July 2021

# Healthcare Data Analytics for Predicting Health Outcomes of Older Adults and Emergency Responses of Aged Care Facilities

Nazmus Sakib
*University of South Florida*

Healthcare Data Analytics for Predicting Health Outcomes of Older Adults and Emergency Responses of Aged Care Facilities

by

Nazmus Sakib

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Industrial and Management Systems Engineering
College of Engineering
University of South Florida

Major Professor: Mingyang Li, Ph.D.
José L. Zayas-Castro, Ph.D.
Carla VandeWeerd, Ph.D.
Hongdao Meng, Ph.D.
Sriram Chellappan, Ph.D.

Date of Approval:
June 23, 2021

Keywords: Data-driven Modeling, Applied Statistics, Cognitive Trajectory Modeling, Length-of-Stay Prediction, Nursing Home Evacuation

**Dedication**

To my parents, Masuda Begum and Md. Mustafizur Rahman Akhand, for your constant love and immense hard work that allowed me to come this far.

To my brother, Sadat, for being my source of strength, hope, and wonder.

To my wife, Nusrat, for your boundless love and relentless support that inspire me to dream bigger for the brightest future for me, you, and everyone.

## Acknowledgments

I would like to express my wholehearted gratitude towards many individuals without whose support this dissertation would not have been possible.

First of all, I am grateful to my major advisor Dr. Mingyang Li. Everything I have learnt in my PhD work is because of him. He allowed me the opportunity to work under his supervision and support, patiently taught me the most difficult concepts, provided hope when I was out of inspiration, and carefully guided and nurtured my growth. It has been an absolute honor to have Dr. Li as my mentor.

I am also grateful to my advisory committee members. Dr. José L. Zayas-Castro, Dr. Carla VandeWeerd, Dr. Hongdao Meng, and Dr. Sriram Chellappan have guided this research effort with crucial domain knowledge and constructive criticism of the results. Moreover, I thank Dr. Kathryn Hyer, who generously supported the research and showed me how analytical tools can be used for reducing risk of the most vulnerable older adults. I am further thankful to Dr. Tapas K. Das for granting me a place in the PhD program and mentoring me to be a better teacher. I also thank Dr. Zayas-Castro for his overall guidance and kind support during my difficult times.

Finally, my deepest gratitude goes to my family and friends, whose limitless love, care, and support allowed me to pursue my dreams and made this dissertation possible.

This dissertation work has been partly supported by the National Science Foundation under Grant No. CMMI-1825761 and the National Institutes of Health under Grant No. R01AG060581.

## Table of Contents

## List of Tables

## List of Figures

**Abstract**

The United States (US) is experiencing rapid growth in its older adult population, who may suffer from multiple chronic diseases, injuries, and impairments. To meet with the excess demand without compromising the quality of care for older adults, the current aged care systems, such as nursing home systems, will face unprecedented challenges of healthcare resource shortage with rising costs. Accurate prediction of health outcomes of individual older adults will facilitate the aged care professionals to better prioritize healthcare resources for the most at-risk individuals with more focused care and provide more proactive and individualized treatment and care delivery. In addition, since older adults are highly vulnerable to natural disasters, accurate prediction of emergency response of aged care facilities, such as evacuation response of nursing homes during hurricane, will further facilitate the emergency operations agencies to provide more proactive and targeted support and coordination with adequate resources to ensure the health outcomes of older adults during hazard scenarios. In this dissertation, a series of healthcare data analytics models, algorithms, and tools are developed to improve modeling and prediction of individual older adults' health outcomes as well as aged care systems' emergency responses. First, a bi-level longitudinal data modeling approach is proposed to characterize the heterogeneous degradation of cognitive performance outcomes among community-dwelling older adults at both sub-population level and individual levels. The proposed model comprehensively investigates both the temporal heterogeneity at sub-population and individual levels with relaxed statistical assumptions and improved prediction accuracy. Second, a discharge outcomes prediction model is proposed to characterize the

heterogeneous length of stays of post-acute care residents in the nursing home with multiple and competing discharge dispositions. The developed modeling approach allows accurate prediction of re/hospitalization risk and community discharge likelihood over time of individual post-acute care residents with various individual risk factors identified and quantified. Third, a GIS data integrated predictive analytics approach is proposed to improve the prediction of nursing home evacuation response under natural disaster scenarios of hurricane. The proposed work improves the prediction accuracy of evacuation response by integrating spatial and temporal rich storm characteristics information with varied facility characteristics and resident characteristics of nursing home facilities at different spatial locations. Real case studies are considered to illustrate the proposed modeling approaches and demonstrate their superior performances over existing benchmark models.

# Chapter 1
# Introduction

## 1.1 Background

### 1.1.1 Population Aging and Health of Older Adults

The world is rapidly aging. We are currently experiencing the highest rate of aging in history, both in absolute numbers and as a proportion of the general population. According to 2015 estimates, people aged 60 or more had a population of 760 million, amounting to approximately 11% of the total world population of over 7 billion (Bloom *et al.*, 2016). By 2050, this number is projected to reach 2 billion, about 18% of the entire world population of estimated 11 billion. The older population is more pronounced in developed countries compared to developing ones. In the United States, the approximate number of older adults (referring to the people aged 65+ from hereon in this dissertation) in 2012 was 43.1 million and projected to nearly double to 83.7 million in 2050 (Ortman *et al.*, 2014). Such inflation of the older adults cohort has been occurring in the past few decades because of three primary reasons. First, longevity of humans has increased due to the better standard of living, the improved public health infrastructure and policy implementation, and the increased public awareness on nutrition and diseases (Fogel, 1997; Preston, 1975; Lleras-Muney, 2005) . Second, fertility has declined over the years due to the diminished desire for children owing to financial and economic costs (Galor *et al.*, 2000). Third, baby boomers, i.e., people born between 1946 through 1964, aged and began entering the older adults cohort in 2011, marking a significant rise in the proportion of the cohort compared with general

population (He *et al.*, 2014). The unprecedented change in population demographics presents an ever-growing demand to healthcare systems caring for the older adults, who are by nature frail and vulnerable due to the aging process.

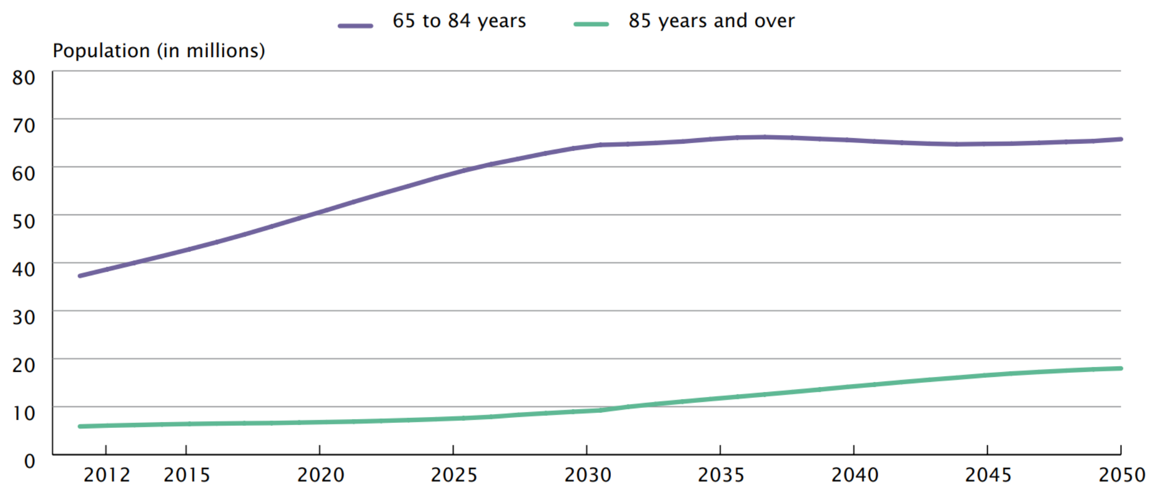



Figure 1.1: U.S. population trend by age group of older adults. Source: Public domain image from Ortman *et al.* (2014)

Due the the rapid population aging, the health of older adults in the U.S. has become a critical concern. Between 2008-2012, it was reported that about 15.7 million, or 38.7%, of all older adults in the U.S. certain types of disabilities, such as mobility challenges, difficulty in hearing and vision, inability of self-care and independent living, and cognitive limitations(He *et al.*, 2014). Furthermore, older adults are highly susceptible towards having different kinds of diseases and multiple chronic conditions (e.g. heart disease, cancer, diabetes, hypertension, arthritis, etc.) (Anderson *et al.*, 2020; Pearson *et al.*, 2011). In 2011, 80% of the older adult population had at least one chronic condition and about 68% of all Medicare beneficiaries in 2010 had at least two or more chronic conditions (CDC, 2011; CMS, 2012). The health complications contribute to multi-dimensional functional limitations and escalates with age leading to disability, and increase risk of critical events, such as falls or stroke. Most older adults thus require some form of healthcare services, such as personal assistance in daily living activities, specialized care and treatment of multiple chronic conditions, different forms of physical, cognitive, and occupational

therapies, and/or clinical interventions associated with emergency events. Provision of such services in the current healthcare system is a significant challenge in terms of managing adequate resources, delivering quality care, and reducing costs, exacerbated by workforce shortage in healthcare professionals and limited public financing (Janiszewski Goodin *et al.*, 2003). Thus, the U.S. healthcare system needs to be better prepared to meet the increasing demand and diverse needs of frail and vulnerable older adults to increase life span without sacrificing their quality of life .

### 1.1.2 Healthcare Systems for Older Adults

Older adults' healthcare needs are highly heterogeneous due to different disease diagnoses, multi-functional (physical, mental, cognitive) performance limitations, prevalence of one or more chronic conditions, and somatic disorders. The range, extent, and quantity of care demand has spawned a complex system of care settings, each providing care differing in service type, staff, equipment, duration, delivery location, etc. Healthcare systems for older adults is comprised of a broad spectrum of care settings. The heterogeneous care settings can be loosely grouped, including community-based setting, acute care setting, and long-term care setting based on the acuity of an older adult. Over the course of his/her life, an older adult can be a consumer of multiple settings, transferring between facilities according to his/her changing health condition and healthcare needs.

Figure 1.2 shows an overview of the care journey of an older adult. An older adult generally prefers aging in place and remaining his/her living at home or in the community, rather than being placed in an institutional setting, such as the nursing home. Community-based setting constitutes various forms of informal care that facilitates aging at home or a non-institutional residence, depending on the intensity of care required. Relatively functionally independent older adults would stay at their home caring for themselves, or rely on informal caregivers, such as family members or friends, for light assistance with daily activities. In absence of family and/or for those who require more assistance which

are available at home, home health agencies provide necessary nursing, care management, and therapy services for limited time of the day. If supervisory support is needed for the entire day, adult day care agencies supply caregivers to aid functionally dependent individuals for extended hours in the day, also acting as a relief for primary informal caregivers at home. Further available at home is hospice service, where volunteers provide various supports, such as emotional support and daily living assistance, to terminally ill older adults (Pratt, 2010; Harris-Kojetin *et al.*, 2019). The community-based setting also includes senior living communities and retirement homes, which accommodate individuals of similar age and offer greater health monitoring, socialization and security services and support than those provided by informal caregivers, while still maintaining the home atmosphere.

In contrast with the residential environment of the home and community-based settings, the acute care setting is comprised of facilities which address healthcare needs of individuals characterized by severe illnesses requiring the most specialized form of treatments and care services from medical professionals. A distinct characteristic of the acute care setting is the element of time-sensitiveness and rapid intervention. Recipients of acute care are usually short-stay patients who have experienced significant injury or critical degradation of illness that requires intense and immediate care response (Hirshon *et al.*, 2013). The acute care setting includes, but not limited to, hospitals, urgent care, and surgery centers. Hospitals respond to urgent health events at their emergency department (ER), treat short-stay inpatients with constant close monitoring from nurses, doctors and other health practitioners, and can provide different forms of technologically modern and specialized diagnostic, trauma, surgery, stabilizing and/or rehabilitative services. As opposed to ERs treating lethal injuries, urgent care facilities treat outpatients with a less serious injury requiring immediate attention by a healthcare professional available outside business hours (Mount Sinai Hospital, 2021). Ambulatory surgery centers, on the contrary, provide advanced operating theatres for less extensive procedures compared to those

performed at hospitals that can be completed within a day for non-complex outpatients (Texas Medical Management, 2020).

In addition to the community based settings and acute care settings which provide care services for older adults with lower and higher acuity, another important class of health setting, namely the long-term care (LTC) setting, provides LTC support for older adults with medium level of acuity. Historically, LTC setting was defined as institution-based, although recently some forms of non-institution-based support services (e.g. home health agency) provided at the community are also considered under the umbrella of the LTC setting. Traditional institution-based LTC setting includes nursing homes (NH) and assisted living facilities (ALFs) which differ in the intensity of care. NH offers 24/7 specialized care for post-acute residents (e.g. individuals who are recently discharged from hospitals and require short-term rehabilitation for functional recovery), or long-stay residents who remain in the NH for years or the rest of their lives to receive personal assistance to postpone and delay their functional decline. The nursing staff and health practitioners caring for the frail older adults at NHs are specially trained to mitigate health issues of clinically complex NH residents. The services offered in a NH also include various restorative and rehabilitative services, assistance with activities of daily living, and end-of life care. Infrequently, NHs also offer special care services for specific diagnoses, such as Alzheimer's, mental disabilities, AIDS, etc (Pratt, 2010). ALFs bridge the gap between NHs and community-based settings. On one side, they provide residential institutional care for individuals who are more independent and less acute than those residing at a NH. On the other side, they provide residents with more home-like services and care to ensure their privacy and comfort. ALFs provide assistance with activities of daily living (i.e., eating, toileting, bathing), meals, medication management, social services, and closer observation in case of emergencies.

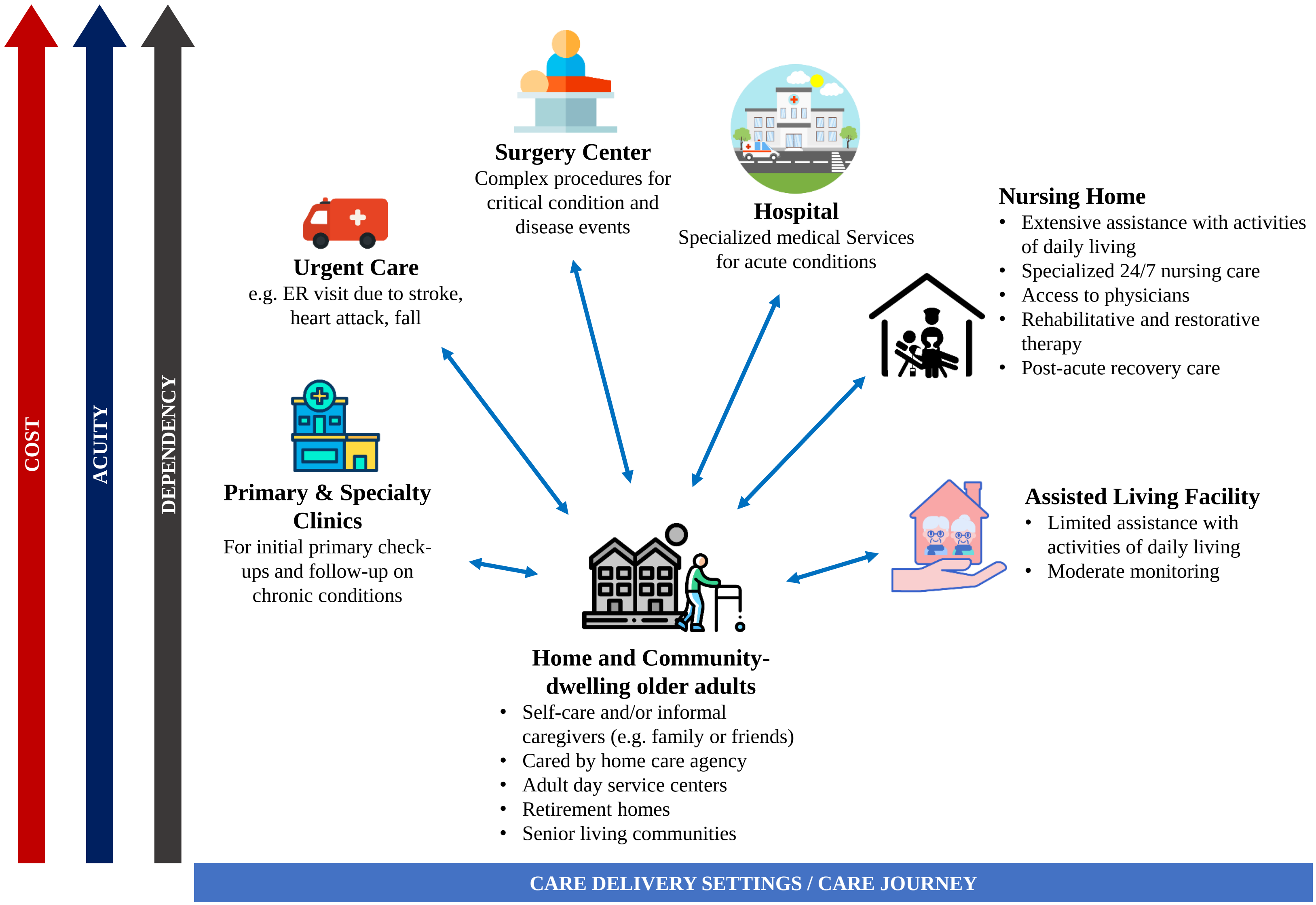



Figure 1.2: Care delivery settings for older adults

### 1.1.3 Healthcare Data Analytics

Fulfilling the rising demand of diverse healthcare needs of older adults and ensuring quality care delivery and treatment require multidimensional solution approach by the healthcare industry. A few avenues for achieving better health outcomes of older adults include techniques such as increasing service capacity, training and hiring skilled staff, strengthening policy and regulations at private and public levels, optimizing costs for better resource allocation, etc. In all of the approaches, incorporating and analyzing information collected by facilities at various granularity levels, e.g., data on residents, caregivers, facilities, etc., can play crucial role in informing better clinical and organizational decisions.

As the healthcare system and facilities evolved technologically over time, the care environment has increasingly become digitized. Higher data storage capacity and development of enhanced medical equipment and computational devices have allowed rapid generation of large volumes of complex healthcare data in diverse forms. For instance, Electronic Health Records (EHR) store information of care recipients at the healthcare facilities and over time. The EHR contain rich clinical information from physician's notes and prescriptions, test results, medical images, and medication history. More technologically advanced forms of data include human genomics data and sensor-generated data, e.g., from vital signs monitors, wearable devices, or physiological sensors. Administrative and financial data is available from insurance claims records. Social media and sentiment data, and journal publications related to health can also be considered valuable unstructured information relevant to care delivery (Raghupathi *et al.*, 2014; Kankanhalli *et al.*, 2016). Facility-related data such as service schedules, staffing allocation, and resource management, are further available from day-to-day operations management.

Leveraging the vast collection of raw data effectively requires substantial data processing, so that useful and explainable insights are obtained to assist health practitioners and administrators make clinical, operational, and business decisions. Healthcare data analyt-

ics becomes an important tool in this effort, where the data is systematically examined with application of various methodologies, such as, statistical, contextual, quantitative, predictive, and other models, to uncover important patterns, associations, and trends within the data (Cortada *et al.*, 2012; Raghupathi *et al.*, 2014). Avoiding the pitfalls of subjectivity, an evidence-based analytics-driven approach directs a smarter decision-making process and has proved to improve health and business outcomes across numerous scenarios. For example, health characteristics in EHR data is analyzed to characterize individual acuity and and segment groups who are at more risk, allowing targeted customization of clinical treatments and regimens (Hersh, 2014; Vuik *et al.*, 2016). Insurance claims data is analyzed for accuracy and consistency to identify current and prevent future fraudulent transactions (Travaille *et al.*, 2011; Dora *et al.*, 2015). Imaging data, e.g., MRI, PET, sMR scans, is analyzed to automate early disease diagnoses which are difficult to detect with human perception (Vemuri *et al.*, 2016; Yadav *et al.*, 2019). Preventive medical intervention can be employed by analyzing sensor data to accurately detect underlying diseases, e.g, using electro-cardiogram (ECG) signals to detect silent heart attacks (Acharya *et al.*, 2017). Real-time physiological data from vital signs monitors in Intensive Care Units (ICUs) of hospitals can be used to anticipate catastrophic complications early, e.g, hypotensive episodes, sepsis, etc., and administer appropriate medical response (Sow *et al.*, 2013). Historical administrative care utilization data, e.g., Length of Stay (LOS), surgery scheduling, is analyzed to determine the best resource management and staffing strategy which would improve care delivery and decreases wastes (Carey, 2002; Beliën *et al.*, 2006). On a more grander scale, evidence-based public health policy formulation, e.g., on pandemic or hurricane preparedness, can be guided by results from computer simulations based on real population health data which would potentially save numerous lives in the older adults and general population (Chen *et al.*, 2006; Squazzoni *et al.*, 2020).

## 1.2 Literature Review

This section will provide a comprehensive overview on the existing healthcare data analytics literature for older adults. Healthcare data analytics can be broadly classified into three categories, namely descriptive analytics, predictive analytics, and prescriptive analytics. Each category will be introduced in the following sections with emphasis on the application of relevant methods in analyzing data of older adults. The relevant research gaps will be also identified to better motivate the contributions of the proposed work described in Section 1.3.

### 1.2.1 Descriptive Healthcare Analytics for Older Adults

Exploration of the dataset to extract meaningful information in order to answer questions such as "What happened?" or "What is happening?" is called descriptive analytics. It consists of application of methodologies that summarizes the dataset and tries to discover important patterns, associations and correlations that may contribute to the evidence-base for the decision-maker (Islam *et al.*, 2018). Descriptive analytics can explain past failures, successes, and historical trends in the service delivery performance and guide focused improvement by identifying their root causes (Ursprung *et al.*, 2010). A significant portion of literature on older adults healthcare are exploratory and use descriptive analytics as the main technique to analyze the data. These studies use one of the following descriptive analytics methods, such as conceptual modeling, graphical visualization and summary statistics. They are often focused on investigating the older adults at a specific settings such as community-dwelling older adults, NH residents and hospital patients. Different diseases and chronic conditions of older adults are also investigated, such as heart disease, diabetes, Alzheimer’s disease, etc.

In conceptual modeling, a real-life system being analyzed is characterized and graphically represented by abstract entities that explain information flow, interactions and

interrelationships within the system. Conceptualization can uncover general influencing factors influencing a health or service outcome, their consequences, and guide the data collection process. For example, Campbell *et al.* (2007) recommends using conceptual models to simplify designing medical interventions for healthcare problems. It rationalizes, for example, that to reduce death rates among people with cardiovascular disease, the patient's disease history, lifestyle choices, care delivery facility, access to care, adherence to public health policy, and socio-demographics can be identified through conceptualization as relevant information/data/evidence to investigate. Similarly, Raina *et al.* (2004) proposed a conceptual model to explain stress outcome of geriatric (and pediatric) caregivers with components, such as care receiver's characteristics, intrapsychic factors, and coping mechanisms. The model eventually guided a roadmap for collection of cross-sectional longitudinal data and subsequent data analysis with Structural Equation Model (SEM) to explain association between the factors with outcome (Hayduk, 1987). Similarly, Hoenig *et al.* (1997) develops individualized geriatric disability rehabilitation plans by establishing a conceptual framework for underlying factors of disability, such as health conditions, and physical and social environment.

In addition to conceptual modeling, many of existing data analysis work in older adults healthcare mainly consider the descriptive statistics to numerically summarize the data (Morr *et al.*, 2019). For example central tendency measures, such as mean, median and mode, and dispersion measures such as minimum, maximum, range, percentiles and variance, are mainly considered to characterize quantitative data, and frequencies and proportions are used to characterize qualitative data (Sims *et al.*, 2018). Furthermore, different types of statistical tests are applied to explain the data, such as, checking the correlation to measure degree of association between the data variables, or, fitting the probability distributions (e.g., Normal distribution, binomial distribution, etc.) on specific data variable and checking their corresponding goodness-of-fit (Cheng *et al.*, 2020; Sharma, 1994). T-tests and analysis of variance (ANOVA) are also used for investigating the

difference between two groups or multiple groups, respectively. Chi-square tests are used to investigate the independence of categorical variables (Gould *et al.*, 2016; Latham *et al.*, 2014; Bravo-José *et al.*, 2018; Shumway-Cook *et al.*, 1997).

Visualization tools are also widely considered in existing healthcare literature for older adults to provide initial visual assessment of the data. Histograms are used to illustrate distribution of the data, and various charts, e.g., pie charts, bar plots, can indicate further segmentation of the dataset. Time-series plots are also generated if the interest is to observe temporal patterns (Lin *et al.*, 2012; Strunk *et al.*, 2006). Moreover, advanced analysis are infrequently applied, such as cause-and-effect or cluster analysis (Liao *et al.*, 2016), which further contribute to the knowledge base and informed decisions. The procedures applied in descriptive analytics also has the benefit of obtaining cleaned and formatted data which can be used in subsequent data analytics steps, such as predictive or prescriptive analytics.

### 1.2.2 Predictive Healthcare Analytics for Older Adults

Although current literature is rich in use of descriptive analytics, it is limited in informing the decision-making process as it provides only retrospective insight that qualitatively informs about the future. Predictive analytics is the next step where data about future events can be forecasted with adequate accuracy, allowing the decision-maker to make better quantitative decisions. In the healthcare analytics literature for older adults, predictive analytics methods have been considered by the healthcare researchers and practitioners in modeling different types of performance data, such as trajectory data and time to critical event data, and further predicting outcomes at different levels, such as individual level and organizational level. Since the scope of this dissertation is within the predictive analytics, a review of the current literature will be given in three subsections, namely, modeling individual level trajectory performance, modeling individual level time-to-event performance or outcomes, and modeling organizational performance at facility level.

#### 1.2.2.1 *Individual-level Trajectory Performance Modeling of Older Adults*

At the individual level, the ability to predict health status of the heterogeneous older adult is important to health practitioners in determining effective medical interventions, rehabilitation plan, and individualized treatment and achieve improved health outcomes. On the other hand, for care delivery administrators, determination of health outcomes of care recipients facilitates optimal resource management and ensure quality service delivery and reduced costs at the facility level. One of many indicators of health condition of older adults is their functional performance scores, which measure an individual's physical, cognitive, and mental abilities to perform activities of daily living (Bonder *et al.*, 2017). Functional performance of older adults change over time as their health conditions change, and the trajectories are highly heterogeneous since an individual's health characteristics is determined by multiple variable factors, such as, diseases, chronic conditions, history of critical events, and, social and physical environment. Characterizing the heterogeneity and predicting future of functional performance trajectory of older adults becomes greatly important in understanding the vulnerable population, their care demand, and relevant decision-making process in care service delivery. Current literature for modeling and predicting trajectory performance of individual older adults can be summarized as follows with their major limitations identified as well.

First, many studies in trajectory modeling assume a homogeneous population based on certain subjective knowledge. For example, Garber *et al.* (2002); Ge *et al.* (2001) proposed a quadratic pattern to characterize trajectories of depressive symptoms (among adolescents). (Blanchard *et al.*, 1996) reported that rate of Post-Traumatic Stress Disorder (PTSD) symptoms among Gulf War veterans decreased over time, while (Southwick *et al.*, 1995) reported that the rate increased over time. There is a need to relax the simplified assumption of homogeneous population and investigate the temporal heterogeneity of trajectory performance data between different sub-populations from the overall heterogeneous population and/or individual differences within each sub-population.

Second, in the existing trajectory modeling literature which considered the heterogeneous population, they either consider heterogeneity at the sub-population level, or at the individual level, but not both and thus there is a lack of literature in comprehensively investigating heterogeneity at both levels. Specifically, Lunney *et al.* (2003) studied the functional measure called Activities of Daily Living (ADL) Score (Katz *et al.*, 1963) of 14456 community-based older adults across the US over time between 1981-1987. They assumed four theoretical sub-groups in the sample, namely those who experienced sudden death, died due to cancer, died due to organ failure, and died due to frailty, and observed characteristically different trajectory patterns for each. Ferrucci *et al.* (1996) also studied ADL scores of 6640 older adults in the US over 7 years and observed three distinct sub-populations of characteristic trajectories, namely those who developed regular, progressive, and catastrophic functional limitations over time. Both of these works assumed the number of sub-populations in a subjective manner led by visual observation, qualitative literature, or domain knowledge instead of determining the sub-population number objectively from the data. Subjectivity in setting the rules of categorization of trajectories can be troublesome because: (i) there is no guarantee the presumed categories actually exist or are correct; and (ii) the categories can lead to failure in identifying important outlier patterns which are supposed to be categories on their own, i.e., the model may overfit or underfit the data (Nagin *et al.*, 2010).

Third, many of existing heterogeneous trajectory modeling approaches also have restrictive modeling assumptions. A popular alternative to subjective categorization in quantifying sub-population heterogeneity is group-based trajectory modeling (GBTM; Nagin (2005)). In GBTM, the data trajectory is modeled by a finite mixture model, and the number of sub-populations is determined in a two-steps fashion. First different models with pre-specified number of sub-populations are built. Then, the best number of sub-populations is determined by certain types of model selection criteria, such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and entropy, or certain

hypothesis test, such as Lo-Mendell-Rubin Likelihood Ration Test (LMR-LRT) (Dodge *et al.*, 2006; Gill *et al.*, 2010; Liang *et al.*, 2011). The incorporation of the two-step procedure in model estimation and selection is computationally inconsistent. A similar modeling approach with comparable issues, consists of modeling the trajectory data by Growth Mixture Modeling (GMM; (Muthén *et al.*, 1999)), where a certain number of sub-population of trajectories is presumed to exist in the population, but each with its own finite mixture model. Both models have drawbacks as they initiate modeling with a set range of the number of sub-populations that have to be tested, which is influenced further influenced by subjectivity. Moreover, in characterizing heterogeneity at the individual level, most common approaches include hierarchical modeling (Bryk *et al.*, 1987; Goldstein, 2011) and latent curve analysis (McArdle *et al.*, 1987; Meredith *et al.*, 1990), referred together as Growth Curve Models (GCM) (Taylor *et al.*, 2004). GCM assumes there is no sub-population in the data sample, and that individual trajectory is defined by variablity from the population mean trajectory based on a set of parameters continuously distributed according to multivariate normal distribution. Such assumption is restrictive in assuming that the population is directly composed of individuals with different trajectories without investigating the sub-population heterogeneity patterns among individual trajectories.

#### *1.2.2.2 Individual-level Critical Event Modeling of Older Adults*

In addition to modeling and predicting individual-level functional performance trajectories, it is also important for health practitioners and facility administrators to anticipate or predict adverse events of the older adults receiving care so that more at-risk individuals can be identified with more targeted care provided. For an acute care setting, such as the hospital, an often studied example of critical event is the readmission or death of the patient within 30- or 60-days of discharge. For a long-term care setting, such as a NH, critical events may refer to hospitalization within 30- or 60-days for short-stay nursing home residents and functional decline for long-stay nursing home residents. These

individual-level critical event measures are often aggregated towards the facility-level to serve as certain quality indicators of the overall facility and becomes the major quality improvement indices considered by government agencies. For instance, the Centers for Medicare and Medicaid Services (CMS) began imposing financial penalties to hospitals with high 30-day readmission rates and NHs with high 30-day hospitalization rates, in 2012 and 2018 respectively (Joynt *et al.*, 2013; Castellucci, 2018). Hence, the ability to accurately predict the individualized risk of critical event becomes important since it would allow the healthcare and facility administrators proactively realign and organize clinical, medical and nursing resources to identify and serve the individuals most at risk, thereby achieving better health outcomes. However, current studies in developing risk prediction models for critical events among older adults have several limitations as follows.

First, many studies in literature are focused on quantification of the risk only at a fixed time period. Kansagara *et al.* (2011) observed that the most common outcome measure in hospital context was 30-day readmission risk and their prediction performances were generally poor. Moreover, Maali *et al.* (2018) used various advanced predictive models to estimate and validate 7-day, 30-day, and 60-day hospital readmission and emphasized that discretization of a continuous measure, i.e., time to the critical event, may have adversely affected prediction performance. Anderson *et al.* (1985) studied over 270000 Medicare beneficiaries to identify 60-days readmission risk factors and further discussed that the choice of 60-days is somewhat arbitrary and the findings can be extended to 5-, 30- or even 365-days. Considering the limitations, avoiding discretization and modeling time to the critical event, such as time-to-rehospitalization or time-to-discharge, in continuous scale directly would provide more robust results.

Second, a majority of studies in predicting critical events/outcomes is based on the acute care settings, e.g., the hospital, and the relevant studies in LTC settings, e.g., NH, is relatively lacking in quantity, the depth of the developed analytics methodologies as well as the comprehensiveness of the influencing factors investigated. Since the NH has

increasingly been considered as a key healthcare setting for caring post-acute care (PAC) older adults (Fashaw *et al.*, 2020) and the majority of hospital readmission is composed of resident from NHs (Kripalani *et al.*, 2014), it is necessary to understand and predict discharge outcomes (e.g., rehospitalization, community discharge) of this segment. A recent review of 77 hospital readmission risk studies by Artetxe *et al.* (2018) presented a prolific research climate, where investigators used techniques ranging from basic statistics to sophisticated machine learning models for prediction. In contrast, a review of current literature on NH discharge to hospital by Ågotnes *et al.* (2016) noted that studies have converged into a narrow field of overused approaches. Specifically, most literature are focused on exploring data retrospectively and failed to investigate potential future re/hospitalization. Further, there is a need to incorporate more facility, staffing, and day-to-day operations characteristics to improve estimation of discharge risk and quantify their influence in characterizing the risks (Miller *et al.*, 2000).

Third, there is lack of analytics methods in modeling the time-to-event quantities, such as length of stay (LOS), of PAC residents in the NH context, and the existing LOS modeling approaches cannot adequately address the complexities of LOS data of PAC residents. PAC residents are admitted from the hospital to the NH for recovery and rehabilitation, and form the majority of NH resident census. They require intensive specialized care and extensive resources, driving the service demand and resource utilization in the overall facility. The main goal of the NH is to return the resident back to community in better health without any critical event occurred. Re/hospitalization or longer LOS often imply non-improving health outcome, financially burdensome, and is detrimental to facility quality indicators. Thus, the multiple and correlated discharge outcomes of PAC residents, such as re/hospitalization and community discharge, with various influencing factors of individuals quantified will be desirable. Existing LOS modeling approaches have one or more of the following challenges and limitations described below.

Many studies in the literature modeled LOS data with distributions, such as exponential, Weibull, lognormal, gamma, etc. For instance, Xie *et al.* (2005) characterized data of 889 individuals residing and transitioning between residential and NH care in London for 4 years with a Markov phase type model, assuming that the LOS of residents in NH were distributed exponentially as a mixture of PAC and long-stay sub-groups. In a similar work, Faddy *et al.* (2009) modeled 1901 patients of a hospital in Australia with exponential, gamma, and lognormal models and compared model fit. Both works attempted to capture heavy right-skewness of the LOS distribution with fully parametric models. A semi-parametric or non-parametric approach would capture the distribution better. Furthermore, in distribution-based studies, none or very limited resident characteristics are incorporated. Heterogeneity of LOS is dependent on various functional performance limitations, diseases, and chronic conditions of the individual, and there is a need to consider as many characteristics as possible in data modeling to improve prediction accuracy.

Many other studies also considered linear regression-based approaches. For example, Carey (2002) applied hierarchical linear model to data of over 360,000 patients of US community hospitals to predict LOS and corresponding costs associating with numerous individual characteristics. Kelly *et al.* (2010) applied multiple linear regression to predict LOS of 1817 US NH residents data from the Health and Retirement Study (HRS) and found association of LOS with social and clinical factors. Lee *et al.* (2003) modeled LOS and discharge data of 560 emergency department (ED) patients in Western Australia with linear mixed regression. Alternatively, many studies considered data mining based approaches to capture nonlinear relationship between LOS, discharge disposition and influential factors. For instance, Liu *et al.* (2006) used decision trees and Naive Bayes Classifiers to model LOS data of 4722 geriatric patients of a hospital in UK. Combes *et al.* (2014) used several tree-based ensemble models, support vector machines (SVM), and neural networks, to predict LOS of over 12,000 pediatric emergency units in France. Similarly, Turgeman *et al.* (2017) applied tree-based and SVM models on data of 4840 patients to a veterans'

hospital in US to predict their LOS. Although regression and data mining approaches incorporate considerable heterogeneous influencing factors, they ignored incorporating multiple discharge dispositions during prediction, mostly focusing on single disposition only. It is important simultaneously consider the multiple and competing discharge dispositions, such as hospital and community of PAC residents in the presence of varied individual characteristics that may potentially affect the LOSs of PACs.

#### *1.2.2.3 Facility-level Performance Modeling of Aged Care Systems Under Normal Operating Conditions*

In addition to predicting individual-level performance outcomes, it is necessary to predict facility-level performance outcomes, such as the average LOS, bed utilization, and overall discharge disposition proportions of the resident cohort. Ability to predict such outcomes may help improve resource planning and care delivery decisions, such as optimizing care delivery according to required service times, allocation of appropriate nursing staff levels, setting cost reduction projects, etc., and evaluating facility-level quality outcomes, such as proportion of residents re/hospitalized or discharged to community, etc, under the care delivery decisions. In the current healthcare systems literature, computer-based simulation is a popular approach in this regard, but application to older adults healthcare is still relatively limited. In computer-based simulation, user-interactive software platforms are developed to virtually imitate operation of the facility and to facilitate prediction and evaluation of aggregated facility-level performance outcomes. Different computer simulation methods, such as discrete event simulation and agent-based simulation, underlie the software design which are used to generate cohorts of virtual residents/patients and simulate their care demand and resource consumption.

In discrete event simulation (DES), the residents/patients are characterized as independent entities who go through a series of "events" occurring over time while consuming facility resources (Allen *et al.*, 2015). For an example in older adults healthcare, El-Darzi *et*

*al.* (1998) used DES to model the flow of residents in a U.K. hospital geriatric department through different sequential stages, namely short-stay, medium-stay, and long-stay, to forecast average LOS of the patients, discharge rates, and bed utilization in each stage. In another example, Schnelle *et al.* (2016) used DES to model resident care utilization for daily-living activities, predicted corresponding care workload, and calculated discrepancies in required nurse aide staffing in NHs in the U.S. DES is also widely applied in the hospital setting to investigate operational policies and "what-if" scenarios to improve various performance outcomes, e.g., reducing LOS in the ED (Wang *et al.*, 2012), minimizing overcrowding and optimizing bed utilization in a general hospital (Holm *et al.*, 2013), and predicting staffing levels in a neonatal unit (DeRienzo *et al.*, 2017).

On the other hand, in agent-based simulation (ABS), residents/patients are modeled as dynamic autonomous entities (agents) who can interact with each other and the virtual environment, and capable of learning and making their own behavioral decisions, such that holistic patterns in the facility can emerge and be observed (Chan *et al.*, 2010). For instance, Cui *et al.* (2018) developed a simulation platform based an a real NH and modeled its residents as agents possessing different emotional states and personal characteristics, and capable of movement and social interaction. At the facility-level, the simulation enabled evaluation of overall movement densities in the NH layout, reflecting physical well-being of residents, and identified residents in risk of psychological afflictions. Similarly, numerous studies in the hospital setting have also used ABS to evaluate ED performance, determine medical staff scheduling in different departments, and prevent of infection-spread in the facility (Gul *et al.*, 2015; Stainsby *et al.*, 2009; Friesen *et al.*, 2014).

An essential element in developing an effective simulation platform is to simulate resident care utilization through accurate prediction of their LOSs and discharge outcomes. For the simulation to run, cohorts of virtual residents (entities) must be generated and their LOSs and discharge dispositions need to be simulated by an accurate sampling algorithm driven by real data. Existing approaches in developing sampling algorithms

have several limitations, as follow. First, many studies model LOS only with simple distributions. For example, New *et al.* (2015) simulated LOS of rehabilitation patients in an Australian hospital by sampling from multi-stage PERT distributions. McGuire (1994) sampled times spent by patients in different stages in a hospital ED using Weibull, log-logistic, and Beta distributions. El-Darzi *et al.* (1998) sampled LOS of patients of a geriatric hospital department using exponential and log-normal distributions. These studies ignore heterogeneity in LOS by failing to consider the multiple factors of the residents/patients, e.g, socio-demographic characteristics, disease diagnoses, chronic conditions, etc., that may influence their LOS. Second, some studies model LOS with regression-based methods considering limited heterogeneous characteristics, e.g, Austin *et al.* (2002) used multiple linear regression to fit and sample LOS of post-operative CABG patients in a hospital in Ontario, Canada, and compared prediction performance with other models. Both of the aforementioned sampling approaches did not consider multiple discharge disposition which may influence the LOS. Hence, there is a need to develop a sampling algorithm that facilitates computer simulation of generating LOS realizations of a heterogeneous population of NH residents in a aged care system with varied individual characteristics.

#### *1.2.2.4 Facility-level Evacuation Response Modeling of Aged Care Systems Under Extreme Event Scenarios*

Apart from predicting facility-level performance of aged care systems under normal operating conditions, it is also important for the emergency operations agencies and health associations to understand how each aged care system respond in extreme event scenarios, e.g., hurricanes, winter storms, fires, earthquakes, etc. When an extreme event strikes, older adults are among the most vulnerable in the general population and face greater risks of functional declines, morbidity and mortality (Morrow, 1999; Dosa *et al.*, 2010). One of the most complex decisions administrators face related to facility performance is whether to evacuate the residents in extreme events or shelter-in-place. Extreme events, such as

hurricanes and winter storms, can cause extended power loss, damage to the building structure, disruption in communication, shortage in medical supply and staffing, etc., which result in disruption of care delivery and decline in resident health. The residents may need to be transferred to a safer location where care service can be continued. However, execution of transfer is challenging since residents in the facility, specially in NHs, are frail older adults, suffer from various functional limitations, and sometimes require specialized medical equipment. Hence, it is desirable to accurately predict the facility-level evacuation response to proactively prepare the facilities, i.e., identify host facility, procure transportation, train staff in transfer operations and psychological counselling for the residents, etc. Furthermore, at the regional-level, evacuation response prediction will allow planning and allocation of necessary resources by the associated government emergency management and response agencies. However, current studies in literature in understanding and predicting evacuation response have several deficiencies are as follows.

First, many works in evacuation response based their studies on behavior of the general population only. For example, Baker (1991) studied 12 hurricanes in the southeast region of the U.S., and reported that being located in low-lying area, evacuation order from the government agencies, housing structures, storm intensity, length of residence, and previous hurricane experience were influential factors in evacuation behavior for the general public. Moreover, Dash *et al.* (2007) conducted a review of then available models on household and individual hurricane evacuation responses and reported that most studies focused on identifying characteristic differentiating evacuees and non-evacuees, explaining important influential factors (e.g., perceived risk, family structure, socio-demographics, etc.), investigating compliance with government orders, and performing studies on traffic response. Kuligowski (2021) summarized over 200 studies on wildfire evacuations and emphasized that the works were mainly focused on general population identifying influential factors of evacuation, travel decisions on evacuation routes, destination, and mode of transport, and response times to the extreme event. Although fewer in number compared the nu-

merous works focusing on the general population, research on older adult highlight that the subgroup is differently and more acutely affected by extreme events. In a study based on community-dwelling older adults in New Orleans, Louisiana, McGuire *et al.* (2007) reported that about one-third of the population segment needed special accommodations for evacuation or sheltering-in-place, one-sixth needed special equipment. Furthermore, Parker *et al.* (2016) in their review on effects of natural disaster on psychology of older adults indicated that mental health services need be prepared with appropriate resources to tackle greater incidence of post-traumatic stress disorder (PTSD) among this segment of population, adversely affected if forced to evacuate. Hence, as advised by Willoughby *et al.* (2017) more research efforts need to be directed towards studying the extreme event response among older adults.

Second, current studies investigating evacuation response of aged care facilities, i.e., NHs and ALFs, are mainly limited to qualitative analysis methods, such as conceptual studies, narratives from interviews and surveys, and basic descriptive statistics. For example, Dobalian *et al.* (2010) summarized current studies in NH evacuation response and proposed a conceptual model to guide evacuation and extreme event preparedness plan, consisting of various factors of threat conditions, community resources, social coordination, and relocation strategies. Hyer *et al.* (2007) conducted a survey on NH administrators based on their experience during several major hurricanes in FL, summarized the narratives, and suggested that the need for availability of appropriate resources such as transportation, generators, and staff at the facilities must be recognized at the state- and community-level. Further, Peterson *et al.* (2020) used bivariate analyses and Chi-square tests on Florida state emergency system data of ALFs affected by hurricane Irma (2017) and reported location and size to be important factors influencing evacuation. In a more advanced example, Brown *et al.* (2012) fitted health outcomes data of residents with dementia in NHs affected by hurricane Gustav (2008) with difference-in-difference model and concluded that evacuation had a deleterious effect on the individuals. Since the majority of the studies

are descriptive in nature and focused on understanding various factors of evacuation, there is a need to utilize more advanced quantitative models available concerning predictive analytics, specifically, to obtain accurate predictions on facility evacuation response.

Third, current studies are limited in the data considered and most works mainly considered a single source of data, such as retrospective surveys or resident health data. There is a lack of research to comprehensively investigate rich environmental data for studying the evacuation response of aged care systems. For instance, Dosa *et al.* (2007) conducted telephone interviews and focus groups with administrators of NHs in Louisiana affected by hurricanes Katrina (2005) and Rita (2005) to investigate consequences of evacuation response on the facility , e.g., staffing, supplies, support, and resident conditions. Similarly, Blanchard *et al.* (2009) conducted further interviews with NH administrators to investigate effects of hurricane Gustav (2008) and corresponding evacuation details such as, transportation procurement, issues with host facility, and resident transfer challenges. In an example using more quantitative approach, Brown *et al.* (2012) joined state-mandated Minimum Data Set (MDS), which contains various socio-demographic, disease diagnoses, and functional limitation scores of NH residents recorded during course of stay, with CMS insurance data of 119 NHs over 3 years following hurricane Gustav (2008) to observe health outcomes of the residents. Moreover, Dosa *et al.* (2020) analyzed MDS data of over 61,000 residents of 640 NHs in Florida and reported increased mortality among residents exposed to hurricane Irma (2017) compared to non-extreme conditions in 2015. The objectives of aforementioned studies in investigating evacuation response can be further improved by augmenting resident data with facility characteristics and environmental data. Specifically, with the availability of advanced forms of technology, highly robust and detailed Geographic Information System (GIS) data is available with rich environmental information of the extreme events, e.g., hurricane data concerning forecast trajectory, wind speed, and potential flooding levels, are produced in real-time by the National Hurricane Center. Furthermore, inspection data detailing facility characteristics, such as staffing

levels, resident census, and organizational structure, are also available from the CMS. Incorporating such rich and several data sources together would highly improve prediction accuracy and also identify important facility and environmental factors that should be considered in response decision-making.

### 1.2.3 Prescriptive Healthcare Analytics for Older Adults

While descriptive analytics is used to discover important patterns in the data and predictive analytics is used to forecast future data/events, prescriptive analytics forms the next step in healthcare analytics where evidence-based optimal decisions are determined with various methodologies (Lepenioti *et al.*, 2020; Evans *et al.*, 2012). Prescriptive analytics is used to decide the best quantitative strategies that allows maximization of service delivery and health outcomes, while minimizing costs. Results from predictive analytics is used to develop "what-if" scenarios, formulate different operational strategies, and evaluate performance outcomes in the care delivery process to guide optimal decisions (Lopes, 2020).

In healthcare literature, different analytics-based models and methods such as stochastic models (Lakshmi *et al.*, 2013; Chan *et al.*, 2017; Schaefer *et al.*, 2005), simulation models (Oh *et al.*, 2016; Bhattacharjee *et al.*, 2016; Cappanera *et al.*, 2014), and mathematical programming models (Punnakitikashem *et al.*, 2013; Kim *et al.*, 2015; Bard *et al.*, 2005a) have been developed for informing and optimizing decisions at various health applications context, such as medical decision-making for screening and treatment of chronic diseases (Zhang *et al.*, 2011; Negoescu *et al.*, 2018; Helm *et al.*, 2015), capacity planning and bed management of healthcare facilities (Green, 2005; Kokangul *et al.*, 2017; Rau *et al.*, 2013), and scheduling and assignment of nursing staff and healthcare professionals (Harper *et al.*, 2003; Burke *et al.*, 2004; Bard *et al.*, 2005b). However, many of prescriptive analytics works have focused on the acute care setting, such as hospital, but for aged care systems mainly utilized by older adults, such as LTC systems, the research is relatively limited.

For instance, Zhang *et al.* (2012) developed an analytical framework to determine optimal bed capacity over multi-year horizon for LTC systems at the regional and facility level by incorporating survival analysis, DES, and optimization techniques and further proposed appropriate capacity planning strategies. Similarly, Li *et al.* (2016) proposed an analytical framework for LTC network consisting of home- and community-based setting and nursing homes to determine optimal bed capacities at the regional level, considering profit maximization under budget constraints. In another example, Ansah *et al.* (2014) applied a Systems Dynamics framework to model a network of LTC facilities and acute care settings to investigate the gap between staff and capacity demand and supply, and proposed that failure to proactively expand LTC services would result in surge in acute care demand and under-staffing. Furthermore, van Eeden *et al.* (2016) used queueing model to simulate a NH's "on-demand" care delivery to its residents, considering various exponential, gamma, and hyper-exponential distributions to response times, and proposed at least 80% service requests should be attended within 10 minutes to achieve better outcomes.

Since the primary objective of this dissertation is predictive analytics in older adults healthcare, application of prescriptive analytics is out of scope for this work. However, it forms a promising future direction based on the results obtained.

### 1.3 Overview and Organization of the Dissertation

This dissertation focuses on developing a series of healthcare data analytics models and algorithms for improving modeling and prediction of performance outputs at both individual level of older adults as well as system level of aged care facilities. The detailed contributions of the developed analytics methods and tools as well as their advancement of the aforementioned literature are elaborated with details as follows.

As described in the previous sections, many of existing individual level trajectory data modeling approach for older adults mainly considered a homogeneous population of individuals. Existing heterogeneity modeling approaches for characterizing temporal

heterogeneity of individual trajectories mainly consider either sub-population level model or individual-level model, but not both. In chapter 2, I proposed a bi-level temporal heterogeneous modeling framework to simultaneously capture the sub-population level and individual level temporal heterogeneity of individual older adults. Further, at each level, there are both technical contributions. For sub-population heterogeneity modeling, a Bayesian non-parametric modeling framework is proposed to relax the conventional parametric modeling assumption of pre-specifying the sub-population number. Further, the developed Bayesian sampling algorithm also allows the joint model estimation and sub-population number identification in a single step, which advances the conventional 2-step approach of estimating different models under different pre-specified number of sub-population in step 1 and selecting the best model in step 2. For individual-level heterogeneity modeling, functional data analysis techniques are considered to improve both the individualized performance prediction as well as model interpretation. To further illustrate the proposed modeling framework and demonstrate its practical importance, a longitudinal survey data of community-dwelling older adults is analyzed to capture their highly heterogeneous cognitive degradation performance over time.

The second type of individual level performance outcome data this dissertation aims to address is to develop analytics model for modeling the heterogeneous time-to-event data, and in particular the heterogeneous LOS data of post-acute care resident in the context of nursing home. As described in the previous section, existing LOS modeling approaches include distribution-based methods, regression-based methods and machine learning methods. There is lack of analytics tools to address the LOS data with multiple and competing discharge dispositions, such as community discharge and re/hospitalization of post-acute care residents in NHs. Further, many of existing LOS data modeling approaches focused on modeling individuals or a cohort in the acute care setting, such as hospital, and there is lack of studies focusing on LOS modeling in long-term care setting, such as NH. To fill the aforementioned research gap, chapter 3 proposes a semi-parametric LOS

modeling framework for post-acute care residents with multiple discharge dispositions by jointly modeling and characterizing the hospital readmission risk over time as well as community discharge likelihood over time. In addition, various individual characteristics are incorporated into the model to improve the individualized performance output prediction. Further, a sampling algorithm is developed for the proposed model to ensure the accurate predicted samples generation for a population of NH residents with varied individual characteristics.

In addition to modeling the individual performance outputs of older adults, such as community-dwelling older adults or NH residents, this dissertation also investigates the organizational level performance output modeling and prediction under extreme event scenarios, such as evacuation response of NHs under the hurricane scenarios. As described in the previous section, existing facility performance modeling works mainly focused on aged care systems under normal operating conditions. On the other hand, existing studies focused on studying extreme event scenarios, such as hurricanes, mainly focused on the general population instead of the vulnerable population of NH residents. Existing NH evacuation research are mainly qualitative and there is a need to develop advanced analytics-based models to improve the understanding and prediction performance of NH evacuation response. To fill the above research gap and address the research need, in chapter 4, a GIS-integrated predictive analytics framework is proposed to improve the modeling and prediction of NH evacuation response. Rich GIS data with both spatial and temporal heterogeneity of storm characteristics are collected and relevant GIS features are extracted to reflect different aspects of environmental conditions that may potentially influence the NH evacuation decision. Further, GIS features are incorporated with varied facility level characteristics as well as resident characteristics to comprehensively investigate the multi-factorial nature of NH evacuation decisions. To illustrate the proposed work and show the benefits of incorporating GIS information for prediction performance

improvement, a real case study of a Florida NH evacuation during recent hurricane Irma in 2017 is considered to demonstrate the practicability of the proposed work.

The dissertation is organized as follows. Chapter 1 introduces the background and significance of older adults healthcare and healthcare analytics, and further presents a literature review on existing analytics methods for modeling individual level performance as well as facility level performance for older adults. Chapter 2 proposes a bi-level heterogeneous performance degradation modeling framework for characterizing temporal heterogeneity of cognitive performance trajectory data of individual community-dwelling older adults by integrating both Bayesian non-parametric as well as functional data analysis techniques. Chapter 3 proposes a discharge outcome prediction model and sampling algorithm with improved individualized risk prediction for capturing heterogeneous length of stay data of individual NH residents with multiple and competing discharge dispositions and varied individual characteristics. Chapter 4 proposes a GIS-integrated predictive analytics method for improving the prediction of NH evacuation response at the facility level by integrating rich spatial and temporal storm characteristics as well as varied individual and facility characteristics. Chapter 5 concludes the dissertation with some future research directions outlined. Figure 1.3 gives an organizational diagram of this dissertation.

All healthcare datasets considered for illustrating the proposed healthcare analytics methods in this dissertation are de-identified secondary datasets, which do not involve human subjects and do not meet the definition of human subject research under USF IRB policy. To further safeguard the confidentiality of data, we already signed the business associate agreement prepared by our industrial collaborator. Such agreement is required by Health Insurance Portability and Accountability (HIPAA) Act and Health Information Technology for Economic and Clinical Health (HITECH) Act.

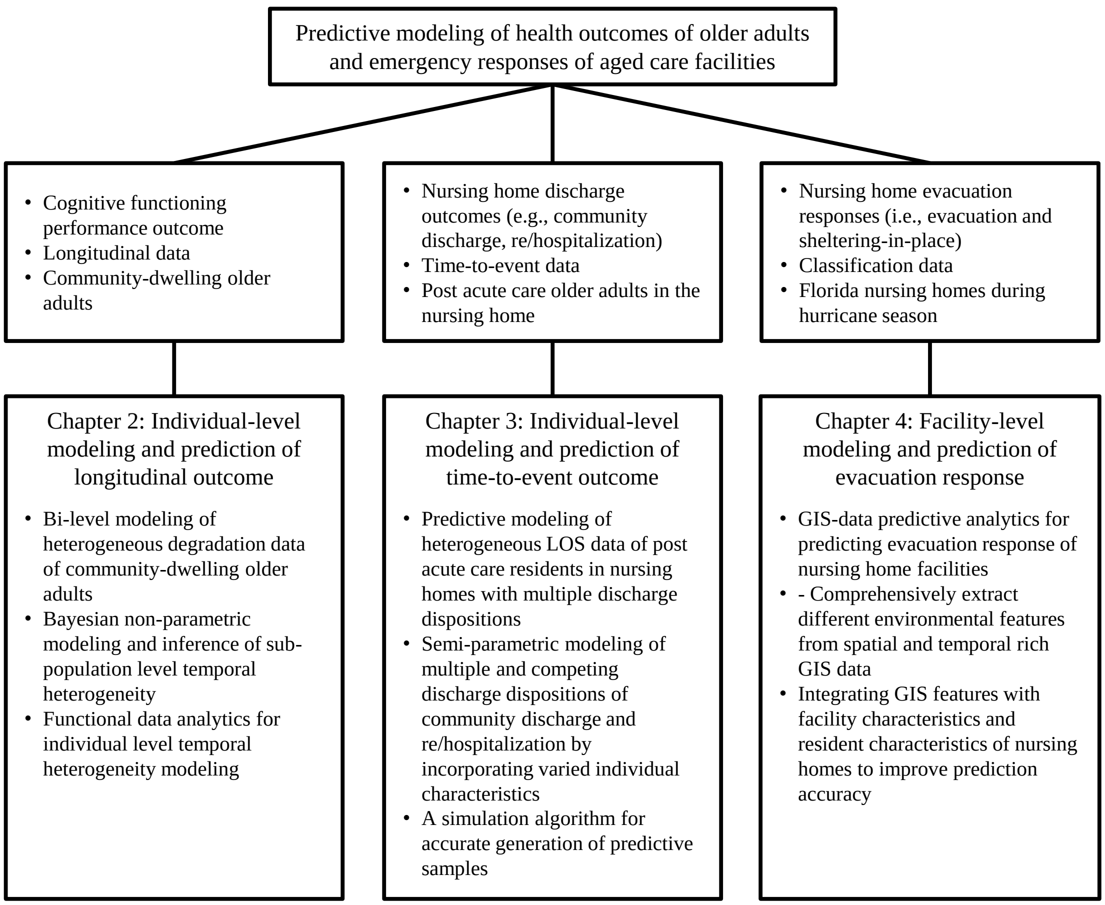



Figure 1.3: Organization of dissertation

## Chapter 2

## Bi-level Heterogeneity Modeling of Functional Performance Degradation for the Aging Population[1]

### 2.1 Introduction

Experiencing functional (e.g., physiological or psychological) limitation and performance degradation, the elderly with disabilities are the majority of the service recipients of the U.S. long-term care system (LTC). National estimates suggest that the number of the elderly with disabilities in 2008-2012 was about 15.6 million, which accounted for 38.7% of the total number of U.S. older adults (He *et al.*, 2014). As the baby boomers began turning 65 in 2011, coupled with decreased fertility and increased longevity, the U.S. will experience considerable growth in the elderly population with disabilities. To meet with the excess demand without compromising the quality of care, successful LTC preparedness and service planning becomes of great importance.

The LTC system encompasses a variety of settings and services to satisfy different needs of the elderly people, which are mainly determined by their functional performance and disability conditions. Due to the highly-varied characteristics among individuals, functional limitation and performance degradation of elderly people are highly heterogeneous, thus generating heterogeneous needs of LTC resources in different aspects. From the LTC network aspect, heterogeneous functional performance of elderly people will greatly affect their demands of different LTC facility settings in a LTC network, ranging from institu-

[1]This chapter is derived in part from my first-authored article published in IISE Transactions on Healthcare Systems Engineering, Volume 7, 2017, copyright Taylor & Francis, available online: https://doi.org/10.1080/24725579.2017.1339147. Copyright information is provided in Appendix A.1

tional settings, e.g., nursing home (NH), to home and community-based settings, e.g., assisted living facility (ALF). According to national estimates in 2013 (Harris-Kojetin *et al.*, 2013), 90.9% and 86.6% of people in NH need dressing and toiling assistance due to major functional declines, compared to 44.9% and 36.8% of people in ALF. In addition, 48.5% of people in NH have depression or cognitive impairment, compared to 24.8% and 39.6% of people in ALF, respectively. People with more severe functional limitations are more likely to need NH, since they can receive 24-hour skilled nursing care and a wide range of personal care assistance. On the contrary, people with minor disability conditions may be more willing to choose home and community-based care settings, such as ALF, because their quality-of-life can be maximized by balancing their wellness, self-independence and privacy. Successful modeling and quantification of heterogeneous performance degradation over time will help develop more proactive and adaptive LTC preparedness decisions, such as better capacity planning and/or resource re-balancing strategies among multiple LTC settings in the region of interest, to meet the heterogeneously evolving needs of the aging population. Neglecting such heterogeneity may result in capacity shortage or surplus of certain LTC setting, causing negative consequences for both care providers and care receivers. For instance, due to capacity shortage of NH, people with severe physical limitations may be placed in ALF with less intensity of care. Their fall incidences and fall-related injuries will dramatically increase, resulting in both increase of health care cost and mortality rate.

From the aspect of a single LTC facility, heterogeneous functional performance will affect the utilization of different types of nursing resources and their service times. According to relevant statistics reported by the Centers for Medicare & Medicaid Service (CMS) (STRIVE Project Study, 2005), for the nursing home residents with reduced physical function, their utilization of nursing resources varies based on their total Activities of Daily Living (ADL) Score (Weiner *et al.*, 1989) on a scale of 0-15. The score measures the resident's self-independence on activities in the nursing home, such as bed mobility, transfer, toilet

use, eating, etc., and a larger score indicates less independence. For the residents with ADL score more than 14, their average nursing time per day is about 19 minutes for the Licensed Practical Nurse (LPN) and about 109 minutes for the Certified Nursing Assistant (CNA). On the contrary, for residents with ADL score less than 2, their average nursing time per day is about 14 minutes for the LPN and about 18 minutes for the CNA. CNAs help residents with tasks, such as eating, dressing, using toilet, etc., and less independent residents have much higher utilization of CNA. LPNs help plan and implement the residents' care and treatment, and less independent residents have slightly higher utilization of LPN. Successful modeling and quantification of heterogeneous performance degradation over time will help develop more targeted and proactive decision making in workforce planning and nursing resources prioritization based on the heterogeneous utilization of residents in a single LTC facility. Neglecting such heterogeneity and implementing one-size-fits-all decisions may lead to inappropriate staffing levels and worsened the quality-of-care for residents who need more LTC resources.

The majority of existing studies in modeling heterogeneous performance degradation of the aging population have primarily focused on single-level heterogeneity modeling, namely the sub-population level or the individual level. Existing sub-population level modeling mainly consists of subjective categorization (Lunney *et al.*, 2003; Ferrucci *et al.*, 1996) and group-based modeling (Liang *et al.*, 2011; Nagin, 2005; Zimmer *et al.*, 2012). The main objective of these methods is to categorize the performance degradation of the overall population into multiple sub-populations/groups with distinct temporal evolving patterns. Subjective categorization qualitatively categorizes sub-populations based on clinical observations, medical experience, and/or graphical visualization. Such *ad hoc* methods lack rigorous justification and are subject to subjective bias and/or visualization error (Nagin *et al.*, 2010). To quantitatively categorize sub-populations, the group-based modeling approaches develop finite mixture models to quantify heterogeneity in a two-step procedure. In the model estimation step, a series of mixture models with different

pre-specified numbers of sub-populations are developed and estimated. In the model selection step, the most appropriate number of sub-populations is determined based on hypothesis tests such as likelihood ratio tests, or information criterion such as Akaike information criterion (AIC) and Bayesian information criterion (BIC). In summary, the aforementioned methods assume a known and fixed number of sub-populations, and either pre-determine it based on subjective knowledge or post-determine the number based on a two-step procedure. It will be desirable to relax such assumption without assuming a known number of sub-populations and learn it objectively from data in a one-step procedure.

Unlike the sub-population level heterogeneity modeling, the individual level heterogeneity modeling approaches, such as discrete time transition models (Li *et al.*, 2013; Wolfson *et al.*, 2014) and growth-curve models (Lynch *et al.*, 2002; Brown, 2010; Gill *et al.*, 2010), characterize the individual level evolving patterns of performance degradation with an implicit assumption of homogeneous population. In the discrete time transition modeling approaches, performance degradation can only be evaluated at pre-determined discrete time points. Growth-curve models are capable of evaluating degradation performance at any continuous time stamp and incorporate random parameters to quantify individual heterogeneity. However, they often utilize individual specific random intercept parameter and/or random slope parameter to quantify heterogeneity of initial degradation performance and/or degradation rate among individuals, respectively. The modeling accuracy may be constrained by the limited temporal evolving information extracted. In addition, there is no further investigation on how different potential influencing factors (e.g., risk and/or protective factors) will explain such individual heterogeneity. It will be desirable to investigate the underlying continuous degradation process of all individuals and extract richer temporal evolving information, such as the higher-order degradation curvature. Moreover, the potential influencing factors that may contribute to explaining such individual heterogeneity need to be identified.

To fill the research gap and systematically investigate the heterogeneous performance degradation of the aging population, this paper proposes a bi-level (i.e., both sub-population level and individual level) heterogeneity modeling and quantification framework. Specifically, at the sub-population level, a Bayesian non-parametric model is proposed to characterize sub-population level heterogeneity by relaxing the assumption of pre-specifying the number of sub-populations in the conventional modeling approaches. The proposed formulation also allows simultaneously identifying the number of sub-populations and estimating sub-population specific parameters in a one-step procedure. In addition, a convenient estimation algorithm is proposed to address a series of technical challenges encountered during the model estimation. At the individual level, a functional data analysis technique, namely Functional Principal Components Analysis (FPCA) (Ramsay *et al.*, 2005), is employed to investigate the individual degradation heterogeneity within each sub-population. FPCA characterizes the individual performance degradation at a finer and continuous time scale and extracts richer temporal evolving information to improve the modeling accuracy. Furthermore, the extracted temporal evolving signatures of individual degradation profiles can be represented with concise and interpretable scores. They will facilitate the identification and quantification of different risk and/or protective factors involved in influencing the individual heterogeneity.

The remaining of the paper is organized as follows. The next section presents the proposed systematic framework with details of its composing sub-modules, namely sub-population level heterogeneity modeling, estimation algorithm, individual heterogeneity modeling and post-analysis. Then, both numerical case study and real-data case study is provided to illustrate the proposed work and demonstrate its performance. Conclusions follow in the end.

## 2.2 Methodology

To quantify the heterogeneous functional performance degradation of the aging population, the proposed systematic framework consists of several interrelated sub-modules. Figure 2.4 gives an overview of the process. First, sub-modules 1 and 2 realize sub-population level heterogeneity modeling by formulating a non-parametric Bayesian model and developing a model estimation algorithm, respectively. Based on the outputs of sub-population level modeling, sub-module 3 investigates individual level heterogeneity using functional data analysis techniques. The extracted rich individual heterogeneity information can be further analyzed to identify possible influencing factors, such as risk and/or protective factors. Technical details of each sub-module are elaborated as follows.

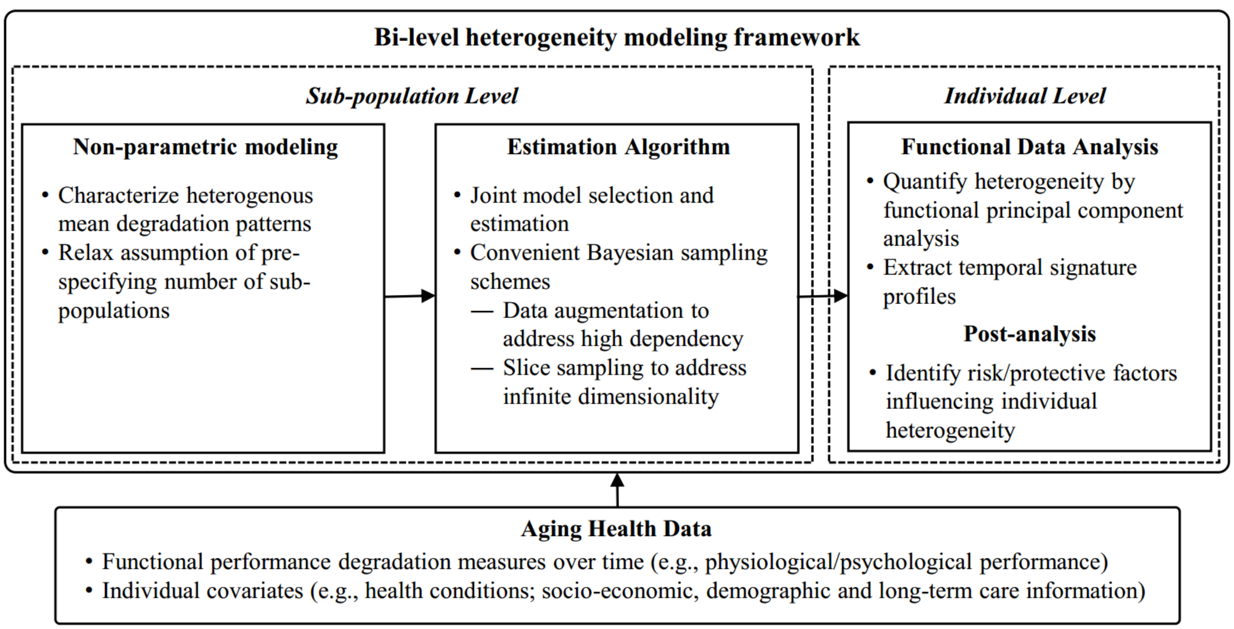



Figure 2.4: Descriptive diagram of the proposed bi-level heterogeneity quantification framework for functional performance degradation

### 2.2.1 Sub-population Level Heterogeneity Modeling

Considering a heterogeneous population of $N$ elderly individuals, the functional performance degradation of individual $i$ is evaluated at multiple discrete time points

$t_{ij}, j = 1, ..., n_i$, with performance measures $y_{ij}$'s. Conventional group-level heterogeneity modeling approaches (Nagin, 2005; Zimmer *et al.*, 2012) assume that a heterogeneous population consists of a fixed number $K$ of sub-populations with distinct evolving patterns characterized by $m_k(t)$, where $m_k(t)$ represents the mean functional degradation of individuals belonging to sub-population $k$. $y_{ij}$ can then be parametrically modeled as

$$f(y_{ij}|\mathbf{\Theta}) = \sum_{k=1}^{K} p_k f_k(y_{ij}|m_k(t_{ij}), \sigma_k^2), \quad \forall i = 1, ..., N; j = 1, ..., n_i, \tag{2.1}$$

where $f(\cdot)$ is the probability density function (pdf) of the overall heterogeneous population governed by a collection of all unknown parameters $\mathbf{\Theta}$. $p_k$ is the proportion of sub-population $k$ in the overall population and $f_k(\cdot)$ is the corresponding pdf. $f_k(\cdot)$ captures the sub-population level heterogeneity and is a pdf of normal distribution, i.e., $\mathcal{N}(m_k(t), \sigma_k^2)$, where $\sigma_k^2$ quantifies the corresponding within-sub-population variability.

A major limitation of the conventional formulation in Equation (2.1) is that a known and fixed number of sub-populations, $K$, needs to be determined. It is either pre-determined based on subjective judgment and/or graphical visualization, or post-determined based on various model selection techniques, such as pairwise hypothesis testing or ranking of information criterion. Pre-determination is less objective and its accuracy is often constrained by visualization errors or subjective bias. Post-determination requires a two-step approach of estimating a series of candidate models by varying the number of sub-populations and selecting the most appropriate one. To relax such assumption, obviate the difficulty of assuming a fixed number of sub-populations, and realize one-step joint model of estimation and selection, a new model is formulated for characterizing the sub-population level functional performance degradation heterogeneity and can be

represented as

$$\begin{aligned} y_{ij}|m_{(i)}, \sigma^2_{(i)} &\sim f_{(i)}(\cdot|m_{(i)}(t_{ij}), \sigma^2_{(i)}), \quad \forall i = 1, ..., N; j = 1, ..., n_i; \\ (m_{(i)}(t), \sigma^2_{(i)})|G &\sim G, \quad \forall i = 1, ..., N; \\ G|\eta, G_0 &\sim DP(\eta, G_0(\cdot)), \end{aligned} \tag{2.2}$$

where $(m_{(i)}(t), \sigma^2_{(i)})$ represents the sub-population specific mean degradation and variability associated with individual $i$. $(m_{(i)}(t), \sigma^2_{(i)})$ has a probability of $p_k$ to be selected as $(m_k(t), \sigma^2_k)$ from a set of countable but infinite number of possible sub-populations, i.e., $\{m_k(t), \sigma^2_k\}_{k=1}^{\infty}$. $G$ is a random distribution generated from a non-parametric prior of Dirichlet process (DP), denoted as $DP(\cdot)$. $G_0$ is the base distribution of DP with a positive diffusion parameter $\eta$. To show the connection and difference between Equation (2.1) and Equation (2.2), a constructive representation of $G$ (Sethuraman, 1994) is considered and represented as $G = \sum_{k=1}^{\infty} p_k \delta_{(m_k, \sigma^2_k)}$, where $\delta_{(m_k, \sigma^2_k)}$ is the Dirac delta measure of a point mass of 1 at $(m_k, \sigma^2_k)$. Thus, the pdf of the overall heterogeneity population based on Equation (2.2) can be explicitly written as

$$f(y_{ij}|\mathbf{\Theta}) = \sum_{k=1}^{\infty} p_k f_k(y_{ij}|m_k(t_{ij}), \sigma^2_k), \quad \forall i = 1, ..., N; j = 1, ..., n_i. \tag{2.3}$$

Compared to Equation (2.1), there is no restriction on the number of sub-populations that needs to be specified before model estimation. The actual number of sub-populations will be automatically determined from the available data jointly with the model estimation.

### 2.2.2 Model Estimation

Estimation of the proposed model in Equation (2.2) is mathematically nontrivial. Specifically, given the performance degradation data of $N$ individuals, denoted as $\mathbf{D} = \{y_{ij}\}_{i=1,...,N}^{j=1,...,n_i}$, and a collection of unknown functions and parameters, denoted as $\mathbf{\Theta} =$

$\{m_k(\cdot), \sigma_k^2\}_{k=1}^{\infty}$, the likelihood function $L(\mathbf{\Theta}|\mathbf{D})$ can be explicitly written as

$$L(\mathbf{\Theta}|\mathbf{D}) = \prod_{i=1}^{N}\prod_{j=1}^{n_i}\left(\sum_{k=1}^{\infty} p_k f_k(y_{ij}|m_k(t_{ij}), \sigma_k^2)\right). \tag{2.4}$$

Since the proposed model is a non-parametric Bayesian hierarchical model, a convenient Bayesian estimation approach needs to be developed. Under the Bayesian framework, the objective is to obtain the joint posterior density $\pi(\mathbf{\Theta}|\mathbf{D})$. Denoting $\pi(\mathbf{\Theta})$ as the joint prior density function, the analytical derivation of $\pi(\mathbf{\Theta}|\mathbf{D}) = L(\mathbf{\Theta}|\mathbf{D}) \cdot \pi(\mathbf{\Theta})/\pi(\mathbf{D})$ involves evaluating $\pi(\mathbf{D}) = \int_{\mathbf{\Theta}} \pi(\mathbf{D}, \mathbf{\Theta}) d\mathbf{\Theta}$, which is mathematically intractable due to the high-dimensional integration. Markov chain Monte Carlo (MCMC) sampling algorithms need to be developed to draw exact samples for the joint posterior density $\pi(\mathbf{\Theta}|\mathbf{D})$. However, due to the complex form of $L(\mathbf{\Theta}|\mathbf{D})$ in Equation (2.4), several estimation challenges need to be addressed: (i) All unknown parameters (e.g., $\sigma_k^2$) and functions $m_k(\cdot)$ are highly dependent. Such high dependency will result in failed convergence of conventional MCMC (Roberts *et al.*, 1997); (ii) There are infinitely many parameters and functions to be estimated in $\mathbf{\Theta}$, which makes model estimation computationally formidable; (iii) Parametric form of $m_k(\cdot)$ needs to be specified by considering both model flexibility and computational convenience.

To address the aforementioned model estimation challenges, a data augmentation technique is first adopted through augmenting auxiliary variables to simplify the complex dependency structure in $L(\mathbf{\Theta}|\mathbf{D})$. It has been successfully considered in both deterministic optimization for maximizing a likelihood function (Watanabe *et al.*, 2003) and stochastic algorithm for sampling a posterior density (Tanner *et al.*, 1987). Specifically, considering auxiliary variables $\mathbf{Z} = \{z_i\}_{i=1}^{N}$, where $z_i$ is a categorical variable taking positive integers to represent the sub-population membership of individual $i$, $z_i = k$ indicates that performance degradation of individual $i$ belongs to sub-population $k$. It is noticed that, in the proposed model, both the sub-population memberships $z_i$'s and the total number of sub-populations are unknown quantities. They will be simultaneously estimated together with the model

parameters during the course of model estimation. Based on the data augmentation, the augmented joint likelihood function $L(\mathbf{\Theta}|\mathbf{Z},\mathbf{D})$ can be written as

$$L(\mathbf{\Theta}|\mathbf{Z},\mathbf{D}) = \prod_{i=1}^{N}\prod_{k=1}^{\infty}\left(\prod_{j=1}^{n_i} f_k(y_{ij}|m_k(t_{ij}),\sigma_k^2)\right)^{\mathbf{I}(z_i=k)}, \quad (2.5)$$

where $\mathbf{I}(\cdot)$ is an indicator function. With data augmentation, unlike Equation (2.4) where all unknown functions and parameters are dependent, $(m_k(\cdot),\sigma_k^2)$ become conditionally independent of $(m_{k'}(\cdot),\sigma_{k'}^2), \forall k \neq k'$, and the dependency structure is greatly simplified.

For sub-population $k$, $m_k(t)$ represents the mean performance degradation. To characterize the potential nonlinear-evolving degradation patterns while maintaining the Bayesian computation convenience, a linear additive functional form is specified, i.e., $m_k(t) \approx \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\phi}_k(t)$. $\boldsymbol{\phi}_k(t)$ is a vector of basis functions, such as polynomial basis and splines, and $\boldsymbol{\beta}_k$ is the corresponding vector of coefficients. Based on such specification, $m_k(t)$ becomes a nonlinear function of $t$ with adequate modeling flexibility in capturing various sub-population specific degradation patterns. On the other hand, $m_k(t)$ is also a linear function of unknown coefficients $\boldsymbol{\beta}_k$, and the conjugate prior distributions are available to facilitate the Bayesian estimation. Specifically, considering the prior distribution for $\boldsymbol{\beta}_k$ as the multivariate normal distribution, i.e., $\boldsymbol{\beta}_k \sim \mathrm{MVN}(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$, the corresponding full conditional posterior distribution can be expressed as

$$\boldsymbol{\beta}_k|\sigma_k^2,\mathbf{Z},\mathbf{D} \sim \mathrm{MVN}(\boldsymbol{\mu}_k',\boldsymbol{\Sigma}_k'), \quad (2.6)$$

where $\boldsymbol{\mu}_k' = \boldsymbol{\Sigma}_k'\left(\frac{1}{\sigma_k^2}\boldsymbol{\Phi}_k^{\mathrm{T}}(t)\mathbf{Y}_k + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right)$ and $\boldsymbol{\Sigma}_k' = \left(\frac{1}{\sigma_k^2}\boldsymbol{\Phi}_k^{\mathrm{T}}(t)\boldsymbol{\Phi}_k(t) + \boldsymbol{\Sigma}_k^{-1}\right)^{-1}$. $\boldsymbol{\Phi}_k(t)$ is a matrix with its $j^{th}$ row defined as $\phi_k(t_j)$. $\mathbf{Y}_k$ is a matrix with its $i^{th}$ column represented as $\mathbf{y}_i, i \in S_k$, where index set $S_k$ is defined as $S_k = \{i : z_i = k, \forall i = 1, ..., N\}$. Similarly, considering the prior distribution for $\sigma_k^2$ as the inverse gamma distribution, i.e., $\sigma_k^2 \sim$

$\mathrm{IG}(a_k, b_k)$, the corresponding full conditional posterior distribution can be expressed as

$$\sigma_k^2 | \boldsymbol{\beta}_k, \mathbf{Z}, \mathbf{D} \sim \mathrm{IG}(a_k', b_k'), \tag{2.7}$$

where $a_k' = \frac{1}{2}|S_k| + a_k$, $b_k' = \frac{1}{2}\left(\mathbf{Y}_k - \Phi_k(t)\mathbf{fi}_k\right)^{\mathrm{T}}\left(\mathbf{Y}_k - \Phi_k(t)\mathbf{fi}_k\right) + b_k$ and $|\cdot|$ denotes the cardinality of set. As shown in Equation (2.6) and Equation (2.7), it is convenient to draw samples for both $\boldsymbol{\beta}_k|\sigma_k^2, \mathbf{Z}, \mathbf{D}$ and $\sigma_k^2|\boldsymbol{\beta}_k, \mathbf{Z}, \mathbf{D}$ since sampling routines for such common distributions are readily available for most of the programming and computing environments.

Another estimation challenge remains to be solved is that for the augmented variables $\mathbf{Z} = \{z_i\}_{i=1}^N$, each $z_i$ could take any positive integer, i.e., $z_i = 1, 2, ....$ In addition, an infinite number of parameters, i.e., $\{\boldsymbol{\beta}_k, \sigma_k^2\}_{k=1}^{\infty}$, are required to be estimated. To address such infinite-dimensionality issue, slice sampling technique (Li *et al.*, 2016; Walker, 2007) is considered. Specifically, introducing uniformly distributed random variables $\{u_i\}_{i=1}^N$, i.e., $u_i|z_i = k \sim \mathrm{Unif}(0, p_k)$, the full conditional posterior for $z_i$ becomes

$$\pi(z_i = k|u_i, \boldsymbol{\Theta}, \mathbf{D}) \propto \mathbf{I}(k \in B(u_i)) \prod_{j=1}^{n_i} f_k(y_{ij}|\boldsymbol{\phi}_k(t_{ij}), \boldsymbol{\beta}_k, \sigma_k^2), \tag{2.8}$$

where $B(u_i) = \{k : p_k > u_i\}$. It is noticed that $p_k$ is the proportion of sub-population $k$ satisfying $\sum_{k=1}^{\infty} p_k = 1$, which can be generated using the stick-breaking procedure as in the work of Sethuraman (1994): $p_k = v_k \prod_{k'=1}^{k-1}(1 - v_{k'}), k > 2$, $p_1 = v_1$ and $v_k \sim \mathrm{Beta}(1, \eta)$. Based on the stick-breaking procedure, it can be shown that $p_k < 1 - \sum_{k'=1}^{k-1} p_{k'}, \forall k$. Thus, when $1 - \sum_{k'=1}^{k_i^*} p_{k'} < u_i$, there is $k \notin B(u_i), \forall k > k_i^*$. Then, the cardinality $|B(u_i)| \leq k_i^*$ is a finite integer and $z_i$ in Equation (2.5) can only take $|B(u_i)|$ number of distinct positive integers. Similarly, when $1 - \sum_{k'=1}^{K^*} p_{k'} < \min\{u_i\}_{i=1}^N$, there is $k \notin \bigcup_{\forall i} B(u_i), \forall k > K^*$. It implies that even though there is an infinite number of sub-populations in the formulation, in actual model estimation, only $K^*$ number of sub-populations needs to be sampled.

Based on $\mathbf{Z} = \{z_i\}_{i=1}^N$ sampled in Equation (2.8), the stick-breaking procedure of generating $p_k|\mathbf{Z}$ can be achieved based on the updated $v_k|\mathbf{Z}$, which is given by

$$v_k|\mathbf{Z} \sim \text{Beta}\left(1 + \sum_{i=1}^{N} \mathbf{I}\left(z_i = k\right), N + \eta - \sum_{l=1}^{k}\sum_{i=1}^{N} \mathbf{I}(z_i = l)\right), \tag{2.9}$$

where $\eta$ is the aforementioned positive diffusion value of DP.

In summary, the proposed estimation procedure for sub-population level heterogeneous degradation modeling is shown in Algorithm 1.

---

**Algorithm 1** Sampling algorithm for proposed model

---

**Initialization:** $K^{(0)}$=1, $z_i^{(0)}$=1, $\forall i$=1, ..., $N$, $(\boldsymbol{\beta}_k^{(0)}, \sigma_k^{2(0)}) \sim \pi(\boldsymbol{\beta}_k, \sigma_k^2)$, $p_1^{(0)}$=$v_1^{(0)} \sim \text{Beta}(N+1, \eta)$

**procedure** DRAWSAMPLES

  **for** $\tau \leftarrow 1, \tau_{\max}$ **do**

    $u_i^{(\tau)}|z_i^{(\tau-1)}$=$k \sim \text{Unif}\left(0, p_k^{(\tau-1)}\right)$

    obtain $K^*$ from $K^*$=min$\left\{K : \sum_{k=1}^{K} p_k > 1 - \min\{u_i\}_{i=1}^N\right\}$, set $K^{(\tau)}$=$K^{(\tau-1)}$

    **if** $K^* > K^{(\tau-1)}$ **then**

      $K^{(\tau)}$=$K^*$

      generate $v_k^{(\tau)} \sim \text{Beta}(N+1, \eta)$ and compute $p_k^{(\tau)}$=$\prod_{k'=1}^{k-1}(1 - v_{k'}^{(\tau)})v_k^{(\tau)}, \forall k > K^{(\tau-1)}$

      generate $(\boldsymbol{\beta}_k^{(\tau)}, \sigma_k^{2(\tau)}) \sim \pi(\boldsymbol{\beta}_k, \sigma_k^2)$

    **end if**

    update $z_i^{(\tau)}$=$k|u_i^{(\tau)}$ by Equation (2.8)

    update $(\boldsymbol{\beta}_k^{(\tau)}, \sigma_k^{2(\tau)})|\mathbf{Z}, \mathbf{D}$ by Equation (2.6) and Equation (2.7)

    update $v_k^{(\tau)}|\mathbf{Z}$ by Equation (2.9) and compute $p_1^{(\tau)}$=$v_1^{(\tau)}$, $p_k^{(\tau)}$=$\prod_{k'=1}^{k-1}(1 - v_{k'}^{(\tau)})v_k^{(\tau)}, \forall k > 2$

  **end for**

**end procedure**

---

### 2.2.3 Individual Level Heterogeneity Modeling

In the previous section, based on the proposed model estimation algorithm, $\hat{K}$ number of sub-populations is identified jointly with the estimated model parameters $\hat{\boldsymbol{\Theta}} = \{\hat{\boldsymbol{\beta}}_k, \hat{p}_k, \hat{\sigma}_k^2, k = 1, ..., \hat{K}\}$. For the overall heterogeneous population, sub-population level

degradation heterogeneity is explicitly quantified by $\hat{m}_k(t)$'s with $\hat{m}_k(t) = \hat{\boldsymbol{\beta}}_k^{\mathrm{T}} \boldsymbol{\phi}_k(t)$. The unknown sub-population membership for any individual $i$ can be calculated as

$$\hat{z}_i = \max_k Pr(z_i = k|\mathbf{D}) = \max_k \frac{\hat{p}_k \prod_{j=1}^{n_i} f_k(y_{ij}|\hat{m}_k(t_{ij}), \hat{\sigma}_k^2)}{\sum_{k=1}^{\hat{K}} \hat{p}_k \prod_{j=1}^{n_i} f_k(y_{ij}|\hat{m}_k(t_{ij}), \hat{\sigma}_k^2)}, \tag{2.10}$$

where $Pr(z_i = k|\mathbf{D})$ is the calculated posterior probability of individual $i$ belonging to sub-population $k$ based on the estimation results in Algorithm 1. $\hat{z}_i$ is the estimated sub-population membership of individual $i$. In this paper, the estimated membership is under the maximum posterior probability criterion. It is the optimal criterion under 0-1 loss function by assuming equal miss-classification cost for each sub-population (Friedman *et al.*, 2001). Considering all individuals $i$'s that belong to sub-population $k$ (i.e., $z_i = k$), $\hat{m}_k(t)$ only represents their mean performance degradation. However, the performance of every individual does not evolve identically by exactly following the mean degradation path. There is still within-sub-population variability and the individual heterogeneity of $y_i(t)$ from $\hat{m}_k(t)$ that cannot be neglected and needs to be explicitly addressed and quantified.

To explicitly investigate individual level heterogeneity of degradation, the observed degradation path $y_i(t)$ of individual $i$ can be represented as

$$y_i(t) = x_i(t) + \epsilon_i(t) = m_{z_i}(t) + \Delta x_i(t) + \epsilon_i(t), \tag{2.11}$$

where $x_i(t)$ is the unknown but true degradation path of individual $i$, and $\epsilon_i(t)$ is the measurement error by assuming $\epsilon_i(t) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \forall i$, at any time $t$. $x_i(t)$ can be further decomposed into the sub-population specific mean degradation path $m_{z_i}(t)$ and individual heterogeneity $\Delta x_i(t)$. In the previous section, $m_{z_i}(t)$ is well quantified by the proposed non-parametric model and this section focuses on employing functional principal components analysis (FPCA) to quantify $\Delta x_i(t)$. FPCA is the important functional data analysis technique developed in recent years for studying variability of random functions (Ramsay *et al.*, 2005). The novel utilization of FPCA in the context of individual heterogeneity

modeling in this paper is because it is possible to further decompose $\Delta x_i(t)$ onto several orthogonal eigenfunctions and each eigenfunction carries an interpretable mode of degradation pattern. Specifically, considering $\Delta x_i(t)$'s are realizations of a random function $\Delta X(t)$ and based on Mercer's theorem (Ramsay *et al.*, 2005), the covariance function for $\Delta X(t)$ can be decomposed by the spectral decomposition as

$$\text{cov}(\Delta X(t), \Delta X(s)) = \textstyle\sum_{q=1}^{\infty} \lambda_q \psi_q(t) \psi_q(s), \tag{2.12}$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ are ordered eigenvalues with $\sum_{q=1}^{\infty} \lambda_q^2 < \infty$ and $\psi_q(t)$'s are orthonormal eigenfunctions satisfying $\int \psi_q^2(t)dt = 1$ and $\int \psi_q(t)\psi_{q'}(t)dt = 0, \forall q \neq q'$. Equation (2.12) indicates that the variability of $\Delta X(t)$ can be projected onto several orthonormal eigenfunctions $\psi_q(t)$'s, where $\psi_1(t)$ carries the largest variability, $\psi_2(t)$ carries the second largest variability, and so forth. As $q$ increases, the corresponding $\psi_q(t)$ carries less variability and is less interpretable. Thus, the first $Q$ number of eigenfunctions are often considered and the covariance function can be approximated as: $\text{cov}(\Delta X(t), \Delta X(s)) \approx \sum_{q=1}^{Q} \lambda_q \psi_q(t) \psi_q(s)$.

Based on a finite number of eigenfunctions and according to Karhunen–Loève theorem (Ramsay *et al.*, 2005), the individual degradation heterogeneity $\Delta x_i(t)$ of individual $i$ can be approximately written as

$$\Delta x_i(t) \approx \textstyle\sum_{q=1}^{Q} \xi_{iq} \psi_q(t), \tag{2.13}$$

where $\xi_{iq}$ is the functional principal component score for the $q^{\text{th}}$ eigenfunction of individual $i$, which is normally distributed with mean 0 and variance $\lambda_q$. It controls the magnitude of $\psi_q(t)$ contributing to $\xi_{iq}$. Equation (2.13) is interpreted as follows. The individual degradation heterogeneity can be expressed as a weighted combination of multiple eigenfunctions while each eigenfunction $\psi_q(t)$ characterizes a specific temporal variation signature of performance degradation.

To estimate the aforementioned eigenvalues, eigenfunctions, and scores of individuals, the performance degradation data $\mathbf{D} = \{y_{ij}, t_{ij}, \forall i = 1, ..., N; j = 1, ..., n_i\}$ of the heterogeneous population is first partitioned into sub-population degradation data $\mathbf{D}_k$'s, i.e., $\mathbf{D}_k = \{y_{ij} : \forall z_i = k, j = 1, ..., n_i\}$, according to Equation (2.10). For sub-population $k$, a set of sub-population specific eigenvalue and eigenfunction pairs $\{\lambda_q^{(k)}, \psi_q^{(k)}(t)\}_{q=1}^{Q}$ can be estimated by maximizing the following likelihood function $L(\{\lambda_q^{(k)}, \psi_q^{(k)}(t)\}_{q=1}^{Q}|\mathbf{D}_k)$ as

$$L(\{\lambda_q^{(k)}, \psi_q^{(k)}(t)\}_{q=1}^{Q}|\mathbf{D}_k) \propto \prod_{\forall z_i = k} \frac{1}{\sqrt{\det(\mathbf{\Sigma}_{\Delta x_i}^{(k)})}} \exp[-\frac{1}{2}(y_i(\boldsymbol{t}_i) - m_k(\boldsymbol{t}_i))^{\mathrm{T}}(\mathbf{\Sigma}_{\Delta x_i}^{(k)})^{-1}(y_i(\boldsymbol{t}_i) - m_k(\boldsymbol{t}_i))], \tag{2.14}$$

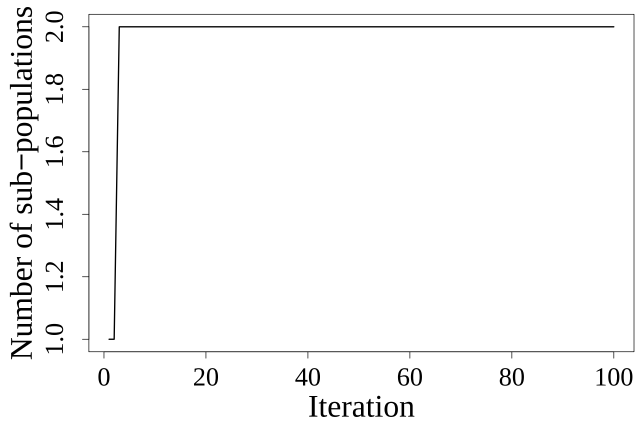
where $y_i(\boldsymbol{t}_i) = [y_{i1}, ..., y_{in_i}]^{\mathrm{T}}$, $m_k(\boldsymbol{t}_i) = [m_k(t_{i1}), ..., m_k(t_{in_i})]^{\mathrm{T}}$ and $\mathbf{\Sigma}_{\Delta x_i}^{(k)} = \mathrm{var}(y_i(\boldsymbol{t}_i) - m_k(\boldsymbol{t}_i))$. The restricted Maximum likelihood estimation method proposed by Peng *et al.* (2009) can be adopted to estimate $\{\lambda_q^{(k)}, \psi_q^{(k)}(t)\}_{q=1}^{Q}$ as $\{\hat{\lambda}_q^{(k)}, \hat{\psi}_q^{(k)}(t)\}_{q=1}^{Q}$. The scores $\xi_{iq}$'s of functional principal components can be further calculated (Yao *et al.*, 2005) as: $\hat{\xi}_{iq} = \sum_{j=1}^{n_i}(y_{ij} - m_k(t_{ij}))\hat{\psi}_q^{(k)}(t_{ij})(t_{ij} - t_{i,j-1})$ and $t_{i0} = 0, \forall i = 1, ..., N, q = 1, ..., Q$.

## 2.3 Case Study

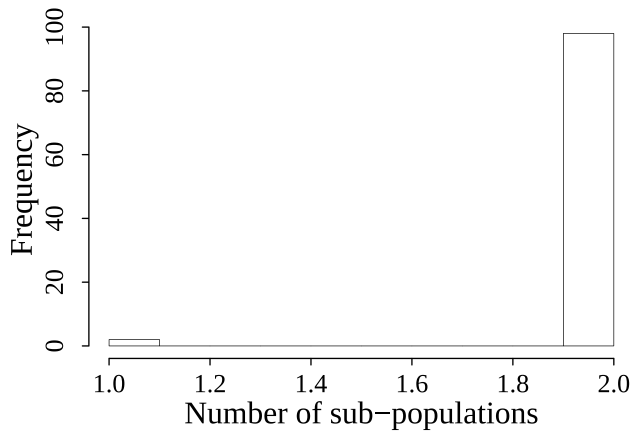
### 2.3.1 Numerical Case S tudy

To demonstrate the effectiveness of the proposed Bayesian sub-population level heterogeneity quantification method, a heterogeneous population with 2 sub-populations is assumed. Third-order polynomial basis functions, i.e., $\boldsymbol{\phi}(t) = [1, t, t^2, t^3]^{\mathrm{T}}$, are considered for the illustration purpose. For simplicity, sub-population $k$ is written as "sub-$k$" in short. Ground-truth values for sub-population $k$'s ($k = 1, 2$) parameters, namely the mixing proportion $p_k$, a vector of basis coefficients $\boldsymbol{\beta}_k = [\beta_{0k}, \beta_{1k}, \beta_{2k}, \beta_{3k}]^{\mathrm{T}}$ and the within-sub-population variability $\sigma_k^2$, are pre-assigned and summarized in Table 2.1. A total of 1000 samples are simulated for this study.

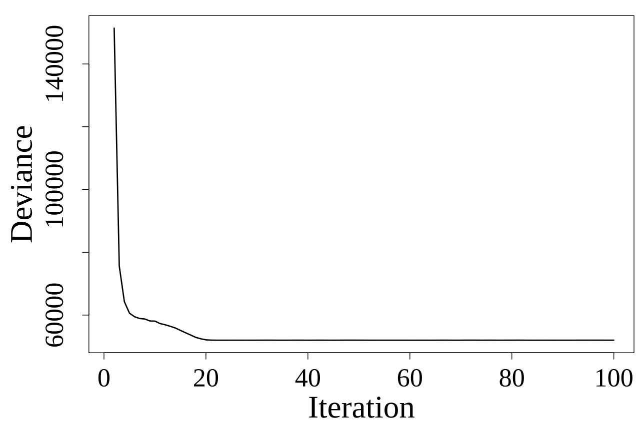
Figure 2.5 shows trace plots and histogram of the results of the estimated number of sub-populations. The estimated number of sub-populations over iterations stabilize at 2 in




(a) Trace plot of $\hat{K}$




(b) Histogram of $\hat{K}$




(c) Trace plot of deviance

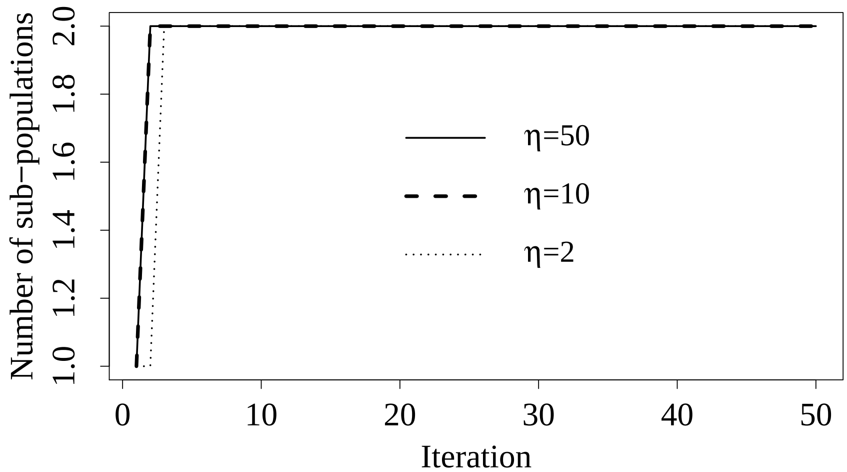
Figure 2.5: Trace plots and histogram of estimation of sub-population number $\hat{K}$

Table 2.1: Estimation results of simulation study

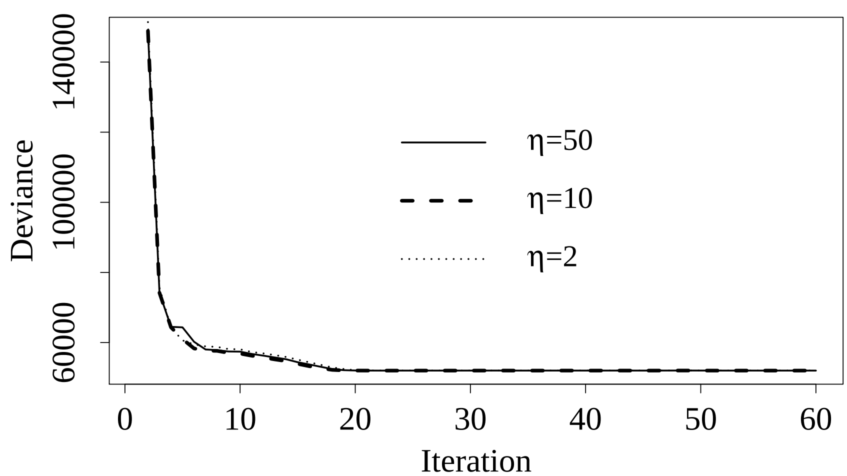
| Parameters | True value | Maximum likelihood estimation | Proposed method | |
|---|---|---|---|---|
| | | | Posterior mean | 95% Credible interval |
| $p_1$ | 0.4 | 0.39 | 0.39 | [0.36,0.42] |
| $p_2$ | 0.6 | 0.61 | 0.61 | [0.58,0.64] |
| $\beta_{01}$ | 1 | 1.09 | 1.08 | [0.11,2.02] |
| $\beta_{02}$ | 1 | 0.33 | 0.33 | [-1.22,1.86] |
| $\beta_{11}$ | 0.2 | 0.10 | 0.11 | [-0.60,0.83] |
| $\beta_{12}$ | 0.4 | 0.96 | 0.96 | [-0.17,2.13] |
| $\beta_{21}$ | 0.3 | 0.33 | 0.33 | [0.18,0.48] |
| $\beta_{22}$ | 0.2 | 0.07 | 0.07 | [-0.17,0.30] |
| $\beta_{31}$ | -0.4 | -0.40 | -0.40 | [-0.41,-0.39] |
| $\beta_{32}$ | -0.45 | -0.44 | -0.44 | [-0.46,-0.43] |
| $\sigma_1^2$ | 25 | 24.70 | 24.78 | [23.70,25.89] |
| $\sigma_2^2$ | 100 | 101.00 | 100.98 | [97.45,104.57] |

Figure 2.5a and the histogram highly concentrates at 2 sub-populations in Figure 2.5b. They both indicate that the identified number of sub-population is 2, which is consistent with the ground-truth specification of 2 sub-populations. Deviance values are further displayed in Figure 2.5c to evaluate the goodness-of-fit of the identified model over iterations and to monitor the convergence (Green *et al.*, 2001) of the sampling algorithm. The algorithm converges fast within 20 iterations since conjugate priors greatly facilitate the estimation convenience and performance. The corresponding point (e.g., posterior mean) and interval (e.g., 95% credible intervals) estimation results are summarized in Table 2.1. Maximum likelihood estimation (MLE) results assuming a known ground-truth number of sub-populations and known sub-population membership for each observation are also shown in Table 2.1. The posterior means of all parameters are almost identical to the MLE results since Bayesian estimation is dominated by the available data rather than prior knowledge. Both estimation results are also close to the pre-assigned ground-truth values. Some discrepancies can be seen for sub-population 2 due to its large within-sub-population variability (i.e., $\sigma_2^2 = 100$) but a finite sample size of data. Bayesian 95% credible intervals

successfully quantify such estimation uncertainty and fully cover all ground-truth values. The numerical case study can also help investigate an important issue regarding the miss-clustering of the sub-populations memberships. Unlike the real case study where the ground-truth values of sub-population memberships are unavailable, the ground-truth membership values of simulated data are known and it is possible to quantify the miss-clustering error as the miss-classification error. By comparing the predetermined ground truth values $z_i$'s and the estimated values $\hat{z}_i$'s in Equation (2.10), the proposed method achieves 100% accuracy of sub-population memberships.




(a) Trace plot comparison




(b) Deviance comparison

Figure 2.6: Comparison of results with different $\eta$

To further investigate the influence of hyper-parameter $\eta$ on the estimation performance, Figure 2.6 shows both trace plots and deviance monitoring results under scenarios of $\eta = 2$, $\eta = 10$ and $\eta = 50$, respectively. They both correctly identify the ground-truth number of sub-populations, and the numbers of iterations required to converge are also similar. Since the influence of hyper-parameter $\eta$ is negligible, $\eta = 2$ is chosen to show the estimation results of Table 2.1 for illustration purpose.

### 2.3.2 Real Case Study

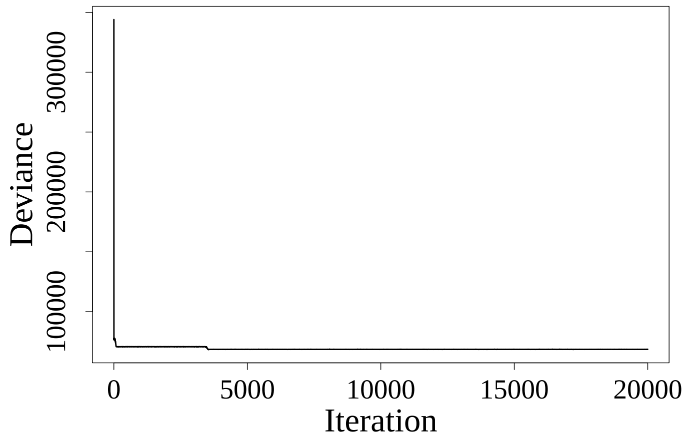
#### *2.3.2.1 Data Description*

To demonstrate the performance of the proposed heterogeneity quantification framework, the biennial longitudinal survey data of the Health and Retirement Study (HRS) is considered. The HRS data contains a wealth of information on disability and health conditions, socio-demographics, and health service utilization of a representative sample of community-dwelling older adults in the United States (The Health and Retirement Study, 2016). In this paper, a subset of cognitive degradation data of 2457 individuals with complete information is investigated since cognitive degradation is a hallmark of aging-related functional impairment and development of disability. Measurement metric of cognition of the HRS data was developed based on derivatives of the MMSE score (Folstein *et al.*, 1975). A telephone questionnaire containing 35 questions with binary outcomes (i.e., 1 indicating affirmative response of good health and 0 indicating otherwise) was asked to elderly individuals at 8 waves from 1998 to 2012 on a biennial basis. Such questions evaluate the multi-dimensional cognitive capabilities of the older adults, such as memory, information remembrance, comparative reasoning, judgment and behavioral consistency. The total score (ranging from 0 - 35) of the questionnaire is calculated as a cognition index to assess the overall cognitive performance (Ofstedal *et al.*, 2005). The proposed work aims to investigate both the sub-population level and the individual level heterogeneity of cognitive degradation.

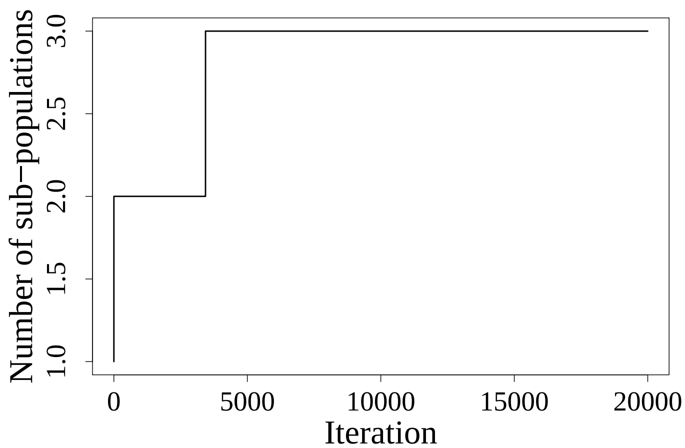
#### 2.3.2.2 *Sub-population Level Cognitive Degradation Heterogeneity*

To model the sub-population level heterogeneity of cognitive degradation, the proposed non-parametric modeling approach is implemented to obviate the requirement of pre-specifying the number of sub-populations. Three-order basis functions are considered to capture the sub-population mean degradation trajectory. According to existing literature (Nagin, 2005; Nagin *et al.*, 2010) and visual inspection of real data, three-order basis functions are reasonable choices in representing the overall trend of mean degradation trajectory. Modeling degradation heterogeneity at the individual level may require higher-order basis functions to capture detailed evolving patterns among individuals, which will be addressed by FPCA in Section 3.2.3. Compared to individual degradation trajectory, the mean degradation trajectory is at a more aggregate level to explore the similar evolving trends among individuals, thus requiring less modeling fidelity. Figure 2.7 shows the estimation results of the number of sub-populations $K$. The algorithm converges approximately at 4,500 iterations, and the identified number of sub-populations is 3. Figure 2.8 further displays the mean degradation trajectories (i.e., $\hat{m}_k(t) = \hat{\boldsymbol{\beta}}_k^{\mathrm{T}} \phi_k(t)$) and their corresponding variability bounds (i.e., $\hat{m}_k(t) \pm 2\hat{\sigma}_k$). The mean trajectory of each sub-population describes the average degradation of cognitive condition and the corresponding variability bounds describe the within-sub-population variability. As shown in Figure 2.8, the average cognition index values for sub-populations 1, 2 and 3 at the baseline wave (i.e., wave 1) are 27.5, 24.5 and 20.5, while their corresponding average absolute decrements after 8 waves are 1.5, 2 and 5, respectively. Thus, the cognitive degradation of the overall heterogeneous aging population can be characterized by three sub-populations with low, median and high levels of cognitive impairment in sub-1, 2 and 3, respectively.

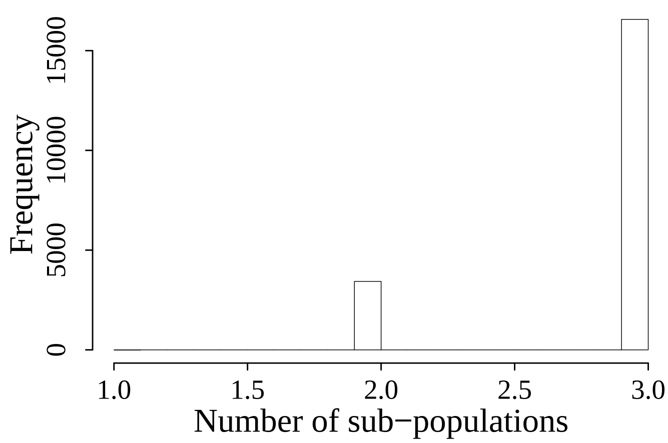
The proposed work is also compared with some of the existing parametric modeling approaches (Jones *et al.*, 2007; Nagin *et al.*, 2010) in investigating the sub-population level heterogeneity. Parametric models require a two-step model estimation and selection procedure. In the model estimation step, candidate models which pre-specify a fixed




(a) Deviance convergence results




(b) Trace plot of $\hat{K}$




(c) Posterior histogram of $\hat{K}$

Figure 2.7: Estimation results of number of sub-populations $\hat{K}$

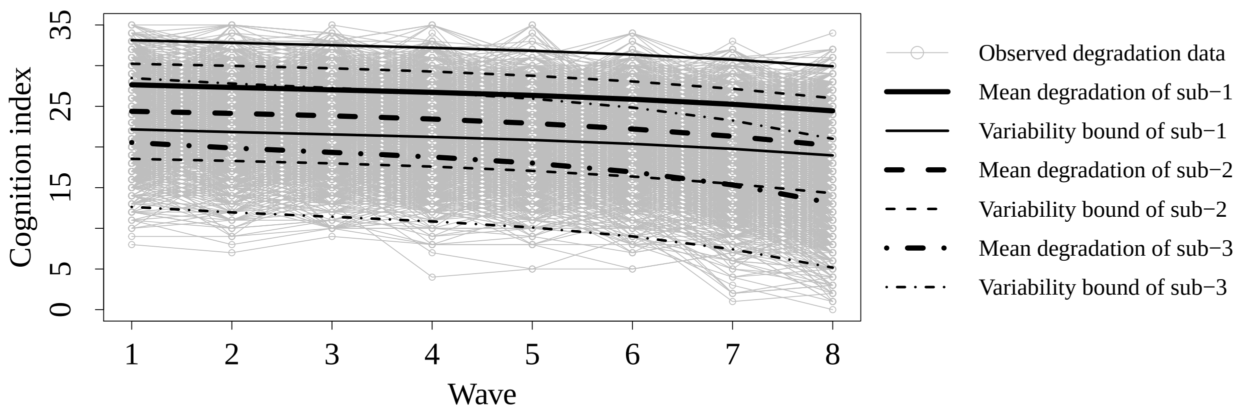



Figure 2.8: Estimation results of the identified 3 sub-populations

Table 2.2: Comparison of the estimation results with AIC-based and BIC-based methods

| Variables | Proposed method | | AIC based | BIC based |
|---|---|---|---|---|
| | Posterior mean | 95% Credible interval | | |
| $p_k$ | sub-1: 0.226<br>sub-2: 0.505<br>sub-3: 0.261 | sub-1: [0.198,0.256]<br>sub-2: [0.480,0.531]<br>sub-3: [0.238,0.285] | sub-1: 0.310, sub-2: 0.517<br>sub-3: 0.092, sub-4: 0.100 | sub-1: 0.691<br>sub-2: 0.315 |
| $\boldsymbol{\beta}_{0k}$ | sub-1: 28.060<br>sub-2: 24.803<br>sub-3: 21.267 | sub-1: [27.463,28.701]<br>sub-2: [24.365,25.262]<br>sub-3: [20.479,22.045] | sub-1: 23.933, sub-2: 26.537<br>sub-3: 27.468, sub-4: 21.227 | sub-1: 25.679<br>sub-2: 23.580 |
| $\boldsymbol{\beta}_{1k}$ | sub-1: -0.561<br>sub-2: -0.536<br>sub-3: -0.952 | sub-1: [-1.087,-0.044]<br>sub-2: [-0.913,-0.160]<br>sub-3: [-1.639,-0.242] | sub-1: -2.276, sub-2: -1.029<br>sub-3: -2.418, sub-4: 2.855 | sub-1: -0.482<br>sub-2: -1.038 |
| $\sigma_k^2$ | sub-1: 7.458<br>sub-2: 8.560<br>sub-3: 15.711 | sub-1: [7.082,7.853]<br>sub-2: [8.265,8.847]<br>sub-3: [15.016,16.298] | sub-1: 22.604, sub-2: 6.410<br>sub-3: 12.284, sub-4: 15.953 | sub-1: 9.321<br>sub-2: 23.141 |

*Note: sub-1, sub-2, sub-3 and sub-4 are abbreviations for sub-population 1, 2, 3 and 4 respectively*

number of sub-populations are developed and estimated. Then, in the model selection step, the final model is selected among the candidate models based on an information criterion, such as AIC and BIC. Table 2.2 summarizes the comparative results. All three methods

indicate the existence of sub-population level heterogeneity in cognitive degradation. Different methods select different numbers of sub-populations, or equivalently, admit different model complexities in the context of heterogeneity modeling. The BIC tends to select a simpler model with a smaller number of sub-populations since more penalty is imposed to control the model complexity. On the contrary, the AIC tends to select a more complex model with a larger number of sub-populations since less penalty term for controlling the model complexity is involved. The model complexity (or the number of sub-populations) of the proposed method is automatically determined by the data. There is no need to pre-specify the number of sub-populations or pre-design the penalty structure in controlling the model complexity. Moreover, unlike two-step modeling approaches, model estimation and selection can be accomplished simultaneously in a single step in the proposed work.

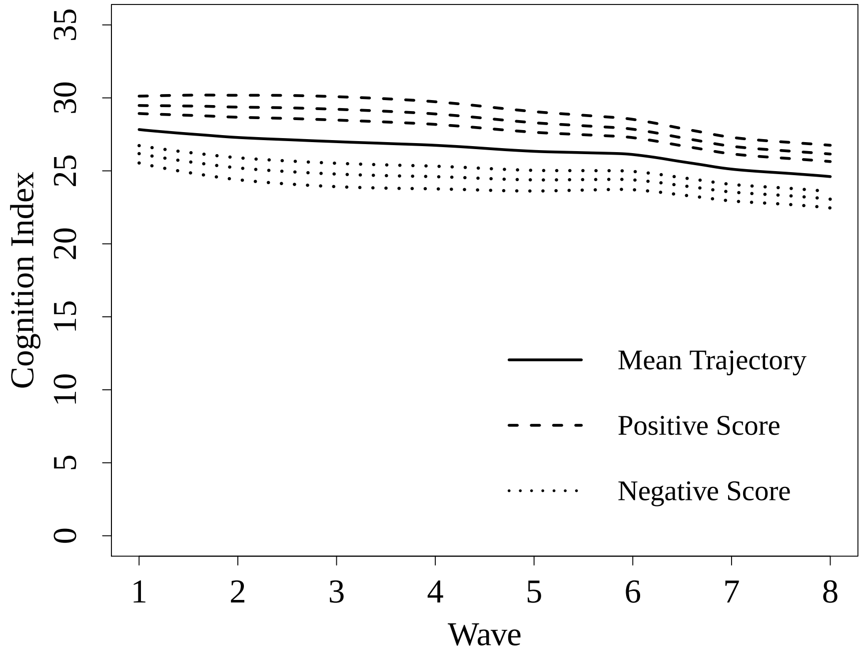
#### *2.3.2.3 Individual Level Cognitive Degradation Heterogeneity*

In the previous section, a non-parametric modeling approach is proposed and applied to characterize the sub-population heterogeneity of cognitive degradation. The overall heterogeneous aging population is decomposed into three groups of individuals with different average levels (i.e., low, median and high) of cognitive performances and evolving patterns. As shown in Figure 2.8, for each of the identified sub-population, there is a significant amount of within-sub-population variability that has not been further explained and quantified. Such within-sub-population variability is caused by individual level degradation heterogeneity nested within each sub-population. Based on the modeling outputs from the previous section, FPCA is employed in this section to investigate the individual level heterogeneity by treating each individual degradation trajectory as a functional data object.

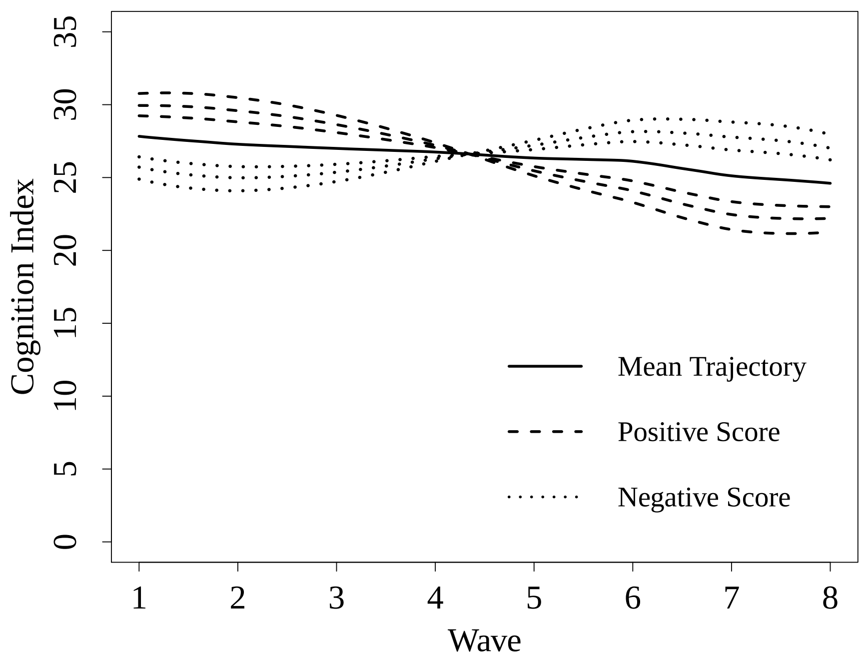
In the context of cognitive degradation modeling, FPCA essentially projects the within-sub-population variability into a set of orthonormal eigenfunctions. Each individual

degradation trajectory can be represented by a weighted combination of these eigenfunctions. The individual specific weights, defined as FPCA scores, can be utilized to quantify the individual degradation heterogeneity. Considering the most parsimonious number of eigenfunctions which gives desirable interpretation and retains a reasonable portion of variability, 2 eigenfunctions are selected for each sub-population. Based on the approximate leave-one-curve-out cross-validation scores (Peng *et al.*, 2009), four-order basis functions of splines are selected for eigenfunctions within sub-population 1 and 3, and six-order for sub-population 2 to avoid over-fitting the data. Figures 2.9, 2.10, and 2.11 summarize the sub-population specific eigenfunctions and visualizes their meaningful interpretations through comparison with the mean degradation trajectories (in solid curves). For simplicity, the eigenfunction of the first principal component for sub-population 1 is termed as EPC1-1 in short and similar abbreviations are applied to other eigenfunctions.

As shown in Figure 2.9a, a positive/negative score of EPC1-1 indicates an above-average/below-average cognitive degradation over the entire time period. Its degradation rate is similar to the average degradation of sub-population 1. As shown in Figure 2.9b, a positive score of EPC1-2 represents the above-average degradation performance value at the beginning but below-average performance value in the end. In addition, its degradation first accelerates and then decelerates compared to the mean degradation rate of sub-population 1. As opposed to EPC1-1, for EPC2-1 in Figure 2.10a, a positive/negative score indicates a below-average/above-average cognitive degradation. For EPC1-2 and EPC2-2, after above-average degradation transits into below-average degradation between waves 4 and 5, the cognition index value will first decrease then become stabilized. However, for EPC3-2 in Figure 2.11b, the cognition index value will continue to decrease monotonically, indicating a higher level of cognitive impairment and deterioration of individuals from sub-population 3. In addition, the transition times for EPC2-2 and EPC3-2 are slightly delayed than that of EPC1-2.

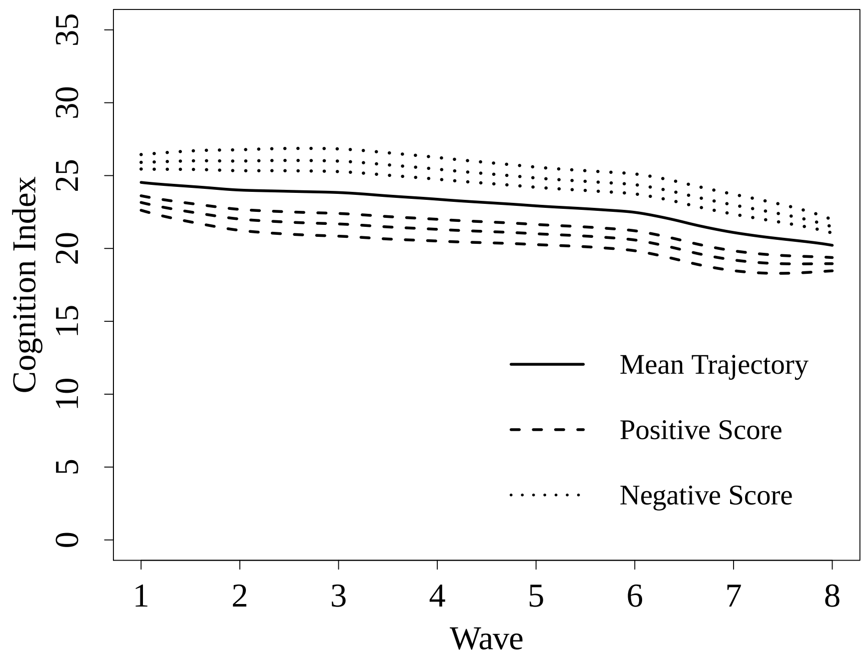



(a) EPC1-1

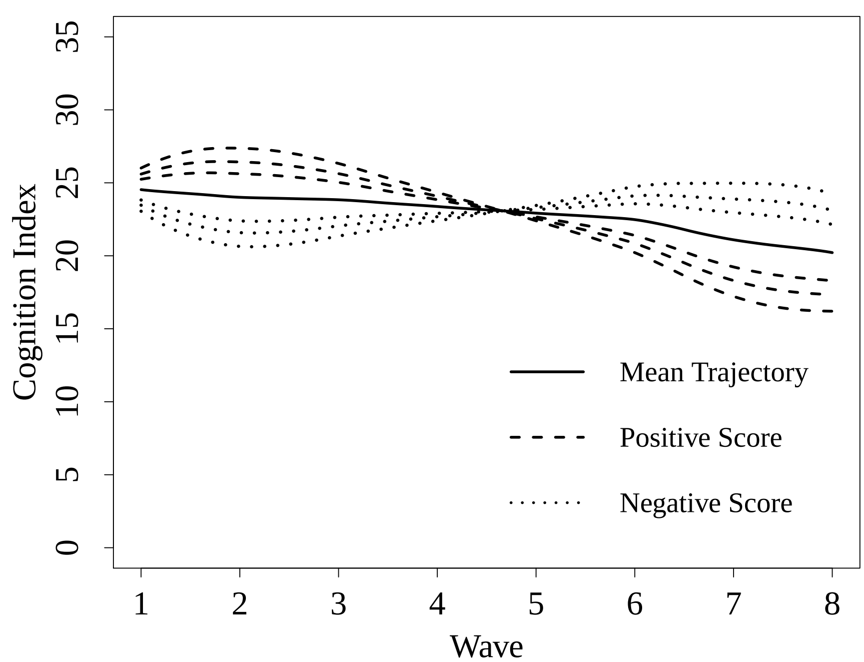



(b) EPC1-2

Figure 2.9: Temporal variation signatures of cognitive degradation represented by eigenfunctions for sub-population 1

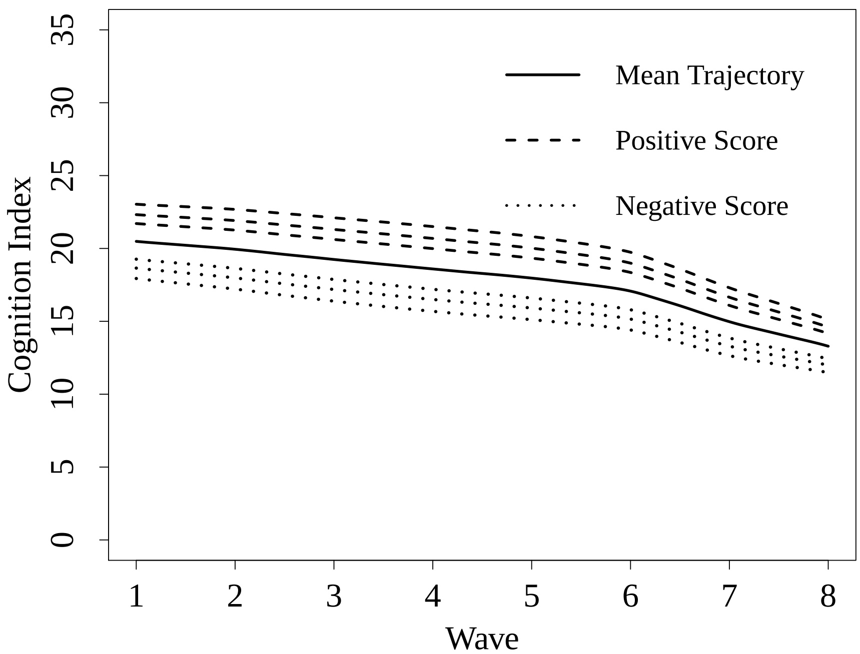



(a) EPC2-1

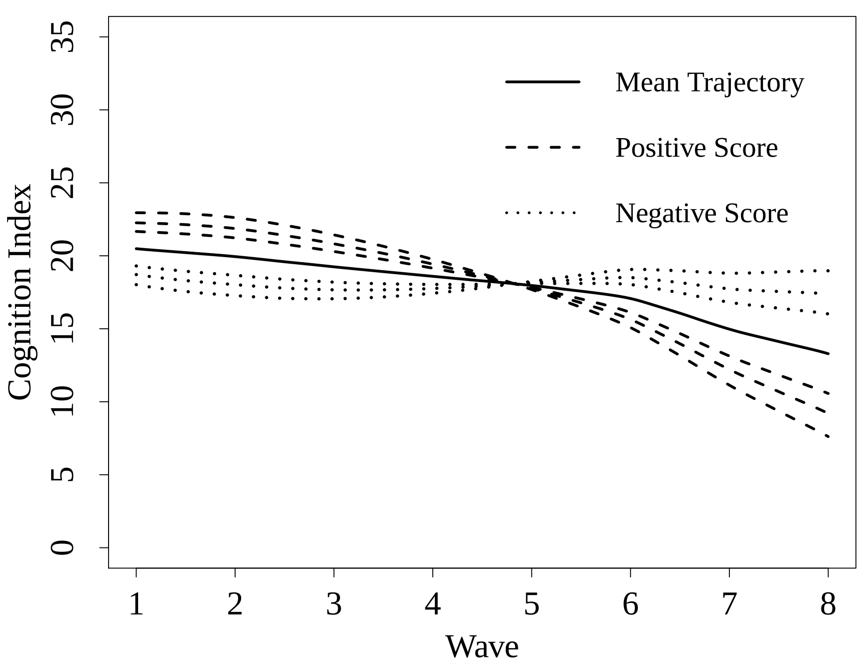



(b) EPC2-2

Figure 2.10: Temporal variation signatures of cognitive degradation represented by eigenfunctions for sub-population 2




(a) EPC3-1




(b) EPC3-2

Figure 2.11: Temporal variation signatures of cognitive degradation represented by eigenfunctions for sub-population 3

Based on FPCA, the individual level degradation heterogeneity can be directly quantified by the lower-dimensional scores associated with different eigenfunctions. Figure 2.12 visualizes the scores of EPC1-1 and EPC1-2 for all individuals belonging to sub-population 1. Individual cognitive degradation heterogeneity can be concisely summarized and visualized in a two-dimensional plane. These concise scores will serve as informative and quantitative indices in helping LTC decision-makers better allocate available nursing resources for individuals. Taking individuals A, B, C and D as four extreme cases, the positive score of EPC1-1, the negative score of EPC1-1, the positive score of EPC1-2, and the negative score of EPC1-1 are dominated, respectively. Based on the interpretation of eigenfunctions, the extreme scores indicate that individual A will have an above-average degradation, individual B will have a below-average degradation, individual C will have a first above-average then below-average degradation, and individual D will have a first below-average then above-average degradation.

In addition to the aforementioned concise and meaningful interpretations offered by FPCA, the modeling accuracy can be improved over the existing individual heterogeneity modeling approaches, such as growth-curve models. Specifically, we compare the modeling accuracy among FPCA and different growth-curve models with different individual-specific random parameters, such as the random intercept (RI), the random slope (RS), as well as the random intercept and slope (RIS). The overall mean square errors between predicted and observed degradation performances based on FPCA, RI, RS and RIS are 5.456, 7.273, 7.908 and 5.839, respectively. The superior modeling accuracy of the FPCA results is mainly because eigenfunction captures more temporal information, such as higher-order degradation curvature. Figures 2.13, 2.14, 2.15, 2.16 further visualizes the comparison results for the above representative individuals and demonstrates the satisfactory modeling accuracy of FPCA. The interpretation findings of FPCA can also be justified from the figures.

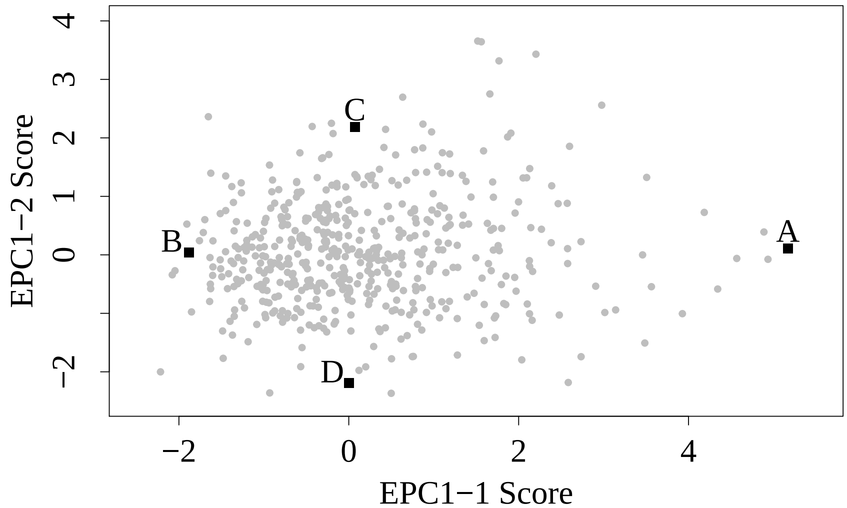



Figure 2.12: Scatter plot of EPC1-1 vs. EPC1-2 scores with extreme individuals identified

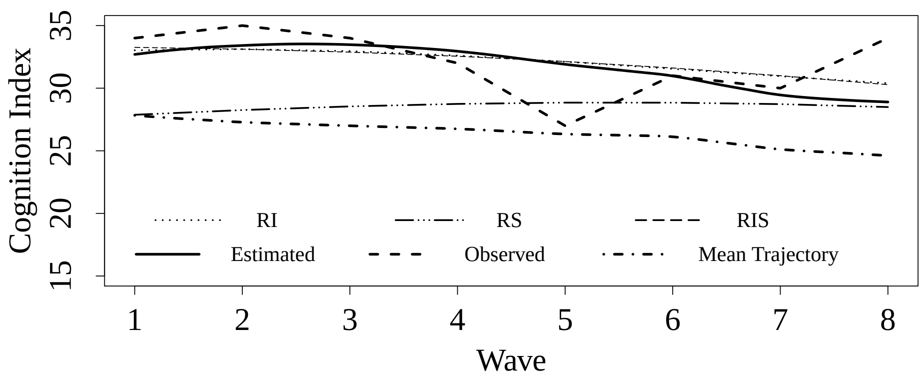



Figure 2.13: Observed, estimated, and mean trajectories with growth-curve model results for individual A with high positive EPC1-1 score

To further explain the individual level heterogeneity of cognitive degradation within each sub-population, it will be desirable to investigate how FPCA scores are correlated with different observed individual covariates, such as health conditions, socio-demographics and healthcare financing status. Table 2.3 summarizes the significant and interpretable individual covariates for all three sub-populations based on the regression analysis. The maximum allowable significance level is set as 0.1.

For sub-population 1, a higher value of EPC1-1 indicates a higher-than-average cognitive performance degradation over time. A risk/protective factor tends to decrease/increase

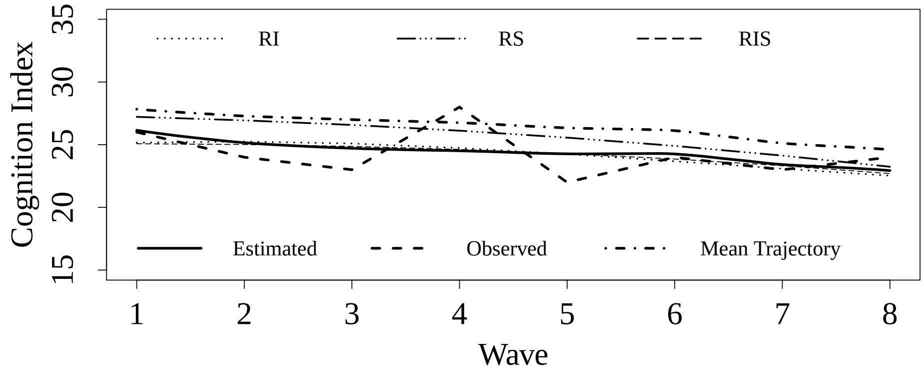



Figure 2.14: Observed, estimated, and mean trajectories with growth-curve model results for individual B with high negative EPC1-1 score

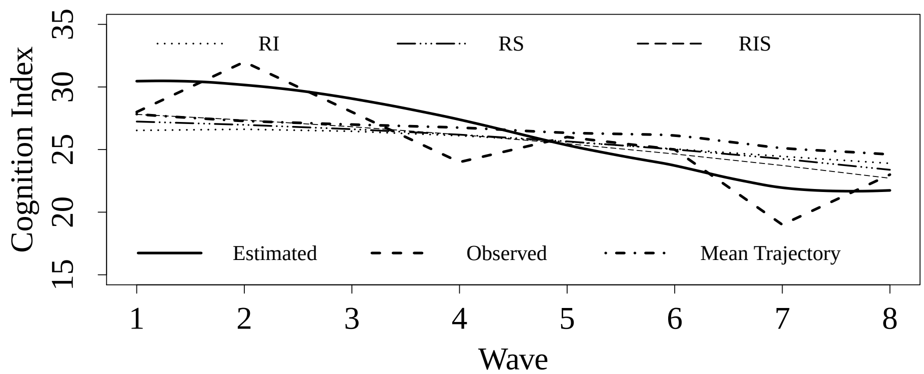



Figure 2.15: Observed, estimated, and mean trajectories with growth-curve model results for individual C with high positive EPC1-2 score

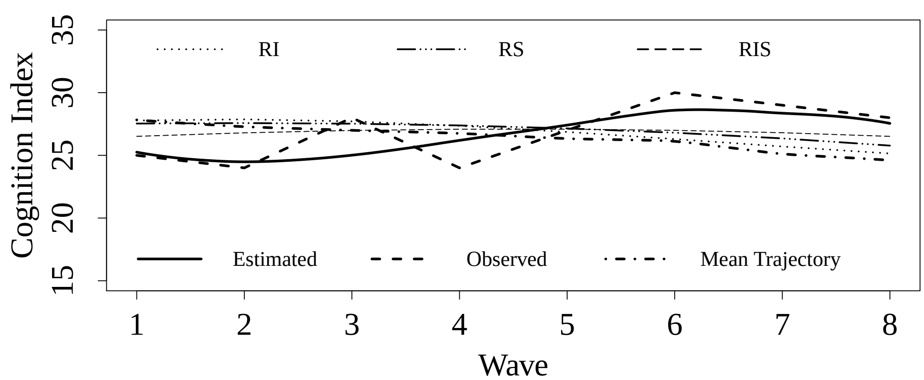



Figure 2.16: Observed, estimated, and mean trajectories with growth-curve model results for individual D with high negative EPC1-2 score

the EPC1-1 value. All significant covariates for EPC1-1 have negative effects and can be treated as risk factors. They can be interpreted as: an individual, who is older at initial interview, has a higher body mass index (i.e., higher level of obesity), covered under the Government Medicare insurance, and/or is not African America, Caucasian or Hispanic, will tend to experience a lower-than-average cognitive degradation over time. A higher value of EPC1-2 indicates a first higher-than-average then lower-than-average cognitive degradation over time. A risk/protective factor tends to increase/decrease the EPC1-2 value. Economic status and heart problem have negative and positive effects, respectively, and thus they can be treated as a protective factor and a risk factor, respectively.

For sub-population 2, a lower value of EPC2-1 indicates a higher-than-average cognitive performance degradation over time. Therefore, a risk/protective factor tends to increase/decrease the EPC2-1 value. Significant covariates with positive coefficients can be treated as risk factors, including age, smoking status and different ethnicity indicators. Economic status, doctors visits and long-term care insurance indicator can be treated as protective factors. A higher value of EPC2-2 indicates a first higher-than-average then lower-than-average cognitive degradation over time. Thus, home care and age can be treated as a protective factor and a risk factor, respectively.

For sub-population 3, a higher value of EPC3-1 indicates a higher-than-average cognitive performance degradation over time. Therefore, significant covariates with negative coefficients can be treated as risk factors, including age, hospital stay and different ethnicity indicators. Economic status and partial retiree indicator can be treated as protective factors. A higher value of EPC3-2 indicates a first higher-than-average then lower-than-average cognitive degradation over time. As observed, the number of living children (e.g., a indicator of informal care) and age can be treated as a protective factor and a risk factor, respectively. The above identified risk/protective factors and the evidence of economic, racial and long-term care support disparities provide important information for profes-

Table 2.3: Post-analysis results of identified covariates explaining individual heterogeneity (Standard Errors, SE, in parentheses)

| ***Sub-population 1 covariates*** | ***EPC1-1*** | ***EPC1-2*** |
|---|---|---|
| Age at initial interview in years | -0.061 (0.015)*** | |
| Body mass index in KG/m$^2$ | -0.023 (0.013). | |
| Government Medicare indicator (1 if insurance covered by Medicare, 0 otherwise) | -0.857 (0.289)** | |
| Other Race Indicator (1 if not African American Caucasian or Hispanic, 0 Otherwise) | -1.199 (-0.447)** | |
| Combined income of individual and spouse in USD | | -1.05E-06 (5.62E-07). |
| Heart problem indicator (1 if ever had heart problem, 0 otherwise) | | 0.225 (0.128). |
| ***Sub-population 2 covariates*** | ***EPC2-1*** | ***EPC2-2*** |
| Age at initial interview in years | 0.034 (0.008)*** | 0.022 (0.007)** |
| Combined income of individual and spouse in USD | -1.4E-06 (8.46E-07). | |
| Smoking indicator (1 if ever smoked, 0 otherwise) | 0.154 (0.071)* | |
| Number of Doctor visits in previous 2 years | -0.481 (0.164)** | |
| Long-term care insurance indicator (1 if covered by LTC insurance, 0 otherwise) | -0.240 (0.097)* | |
| Hispanic Indicator (1 if Hispanic, 0 otherwise) | 0.445 (0.190)* | |
| African American Indicator (1 if African American, 0 otherwise) | 0.391 (0.143)** | |
| Other Race Indicator (1 if not African American Caucasian or Hispanic, 0 Otherwise) | 0.642 (0.299)* | |
| Home healthcare indicator (1 if ever had healthcare at home in previous 2 years, 0 otherwise) | | -0.263 (-1.74). |
| ***Sub-population 3 covariates*** | ***EPC3-1*** | ***EPC3-2*** |
| Age at initial interview in years | -0.034 (0.018). | 0.079 (0.011)*** |
| Total of all assets owned in USD | 7.31E-07 (2.21E-07)*** | |
| Hospital stay indicator (1 if stayed overnight in hospital in previous 2 years, 0 otherwise) | -0.453 (0.218)* | |
| Hispanic Indicator (1 if Hispanic, 0 otherwise) | -1.318 (0.303)*** | |
| Partial retiree indicator (1 if retired partially, 0 otherwise) | 0.791 (0.414). | |
| African American Indicator (1 if African American, 0 otherwise) | -1.82 (0.234)*** | |
| Other Race Indicator (1 if not African American Caucasian or Hispanic, 0 Otherwise) | -0.908 (0.413)* | |
| Number of living children | | -0.06 (0.024)* |

***Notes:*** *1) . p<0.1; * p<0.05; ** p<0.01; *** p<0.001* .

*2) 95% Confidence Intervals for each parameter estimate are calculated by* $\hat{\kappa}_j \pm 1.96 \times SE(\hat{\kappa}_j)$*,*
*where* $\hat{\kappa}_j$ *is the respective estimated covariate coefficient.*

sionals in aging studies to develop effective health promotion programs and adjust the current long-term care policies.

## 2.4 Conclusion

A bi-level heterogeneity modeling and quantification framework is presented in this paper to systematically and comprehensively characterize the heterogeneous functional performance degradation of older adults. A Bayesian non-parametric model is first presented to characterize heterogeneous degradation at the sub-population level which bypasses the limitation of pre-specifying a known and fixed number of sub-populations in the conventional models. The resulting model is complex and makes conventional Bayesian estimation strategies challenging. An estimation algorithm is further developed to address the challenging issues of high dependency and infinite dimensionality in model estimation. Based on the proposed estimation algorithm, the number of sub-populations can be jointly identified with the parameters estimation in a one-step procedure. A numerical case study is provided to demonstrate the effectiveness of the proposed non-parametric model and compares it with parametric group-based modeling approaches. After quantifying the sub-population level heterogeneity, FPCA is employed to extract temporal variation signatures of degradation profiles for individuals within each sub-population. The scores of FPCA quantify the individual heterogeneity and allow post-analysis in investigating the influence of possible individual covariates. To illustrate the proposed work, a real case study of heterogeneous cognitive degradation is provided, using the HRS data. The sub-population level modeling identifies 3 distinct sub-populations with high, median and low levels of cognitive degradation. Within each sub-population, temporal variation signatures of individual heterogeneity are extracted and explained by several externals factors, such as demographic and socioeconomic factors, health conditions and long-term care supports. The proposed work successfully characterizes the heterogeneous functional performance degradation of the aging population. The modeling outputs will facilitate

LTC policy-makers and administrators to better project heterogeneous LTC demand and utilization over time of older adults in a region of interest or in a LTC facility and to further develop more proactive, targeted and adaptive LTC preparedness and service delivery decision-making strategies at both LTC network and facility levels. In this paper, the proposed Bayesian non-parametric model automatically identifies the number of sub-populations while the number of basis functions still needs to be pre-specified. One interesting future work is to further realize automatic and simultaneous identification of both the number of sub-populations and the number of basis functions required.

# Chapter 3

# Heterogeneous Length-of-stay Modeling of Post-acute Care Residents in the Nursing Home with Competing Discharge Dispositions

## 3.1 Introduction

Nursing homes (NHs), or skilled nursing facilities, are mainly responsible for caring the frail and vulnerable population of older adults with 24/7 personal care and assistance. Historically, NHs have been considered a major healthcare setting for providing custodial care for long-term care (LTC) residents. In recent decades, while maintaining their conventional role as LTC providers, modern NHs have increasingly been responsible for caring for post-acute care (PAC) residents who are recently hospitalized patients and require extended rehabilitation and recovery after acute care hospital stay. Recent studies reported that the percentage of residents (within NHs) admitted from the hospital increased from 67% in 2000 to 85% in 2015 (Fashaw *et al.*, 2020). Several policy and market changes contribute to the shifts of NH population composition. First, the rising trend towards rapid hospital discharge with the reduced hospital length-of-stay generated a growing population of quicker-and-sicker patients and drove many NHs to expand their sub-acute care and rehabilitation services (Murad, 2011). Second, Medicare covers qualified PAC residents with a higher reimbursement rate than Medicaid in NHs, the latter of which is the primary payer for qualified LTC residents who mainly require custodial care. In 2017, the national average reimbursement rate of Medicaid was $206 per resident day, less than half the rate paid by Medicare, $503 per resident day (NIC, 2018). Thus, there is a strong financial incentive for NH providers to accept more PAC Medicare beneficiaries.

Finally, due to the impact of a series of laws and initiatives (Ginsburg *et al.*, 1998; Moon, 1997; Eiken *et al.*, 2014) in recent decades, there has been a considerable increase in home- and community-based services for the overall LTC population to advocate age-in-place and divert or delay the expensive NH placement.

With the growing demand of PAC residents in NHs, the major goal of NH in caring PAC residents is to return them back to community successfully and efficiently with lower re/hospitalization rate. Thus, successful prediction of how long every individual PAC resident will stay in the NH and what will be his/her discharge disposition is of great importance for both NH administrators and individual residents (and their family). To NH administrators, successfully predicting length-of-stay (LOS) and the discharge disposition of each resident at individual level with individual risk factors identified will help administrators identify the most at-risk residents (e.g., residents with shorter LOS and higher re/hospitalization risk, residents with longer LOS and lower community discharge likelihood) and more targeted care can be provided to improve the care quality of the overall facility. At the facility level, an accurate predictive model of LOS with incorporation of varied individual characteristics will further allow accurate evaluation of the NH utilization (measured in average LOS) of a heterogeneous population of PAC residents with varied individual characteristics. To individual resident and his/her family, accurate prediction of how long he or she will stay in a NH will improve the communication between caregivers and care recipients, and will further assist the family of residents to better prepare the informal care resources to accommodate the needs of residents to be discharged.

Accurate prediction of LOS and discharge disposition of PAC residents is challenging. First, PAC residents admitted in NHs are often medically complex with a high level of functional dependence and with a variety of clinical diagnoses, ranging from severe orthopedic injuries (e.g., hip and pelvic fractures) to cardiovascular diseases (e.g., stroke, myocardial infarction, etc.). It is unclear what individual characteristics will affect LOS and discharge

disposition. Many of the existing LOS models in literature are distribution-based methods and considered various distributions, such as exponential, phase-type, log-normal, gamma, etc. (Xie *et al.*, 2005; Faddy *et al.*, 2009) , to characterize LOS of patients. They failed to take into account and quantify the influence of various possible individual characteristics for improving LOS prediction. Second, PAC residents have multiple possible discharge dispositions. They may be discharged to residential community for further recovery or transferred to hospital due to occurrence of critical events (e.g., infection, fall). Community discharge and re-hospitalization are mutually exclusive events; whichever comes first will terminate the NH dwelling duration of a resident. Existing LOS modeling approaches, such as regression-based methods (Carey, 2002; Kelly *et al.*, 2010; Kramer *et al.*, 2010) and machine learning methods (Hachesu *et al.*, 2013; Pendharkar *et al.*, 2011; Turgeman *et al.*, 2017) , mainly focused on predicting time-to-discharge without differentiating the disposition difference and overlooked the complexity arising in the competing risks between the dispositions. Thus, there is a need to develop an advanced LOS modeling approach for PAC residents that incorporates both individual characteristics and considers multiple competing discharge dispositions.

After realizing superior individual prediction of LOS and discharge disposition with relevant factors identified, there is still a need to evaluate the facility-level LOS and discharge outcome performance for a population of residents with varied individual characteristics. This will better inform NH or resource preparedness and evaluate the facility-level quality outcome. Computer simulation, such as discrete event simulation and agent-based simulation, are often considered in healthcare system engineering by modeling and simulating each individual patient as a discrete event or agent, and further evaluates the system level performance (e., average LOS of a population of individual at the facility level) (Hoot *et al.*, 2008; Taboada *et al.*, 2011; Wang *et al.*, 2012). Among these simulation approaches, a key step is to develop sampling algorithm for simulating LOS observations for a population of individuals. Existing sampling algorithms are only applicable to

simulating LOS observations characterized by distribution-based models ignoring various resident characteristics influencing the LOS (Zhang *et al.*, 2019; Cappanera *et al.*, 2014), or regression models (Austin *et al.*, 2002) without taking into account the multiple discharge dispositions. There is a need to develop a sampling algorithm which allows simulations of LOS observations of individuals with multiple competing discharge dispositions as well as varied individual characteristics.

To fill the aforementioned research gap, in this paper, we propose a heterogeneous LOS modeling approach for PAC residents by taking into account multiple discharge dispositions and incorporating the varied individual characteristics that may potentially influence each discharge disposition. At individual level, a semi-parametric hazard regression is considered to characterize the heterogeneous LOS observations of PAC residents with improved prediction accuracy on individual re/hospitalization risk and community discharge likelihood. Various factors affecting re/hospitalization risk and/or community discharge likelihood are also identified with their influence quantified. At facility level, a simulation algorithm is further developed to realize the simulation of both LOS observations and multiple discharge dispositions of a heterogeneous population of PAC residents with accurate predicted samples obtained. A real case study is provided to demonstrate the superior prediction performance of the proposed work at both individual and facility levels.

The remaining of the paper is organized as follows. The next section presents the proposed LOS modeling framework with individual-level LOS model characterizing influencing factors and incorporating competing discharge dispositions, and subsequent proposed sampling algorithm for facility-level performance evaluation. Then, a real-data case study is provided to demonstrate efficacy of the proposed work. Conclusive remarks are provided in the end.

## 3.2 Methodology

### 3.2.1 Model Formulation

Considering a cohort of $N$ PAC residents in a NH facility, each PAC resident may be discharged to residential community for further recovery or transferred to hospital due to occurrence of critical events (e.g., infection, fall, etc.). Community discharge and re/hospitalization are mutually exclusive events; whichever comes first will terminate the NH dwelling duration of a resident. Unlike many of existing LOS modeling works (Kelly *et al.*, 2010; Faddy *et al.*, 2009) which focused on modeling a single discharge disposition, this proposed model formulation aims to take into account the multiple and competing discharge dispositions of PAC residents, namely re/hospitalization and community discharge. Moreover, unlike many of existing discharge outcome prediction models, such as hospital re/admission models, which focused on predicting the risk of critical outcomes at a fixed time period, e.g, 30-days or 90-days re/hospitalization risk (Incalzi *et al.*, 1992), the proposed model will capture the re/hospitalization risk as well as community discharge likelihood of each individual PAC resident over time. To begin with, the instantaneous discharge rate of the $i$-th PAC resident with discharge disposition type $s$ can be characterized as,

$$d_{i,s}(t|\boldsymbol{x}_i) = \lim_{\Delta t \to \infty} \frac{\Pr(t \leq T_{i,s} \leq t + \Delta t | T_{i,min} \geq t, \boldsymbol{x}_i)}{\Delta t}, \quad i = 1, ..., n;\ s \in \{C, H\} \tag{3.15}$$

where $T_{i,min} = \min\{T_{i,C}, T_{i,H}\}$, $T_{i,min}$ is the LOS of resident $i$, $T_{i,C}$ and $T_{i,H}$ are latent time-to-discharge quantities with discharge disposition of community and hospital, respectively. $\boldsymbol{x}_i$ is a $\mathrm{p}_s$-dimensional observed vector which contains varied individual covariates that may potentially influence $d_{i,s}(t)$, such as individual demographics, clinical diagnoses,

cognitive deficits and physical functional performance. To explicitly associate the varied individual characteristics with $d_{i,s}(t)$, the hazard regression is considered as follows.

$$d_{i,s}(t|\boldsymbol{x}_i) = d_s^b(t)\exp(\boldsymbol{\beta}_s^T \boldsymbol{x}_i), \quad i = 1, ..., n;\ s \in \{C, H\} \tag{3.16}$$

where $d^b_{s(t)}$ is the population average instantaneous discharge rate with disposition $s$ in the absence of the influence of $\boldsymbol{x}_i$. $\boldsymbol{\beta}_s$ is $\mathrm{p}_s$-dimensional disposition-specific coefficient vector that quantify the influence of $\boldsymbol{x}_i$ on $d_{i,s}(t)$.

A benefit of the model in Equation (3.16) is that, for any time $t$, individual hospital re/admission risk till time $t$, i.e., $\Pr(T_{i,H} \leq t)$, and community discharge likelihood till time $t$, i.e., $\Pr(T_{i,C} \leq t)$, can be written as,

$$\Pr(T_{i,H} \leq t) = \int_0^t d_{i,H}(\tau|\boldsymbol{x}_i) \exp\left[-\int_0^\tau \sum_{s\in\{C,H\}} d_{i,s}(y|\boldsymbol{x}_i)\ \delta y\right]\delta t \tag{3.17a}$$

$$\Pr(T_{i,C} \leq t) = \int_0^t d_{i,C}(\tau|\boldsymbol{x}_i) \exp\left[-\int_0^\tau \sum_{s\in\{C,H\}} d_{i,s}(y|\boldsymbol{x}_i)\ \delta y\right]\delta t \tag{3.17b}$$

In other words, given a specific time period for $t$, the proposed model can always be converted into evaluating both the re/hospitalization risk and community discharge likelihood of resident $i$ at a fixed time period based on Equations (3.17a) and (3.17b). It bypasses the conventional discharge outcomes modeling approach which requires the discretization of discharge outcomes data in advance based on a specific time period and formulates a classification model for outcome prediction. It also allows the comparison of re/hospitalization risk and community discharge likelihood among different individuals with varied individual characteristics $\boldsymbol{x}_i$ over time.

### 3.2.2 Model Estimation

Given observed data $\boldsymbol{D} = \{t_i, z_{i,s}, \boldsymbol{x}_i\}_{i=1}^n$, where $z_{i,s} = 1$ if PAC resident $i$ is discharged to disposition $s$; 0 otherwise, $s \in \{C, H\}$, with unknown parameters/functions $\boldsymbol{\theta} = \cup_{s\in\{C,H\}}\boldsymbol{\theta}_s$, where $\boldsymbol{\theta}_s = \{d_s^b(t),\ \boldsymbol{\beta}_s\}$. The joint likelihood function $L(\boldsymbol{\theta}|\boldsymbol{D})$ can be written as

$$L(\boldsymbol{\theta}|\boldsymbol{D}) = \prod_{i=1}^{n} \prod_{s\in\{C,H\}} \left\{ d_{i,s}^b(t_i)\exp(\boldsymbol{\beta}_s^T\boldsymbol{x}_i)\exp\left[-\sum_{s\in\{C,H\}}\int_0^{t_i} d_{i,s}^b(\tau)\exp(\boldsymbol{\beta}_s^T\boldsymbol{x}_i)\delta\tau\right]\right\}^{z_{i,s}} \tag{3.18}$$

Let index set $A_s = \{i : z_{i,s} = 1\}$, $s \in \{C, H\}$, the joint likelihood function can be simplified as

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{D}) = \prod_{i\in A_C} d_{i,C}^b(t_i)\exp(\boldsymbol{\beta}_C^T\boldsymbol{x}_i) \cdot \prod_{i\in A_H} d_{i,H}^b(t_i)\exp(\boldsymbol{\beta}_H^T\boldsymbol{x}_i) \\ \cdot \prod_{i=1}^{n} \exp\left[-\sum_{s\in\{C,H\}}\int_0^{t_i} d_{i,s}^b(\tau)\exp(\boldsymbol{\beta}_s^T\boldsymbol{x}_i)\delta\tau\right] \end{aligned} \tag{3.19}$$

When $\boldsymbol{\theta}$'s are mutually exclusive, $L(\boldsymbol{\theta}|\boldsymbol{D})$ can be multiplicatively decomposed into $L(\boldsymbol{\theta}|\boldsymbol{D}) = \prod_{s\in\{C,H\}} L_s(\boldsymbol{\theta}_s|\boldsymbol{D})$, where $L_s(\boldsymbol{\theta}_s|\boldsymbol{D})$ can be expressed as

$$L_s(\boldsymbol{\theta}_s|\boldsymbol{D}) = \prod_{i\in A_s} d_{i,s}^b(t_i)\exp(\boldsymbol{\beta}_s^T\boldsymbol{x}_i) \cdot \prod_{i=1}^{n}\exp\left[-\int_0^{t_i} d_{i,s}^b\exp(\boldsymbol{\beta}_s^T\boldsymbol{x}_i)\,\delta t\right], \quad s \in \{C, H\} \tag{3.20}$$

Thus, maximizing $L(\boldsymbol{\theta}|\boldsymbol{D})$ can be equivalent to maximizing $L_s(\boldsymbol{\theta}_s|\boldsymbol{D})$ separately. To maximize $L_s(\boldsymbol{\theta}_s|\boldsymbol{D})$ by treating $d_s^b(t)$ as unknown function, we will first maximize the partial likelihood $L_s(\boldsymbol{\beta}_s|\boldsymbol{D})$ written as (Cox, 1972)

$$L_s(\boldsymbol{\beta}_s|\boldsymbol{D}) = \prod_{i\in A_s} \frac{\exp(\boldsymbol{\beta}_s^T\boldsymbol{x}_i)}{\sum_{j\in B(t_i)}\exp(\boldsymbol{\beta}_s^T\boldsymbol{x}_j)}, \quad s \in \{C, H\} \tag{3.21}$$

where $B(t_i)$ is a set of residents who are still in the NH before $t_i$.

Maximum likelihood estimation will be considered by solving $\max_{\boldsymbol{\beta}_s}\ log\ L_s(\boldsymbol{\beta}_s|\boldsymbol{D})$, which can be realized by numerical optimization algorithm, such as Newton-Raphson method (Gonzàlez *et al.*, 2008).

Based on the estimated $\boldsymbol{\beta}_s$, we will estimate $d_s^b(t)$ by maximizing the profile likelihood $L_s(d_s^b|\boldsymbol{D})$ written as

$$L_s(d_s^b|\boldsymbol{D}) \propto \prod_{i\in A_s} d_{s,i}^b \exp\left[-d_{s,i}^b \sum_{j\in B(t_i)} \exp(\boldsymbol{\beta}_s^T \boldsymbol{x}_j)\right], \quad s\in\{C,H\} \tag{3.22}$$

where $d_{s,i}^b = d_s^b(t_i),\ i\in A_s$, and $\hat{d}_s^b(t) = 0\ \forall t\ \notin \{t_i\}_{i\in A_s}$.

Thus, the profile maximum likelihood estimator can be obtained as (Cole *et al.*, 2014)

$$\hat{d}_{s(t_i)}^b = \frac{1}{\sum_{j\in B(t_i)} \exp(\boldsymbol{\beta}_s^T \boldsymbol{x_j})}, \quad s\in\{C,H\} \tag{3.23}$$

### 3.2.3 Sampling Algorithm

Based on the proposed model formulation and developed estimation algorithms as illustrated in Sections 3.2.1 and 3.2.2, the re/hospitalization risk and community discharge likelihood over time of any individual $i$ with individual characteristics $\boldsymbol{x}_i$ can be predicted. To further investigate the service utilization of the sample of PAC residents in a NH facility and evaluate the facility level performance given a heterogeneous population of PAC residents with varied individual characteristics, it becomes important to utilize computer simulation to mimic the patient flow of individual PAC residents and then evaluate the system level performance at the aggregate level. An essential basis of such computer simulation requires simulating the realization of LOS for each PAC resident. Existing simulation algorithm for LOS models may focus on simulation realization based on distribution-based models (New *et al.*, 2015; McGuire, 1994; El-Darzi *et al.*, 1998), such as Weibull distribution and log-normal distribution. For the developed semi-parametric

regression models with multiple competing discharge dispositions in Section 3.2.1, existing sampling algorithm is not applicable and there is a need to develop the corresponding simulation algorithm to facilitate generating predictive samples of LOS realizations given a heterogeneous population of PAC residents with varied individual characteristics $\boldsymbol{x}_i$'s. The developed sampling algorithm is summarized below.

$T_{i,s}$ of resident $i$ with disposition $s$ will be simulated as follows

---

**Algorithm 2** Proposed LOS sampling algorithm

---

**Step 1:** Compute $\phi_i(t|\boldsymbol{x}_i) = \sum_{s\forall\{C,H\}} \hat{d}_s^b(t) \exp(\hat{\boldsymbol{\beta}}_s^T \boldsymbol{x}_i)$, where $\phi_i(t|\boldsymbol{x}_i)$ is the instantaneous probability of resident $i$ with $\boldsymbol{x}_i$ being discharged at time $t$;

**Step 2:** Compute $S_i^{(l)} = \exp[-\sum_{p=0}^{l} \phi_i(t_{(p)}|\boldsymbol{x}_i)],\ l = 0, ..., N, N+1$, where $S_i^{(l)}$ is the probability of resident $i$ still residing in NH at time $t_{(l)}$; $t_{(0)} < t_{(1)} < ... < t_{(l)} < ... < t_{(N)} < t_{(N+1)}$ and $t_{(0)} = 0$, $t_{(N+1)} = +\infty$, $\{t_{(l)}\}_{l=1}^{n}$ are ordered distinct historical LOS observations;

**Step 3:** Randomly generate $\mu \sim \text{Unif}(0, 1)$

**Step 4:** Compute $l_1 = \max\{l : \mu \leq S_i^{(l)}\}$,

$l_2 = \min\{l : S_i^{(l)} \leq \mu\}$

and get simulated LOS, $T_i$ as:

$T_i = \frac{S_i^{(l_1)} - \mu}{S_i^{(l_1)} - S_i^{(l_2)}} \cdot (t_{(l_2)} - t_{(l_1)}) + t_{(l_1)}$

**Step 5:** Determine disposition state $s$ by drawing for categorical distribution as

$$w_s = \frac{\hat{d}_s^b(t)\exp(\hat{\boldsymbol{\beta}}_s \boldsymbol{x}_i)}{\phi_i(t|(x)_i)}, \quad s \in \{C, H\}$$

i.e., $s \sim Categorical(\boldsymbol{w})$, where $\boldsymbol{w} = [w_C, w_H]^T$

---

## 3.3 Case Study

### 3.3.1 Data Description

To demonstrate the performance of the proposed model and sampling algorithm, the Minimum Data Set (MDS) 3.0 of a certified nursing home (NH) in Tampa Bay Area, Florida, is considered. The MDS 3.0 is a rich data set containing comprehensive assessment

of clinical and functional status of all residents in a Medicare/Medicaid-certified NH during their stays. The data set is mandated federally and required by the CMS (CMS, 2017b). Each resident is assessed upon admission, periodically during the stay, and upon discharge or in case of any event causing significant change in status. Each assessment contains over 680 data covariates representing information on identification, admission and discharge dates, socio-demographics, financial details, various functional performance metrics, diseases and chronic conditions, medication and therapy information, discharge outcomes/dispositions of each resident, and facility administrative details.

The data collected includes stays of all residents admitted to the NH in the 1-year period. For this case study, a sub-cohort of PAC residents are selected according to "short-stay" criteria defined by the CMS (CMS, 2013). The CMS differentiates between "short-stay" and "long-stay" residents by examining episodes of care of the resident. An episode consists of one or more consecutive stays with breaks no more than 30 days. If the cumulative length of stay(s) in the NH is equal to or less than 100 days, the resident is labelled as a "short-stayer," otherwise considered as a "long-stayer." Most recent episode coinciding with the end-date of the consideration time period is used for the categorization. In the selected data, each data instance refers to the LOS observation of a short-stay resident with his/her individual characteristics.

A total of 710 LOS observations with complete information from 611 individual residents is selected for analysis. 98.02% of the LOS observations may be considered post-acute, meaning the resident was either admitted directly from the hospital to the NH, and/or covered under Medicare Part A insurance plan (Holup *et al.*, 2017). LOS observations with discharges to community or hospital disposition are included. LOS observations with other discharge dispositions, such as death, another facility, etc., are excluded since they form a very small portion of the dataset and inadequate for significant model building. Left-truncated and/or right-censored observations are also neglected due to a small portion.

Table 3.4 provides a summary of descriptive statistics of the selected cohort and stays, which includes socio-demographics (e.g., age, gender, ethnicity, marital status, etc.), care utilization details of the stay (e.g., length of stay (LOS), admission origin, discharge disposition, payment source, etc.), and health characteristics (e.g., body measures, various functional performance scores, disease and chronic conditions, etc.). The calculated mean LOS of the short-stay residents was 20.33 days, with a majority of 97.3% being admitted from hospital, and a majority of 79.9% being discharged to community and the rest 20.1% being readmitted to hospital. The discharge disposition, socio-demographics and health characteristics form possible covariates influencing LOS of the resident and consequentially, their care service utilization.

### 3.3.2 Feature Selection

Since the MDS dataset contains numerous elements of data, it is necessary to consider information directly relevant to the LOS. Guided by domain knowledge in NH care, a subset of data most related to care utilization i.e., socio-demographics, functional performance scores, disease diagnoses and chronic conditions observed on admission are considered. Although MDS data monitors the stay over time, only assessments upon admission are relevant to predict LOS of an unknown cohort of residents in the facility, which are also known as baseline observations.

After summarizing the data (i.e., calculating LOS, deriving various functional performance scores and converting categorical covariates into dummy variables) and preprocessing (i.e., removing low-frequency covariates, removing one from each highly correlated pair, and checking for multicollinearity), 68 covariates relevant to LOS are selected. It is still a large number of covariates and incorporating all of them for developing a predictive model may yield data over-fitting (Hughes *et al.*, 1968). To further reduce the dimensionality of the input variables, a collection of 10 popular feature selection algorithms is applied to the dataset.

Table 3.4: Descriptive summary statistics of the selected resident cohort

| *Characteristics* | *Mean (SD) or %* |
|---|---|
| Number of stays | 710 |
| Number of residents | 611 |
| Demographics: | |
| Age at admission, years | 76.68 (10.66) |
| Gender: Female | 64.40% |
| Race: | |
| Black or African American | 6.90% |
| White | 90.30% |
| Marital status: | |
| Never married | 14.20% |
| Married | 35.60% |
| Widowed | 33.20% |
| Divorced | 15.60% |
| Care utilization | |
| Length of stay (LOS), days | 20.33 (15.72) |
| Admission Origin: | |
| Community | 2.00% |
| Hospital | 97.30% |
| Discharge disposition: | |
| Community | 79.90% |
| Hospital | 20.10% |
| Primary payer: Medicare Part A | 59.00% |
| Health Characteristics | |
| Height, inches | 65.36 (4.13) |
| Weight, pounds | 170.95 (54.45) |
| ADL score | 6.48 (3.85) |
| Mood/depression score | 0.91 (1.24) |
| Cognitive score | 13.17 (2.96) |
| Visual impairment | 15.90% |
| Hearing impairment | 22.70% |
| Incontinence – urinary | 54.50% |
| Incontinence – fecal | 46.20% |
| Fall within past 180 days | 32.70% |
| Fracture within past 180 days | 18.00% |
| Diseases | |
| Cancer | 8.00% |
| Heart/Circulation | 79.40% |
| Gastrointestinal | 38.60% |
| Genitourinary | 25.20% |
| Metabolic | 73.80% |
| Musculoskeletal | 42.00% |
| Neurological | 26.20% |
| Psychiatric/Mood Disorder | 45.90% |
| Pulmonary | 34.60% |

SD: Standard Deviation; ADL: Activities of Daily Living

The feature selection methods include 4 linear feature selection algorithms, namely, stepwise regression (Stepwise AIC), recursive feature elimination (RFE), simulated annealing (SA), and regularized linear regression (LASSO). Each selection algorithm implements different procedures to find the best subset of covariates. For example, Stepwise AIC trains linear regression models by progressively adding covariates and evaluating model performance with Akaike Information Criterion (AIC). In contrast, RFE ranks all covariates and progressively removes unimportant ones, training and reevaluating a linear model at each step. Simulated annealing performs a random heuristic search for best combination in the covariate space. LASSO regression penalizes unimportant covariates to zero coefficient value with L1-norm regularization.

Moreover, 6 non-linear tree-based and ensemble feature selection algorithms are applied, namely, Filtering: Random Forest, RFE: Bagged Trees, RFE Random Forest, Genetic algorithm: Random Forest, SA: Random Forest, and Boruta. Filtering employs a preprocessing step to test strength of individual relationship between each covariate and the response variable before training a predictive model. RFE, Genetic and SA employ subset selection heuristic similar to that applied in training linear models. However, tree-based algorithms are trained instead of linear models at each step. In each case, the tree-based model with highest accuracy evaluated identifies the best subset of covariates.

The most frequent significant covariates influencing LOS are identified by all the above models, which results in a union set of 35 covariates that can be utilized in further predictive modeling. Table 3.5 displays the final selected covariates, and Table 3.6 shows significant covariates identified by each feature selection algorithm.

Table 3.5: List of selected covariates (35) from feature selection and corresponding descriptions

| *Covariate Short Name* | *MDS 3.0 Covariate Description* |
|---|---|
| **Age** | Age at admission, years |
| **ADL score** | Activities of daily living score at admission |
| **Cognitive score** | Brief Interview for Mental Status (BIMS) summary score at admission |
| **Mood score** | Resident mood interview PHQ-9© total severity score at admission |
| | **Active diagnosis indicator of resident at admission:** |
| **Cancer** | Cancer (with or without metastasis) |
| **Anemia** | Anemia (e.g., aplastic, iron deficiency, pernicious, and sickle cell) |
| **Atrial Fibrillation** | Atrial Fibrillation or Other Dysrhythmias (e.g., bradycardias and tachycardias) |
| **Coronary Artery Disease** | Coronary Artery Disease (CAD) (e.g., angina, myocardial infarction, and atherosclerotic heart disease (ASHD)) |
| **DVT, PE, PTE** | Deep Venous Thrombosis (DVT), Pulmonary Embolus (PE), or Pulmonary Thrombo-Embolism (PTE) |
| **Heart Failure** | Heart Failure (e.g., congestive heart failure (CHF) and pulmonary edema) |
| **Hypertension** | Hypertension |
| **GERD or Ulcer** | Gastroesophageal Reflux Disease (GERD) or Ulcer (e.g., esophageal, gastric, and peptic ulcers) |
| **UC, CD, IBD** | Ulcerative Colitis, Crohn's Disease, or Inflammatory Bowel Disease |
| **BPH** | Benign Prostatic Hyperplasia (BPH) |
| **Renal disease** | Renal Insufficiency, Renal Failure, or End-Stage Renal Disease (ESRD) |
| **Neurogenic Bladder** | Neurogenic Bladder |
| **Obstructive Uropathy** | Obstructive Uropathy |
| **MDRO** | Multidrug-Resistant Organism (MDRO) |
| **Pneumonia** | Pneumonia |
| **Septicemia** | Septicemia |
| **Diabetes Mellitus** | Diabetes Mellitus (DM) (e.g., diabetic retinopathy, nephropathy, and neuropathy) |
| **Hyponatremia** | Hyponatremia |
| **Hyperlipidemia** | Hyperlipidemia (e.g., hypercholesterolemia) |

Table 3.5: Continued

| *Covariate Short Name* | *MDS 3.0 Covariate Description* |
|---|---|
| **Arthritis** | Arthritis (e.g., degenerative joint disease (DJD), osteoarthritis, and rheumatoid arthritis (RA)) |
| **Hip fracture** | Hip Fracture (e.g., sub-capital fractures, and fractures of the trochanter and femoral neck) |
| **Other Fracture** | Other Fracture |
| **Alzheimer's Disease** | Alzheimer's Disease |
| **Non-Alzheimer's Dementia** | Non-Alzheimer's Dementia (e.g. Lewy body dementia, vascular or multi-infarct dementia; mixed dementia; frontotemporal dementia such as Pick's disease; and dementia related to stroke, Parkinson's or Creutzfeldt-Jakob diseases) |
| **Hemiplegia or Hemiparesis** | Hemiplegia or Hemiparesis |
| **Malnutrition** | Malnutrition (protein or calorie) or at risk for malnutrition |
| **Anxiety Disorder** | Anxiety Disorder |
| **Depression** | Depression (other than bipolar) |
| **Schizophrenia** | Schizophrenia (e.g., schizoaffective and schizophreniform disorders) |
| **PTSD** | Post-Traumatic Stress Disorder (PTSD) |
| **Respiratory Failure** | Respiratory Failure |

Table 3.6: Frequency of covariates identified as significant by corresponding feature selection algorithms

| *Covariate* | *Vote* | ***Feature Selection Models*** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Linear** | | | | **Tree-based** | | | | | |
| | | Stepwise AIC | RFE: Linear | SA: Linear | LASSO | Filtering: Random Forest | RFE: Bagged Trees | RFE: Random Forest | Genetic: Random Forest | SA: Random Forest | Boruta |
| | | | | | | **Community** | | | | | |
| **ADL score** | 80% | • | | • | • | • | • | • | • | | • |
| **Cognitive score** | 70% | • | | | | • | • | • | • | • | • |
| **Mood score** | 80% | • | | • | • | • | • | • | • | | • |
| **Cancer** | 70% | • | • | | • | | • | • | • | • | |
| **Anemia** | 70% | | | • | • | • | • | • | • | | • |
| **Atrial Fibrillation** | 60% | | | • | | • | • | • | • | • | |
| **Coronary Artery Disease** | 30% | | | | | • | • | | • | | |
| **DVT, PE, PTE** | 80% | • | • | | • | • | • | • | • | | • |
| **Heart Failure** | 70% | | | | • | • | • | • | • | • | • |
| **Hypertension** | 100% | • | • | • | • | • | • | • | • | • | • |
| **GERD or Ulcer** | 40% | | | | | • | • | • | | | • |
| **UC, CD, IBD** | 30% | • | • | | • | | | | | | |
| **BPH** | 50% | | • | | • | • | • | • | | | |
| **Renal disease** | 80% | • | • | • | • | • | • | • | | | • |
| **Neurogenic Bladder** | 60% | | • | • | • | | | • | • | | • |
| **Obstructive Uropathy** | 50% | | • | • | • | | • | | • | | |
| **MDRO** | 40% | | • | | • | | | • | | | • |
| **Pneumonia** | 50% | | • | | • | • | • | | • | | |
| **Diabetes Mellitus** | 40% | | | | • | | • | • | • | | |
| **Hyponatremia** | 90% | • | • | | • | • | • | • | • | • | • |
| **Hyperlipidemia** | 60% | | | | • | • | • | • | • | | • |
| **Arthritis** | 30% | | | | | • | • | • | | | |
| **Hip fracture** | 90% | • | • | • | • | • | • | • | | • | • |
| **Other Fracture** | 60% | • | • | | • | • | • | • | | | |
| **Alzheimer's Disease** | 60% | • | • | | • | | • | • | | | • |
| **Non-Alzheimer's Dementia** | 90% | • | • | | • | • | • | • | • | • | • |
| **Hemiplegia or Hemiparesis** | 80% | • | • | | • | • | • | • | • | | • |
| **Malnutrition** | 30% | | | | | | • | • | • | | |
| **Anxiety Disorder** | 30% | | | • | | | • | | • | | |
| **Depression** | 40% | | | | | • | • | • | | | • |
| **Schizophrenia** | 60% | • | • | • | • | • | • | | | | |
| **PTSD** | 40% | | | | • | • | • | | | • | |
| **Respiratory Failure** | 30% | | • | • | • | | | | | | |

Table 3.6: Continued

| *Covariate* | *Vote* | ***Feature Selection Models*** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Linear** | | | | **Tree-based** | | | | | |
| | | Stepwise AIC | RFE: Linear | SA: Linear | LASSO | Filtering: Random Forest | RFE: Bagged Trees | RFE: Random Forest | Genetic: Random Forest | SA: Random Forest | Boruta |
| | | | | | | **Hospital** | | | | | |
| **Age** | 30% | • | | | • | | • | | | | |
| **ADL score** | 50% | • | | • | | | • | • | | • | |
| **Cognitive score** | 20% | | | | | | • | • | | | |
| **Hypertension** | 50% | • | | • | | | | • | | • | • |
| **GERD or Ulcer** | 30% | | | • | | | • | | | • | |
| **Obstructive Uropathy** | 20% | | | • | | | | | | • | |
| **Septicemia** | 20% | | | | | | | | • | | • |
| **Hyponatremia** | 20% | • | | • | | | | | | | |
| **Hyperlipidemia** | 30% | | | | | | • | • | | | • |
| **Arthritis** | 30% | • | | | | | • | | • | | |
| **Non-Alzheimer's Dementia** | 40% | • | | | | | | • | | • | • |
| **Hemiplegia or Hemiparesis** | 20% | | | • | | | | | | • | |
| **PTSD** | 30% | | | | | | • | | • | | • |
| **Respiratory Failure** | 40% | • | | • | | | | | | • | • |

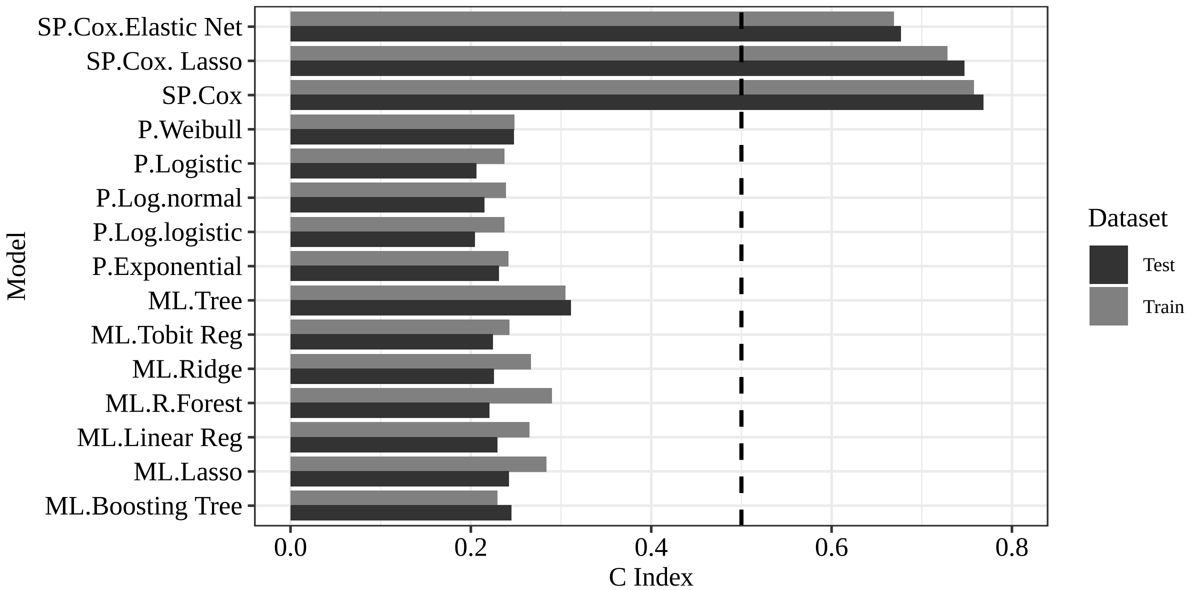
### 3.3.3 Prediction Performance Comparison

To compare the prediction performance between the proposed model and alternative prediction methods in literature, the dataset is split randomly into 90% training and 10% test sets of observations with stratified sampling to preserve proportion of discharge dispositions. It is to be noted that the previous section of feature selection is conducted based on the training data set without touching the independent test data set. The proposed model is compared to others by evaluating the C-index of training and testing sets for each discharge dispositions, namely community and hospital (Harrell *et al.*, 1996; D'Agostino *et al.*, 2003). A C-index value beyond 0.5 indicates the model is consistently satisfactory in predicting discharge risks rather than making random predictions. A higher C-index value indicates the better predictive capability. Several semi-parametric and parametric survival models are compared under the competing risk framework, where the characteristic of competing discharge dispositions, i.e. community and hospital, is incorporated. Semi-parametric models include Cox regression with LASSO, or Elastic Net regularization, where the baseline hazard is non-parametric, and regularization attempts to avoid over-fitting. Parametric models include Weibull, Logistic, Log-normal, Log-logistic, and Exponential hazard regression, where the baseline hazards are parameterized based on the named distribution. Furthermore, several alternative regularized/unregularized linear and tree-based data-mining methods independent of competing risk is considered, namely, Linear Regression, LASSO regression, Ridge regression, Tobit regression, Decision Tree, Boosting Tree, and Random Forest. Thus, a total of 15 different models are evaluated for each of the discharge dispositions. Table 3.7 provides the list of models considered with their abbreviations and corresponding training and testing C-index values. Figure 3.17 further visualizes the results.

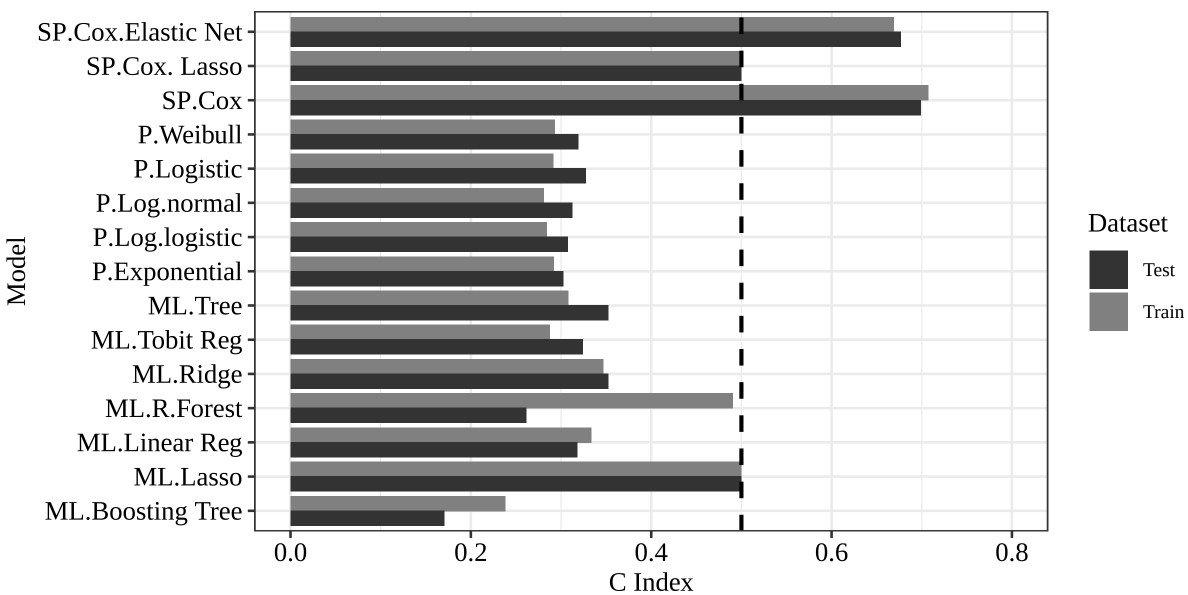
As observed in Figure 3.17a, for predicting LOS before discharge to community, the proposed model outperforms all other models with a training and testing C-index values of about 0.76. Regularized Cox regression models, i.e., LASSO and Elastic Net, yield lower

Table 3.7: Prediction performance (C-index) comparison between proposed and alternative models

| Model Family | Model Short Name | Model Description | Discharge Dispositions: Community Train | Community Test | Hospital Train | Hospital Test |
|---|---|---|---|---|---|---|
| With competing risks assumption | | | | | | |
| Semi-parametric survival | **SP.Cox** | **Proposed:** Cox regression with non-parametric baseline hazard | 0.758 | 0.768 | 0.707 | 0.699 |
| | **SP.Cox. Lasso** | Cox regression with L1-regularization | 0.728 | 0.747 | 0.5 | 0.5 |
| | **SP.Cox.Elastic Net** | Cox regression with L1/L2-mixed regularization | 0.669 | 0.677 | 0.669 | 0.677 |
| Parametric survival | **P.Exponential** | Survival regression with exponential baseline hazard | 0.241 | 0.231 | 0.292 | 0.303 |
| | **P.Weibull** | Survival regression with Weibull baseline hazard | 0.248 | 0.248 | 0.293 | 0.319 |
| | **P.Logistic** | Survival regression with Logistic baseline hazard | 0.237 | 0.206 | 0.291 | 0.327 |
| | **P.Log.logistic** | Survival regression with Log-logistic baseline hazard | 0.237 | 0.204 | 0.284 | 0.308 |
| | **P.Log.normal** | Survival regression with Log-normal baseline hazard | 0.239 | 0.215 | 0.281 | 0.313 |
| Independent of competing risks | | | | | | |
| Data mining: Linear | **ML.Tobit Reg** | Tobit regression | 0.243 | 0.224 | 0.288 | 0.324 |
| | **ML.Linear Reg** | Linear regression | 0.265 | 0.229 | 0.334 | 0.318 |
| | **ML.Lasso** | Linear regression with L1-regularization | 0.284 | 0.242 | 0.5 | 0.5 |
| | **ML.Ridge** | Linear regression with L2-regularization | 0.266 | 0.225 | 0.347 | 0.352 |
| Data mining: Tree-based | **ML.R.Forest** | Random forest regression | 0.29 | 0.22 | 0.491 | 0.261 |
| | **ML.Boosting Tree** | Boosting tree regression | 0.229 | 0.245 | 0.238 | 0.17 |
| | **ML.Tree** | Decision tree regression | 0.305 | 0.311 | 0.308 | 0.352 |




(a) Time to community




(b) Time to community

Figure 3.17: Prediction performance (C-Index) comparison between models with respect to discharge dispositions

C-index values. Both their values are still above 0.5, indicating that Cox baseline hazard is flexible in representing the LOS data with improved prediction performance. The reduced performance with regularization suggests that penalization of covariate coefficients in the Cox model is not necessary, possibly since an optimal set of covariates have been chosen based on the previous step of feature selection. Parametric survival models have poorer performances than Cox model family with C-index values ranging from 0.21 to 0.25, much lower than 0.5, indicating the models are consistently poor at prediction than random chance. Cox model family outperforms parametric survival models since its baseline hazard is non-parametric and able to capture LOS data with more flexibility. Regularized/unregularized linear models perform slightly better than survival models with C-index values ranging between 0.22-0.28. Conversely, tree-based methods perform better than linear models with decision tree and random forest producing C-index around 0.3. The improvement achieved from tree-based methods indicate non-linear relationship between LOS and the covariates.

Similarly, from Figure 3.17b, the proposed model outperforms all the others for predicting LOS before transferring to hospital. The performance patterns are similar to that of community discharge likelihood with a few irregularities. LASSO regularization in Cox regression performs poorly for predicting hospitalization risk, since a minimum penalty term was not found that improves prediction better than random chance. Within linear models, LASSO regularization produces improved results but still inadequate for prediction. Within tree-based models, decision tree produces the best result. In general, the testing C-index value is lower than the training C-index value, since the test data set is serving as an independent data set untouched during the model development phase to evaluate future prediction performance of the model.

### 3.3.4 Identification of Risk/Protective Factors

Apart from producing superior prediction performance, the proposed competing risk Cox regression model identifies important risk/protective factors that influence the resident's LOS. Table 3.8 show significant covariates identified by the proposed model influencing discharge risk to community and hospital respectively. The significance level, $\alpha$, is set at 0.05.

Table 3.8: Significant covariates influencing LOS identified by Cox regression

| ***Covariate*** | $\hat{\kappa}_j$ | $SE(\hat{\kappa}_j)$ | ***p-value*** | |
|---|---|---|---|---|
| **Time to community** | | | | |
| ADL score | -0.113 | 0.014 | 1.11E-15 | *** |
| Mood score | -0.143 | 0.041 | 0.0005 | *** |
| Cancer | -0.436 | 0.162 | 0.0072 | ** |
| Anemia | -0.203 | 0.097 | 0.0367 | * |
| Hypertension | -0.559 | 0.1 | 0 | *** |
| BPH | -0.515 | 0.141 | 0.0003 | *** |
| Renal disease | -0.33 | 0.134 | 0.0136 | * |
| MDRO | -0.628 | 0.295 | 0.0331 | * |
| Hip fracture | -0.604 | 0.233 | 0.0096 | ** |
| Other Fracture | -0.421 | 0.127 | 0.0009 | *** |
| Non-Alzheimer's Dementia | -0.448 | 0.14 | 0.0014 | ** |
| Hemiplegia or Hemiparesis | -0.846 | 0.239 | 0.0004 | *** |
| Malnutrition | -0.556 | 0.196 | 0.0045 | ** |
| **Time to hospital** | | | | |
| ADL score | 0.087 | 0.024 | 0.00026 | *** |
| Anemia | 0.482 | 0.18 | 0.00727 | ** |
| Obstructive Uropathy | 1.028 | 0.307 | 0.0008 | *** |
| Diabetes Mellitus | 0.503 | 0.17 | 0.00311 | ** |

**Notes:** 1) * p<0.05; ** <0.01; 2) $SE$: Standard Error;
3) 95% Confidence Intervals for each parameter estimate are calculated by $\hat{\kappa}_j \pm 1.96 \times SE(\hat{\kappa}_j)$, where $\hat{\kappa}_j$ is the respective estimated covariate coefficient.

The magnitude and sign of the coefficient values quantifies influence of the covariate on the probability of being discharged. A higher probability of being discharged implies a shorter LOS, and vice versa. For a resident being ultimately discharged to community, if

s/he has higher ADL score, higher Mood score, or any of the disease diagnoses, her/his discharge likelihood decreases due to the negative sign and increases the LOS. Alternately, if a resident is ultimately transferred to hospital, having a higher ADL, or being diagnosed with Anemia, Uropathy or Diabetes increases their hospitalization risk due to their positive signs, implying a shorter LOS before being admitted to hospital. For both community and hospital dispositions, ADL is the most significant factor on influencing LOS, confirming the intuition that residents with high dependency for daily living activities require greater care. ADL also has opposite effect between dispositions, illustrating the importance of incorporating multiple discharge dispositions. Such identified risk/protective factors are valuable for the healthcare provider to better identify and target on the most "at-risk" NH residents with more focused care and resources.

### 3.3.5 Marginal Effects of Covariates on Community Discharge Likelihood and Hospitalization Risk

As opposed to a single value to quantify predicted LOS obtained from many existing predictive models, the proposed competing risk Cox regression model further provides information on disposition-specific probability of being discharged. Such information can be visualized and compared by plotting survival probability (1-the probability of being discharged) against time. Furthermore, since Cox regression is a proportional hazards model, marginal effects of survival curves can be visualized under different values of covariates. Based on such survival curves, the influence of each individual covariate on the probability of being discharged can be visualized. Further, the probability of being discharged over time among different individuals with different individual characteristics can be visualized and compared as well. Figures 3.18 and 3.19 provide examples of marginal effects of different baseline ADL values on the LOS of an example resident over time for specific discharge dispositions. All the variables other than ADL score is fixed at the mean level of the observed sample.

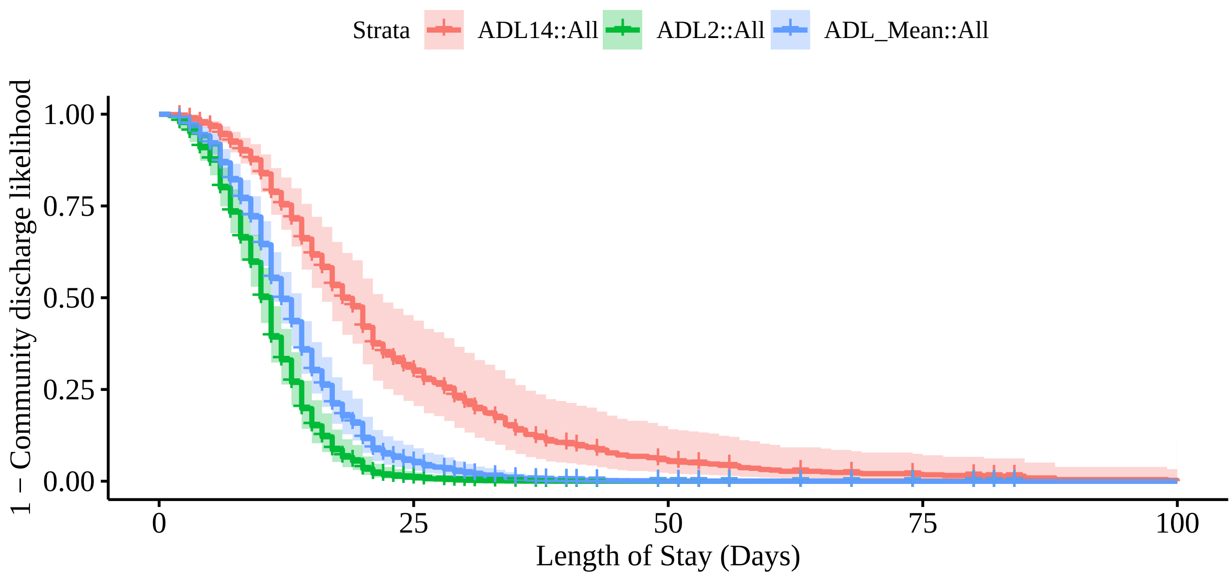



Figure 3.18: Marginal effect of ADL score on survival curves for community discharge disposition

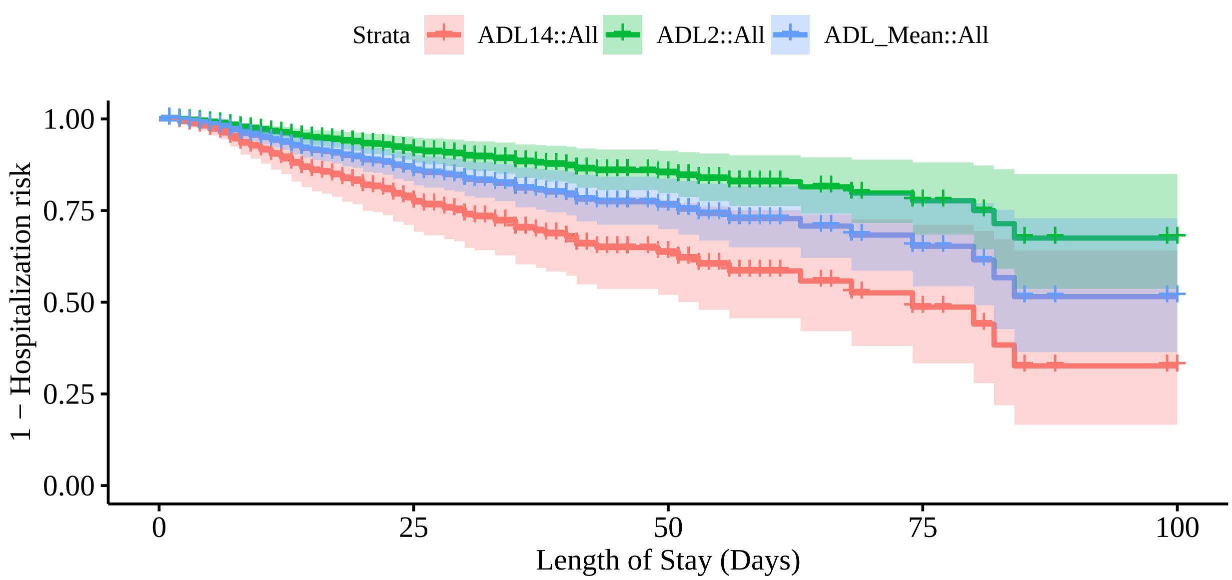



Figure 3.19: Marginal effect of ADL score on survival curves for hospital discharge disposition

As observed in Figure 3.18, a resident with disposition to community with higher baseline ADL score (more physical functional dependency) has a curve (red) higher than the average (blue) resident. In such case, probability to remain in the facility is higher than average at any point in time, increasing the LOS. In contrast, a resident with lower baseline ADL score (more functionally independent), has a survival curve (green) lower than average and tends to have shorter LOS. Figure 3.19 shows survival curves for resident with disposition to the hospital. ADL score has opposite effect on the curves, reaffirming the competing risk assumption. A higher baseline ADL score results in shorter stay, and a lower one increases the stay. Since the curves evolves differently over time, it is possible to assess the probability of being discharged at any time point during the resident's stay.

Similarly, a resident's disposition outcome can also be examined over time for different combinations of different individual characteristics. For instance, in Figure 3.20, a hypothetical resident with better health (low ADL score, mood/depression score, and low number of diseases diagnosed at baseline) tend to have shorter LOS with disposition as community (red curve) but longer LOS with disposition as the hospital (blue). In contrast, in Figure 3.21 a hypothetical resident with worse health (high ADL, mood/depression score, and higher number of diseases diagnosed at baseline, tend to remain in the facility for a very long time for recovery before being discharged to community (red), and relatively short stays if being transferred to hospital (blue).

### 3.3.6 Performance of the Proposed Sampling Algorithm for Generating LOS Predictive Samples

Survival models are different in predicting response than conventional data mining models. The former model family provides the probability of being discharged over time for each of the resident, while the latter models provide single LOS response values. To simulate residents flow in a typical NH facility using computer simulation, an important step is to accurately simulate LOS predictive samples. Our proposed sampling algorithm

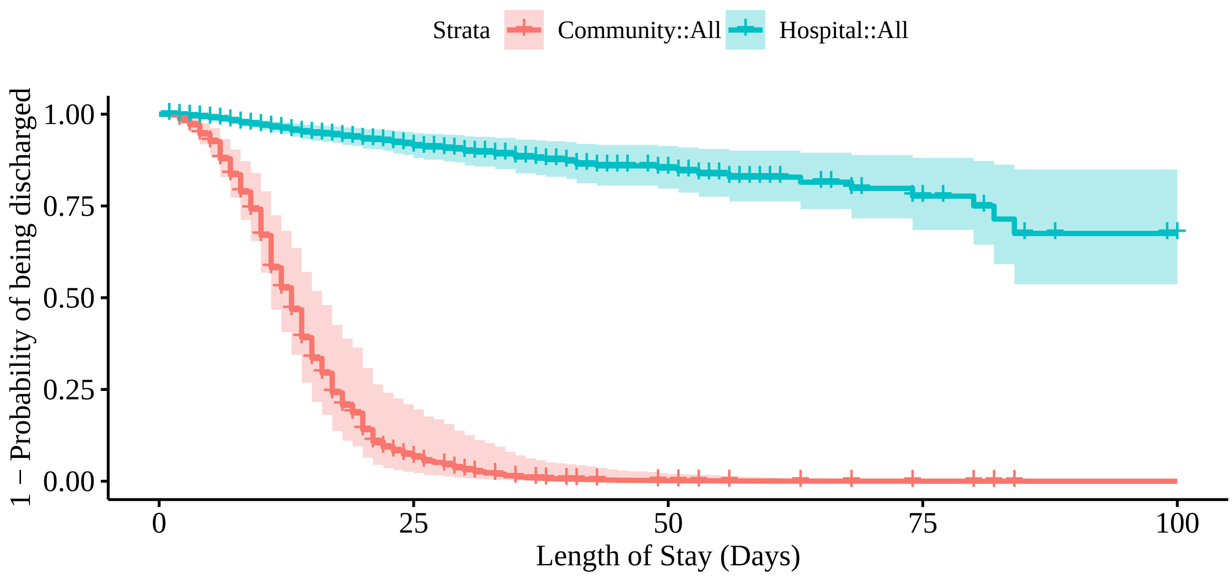



Figure 3.20: Marginal effect of health on survival curves of a healthy resident for different discharge dispositions

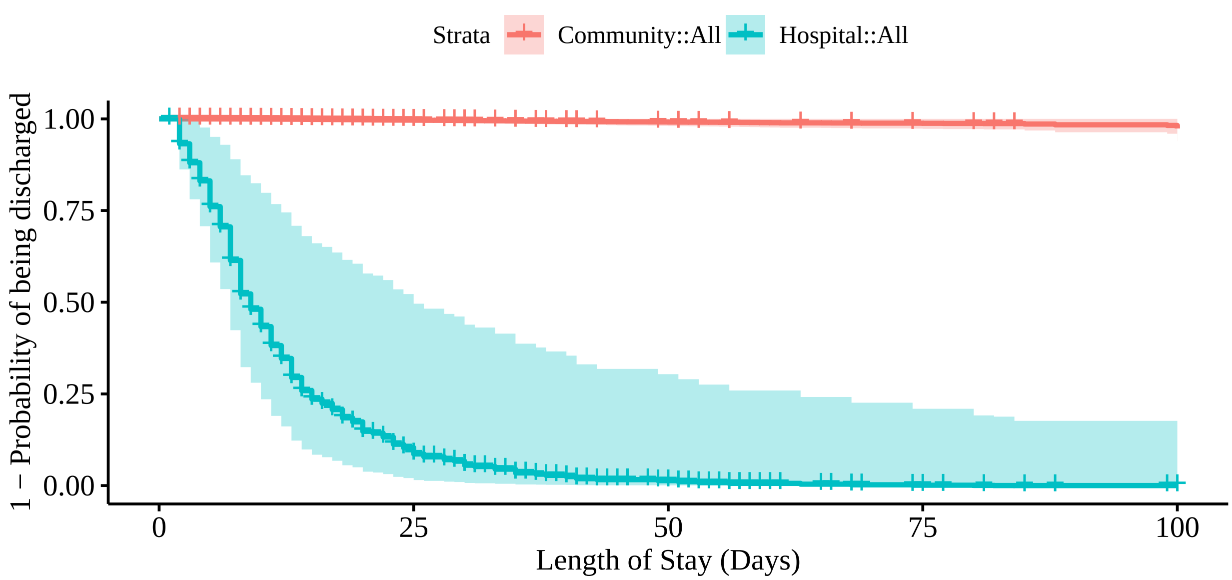



Figure 3.21: Marginal effect of health on survival curves of an unhealthy resident for different discharge dispositions

is able to generate predictive LOS samples accurately and simultaneously provide the corresponding discharge dispositions as well. Sampling performance can be evaluated by comparing survival plots of observed LOS samples and simulated LOS samples. The survival curves are calculated by the Kaplan-Meier curves which provide the disposition-specific observed and simulated survival probability of the sample over time. In Figure 3.22, survival curves are compared for each disposition and full dataset. The sampling algorithm is very effective in generating predictive samples of LOS, since the simulated (light-colored) curves are very close to their observed counterparts (dark-colored). The green curves (light and dark) are slightly lower than full dataset (black and grey), indicating residents transferred to hospital tend to have shorter LOSs than average, as opposed to blue curves indicating residents discharged to community have slightly longer LOSs than average. Figure 3.23 shows performance of the sampling algorithm in predicting discharge dispositions with 100% classification accuracy.

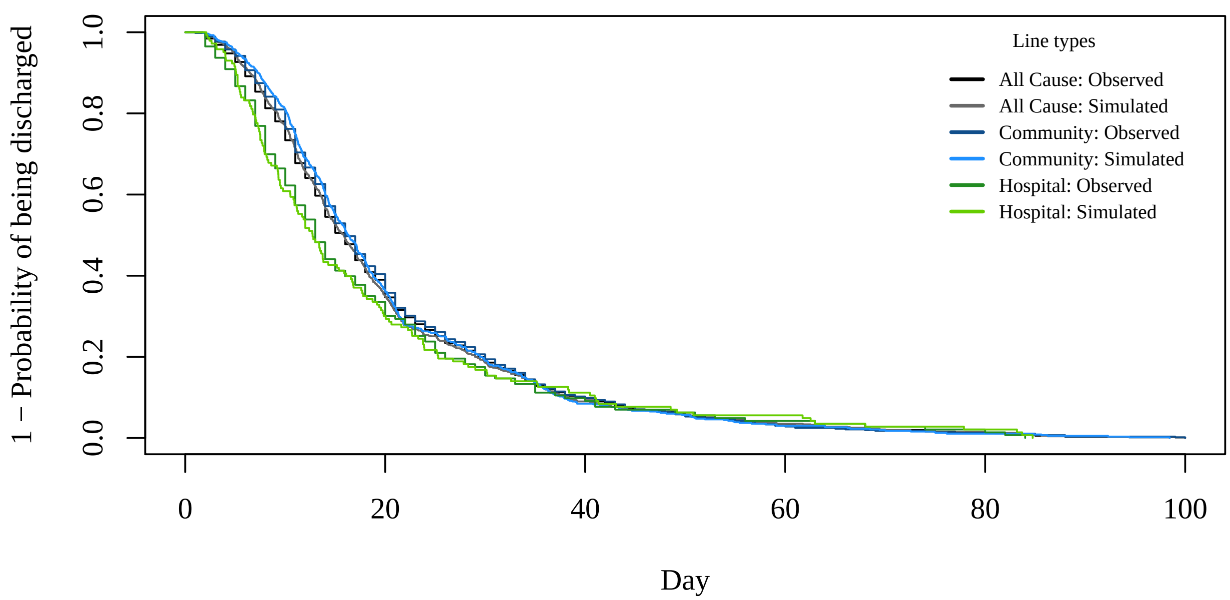



Figure 3.22: Sampling algorithm prediction performance

The accuracy of the proposed algorithm is compared with several other survival simulation models and data mining methods. Simulation is performed Cox Weibull regression and Lognormal regression, where the baseline hazard functions are fitted with

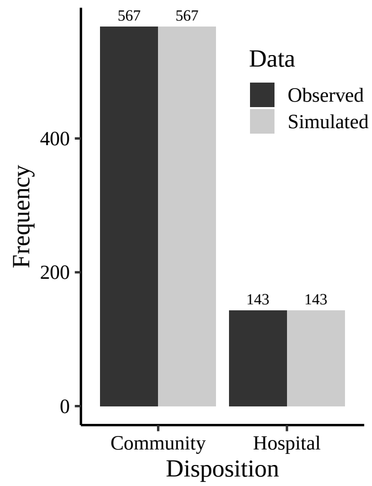



Figure 3.23: Discharge disposition prediction

Weibull and Lognormal distributions respectively and under competing risk paradigm. As shown in the survival curves in Figures 3.24, 3.25, and 3.26, Cox Lognormal regression performs poorly in generating samples as compared to the observed data, while Cox Weibull performs better due to increased flexibility in fitting the LOS data. The prediction performance of the proposed work is also compared with popular linear and tree-based data mining models, namely Linear regression, L1-regularized linear regression (LASSO), Decision tree, and Random Forest. Overall, the proposed work generates the most accurate LOS samples as compared to other methods due to the incorporation of non-parametric baseline hazard as well as the proposed simulation algorithm.

### 3.3.7 Simulation-based Facility-level Performance Evaluation

The proposed sampling algorithm is not only able to predicting probability of being discharged to a specific discharge disposition for a specific individual resident, it can be further utilized to generate predictive LOS samples of a heterogeneous population of NH residents with varied individual characteristics. This will allow the users to evaluate the system level performance of a NH facility given a census composition scenario of a heterogeneous population of NH residents.

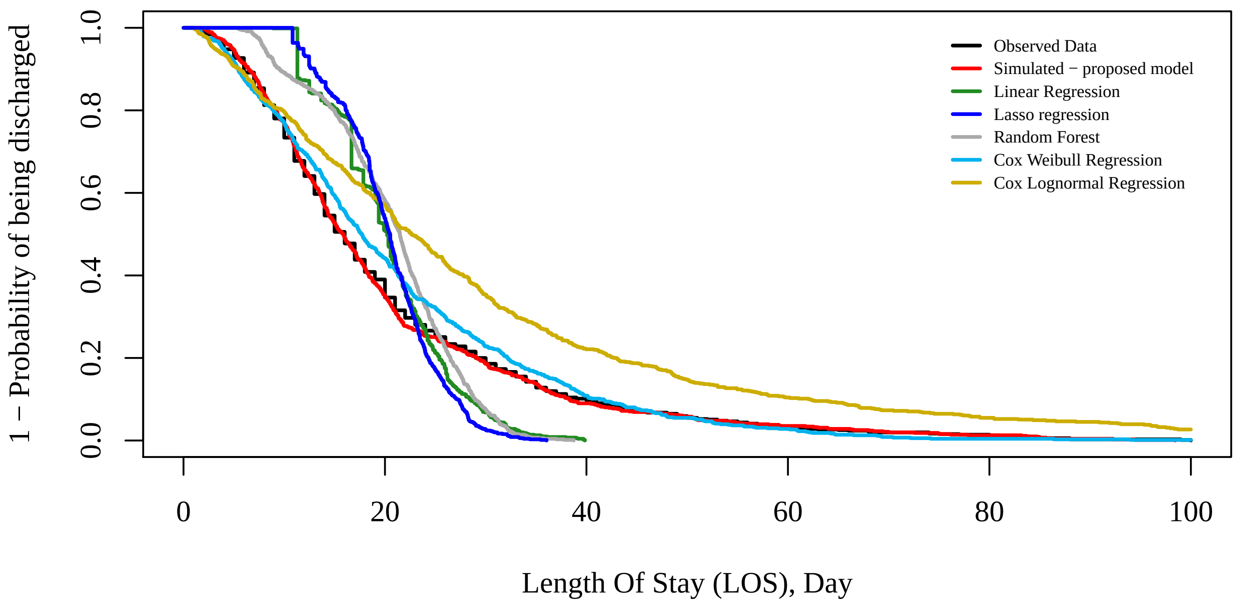



Figure 3.24: Comparison of prediction performance between proposed sampling algorithm and alternative models for all discharge dispositions

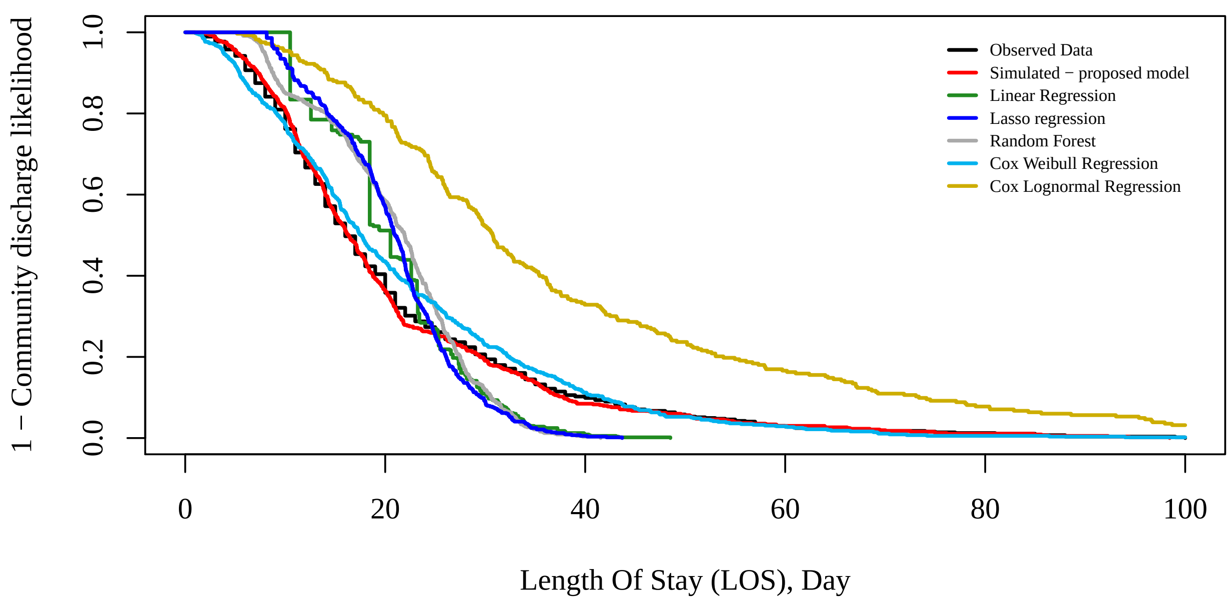



Figure 3.25: Comparison of prediction performance between proposed sampling algorithm and alternative models for community discharge disposition

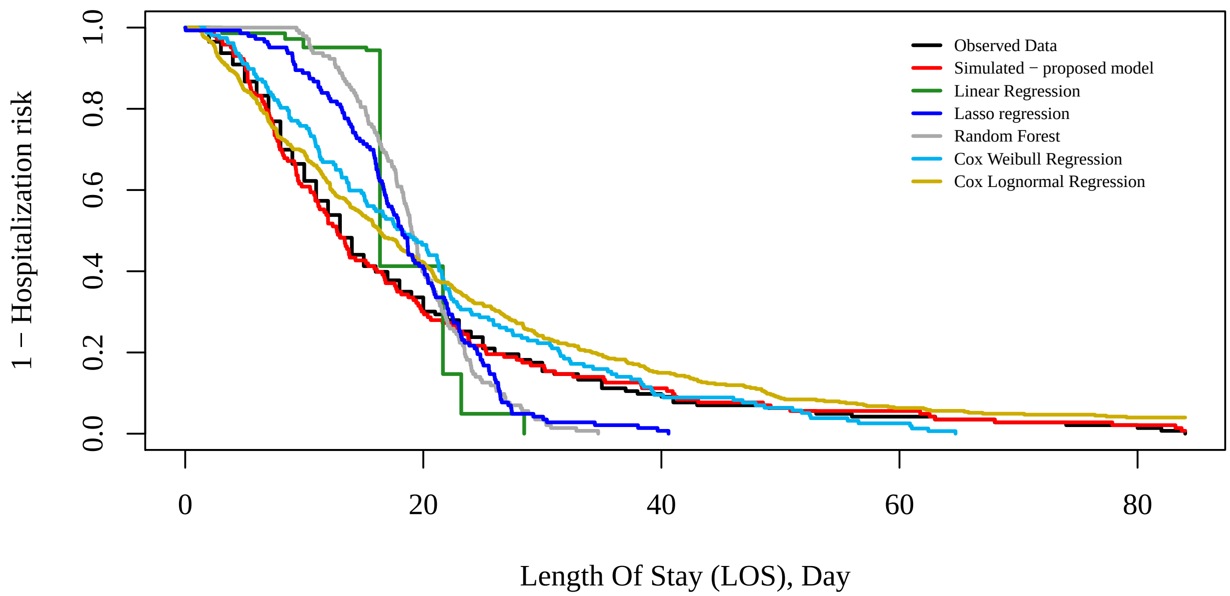



Figure 3.26: Comparison of prediction performance between proposed sampling algorithm and alternative models for hospital discharge disposition

To explain the above functionality of the proposed work, 8 different cohorts of residents are defined in an increasing order of acuity. Simulation data is generated for each cohort using different segmentation and distributions of the significant covariates identified by Cox regression in Table 3.8. The setting for each cohort is provided in Table 3.9. For each acuity scenario, 1000 random admission observations are generated in the following process: 1) ADL score is randomly generated with a truncated normal distribution with mean fixed at desired level corresponding to acuity and upper and lower limits set with range of observed data, 2) Mood score is similarly randomly generated with truncated normal distribution, 3) For each observation, any of the 13 diseases are randomly selected and binary value (0 or 1) is generated by Bernoulli trial, where the rate of success is sampled from a Beta prior distribution with shape parameters set according to desired acuity level. The sampling process is repeated 20 times for each acuity scenario. Mean LOS and disposition-specific discharge rates are estimated with standard errors for each. The results obtained is summarized in Table 3.10 and Figure3.27.

Table 3.9: Experimental settings of acuity scenarios in simulation study

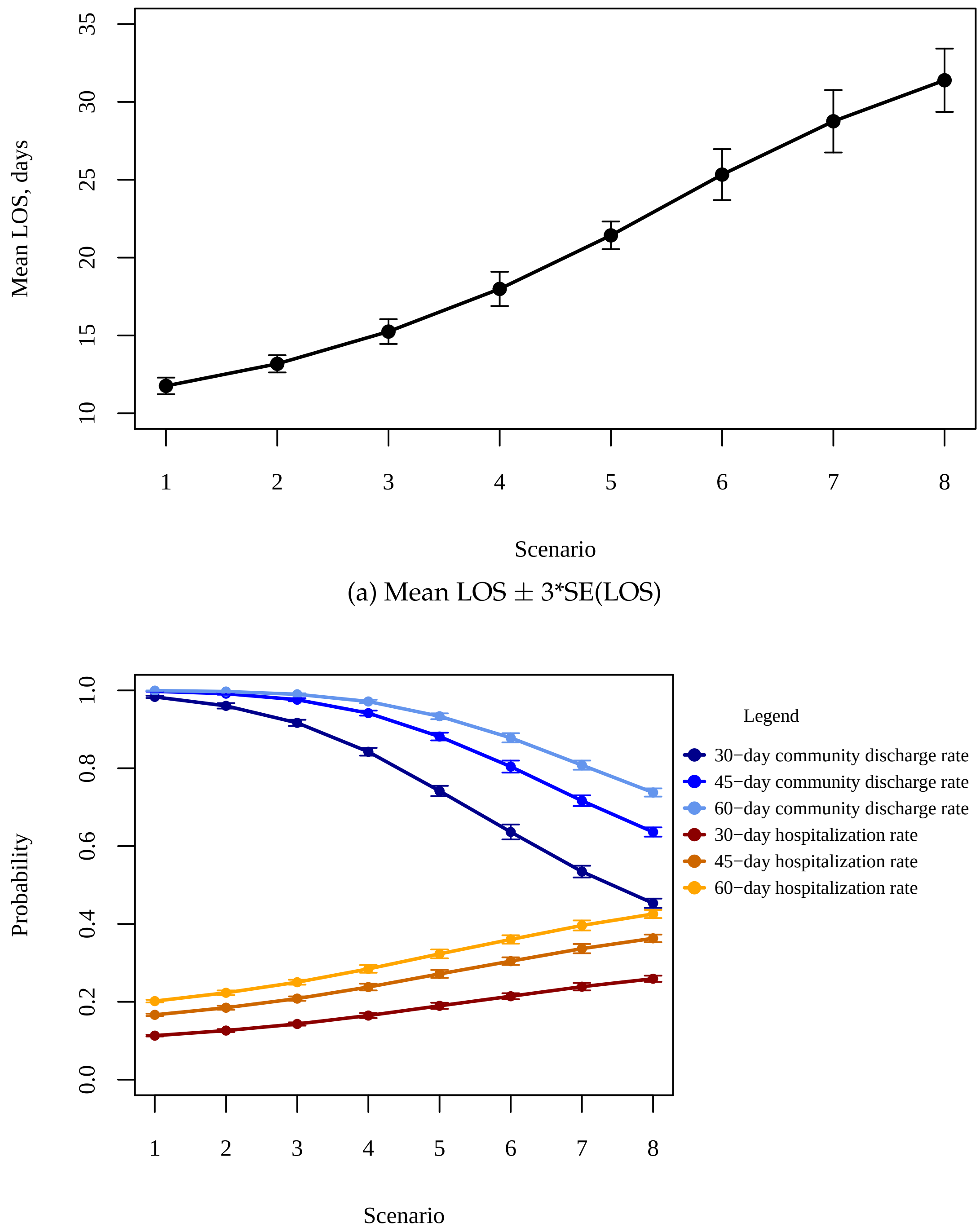
| ***Covariate*** | ***Distribution*** | ***Parameter*** | ***Acuity scenario*** **Less acute** ***1*** | ***2*** | ***3*** | ***4*** | ***5*** | ***6*** | **More acute** ***7*** | ***8*** | **Limits** | **Std. Dev.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ADL score | Truncated normal | Mean | 1 | 3 | 5 | 7 | 9 | 11 | 13 | 15 | [0,16] | 4 |
| Mood score | Truncated normal | Mean | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | [1,8] | 2 |
| Disease incidence prior | Beta | Mean | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | [0,1] | |
| | | Std. Dev. | 0.09 | 0.121 | 0.138 | 0.148 | 0.151 | 0.148 | 0.138 | 0.121 | | |

**Note:** Diseases include – Cancer, Anemia, Hypertension, BPH, Renal disease, MDRO, Hip fracture, Other Fracture, Non-Alzheimer's Dementia, Hemiplegia or Hemiparesis, Malnutrition, Obstructive Uropathy, Diabetes Mellitus

Table 3.10: Results of acuity scenarios in simulation study

| ***Metric*** | | ***Measure*** | ***Acuity scenario*** **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|---|---|
| LOS | | Mean | 11.77 | 13.21 | 15.25 | 18.02 | 21.37 | 25.4 | 28.74 | 31.36 |
| | | SE | 0.2 | 0.18 | 0.23 | 0.32 | 0.33 | 0.54 | 0.72 | 0.66 |
| Disposition: Community | 30-day discharge rate | Mean | 0.98 | 0.96 | 0.92 | 0.84 | 0.74 | 0.64 | 0.53 | 0.45 |
| | | SE | 0.0009 | 0.0023 | 0.0024 | 0.0042 | 0.0051 | 0.0066 | 0.0055 | 0.0043 |
| | 45-day discharge rate | Mean | 1 | 0.99 | 0.98 | 0.94 | 0.88 | 0.8 | 0.72 | 0.64 |
| | | SE | 0.0002 | 0.001 | 0.0012 | 0.0026 | 0.0038 | 0.0056 | 0.005 | 0.0045 |
| | 60-day discharge rate | Mean | 1 | 1 | 0.99 | 0.97 | 0.93 | 0.88 | 0.81 | 0.74 |
| | | SE | 9.43E-05 | 0.0005 | 0.0007 | 0.0017 | 0.0027 | 0.0045 | 0.0042 | 0.0041 |
| Disposition: Hospital | 30-day discharge rate | Mean | 0.11 | 0.13 | 0.14 | 0.16 | 0.19 | 0.21 | 0.24 | 0.26 |
| | | SE | 0.0007 | 0.0013 | 0.0015 | 0.0025 | 0.003 | 0.0029 | 0.0037 | 0.003 |
| | 45-day discharge rate | Mean | 0.17 | 0.18 | 0.21 | 0.24 | 0.27 | 0.3 | 0.34 | 0.36 |
| | | SE | 0.001 | 0.0018 | 0.0021 | 0.0034 | 0.0038 | 0.0037 | 0.0047 | 0.0038 |
| | 60-day discharge rate | Mean | 0.2 | 0.22 | 0.25 | 0.28 | 0.32 | 0.36 | 0.39 | 0.43 |
| | | SE | 0.0012 | 0.0021 | 0.0024 | 0.0039 | 0.0042 | 0.004 | 0.0052 | 0.0041 |

SE: Standard Error




(a) Mean LOS ± 3*SE(LOS)



(b) Mean disposition-specific discharge rates ± 3*SE(•)

Figure 3.27: Performance metrics of acuity simulation

The proposed sampling algorithm generates LOS and predicts discharge disposition for each simulated resident in each cohort. As seen in Figure 3.27a, as acuity increases,

mean LOS increases in the samples. More residents are discharged to hospital; however, residents being discharged to community have increasingly longer stays, increasing the mean LOS of the samples across acuity scenarios. The whiskers represent dispersion among samples in each acuity and are mostly non-overlapping indicating significant difference of LOS between acuity scenarios. As evident in Figure 3.27b, increasing acuity has a sharper decreasing effect on community discharge rates over time, while hospital discharge rates increase at a more gradual rate. The phenomenon occurs because since a greater number of diseases influence community discharge likelihood than hospitalization risk. The results further emphasize the competing nature of the dispositions – as acuity increases, LOS tends to increase for residents discharged to community, but LOS tends to decrease for residents transferred to hospital. Depending on the proportion of the residents finally discharged to community or transferred to hospital, the mean LOS varies accordingly. The sampling algorithm successfully mimics the phenomenon. Furthermore, the algorithm can also provide disposition-specific probability of being discharged over time in a continuous time scale for the collective cohort and individual resident, allowing greater understanding of facility utilization and resident outcome (i.e. re/hospitalization risk) over the course of the stay. Figure 3.27b shows a few of the discharge rate curves at discrete times of 30-, 45-, and 60-days.

### 3.4 Conclusion

In this paper, a heterogeneous LOS modeling approach was proposed by considering multiple discharge dispositions and incorporating varied individual characteristics for NH PAC residents. At individual level, several popular predictive models, such as machine learning and survival models, are considered to predict LOS and their performances are compared with the proposed model. The proposed model outperformed other models by jointly predicting the re/hospitalization risk and community discharge likelihood over time. It is also capable of identifying disposition-specific risk/protective factors for

influencing the disposition-specific probability of being discharged over time. Furthermore, to enable facility-level performance evaluation of the NH, a novel simulation algorithm is proposed for generating LOS predictive samples of residents by incorporating varied individual characteristics and competing discharge dispositions. The proposed algorithm is capable of accurately generating samples a heterogeneous population of NH residents with varied individual characteristics, which allows the evaluation of facility performance measures such as re/hospitalization rate and mean LOS. A real case study based on de-identified data from a NH in Tampa Bay area is considered to illustrate the proposed work and demonstrate its superior prediction performance. The proposed approach will allow NH administrators and health practitioners to identify the most at-risk residents and design more targeted care delivery, facilitate optimal resource allocation strategies at the facility level for achieving greater quality outcomes at reduced costs, and further improve communication of prognostic information among everyone involved in the care delivery process.

# Chapter 4

# A Predictive Analytics Approach for Nursing Home Hurricane Evacuation

## 4.1 Introduction

Skilled nursing facilities, or nursing homes (NHs), are responsible for caring the frailer, older adults by providing 24/7 personal medical care and daily living assistance. Most older adults in the NHs suffer from significant functional (e.g., physical, cognitive, social) limitations, aging-related disabilities, vision/hearing impairments and multiple chronic disease, making them highly vulnerable to natural disasters, such as hurricanes (Morrow, 1999; Fernandez *et al.*, 2002; Shaughnessy *et al.*, 1990). Their impaired mobility, diminished sensory awareness, and chronic health conditions make them less likely to respond and adapt appropriately during hurricanes, leaving their lives clearly at risk to the aftermath of hurricanes, such as physical damage of NH infrastructures, storm surge and massive flooding, power outage, and disruption of medical supplies. Existing studies show that both the mortality and morbidity of NH residents significantly increase during hurricanes (Dosa *et al.*, 2010, 2012).

Due to devastating threats and negative consequences of hurricanes on vulnerable NH residents, many NHs have to evacuate and move their frail residents away from hazard regions to safer places. However, whether to evacuate a NH or not is one of the most complex and difficult decisions encountered by NH administrators. The prospects of not evacuating in response to hurricanes can be tragic. For instance, 34 residents were presumed to have drowned at St Rita's NH in Chalmette, LA, after its facility owners refused to evacuate before landfall of Hurricane Katrina (Harris *et al.*, 2005). In recent

Hurricane Irma, 14 residents died from overheat after the hurricane knocked out power to the air conditioner at Hollywood Hills NH in Hollywood, FL. On the other hand, however, evidence also shows that evacuation had an adverse effect on health and well-being of many frail and impaired residents since the disruption associated with evacuation, changes of environment and care routines, and the trauma of moving itself may result in physical injuries, functional declines and depression, which further complicates NH evacuation (Thomas *et al.*, 2012). Successful modeling and prediction of NH evacuation response (i.e., evacuating or sheltering in place) of NH administrators is of great importance. It will enrich the understanding of the multi-factorial complexity of NH evacuation decision by identifying and quantifying the effects of different internal and external factors with an evidence base that is informative and critical to disaster preparedness and response. It will also assist the local emergency authorities better plan and manage healthcare resources to meet with the NH evacuation demand surge in a more proactive manner.

In the existing literature of both qualitative studies and quantitative studies in investigating evacuation response, many of studies mainly focused on studying the evacuation choices of community-dwelling households from general population (Baker, 1991; Wolshon *et al.*, 2005; Whitehead *et al.*, 2000; Hasan *et al.*, 2011). Baker (1991) studied a number of hurricanes in the Atlantic states from Texas through Massachusetts occurring between 1961 and 1989, where sample surveys from the general population was used to identify characteristics, such as, hazardousness of the region, public service, residence type, perceived risk, and general storm severity, etc., influencing aggregate-level evacuation rates. Wolshon *et al.* (2005) combined a survey on state evacuation plans performed by Louisiana State University and then current literature to summarize evacuation policies and procedures implemented by state authorities from a transportation service utilization and response perspective for the general population. Whitehead *et al.* (2000) examined prospective hurricane evacuation behavior of North Carolina coastal residents following occurrence of Hurricane Bonnie through telephone surveys, and concluded that storm

severity, reception of evacuation order, possibility of flooding, housing structure, and socio-demographic disparity to important determinants of evacuation. Hasan *et al.* (2011) considered post-storm damage assessment data of households affected by Hurricane Ivan and characterized various factors affecting evacuation behavior.

Many of these studies considered a single source of data in the context of non-disaster conditions by extracting aggregated individual characteristics without explicitly considering the rich information form actual storms. Unlike the above studies which focused on studying the healthy individuals from general population, we will focus on studying the evacuation response of NH population at the organization level and each organization consisting of frail older adults with complex health conditions and functional limitations. There is a need to incorporate both internal factors that could comprehensively describe the different aspects (e.g., staffing, dwelling residents) of an organization and further integrate them with the highly heterogeneous geo-spatial characteristics of NHs in the context of actual disaster conditions. In the existing long-term care literature for studying NH evacuation, many focused on conceptual and qualitative studies (Dosa *et al.*, 2007, 2008; Blanchard *et al.*, 2009) based on descriptive statistics or narrative summary, and they often utilize a single source of data, such as retrospective surveys and telephone questionnaire, without taking in to account the spatial heterogeneity of environmental characteristics of NHs and quantifying the influence of environmental conditions on evacuation response. Some of existing qualitative studies consider linear statistical models (Brown *et al.*, 2012) in quantifying the influence of different inputs factors for prediction performance output of evacuating or sheltering-in-place. However, they only consider limited environmental characteristics at aggregate level. Further, the linear models developed in these quantitative studies may not be appropriate in capturing the potential nonlinear relationship between various inputs and the evacuation response and the prediction performance accuracy will be greatly undermined.

To fill the aforementioned research gap, in this paper, we propose a GIS-integrated predictive analytics framework for evacuation response prediction of NHs by integrating multi-source data from NH residents, NH facilities and environmental conditions in the context of real disaster scenario. In particular, we extract multiple GIS features to comprehensively characterize the spatially heterogeneous environmental conditions (e.g., both geographical condition and storm conditions) of NHs at different spatial locations in the state of Florida. With the incorporation and integration of such rich GIS information with NH resident and staffing characteristics, we considered different linear and nonlinear machine learning methods to achieve the improved evacuation response prediction of NHs in real disaster scenarios. The influence of environmental conditions on NHs evacuation are further quantified explicitly in the presence of varied characteristics of individual NH facility.

The remaining of the paper is organized as follows. In the next section, we will introduce the proposed methodology of extracting various GIS features and NH features as well as the development of different machine learning models to integrate the extracted features. Then, we will give a concrete real-world example using the recent disaster scenario of Hurricane Irma to compare the prediction performance of different machine learning models and emphasize the prediction performance benefits of incorporating multiple GIS features extracted. The model interpretation results based on linear classification model will be also discussed. Conclusions are provided in the end.

### 4.2 Methodology

To develop an predictive analytics method for investigating the multifactorial nature of NH evaluation response and further predicting the evaluation response of NH facilities, the features extracted (to be explained below) are based on the following conceptual model described in Figure 4.28.

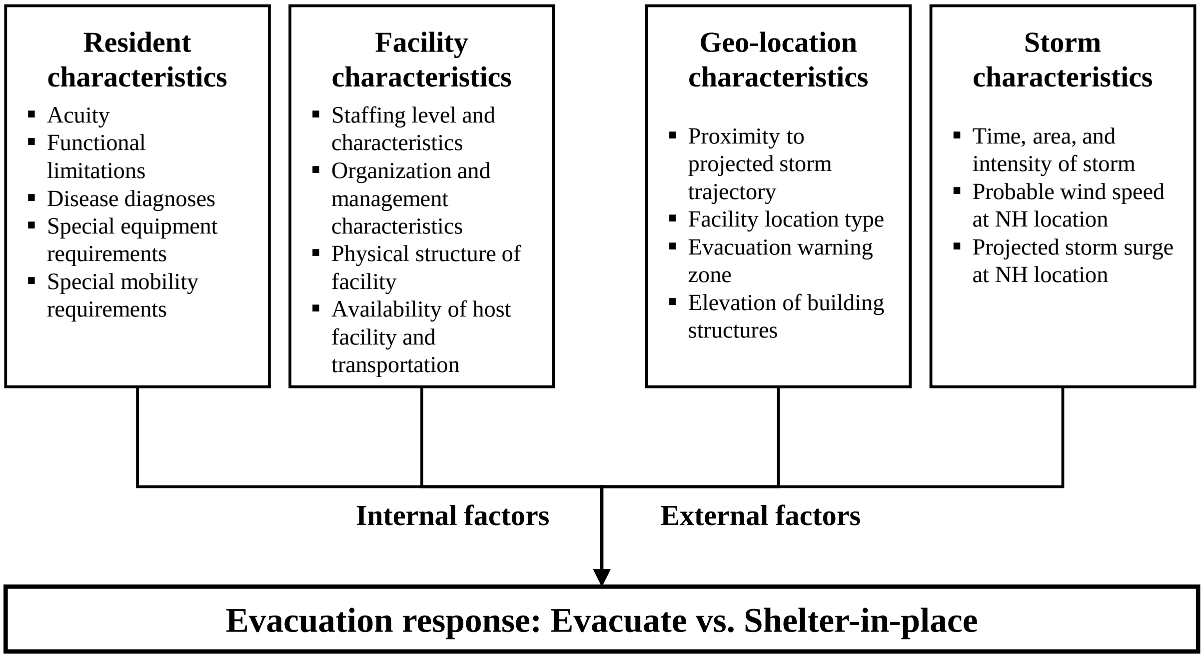



Figure 4.28: A conceptual model for evacuation decision-making criteria for NHs during hurricanes

### 4.2.1 Extracting Environmental Characteristics

Due to the devastating threats and negative consequences of hurricanes on vulnerable NH residents, such as drowning, shutdown of life-sustaining devices, and shortage of medical supplies, many NHs have to evacuate and move their frail residents away from hazard regions to safer places. The severity of actual damage on NHs at different storm affected regions may vary significantly. Thus, the anticipated geographical conditions in the neighboring area of each NH before storm's arrival tend to become important external factors (beyond the facility's control) that may influence NH administrators' evacuation decisions. In this section, we extracted a series of environmental features to represent various environmental characteristics (e.g., storm characteristics, geographic characteristics of each NH) that may affect NH evacuation decision. They will be incorporated as predictors while developing different predictive models. Before extracting environmental features, we first obtained the actual evacuation responses (i.e., evacuation or shelter-in-place) of

all NHs in the state of Florida during Hurricane Irma from the Florida Agency for Health Care Administration (AHCA) (FL AHCA, 2021). AHCA is the statutory organization responsible for health and policy planning in Florida. The agency also reports emergency response information of long-term care providers during extreme event scenarios, such as hurricanes. Figure 4.29 visualizes geolocation of all NHs in operation in the state of Florida and individual evacuation status during hurricane Irma. Such response data will further be utilized as labeled outputs in Sections 4.2.3 and 4.2.4 for predictive models training and prediction performance evaluation. Furthermore, we also extracted the geolocation of each NH to facilitate the calculation of NH-specific environmental features. The latitudes and longitudes of each NH location were extracted using ArcGIS Online World Geocoding Service (ESRI, 2021).

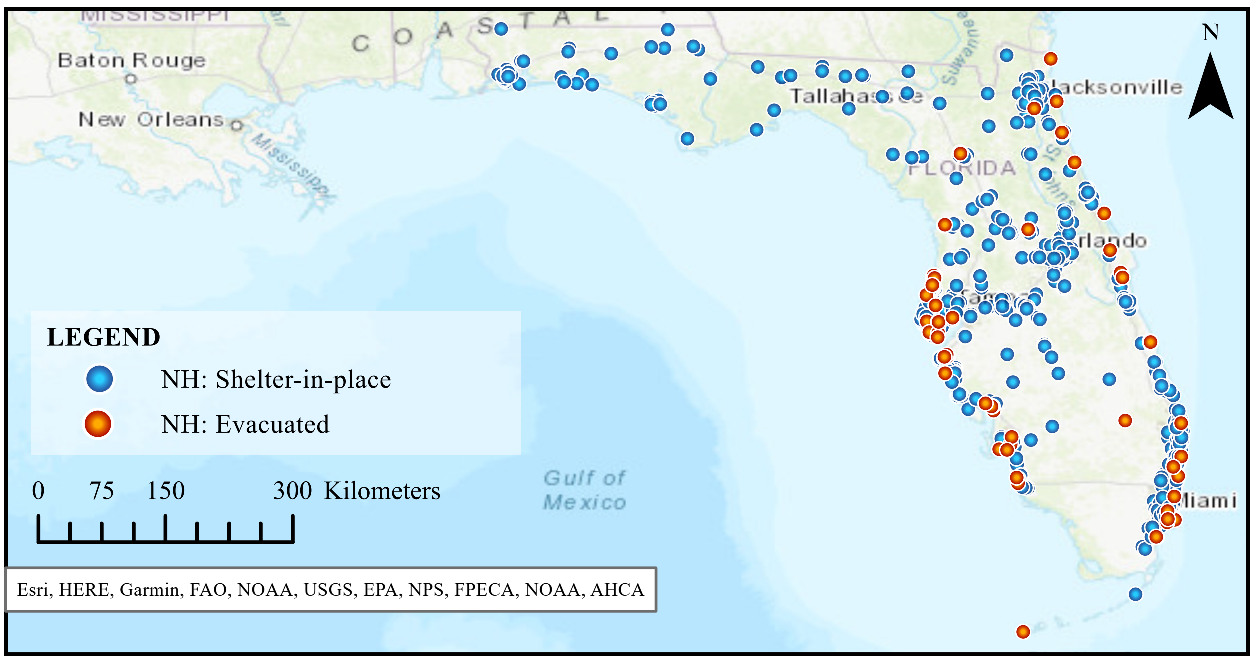



Figure 4.29: Extracted geolocations and evacuation status of NHs in FL during hurricane Irma

We began with extracting the storm characteristics affecting NH evacuation decision. The storm GIS data was extracted from the National Hurricane Center (NHC) of the National Oceanic and Atmospheric Administration (NOAA) (NOAA, 2021). During a

hurricane event, NHC monitors and records rich spatial-temporal information of a storm every 3-6 hours, including current storm location, projected trajectory, spatial probability distribution of wind speeds, and potential storm surge areas. It allows us to extract and calculate different NH-specific storm features and investigate their impacts on NH evacuation. As a storm approaches, the closeness between the projected storm path and the location of a NH may reflect the level of storm threat and can be potentially relevant to evacuation decision of a NH. We extracted the projected storm path, which represented the forecast trajectory of the center locations of a storm. To quantify proximity of a NH to the projected storm path, we calculated the shortest Euclidean distance from each NH geolocation point to the projected storm path, as shown in Figure 4.30.

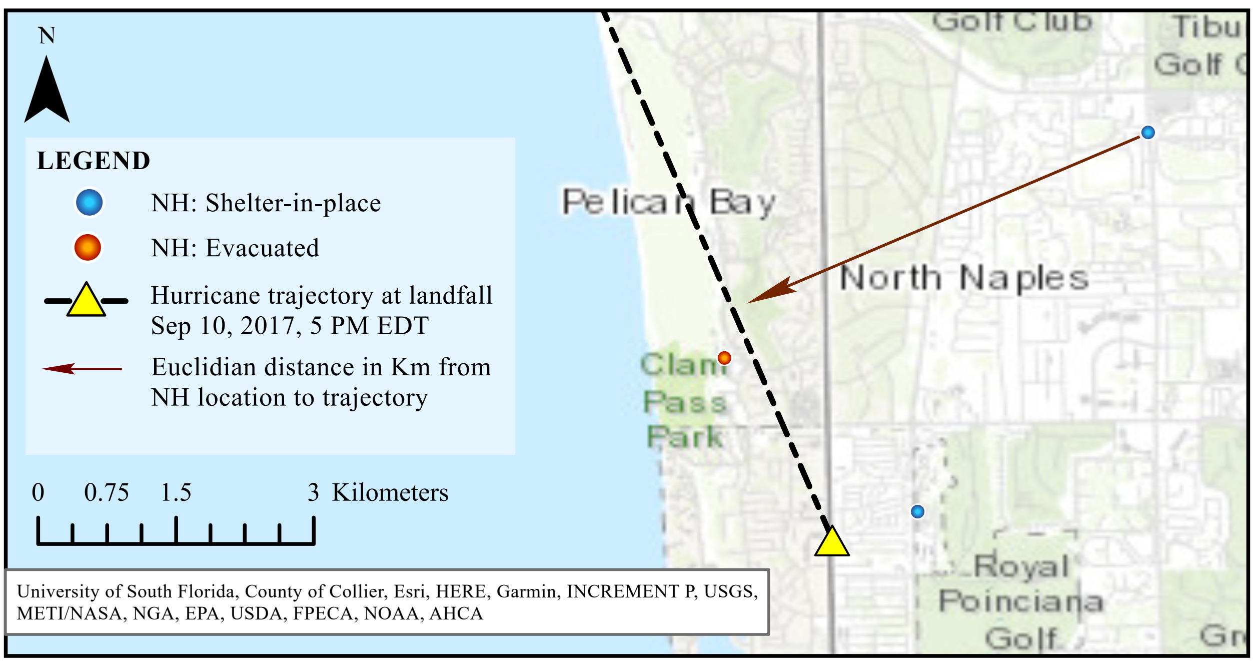



Figure 4.30: NH-specific proximity distance to the projected storm path

The building damage and power outage resulting from high and sustained wind speeds may greatly affect the NH administrators' decisions in evaluating the facility or not. To investigate the impact of the projected wind speed on NH evacuation decision, we extracted the spatial probability distribution map of wind speeds over a regularly spaced (5 km) grid of points, as shown in Figure 4.31. The projected wind speed probability

at a specific 5km-by-5km grid area represented the cumulative probability of sustained (1-minute) surface (10-meter altitude) wind speeds equal to or exceeding 50-knot (i.e., 57.5 mph) within a 120-hour time period. According to the Beaufort Wind Scale (Barua, 2005), 50-knot winds are classified officially as storm-force winds which can cause significant structural damage. As shown in Figure 4.31, NHs located closer to the projected storm path tend to have a higher probability of experiencing higher winds, and vice versa.

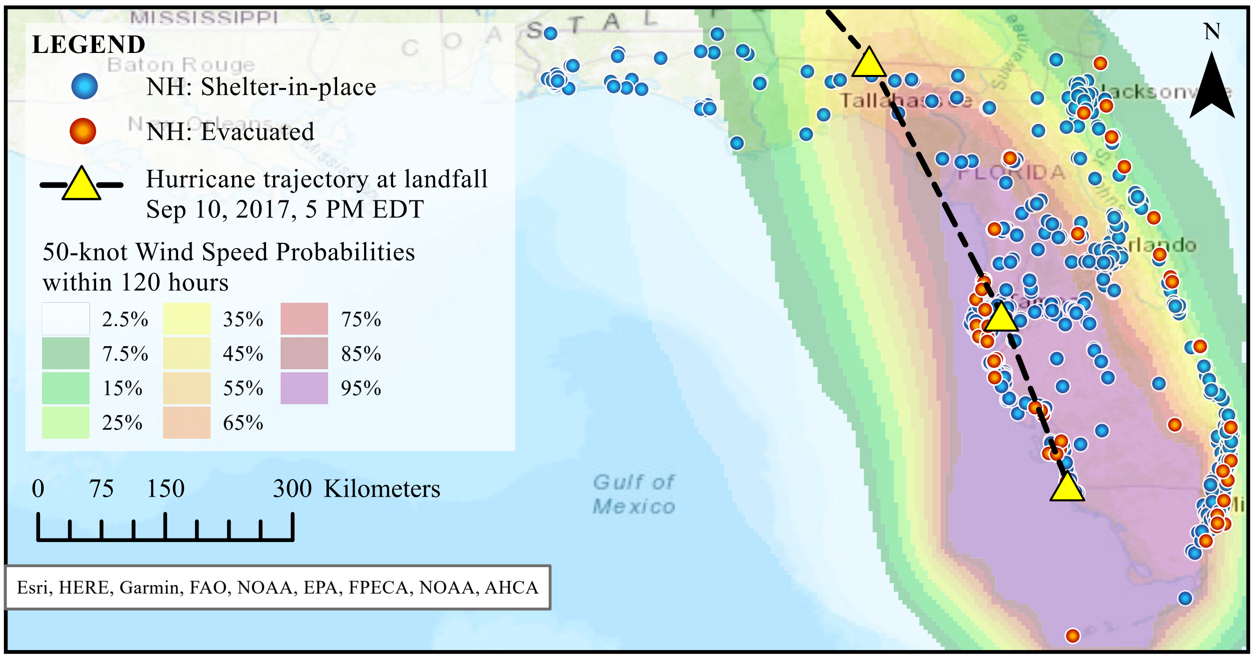



Figure 4.31: 50-knots cumulative wind speed probability map

As the adage "hide from the wind, run from the water" suggests, another important aspect which may considerably affect the vulnerability and safety of a NH location during hurricane is the projected flood risk. We considered the potential storm surge and elevation at each NH location as external and inherent features respectively to characterize the potential flood risk. Figure 4.32 shows the spatial map of potential storm surge associated with the storm, which describes the risk of potential coastal flooding due to a storm. Both the predicted areas (where inundation from storm surge could occur) and the predicted heights (that water could rich in those areas) were numerically determined by the Sea, Lake, and Overland Surges from Hurricanes (SLOSH) model developed by National Weather

Service (Jelesnianski *et al.*, 1992). The tidal mask region refers to the area usually submerged during daily or seasonal high tides. As shown in Figure 4.32, several NHs located in coastal regions at high potential storm surge evacuated before hurricane landfall. Apart from examining the food risk resulting from potential storm surge, we further considered the inherent geographic characteristics of each NH facility, namely, the elevation. Inland NHs in low-lying regions may also potentially experience flooding due to the rain-water deposition and/or the rise of water-level in nearby pond, lake, river, or other water reservoirs. To extract elevation of each NH, we considered the Florida Digital Elevation Model (DEM) (University of Florida, 2012) developed by the University of Florida GeoPlan Center. Figure 4.33 show the spatial map of elevation values recorded on a 5m-by-5m statewide grid. The elevation value from a grid area which is the closest to a NH's geolocation point has been selected as the approximate elevation value for that NH.

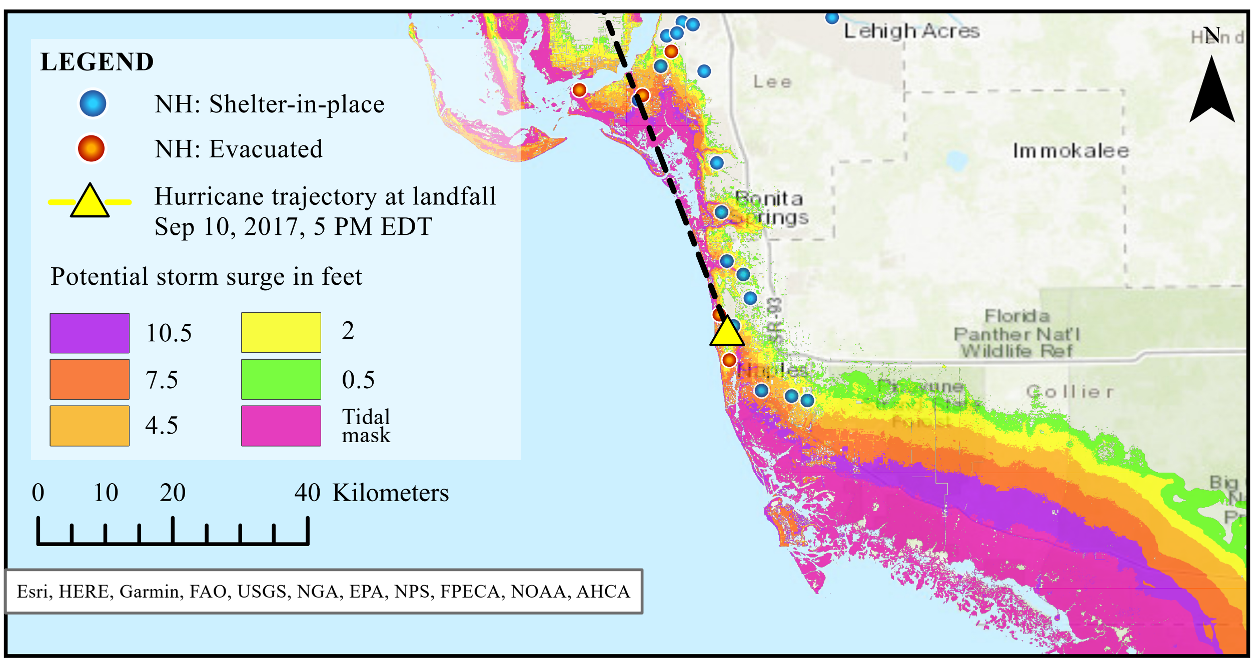



Figure 4.32: Potential storm surge heights near coastal region

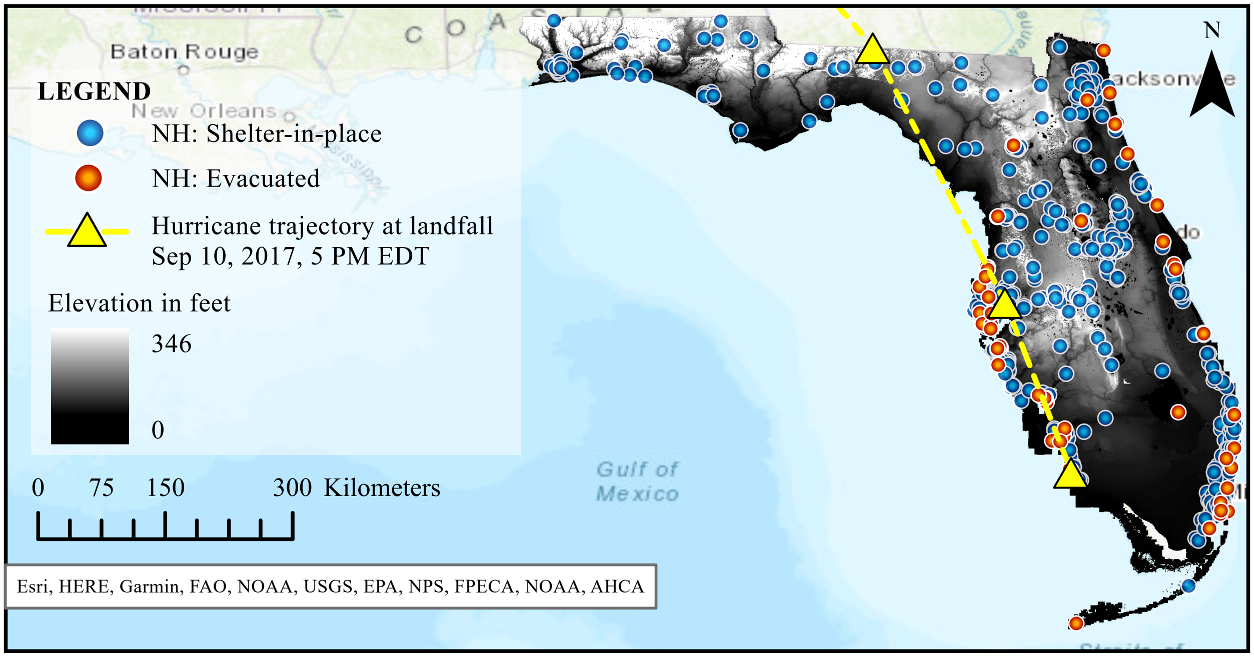



Figure 4.33: Elevation of NH location as mapped by FL DEM

### 4.2.2 Extracting NH Characteristics

The NH administrator's decision of evacuating or sheltering-in-place may not only be affected by external factors of environmental characteristics as described above, it may also be affected by various internal factors related each NH facility. In this section, we will further investigate various internal factors' influence by comprehensively extracting various aspects of NH characteristics, such as organizational characteristics, staffing characteristics and resident characteristics of each NH. To comprehensively evaluate the NH characteristics from different aspects, we consider the most updated Certification and Survey Provider Enhanced Reports (CASPER) data of each NH closest to the storm season. CASPER data, originally known as Online Survey Certification And Reporting (OSCAR) data, is the annual regulatory inspection data collected by state survey agencies and maintained by the Centers for Medicare and Medicaid Services (CMS) (CMS, 2017a). It contains rich facility level NH data related to the overall organization, such as size and ownership, as well as the aggregate characteristics of caregivers and residents within each NH. Based on the domain knowledge of expertise as well as national guidelines for NH evacuation

(FLHCA, 2008) we extract the NH characteristics based on the CAPSER data from three aspects, namely, (i) organizational properties, (ii) aggregated staffing characteristics, and (iii) aggregated resident characteristics, which will be elaborated with details as follows.

From the organizational perspective, existing studies indicate that the structural characteristics of an organization, such as ownership type, may have major implications in the extent of challenges the NH administrators would face in making decisions during the storm. For instance, government-owned facilities may have greater access to financial and/or transportation resources/support from government agencies than for-profit facilities, which would lower for-profit facilities' logistics capabilities and increase financial concerns in initializing evacuation (Brown *et al.*, 2007; Dosa *et al.*, 2007). Further, NHs of different sizes may have different likelihoods of exhausting their own organizational resources and may decide to evacuate due to their self-insufficiency. For a NH within a larger NH chain, it may be easier to identify and prepare the hosting facility within the chain to receive the evacuees, making the evacuation more convenient. To comprehensively quantify various organizational characteristics of each NH, we extract and calculate various organization level features based on CASPER data of each NH, such as the type of ownership, the overall size and the average occupancy rate. To quantify detailed organizational structure, we also introduce the binary indicators, "Any special care unit", to indicate whether the facility contains a special care unit (e.g., special units for caring residents with Alzheimer's Disease and Related Diseases) and various binary indicators under "Medical team structure", to indicate whether the medical team contains senior leadership and advanced medical personnel in the facility. The set of organizational features extracted are summarized in Table 4.11.

NH staffing also plays an important role in disaster preparedness and response (e.g., evacuating or sheltering-in-place) against hurricane from the following two aspects. First, adequate staffing is required in order to ensure the care continuity and the success of disaster preparedness to weather out the storm or evacuate safely. Second, NH caregivers,

such as nurses and aides, must be trained or have the right skills mix to tackle unique challenges during extreme hazard scenarios, such as hurricanes (Dosa *et al.*, 2007). If the facility is sheltering-in-place, adequate staff is crucial in avoiding increased morbidity and mortality of residents, as the staff would provide formal care, emotional support to residents, and also complete preparatory tasks such as strengthening building structures and storing supplies. For evacuation, the staff needs to coordinate transfer efforts, carry residents onto vehicles and transport them, and help them relocate into new hosting facility. Many NHs may face the challenges of staffing shortage and caregiver absenteeism during hurricane since many staff members may evaluate by themselves or have concerns for their own family members. The NH may have less self-sufficiency accordingly to shelter-in-place successfully with adequate staffing. To comprehensively investigate the influence of staffing characteristics affecting evacuation response, staffing levels of 3 different types of direct caregivers, such as Registered nurses, Licensed practical nurses, Certified nursing assistants, and 6 types of non-direct caregivers, such as Administrative Nurse, Occupational therapy services, Physical therapy services, Activities staff, Social services, and Housekeeping staff, are extracted and calculated based on CASPER data. Hours per resident per day (HPRD) (Konetzka *et al.*, 2008) is considered as the aggregate measure to characterize the staffing level of each type of caregiver. Specifically, for NH $i$, the HPRD of the $k$-th type of caregiver can be calculated as

$$HPRD_{ik} = \frac{Total\ FTE\ of\ caregiver\ type\ k * 70\ hours\ bi-weekly}{14\ days * Number\ of\ residents\ in\ NH\ i} \tag{4.24}$$

For each type of caregiver, part-time and temporary employees are converted into full-time equivalent (FTE) to facilitate the calculation of total FTE. For physical therapists, both the therapists and therapist assistants are taken into account. The extracted staffing characteristics features of each NH are summarized in Table 4.11.

In addition to the organizational and staffing characteristics, characteristics of the vulnerable residents in a NH are also important aspects that NH administrators need to

take into account and thus may potentially influence the NH evacuation decision. Many NH residents are non-ambulatory and bed-ridden, and evacuating them safely is more challenging since it requires more efforts from nursing staff as well as special transportation means, such as wheelchair conversion vans. For those residents with morbid obesity, specific equipment, such as lift and transfer equipment, need to be prepared. Many of NH residents may have complex medical conditions, such as renal and respiratory diseases, and they may either need special care, such as dialysis, or be highly oxygen dependent. If the NH is sheltering-in-place, power outages and inadequate medical supplies (due to road disruption) may be devastating to their residents. Existing studies also show that for those residents with mental conditions, such as, dementia or anxiety, NH evacuation and relocation availability may have detrimental effects on their health outcomes and induce post-traumatic stress (Dosa *et al.*, 2008). To comprehensively investigate the influence of resident characteristics from different aspects, several aggregate features in a NH facility, such as the percentage of residents having aforementioned conditions, the percentage of residents receiving different types of medications (e.g., antipsychotics, antianxiety, antidepressants, sedative/hypnotics), the percentage of residents who require physical restraints and the percentage of residents covered under Medicare, Medicaid, or paying by themselves, are extracted and calculated from CASPER data. To further quantify the highly varied health conditions and acuities of NH residents, a composite index feature called "Acuindex" is employed to summarize the overall resident acuity in the facility (Cowles, 2002; Hyer *et al.*, 2011). Acuindex is a numeric measure calculated by first combining the percentage of residents requiring nursing staff assistance with different activities of daily living, such as eating, toileting, bed transferring, and the percentage of residents requiring special treatments, such as respiratory treatment, suctioning, intravenous therapy, tube feeding, etc., and then further dividing by the total number of residents in the facility. A higher Acuindex value indicates that facility has a frailer population of residents with

more extensive care needs and vice versa. The extracted resident characteristics related features are summarized in Table 4.11.

### 4.2.3 Classification Models

After extracting various features that may potentially affect the NH evacuation decision (as described above), in this section, we will develop data-driven predictive models to predict the binary decision of "evacuating" or "sheltering-in-place" by comprehensively investigating different classification algorithms. For NH facility $i$, the binary response variable $y_i$ is labelled as "1" if the facility has evacuated, and "0" if the facility has sheltered-in-place. Different features that represent different aspects of the $i$-th NH, such as environmental characteristics, NH facility characteristics and NH dwelling-residents' characteristics, will serve as input variables $\boldsymbol{x}_i$ for the developed predictive models. Since the sheltered-in-place NHs account for about 90% of the total number of NHs in the dataset, there is considerable classification imbalance issue, which will significantly affect the modeling accuracy (Chawla *et al.*, 2004). To address such class imbalance issue, over-sampling technique is performed for the minority class to create an equally balanced dataset for model development (Chawla *et al.*, 2002). With the balanced dataset, it is further divided into training data set and test data sets for model training and model assessment, respectively.

We first investigate the linear classification model of logistic regression (LR) and its regularized variants. These models aim to find the optimal model parameters $\boldsymbol{\theta}_{LR}$ by minimizing the loss function

$$l(\boldsymbol{\theta}_{LR}) = \sum_{i=1}^{n}\{-y_i\boldsymbol{\theta}_{LR}^T\boldsymbol{x}_i + \log[1+\exp\boldsymbol{\theta}_{LR}^T\boldsymbol{x}_i]\} + \lambda_1\|\boldsymbol{\theta}_{LR}\|_2^2 + \lambda_2\|\boldsymbol{\theta}_{LR}\|_1 \tag{4.25}$$

where the first term is the negative log-likelihood function of LR, and the last two terms contain L2-norm and L1-norm penalties, respectively. The former terms represent the

goodness-of-fit and a smaller negative log-likelihood function value indicates a better goodness-of-fit. The latter terms control the complexity by shrinking the irrelevant model parameters towards zero (in L2-norm) and exactly equal to zero (in L1-norm), respectively. When both $\lambda_1 = 0$ and $\lambda_2 = 0$, the model is conventional LR; when $\lambda_1 > 0$ and $\lambda_2 = 0$, it becomes the LR model with Ridge penalty, and when $\lambda_1 = 0$ and $\lambda_2 > 0$, it becomes the LR model with LASSO penalty. Regularization-based LR models are considered to address the potential overfitting issues of conventional LR for prediction performance improvement. The tuning parameter $\lambda$ in both ridge and LASSO penalties are determined based on the cross-validation (CV).

After investigating the linear classification model, we further investigate different tree-based nonlinear classification modeling approaches, namely classification and regression tree (CART) and tree-based assembling methods. For CART, the impurity measure of GINI index is considered for tree plotting and a tree model will stop growing once all its leaf nodes only contain a single class of ether "evacuation" or "sheltering-in-place". To mitigate the overfitting issue of CART, pruning is further considered based on cross-validation to merge some of the branches to form a smaller tree. Due to the high variance of CART, we further considered the tree-based ensemble methods, namely, random forest and gradient-boosted tree, to further strengthen the prediction accuracy. The former ensemble learning methods generate a large number of deep trees in the parallel structure while the latter generates a large number of simple trees in sequential manner. The tuning parameters, such as size of trees in random forests or the depth of tree in gradient-boosted tree, are also determined based on the cross-validation.

Although tree-based classification methods give nonlinear decision boundaries, they are established based on the rectangular-shaped partitions of the feature space. We further investigate other nonlinear classification models, such as memory-based method of nearest-neighbor classification, optimization-based method of support vector machines (SVM) and network-based method of artificial neural network (ANN), which constructs nonlinear

decision boundaries based on varies assumptions and criteria. K-Nearest Neighbor (KNN) directly performs the prediction based on the major vote of its K-neighbor data points, i.e., $\hat{y} = \text{argmax}_{k\in\{0,1\}} \frac{1}{K} \sum_{N_x} \mathrm{I}(\mathrm{y_i} = \mathrm{k})$, where $N_x$ is an index set of K-nearest neighboring observations for input variables $\boldsymbol{x}$ and $\mathrm{I}(\bullet)$ is an indicator function. SVM and ANN are more computationally demanding nonlinear classification methods which either formulates the classification problem as an optimization model, or capture the nonlinear mapping among inputs and outputs with a multi-layer network structure, respectively. The tuning parameters and settings of each method, such as choice of $K$ in KNN, kernel type and cost settings in SVM, and number of neurons in layer in ANN, are also determined based on cross-validation (Hastie *et al.*, 2009; Bishop, 2006).

### 4.2.4 Performance Evaluation

During the model development stage, 10-fold cross-validation (CV) is employed to: (i) obtain the expected estimate of the test accuracy, (ii) tune model parameter values such that overfitting can be avoided. CV is achieved by partitioning the training dataset into several approximately equal-sized folds and building a model on the dataset by progressively holding one fold out for validation. CV-accuracy is the calculated as the average accuracy obtained over all the validation set predictions, i.e., $Acc_{CV} = \frac{1}{10*n_m} \sum_{m=1}^{10} \sum_{i=1}^{n_m}$ $\mathrm{I}(\mathrm{y_{i,m}} = \hat{\mathrm{y}}_{\mathrm{i,m}})$, where $y_{i,m}$ $\hat{y}_{i,m}$ are observed and predicted values in the $m$-th validation set with sample size $n_m$ , respectively, and 10 is the number of folds in the dataset. Limiting exposure to the full training dataset allows selection of model parameters that do not perfectly fit the training data but are generalized adequately resulting in optimal performance over unseen test data, hence reducing overfitting.

Following model development and tuning, each optimized model is utilized to generate prediction on previously unseen test dataset and prediction accuracy is evaluated. Several metrics are employed to assess model effectiveness from different aspects, namely, test accuracy, test sensitivity, test specificity, and test balanced accuracy. Test accuracy is simply

the proportion of correctly predicted class labels against the total number of observations in the test data, i.e., $Acc_{test} = \frac{1}{n_t} \sum_{j=1}^{n_t} \mathrm{I}(\mathrm{y_j} = \mathrm{\hat{y}_j})$, where $y_j$ $\hat{y}_j$ are observed and predicted values, respectively, and $n_t$ is the total number of test observations. Test accuracy is an overall metric and may not fully explain individual class-specific prediction performance. Test sensitivity is used to measure prediction performance of the model for the minority class (evacuated) as a proportion of the number of correctly predicted evacuated NHs against the total number of observed evacuated NHs, i.e., $Acc_{test} = \frac{1}{n_{evac}} \sum_{j=1}^{n_{evac}} \mathrm{I}(\mathrm{y_{j,evac}} = \mathrm{\hat{y}_{j,evac}})$, where $y_{j,evac}$ and $\hat{y}_{j,evac}$ are observed and predicted values for evacuation, respectively, and $n_{evac}$ is the total number of evacuated NHs in test dataset. Similarly, test specificity measures prediction performance of the model for the majority class (shelter-in-place) as a proportion of the number of correctly predicted shelter-in-place NHs against the total number of observed shelter-in-place NHs, i.e., $Acc_{spec} = \frac{1}{n_{shelter}} \sum_{j=1}^{n_{shelter}} \mathrm{I}(\mathrm{y_{j,shelter}} = \mathrm{\hat{y}_{j,shelter}})$, where $y_{j,shelter}$ and $\hat{y}_{j,shelter}$ are observed and predicted values for shelter-in-place, respectively, and $n_{shelter}$ is the total number of shelter-in-place NHs in test dataset. Assessing sensitivity and specificity of the predictions produced by each model enhances the ability to compare models in terms of their flexibility in detecting rare classes and differentiating data from the majority class, and also allows determining the effect of oversampling in performance improvement. Since test accuracy may indicate greater model performance even if the model predicts all majority class in the test data correctly and misses all minority class, an improved measure is desirable. Test balanced accuracy is the average of test sensitivity and test specificity, i.e., $Acc_bal = \frac{Acc_{sens}+Acc_{spec}}{2}$, and provides a better representation of the prediction performance (Hastie *et al.*, 2009).

### 4.3 Real Case Study

#### 4.3.1 Data Description and Preprocessing

Hurricane Irma was one of the major hurricanes in history over the open Atlantic Ocean. The storm made initial landfall in Florida near Cudjoe Key as a Category 4 (130 mph)

hurricane on September 10, 2017 9 AM and afterwards made final landfall at Marco Island as Category 3 (115 mph) on September 10, 2017 3:35 PM, moving up the state dissipating over the next day. An estimated 6.5 million people were ordered to evacuate causing scarcity of supplies and fuel, and heavy traffic along evacuation routes. One direct and 33 indirect deaths were reported in South Florida. The storm caused significant destruction by uprooting trees, damaging building roofs and structures, excessive inland flooding and coastal surge, and heavy rainfall. More than 75% clients in the state lost power for almost a week, and half of all crops in Miami-Dade county was ruined (NWS, 2017). The estimated cost of damages in flood loss to homes in all storm-affected state was between 25-38 billion USD (LaVito, 2017), with state property damages costing hundreds of millions USD in different counties. As with any extreme event, long-term care service residents and staff were at greater risk. 684 NHs were in operation during the hurricane, among which a total of 85 facilities decided to evacuate pre- or post-landfall.

To maintain consistency with the scope of the study, several inclusion/exclusion criteria were applied to the list of operating NHs. NHs which evacuated after landfall were not considered since the decision was based on post-storm damages and after-effects, rather than evaluation of pre-storm anticipated risks. NHs which evacuated excessive time in advance (on or before 96 hours of landfall) were excluded since the storm was considerable distance away from FL and related environmental data was not yet available. NHs with facility characteristics data missing entirely or with data recorded on inappropriate survey dates (i.e., survey significantly predating the storm), categorized as hospital swing beds, and/or with incorrectly reported nursing staff levels (i.e. greater than 24, or 0 HPRD) were removed. The resulting dataset contained a total of 653 NHs, of which 59 (9.04%) evacuated and the rest 594 (90.96%) sheltered-in-place. The evacuation status was encoded as binary numbers, where 1 indicated evacuated and 0 shelter-in-place, such that it can be used as a numerical response during modeling. Table 4.11 presents descriptive statistics of the selected NHs with various characteristics stratified by evacuation status.

Following extraction of NH-specific environmental features as described in Section 4.2.1 and facility structural, staffing and resident characteristics data as described in Section 4.2.2, a joint dataset was created including all extracted features of each NH. To prepare for predictive modeling, the dataset is treated with several preprocessing measures. The data was randomly split into 80-20 train-test subsets. The training set is intended to be utilized in model estimation, while the test set remained as unseen data for later predictions. Based on the training set, redundant facility characteristics features were removed by evaluating correlation coefficients between all possible pairs of features and setting a cutoff of 0.6 and guided by domain knowledge. For instance, Social services HPRD feature was removed as it was highly correlated with Registered nurses HPRD. Existence of multicollinearity among undetected feature combinations was determined with calculation of Variance Inflation Factor (VIF) with cutoff set at 5. According to disaster management timelines, NHs need to take decision on whether to evacuate or shelter-in-place at least 24 to 36 hours prior storm occurrence to allow sufficient time for clearing the area or completing preparations, respectively (Natarajan, 2013; Chen *et al.*, 2006). The environmental features recorded between 24-hours and 36-hours prior decision were highly correlated and repetitive as the storm's projected trajectory changed little. Hence, the environmental features extracted from geographic observations were recorded 24-hours prior evacuation decision of each NH, assuming it is the last time NHs can make decision. The final set of features included 4 environmental features and 32 facility characteristics features in the joint dataset. Since NHs sheltering in place substantially outnumbered evacuated facilities, up-sampling was applied to training dataset to balance proportion of each class and ease estimation of predictive models.

### 4.3.2 Prediction Performance Comparison

To investigate how different environmental features impact predictive performance individually and altogether, 1 baseline and 5 different proposed modeling strategies

Table 4.11: Descriptive statistics summary of the integrated NH evacuation data

| Statistic/Feature | All Facilities* N = 653 | Evacuated Facilities N = 59 (9.04%) | Sheltered Facilities N = 594 (90.96%) |
|---|---|---|---|
| **Organizational structure (Y/N)** | | | |
| For profit facility | 72% | 53% | 74% |
| Not for-profit facility | 25% | 44% | 24% |
| Government facility | 2% | 3% | 2% |
| Chain-facility | 61% | 58% | 61% |
| Part of a CCRC | 9% | 20% | 8% |
| Size (# beds) | 123.3 (48.6) | 118.8 (48.47) | 123.74 (48.63) |
| Resident count | 107.47 (44.33) | 99.54 (44.53) | 108.26 (44.27) |
| Occupancy rate | 87.09% (11.36) | 83.96% (15.84) | 87.4% (10.79) |
| ADRD special care unit | 13% | 14% | 13% |
| Non-ADRD special care unit | 5% | 3% | 5% |
| Any special care unit | 17% | 17% | 17% |
| Has organized resident group | 98% | 97% | 98% |
| Has organized family members group | 42% | 37% | 42% |
| **Payer mix (% residents)** | | | |
| Medicare | 54.98% (22.2) | 48.71% (26.39) | 55.6% (21.66) |
| Medicaid | 20.28% (14.37) | 19.46% (13.83) | 20.36% (14.43) |
| Private pay and other | 24.75% (17.98) | 31.83% (24.55) | 24.04% (17.05) |
| **Environmental GIS features** | | | |
| Distance from projected trajectory 24 hours prior decision (Km) | 112.56 (78.04) | 66.11 (55.48) | 117.17 (78.48) |
| 50 knots wind speed probability 24hrs prior (%) | 61.14 (28.37) | 55.17 (15.37) | 61.74 (29.29) |
| Potential storm surge 24 hrs prior (feet) | 0.24 (1.21) | 0.3 (1.17) | 0.23 (1.21) |
| Elevation of facility (feet) | 45.1 (47.71) | 12.36 (13.05) | 48.35 (48.67) |
| **Staffing characteristics (HPRD)** | | | |
| Registered nurses | 0.49 (0.58) | 0.54 (0.84) | 0.49 (0.55) |
| Licensed practical nurses | 0.95 (0.37) | 1.02 (0.71) | 0.95 (0.32) |
| Certified nursing assistants | 2.82 (0.77) | 2.99 (1.19) | 2.81 (0.72) |
| Direct care nurse staffing** | 4.27 (1.37) | 4.56 (2.63) | 4.24 (1.17) |
| Administrative Nurse | 0.28 (0.22) | 0.31 (0.39) | 0.27 (0.19) |
| Occupational therapy services | 0.26 (0.16) | 0.24 (0.12) | 0.27 (0.16) |
| Physical therapy services | 0.31 (0.2) | 0.28 (0.19) | 0.31 (0.21) |
| Activities staff | 0.21 (0.16) | 0.26 (0.37) | 0.21 (0.13) |
| Social services | 0.11 (0.13) | 0.15 (0.36) | 0.11 (0.08) |
| Housekeeping staff | 0.58 (0.57) | 0.79 (1.64) | 0.56 (0.3) |

Note: * Mean (SD); ** Sum of RN, LPN, CNA.

Table 4.11: Continued

| Statistic/Feature | All Facilities* N = 653 | Evacuated Facilities N = 59 (9.04%) | Sheltered Facilities N = 594 (90.96%) |
|---|---|---|---|
| **Medical team structure (Y/N)** | | | |
| Medical director only | 18% | 19% | 18% |
| Physician extender only | 0.5% | 0% | 1% |
| Full medical team | 45% | 39% | 45% |
| No medical team | 4% | 5% | 4% |
| **Resident Characteristics (% residents)** | | | |
| Acuindex (patient acuity) | 10.94 (1.2) | 11.01 (1.34) | 10.93 (1.18) |
| Behavioral healthcare needs | 18.13% (17.31) | 18.69% (15.1) | 18.07% (17.53) |
| Dementia or Alzheimer's | 42.98% (17.26) | 43.91% (17.17) | 42.88% (17.28) |
| Depression | 33.73% (21.1) | 31.77% (21.31) | 33.93% (21.08) |
| Intellectual disability | 1.19% (3.6) | 1.03% (1.56) | 1.21% (3.74) |
| Physical restraint use | 0.63% (1.74) | 0.23% (0.62) | 0.67% (1.81) |
| Serious Mental Illness | 29.6% (17.46) | 27.96% (14.31) | 29.77% (17.75) |
| **Medication utilization (% residents)** | | | |
| Antipsychotics | 18.11% (11.31) | 16.93% (9.59) | 18.23% (11.46) |
| Antianxiety | 25.31% (10.71) | 24.64% (9.69) | 25.37% (10.81) |
| Antidepressants | 48.53% (13.05) | 47.48% (12.8) | 48.63% (13.08) |
| Sedative/hypnotics | 7.36% (5.96) | 7.05% (5.07) | 7.39% (6.04) |

Note: * Mean (SD)

were adopted as detailed in Table 4.12. As the conventional approaches consider only facility characteristics influencing evacuation decision in literature, D1-BASE strategy was set up for model estimation using only the facility characteristics dataset and no GIS features. Hence D1-BASE was the baseline for comparison against proposed strategies. The proposed D2-DIST, D3-WSP, D4-SURG, D5-ELEV strategies were set up where each dataset contained only one GIS-feature in addition to facility characteristics – distance between facility and storm trajectory, 50-knot wind speed probabilities at facility location, potential storm surge at facility location, and elevation of facility location, respectively. This allowed evaluation of marginal change in prediction performance of the models compared to baseline and determination of the best individual GIS-feature. D6-FULL

contained all 4 GIS-features together in addition to NH facility characteristics to include all available information in modeling.

Table 4.12: Different predictive modeling strategies

| *Short name* | *Modeling strategy description* | *Strategy type* |
|---|---|---|
| D1-BASE | NH facility characteristics without any GIS features | Conventional |
| D2-DIST | NH facility characteristics with 1 GIS feature only: distance between facility and storm trajectory | Proposed |
| D3-WSP | NH facility characteristics with 1 GIS feature only: probable wind speed at facility location | Proposed |
| D4-SURG | NH facility characteristics with 1 GIS feature only: potential storm surge at facility location | Proposed |
| D5-ELEV | NH facility characteristics with 1 GIS feature only: elevation at facility location | Proposed |
| D6-FULL | NH facility characteristics with all 4 GIS features | Proposed |

For each of the modeling strategies, 9 different linear and non-linear predictive classification models were employed, as listed in Table 4.13, to establish functional relationship between evacuation response and heterogenous facility characteristics and GIS-features. Each predictive model differs in mathematical formulation and estimation process. Some involve pre-setting hyperparameters of the model configuration to maximize prediction accuracy suitable for respective data. Unknown values of hyperparameters which maximize prediction accuracy are found by searching through the numerical space by trial-and-error. In this case, optimal hyperparameters of each model were determined with 10-fold cross-validation (CV), as described in Section 4.2.4, on training dataset to maximize prediction performance, i.e., CV-accuracy. For instance, regularization parameter was optimized for LASSO and Ridge logistic regression, number of trees and number of features randomly chosen at each split was tuned for Random Forest, number of trees for Gradient Boosted Trees, cost and kernel parameters for Support Vector Machines, number of nearest neighbors in K-nearest Neighbor, number of units in hidden layer for Artificial Neural Network, etc. Features of test dataset was fed to each tuned model to predict evacuation

response and compared with observed responses to evaluate test classification accuracy. Furthermore, test sensitivity, test specificity and test balanced accuracy are evaluated to assess prediction performance of each category of evacuation response. Comparison of prediction performance obtained over each modeling strategy and predictive model are visualized in Figures 4.34, 4.35, and 4.36, and numerically reported in Table 4.14.

Table 4.13: Different machine learning methods considered

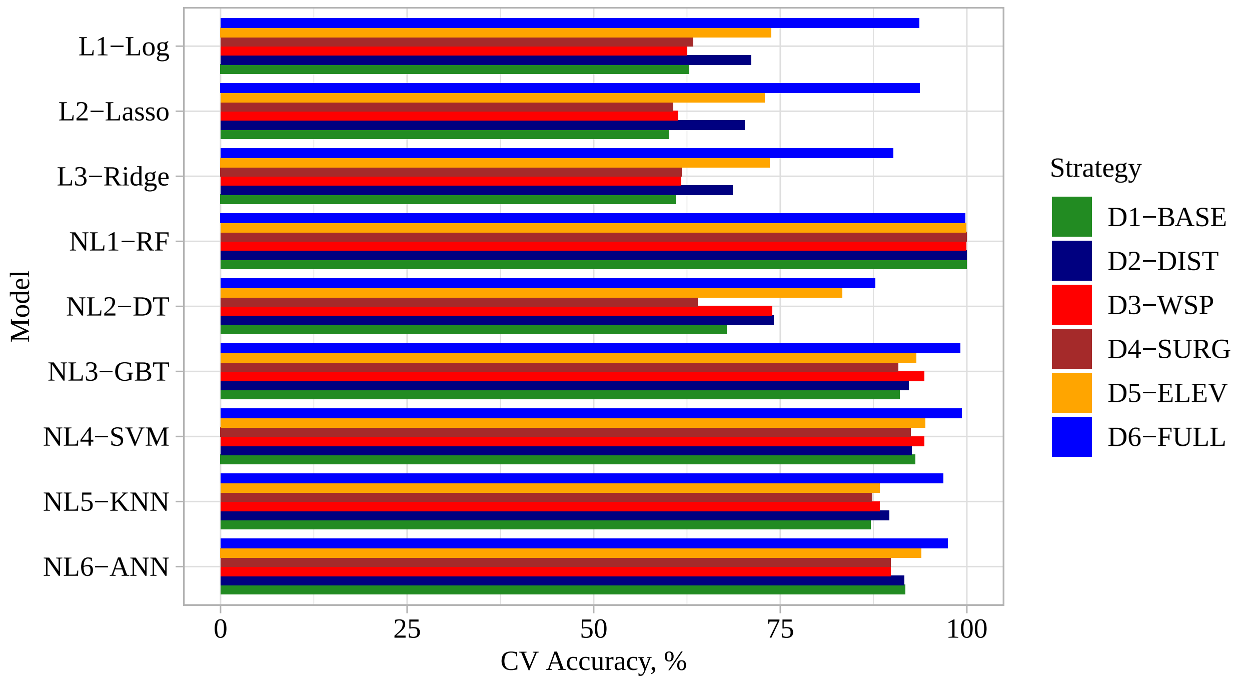
| *Model Name* | *Model Description* | *Model Type* |
|---|---|---|
| L1-Log | Logistic Regression (LR) | Linear |
| L2-Lasso | LASSO LR | Linear |
| L3-Ridge | Ridge LR | Linear |
| NL1-RF | Random Forest | Non-linear |
| NL2-DT | Decision Tree | Non-linear |
| NL3-GBT | Gradient Boosted Trees | Non-linear |
| NL4-SVM | Support Vector Machines | Non-linear |
| NL5-KNN | K-Nearest Neighbor | Non-linear |
| NL6-ANN | Artificial Neural Network | Non-linear |

Several insights were obtained from the results through perspectives of each performance metrics. From CV-accuracy and test accuracy in Figure 4.34a and 4.34b, it is evident that incorporation of one or more GIS-features improved performance significantly for most models compared to baseline strategy. Incorporating all 4 GIS-features improved performance for all models the most. Among individual features, elevation of facility (D5-ELEV strategy) and distance of facility to storm trajectory (D2-DIST) interchangeably provided the strongest improvement in prediction accuracy. Non-linear models in general provided increased CV- and test-accuracies for all strategies since they are more capable of capturing non-linear relationship between the features and evacuation responses, and their greater model complexity allows optimal generalization over the data. In contrast, linear models were largely dependent on available information in the training dataset and higher margins of accuracy improvements were obtained with inclusion of GIS-features compared with non-linear models. Sensitivity and Specificity in Figure 4.35b and 4.35a

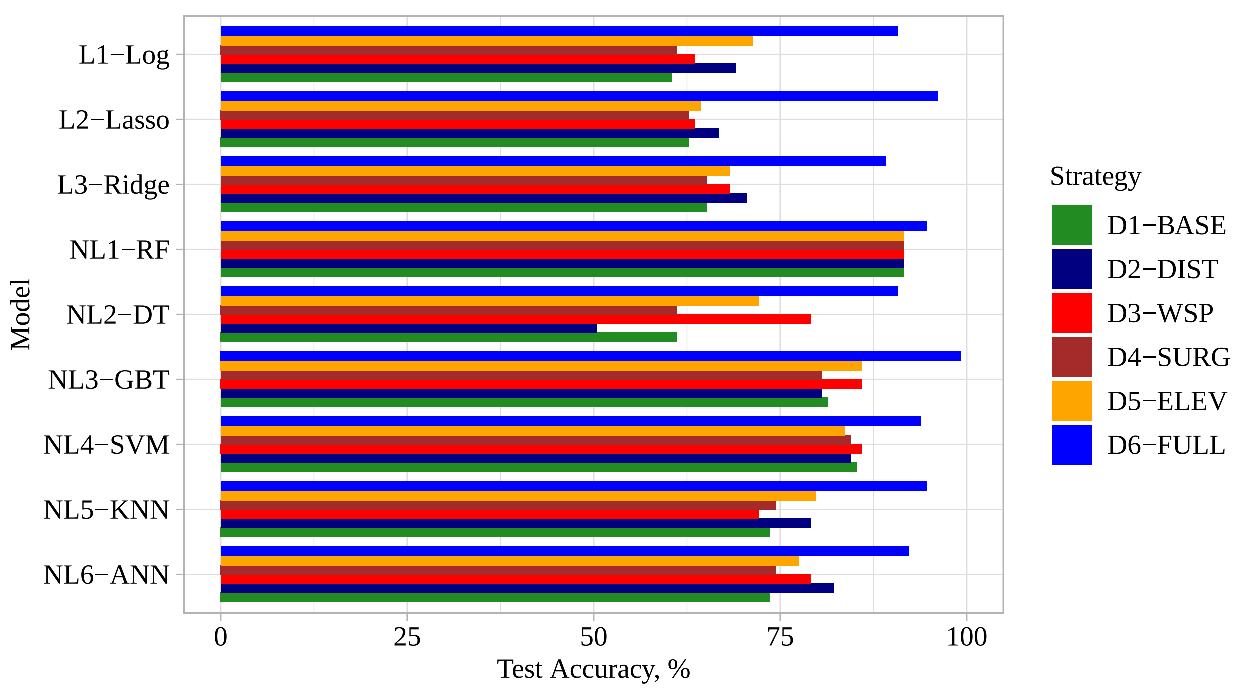
show prediction accuracies achieved for minority and majority classes respectively. Incorporation of GIS-features greatly improved predictive capacity of minority class, since more information was available in informing class separation. Balanced accuracy provided a better criterion in discerning model efficacy for individual class prediction. Especially for a few models, such as Random Forest and SVM, incorporating GIS-features was the only way to obtain any correct prediction for minority class.

Comparing all metrics across models and strategies, Gradient Boosted Tree (NL3-GBT) with incorporation of all 4 GIS-features (D6-FULL) was the best model for the dataset in predicting NH evacuation decision response. With high CV-accuracy (0.877), it gave the highest performance on unseen data with test accuracy of 0.992. It could detect both minor and major classes with high accuracy (test sensitivity of 1 and test specificity of 0.992), showing the best performance at test balanced accuracy of 0.966. A close contender was LASSO Logistic model (L2-Lasso), which was also the best among all linear models. Contrary to expectations from a complex model and despite achieving high CV- and test accuracies, Random Forest failed at predicting minority class and was overly biased towards majority class. Decision Tree (NL2-DT) performance was often unstable across different strategies since the modeling method of training single trees leads to high variance in predictions.

Up-sampling played an important role in improving prediction accuracy and it is illustrated by a further case study as follows. The previously best performing model NL3-GBT was applied to the dataset with all GIS-features (D6-FULL) before and after up-sampling. As observed in Figure 4.37 and Table 4.15, the unbalanced dataset results in poor accuracies and balancing the classes enhanced performance across all metrics. Particularly as seen in test sensitivity, up-sampling drastically improved the model's ability to detect minority class. The model became more nuanced towards class distinctions in the feature space resulting in higher prediction accuracy.

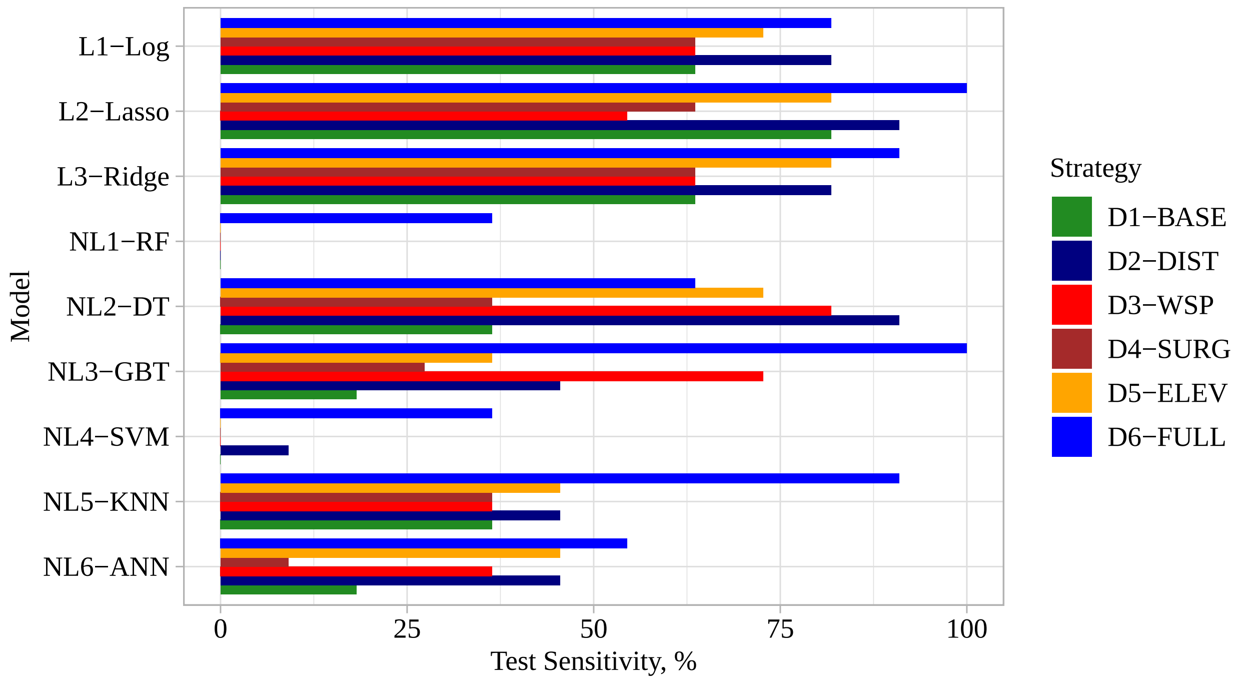



(a) CV accuracy

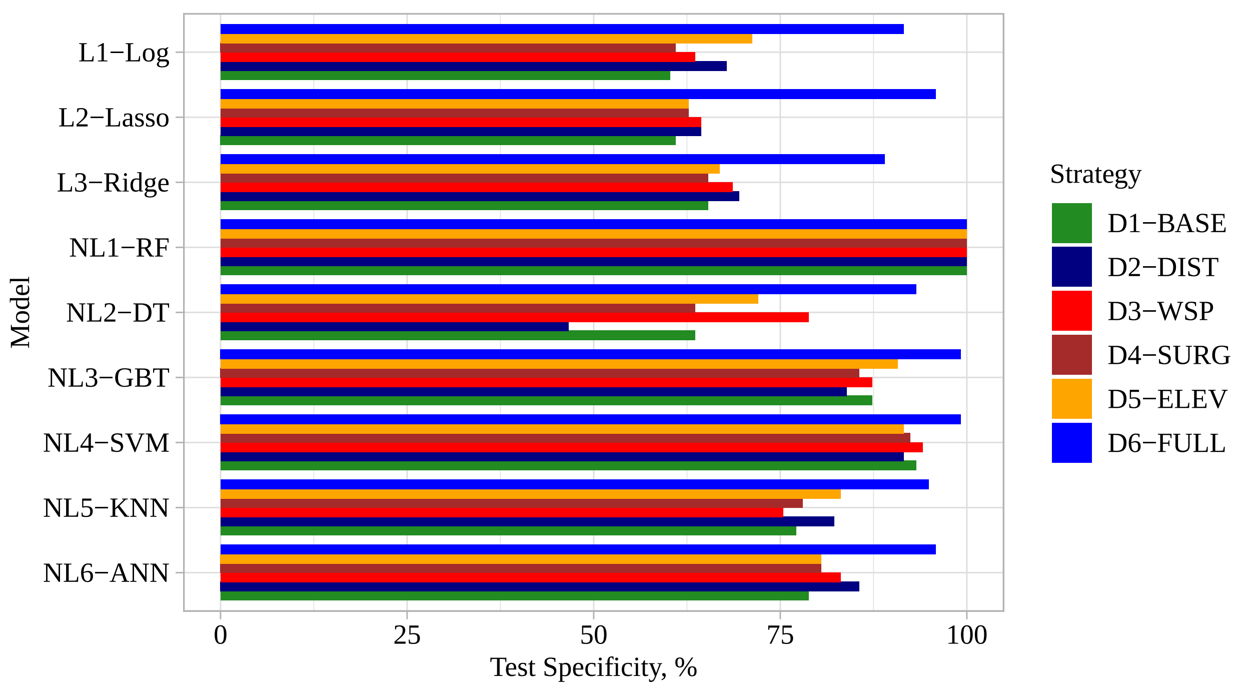



(b) Test accuracy

Figure 4.34: Prediction performance comparison among different models based on CV and Test accuracy

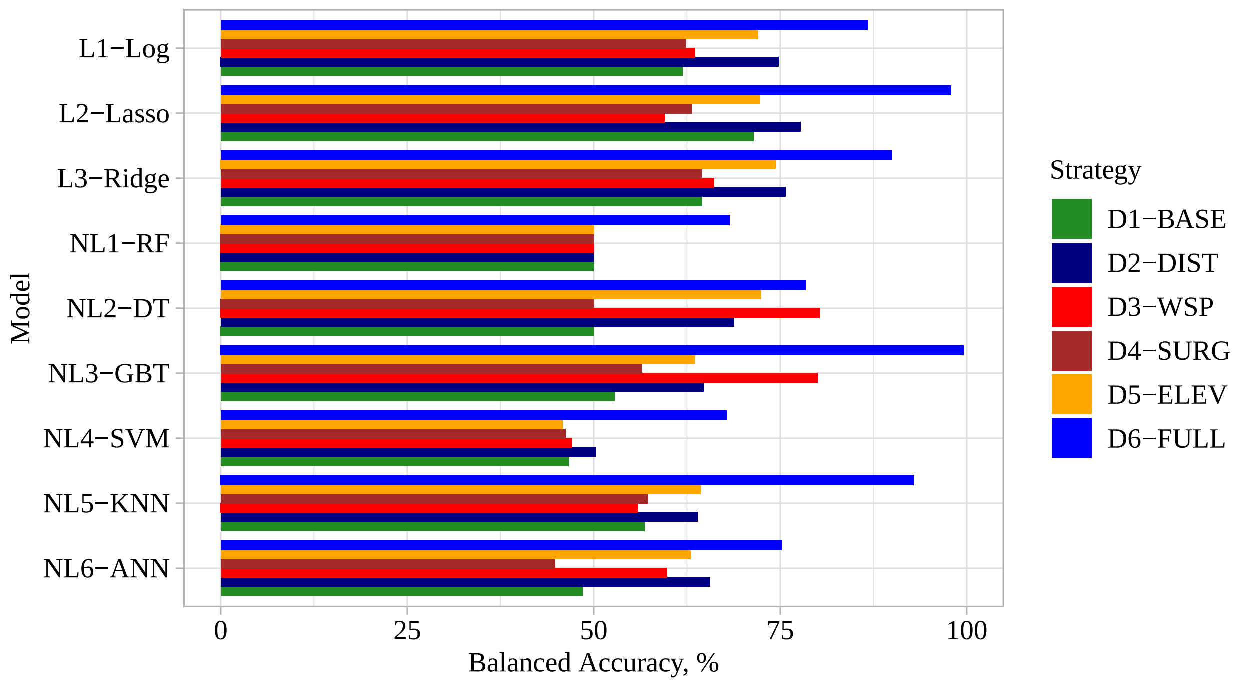



(a) Test sensitivity




(b) Test specificity

Figure 4.35: Prediction performance comparison among different models based on Test sensitivity and specificity




(a) Balanced accuracy

Figure 4.36: Prediction performance comparison among different models based on Test balanced accuracy

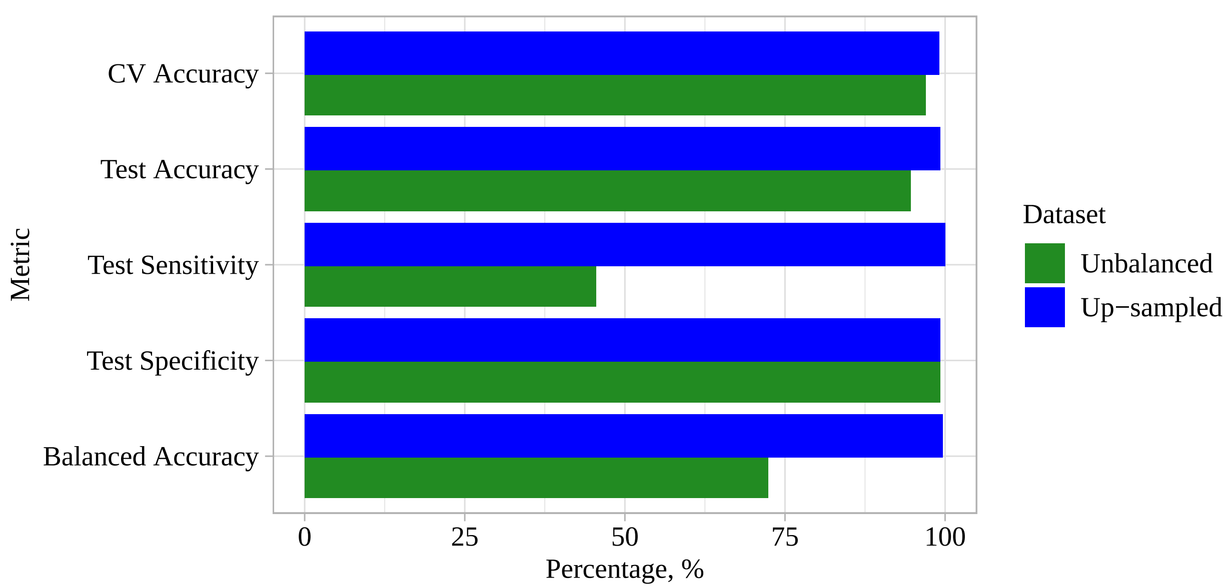



Figure 4.37: The influence of up-sampling on prediction

Table 4.14: Numerical summary of prediction performance comparison results

| *Metric\ Strategy* | *Model* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *L1-Log* | *L2-Lasso* | *L3-Ridge* | *NL1-RF* | *NL2-DT* | *NL3-GBT* | *NL4-SVM* | *NL5-KNN* | *NL6-ANN* |
| **CV Accuracy** | | | | | | | | | |
| D1-BASE | 0.628 | 0.601 | 0.61 | 1 | 0.678 | 0.91 | 0.931 | 0.871 | 0.917 |
| D2-DIST | 0.711 | 0.702 | 0.686 | 1 | 0.741 | 0.922 | 0.926 | 0.896 | 0.916 |
| D3-WSP | 0.625 | 0.613 | 0.617 | 0.999 | 0.739 | 0.943 | 0.943 | 0.883 | 0.898 |
| D4-SURG | 0.633 | 0.606 | 0.618 | 1 | 0.639 | 0.908 | 0.925 | 0.873 | 0.898 |
| D5-ELEV | 0.738 | 0.729 | 0.736 | 0.999 | 0.833 | 0.932 | 0.944 | 0.883 | 0.939 |
| D6-FULL | 0.936 | 0.937 | 0.901 | 0.998 | 0.877 | 0.991 | 0.993 | 0.968 | 0.974 |
| **Test Accuracy** | | | | | | | | | |
| D1-BASE | 0.605 | 0.628 | 0.651 | 0.915 | 0.612 | 0.814 | 0.853 | 0.736 | 0.736 |
| D2-DIST | 0.69 | 0.667 | 0.705 | 0.915 | 0.504 | 0.806 | 0.845 | 0.791 | 0.822 |
| D3-WSP | 0.636 | 0.636 | 0.682 | 0.915 | 0.791 | 0.86 | 0.86 | 0.721 | 0.791 |
| D4-SURG | 0.612 | 0.628 | 0.651 | 0.915 | 0.612 | 0.806 | 0.845 | 0.744 | 0.744 |
| D5-ELEV | 0.713 | 0.643 | 0.682 | 0.915 | 0.721 | 0.86 | 0.837 | 0.798 | 0.775 |
| D6-FULL | 0.907 | 0.961 | 0.891 | 0.946 | 0.907 | 0.992 | 0.938 | 0.946 | 0.922 |
| **Test Sensitivity** | | | | | | | | | |
| D1-BASE | 0.636 | 0.818 | 0.636 | 0 | 0.364 | 0.182 | 0 | 0.364 | 0.182 |
| D2-DIST | 0.818 | 0.909 | 0.818 | 0 | 0.909 | 0.455 | 0.091 | 0.455 | 0.455 |
| D3-WSP | 0.636 | 0.545 | 0.636 | 0 | 0.818 | 0.727 | 0 | 0.364 | 0.364 |
| D4-SURG | 0.636 | 0.636 | 0.636 | 0 | 0.364 | 0.273 | 0 | 0.364 | 0.091 |
| D5-ELEV | 0.727 | 0.818 | 0.818 | 0 | 0.727 | 0.364 | 0 | 0.455 | 0.455 |
| D6-FULL | 0.818 | 1 | 0.909 | 0.364 | 0.636 | 1 | 0.364 | 0.909 | 0.545 |
| **Test Specificity** | | | | | | | | | |
| D1-BASE | 0.602 | 0.61 | 0.653 | 1 | 0.636 | 0.873 | 0.932 | 0.771 | 0.788 |
| D2-DIST | 0.678 | 0.644 | 0.695 | 1 | 0.466 | 0.839 | 0.915 | 0.822 | 0.856 |
| D3-WSP | 0.636 | 0.644 | 0.686 | 1 | 0.788 | 0.873 | 0.941 | 0.754 | 0.831 |
| D4-SURG | 0.61 | 0.627 | 0.653 | 1 | 0.636 | 0.856 | 0.924 | 0.78 | 0.805 |
| D5-ELEV | 0.712 | 0.627 | 0.669 | 1 | 0.72 | 0.907 | 0.915 | 0.831 | 0.805 |
| D6-FULL | 0.915 | 0.958 | 0.89 | 1 | 0.932 | 0.992 | 0.992 | 0.949 | 0.958 |
| **Balanced Accuracy** | | | | | | | | | |
| D1-BASE | 0.619 | 0.714 | 0.645 | 0.5 | 0.5 | 0.528 | 0.466 | 0.568 | 0.485 |
| D2-DIST | 0.748 | 0.777 | 0.757 | 0.5 | 0.688 | 0.647 | 0.503 | 0.639 | 0.656 |
| D3-WSP | 0.636 | 0.595 | 0.661 | 0.5 | 0.803 | 0.8 | 0.471 | 0.559 | 0.598 |
| D4-SURG | 0.623 | 0.632 | 0.645 | 0.5 | 0.5 | 0.565 | 0.462 | 0.572 | 0.448 |
| D5-ELEV | 0.72 | 0.723 | 0.744 | 0.5 | 0.724 | 0.636 | 0.458 | 0.643 | 0.63 |
| D6-FULL | 0.867 | 0.979 | 0.9 | 0.682 | 0.784 | 0.996 | 0.678 | 0.929 | 0.752 |

Table 4.15: Numerical summary of the influence of up-sampling on prediction performance of model NL3-GBT under D6-FULL strategy

| *Type* | *Metric* | | | | |
|---|---|---|---|---|---|
| | *CV Accuracy* | *Test Accuracy* | *Test Sensitivity* | *Test Specificity* | *Balanced Accuracy* |
| **Unbalanced** | 0.97 | 0.946 | 0.455 | 0.992 | 0.724 |
| **Up-sampled** | 0.991 | 0.992 | 1 | 0.992 | 0.996 |

### 4.3.3 Interpretation of Linear Model

As described in the previous section, in general, nonlinear predictive model exhibits superior prediction performance than linear predictive model due to the nonlinear nature between evacuation response output and different input variables. However, this does not imply that linear model, such as LR, has no usefulness. As compared to nonlinear predictive model, linear predictive model has more meaningful model interpretation, which will help enrich the understanding and evidence base of evacuation process for healthcare professionals. For instance, GBT is the best nonlinear predictive model obtained. However, compared to LR model, it cannot quantify the influence of different input variables and evaluate their statistical significance (e.g., using p-value) due to the data uncertainties. The linear and additive model structure of LR makes all the above model interpretation capability possible. Table 4.16 summarizes the model estimation results of LR by including all GIS-features. For significant features, both the signs and magnitude of their estimated coefficients, e.g., $\hat{\beta}$, have meaningful interpretations. Positive (or negative) sign of a feature indicates that the increased value of that feature will increase (or decrease) the probability of a NH to be evacuated. Further, the actual influence of a feature can be quantified by the adjusted odds ratio value, which is a ratio of the evacuation probability over the shelter-in-place probability of a NH by holding other features constant.

Based on the typical choice of the significance level of 0.05, significant features are available from different aspects, such as NH organizational structure, environmental

Table 4.16: LR model interpretation

| *Feature* | $\hat{\beta}$ | *Odds Ratio* | *Std. Error($\hat{\beta}$)* | *p-value* |
|---|---|---|---|---|
| **Organizational structure** | | | | |
| Not for-profit facility (Y/N) | 2.049 | 7.76 | 0.949 | 0.031* |
| Government facility (Y/N) | 1.908 | 6.74 | 2.507 | 0.447 |
| Chain-facility (Y/N) | 0.481 | 1.618 | 0.781 | 0.538 |
| Part of a CCRC (Y/N) | 0.391 | 1.478 | 1.057 | 0.712 |
| Occupancy rate (% beds) | -0.008 | 0.992 | 0.038 | 0.821 |
| ADRD special care unit (Y/N) | 2.142 | 8.516 | 1.252 | 0.087 |
| Non-ADRD special care unit (Y/N) | -2.015 | 0.133 | 1.857 | 0.278 |
| Has organized resident group (Y/N) | -2.597 | 0.074 | 1.745 | 0.137 |
| Has organized family members group (Y/N) | -0.904 | 0.405 | 0.707 | 0.201 |
| Medicare (% residents) | -0.01 | 0.99 | 0.013 | 0.419 |
| Medicaid (% residents) | 0.062 | 1.064 | 0.038 | 0.109 |
| **Environmental GIS characteristics** | | | | |
| Distance from projected trajectory 24 hours prior decision (Km) | -0.142 | 0.868 | 0.027 | 1.17E-07*** |
| 50 knots wind speed probability 24hrs prior (%) | -0.375 | 0.687 | 0.07 | 6.94E-08*** |
| Potential storm surge 24 hrs prior (feet) | -0.492 | 0.611 | 0.35 | 0.16 |
| Elevation of facility (feet) | -0.142 | 0.868 | 0.032 | 8.65E-06*** |
| **Staffing characteristics** | | | | |
| Nurse with admin duties (HPRD) | 0.838 | 2.312 | 2.424 | 0.729 |
| Registered nurses (HPRD) | -2.58 | 0.076 | 1.811 | 0.154 |
| Licensed practical nurses (HPRD) | -3.358 | 0.035 | 1.504 | 0.026* |
| Certified nursing assistants (HPRD) | 1.119 | 3.062 | 0.964 | 0.246 |
| Occupational therapy services (HPRD) | -0.197 | 0.821 | 3.782 | 0.958 |
| Activities staff (HPRD) | 2.501 | 12.2 | 2.903 | 0.389 |
| Housekeeping staff (HPRD) | 0.592 | 1.808 | 1.376 | 0.667 |
| Medical director only (Y/N) | -0.363 | 0.696 | 1.097 | 0.741 |
| Full medical team (Y/N) | -0.299 | 0.742 | 0.867 | 0.73 |
| No medical team (Y/N) | 2.49 | 12.06 | 2.018 | 0.217 |

**Notes:** 1) * p<0.05; ** p<0.01; *** p<0.001;
2) 95% Confidence Intervals for each parameter estimate are calculated by $\hat{\beta}_j \pm 1.96 \times SE(\hat{\beta})$ , where is the respective estimated covariate coefficient and SE is the Std. Error.

Table 4.16: Continued

| *Feature* | $\hat{\beta}$ | *Odds Ratio* | *Std. Error($\hat{\beta}$)* | *p-value* |
|---|---|---|---|---|
| **Resident Characteristics (% residents)** | | | | |
| Acuindex (patient acuity) | 0.579 | 1.784 | 0.34 | 0.088 |
| Behavioral healthcare needs | 0.02 | 1.02 | 0.025 | 0.431 |
| Dementia or Alzheimer's | -0.041 | 0.96 | 0.027 | 0.135 |
| Depression | -0.004 | 0.996 | 0.018 | 0.835 |
| Intellectual disability | 0.064 | 1.066 | 0.123 | 0.605 |
| Physical restraint use | 0.131 | 1.14 | 0.358 | 0.715 |
| Serious Mental Illness | 0.016 | 1.016 | 0.032 | 0.615 |
| Antipsychotics medication | -0.033 | 0.968 | 0.046 | 0.478 |
| Antianxiety medication | -0.106 | 0.899 | 0.051 | 0.039* |
| Antidepressants medication | 0.034 | 1.035 | 0.031 | 0.277 |
| Sedative/hypnotics medication | 0.118 | 1.125 | 0.075 | 0.118 |

**Notes:** 1) * p<0.05; ** p<0.01; *** p<0.001;
2) 95% Confidence Intervals for each parameter estimate are calculated by $\hat{\beta}_j \pm 1.96 \times SE(\hat{\beta})$ , where is the respective estimated covariate coefficient and SE is the Std. Error.

conditions, caregivers working in a NH and dwelling NH residents. It confirms the need of fusing multi-source data to investigate and identify the multi-factorial determinants for NH evacuation. From organizational structure perspective, the type of ownership is a significant factor and a not-for-profit NH is more likely to evacuate (e.g., AOR = 7.76) than a for-profit NH by holding other features the same. It could be explained due to several following reasons. First, compared to not-for-profit NHs, for-profit NHs may have a less well-prepared evacuation plan, making evacuation on their own challenging. Existing studies indicated that for-profit NHs tend to have a less effective and adequate evacuation plan with higher chance of being cited for evacuation plan deficiencies (Castle, 2008). NH evacuation is a complex process involving moving frail residents to the designated receiving facilities with adequate medical equipment, food, water, medication, medical record and caregivers. A thoughtful and adequate evacuation plan includes detailed evacuation procedures, transportation logistics and evacuation provisions, and will be an

essential basis for ensuring successful NH evacuation. Further, for-profit NHs may also have other barriers (Brown *et al.*, 2007; Thomas *et al.*, 2012), such as limited logistics and financial support from public agencies and/or a lack of economic incentives for moving due to costly transportation.

From environmental condition perspective, three GIS features, namely, distance from a NH to the projected storm path, the neighboring wind speed of a NH and NH elevation, play significant roles in influencing evacuation decision. Specifically, the farther distance a NH to the projected storm trajectory, the less likelihood the NH will be evacuated due to a lower chance of experiencing a hurricane threat. In particular, an unit increase (in Km) of distance from a NH's location to the projected storm trajectory will decrease odds ratio between evacuation and shelter-in-place by a factor of 0.868 by holding other features fixed. Similarly, the greater probability of anticipated wind speed exceeding 50 knots within a 120-hour time period at a NH location, the less likely that NH will evacuate with an AOR of 0.687. The severe weather conditions may present significant disruption in transfer efforts and evacuation safety since 50 knots winds and gusts or above may break branches, uproot trees, or tip or veer high profile vehicles off course (Lindell *et al.*, 2007; Barua, 2005). In addition, a NH situated on higher ground will be less likely to evacuate. For an unit elevation increase (in feet), the odds ratio between evacuation and shelter-in-place will be decreased by a factor of 0.868. This is also intuitive and self-explanatory since NH at higher elevation will be less likely experiencing potential flooding from coastal surge or inland inundation.

From NH staffing and residents' characteristics perspective, two features, namely the staffing level of licensed practical nurses and the percentage of NH residents who received antianxiety medication, play significant roles in affecting whether a NH will be evacuated or not. Specifically, a NH with a higher staffing level of LPN (quantified in HPRD) indicates a low likelihood of evacuation, which can be explained from two aspects. First, LPNs are nursing staff who provide basic routine medical care, such as monitoring and recording

vital signs (e.g., blood pressure, heart rate, respiration, etc.) of patients, giving injections, changing bandages and administering medications. Adequate LPNs implies that the NH has adequate workforce to self-sufficiently take care of NH residents during hurricane and sheltering-in-place requires such self-sufficiency (FLHCA, 2008). Second, adequate LPNs can also mitigate the potential shortage of other types of nursing staff (e.g., RN, CNA) due to hurricane, which further strengthens the workforce self-sufficiency. For instance, many of LPNs also perform similar tasks as CNAs on providing direct bedside care (e.g., feeding, dressing, personal hygiene, walking, etc.) for NH residents. Experienced LPNs may be also responsible for advanced nursing activities, such as developing care plans, which are often conducted by RNs. In addition, a NH with a higher percentage of residents who receive antianxiety medication is less likely to evacuate. It is because that NH residents with pre-existing mental disorders are more vulnerable to evacuation. The changing environment in the new hosting facility, the discontinuity of care and moving itself due to evacuation will exacerbate their mental health conditions, such as anxiety, depression and post-traumatic stress disorders (Blanchard *et al.*, 2009). Besides, releasing stress, providing reassurance and persuasion to manage mental disorder symptoms of these residents during evacuation is also challenging for caregivers with limited time but overwhelming workload (Laditka *et al.*, 2008; Claver *et al.*, 2013).

## 4.4 Conclusion

In this paper, a GIS-integrated predictive analytics framework is proposed for predicting evacuation response of NHs in hurricane disaster scenario. Data from multiple sources, such as environmental conditions, resident census in the facility, and facility staffing and organizational characteristics during the time of disaster are considered and integrated for improving the prediction performance. Specifically, several important spatial and temporal heterogeneous environmental GIS features are extracted for NHs at different spatial locations, e.g., distance to storm trajectory, projected wind speed, potential storm surge,

and elevation of the facility. A number of linear and nonlinear machine learning models are applied and optimized for predicting the evacuation response and compared based on different prediction performance measures identify the final best predictive model. A case study on hurricane Irma impacting NHs in FL is considered to demonstrate effectiveness of the framework, comparing prediction performance among models with and without incorporating GIS features. Furthermore, the influence of the GIS-features are quantified, together with several resident and facility characteristics identified as influential factors for evacuation response. The proposed framework will allow NH administrators understand the multifactorial complex nature of evacuation response and the predictive capability with improved accuracy will assist emergency management agencies plan proactive resource management strategies for evacuation demand surge during disasters, such as hurricanes.

# Chapter 5
# Conclusion

In this dissertation, a series of predictive analytics approaches with both models and computational tools are developed for analyzing the complex healthcare data for older adults to improve prediction of both health outcomes of individual older adults as well as emergency response of aged care facilities. At the individual level, the improved prediction of health outcomes of individual older adults will help inform the resource prioritization decision of healthcare administrators by identifying the most at-risk individual with worse health conditions. The developed individualized prediction models will also serve as essential basis for proactive and individualized care delivery to optimize the health outcomes of individuals. At the facility level, the improved prediction of emergency responses, such as evacuation response, of aged care facilities (e.g., nursing homes), will help inform the emergency response planning activities of emergency operations agencies as well as public health associations to provide more proactive support in meeting with the surge demand of nursing home evacuees.

Chapter 2 focused on modeling the heterogeneous degradation of cognitive performance outcomes among community-dwelling older adults at both sub-population level and individual levels. At sub-population level, the developed Bayesian non-parametric model relaxes the conventional sub-population level modeling assumptions of pre-specifying a fixed number of sub-populations and allows the joint model estimation and identification of sub-population numbers. At individual level, the functional principal components analysis techniques were adopted to realize both individual prediction accuracy improvement

and meaningful model interpretations. For future research, multidimensional degradation of functional performance outcomes, such as both physical and cognitive functioning performance, will be investigated. In addition, the relationship between the heterogeneous functional performance of older adults and their associated healthcare utilizations, such as nursing home visits, will be also investigated.

Chapter 3 focused on modeling the heterogeneous LOS data of post-acute care residents with multiple discharge dispositions (e.g., community discharge, re/hospitalization) and varied individual characteristics. A semi-parametric hazard regression model was considered to jointly predict the individualized re/hospitalization risk and community discharge likelihood over time in the presence of varied individual characteristics. A simulation algorithm was further developed to generate accurate predictive samples of a heterogeneous population of post-acute care individuals with varied individual characteristics and multiple competing discharge dispositions. For further research, staffing characteristics and other facility level environmental characteristics will be incorporated to improve the prediction accuracy of the developed predictive model. Also, multiple nursing home facilities data will be considered, and the influence of unobserved factors shared within each facility will be further incorporated to improve the prediction accuracy.

Chapter 4 focused on predicting the evacuation response, namely evacuating or sheltering-in-place, for nursing home during natural disaster scenarios of hurricanes, by systematically extracting multiple storm features and integrating them with varied facility and resident characteristics to improve the overall prediction accuracy. Various spatial and temporal heterogeneous GIS features relevant to storm and environmental conditions of each nursing home facility are extracted, such as proximity to hurricane trajectory, anticipated wind speed and flood height, elevation, etc. A case study of recent hurricane Irma is considered to demonstrate the benefits of the predictive performance improvement of the proposed model with considering GIS features with various linear and non-linear machine learning models without considering GIS features. Both the model

prediction performance and interpretation capability are also investigated. For further research, the appropriateness of nursing home evacuation will be also investigated. In addition, based on the developed predictive model with improved surge demand prediction of evacuees, proactive evacuation planning models will be investigated to better help emergency operation agencies to realize cost-efficient network-wide coordination and management of multiple nursing facility during hazard scenarios.

## References

Acharya, U. R., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H., and Adam, M. (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. *Information Sciences*, 415, 190-198.

Ågotnes, G., Jacobsen, F. F., Harrington, C., and Petersen, K. A. (2016). A critical review of research on hospitalization from nursing homes; what is missing?. *Ageing International*, 41(1), 3-16.

Allen, M., Spencer, A., Gibson, A., Matthews, J., Allwood, A., Prosser, S., and Pitt, M. (2015). Right cot, right place, right time: improving the design and organisation of neonatal care networks.

Anderson, G. F., and Steinberg, E. P. (1985). Predicting hospital readmissions in the Medicare population. *Inquiry*, 251-258.

Anderson, L. A., Goodman, R. A., Holtzman, D., Posner, S. F., and Northridge, M. E. (2012). Aging in the United States: opportunities and challenges for public health. *American journal of public health*, 102(3), 393-395.

Ansah, J. P., Eberlein, R. L., Love, S. R., Bautista, M. A., Thompson, J. P., Malhotra, R., and Matchar, D. B. (2014). Implications of long-term care capacity response policies for an aging population: A simulation analysis. *Health Policy*, 116(1), 105-113.

Artetxe, A., Beristain, A., and Grana, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Computer methods and programs in biomedicine*, 164, 49-64.

Austin, P. C., Rothwell, D. M., and Tu, J. V. (2002). A comparison of statistical modeling strategies for analyzing length of stay after CABG surgery. *Health Services and Outcomes Research Methodology*, 3(2), 107-133.

Baker, E. J. (1991). Hurricane evacuation behavior. *International journal of mass emergencies and disasters*, 9(2), 287-310.

Bard, J. F., and Purnomo, H. W. (2005a). Preference scheduling for nurses using column generation. *European Journal of Operational Research*, 164(2), 510-534.

Bard, J. F., and Purnomo, H. W. (2005b). Short-term nurse scheduling in response to daily fluctuations in supply and demand. *Health Care Management Science*, 8(4), 315-324.

Barua, D. K. (2005). Beaufort wind scale. *Encyclopedia of coastal science Encyclopedia of earth science series*. Springer, Dordrecht. Accessed at: https://doi.org/10.1007/1-4020-3880-1_45

Beliën, J., Demeulemeester, E., and Cardoen, B. (2006). Visualizing the demand for various resources as a function of the master surgery schedule: A case study. *Journal of medical systems*, 30(5), 343-350.

Bhattacharjee, P., and Ray, P. K. (2016). Simulation modelling and analysis of appointment system performance for multiple classes of patients in a hospital: a case study. *Operations Research for Health Care*, 8, 71-84.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer-Verlag, NY

Blanchard, E. B., Jones-Alexander, J., Buckley, T. C., and Forneris, C. A. (1996). Psychometric properties of the PTSD Checklist (PCL). *Behaviour research and therapy*, 34(8), 669-673.

Blanchard, G., and Dosa, D. (2009). A comparison of the nursing home evacuation experience between Hurricanes Katrina (2005) and Gustav (2008). *Journal of the American Medical Directors Association*, 10(9), 639-643.

Bloom, D. E., and Luca, D. L. (2016). The global demography of aging: facts, explanations, future. In *Handbook of the economics of population aging* (Vol. 1, pp. 3-56). North-Holland.

Bonder, B. R., and Dal Bello-Haas, V. (2017). *Functional performance in older adults*. FA Davis.

Bravo-José, P., Moreno, E., Espert, M., Romeu, M., Martínez, P., and Navarro, C. (2018). Prevalence of sarcopenia and associated factors in institutionalised older adult patients. *Clinical nutrition ESPEN*, 27, 113-119.

Brown, L. M., Hyer, K., and Polivka-West, L. (2007). A comparative study of laws, rules, codes and other influences on nursing homes' disaster preparedness in the Gulf Coast states. *Behavioral sciences and the law*, 25(5), 655-675.

Brown, M. T. (2010) Early-life characteristics, psychiatric history, and cognition trajectories in later life. *The Gerontologist*, 50(5), 646-656.

Brown, L. M., Dosa, D. M., Thomas, K., Hyer, K., Feng, Z., and Mor, V. (2012). The effects of evacuation on nursing home residents with dementia. *American Journal of Alzheimer's Disease and Other Dementias* ®, 27(6), 406-412.

Bryk, A. S., and Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological bulletin*, 101(1), 147.

Burke, E. K., De Causmaecker, P., Berghe, G. V., and Van Landeghem, H. (2004). The state of the art of nurse rostering. *Journal of scheduling*, 7(6), 441-499.

Campbell, N. C., Murray, E., Darbyshire, J., Emery, J., Farmer, A., Griffiths, F., Guthrie, B., Lester, H., Wilson, P., and Kinmonth, A. L. (2007). Designing and evaluating complex interventions to improve health care. *BMJ*, 334(7591), 455-459.

Cappanera, P., Visintin, F., and Banditori, C. (2014). Comparing resource balancing criteria in master surgical scheduling: A combined optimisation-simulation approach. *International Journal of Production Economics*, 158, 179-196.

Carey, K. (2002). Hospital length of stay and cost: a multilevel modeling analysis. *Health Services and Outcomes Research Methodology*, 3(1), 41-56.

Castellucci, M. (2018). Most skilled-nursing facilities penalized by CMS for readmission rates. *Modern Healthcare*.

Castle, N. G. (2008). Nursing home evacuation plans. *American journal of public health*, 98(7), 1235-1240.

Centers for Disease Control and Prevention(CDC). (2011). Healthy Aging at a Glance 2011. Accessed at: http://stacks.cdc.gov/view/cdc/22022

Centers for Medicare and Medicaid Services (CMS). (2012). Chronic Conditions among Medicare Beneficiaries, Chartbook, 2012 Edition. Baltimore, MD.

Centers for Medicare and Medicaid Services. (2013). MDS 3.0 quality measures: user's manual. Research Triangle Park, NC: RTI International.

Centers for Medicare and Medicaid Services (CMS). (2017). Certification and Survey Provider Enhanced Reports (CASPER). Accessed at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/NursingHomeQualityInits/Skilled-Nursing-Facility-Quality-Reporting-Program/SNF-Quality-Reporting-Program-Public-Reporting

Centers for Medicare and Medicaid Services. (2017). Long-term care facility resident assessment instrument 3.0 user's manual version 1.13.

Chan, W. K. V., Son, Y. J., and Macal, C. M. (2010). Agent-based simulation tutorial-simulation of emergent behavior and differences between agent-based simulation and discrete-event simulation. In *Proceedings of the 2010 winter simulation conference*. IEEE.

Chan, C. W., Dong, J., and Green, L. V. (2017). Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research*, 65(2), 469-495.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1), 1-6.

Chen, X., Meaker, J. W., and Zhan, F. B. (2006). Agent-based modeling and analysis of hurricane evacuation procedures for the Florida Keys. *Natural Hazards*, 38(3), 321.

Cheng, Y., Goodin, A. J., Pahor, M., Manini, T., and Brown, J. D. (2020). Healthcare utilization and physical functioning in older adults in the United States. *Journal of the American Geriatrics Society*, 68(2), 266-271.

Claver, M., Dobalian, A., Fickel, J. J., Ricci, K. A., and Mallers, M. H. (2013). Comprehensive care for vulnerable elderly veterans during disasters. *Archives of gerontology and geriatrics*, 56(1), 205-213.

Cole, S. R., Chu, H., and Greenland, S. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *American journal of epidemiology*, 179(2), 252-260.

Combes, C., Kadri, F., and Chaabane, S. (2014). Predicting hospital length of stay using regression models: Application to emergency department. In 10ème Conférence Francophone de Modélisation, *Optimisation et Simulation-MOSIM'14*.

Cortada, J. W., Gordon, D., and Lenihan, B. (2012). The value of analytics in healthcare: From insights to outcomes. *IBM Global Business Services, Executive Report.*

Cowles, C. M. (2002). *Nursing home statistical yearbook, 2002*. Montgomery Village, MD: Cowles Research Group.

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

Cui, F., Ma, L., Hou, G., Pang, Z., Hou, Y., and Li, L. (2018). Development of smart nursing homes using systems engineering methodologies in industry 4.0. *Enterprise Information Systems*, 14(4), 463-479.

D'Agostino, R. B., and Nam, B. H. (2003). Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of statistics*, 23, 1-25.

DeRienzo, C. M., Shaw, R. J., Meanor, P., Lada, E., Ferranti, J., and Tanaka, D. (2017). A discrete event simulation tool to support and predict hospital and clinic staffing. *Health informatics journal*, 23(2), 124-133.

Dash, N., and Gladwin, H. (2007). Evacuation decision making and behavioral responses: Individual and household. *Natural hazards review*, 8(3), 69-77.

Dobalian, A., Claver, M., and Fickel, J. J. (2010). Hurricanes Katrina and Rita and the Department of Veterans Affairs: a conceptual model for understanding the evacuation of nursing homes. *Gerontology*, 56(6), 581-588.

Dodge, H. H., Du, Y., Saxton, J. A., and Ganguli, M. (2006). Cognitive domains and trajectories of functional independence in nondemented elderly persons. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 61(12), 1330-1337.

Dora, P., and Sekharan, G. H. (2015). Healthcare insurance fraud detection leveraging big data analytics. *IJSR*, 4(4), 2073-2076.

Dosa, D. M., Grossman, N., Wetle, T., and Mor, V. (2007). To evacuate or not to evacuate: lessons learned from Louisiana nursing home administrators following Hurricanes Katrina and Rita. *Journal of the American Medical Directors Association*, 8(3), 142-149.

Dosa, D. M., Hyer, K., Brown, L. M., Artenstein, A. W., Polivka-West, L., and Mor, V. (2008). The controversy inherent in managing frail nursing home residents during complex hurricane emergencies. *Journal of the American Medical Directors Association*, 9(8), 599-604.

Dosa, D., Feng, Z., Hyer, K., Brown, L. M., Thomas, K., and Mor, V. (2010). Effects of Hurricane Katrina on nursing facility resident mortality, hospitalization, and functional decline. *Disaster medicine and public health preparedness*, 4(01), S28.

Dosa, D., Hyer, K., Thomas, K., Swaminathan, S., Feng, Z., Brown, L., and Mor, V. (2012). To evacuate or shelter in place: implications of universal hurricane evacuation policies on nursing home residents. *Journal of the American Medical Directors Association*, 13(2), 190-e1.

Dosa, D. M., Skarha, J., Peterson, L. J., Jester, D. J., Sakib, N., Ogarek, J., Thomas, K. S., Andel, R., and Hyer, K. (2020). Association between exposure to Hurricane IRMA and mortality and hospitalization in Florida nursing home residents. *JAMA network open*, 3(10), e2019460-e2019460.

Eiken, S., Sredl, K., Burwell, B., and Amos, A. (2014). Medicaid expenditures for long-term services and supports in FFY 2012. Bethesda, MD: Truven Health Analytics.

El-Darzi, E., Vasilakis, C., Chaussalet, T., and Millard, P. H. (1998). A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health care management science*, 1(2), 143-149.

Environmental Systems Research Institute (ESRI). (2021). ArcGIS Desktop. Redlands, CA.

Faddy, M., Graves, N., and Pettitt, A. (2009). Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2), 309-314.

Evans, J. R., and Lindner, C. H. (2012). Business analytics: the next frontier for decision sciences. *Decision Line*, 43(2), 4-6.

Fashaw, S. A., Thomas, K. S., McCreedy, E., and Mor, V. (2020). Thirty-year trends in nursing home composition and quality since the passage of the Omnibus Reconciliation Act. *Journal of the American Medical Directors Association*, 21(2), 233-239.

Fernandez, L. S., Byard, D., Lin, C. C., Benson, S., and Barbera, J. A. (2002). Frail elderly as disaster victims: emergency management strategies. *Prehospital and disaster medicine*, 17(2), 67-74.

Ferrucci, L., Guralnik, J. M., Simonsick, E., Salive, M. E., Corti, C., and Langlois, J. (1996) Progressive versus catastrophic disability: a longitudinal view of the disablement process. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 51(3), M123-M130.

Florida Agency for Healthcare Administration. (2021). Accessed at: https://ahca.myflorida.com/

Florida Health Care Association. (2008). National criteria for evacuation decision-making in nursing homes. 2008.

Fogel, R. W. (1997). New findings on secular trends in nutrition and mortality: some implications for population theory. *Handbook of population and family economics*, 1, 433-481.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3), 189-198.

Friedman, J., Hastie, T., and Tibshirani, R. (2001) *The elements of statistical learning (Vol. 1)*, Springer, Berlin: Springer series in statistics.

Friesen, M. R., and McLeod, R. D. (2014). A survey of agent-based modeling of hospital environments. *IEEE Access*, 2, 227-233.

Galor, O., and Weil, D. N. (2000). Population, technology, and growth: From Malthusian stagnation to the demographic transition and beyond. *American economic review*, 90(4), 806-828.

Garber, J., Keiley, M. K., and Martin, N. C. (2002). Developmental trajectories of adolescents' depressive symptoms: predictors of change. *Journal of consulting and clinical psychology*, 70(1), 79.

Ge, X., Conger, R. D., and Elder Jr, G. H. (2001). Pubertal transition, stressful life events, and the emergence of gender differences in adolescent depressive symptoms. *Developmental psychology*, 37(3), 404.

Gill, T. M., Gahbauer, E. A., Han, L., and Allore, H. G. (2010) Trajectories of disability in the last year of life. *New England Journal of Medicine*, 362(13), 1173-1180.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). John Wiley & Sons.

Gonzàlez, D., Piña, M., and Torres, L. (2008). Estimation of Parameters in Cox's Proportional Hazard Model: Comparisons between Evolutionary Algorithms and the Newton-Raphson Approach. In *Mexican International Conference on Artificial Intelligence*. pp. 513-523. Springer, Berlin, Heidelberg.

Gould, C. E., O'Hara, R., Goldstein, M. K., and Beaudreau, S. A. (2016). Multimorbidity is associated with anxiety in older adults in the Health and Retirement Study. *International journal of geriatric psychiatry*, 31(10), 1105-1115.

Green, P. J., and Richardson, S. (2001) Modelling heterogeneity with and without the Dirichlet process. *Scandinavian journal of statistics*, 28(2), 355-375.

Green, L. V. (2005). Capacity planning and management in hospitals. In *Operations research and health care* (pp. 15-41). Springer, Boston, MA.

Gul, M., and Guneri, A. F. (2015). A comprehensive review of emergency department simulation applications for normal and disaster conditions. *Computers & Industrial Engineering*, 83, 327-344.

Hachesu, P. R., Ahmadi, M., Alizadeh, S., and Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare informatics research*, 19(2), 121.

Harper, P. R., and Gamlin, H. M. (2003). Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *OR Spectrum*, 25(2), 207-222.

Harrell Jr, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4), 361-387.

Harris, G. (2005). In NH, a fight lost to rising waters. *New York Times*, 7, A1.

Harris-Kojetin, L., Sengupta, M., Park-Lee, E., and Valverde, R. (2013) Long-Term Care Services in the United States: 2013 Overview. *Vital & health statistics. Series 3, Analytical and epidemiological studies/[US Dept. of Health and Human Services, Public Health Service, National Center for Health Statistics]*, (37), 1-107.

Harris-Kojetin, L., Sengupta, M., Lendon, J.P., Rome, V., Valverde, R., and Caffrey, C. Long-term care providers and services users in the United States, 2015–2016. (2019). National Center for Health Statistics. *Vital Health Stat*. 3(43).

Hasan, S., Ukkusuri, S., Gladwin, H., and Murray-Tuite, P. (2011). Behavioral model to understand household-level hurricane evacuation decision making. *Journal of Transportation Engineering*, 137(5), 341-348.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hayduk, L. A. (1987). Structural equation modeling with LISREL: Essentials and advances. Jhu Press.

He, W., Larsen L. J. (2014) Older Americans With a Disability: 2008-2012. *U.S. Census Bureau and National Institute on Aging report*.

Helm, J. E., Lavieri, M. S., Van Oyen, M. P., Stein, J. D., and Musch, D. C. (2015). Dynamic forecasting and control algorithms of glaucoma progression for clinician decision support. *Operations Research*, 63(5), 979-999.

Hersh, W. R. (2014). Healthcare data analytics. *Health informatics: practical guide for healthcare*.

Hirshon, J. M., Risko, N., Calvello, E. J., Ramirez, S. S. D., Narayan, M., Theodosis, C., and O'Neill, J. (2013). Health systems and services: the role of acute care. *Bulletin of the World Health Organization*, 91, 386-388.

Hoenig, H., Nusbaum, N., and Brummel-Smith, K. (1997). Geriatric rehabilitation: state of the art. *Journal of the American Geriatrics Society*, 45(11), 1371-1381.

Holm, L. B., Lurås, H., and Dahl, F. A. (2013). Improving hospital bed utilisation through simulation and optimisation: with application to a 40% increase in patient volume in a Norwegian General Hospital. *International journal of medical informatics*, 82(2), 80-89.

Holup, A. A., Hyer, K., Meng, H., and Volicer, L. (2017). Profile of nursing home residents admitted directly from home. *Journal of the American Medical Directors Association*, 18(2), 131-137.

Hoot, N. R., LeBlanc, L. J., Jones, I., Levin, S. R., Zhou, C., Gadd, C. S., and Aronsky, D. (2008). Forecasting emergency department crowding: a discrete event simulation. *Annals of emergency medicine*, 52(2), 116-125.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1), 55-63.

Hyer, K., Polivka-West, L., and Brown, L. (2007). Nursing homes and assisted living facilities: Planning and decision making for sheltering in place or evacuation. *Generations*, 31(4), 29-33.

Hyer, K., Thomas, K. S., Branch, L. G., Harman, J. S., Johnson, C. E., and Weech-Maldonado, R. (2011). The influence of nurse staffing levels on quality of care in nursing homes. *The Gerontologist*, 51(5), 610-616.

Incalzi, R. A., Gemma, A., Capparella, O., Terranova, L., Porcedda, P., Tresalti, E., and Carbonin, P. (1992). Predicting mortality and length of stay of geriatric patients in an acute care general hospital. *Journal of gerontology*, 47(2), M35-M39.

Islam, M. S., Hasan, M. M., Wang, X., and Germack, H. D. (2018). A systematic review on healthcare analytics: application and theoretical perspective of data mining. In *Healthcare* (Vol. 6, No. 2, p. 54). Multidisciplinary Digital Publishing Institute.

Janiszewski Goodin, H. (2003). The nursing shortage in the United States of America: an integrative review of the literature. *Journal of advanced nursing*, 43(4), 335-343.

Jelesnianski, C. P. (1992). *SLOSH: Sea, lake, and overland surges from hurricanes* (Vol. 48). US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service.

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

Joynt, K. E., and Jha, A. K. (2013). A path forward on Medicare readmissions. *N Engl J Med*, 368(13), 1175-1177.

Kankanhalli, A., Hahn, J., Tan, S., and Gao, G. (2016). Big data and analytics in healthcare: introduction to the special section. *Information Systems Frontiers*, 18(2), 233-235.

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., and Kripalani, S. (2011). Risk prediction models for hospital readmission: a systematic review. *Jama*, 306(15), 1688-1698.

Katz, S., Ford, A. B., Moskowitz, R. W., Jackson, B. A., and Jaffe, M. W. (1963). Studies of illness in the aged: the index of ADL: a standardized measure of biological and psychosocial function. *Jama*, 185(12), 914-919.

Kelly, A., Conell-Price, J., Covinsky, K., Cenzer, I. S., Chang, A., Boscardin, W. J., and Smith, A. K. (2010). Length of stay for older adults residing in nursing homes at the end of life. *Journal of the American Geriatrics Society*, 58(9), 1701-1706.

Kim, K., and Mehrotra, S. (2015). A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. *Operations Research*, 63(6), 1431-1451.

Kokangul, A., Akcan, S., and Narli, M. (2017). Optimizing nurse capacity in a teaching hospital neonatal intensive care unit. *Health care management science*, 20(2), 276-285.

Konetzka, R. T., Stearns, S. C., and Park, J. (2008). The staffing–outcomes relationship in nursing homes. *Health services research*, 43(3), 1025-1042.

Kramer, A. A., and Zimmerman, J. E. (2010). A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. *BMC medical informatics and decision making*, 10(1), 1-16.

Kripalani, S., Theobald, C. N., Anctil, B., and Vasilevskis, E. E. (2014). Reducing hospital readmission rates: current strategies and future directions. *Annual review of medicine*, 65, 471-485.

Kuligowski, E. (2021). Evacuation decision-making and behavior in wildfires: Past research, current challenges and a future research agenda. *Fire Safety Journal*, 120, 103129.

Laditka, S. B., Laditka, J. N., Xirasagar, S., Cornman, C. B., Davis, C. B., and Richter, J. V. (2008). Providing shelter to nursing home evacuees in disasters: lessons from Hurricane Katrina. *American journal of public health*, 98(7), 1288-1293.

Lakshmi, C., and Iyer, S. A. (2013). Application of queueing theory in health care: A literature review. *Operations research for health care*, 2(1-2), 25-39.

Latham, L. P., and Ackroyd-Stolarz, S. (2014). Emergency department utilization by older adults: a descriptive study. *Canadian Geriatrics Journal*, 17(4), 118.

LaVito, A. (2017). Irma may have caused 42.5$billionto$65 billion in property damage, report says. CNBC. Accessed at: https://www.cnbc.com/2017/09/19/irma-may-have-caused-42-point-5-billion-to-65-billion-in-property-damage-report.html

Lee, A. H., Fung, W. K., and Fu, B. (2003). Analyzing hospital length of stay: mean or median regression? *Medical care*, 681-686.

Lepenioti, K., Bousdekis, A., Apostolou, D., and Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, 50, 57-70.

Li, M., Han, J., and Liu, J. (2016) Bayesian Nonparametric Modeling of Heterogeneous Time-to-event Data with an Unknown Number of Sub-populations. *IISE Transactions*, (accepted).

Li, Y., Zhang, Y., Kong, N., and Lawley, M. (2016). Capacity planning for long-term care networks. *IIE Transactions*, 48(12), 1098-1111.

Li, Y., Swift, S., and Tucker, A. (2013) Modelling and analysing the dynamics of disease progression from cross-sectional studies. *Journal of biomedical informatics*, 46(2), 266-274.

Liang, J., Xu, X., Quiñones, A. R., Bennett, J. M., and Ye, W. (2011) Multiple trajectories of depressive symptoms in middle and late life: racial/ethnic variations. *Psychology and aging*, 26(4), 761.

Liao, M., Li, Y., Kianifard, F., Obi, E., and Arcona, S. (2016). Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. *BMC nephrology*, 17(1), 1-14.

Lin, S. F., Beck, A. N., Finch, B. K., Hummer, R. A., and Master, R. K. (2012). Trends in US older adult disability: exploring age, period, and cohort effects. *American journal of public health*, 102(11), 2157-2163.

Lindell, M. K., Prater, C. S., and Peacock, W. G. (2007). Organizational communication and decision making for hurricane emergencies. *Natural Hazards Review*, 8(3), 50-60.

Liu, P., Lei, L., Yin, J., Zhang, W., Naijun, W., and El-Darzi, E. (2006). Healthcare data mining: Prediction inpatient length of stay. In *2006 3rd International IEEE Conference Intelligent Systems* (pp. 832-837). IEEE.

Lleras-Muney, A. (2005). The relationship between education and adult mortality in the United States. *The Review of Economic Studies*, 72(1), 189-221.

Lopes, J., Guimarães, T., and Santos, M. F. (2020). Predictive and prescriptive analytics in healthcare: a survey. *Procedia Computer Science*, 170, 1029-1034.

Lunney, J. R., Lynn, J., Foley, D. J., Lipson, S., and Guralnik, J. M. (2003) Patterns of functional decline at the end of life. *Jama*, 289(18), 2387-2392.

Lynch, S. M., and George, L. K. (2002) Interlocking trajectories of loss-related events and depressive symptoms among elders. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 57(2), S117-S125.

Maali, Y., Perez-Concha, O., Coiera, E., Roffe, D., Day, R. O., and Gallego, B. (2018). Predicting 7-day, 30-day and 60-day all-cause unplanned readmission: a case study of a Sydney hospital. *BMC medical informatics and decision making*, 18(1), 1-11.

McArdle, J. J., and Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child development*, 110-133.

McGuire, F. (1994). Using simulation to reduce length of stay in emergency departments. In *Proceedings of Winter Simulation Conference*. *IEEE*.

McGuire, L. C., Ford, E. S., and Okoro, C. A. (2007). Natural disasters and older US adults with disabilities: implications for evacuation. *Disasters*, 31(1):, 49-56.

Meredith, W., and Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55(1), 107-122.

Miller, E. A., and Weissert, W. G. (2000). Predicting elderly people's risk for nursing home placement, hospitalization, functional impairment, and mortality: a synthesis. *Medical care research and review*, 57(3), 259-297.

Moon, M. (1997). An examination of key Medicare provisions in the Balanced Budget Act of 1997.

Mount Sinai Hospital - Icahn School of Medicine. (2021). What is Urgent Care and When Should You Use It? Accessed at: https://www.mountsinai.org/locations/urgent-care/what-is-urgent-care

El Morr, C., and Ali-Hassan, H. (2019). *Analytics in Healthcare: A Practical Introduction*. Springer.

Morrow, B. H. (1999). Identifying and mapping community vulnerability. *Disasters*, 23(1), 1-18.

Murad, Y. (2011). Skilled nursing facilities and post-acute care. J Gerontol Geriatr Res, 1(101), 1-4.

Muthén, B., and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55(2), 463-469.

Nagin, D. S. (2005) *Group-based modeling of development*, Harvard University Press.

Nagin, D. S., and Odgers, C. L. (2010) Group-based trajectory modeling in clinical research. *Annual review of clinical psychology*, 6, 109-138.

Natarajan, N. (2013). Coastal Storm Planning, the Healthcare Facility Evacuation Center (HEC), and Patient Tracking. New York State

New, P. W., Racp, F., Stockman, K., Cameron, P. A., Olver, J. H., and Stoelwinder, J. U. (2015). Computer simulation of improvements in hospital length of stay for rehabilitation patients. *Journal of rehabilitation medicine*, 47(5), 403-411.

National Oceanic and Atmospheric Administration (NOAA). (2021). NHC Data in GIS Formats. Accessed at: https://www.nhc.noaa.gov/gis/

Negoescu, D. M., Bimpikis, K., Brandeau, M. L., and Iancu, D. A. (2018). Dynamic learning of patient response types: An application to treating chronic diseases. *Management science*, 64(8), 3469-3488.

National Investment Center for Seniors Housing & Care (NIC). (2018) Skilled Nursing Data Report: Key Occupancy & Revenue Trends. 4Q2017.

National Weather Service (NWS). (2017). Hurricane Irma Local Report/Summary. Accessed at: https://www.weather.gov/mfl/hurricaneirma

Ofstedal, M. B., Fisher, G. G., and Herzog, A. R. (2005). Documentation of cognitive functioning measures in the Health and Retirement Study, *HRS Documentation Report*.

Oh, C., Novotny, A. M., Carter, P. L., Ready, R. K., Campbell, D. D., and Leckie, M. C. (2016). Use of a simulation-based decision support tool to improve emergency department throughput. *Operations Research for Health Care*, 9, 29-39.

Ortman, J. M., Velkoff, V. A., and Hogan, H. (2014). An aging nation: the older population in the United States.

Parker, G., Lie, D., Siskind, D. J., Martin-Khan, M., Raphael, B., Crompton, D., and Kisely, S. (2016). Mental health implications for older adults after natural disasters-a systematic review and meta-analysis. *International Psychogeriatrics*, 28(1), 11.

Pearson, W. S., Bhat-Schelbert, K., and Probst, J. C. (2012). Multiple chronic conditions and the aging of America: challenge for primary care physicians. *Journal of Primary Care & Community Health*, 3(1), 51-56.

Pendharkar, P. C., and Khurana, H. (2014). Machine learning techniques for predicting hospital length of stay in pennsylvania federal and specialty hospitals. *International Journal of Computer Science & Applications*, 11(3).

Peng, J., and Paul, D. (2009) A Geometric Approach to Maximum Likelihood Estimation of the Functional Principal Components From Sparse Longitudinal Data. *Journal of Computational and Graphical Statistics*, 18(4), 995-1015.

Peterson, L. J., June, J., Sakib, N., Dobbs, D., Dosa, D. M., Thomas, K. S., Jester, D. J., and Hyer, K. (2020). Assisted living communities during Hurricane Irma: The decision to evacuate or shelter in place and resident acuity. *Journal of the American Medical Directors Association*, 21(8), 1148-1152.

Punnakitikashem, P., Rosenberber, J. M., and Buckley-Behan, D. F. (2013). A stochastic programming approach for integrated nurse staffing and assignment. *IIE Transactions*, 45(10), 1059-1076.

Pratt, J. R. (2010). Long-term care: Managing across the continuum. Jones & Bartlett Publishers.

Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population studies*, 29(2), 231-248.

Raghupathi, W., and Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1), 1-10.

Raina, P., O'Donnell, M., Schwellnus, H., Rosenbaum, P., King, G., Brehaut, J., Russel, D., Swinton, M., King, S., Wong, M., Walter, S. D., and Wood, E. (2004). Caregiving process and caregiver burden: conceptual models to guide research and practice. *BMC pediatrics*, 4(1), 1-13.

Ramsay, J. O., Silverman B. W. (2005) *Functional Data Analysis*, Springer, New York, NY.

Rau, C. L., Tsai, P. F. J., Liang, S. F. M., Tan, J. C., Syu, H. C., Jheng, Y. L., Ciou, T., and Jaw, F. S. (2013). Using discrete-event simulation in strategic capacity planning for an outpatient physical therapy service. *Health Care Management Science*, 16(4), 352-365.

Roberts, G. O., and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2), 291-317.

Schaefer, A. J., Bailey, M. D., Shechter, S. M., and Roberts, M. S. (2005). Modeling medical treatment using Markov decision processes. In *Operations research and health care* (pp. 593-612). Springer, Boston, MA.

Schnelle, J. F., Schroyer, L. D., Saraf, A. A., and Simmons, S. F. (2016). Determining nurse aide staffing requirements to provide care based on resident workload: a discrete event simulation model. *Journal of the American Medical Directors Association*, 17(11), 970-977.

Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statistica sinica*, 639-650.

Sharma, A. (2021). Estimating older adult mortality from COVID-19. *The Journals of Gerontology: Series B*, 76(3), e68-e74.

Shaughnessy, P. W., and Kramer, A. M. (1990). The increased needs of patients in nursing homes and patients receiving home health care. *New England Journal of Medicine*, 322(1), 21-27.

Shumway-Cook, A., Baldwin, M., Polissar, N. L., and Gruber, W. (1997). Predicting the probability for falls in community-dwelling older adults. *Physical therapy*, 77(8), 812-819.

Sims, O. T., Oh, H., Noh, H., Melton, P. A., Sheffield, S., Ingram, K., and Sawyer, P. (2018). Prevalence and predictors of co-occurring diabetes and hypertension in community-dwelling older adults. *Geriatrics & gerontology international*, 18(9), 1356-1360.

Southwick, S. M., Morgan, C. A., Darnell, A., Bremner, D., Nicolaou, A. L., Nagy, L. M., and Charney, D. S. (1995). Trauma-related symptoms in veterans of Operation Desert Storm: a 2-year follow-up. *The American journal of psychiatry*.

Sow, D., Turaga, D. S., and Schmidt, M. (2013). Mining of sensor data in healthcare: a survey. *Managing and mining sensor data*, 459-504.

Squazzoni, F., Polhill, J. G., Edmonds, B., Ahrweiler, P., Antosz, P., Scholz, G., Chappin, É., Borit, M., Vergahen, H., Giardini, F., and Gilbert, N. (2020). Computational models that matter during a global pandemic outbreak: a call to action. *Journal of Artificial Societies and Social Simulation*, 23(2), 10.

Stainsby, H., Taboada, M., and Luque, E. (2009). Towards an agent-based simulation of hospital emergency departments. In *2009 IEEE International Conference on Services Computing*. IEEE.

*Staff Time and Resource Intensity Verification (STRIVE) Project* (2005), Accessed at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/SNFPPS/TimeStudy.html

Strunk, B. C., Ginsburg, P. B., and Banker, M. I. (2006). The Effect Of Population Aging On Future Hospital Demand: A simulation of future spending finds that aging will not be the strongest influence on inpatient hospital use. *Health Affairs*, 25(Suppl1), W141-W149.

Taboada, M., Cabrera, E., Iglesias, M. L., Epelde, F., and Luque, E. (2011). An agent-based decision support system for hospitals emergency departments. *Procedia Computer Science*, 4, 1870-1879.

Tanner, M. A., and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398), 528-540.

Taylor, M. G., and Lynch, S. M. (2004). Trajectories of impairment, social support, and depressive symptoms in later life. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 59(4), S238-S246.

Texas Medical Management. (2020). Hospitals vs Surgery Centers - Where should I have my surgery? Accessed at: https://texasfreemarketsurgery.com/hospitals-vs-surgery-centers-where-should-i-have-my-surgery/

*The Health and Retirement Study (HRS)* (2016), Accessed at: https://hrs.isr.umich.edu/about

Thomas, K. S., Dosa, D., Hyer, K., Brown, L. M., Swaminathan, S., Feng, Z., and Mor, V. (2012). Effect of forced transitions on the most functionally impaired nursing home residents. *Journal of the American Geriatrics Society*, 60(10), 1895-1900.

Turgeman, L., May, J.H., and Sciulli, R. (2017). Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. *Expert Systems with Applications*. 78: p. 376-385

Travaille, P., Müller, R. M., Thornton, D., and Van Hillegersberg, J. (2011). Electronic Fraud Detection in the US Medicaid Healthcare Program: Lessons Learned from other Industries. In *AMCIS*.

University of Florida GeoPlan Center. (2012). Florida Digital Elevation Model (DEM) Mosaic. Accessed at: https://www.fgdl.org/metadata/fgdl_html/flidar_mosaic_m.htm

Ursprung, R., and Gray, J. (2010). Random safety auditing, root cause analysis, failure mode and effects analysis. *Clinics in perinatology*, 37(1), 141-165.

Ginsburg, R. B., and Supreme Court Of The United States. (1998) U.S. Reports: Olmstead v. L.C., 527 U.S. 581. [Periodical] Retrieved from the Library of Congress, https://www.loc.gov/item/usrep527581/

van Eeden, K., Moeke, D., and Bekker, R. (2016). Care on demand in nursing homes: a queueing theoretic approach. *Health care management science*, 19(3), 227-240.

Vemuri, P., Gunter, J. L., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S., Boeve, B. F., Peterson, R. C., and Jack Jr, C. R. (2008). Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *Neuroimage*, 39(3), 1186-1197.

Vuik, S. I., Mayer, E. K., and Darzi, A. (2016). Patient segmentation analysis offers significant benefits for integrated care and support. *Health Affairs*, 35(5), 769-775.

Walker, S. G. (2007) Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1), 45-54.

Wang, J., Li, J., Tussey, K., and Ross, K. (2012). Reducing length of stay in emergency department: A simulation study at a community hospital. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(6), 1314-1322.

Watanabe, M., and Yamaguchi, K. (eds.) (2003) *The EM algorithm and related statistical models*, Marcel Dekker, New York.

Weiner J. M., Hanley R. J. (1989) *Measuring the activities of daily living among the elderly: A guide to National Surveys*, US Department of Health and Human Services.

Whitehead, J. C., Edwards, B., Van Willigen, M., Maiolo, J. R., Wilson, K., and Smith, K. T. (2000). Heading for higher ground: Factors affecting real and hypothetical hurricane evacuation behavior. *Global Environmental Change Part B: Environmental Hazards*, 2(4), 133-142.

Willoughby, M., Kipsaina, C., Ferrah, N., Blau, S., Bugeja, L., Ranson, D., and Ibrahim, J. E. (2017). Mortality in nursing homes following emergency evacuation: a systematic review. *Journal of the American Medical Directors Association*, 18(8), 664-670.

Wolfson, M., and Rowe, G. (2014) HealthPaths: Using functional health trajectories to quantify the relative importance of selected health determinants. *Demographic Research*, 31, 941-974.

Wolshon, B., Urbina, E., Wilmot, C., and Levitan, M. (2005). Review of policies and practices for hurricane evacuation. I: Transportation planning, preparedness, and response. *Natural hazards review*, 6(3), 129-142.

Xie, H., Chaussalet, T. J., and Millard, P. H. (2005). A continuous time Markov model for the length of stay of elderly people in institutional long-term care. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1), 51-61.

Yadav, S. S., and Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6(1), 1-18.

Yao, F., Müller, H. G., and Wang, J. L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470), 577-590.

Zimmer, Z., Martin, L. G., Nagin, D. S., and Jones, B. L. (2012) Modeling disability trajectories and mortality of the oldest-old in China. *Demography*, 49(1), 291-314.

Zhang, J., Mason, J. E., Denton, B. T., and Pierskalla, W. P. (2011). Applications of operations research to the prevention, detection, and treatment of disease. *Encyclopedia of Operations Research and Management*, Wiley.

Zhang, Y., Puterman, M. L., Nelson, M., and Atkins, D. (2012). A simulation optimization approach to long-term care capacity planning. *Operations research*, 60(2), 249-261.

Zhang, X., Barnes, S., Golden, B., Myers, M., and Smith, P. (2019). Lognormal-based mixture models for robust fitting of hospital length of stay distributions. *Operations Research for Health Care*, 22, 100184.

# Appendix A Copyright Permissions

## A.1 Permission for Chapter 2

Copyright Clearance Center RightsLink®

Home | Help | Email Support | Sign in | Create Account

Taylor & Francis
Taylor & Francis Group

Journal Reprints

**Bi-level heterogeneity modeling of functional performance degradation for the aging population**

**Author:** Nazmus Sakib, , Xuxue Sun, et al

**Publication:** IISE Transactions on Healthcare Systems Engineering

**Publisher:** Taylor & Francis

**Date:** Jul 3, 2017

*Rights managed by Taylor & Francis*

**Thesis/Dissertation Reuse Request**

Taylor & Francis is pleased to offer reuses of its content for a thesis or dissertation free of charge contingent on resubmission of permission request if work is published.

BACK CLOSE