Both tables only included individuals aged 65 or older, and the outcomes were calculated by year. In 2008, the pneumonia incidence rate was 0.9%, and in other years it was above 1%, with a maximum of 1.9% (Table 4.1). When we compared the pneumonia incidence rate in the geriatric population to the overall population, the data showed that individuals aged 65 or older have a higher percentage of being diagnosed with pneumonia than the overall participants (Figure 4.2). This reflects the fact that older age is an increasing risk for pneumonia.

Table 4.1 Binary outcome of geriatric pneumonia from 2008 to 2017, N(%)

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
Yes	26 (0.9)	39 (1.5)	29 (1.1)	38 (1.6)	36 (1.6)	31 (1.5)	52 (1.7)	52 (1.8)	47 (1.8)	49 (1.9)	399 (1.5)
No	2817 (99.1)	2610 (98.5)	2492 (98.9)	2371 (98.4)	2212 (98.4)	2093 (98.5)	2975 (98.3)	2774 (98.2)	2623 (98.2)	2497 (98.1)	25464 (98.5)
Total	2843	2649	2521	2409	2248	2124	3027	2826	2670	2546	25863

The majority of the participants didn't visit the hospital due to pneumonia, so we can assume that the count outcome data is zero-inflated. Of the participants who visited the hospital, 117 patients visited only once, and 100 patients visited twice. Only eight patients visited the hospital more than 21 times, and the patient who visited the most had 41 visits due to pneumonia in one year (Table 4.2).

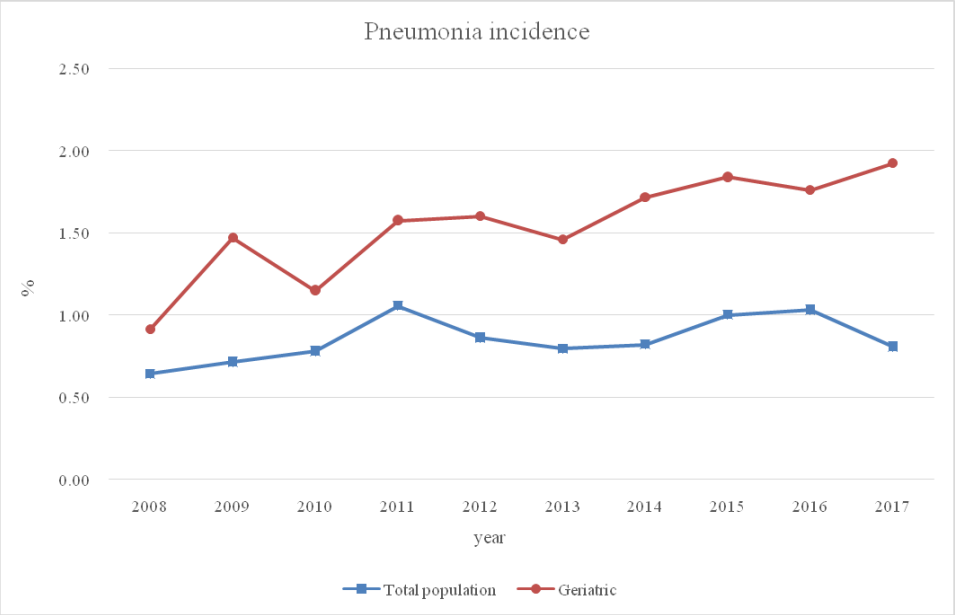


Figure 4.2 Pneumonia incidence from 2008 to 2017

Table 4.2 Frequency of hospital visits due to pneumonia from 2008 to 2017 (N)

year / visit	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
0	2817	2610	2492	2371	2212	2093	2975	2774	2623	2497	25464
1	6	8	6	14	8	8	20	15	13	19	117
2	5	13	9	6	11	14	11	14	12	5	100
3	5	7	1	8	3	2	5	7	6	8	52
4	1	3	6	1	5	2	7	5	7	5	42
5	5	3	1	0	3	1	2	3	4	4	26
6	1	0	2	4	2	0	2	2	0	3	16
7	0	2	1	1	1	1	1	1	2	0	10
8-10	1	1	1	2	1	2	1	2	1	2	14
11-20	2	2	2	1	2	0	2	0	1	2	14
21+	0	0	0	1	0	1	1	3	1	1	8
Total	2843	2649	2521	2409	2248	2124	3027	2826	2670	2546	25863

The baseline characteristics for 4,007 participants aged 65 or older are provided in Table 4.3. 2,286(57%) participants were female, and 1,721(43%) were male. Of the individuals aged 65 or older, 1,529(38.2%) were between 65 and 69, and the average age was 72.5 (s.d=±6.0). More than 60% had never received any education or only received it until elementary school. Among participants, 50% had two members in their household, and 57.0% lived in a detached house. While 1,667(41.6%) lived in a multi-unit house/ townhouse or apartment. 2,216(59.8%)

answered that they had never smoked, while 514(13.9%) answered that they were currently smoking. Among the smokers, the average smoking amount per day was 14.24. For chronic diseases, 3,509(88.1%) answered that they have more than three chronic diseases. From the baseline, 700(17.5%) were diagnosed with diabetes, 353(8.8%) were diagnosed with chronic heart disease, and 254(6.3%) had chronic respiratory disease. During the study period, 12(0.3%) people died.

Table 4.3 Baseline Characteristics (N=4007)

Variable	**Number (%)**
Gender	
Female	2286 (57.0)
Male	1721 (43.0)
Age	
65-69	1529 (38.2)
70-74	1220 (30.4)
75-79	749 (18.7)
80+	509 (12.7)
Age (continous)	
Mean (sd)	72.48 (6.0)
Median (IQR)	71 (68, 76)
Education	
None	806 (20.1)
Elementary	1712 (42.7)
Middle/High	1188 (29.7)
University	301 (7.5)
Number of household members	
1	690 (17.2)
2	2024 (50.5)
3	588 (14.7)
4	280 (7.0)
More than 5	425 (10.6)
Housing	
Detached House	2283 (57.0)
Multi-unit/Town house	557 (13.9)
Apartment	1110 (27.7)
Others	57 (1.4)
Smoking	
Current	514 (13.9)
Previous	976 (26.3)
No	2216 (59.8)
Drinking	
Never	1509 (40.5)
Didn’t drink for past 1 year	566 (15.2)
< 2 days/week	1076 (28.9)

2-3 days/week	280 (7.5)
Almost daily	294 (7.9)
Disability	
No	3457 (86.3)
Yes	550 (13.7)
Income quantile	
<20	1583 (39.7)
20-40	1041 (26.1)
40-60	664 (16.7)
60-80	394 (9.9)
80-100	303 (7.6)
Economic Activity	
No	2543 (63.5)
Yes	1464 (36.5)
>3 Chronic diseases	
No	474 (11.9)
Yes	3509 (88.1)
Walking	
None	699 (18.8)
≤ 3days/week	478 (12.8)
>3 days/week	2538 (68.4)
Medium physical activity	
None	2602 (69.9)
≤ 3days/week	331 (8.9)
>3 days/week	792 (21.2)
Diabetes	
No	3307 (82.5)
Yes	700 (17.5)
Chronic heart disease	
No	3654 (91.2)
Yes	353 (8.8)
Chronic respiratory disease	
No	3753 (93.7)
Yes	254 (6.3)
Death	
No	3995 (99.7)
Yes	12 (0.3)

*Chronic heart disease: myocardial infarction, ischemic heart disease, angina, pulmonary embolism, arrhythmia, conduction disorder, heart failure, heart valve syndrome, mitral stenosis, and other heart diseases

*Chronic respiratory disease: chronic obstructive pulmonary disease, bronchitis, asthma, pulmonary edema, and any other disease that was classified as 'disease of respiratory system'

4.4.2 Group-based trajectory modeling (GBTM) with binary outcome

4.4.2.1 Development of group-based trajectory modeling

The first step in building the trajectory model is determining the number of trajectory groups. The Bayesian Information Criterion (BIC) is the most commonly used criteria for deciding the optimal model. However, it should not be the only criterion when selecting the number of groups for trajectory modeling (Nagin, 2005). We compared multiple models with no starting points, each estimated with a different number of trajectory groups. There were no previous studies about geriatric pneumonia trajectories, so we generated the trajectories as quadratic. Table 4.4 shows the goodness-of-fit tests to select the optimal number of trajectories group, and we found that BIC monotonically decreased when more groups were added. Thus, BIC could not be the sole criterion for determining the number of groups. Also, Akaike's Information Criteria (AIC) and p_j were presented and considered. Here p_j denotes the probability that a model with j groups is the correct model from a set of J different models. The model with the largest BIC, AIC value, and p_j close to 1 would be considered the best-fitted model (Nagin, 2005). Based on BIC, AIC, and p_j criteria, trajectory with two groups was preferred. However, we considered whether the trajectory groups captured distinct and potentially meaningful patterns in the data with clinical aspects. Therefore, three groups of trajectories were selected for our trajectory analysis of geriatric pneumonia.

Table 4.4 Goodness of model fit to select the number of trajectory groups

Number of trajectories	**BIC**	**AIC**	p_j
2	-1971.01	-1948.97	0.89

3	-1973.03	-1938.40	0.12
4	-1978.97	-1931.76	<.0001
5	-1992.62	-1932.81	<.0001
6	-2000.94	-1928.54	<.0001

*BIC=Bayesian Information Criterion, AIC= Akaike's Information Criteria, p_j = probability that a model with j groups is the correct model, $p_j = \frac{e^{BIC_j - BIC_{max}}}{\sum_{j=1}^{J} e^{BIC_j - BIC_{max}}}$

After selecting the number of trajectory groups, the final model was selected with the best polynomial trajectory functions. Group-based trajectory modeling for geriatric pneumonia with binary outcomes included three trajectory groups: "low-flat" (Group1), "low-to-high" (Group2), and "high-to-low" (Group3). Group1 had a flat trajectory, Group2 had a linear trajectory, and Group3 had a quadratic trajectory. The final trajectory model is presented in Figure 4.3. In Figure 4.3, solid lines represent the group means, and the dashed lines represent the predictions.

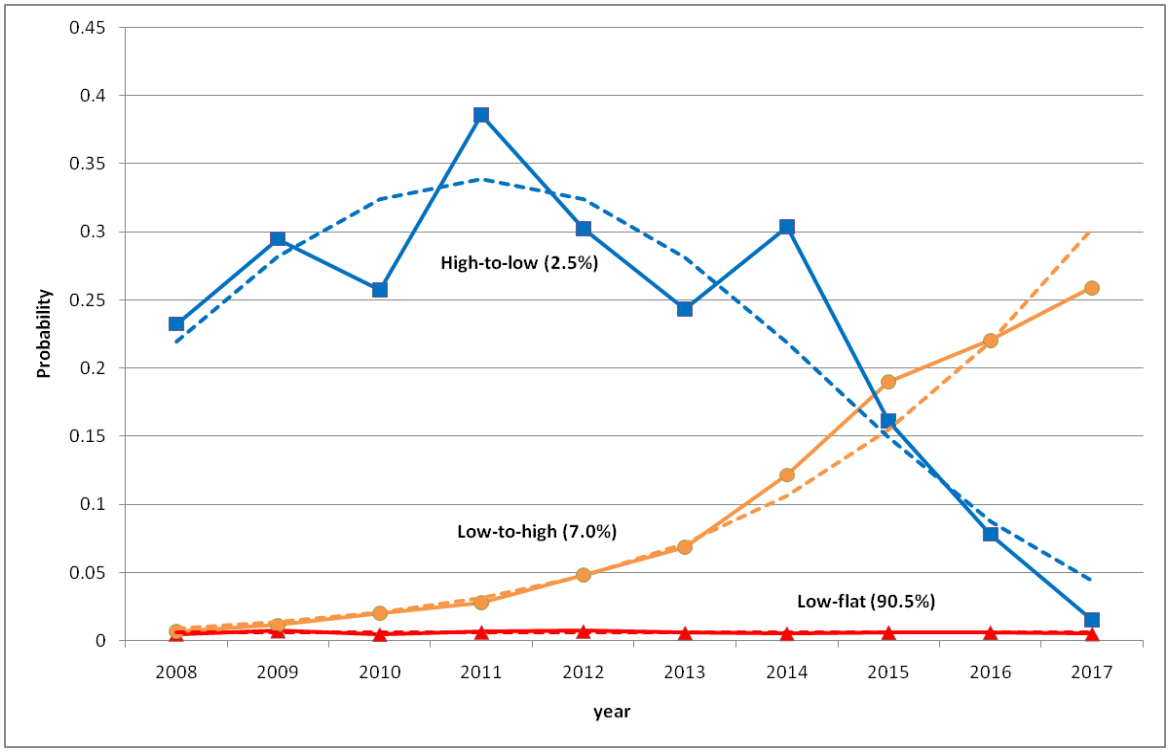


Figure 4.3 Pneumonia trajectories for group-based trajectory modeling with binary outcome

The first trajectory group included 90% of the participants, Group1 (n=3858; 90.5%), which indicates that the majority of participants were not diagnosed with pneumonia during the 10-year follow-up time. It showed a low-flat shape with a probability close to zero. The second trajectory, Group2 (n=90; 7.0%), showed an increasing trend. The probability started close to zero and consistently increased over time. The increase was more rapid after six years from the beginning. The third trajectory group, Group3 (n=59; 2.5%), showed a parabola shape. The probability of being diagnosed with pneumonia started high compared to the other groups, and it remained higher until seven years. But after that point, the probability started to decrease consistently. After ten years, the probability was close to zero.

Parameter estimates for the final model, including predictor variables, are presented in Table 4.5. For every group, all p-values were lower than our significance level of 5%. The null hypothesis is that the parameter equals zero. Hence, we can reject the null hypothesis and conclude that the parameter is not zero. Table 4.6 represents the estimates of group membership proportions, which also show significance in every group. The null hypothesis is that the proportion equals zero. In Table 4.6, two different percentages are reported. Sample membership classification percentages represent the percentage in each trajectory based on our total participants (N=4007), and estimated proportions are the percentage of the average probability of being in the trajectory (Wang et al., 2008).

Table 4.5 Parameter estimates for trajectory shapes

Group	**Parameter**	**Estimate**	**Standard Error**	**p-value**
Low-flat (Group1)	Intercept	-5.136	0.1890	<.0001
Low-to-high (Group2)	Intercept	-5.148	0.7767	<.0001
	Linear	0.431	0.1075	0.0001
High-to-low (Group3)	Intercept	-1.733	0.5111	0.0007
	Linear	0.533	0.2138	0.0126
	Quadratic	-0.067	0.0226	0.0031

Table 4.6 Estimates of group membership proportions

Group	Sample membership classification (%)	Estimate Proportion (%)	Standard Error	p-value
Low-flat (Group1)	96.282	90.484	2.3089	<.0001
Low-to-high (Group2)	2.246	7.054	2.2855	0.0020
High-to-low (Group3)	1.472	2.462	0.7696	0.0014

4.4.2.2 Characteristics of trajectory groups

A total of 4,007 participants were assigned to their trajectory group by group-based trajectory modeling, and the baseline characteristics were compared and represented in Table 4.7. There were no significant differences in age, level of education, housing type, smoking, alcohol intake, income quantile, current economic activity, more than three chronic diseases, walking, physical activity, baseline comorbidities (diabetes and chronic heart disease), and death among the pneumonia trajectory groups. Meanwhile, gender (p-value = 0.0260), number of household members (p-value = 0.0187), disability (p-value = 0.0046), and chronic respiratory disease (p-value = 0.0037) showed a significant difference among the groups from the chi-square test or Fisher's exact test.

The "low-flat" trajectory group (Group1) had the lowest male rate compared to other groups. While other groups had an over 50% proportion of males, the "low-flat" group had 42.5% males. Also, the percentage having a chronic respiratory disease was lower. About 6% of the participants in the "low-flat" group had chronic respiratory disease, while others had more than 10%. The "low-to-high" trajectory group (Group2) showed the highest rate of smokers. If we also consider smoking in the past, 53.3% of the participants were smokers or current smokers. Meanwhile, other trajectory groups' percentages of previous or current smokers were lower

than 50%. Lastly, the "high-to-low" trajectory group (Group3) had the highest mean age and the highest proportion of people over 80 years old. Also, they had a higher percentage of people with disability and the highest number of household members.

Table 4.7 Distribution of baseline characteristics by trajectory groups (N, %)

Variable	Low-flat Group1 (n=3858)	Low-to-high Group2 (n=90)	High-to-low Group3(n=59)	p-value
Gender				
Female	2217 (57.5)	42 (46.7)	27 (45.8)	0.0260
Male	1541 (42.5)	48 (53.3)	32 (54.2)	
Age	72.46 (6.0)	72.24 (5.7)	74.15 (6.8)	0.0922
Age (Categorical)				
65-69	1477 (38.3)	35 (38.9)	17 (28.8)	0.7849
70-74	1175 (30.5)	26 (28.9)	19 (32.2)	
75-79	717 (18.5)	19 (21.1)	13 (22.0)	
80+	489 (12.7)	10 (11.1)	10 (17.0)	
Education				
None	778 (20.2)	21 (23.3)	7 (11.9)	0.3055
Elementary	1639 (42.5)	39 (43.3)	34 (57.6)	
Middle/High	1151 (29.8)	22 (24.5)	15 (25.4)	
University	290 (7.5)	8 (8.9)	3 (5.1)	
# of household members				
1	677 (17.5)	8 (8.9)	5 (8.5)	0.0187
2	1953 (50.6)	42 (46.7)	29 (49.2)	
3	556 (14.4)	23 (25.5)	9 (15.2)	
4	269 (7.0)	6 (6.7)	5 (8.5)	
5+	403 (10.5)	11 (12.2)	11 (18.6)	
Housing				
Detached house	2194 (56.9)	56 (62.2)	33 (55.9)	0.5761
Multi-unit/Town house	537 (13.9)	11 (12.2)	9 (15.3)	
Apartment	1074 (27.8)	21 (23.3)	15 (25.4)	
Others	53 (1.4)	2 (2.2)	2 (3.4)	
Smoking				
No	2148 (60.2)	42 (46.7)	26 (54.2)	0.1147
Current	490 (13.7)	17 (18.9)	7 (14.6)	
Previous	930 (26.1)	31 (34.4)	15 (31.2)	
Disability				
No	3332 (86.4)	82 (91.1)	43 (72.9)	0.0046
Yes	526 (13.6)	8 (8.9)	16 (27.1)	
Drinking				
Never	1447 (40.3)	41 (45.5)	21 (43.7)	0.1645
Didn't drink for past 1 year	539 (15.0)	17 (18.9)	10 (20.8)	
< 2 days/week	1047 (29.2)	15 (16.7)	14 (29.2)	
2-3 days/week	271 (7.6)	8 (8.9)	1 (2.1)	

Almost daily	283 (7.9)	9 (10.0)	2 (4.2)	
Income quantile				
<20	1527 (39.8)	32 (35.6)	24 (40.7)	0.7687
20-40	1002 (26.1)	24 (26.7)	15 (25.4)	
40-60	637 (16.6)	17 (18.9)	10 (16.9)	
60-80	377 (9.8)	13 (14.4)	4 (6.8)	
80-100	293 (7.7)	4 (4.4)	6 (10.2)	
Economic activity				
No	2445 (63.4)	56 (62.2)	42 (71.2)	0.4515
Yes	1413 (36.6)	34 (37.8)	17 (28.8)	
>3 chronic disease				
No	462 (12.0)	8 (8.9)	4 (6.8)	0.3781
Yes	3372 (88.0)	82 (91.1)	55 (93.2)	
Walking				
None	672 (18.7)	16 (17.8)	11 (22.9)	0.7129
≤ 3days/week	463 (12.9)	12 (13.3)	3 (6.3)	
>3 days/week	2452 (68.4)	62 (68.9)	34 (70.8)	
Medium physical activity				
None	2500 (69.7)	62 (68.9)	40 (83.3)	0.0883
≤ 3days/week	325 (9.1)	6 (6.7)	0 (0.0)	
>3 days/week	762 (21.2)	22 (24.4)	8 (16.7)	
Diabetes				
No	3180 (82.4)	79 (87.8)	48 (81.4)	0.4057
Yes	678 (17.6)	11 (12.2)	11 (18.6)	
Chronic heart disease				
No	3522 (91.3)	78 (86.7)	54 (91.5)	0.3089
Yes	336 (8.7)	12 (13.3)	5 (8.5)	
Chronic respiratory disease				
No	3626 (94.0)	78 (86.7)	49 (83.1)	<.0001
Yes	232 (6.0)	12 (13.3)	10 (16.9)	
Death				
No	3847 (99.7)	89 (98.9)	59 (100.0)	0.1900
Yes	11 (0.3)	1 (1.1)	0 (0.0)	

*Chronic heart disease: myocardial infarction, ischemic heart disease, angina, pulmonary embolism, arrhythmia, conduction disorder, heart failure, heart valve syndrome, mitral stenosis, and other heart diseases
*Chronic respiratory disease: chronic obstructive pulmonary disease, bronchitis, asthma, pulmonary edema, and any other disease that was classified as 'disease of respiratory system'
*Chi-square test or Fisher's exact test was conducted for categorical variables. Student's t-test was conducted for continuous variables.

Logistic regression was conducted to identify relevant risk factors that may influence the trajectory groups. Univariate logistic regression was first performed, and the variables with p-value smaller than 0.1 were selected for multivariate logistic regression. Also, multicollinearity was checked. However, none of the variables had a variance influence factor (VIF) larger than

3, so we considered there was no multicollinearity issue in the analysis. The logistic regression analysis table (Table 4.8, Table 4.9) presents the odds ratio with 95% confidence interval (CI) and p-value. The "low-flat" trajectory group was set as the reference group for both logistic regressions.

To select the variables from the univariate logistic regression analysis, we applied a significance level of 10%. In the univariate logistic analysis, gender, number of household members, smoking status, and having a chronic respiratory disease were the significant factors in the "low-to-high" trajectory group (Group2). For the "high-to-low" group (Group3), gender, having a disability, and having a chronic respiratory disease were selected for multivariate logistic analysis.

Table 4.8 Univariate logistic regression analysis ("low-flat" group as reference group)

Variable	**Low to High (n=90)**		**High to Low (n=59)**	
	OR (95% CI)	**p-value**	**OR (95% CI)**	**p-value**
Gender				
Female	-	-	-	-
Male	1.54 (1.02, 2.35)	0.0422	1.60 (1.96, 2.68)	0.0738
Age				
65-69	-	-	-	-
70-74	0.93 (0.56, 1.56)	0.7936	1.41 (0.73, 2.72)	0.3118
75-79	1.12 (0.64, 1.97)	0.6984	1.58 (0.76, 3.26)	0.2209
80+	0.86 (0.42, 1.76)	0.6843	1.78 (0.81, 3.91)	0.1527
Education				
None	-	-	-	-
Elementary	0.88 (0.52, 1.51)	0.6455	2.31 (1.02, 5.22)	0.0453
Middle/High	0.71 (0.39, 1.30)	0.2635	1.45 (0.59, 3.57)	0.4207
University	1.02 (0.45, 2.33)	0.9587	1.15 (0.30, 4.48)	0.8405
# of household members				
1	-	-	-	-
2	1.82 (0.85, 3.90)	0.1231	2.01 (0.78, 5.21)	0.1510
3	3.50 (1.55, 7.89)	0.0025	2.19 (0.73, 6.58)	0.1617
4	1.89 (0.65, 5.49)	0.2437	2.52 (0.72, 8.76)	0.1471
More than 5	2.31 (0.92, 5.79)	0.0742	3.70 (1.28, 10.7)	0.0161
Housing				
Detached house	-	-	-	-
Multi-unit/Town house	0.80 (0.42, 1.54)	0.5093	1.11 (0.53, 2.34)	0.7753
Apartment	0.77 (0.46, 1.27)	0.3027	0.93 (0.50, 1.72)	0.8132
Others	1.48 (0.35, 6.22)	0.5937	2.51 (0.59, 10.7)	0.2147
Smoking				
No	-	-	-	-
Current	1.78 (1.00, 3.14)	0.0494	1.18 (0.51, 2.74)	0.6992
Previous	1.71 (1.07, 2.73)	0.0263	1.33 (0.71, 2.53)	0.3792

Drinking				
Never	-	-	-	-
Didn't drink for past 1 year	1.11 (0.63, 1.98)	0.7144	1.28 (0.60, 2.73)	0.5262
< 2 days/week	0.51 (0.28, 0.92)	0.0251	0.92 (0.47, 1.82)	0.8136
2-3 days/week	1.04 (0.48, 2.25)	0.9167	0.25 (0.03, 1.90)	0.1818
Almost daily	1.12 (0.54, 2.34)	0.7574	0.49 (0.11, 2.09)	0.3327
Disability				
No	-	-	-	-
Yes	0.62 (0.30, 1.29)	0.1974	2.36 (1.32, 4.22)	0.0038
Income quantile				
<20	-	-	-	-
20-40	1.14 (0.67, 1.95)	0.6246	0.95 (0.50, 1.83)	0.8333
40-60	1.27 (0.70, 2.31)	0.4262	1.00 (0.48, 2.10)	0.9975
60-80	1.65 (0.86, 3.17)	0.1358	0.68 (0.23, 1.96)	0.4694
80-100	0.65 (0.23, 1.86)	0.4224	1.30 (0.53, 3.22)	0.5659
Economic activity				
No	-	-	-	-
Yes	1.05 (0.68, 1.62)	0.8225	0.70 (0.40, 1.24)	0.2188
>3 chronic disease				
No	-	-	-	-
Yes	1.40 (0.69, 2.92)	0.3635	1.99 (0.68, 5.22)	0.2235
Walking				
None	-	-	-	-
≤ 3days/week	1.09 (0.51, 2.32)	0.8263	0.40 (0.11, 1.43)	0.1566
>3 days/week	1.06 (0.61, 1.85)	0.8321	0.85 (0.43, 1.68)	0.6350
Medium physical activity				
None	-	-	-	-
≤ 3days/week	0.75 (0.32, 1.74)	0.4951	0.10 (0.01, 1.55)	0.0986
>3 days/week	1.16 (0.71, 1.91)	0.5456	0.69 (0.33, 1.45)	0.3252
Diabetes				
No	-	-	-	-
Yes	0.65 (0.35, 1.23)	0.1893	1.08 (0.56, 2.08)	0.8304
Chronic heart disease				
No	-	-	-	-
Yes	1.61 (0.87, 2.99)	0.1296	0.97 (0.39, 2.44)	0.9500
Chronic respiratory disease				
No	-	-	-	-
Yes	2.41 (1.29, 4.48)	0.0057	3.19 (1.60, 6.38)	0.0010

*Chronic heart disease: myocardial infarction, ischemic heart disease, angina, pulmonary embolism, arrhythmia, conduction disorder, heart failure, heart valve syndrome, mitral stenosis, and other heart diseases

*Chronic respiratory disease: chronic obstructive pulmonary disease, bronchitis, asthma, pulmonary edema, and any other disease that was classified as 'disease of respiratory system'

Multivariate logistic regression was conducted with the selected variables from univariate logistic regression. Compared to the "low-flat" group (Group1), members from the "low-to-high" group (Group2) were more likely to have three members in their household (OR = 3.51, 95% CI: 1.56 – 7.92, p-value = 0.0024), and to have chronic respiratory disease (OR = 2.42, 95% CI: 1.30 – 4.51, p-value = 0.0055). More specifically, it could be interpreted that people with chronic respiratory disease had 2.42 times higher odds of being diagnosed with pneumonia

than those who don't have chronic respiratory disease. Interaction between the factors was checked, but it didn't show significance. Thus, it was not included in the final model. For the "high-to-low" group (Group3), having a disability (OR = 2.34, 95% CI: 1.31 – 4.19, p-value = 0.0042), and having a chronic respiratory disease (OR = 3.17, 95% CI 1.58 – 6.34, p-value = 0.0012) were the significant factors compared to the "low-flat" group. Also, there was no interaction effect between the factors.

Table 4.9 Multivariate logistic regression analysis ("low-flat" group as reference group)

Variable	**Low to High (n=90)**		**High to Low (n=59)**	
	OR (95% CI)	**p-value**	**OR (95% CI)**	**p-value**
# of household members				
1	-	-		
2	1.82 (0.85, 3.90)	0.1234		
3	3.51 (1.56, 7.92)	0.0024		
4	1.89 (0.65, 5.50)	0.2438		
5+	2.42 (1.30, 4.51)	0.0746		
Disability				
No			-	-
Yes			2.34 (1.31, 4.19)	0.0042
Chronic respiratory disease				
No	-	-	-	-
Yes	2.42 (1.30, 4.51)	0.0055	3.17 (1.58, 6.34)	0.0012

*Chronic respiratory disease: chronic obstructive pulmonary disease, bronchitis, asthma, pulmonary edema, and any other disease that was classified as 'disease of respiratory system'

4.4.3 Group-based trajectory modeling (GBTM) with zero inflated count outcome

4.4.3.1 Development of group-based trajectory modeling

In Table 4.2, about 98% of the elderly participants didn't have any hospital visits due to pneumonia. 98.5% of the participants had zero visits, and 0.9% had one or two visits. In contrast, some people had more visits than average, up to 41 visits in one year. Yet, over 99% of the participants had zero to two visits, so it was hard to find a distinctive trajectory model

that contains clinical relevance without categorizing the outcome data. The outcome variable that exceeded a certain point was categorized when modeling trajectory with zero-inflated count outcomes for group-based trajectory modeling (Wojciechowski, 2017). Therefore, we tried to categorize data in various ways and decided to categorize it over eight visits. From 1 to 7 were recorded as their original count, and others were coded as: 8 to 10 visits = 8, 11 to 20 visits = 9, and over 20 visits = 10. So, we used total ten categories for group-based trajectory modeling.

Also, we decided on the number of trajectory groups first for group-based trajectory modeling with count outcomes. Similar to binary outcomes, several criteria were considered for the optimal number of groups. Table 4.10 represents the goodness of fit test results for deciding the optimal number of trajectory groups for geriatric pneumonia with zero-inflated count outcomes. Three groups had the largest BIC, AIC, and p_j. Therefore, three trajectory groups were selected for geriatric pneumonia trajectory analysis with zero-inflated count outcomes.

Table 4.10 Goodness of model fit to select the number of trajectory groups

Number of trajectories	**BIC**	**AIC**	p_j
2	-2811.23	-2782.90	<.0001
3	-2783.80	-2742.88	>.9999
4	-2800.34	-2746.82	<.0001
5	-2810.97	-2744.86	<.0001
6	-2828.64	-2749.94	<.0001

*BIC=Bayesian Information Criterion, AIC= Akaike's Information Criteria, p_j = probability that a model with j groups is the correct model, $p_j = \frac{e^{BIC_j - BIC_{max}}}{\sum_{j=1}^{J} e^{BIC_j - BIC_{max}}}$

The final model with the best polynomial trajectory functions included three trajectory groups: "low-flat" (Group1), "low-to-high" (Group2), and "high-to-low" (Group3). Group1 had a flat

trajectory, and both Group2 and Group3 had a quadratic trajectory. Unlike the group-based trajectory model with binary outcomes, there was extra zero probability in the model. It was specified as linear. The final trajectory model is presented in Figure 4.4.

Figure 4.4 Pneumonia trajectories for group-based trajectory modeling with count outcome

Similar to the trajectory analysis results with binary outcomes, the majority of the participants were included in the first trajectory group, Group1 (n=3747; 84.8%). This group showed a low-flat shape with a count close to zero. The second trajectory, Group2 (n=180; 9.1%), showed an increasing trend. The hospital visits due to pneumonia started slightly higher than Group1 at 0.13, and after ten years of follow-up time, Group2 had the highest count. The third trajectory, Group3 (n=80; 6.1%), started with the highest count and constantly decreased after seven years.

Table 4.11 represents the parameter estimates for the final model. In Group1, the p-value was significant. In Group2, the intercept and linear predictor were not significant with a significance

level of 5%, but the quadratic predictor was significant (p-value = 0.0450). Furthermore, the linear predictor in Group3 also didn't show significance. However, the quadratic predictor was significant (p-value = 0.0085). As we applied the zero-inflated Poisson distribution in this model, the parameters for the extra zero probability part were estimated, and the shape of the polynomial function was determined. Both intercept and linear predictor were significant for the extra zero probability part (p-value = <.0001). Thus, this model shared the linear zero-inflation probability. Table 4.12 represents the estimates of group membership proportions, which show significance in every group.

Table 4.11 Parameter estimates for trajectory shapes

Group	**Parameter**	**Estimate**	**Standard Error**	**p-value**
Low-flat (Group1)	Intercept	-4.100	0.2852	<.0001
Low-to-high (Group2)	Intercept	0.590	0.3194	0.0647
	Linear	-0.096	0.1162	0.4090
	Quadratic	0.019	0.0097	0.0450
High-to-low (Group3)	Intercept	1.576	0.1833	<.0001
	Linear	0.107	0.0848	0.2047
	Quadratic	-0.025	0.0094	0.0085
Extra zero	Intercept	2.567	0.1460	<.0001
	Linear	-0.092	0.0207	<.0001

Table 4.12 Estimates of group membership proportions

Group	**Sample membership classification (%)**	**Estimate Proportion (%)**	**Standard Error**	**p-value**
Low-flat (Group1)	93.511	84.773	1.4763	<.0001
Low-to-high (Group2)	4.492	9.143	1.3019	<.0001
High-to-low (Group3)	2.000	6.084	1.1427	<.0001

4.4.3.2 Characteristics of trajectory groups

Baseline characteristics were compared and represented in Table 4.13. There were no significant differences in age, level of education, number of household members, housing type, alcohol intake, having a disability, income quantile, current economic activity, having more than three chronic diseases, walking, physical activity, baseline comorbidities (diabetes and chronic heart disease), and death among the three geriatric pneumonia trajectory groups. On the other hand, gender (p-value = <.0001), age (categorical) (p-value = 0.0064), smoking (p-value = 0.0001), and chronic respiratory disease (p-value = 0.0005) showed significant differences with a 5% significance level among the groups from the overall chi-square test.

When we compared the three trajectory groups, the remarkable point for the "low-flat" trajectory group (Group1) was that it had the lowest male and the highest non-smoker proportion. Also, the rate of people with chronic respiratory disease was the lowest. On the contrary, the "low-to-high" group (Group2) had the highest rate for male and smoker. In addition, in Group2, about 16.7% were older than 80 years old. Compared to other groups, the figure was relatively high. Lastly, the "high-to-low" trajectory group (Group3) had the highest rate of people with chronic respiratory disease.

Table 4.13 Distribution of baseline characteristics by trajectory groups (N, %)

Variable	**Low-flat Group1 (n=3747)**	**Low-to-high Group2 (n=180)**	**High-to-low Group3 (n=80)**	**p-value**
Gender				
Female	2176 (58.1)	75 (41.7)	35 (43.7)	<.0001
Male	1571 (41.9)	105 (58.3)	45 (56.3)	
Age	72.44 (6.0)	73.28 (6.7)	72.41 (4.8)	0.1887
Age (Categorical)				
65-69	1438 (38.4)	69 (38.3)	22 (27.5)	0.0064
70-74	1147 (30.6)	40 (22.2)	33 (41.3)	
75-79	688 (18.4)	41 (22.8)	20 (25.0)	
80+	474 (12.6)	30 (16.7)	5 (6.2)	
Education				
None	757 (20.2)	35 (19.4)	14 (17.5)	0.7287
Elementary	1598 (42.7)	73 (40.6)	41 (51.3)	

Middle/High	1113 (29.7)	55 (10.6)	20 (25.0)	
University	279 (7.4)	17 (9.4)	5 (6.2)	
# of household members				
1	665 (17.8)	18 (10.0)	7 (8.7)	0.0636
2	1885 (50.3)	91 (50.5)	48 (60.0)	
3	541 (14.4)	34 (18.9)	13 (16.3)	
4	259 (6.9)	16 (8.9)	5 (6.2)	
5+	397 (10.6)	21 (11.7)	7 (8.7)	
Housing				
Detached house	2145 (57.2)	93 (51.7)	45 (56.2)	0.3950
Multi-unit/Town house	524 (14.0)	22 (12.2)	11 (13.8)	
Apartment	1025 (27.4)	63 (35.0)	22 (27.5)	
Others	53 (1.4)	2 (1.1)	2 (2.5)	
Smoking				
No	2102 (60.7)	77 (46.1)	37 (48.0)	0.0001
Current	479 (13.8)	25 (15.0)	30 (39.0)	
Previous	881 (25.5)	65 (38.9)	10 (13.0)	
Drinking				
Never	1413 (40.6)	59 (34.9)	37 (48.0)	0.0546
Didn't drink for past 1 year	515 (14.8)	40 (23.7)	11 (14.3)	
< 2 days/week	1018 (29.3)	38 (22.5)	20 (26.0)	
2-3 days/week	258 (7.4)	17 (10.0)	5 (6.5)	
Almost daily	275 (7.9)	15 (8.9)	4 (5.2)	
Disability				
No	3244 (86.6)	149 (82.8)	64 (80.0)	0.0905
Yes	503 (13.4)	31 (17.2)	16 (20.0)	
Income quantile				
<20	1481 (39.8)	71 (39.4)	31 (38.8)	0.8106
20-40	981 (26.3)	41 (22.8)	19 (23.8)	
40-60	618 (16.6)	33 (18.3)	13 (16.2)	
60-80	360 (9.7)	23 (12.8)	11 (13.7)	
80-100	285 (7.6)	12 (6.7)	6 (7.5)	
Economic activity				
No	2372 (63.3)	121 (67.2)	50 (62.5)	0.5571
Yes	1375 (36.7)	59 (32.8)	30 (37.5)	
>3 chronic disease				
No	453 (12.2)	14 (7.8)	7 (8.7)	0.1402
Yes	3270 (87.8)	166 (92.2)	73 (91.3)	
Walking				
None	647 (18.6)	40 (23.7)	12 (15.6)	0.4660
≤ 3days/week	447 (12.8)	22 (13.0)	9 (11.7)	
>3 days/week	2385 (68.6)	107 (63.3)	56 (72.7)	
Medium physical activity				
None	2421 (69.6)	122 (72.2)	59 (76.6)	0.4905
≤ 3days/week	316 (9.1)	12 (7.1)	3 (3.9)	
>3 days/week	742 (21.3)	35 (20.7)	15 (19.5)	
Diabetes				
No	3094 (82.6)	144 (80.0)	69 (86.3)	0.4557
Yes	653 (17.4)	36 (20.0)	11 (13.7)	
Chronic heart disease				
No	3417 (91.2)	163 (90.6)	74 (92.5)	0.8776
Yes	330 (8.8)	17 (9.4)	6 (7.5)	
Chronic respiratory disease				
No	3521 (94.0)	165 (91.7)	67 (83.8)	0.0005
Yes	226 (6.0)	15 (8.3)	13 (16.2)	
Death				
No	3736 (99.7)	179 (99.4)	80 (0.0)	0.5563

Yes	11 (0.3)	1 (0.6)	0 (0.0)	

*Chronic heart disease: myocardial infarction, ischemic heart disease, angina, pulmonary embolism, arrhythmia, conduction disorder, heart failure, heart valve syndrome, mitral stenosis, and other heart diseases

*Chronic respiratory disease: chronic obstructive pulmonary disease, bronchitis, asthma, pulmonary edema, and any other disease that was classified as 'disease of respiratory system'

*Chi-square test or Fisher's exact test was conducted for categorical variable. Student's t-test was conducted for continuous variable

According to the univariate logistic regression analysis, gender, age, number of household members, smoking status, drinking status, and having more than three chronic diseases were the significant factors in the "low-to-high" trajectory group (Group2) at a significance level of 10%. For the "high-to-low" group (Group3), gender, age, smoking status, having a disability, and having a chronic respiratory disease were selected for multivariate logistic analysis. The reference group was set as the "low-flat" group. Multicollinearity was checked and confirmed that it had no issue. The overall univariate logistic regression results are presented in Table 4.14.

Table 4.14 Univariate logistic regression analysis ("low-flat" group as reference group)

Variable	**Low-to-high (n=180)**		**High-to-low (n=80)**	
	OR (95% CI)	**p-value**	**OR (95% CI)**	**p-value**
Gender				
Female	-	-	-	-
Male	1.94 (1.43, 2.63)	<.0001	1.78 (1.14, 2.78)	0.0113
Age				
65-69	-	-	-	-
70-74	0.73 (0.49, 1.08)	0.1153	1.88 (1.09, 3.24)	0.0231
75-79	1.24 (0.84, 1.85)	0.2847	1.90 (1.03, 3.51)	0.0399
80+	1.32 (0.85, 2.05)	0.2185	0.69 (0.26, 1.83)	0.4557
Education				
None	-	-	-	-
Elementary	0.99 (0.65, 1.49)	0.9544	1.39 (0.75, 2.56)	0.2949
Middle/High	1.07 (0.69, 1.65)	0.7637	0.97 (0.49, 1.94)	0.9349
University	1.32 (0.73, 2.39)	0.3630	0.97 (0.35, 2.72)	0.9523
# of household members				
1	-	-	-	-
2	1.78 (1.07, 2.09)	0.0272	2.42 (1.09, 5.37)	0.0300
3	2.32 (1.30, 4.16)	0.0046	2.28 (0.90, 5.76)	0.0806
4	2.28 (1.15, 4.54)	0.0188	1.83 (0.58, 5.83)	0.3042
More than 5	1.95 (1.03, 3.71)	0.0407	1.68 (0.58, 4.81)	0.3380
Housing				
Detached house	-	-	-	-
Multi-unit/Town house	0.97 (0.60, 1.56)	0.8943	1.00 (0.51, 1.95)	0.9985
Apartment	1.42 (1.02, 1.97)	0.0372	1.02 (0.61, 1.71)	0.9308
Others	0.87 (0.21, 3.63)	0.8488	1.80 (0.43, 7.61)	0.4249
Smoking				

No	-	-	-	-
Current	1.43 (0.90, 2.26)	0.1331	1.19 (0.59, 2.40)	0.6355
Previous	2.01 (1.43, 2.83)	<.0001	1.94 (1.19, 3.15)	0.0080
Drinking				
Never	-	-	-	-
Didn't drink for past 1 year	0.77 (0.43, 1.37)	0.3676	0.82 (0.41, 1.61)	0.5574
< 2 days/week	1.42 (0.77, 2.62)	0.2570	0.75 (0.43, 1.30)	0.3058
2-3 days/week	0.68 (0.37, 1.26)	0.2247	0.74 (0.29, 1.90)	0.5317
Almost daily	1.21 (0.59, 2.57)	0.6043	0.56 (0.20, 1.57)	0.2677
Disability				
No	-	-	-	-
Yes	1.34 (0.90, 2.00)	0.1472	1.61 (0.93, 2.81)	0.0921
Income quantile				
<20	-	-	-	-
20-40	0.87 (0.59, 1.29)	0.4936	0.93 (0.52, 1.65)	0.7919
40-60	1.11 (0.73, 1.70)	0.6178	1.01 (0.52, 1.93)	0.9882
60-80	1.33 (0.82, 2.16)	0.2450	1.46 (0.73, 2.93)	0.2878
80-100	0.88 (0.47, 1.64)	0.6839	1.01 (0.42, 2.43)	0.9898
Economic activity				
No	-	-	-	-
Yes	0.84 (0.61, 1.16)	0.2868	1.04 (0.66, 1.64)	0.8821
>3 chronic disease				
No	-	-	-	-
Yes	1.64 (0.94, 2.86)	0.0793	1.45 (0.66, 3.16)	0.3563
Walking				
None	-	-	-	-
≤ 3days/week	0.80 (0.47, 1.36)	0.4019	1.09 (0.45, 2.60)	0.8538
>3 days/week	0.73 (0.50, 1.05)	0.0920	1.27 (0.68, 2.38)	0.4629
Medium physical activity				
None	-	-	-	-
≤ 3days/week	0.75 (0.41, 1.38)	0.3594	0.39 (0.12, 1.25)	0.1131
>3 days/week	0.94 (0.64, 1.38)	0.7364	0.83 (0.47, 1.47)	0.5224
Diabetes				
No	-	-	-	-
Yes	1.19 (0.81, 1.72)	0.3758	0.76 (0.40, 1.44)	0.3921
Chronic heart disease				
No	-	-	-	-
Yes	1.08 (0.65, 1.80)	0.7686	0.84 (0.36, 1.94)	0.6832
Chronic respiratory disease				
No	-	-	-	-
Yes	1.42 (0.82, 2.44)	0.2110	3.02 (1.64, 5.56)	0.0004

*Chronic heart disease: myocardial infarction, ischemic heart disease, angina, pulmonary embolism, arrhythmia, conduction disorder, heart failure, heart valve syndrome, mitral stenosis, and other heart disease

*Chronic respiratory disease: chronic obstructive pulmonary disease, bronchitis, asthma, pulmonary edema, and any other disease that was classified as 'disease of respiratory system'

With the selected variables from the univariate logistic regression analysis results, multivariate logistic regression was performed. Final variables were selected at a significance level of 5%. Table 4.15 represents the results for multivariate logistic regression. Compared to the "low-flat" group (Group1), members from "low-to-high" group (Group2) were more likely to be male (OR = 1.94, 95% CI: 1.43 – 2.63, p-value = <.0001). For "high-to-low" group (Group3),

the odds of having hospital visits due to pneumonia was higher for male (OR = 1.72, 95% CI: 1.10 – 2.69, p-value = 0.0177). Additionally, having chronic disease (OR = 2.88, 95% CI 1.56 – 5.31, p-value = 0.0007) was significant compared to the "low-flat" group. Interaction between the factors was not significant.

Table 4.15 Multivariate logistic regression analysis ("low-flat" group as reference group)

Variable	**Low-to-high (n=180)**		**High-to-low (n=80)**	
	OR (95% CI)	**p-value**	**OR (95% CI)**	**p-value**
Gender				
Female	-	-	-	-
Male	1.94 (1.43, 2.63)	<.0001	1.72 (1.10, 2.69)	0.0177
Chronic respiratory disease				
No			-	-
Yes			2.88 (1.56, 5.31)	0.0007

*Chronic respiratory disease: chronic obstructive pulmonary disease, bronchitis, asthma, pulmonary edema, and any other disease that was classified as 'disease of respiratory system'

4.4.4 Comparison of group-based trajectory modeling with binary and zero-inflated count outcomes

The results of group-based trajectory modeling with binary outcomes and zero-inflated count outcomes are compared in Table 4.16. Upon visual inspection, both had three groups of trajectory with similar shapes: "low-flat," "low-to-high," and "high-to-low." Even though the percentage of each group was different, the "low-flat" group had the highest proportion, followed by the "low-to-high" and "high-to-low" groups. According to Table 4.17, 94.8% of participants were in the same group in both models. However, the kappa statistic was 0.465 (95% CI: 0.411 – 0.519), suggesting moderate agreement (Byrt, 1996).

The risk factors that influenced the trajectory groups were also different. Having a chronic respiratory disease appeared as a risk factor for most cases. When comparing the "low-to-high"

groups from each trajectory model, none of the factors were common. For "high-to-low" groups, having a chronic respiratory was the same risk factor for both trajectory models with binary outcomes and zero-inflated count outcomes.

Table 4.16 Comparison of group-based trajectory modeling with binary outcome and zero inflated count outcome

Characteristic	**Binary**	**Zero-inflated count**
Number of groups	3	3
Composition of groups	Low-flat (90.5%) Low-to-high (7.0%) High-to-low (2.5%)	Low-flat (84.8%) Low-to-high (9.1%) High-to-low (6.1%)
BIC AIC	-1960.07 -1934.89	-2788.02 -2753.40
Risk factors (ref = low-flat)	<Low-to-high> Number of household members Chronic respiratory disease <High-to-low> Disability Chronic respiratory disease	<Low-to-high> Gender <High-to-low> Gender Chronic respiratory disease

Table 4.17 Group assignment for each model (N, %)

Group	Low-flat (Count)	Low-to-high (Count)	High-to-low (Count)
Low-flat (Binary)	3718 (92.8)	97 (2.4)	43 (1.1)
Low-to-high (Binary)	29 (0.7)	51 (1.3)	10 (0.2)
High-to-low (Binary)	0 (0.0)	32 (0.8)	27 (0.7)

CHAPTER 5. DISCUSSION

In this thesis, we applied group-based trajectory modeling to identify the trajectories for geriatric pneumonia with binary and zero-inflated count outcomes. Trajectory shape and membership differences were compared, and the risk factors for models with binary and zero-inflated count outcomes were identified. This thesis helps to explain the development of pneumonia among older adults and identifies which individual subgroups are at risk for pneumonia.

We observed similarities and differences between the group-based trajectory modeling with binary outcomes and zero-inflated count outcomes. Both models had three trajectory groups that appeared to have some resemblance. The majority of the participants were classified into the “low-flat” group. However, the percentage of the “low-flat” group was higher in the binary model. As the modeling follows the maximum assignment rule, there is a possibility that an individual will be assigned to the “low-flat” group even though they were diagnosed with pneumonia. Among the participants, 140 people were assigned to the “low-flat” group in the binary outcome model but not in the zero-inflated count outcome model. This may account for a higher percentage of the “low-flat” group in the binary outcome trajectory model than the count outcome model.

Identified risk factors for each model were also different for the “low-to-high” group and the “high-to-low” group compared to the “low-flat” group. For the “low-to-high” group, the risk factors were having three household members and having a chronic respiratory disease in the binary outcome model. In comparison, male sex was the only predictor in the zero-inflated count trajectory model. For the “high-to-low” group compared to the “low-flat” group, the risk factors were having a disability and having a chronic respiratory disease in the binary model.

The risk factors for the "high-to-low" group in the zero-inflated count outcome trajectory model were male sex and having a chronic respiratory disease.

The choice between models with different types of outcomes could be based on various decision points. Model diagnostics statistics, such as BIC could be used. Based on the BIC and AIC values, the binary outcome model had a larger value than the zero-inflated count outcome model in this thesis, which can be considered a better fit model. In a trajectory study that compared the trajectory models with different type of outcomes, the final outcome type was also determined by BIC value (Elmer et al., 2019). The study chose the count outcome model with the zero-inflated Poisson distribution applied when they generated trajectories of prescription opioids filled over time with three different types of outcome: binary, count, and continuous. This study also added visual inspection to the decision of selecting the outcome type and concluded that the count outcome model would be the best fit and of greatest clinical interest (Elmer et al., 2019). However, this study did not include individuals who didn't have an opioid prescription. It only included individuals who filled at least one prescription for an opioid analgesic during the nine-year study period. They utilized over four years of data after entry in the model during the period. Regardless of the similar decision points of the outcome types for group-based trajectory modeling, we chose a different outcome model from this study.

Consistent with our study, Ferraro and Wilmoth (2000) compared the use of the binary disease variables with counts of the same condition in models for measuring morbidity and concluded that the binary variable provided better fitting models in both cross-sectional and longitudinal analyses. Overall they suggested that using binary variables gave advantages in explanatory power and model fit (Ferraro & Wilmoth, 2000). However, they suggested testing the statistical

power when using the binary variable, and if parsimony is not a major concern, the binary variable approach may be preferred (Ferraro & Wilmoth, 2000).

As Nagin suggested, the model's decision shouldn't be solely based on one statistic. Nagin recommended that content knowledge is crucial for model selection (Nagin, 2005). Thus, the decision on the outcome and model should be guided by study objectives and the role that it captures for the clinical aspect of the study (Hickson, 2021). The interpretations of each model for this thesis are different. The binary model estimates the probability of being diagnosed with pneumonia, and the zero-inflated count model estimates the number of hospital visits due to pneumonia. Therefore, even though the binary model showed a better fit in this thesis, the choice between the two models might vary depending on the type of information that we would like to obtain. So, if we are interested in the trajectory and the factors that make the risk higher for being diagnosed with pneumonia, the binary model would be more helpful. Also, it could be easier to interpret. On the other hand, analysis according to the number of hospital visits enables us to infer more diverse aspects. Additional health care-interaction due to pneumonia can mean more severe and complex medical conditions, lack of understanding of pneumonia follow up after treatment, cost-ineffective use of inpatient beds, and many other related things (Adamuz et al., 2011).

Overall, even though the two trajectory models had some common features, according to the statistical fit and our primary interest, group-based trajectory modeling with binary outcomes may be more helpful in seeing the trajectories of geriatric pneumonia among the population.

As there were no previous studies about pneumonia trajectories, we couldn't compare our pneumonia trajectory results with other studies. However, similar to other disease trajectory studies, regardless of the number of groups, most of the participants were in the low probability or low count group. This was observed in various trajectory studies with diseases such as

depression, anxiety, asthma, or cardiovascular diseases (Lim et al., 2020; Cheng et al., 2021; Pape et al., 2021; Koohi et al., 2021).

Our study found that having a chronic respiratory disease, the number of household members, and having a disability were predictors of geriatric pneumonia trajectory membership in the binary outcome model. In the zero-inflated count outcome model, having a chronic respiratory disease and gender were predictors of geriatric pneumonia trajectory membership.

Having a chronic respiratory disease was the only factor that appeared as a predictor in both models. Except for the "low-to-high" group in the zero-inflated count trajectory model, having a chronic respiratory disease was stated as a risk factor in every trajectory membership group when compared to the "low-flat" group. This finding was consistent with various studies that have shown that people with chronic respiratory disease are more likely to have pneumonia (Koivular et al., 1994; Chang, 2010; Vila-Corcoles et al., 2008; Jackson et al., 2004; Jackson et al., 2009; Kaysin & Viera, 2016; Kline et al., 2015; Yoshikawa & Marrie, 2000; Gau et al., 2010; Loeb et al., 2009; Skull et al., 2009).

In our study, gender was stated as a risk factor only in the model that used the hospital visit counts as an outcome variable. Thus, we can consider that men are more likely to have more hospital visits due to pneumonia. However, there were some conflicting studies about gender as a risk factor. Some studies stated male sex as a risk factor, which is consistent with our study (Jackson et al., 2004; Vila-Corcoles et al., 2008; Yoshikawa & Marrie, 2000; Skull et al., 2009). On the other hand, many studies did not mention gender as a risk factor. In Koivular et al. (1994), gender was not associated with pneumonia nor any pneumonia-related hospitalization or death.

Crowded living conditions, such as living in nursing homes for the elderly, can increase the risk of contracting pneumonia (WHO, 2020). However, there were not enough samples to

identify if living in a nursing home can increase the risk of pneumonia in our study. Therefore, the number of household members was analyzed, and it was identified as a risk factor in the binary outcome model. Specifically, people with three members in their household were more likely to have pneumonia. This result may be that in our study, 67.7% had one or two household members, and the proportion of four or more household members was not very high. Thus, only having three household members showed significance. However, further research is needed on why having more household members was not associated with the pneumonia risk.

Having a disability was another predictor associated with pneumonia in the binary outcome model. This factor wasn't included in many geriatric pneumonia risk factor studies. However, some studies showed an association between specific impairments and pneumonia. For example, cognitive impairments and swallowing impairments increase the risk of pneumonia in older adults (Naruishi et al., 2018; Hollar et al., 2016; Ohrui, 2005; Nakajoh et al., 2000). Also, Centers for Disease Control and Prevention (CDC) (2020) reported that people with certain types of disability have a higher risk of pneumonia, which is consistent with our study. In our study, disability was utilized as a binary variable. We divided our participants by whether they had a disability or not, but to analyze the reason for the risk factor in more detail, further analysis would be required by type or level of disabilities.

Among pneumonia studies, age and smoking status were stated numerous times as risk factors. Many geriatric pneumonia studies showed an association with age (Koivular et al., 1994; Jackson et al., 2004; Vila-Corcoles et al., 2008; Yoshikawa & Marrie, 2000; Loeb et al., 2009; Skull et al., 2009). However, age was not associated with pneumonia in our study, which was consistent with some studies (Jackson et al., 2009; Gau et al., 2010). Also, smoking was not considered a risk factor in our study. Smoking had conflicting results among various geriatric pneumonia studies. In Gau et al. (2010) and Loeb et al. (2009), smoking status was a

significant risk factor among hospitalized pneumonia patients. Meanwhile, Skull et al. (2009) didn't find an association between smoking and hospitalized pneumonia patients.

5.1 Strength and limitations

There are several strengths in this study. One of the strengths is that no study has been conducted on pneumonia trajectories to the best of my knowledge. Our research suggested new information about geriatric pneumonia, which can act as a guideline for future pneumonia trajectory studies. Additionally, we generated two pneumonia trajectories with different types of outcomes (binary/count), from which we can derive different information about geriatric pneumonia trajectories. Also, we utilized the KHPS data, which is national-level large-scale data collected for ten years. Ten years of the study period would be considered sufficient time to study the development of pneumonia. As additional people were recruited in 2014, we conducted group-based trajectory modeling only with the original data collected from 2008 to check the sensitivity. The trajectory results and relevant risk factors were similar to our final trajectory model that used the combined data. Thus, combining the original and the additional data would not be an issue. The results are attached in Appendix A. Moreover, in this data, pneumonia outcomes and other comorbidities were clinically diagnosed, reducing the bias of using self-reported data.

This study also has several limitations. First, as the data were collected for ten consecutive years, missing data was unavoidable. Even though we used maximum likelihood estimation for parameter estimation, missing data bias could still exist. Second, there are many types of pneumonia, and the risk factors vary accordingly. However, in our data, the types of pneumonia were not distinguished. Also, the risk factors for hospital-acquired pneumonia (HAP) and ventilator-associated pneumonia (VAP), such as mechanical ventilation and residence in

intensive care units (ICU) were unavailable in our data. Additionally, history of hospitalization for pneumonia was also unavailable, and there were not enough participants living in nursing homes. Lastly, for geriatric pneumonia, vaccination could be a preventive factor. Many studies suggest getting vaccinated for older adults (Vila-Corcoles et al., 2008; Furman et al., 2021; Kline et al., 2016; Kaysin & Viera, 2016). However, vaccination information was not available in our data.

CHAPTER 6. CONCLUSION AND FUTURE RESEARCH

6.1 Conclusion

This thesis conducted group-based trajectory modeling for geriatric pneumonia with binary and count outcomes. Overall, the binary outcome model and the count outcome model had similar trajectory shapes. Three trajectory groups were generated for both models: “low-flat,” “low-to-high,” and “high-to-low.” The binary outcome trajectory model had larger BIC and AIC than the count outcome model when we compared the goodness of fit statistics. Also, the binary outcome trajectory model could be easier to interpret and apply to future studies. Thus we preferred the binary outcome trajectory model. However, as they generated similar results on the trajectory groups and shapes, the choice of the outcome type could depend on the researchers’ study objective.

The majority of the participants were included in the “low-flat” group. This finding indicates that most older adults did not suffer from pneumonia during the study period. People with chronic respiratory disease, three household members, and disability were more likely to get pneumonia. Furthermore, being male and having a chronic respiratory disease increased the risk of having more hospital visits due to pneumonia.

As different approaches and management are required for geriatric pneumonia, our findings can assist. We utilized a large national-level longitudinal data and generated geriatric pneumonia trajectories. There were no previous studies about pneumonia trajectories. Therefore, our new findings may support experts on their intervention programs for pneumonia, especially in community-living older adults.

6.2 Future research

Both binary outcome model and zero-inflated count model identified three trajectory groups for geriatric pneumonia: “low-flat,” “low-to-high,” and “high-to-low.” As expected, most of the participants belonged to the “low-flat” group like in other disease trajectory studies (Lim et al., 2020; Cheng et al., 2021; Pape et al., 2021; Koohi et al., 2021). Additionally, as age was a risk factor in many geriatric pneumonia studies, an increasing trend was also expected. However, the “high-to-low” group was remarkable. The predictors for the “high-to-low” group compared to the “low-flat” group were having a chronic respiratory disease and disability in the binary outcome model, and having a chronic respiratory disease and being male in the zero-inflated count outcome trajectory model. Having a chronic respiratory disease was a common factor among the two models. As nobody was dead in the “high-to-low” group during the study period, death was not considered for the reason of the “high-to-low” shape. Sometimes disease incidence rates can have a decreasing trend after an ascending trend when some intervention was effective (Amini et al., 2021; UNAIDS, 1999). To better understand this trajectory group, further analysis and detailed investigation would be needed.

In this thesis, we didn’t include any time-dependent factors. Pneumonia incidence can be affected by some special occasions. For instance, pneumonia was a common complication among patients with H1N1 influenza during the 2009 H1N1 pandemic (Jain et al., 2012). Likewise, in our data, the pneumonia incidence rate increased by 0.6% in older adults, while it increased by only 0.1% in the total population during the 2009 H1N1 pandemic. Considering that the COVID-19 pandemic has hugely affected the world since 2019, we should consider adding time-dependent covariates in future research.

Lastly, pediatric pneumonia trajectory studies are not yet published. As pneumonia is a leading cause of morbidity and mortality in young children (Wardlaw et al., 2006), pediatric pneumonia trajectory research would be informative.

CHAPTER 7. REFERENCES

Adamuz, J., Viasus, D., Campreciós-Rodríguez, P., et al (2011). A prospective cohort study of healthcare visits and rehospitalizations after discharge of patients with community-acquired pneumonia. *Respirology, 16(7),* 1119-26.

Almirall, J., Gonzalez, C. A., Balanzo, X., et al (1999). Proportion of Community-Acquired Pneumonia Cases Attributable to Tobacco Smoking. *Chest, 116(2),* 375-379.

American Thoracic Society (2019). *Fact sheets: Top 20 Pneumonia Facts 2019.* https://www.thoracic.org/patients/patient-resources/resources/top-pneumonia-facts.pdf

American Thoracic Society; Infectious Diseases Society of America. (2005). Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. *Am J Respir Crit Care Med, 171(4),* 388-416

Amini, M., Zayeri, F., Salehi, M. (2021). Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017. *BMC Public Health, 21*(1), 1-12.

Attwood, G., Dyer, G., Skipworth, G. (2000). *Statistics* (Vol. 1). Heinemann.

Böhning, D., Dietz, E., Schlattmann, P., et al (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of Royal Statistical Society, A 162,* 195–209.

Broadbent, J. M., Thomson, W. M., Poulton, R. (2008). Trajectory patterns of dental caries experience in the permanent dentition to the fourth decade of life. *Journal of dental research, 87(1),* 69–72.

Byrt, T. (1996). How good is that agreement? *Epidemiology, 7(5),* 561.

CDC. (2020). *Disability and Health Related Conditions.* https://www.cdc.gov/ncbddd/disabilityandhealth/relatedconditions.html

CDC. (2021). *National Center for Health Statistics. Faststats: Pneumonia.* https://www.cdc.gov/nchs/fastats/pneumonia.htm

Murphy, S. L., Kochanek, K. D., Xu, J. Q., et al (2021). Mortality in the United States, 2020. NCHS Data Brief, no 427. *Hyattsville, MD: National Center for Health Statistics*.

Chang, H. H. (2010). Community-acquired pneumonia in elderly patients. *Korean J Med, 79(4),* 346-355

Cheng, Y., Thorpe, L., Kabir, R., et al (2021). Latent class growth modeling of depression and anxiety in older adults: an 8-year follow-up of a population-based study. *BMC Geriatr, 21*, 550.

Chu, M. K. M., & Koval, J. J. (2014). Trajectory modeling of longitudinal binary data: application of the EM algorithm for mixture models. *Communications in Statistics-Simulation and Computation*, *43*(3), 495-519.

Coughlin, A. M. (2007). Combating community-acquired pneumonia. *Nursing, 37 (2),* 64hn1-64hn3.

Cunha, B. A. (2010). *Pneumonia Essentials.* Jones & Bartlett Learning

Diggle, P. J., Heagerty, P., Liang, K. Y., et al (2002). *Analysis of longitudinal data*. Oxford university press.

Edwards, L. J. (2000). Modern statistical techniques for the analysis of longitudinal data in biomedical research. *Pediatric pulmonology*, *30*(4), 330-344.

Elmer, J., Fogliato, R., Setia, N., et al (2019). Trajectories of prescription opioids filled over time. *PLoS One, 14(10),* e0222677.

Fedorov, V., Mannino, F., Zhang, R. (2009). Consequences of dichotomization. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry, 8*(1), 50-61.

Ferraro, K. F. & Wilmoth, J. M. (2000) Measuring Morbidity: Disease Counts, Binary Variables, and Statistical Power. *The Journals of Gerontology: Series B, 55 (3),* S173–S189.

File, T. M. (2003). Community-acquired pneumonia. *Lancet, 362(9400),* 1991-2001.

Flint, K. M., Schmiege, S. J., Allen, L. A, et al. (2017). Health Status Trajectories Among Outpatients With Heart Failure. *J Pain Symptom Manage, 53(2),* 224-231.

Furman, C. D., Leinenbach, A., Usher, R., et al (2021). Pneumonia in older adults. *Current opinion in infectious diseases, 34(2),* 135–141.

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

Garcia, T. P., & Marder, K. (2017). Statistical approaches to longitudinal data analysis in neurodegenerative diseases: Huntington's disease as a model. *Current neurology and neuroscience reports, 17*(2), 1-9.

Gau, J. T., Acharya, U., Khan, S., et al (2010). Pharmacotherapy and the risk for community-acquired pneumonia. *BMC Geriatr, 6,* 10:45.

Gritly, S. M., Elamin, M. O., Rahimtullah, H., et al (2018). Risk factors of pneumonia among children under 5 years at a pediatric hospital in Sudan. *International Journal of Medical Research & Health Sciences, 7*(4), 60-68.

Gupta, P. L., Gupta, R. C., Tripathi R. C. (1996). Analysis of zero-adjusted count data. *Computational Statistics & Data Analysis, 23*, 531-547.

Hickson, R. P., Annis, I. E., Killeya-Jones, L. A., et al (2021). Comparing Continuous and Binary Group-based Trajectory Modeling Using Statin Medication Adherence Data. *Med Care, 59(11),* 997-1005.

Hollaar, V., van der Maarel-Wierink, C., van der Putten, G. J., et al (2016). Defining characteristics and risk indicators for diagnosing nursing home-acquired pneumonia and aspiration pneumonia in nursing home residents, using the electronically-modified Delphi Method. *BMC geriatrics, 16*(1), 1-10.

Hu, M. C., Pavlicova, M., Nunes, E.V. (2011). Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial, *Am J Drug Alcohol Abuse, 37,* 367-75.

Huang, S. T., Wen, Y. W., Shur-Fen Gau, S., et al (2019). A Group-based Trajectory Analysis of Longitudinal Psychotropic Agent Use and Adverse Outcomes Among Older People. *J Am Med Dir Assoc, 20(12),* 1579-1586.

Inoue, Y., Koizumi, A., Wada, Y., et al (2007). Risk and protective factors related to mortality from pneumonia among middle-aged and elderly community residents: the JACC Study. *Journal of epidemiology, 17(6),* 194–202.

Jackson, M.L., Nelson, J.C., Jackson, L.A. (2009). Risk Factors for Community-Acquired Pneumonia in Immunocompetent Seniors. *Journal of the American Geriatrics Society, 57,* 882-888.

Jackson, M. L., Neuzil, K. M., Thompson, W. W., et al (2004). The burden of community-acquired pneumonia in seniors: results of a population-based study. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 39(11),* 1642–1650.

Jain, S., Benoit, S. R., Skarbinski, J., et al. (2012). Influenza-associated pneumonia among hospitalized patients with 2009 pandemic influenza A (H1N1) virus—United States, 2009. *Clinical infectious diseases*, *54*(9), 1221-1229.

Janssens, J. P. & Krause, K. H. (2004). Pneumonia in the very old. *The Lancet Infectious Diseases, 4 (2),* 112-124.

Jones, B. L. (2001). *Analyzing longitudinal data with mixture models: a trajectory approach* (Doctoral dissertation, Carnegie Mellon University).

Jones, B. L. (2020). *traj, group-based modeling of longitudinal data*. https://www.andrew.cmu.edu/user/bjones/

Jones, B. L., Nagin, D. S. (2007). Advances in group-based trajectory modeling and an sas procedure for estimating them. *Sociological methods & research, 35 (4),* 542-571.

Jones, B. L., Nagin, D. S., Roeder, K. (2001). A sas procedure based on mixture models for estimating developmental trajectories. *Sociological methods & research, 29 (3),* 374-393.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association, 90 (430),* 773–795.

Kass, R. E., & Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association, 90 (431),* 928–934.

Kaysin, A. & Viera, A. J. (2016). Community-Acquired Pneumonia in Adults: Diagnosis and Management. *Am Fam Physician, 94(9),* 698-706.

Kernic, M. A., Holt, V. L., Wolf. M. E., et al (2002). Academic and School Health Issues Among Children Exposed to Maternal Intimate Partner Abuse. *Arch Pediatr Adolesc Med. 156(6),* 549–555.

Kieninger, A. N. & Lipsett, P. A. (2009). Hospital-acquired pneumonia: pathophysiology, diagnosis, and treatment. *Surg Clin North Am, 89(2),* 439-61

Kline, K. A., Bowdish D. M., (2016). Infection in an aging population. *Current Opinion in Microbiology, 29,* 63-67

Koivula, I., Sten, M., Mäkelä, P. H. (1994). Risk factors for pneumonia in the elderly. *Am J Med, 96(4),* 313-20.

Koohi, F., Ahmadi, N., Hadaegh, F., et al (2021). Trajectories of cardiovascular disease risk and their association with the incidence of cardiovascular events over 18 years of follow-up: The Tehran Lipid and Glucose study. *J Transl Med, 19,* 309.

Lambert, D., (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34,* 1–14.

Lanks, C. W., Musani, A. I., Hsia, D. W. (2019). Community-acquired Pneumonia and Hospital-acquired Pneumonia. *Medical Clinics of North America, 103 (3),* 487-501.

Lim, H. J., Cheng, Y., Kabir, R., et al (2020). Trajectories of depression and their predictors in a population-based study of Korean older adults. *The International Journal of Aging and Human Development,* 0091415020944405.

Loeb, M., Neupane, B., Walter, S.D., et al (2009). Environmental Risk Factors for Community-Acquired Pneumonia Hospitalization in Older Adults. *Journal of the American Geriatrics Society, 57,* 1036-1040.

Loeys, T., Moerkerke, B., De Smet, O., et al. (2012). The analysis of zero-inflated count data: beyond zero-inflated Poisson regression. *The British journal of mathematical and statistical psychology, 65(1),* 163–180.

MacCallum, R. C., Zhang, S., Preacher, K. J., et al. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods, 7*(1), 19.

Massaro, T. (2017). Poisson mixture distribution analysis for North Carolina SIDS counts using information criteria. *Epidemiology, Biostatistics and Public Health, 14*(3).

Matheson, F. I., White, H. L., Moineddin, R., et al (2012). Drinking in context: the influence of gender and neighbourhood deprivation on alcohol consumption. *Journal of epidemiology and community health, 66(6),* e4.

McLachlan, G. J., Lee, S. X., Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application, 6,* 355-378.

McLachlan, G. J., & Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, Inc.

Micek, S. T., Kollef, K. E., Reichley, R. M., et al (2007). Health care-associated pneumonia and community-acquired pneumonia: a single-center experience. *Antimicrob Agents Chemother, 51(10),* 3568-73.

Musher, D. M. & Thorner, A. R. (2014). Community-Acquired Pneumonia. *The New England Journal of Medicine, 371,* 1619-1628.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods, 4 (2),* 139.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, Massaachusetts: Havard University Press.

Nagin, D. S. (2014). Group-Based Trajectory Modeling: An Overview. *Ann NutrMetab, 65,* 205-210.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specifcation and estimation of a nonparametric, mixed poissonmodel. *Criminology, 31 (3),* 327-362.

Nagin, D. S., & Odgers, C. L. (2010). Group-based trajectory modeling in Clinical Research. *Annual review of clinical psychology, 6,* 109-38.

Nakajoh, K., Nakagawa, T., Sekizawa, K., et al (2000). Relation between incidence of pneumonia and protective reflexes in post-stroke patients with oral or tube feeding. *Journal of internal medicine*, *247*(1), 39-42.

Naruishi, K., Nishikawa, Y., Kido, J. I., et al (2018). Relationship of aspiration pneumonia to cognitive impairment and oral condition: a cross-sectional study. *Clinical oral investigations*, *22*(7), 2575-2580.

Nawa, V. M. (2014). A Mixture Model for Longitudinal Trajectories. *International Journal of Statistics and Applications, 4(4),* 181-191.

Neil, R., Sampson, R. J., Nagin, D. S. (2021). Social change and cohort differences in group-based arrest trajectories over the last quarter-century. *Proceedings of the National Academy of Sciences*, *118*(31).

Nest, G., Passos, V. L., Candel, M. J. J. M., et al (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modeling approaches, fit statistics, and software. *Advances in Life Course Research, 43,* 100323

Nguena Nguefack, H. L., Pagé, M. G., Katz, J., et al (2020). Trajectory Modelling Techniques Useful to Epidemiological Research: A Comparative Narrative Review of Approaches. *Clin Epidemiol, 2020(12),* 1205-1222

Niederman, M. S., & Brito, V. (2007). Pneumonia in the older patient. *Clinics in chest medicine, 28(4),* 751–vi.

Ohrui, T. (2005). Preventive strategies for aspiration pneumonia in elderly disabled persons. *The Tohoku journal of experimental medicine*, *207*(1), 3-12.

Orui, M. (2020). Re-Increased Male Suicide Rates in the Recovery Phase Following the Great East Japan Earthquake. *Crisis, 41(6),* 422-428.

Pape, K., Cowell, W., Sejbaek, C. S., et al (2021). Adverse childhood experiences and asthma: trajectories in a national cohort. *Thorax,76,* 547-553.

Pitman, J. (1993). *Probability.* Springer-Verlag New York.

Raghavendran, K., Mylotte, J. M., Scannapieco, F. A. (2007). Nursing home-associated pneumonia, hospital-acquired pneumonia and ventilator-associated pneumonia: the contribution of dental biofilms and periodontal inflammation. *Periodontology 2000, 44,* 164–177.

Raymer, K. & Yang, H. (1998). Patients with aortic stenosis: Cardiac complications in non-cardiac surgery. *Can J Anaesth 45,* 855–859.

Salonen, J. (2020). *New methods in pension evaluation: Applications of trajectory analysis and dynamic microsimulation*. Finnish Centre for Pensions.

Smith, L. G. (2010). Mycoplasma pneumonia and its complications. *Infect Dis Clin North Am, 24(1),* 57-60.

Statistics Canada (2022). *Table 13-10-0394-01 Leading causes of death, total population, by age group.* https://doi.org/10.25318/1310039401-eng

Statistics Canada (2022). *Deaths: 2020*. https://www150.statcan.gc.ca/n1/daily-quotidien/220124/dq220124a-eng.htm

Statistics Korea (2021). *Causes of Death Statistics in 2020.* http://kostat.go.kr/portal/eng/pressReleases/8/10/index.board?bmode=read&bSeq=&aSeq=414516&pageNo=1&rowNum=10&navCount=10&currPg=&searchInfo=&sTarget=title&sTxt=

Sutriana, V. N., Sitaresmi, M. N., Wahab, A. (2021). Risk factors for childhood pneumonia: a case-control study in a high prevalence area in Indonesia. *Clinical and experimental pediatrics*, *64*(11), 588.

Tigistu, M., Azale, T., Kebebe, H., et al (2018). Frequency of seizure attack and associated factors among patients with epilepsy at University of Gondar Referral Hospital: a cross-sectional study, Gondar, North West Ethiopia, 2017. *BMC research notes, 11(1),* 652.

Triola, M. M., Tiola, M. F., Roy, J. (2006). *Biostatistics for the Biological and Health Sciences,* Pearson.

Tucker-Seeley, R.D., Li, Y., Sorensen, G., et al. (2011). Lifecourse socioeconomic circumstances and multimorbidity among older adults. *BMC Public Health, 11,* 313.

United Nations Department of Economic and Social Affairs, Population Division (2020). *World Population Ageing 2020 Highlights: Living arrangements of older persons*. https://www.un.org/development/desa/pd/news/world-population-ageing-2020-highlights

UNAIDS (1999). *Trends in HIV incidence and prevalence: natural course of the epidemic or results of behavioural change?* https://data.unaids.org/publications/irc-pub04/una99-12_trends-hiv-incidence_en.pdf

Victora, C. G., Fuchs, S. C., Flores, J. A. C., et al (1994). Risk factors for pneumonia among children in a Brazilian metropolitan area. *Pediatrics, 93(6),* 977-985.

Vila-Corcoles A., Ochoa-Gondar, O., Rodriguez-Blanco, T., et al (2009). Epidemiology of community-acquired pneumonia in older adults: A population-based study. *Respiratory Medicine, 103(2),* 309-316

Vinogradova, Y., Hippisley-Cox, J., Coupland, C. (2009). Identification of new risk factors for pneumonia: population-based case-control study. *Br J Gen Pract, 59,* e329–e338.

Walsh, C. A., Mucherino, S., Orlando, V., et al (2020). Mapping the use of Group-Based Trajectory Modelling in medication adherence research: A scoping review protocol. *HRB open research, 3*, 25.

Wang, J., Xie, H., Fisher, J. (2012). *Multilevel Models: Applications Using SAS*. Berlin: Walter de Gruyter.

Wardlaw, T. M., Johansson, E. W., Hodge, M. J. (2006). *Pneumonia: the forgotten killer of children*. Unicef.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of mathematical psychology, 44 (1),* 92–107.

Welte, T. (2011). Community-acquired pneumonia: a disease of the elderly. *Zeitschrift fur Gerontologie und Geriatrie, 44*(4), 221-228.

WHO (2010). 10 facts on gender and tobacco, https://www.who.int/gender/documents/10facts_gender_tobacco_en.pdf

WHO (2019). *Fact sheets: Pneumonia,* https://www.who.int/news-room/fact-sheets/detail/pneumonia

WHO (2020). *The top 10 causes of death*. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Williams, B., Mandrekar, J., Mandrekar, S., et al (2006). Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes. *Technical Report Series.* 79.

Wojciechowski, T. W. (2017). PTSD as a risk factor for the development of violence among juvenile offenders: a group based trajectory modeling approach. *Journal of Interpersonal* Violence. 0886260517704231.

Wong, I. T., & Worrall, J. L. (2021). Social disorganization and police arrest trajectories. *The Police Journal*. 0032258X211032116.

Xia, Y., Morrison-Beedy, D., Ma, J., et al (2012). Modeling count outcomes from HIV risk reduction interventions: A comparison of competing statistical models for count responses. *AIDS Research and Treatment,* 593569.

Yang, S., Harlow, L. I., Puggioni, G., et al (2017). A comparison of different methods of zero-inflated data analysis and an application in health surveys. *Journal of Modern Applied Statistical Methods, 16(1),* 518–543.

Yoshikawa, T. T. & Marrie, T. J. (2000). Community-Acquired Pneumonia in the Elderly, *Clinical Infectious Diseases, 31 (4),* 1066–1078.

Ziss, D. R., Stowers, A., & Feild, C. (2003). Community-acquired pneumonia: compliance with centers for Medicare and Medicaid services, national guidelines, and factors associated with outcome. *Southern medical journal*, *96*(10), 949-960.

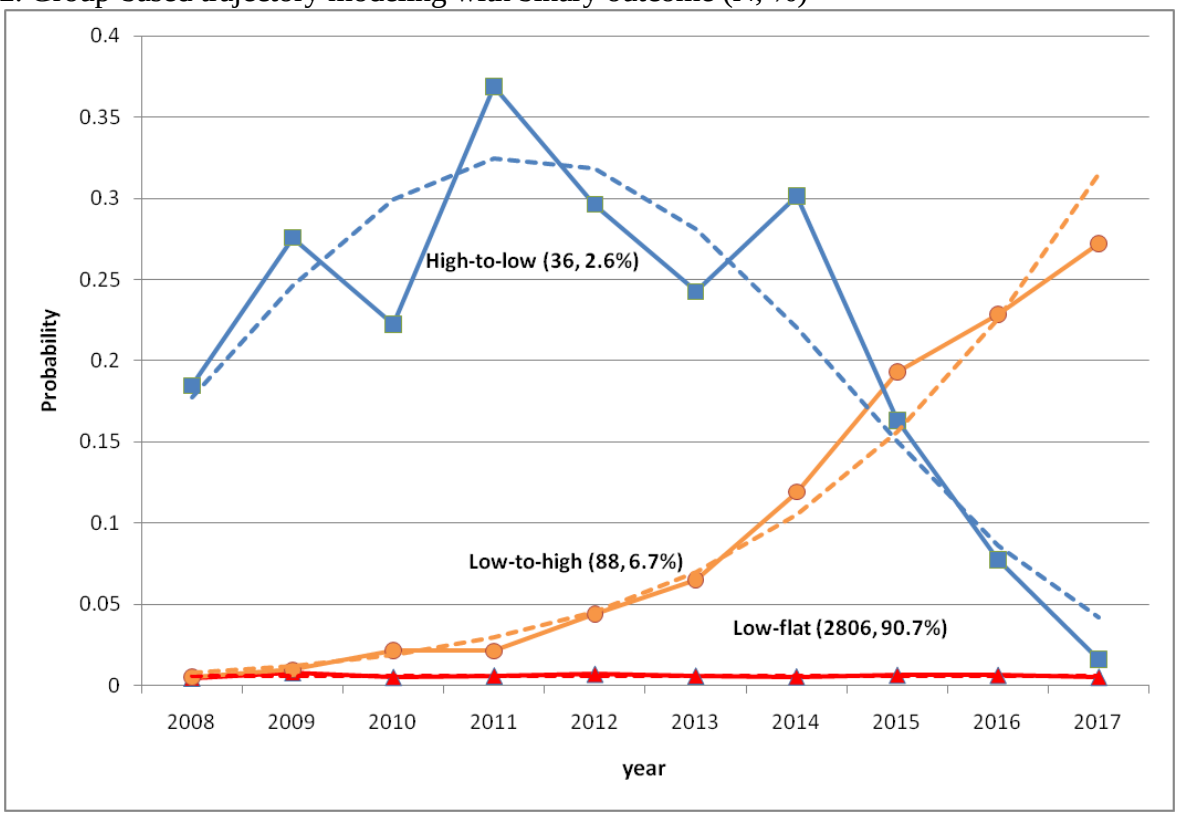
APPENDIX A: GROUP-BASED TRAJECTORY MODELING WITH BINARY OUTCOME USING DATA WITH THE STUDY PARTICIPANTS ONLY RECRUITED IN 2008

1. Goodness of model fit to select the number of trajectory groups

Number of trajectories	**BIC**	**AIC**	p_j
2	-1699.68	-1678.74	0.21
3	-1698.33	-1665.42	0.79
4	-1712.90	-1668.03	0
5	-1712.40	-1662.56	0
6	-1726.71	-1657.91	0

*BIC=Bayesian Information Criterion, AIC= Akaike's Information Criteria, p_j = probability that a model with j groups is the correct model

2. Group-based trajectory modeling with binary outcome (N, %)



3. Parameter estimates for trajectory shapes

Group	Parameter	Estimate	Standard Error	p-value
Low-flat (Group1)	Intercept	-5.123	0.1958	<.0001
Low-to-high (Group2)	Intercept	-5.327	0.8464	<.0001
	Linear	0.455	0.1157	0.0001
High-to-low (Group3)	Intercept	-2.094	0.5713	0.0002
	Linear	0.636	0.2322	0.0062
	Quadratic	-0.074	0.0239	0.0020

4. Selected variables from the univariate logistic regression analysis results with significance level of 10%. ("low-flat" group as reference group)

Low-to-high	High-to-low
Gender	Disability
Number of household members	Chronic respiratory disease
Smoking	
Chronic heart disease	
Chronic respiratory disease	

5. Multivariate logistic regression analysis. Final variables were selected at a significance level of 5%. ("low-flat" group (n=2806, 90.7%) as reference group)

Variable	Low to High (n=88, 6.7%) OR (95% CI)	p-value	High to Low (n=36, 2.6%) OR (95% CI)	p-value
# of household members				
1	-	-		
2	1.73 (0.80, 3.72)	0.1623		
3	3.29 (1.45, 7.49)	0.0045		
4	1.81 (0.62, 5.31)	0.2776		
5+	2.07 (0.82, 5.21)	0.1237		
Disability				
No			-	-
Yes			2.84 (1.38, 5.83)	0.0046
Chronic respiratory disease				
No	-	-	-	-
Yes	2.55 (1.35, 4.81)	0.0038	351 (1.51, 8.17)	0.0035

*Chronic respiratory disease: chronic obstructive pulmonary disease, bronchitis, asthma, pulmonary edema, and any other disease that was classified as 'disease of respiratory system'

APPENDIX B: SAS CODE

-Using binary outcome

Group-based trajectory modeling

```
PROC TRAJ DATA=pn.binary OUTPLOT=OP OUTSTAT=OS OUT=OF OUTEST=OE ITDETAIL;
 ID pidwon; VAR in01-in10; INDEP t01-t10;
 MODEL LOGIT; NGROUPS 3 ; ORDER 1 2 0;
 start   -3.696577     0.145958    -0.130239    -0.122261 0   -7.794095
 24.697768     1.243062    74.059170;
 RUN;
%TRAJPLOT(OP,OS)
```

Baseline characteristics by trajectory groups

```
data pn.bingroup; set of; keep pidwon group; run;
proc sort data=pn.binary; by pidwon; run;
proc sort data=pn.bingroup; by pidwon; run;
data pn.binarybygrp; merge pn.binary pn.bingroup; by pidwon; run;

proc freq data=pn.binarybygrp; tables c3*group / chisq; run;
proc univariate data=pn.binarybygrp; where group=1; var age; run;
proc univariate data=pn.binarybygrp; where group=2; var age; run;
proc univariate data=pn.binarybygrp; where group=3; var age; run;
proc freq data=pn.binarybygrp; table hou*group ; exact fisher; run;
proc freq data=pn.binarybygrp; table num*group / chisq ; run;
proc freq data=pn.binarybygrp; table smo*group / chisq ; run;
proc freq data=pn.binarybygrp; table dis*group / chisq ; run;
proc freq data=pn.binarybygrp; table inc*group ; exact fisher;  run;
proc freq data=pn.binarybygrp; table c40*group; exact fisher;  run;
proc freq data=pn.binarybygrp; table agecat*group / chisq ; run;
proc freq data=pn.binarybygrp; table mar*group / chisq ; run;
proc freq data=pn.binarybygrp; table dri*group / chisq ; run;
proc freq data=pn.binarybygrp; table c24*group / chisq ; run;
proc freq data=pn.binarybygrp; table walk*group; exact fisher;  run;
proc freq data=pn.binarybygrp; table phy*group; exact fisher; run;
proc freq data=pn.binarybygrp; table dia*group / chisq ; run;
proc freq data=pn.binarybygrp; table car*group / chisq ; run;
proc freq data=pn.binarybygrp; table res*group / chisq ; run;
```

Univariate logistic regression analysis

```
proc logistic data=pn.binarybygrp;
class group (ref='3') c3 (ref='1') / param=ref;
model group=c3 / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') agecat (ref='1') / param=ref;
model group=agecat / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') edu (ref='0') / param=ref;
model group=edu / link=glogit;run;
```

```
proc logistic data=pn.binarybygrp;
class group (ref='3') num (ref='1') / param=ref;
model group=num / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') hou (ref='1') / param=ref;
model group=hou / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') smo (ref='3') / param=ref;
model group=smo / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') dri(ref='1') / param=ref;
model group=dri / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') dis(ref='0') / param=ref;
model group=dis / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') inc(ref='1') / param=ref;
model group=inc / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') c24(ref='2') / param=ref;
model group=c24 / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') c40(ref='2') / param=ref;
model group=c40 / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') walk(ref='0') / param=ref;
model group=walk / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') phy(ref='0') / param=ref;
model group=phy / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') dia (ref='0') / param=ref;
model group=dia / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') car(ref='0') / param=ref;
model group=car / link=glogit;run;

proc logistic data=pn.binarybygrp;
class group (ref='3') res(ref='0') / param=ref;
model group=res / link=glogit;run;
```

Multicollinearity

```
proc reg data=pn.binarybygrp ;
where group in (1,3);
model group = c3  num smo res / vif ; run; quit;
```

```
proc reg data=pn.binarybygrp ;
where group in (2,3);
model group = c3  dis res / vif ; run; quit;
```

Multivariate logistic regression analysis

```
proc logistic data= pn.binarybygrp;
where group in (1,3);
class group (ref='3') c3 (ref='1') num (ref='1') smo (ref='3') res (ref='0')
/ param=ref;
model group = c3 num smo res / lackfit  ; run;

proc logistic data= pn.binarybygrp;
where group in (1,3);
class group (ref='3') num (ref='1') smo (ref='3') res (ref='0') / param=ref;
model group = num smo res / lackfit  ; run;

proc logistic data= pn.binarybygrp;
where group in (1,3);
class group (ref='3') num (ref='1')  res (ref='0') / param=ref;
model group = num  res / lackfit  ; run;

proc logistic data= pn.binarybygrp;
where group in (1,3);
class group (ref='3') num (ref='1')  res (ref='0') / param=ref;
model group = num res num*res / lackfit  ; run;

proc logistic data= pn.binarybygrp;
where group in (2,3);
class group (ref='3') c3 (ref='1') dis (ref='0') res (ref='0') / param=ref;
model group = c3 dis res /lackfit; run;

proc logistic data= pn.binarybygrp;
where group in (2,3);
class group (ref='3')  dis (ref='0') res (ref='0') / param=ref;
model group = dis res   /lackfit; run;

proc logistic data= pn.binarybygrp;
where group in (2,3);
class group (ref='3')  dis (ref='0') res (ref='0') / param=ref;
model group = dis res dis*res  /lackfit; run;
```

-Using zero-inflated count outcome

Group-based trajectory modeling

```
PROC TRAJ DATA=pn.zip2 OUTPLOT=OP OUTSTAT=OS OUT=OF OUTEST=OE ITDETAIL;
 ID pidwon; VAR in01-in10; INDEP t01-t10;
 MODEL zip; NGROUPS 3; ORDER 2 0 2 ; iorder 1; RUN;
%TRAJPLOT(OP,OS)
```

Baseline characteristics by trajectory groups

```
data pn.zipbygrp; set of; keep pidwon group; run;
proc sort data=pn.zip2; by pidwon; run;
proc sort data=pn.zipbygrp; by pidwon; run;
data pn.zipbygrp2; merge pn.zip2 pn.zipbygrp; by pidwon; run;
```

```
proc freq data=pn.zipbygrp2; tables c3*group / chisq; run;
proc univariate data=pn.zipbygrp2; where group=1; var age; run;
proc univariate data=pn.zipbygrp2; where group=2; var age; run;
proc univariate data=pn.zipbygrp2; where group=3; var age; run;
proc freq data=pn.zipbygrp2; table hou*group ; exact fisher; run;
proc freq data=pn.zipbygrp2; table num*group / chisq ; run;
proc freq data=pn.zipbygrp2; table smo*group / chisq ; run;
proc freq data=pn.zipbygrp2; table dis*group / chisq ; run;
proc freq data=pn.zipbygrp2; table inc*group / chisq ;  run;
proc freq data=pn.zipbygrp2; table c40*group / chisq ; run;
proc freq data=pn.zipbygrp2; table agecat*group / chisq ; run;
proc freq data=pn.zipbygrp2; table mar*group / chisq ; run;
proc freq data=pn.zipbygrp2; table dri*group / exact fisher ; run;
proc freq data=pn.zipbygrp2; table c24*group / chisq ; run;
proc freq data=pn.zipbygrp2; table walk*group / chisq ; run;
proc freq data=pn.zipbygrp2; table phy*group; exact fisher; run;
proc freq data=pn.zipbygrp2; table dia*group / chisq ; run;
proc freq data=pn.zipbygrp2; table car*group / chisq ; run;
proc freq data=pn.zipbygrp2; table res*group / chisq ; run;
```

Univariate logistic regression analysis

```
proc logistic data=pn.zipbygrp2;
class group (ref='2') c3 (ref='2') / param=ref;
model group=c3 / link=glogit; run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') agecat (ref='1') / param=ref;
model group=agecat / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') edu (ref='0') / param=ref;
model group=edu / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') num (ref='1') / param=ref;
model group=num / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') hou (ref='1') / param=ref;
model group=hou / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') smo (ref='3') / param=ref;
model group=smo / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') dri (ref='5') / param=ref;
model group=dri / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') dis (ref='0') / param=ref;
model group=dis / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') inc (ref='1') / param=ref;
model group=inc / link=glogit;  run;
```

```
proc logistic data=pn.zipbygrp2;
class group (ref='2') c24 (ref='2') / param=ref;
model group=c24 / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') c40 (ref='2') / param=ref;
model group=c40 / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') walk (ref='0') / param=ref;
model group=walk / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') phy(ref='0') / param=ref;
model group=phy / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') dia(ref='0') / param=ref;
model group=dia / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') car(ref='0') / param=ref;
model group=car / link=glogit;  run;

proc logistic data=pn.zipbygrp2;
class group (ref='2') res(ref='0') / param=ref;
model group=res / link=glogit;  run;
```

Multicollinearity

```
proc reg data=pn.zipbygrp2;
where group in (1,2);
model group = c3 agecat num smo dri c40 / vif ; run; quit;

proc reg data=pn.zipbygrp2;
where group in (3,2);
model group = c3 agecat  smo dis res / vif ; run; quit;
```

Multivariate logistic regression analysis

```
proc logistic data=pn.zipbygrp2;
where group in (1,2);
class group (ref='2') c3 (ref='2')  agecat (ref='1') num(ref='1')
smo(ref='3') dri (ref='1') c40(ref='2') / param=ref;
model group=c3 agecat num smo dri c40 / lackfit; run;

proc logistic data=pn.zipbygrp2;
where group in (1,2);
class group (ref='2') c3 (ref='2')  agecat (ref='1') num(ref='1')  dri
(ref='1') c40(ref='2') / param=ref;
model group=c3 agecat num  dri c40 / lackfit; run;

proc logistic data=pn.zipbygrp2;
where group in (1,2);
class group (ref='2') c3 (ref='2')  agecat (ref='1') num(ref='1')
c40(ref='2') / param=ref;
model group=c3 agecat num c40 / lackfit; run;
```

```
proc logistic data=pn.zipbygrp2;
where group in (1,2);
class group (ref='2') c3 (ref='2') agecat (ref='1') c40(ref='2') /
param=ref;
model group=c3 agecat c40 / lackfit; run;

proc logistic data=pn.zipbygrp2;
where group in (1,2);
class group (ref='2') c3 (ref='2') c40(ref='2') / param=ref;
model group=c3 c40 / lackfit; run;

proc logistic data=pn.zipbygrp2;
where group in (1,2);
class group (ref='2') c3 (ref='2')  / param=ref;
model group=c3  / lackfit; run;

proc logistic data=pn.zipbygrp2;
where group in (3,2);
class group (ref='2') c3 (ref='2')  agecat (ref='1')  smo(ref='3') dis
(ref='0') res (ref='0') / param=ref;
model group=c3 agecat  smo dis res /lackfit; run;

proc logistic data=pn.zipbygrp2;
where group in (3,2);
class group (ref='2') c3 (ref='2')  agecat (ref='1')  dis (ref='0') res
(ref='0') / param=ref;
model group=c3 agecat  dis res /lackfit; run;

proc logistic data=pn.zipbygrp2;
where group in (3,2);
class group (ref='2') c3 (ref='2')  agecat (ref='1')  res (ref='0') /
param=ref;
model group=c3 agecat res /lackfit; run;

proc logistic data=pn.zipbygrp2;
where group in (3,2);
class group (ref='2') c3 (ref='2')    res (ref='0') / param=ref;
model group=c3 res /lackfit; run;

proc logistic data=pn.zipbygrp2;
where group in (3,2);
class group (ref='2') c3 (ref='2')    res (ref='0') / param=ref;
model group=c3 res c3*res /lackfit; run;
```

- Kappa statistic for comparing groups

```
data bn; set pn.binarybygrp; keep pidwon group; run;
data zi; set pn.zipbygrp2;keep pidwon group; run;
data zi; rename group=groupz;  set zi; run;
data pn.compare; merge bn zi; by pidwon; run;
data pn.compare; set pn.compare;
if group=1 then group=5; if group=2 then group=6; if group=3 then group=4;
if groupz=1 then groupz=5; if groupz=2 then groupz=4; if groupz=3 then
groupz=6; run;
proc freq data=pn.compare;
table group*groupz / noprint agree plots=none;
ods output KappaStatistics=kappa;
run;
```