DEVELOPMENT AND VALIDATION OF RISK STRATIFICATION MODELS IN A COHORT OF COMMUNITY-LIVING HOMEBOUND OLDER ADULTS, COMPARISON OF THREE METHODS: LOGISTIC REGRESSION, RANDOM FOREST, AND COX PROPORTIONAL HAZARD REGRESSION

By

Mojdeh Nasiriahmadabadi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Epidemiology-Doctor of Philosophy

2019

ABSTRACT

DEVELOPMENT AND VALIDATION OF RISK STRATIFICATION MODELS IN A COHORT OF COMMUNITY-LIVING HOMEBOUND OLDER ADULTS, COMPARISON OF THREE METHODS: LOGISTIC REGRESSION, RANDOM FOREST, AND COX PROPORTIONAL HAZARD REGRESSION

By

Mojdeh Nasiriahmadabadi

Risk stratification (RS) models make predictions of an outcome based on the observed information from predictor variables. Classification of a population into different groups based on their risk of an outcome provides the opportunity for delivering targeted services to each group based on their needs and priorities. Different RS tools have been developed for older adults, but there is a limited number of RS studies developed for use in community-living older adults. This dissertation aims to develop and validate risk stratification models in a cohort of community-living homebound older adults. The study population consisted of older homebound adults who received home-based medical services from the Visiting Physician Association (VPA), which is a part of the United States Medical Management (USMM) Corporation. USMM provides a range of services, including home-based primary care and medical visits, senior home care, palliative care, and hospice services. The cohort had several features indicative of high risk: the average age was 82 years, 50% had ≥ 5 comorbidities, and 45% had a severe disability (defined by a Karnofsky Performance Score KPS ≤40). The population had very high rates of mortality and hospice admission (1-year rates were 32% and 10%, respectively). Given the unique and high-risk nature of this population, a RS approach was developed to help to provide USMM patients with appropriate services aligned with their priorities, as guided by a recent conceptual framework for the care of older adults with multiple comorbidities (Table 1.2). We developed and validated prediction models for two outcomes (death and hospice admission) by using three alternate statistical approaches: logistic regression (LR), random forest (RF), and Cox regression. The performance of these models was compared using the discrimination ability measured by area under the receiver operating curve (AUC).

When developing the LR model we applied different variable selection methods (stepwise, backward, forward, adaptive lasso, elastic net, and manual). We developed a prediction model using a RF algorithm and used Cox regression to model time-to-event for each outcome separately (using the same variable selection methods as used in Logistic regression). All three models were developed in a derivation dataset (consisting of a random 50% of the cohort) and validated by applying to the validation dataset. Because of the large amount of missing data among predictor variables we applied multiple imputation (MI) procedures and compared the performance of LR and RF models in the original data and imputed data. For the prediction of mortality, all of the variable selection methods used in the LR model showed similar predictive performance (AUC 0.762- 0.769). Random forest had the best discrimination ability (AUC=0.83), whereas the LR and Cox models had comparable AUCs (0.76 and 0.74 respectively). We determined that the higher AUC of the RF model was mainly due to its ability to include subjects with missing data because when the subjects with missing data were excluded from the RF cohort, the UAC of the model was similar to the LR model. Also when the RF model was applied to imputed data it has similar predictive performance as the LR model which indicated the basic assumption of multiple imputation (i.e., missing at random) was not met in this data. For hospice admission, all three models had a similar discriminative ability (AUC for RF, LR, and Cox, were 0.70, 0.73, and 0.72, respectively). The variables age, race, KPS, serum albumin, surprise question (SQ), and hyperlipidemia were consistently selected as the important predictors of both outcomes in all three approaches. WE concluded that the RF approach can significantly improve the predictive performance of the RS model but this advantage comes from its ability for the inclusion of observation with missing data. When data are missing not at random use of MI had a limited effect on improving the prediction of models because the basic assumption in MI procedure is missing at random. The quality of data from large electronic health record datasets remains a limitation of developing RS models.

This dissertation is lovingly dedicated to my mom for her thoughts and prayers,
to my family for their unconditional love and support,
to Pooya for relentlessly pushing me to work on it,
to Negar for asking every single day if I'm done yet,

also to my fingernails for surviving the many years of stressful chewing,
and to my many sleepless nights, for making my PhD a truly Permanent Head Damage

ACKNOWLEDGMENTS

Undertaking this PhD has been a truly rewarding and life-altering experience for me and it would not have been possible to do without all the support and guidance that I received.

I would like to express my sincere gratitude to my advisor, Prof. Mathew Reeves, for all the support and encouragement throughout my Ph.D. study and related research. His immense knowledge and experience helped me through the research and writing of this dissertation. I cannot imagine having completed this dissertation without his guidance and support.

I would like to also thank the rest of my dissertation committee: Prof. Joseph Gardiner, Dr. David Todem, and Dr. Erin Sarzynski for their insightful comments and encouragement, and also for the tough questions which incented me to widen my research and perspective.

My sincere thanks goes to Dr. John Strandmark – corporate medical director of the Grace hospice and the representative of US Medical Management Corporation – for the continuous support of this research, and for his patience, motivation, and knowledge. His guidance helped me through each stage of this process. I gratefully acknowledge the funding towards my Ph.D. dissertation and the data for this project from the US Medical Management Corporation.

It has been a privilege and an honor to work with each and every one of these distinguished people.

I would like to say a heartfelt thanks to my family, my mom, and my brothers and sisters for supporting me spiritually throughout my PhD study, as well as my life in general; and to my late father, for instilling the love of reading in me, in those early years. I am also very grateful to Dr. Negar Salehi for all the support and encouragement she gave me.

Last but not the least, a very special thank you to Pooya for his invaluable advice and feedback on my dissertation, and for always being so supportive of my work. This work would have not been possible without you, Pooya.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

This dissertation aims to develop risk stratification (RS) models using a cohort of the US Medical

Management (USMM) patient population, which is a unique population of community-living

homebound older adults. The USMM organization approached my advisor and me in search of a

collaboration with academic partners to develop RS models that can optimally improve their pre-existing

RS approaches. The RS models were needed to identify patients at high risk of death and those at high

risk of hospice admission in the near future and to provide them with appropriate customized palliative

care. The ultimate goal was to improve the quality and timing of the healthcare they provide to the

patients at different risk levels, e.g. different palliative care services including hospice referral. The

collaboration started in 2014, and a cohort of the USMM patients cared for in the calendar year 2015

was assembled for this study. The need to develop RS models specific to the USMM population is based

on the unique characteristics of the USMM population, and the intention of the USMM organization to

implement the most accurate RS in its population.

This chapter is organized around the following sections including a summary of older adults care options

(as currently implemented in the US), a descriptions of the USMM Corporation, its patient population,

and the alternative RS approaches proposed by the USMM providers. The role of RS in population-based

disease management is also explained, and a summary of relevant RS literature is provided. Finally,

three specific aims and analysis plan for each are described.

- *Current care services available for older adults in the US*

As the population of older adults is growing, different types of services for older adults care have been

developed. There are several ways to categorize the older adults care services, for example based on the

location of residence (home or institution), based on services that offered (skilled vs. custodial), or

based on the purpose of health care (long term care facilities, palliative care, or hospice). A continuum

of care for older adults can be described as home-services (such as ADL assistance, IADL assistance), home health services, adult day care, assisted living and retirement communities, skilled nursing facilities and long term care units, palliative care (at home or institutional), and hospice care. Many of these services are offered as specialized care for a specific disease or condition such as dementia or Alzheimer's disease assisted living communities called memory care. As older adults and their families preference is shifting from nursing home admission to living in the community, a range of community based care programs have been developed. (1,2) The program of all-inclusive care for older adults (PACE) and home- and community- based services for older adults (HCBS) are 2 examples of programs that are implemented nationwide to provide care for older adults who are living at home.(1,3)The goal of these programs are to help participants to live in the community as long as it is medically, socially, and financially feasible. (4)

Home health care are skilled medical services that are offered to older adults in their home and can include physician visits, nursing or nursing aide visits, medications, physical therapy, and other services. Patients who are confined to home temporarily (e.g., indicated for a medical reason such as surgery) or permanently (e.g., disability, old age) are often use the home health services. On the other hand palliative care aims to relieve patient's pain and suffering in contrast to the medical services that aim to cure and treat a condition. Palliative services are offered to patients who are dying and so are in the last few months of their life. Hospice care is a kind of palliative care however palliative care is not limited to hospice, i.e., it can be offered at home according to the patients and caregiver preferences.

- ***Description of the USMM Corporation***

United States Medical Management, LLC (USMM) is a management services organization that provides home-based medical services to homebound patients through its Visiting Physicians Association (VPA). The USMM provides medical care to patients across 11 US states including Michigan, Ohio, Texas, Florida, Kansas, Virginia, Illinois, Kentucky, Missouri, Washington, and Wisconsin. There are more than

2

100 USMM offices in these 11 states; the headquarters of the company is located in Troy, Michigan. In December 2011, VPA, in conjunction with the Detroit Medical Center (DMC), was selected as a Pioneering Accountable Care Organization (ACO). This Pioneering ACO was one of only 32 selected from over 4,500 applications in 2012.

The USMM Corporation specializes in home-based health care for homebound older adults and other - mostly disabled- patients unable to access health care through traditional means. Homebound adults are defined as patients who are confined to home according to the criteria in Table 1.1. A patient needs to meet either first or second criteria to be homebound. (5)

Table 1. 1. Definition of homebound patient determined by the Centers for Medicare and Medicaid Services

First Criteria One of the following must be met:	Second Criteria Both of the following must be met:
1. Because of illness or injury, the individual needs the aid of supportive devices such as crutches, canes, wheelchairs, and walkers; the use of special transportation; or the assistance of another person to leave their place of residence.	1. There must exist a normal inability to leave home.
2. Have a condition such that leaving his or her home is medically contraindicated.	2. Leaving home must require a considerable and taxing effort.

For example, a patient who is blind or old senile and need the assistance from another person to leave home, or a patient who recently had surgery and their actions are restricted to specified and limited activities by their physician are considered homebound. Also, a patient with psychiatric illness that is of such a nature that it would not be considered safe for the patient to leave home unattended is another example of a homebound patient. USMM provides comprehensive clinical management, administrative and support services, and has specific expertise in physician house call medicine. The USMM providers include physicians, nurse practitioners, and other allied health professionals that assist in the provision of home-based primary care. These providers include clinical educators as well as personnel from

certified home health agencies, hospice services, and durable medical equipment companies. USMM also owns several health properties and organizations, including hospices and home health agencies.

USMM maintains a large and rich clinical dataset on its population that is drawn from the electronic medical record (EMR) system named APRIMA. This EMR data includes information on demographics, socioeconomic (i.e., living alone, smoking, and insurance), functional status, comorbidities, laboratory tests, and utilization data. The EMR data is collected by USMM clinical staff, including physicians, nurse practitioners, and clinical educators. Another database, named Status Scope, also supplements APRIMA data. It contains supplemental medical information collected by allied health professionals or their assistants during regular home visits. For example the 'surprise question' and 'living alone' are variables that are collected in the Status Scope data. These databases, therefore, contain extensive clinical details from the home visits that USMM patients receive.

In addition to the USMM APRIMA (EMR) and Status Scope (supplemental clinical) data, claims data were also available through a third-party corporation, E-solution, (6) which provides processed claims data from the Centers for Medicare and Medicaid Services (CMS). The processed claims data contained limited information on only 5 types of events: death (date of death), hospice utilization (first and last dates of hospice services specified in 12-week intervals), home health (HH) utilization (first and last dates of HH services specified in 8-week intervals), the most recent hospitalization (admission and discharge dates), and the hospitalization prior to the most recent one. Dates of death and hospice utilization were used as the outcomes of interest in this study.

- ***USMM patient population***

The UMSS patient population is unique in terms of its demographics and functional characteristics. In 2015, more than 50,000 patients in the 11 states received services from USMM. This population is older (mean age of 71 and median of 73 years old; 86% are 65 years and older) and has a more complex comorbidity profile compared to typical Medicare populations. The prevalence of common

comorbidities such as hypertension (81%), hyperlipidemia (50%), chronic kidney diseases (40%), and diabetes (34%) in this cohort are all much higher than the US population of age ≥65 (Table 2.3). (7) The functional status of this population is also different from the typical geriatric population; functional status variables such as Karnofsky Performance Scale (KPS) and Timed Up and Go (TUG) indicate that the USMM population has more severe impaired function and need for assistance and special care. The higher prevalence of comorbidities and impaired functional status is explained by the fact that USMM patients' population are homebound by definition. The CMS criteria (Table 1.1) for a patient to be eligible for home health services includes but not limited to: be confined to the home, need skilled services, and be under the care of a physician. (1)

The USMM population, because of its old age, multiple comorbidities, and significant functional impairments has high levels of vulnerability. The unique and high-risk spectrum of the USMM patient population implies the need to develop a tailored risk stratification tools to effectively and efficiently manage their care and maximize their health outcomes (such as mortality, hospice admission, patient/caregiver satisfaction, and symptom management). In this thesis the two outcomes of interest are 1-year mortality and 1-year hospice admission

Further details of USMM population characteristics and the variables that are used for the model development are presented in chapter 2.

- o Risk stratification approaches proposed by USMM providers

USMM providers proposed two approaches for risk stratification: the surprise question (SQ) and a 3-level risk stratification approach. The surprise question is a simple question answered by the provider, "Would you be surprised if this patient died in the next 6-12 months?". (8) The answer to the SQ is used to find high-risk patients (i.e., where the answer equals no). The second proposed approach is a decision tree that categorizes patients into three risk levels (high, intermediate, or low) based on five variables:

SQ, albumin, a recent fall, hospitalization, or ER visit since their last USMM visit. These two RS approaches are simple and easy to use, but the performance of them in the prediction of the adverse outcomes has not been assessed. The USMM intention for conducting this study was to develop a more refined statistical approach to improve its RS process.

- ***Importance of Risk Stratification***

Risk stratification is a process of using observable/measurable characteristics to predict the risk of an event. It can help to classify a cohort to different levels of risk and then provide them with appropriate care. For example, a patient at high risk of death should prompt early referrals for palliative care or hospice, whereas an old patient at low risk of death should be considered for services – such as home health or other community programs– designed to maintain and improve their functional abilities, and their physical and mental health status.

Similar to many developed countries, the US population is aging faster than any other time in history. (1,2) People are living longer and experiencing more comorbidities. The number of patients with multiple-conditions has significantly increased in the past few decades. (3,4) Chronic diseases often require long-term health care and result in frequent utilization of services. The combination of the aging population and higher prevalence of multi-morbidity in older adults imposes a considerable burden of increased health care expenditure on governments, especially in developed countries. (5–8) The increasing need for health care services and limited resources has brought about an essential need to identify patients who have the most need for different types of services and to allocate services to those who will benefit the most. Risk stratification methods are commonly used for this purpose. Many studies have developed and evaluated risk stratification approaches in different cohorts of older patients, for example patients with atrial fibrillation, syncope, older adults discharged from emergency department, and patients with acute coronary syndrome. These studies often illustrated that the performance of the developed risk stratification model was superior to the prior approaches. (16–21)

6

As population ages, RS (i.e., prognostic) models are becoming increasingly important in clinical decision making. (22) Clinicians, researchers, and policymakers are using prognostic models to make decisions about preventive services or treatment strategies. Mortality based prognostic models are used to influence decisions for screening procedures in the population. An example of these decision making tools is the ePrognosis calculator by UCSF that serves as a repository of published geriatric prognostic indices where clinicians can obtain estimates of their patients' prognosis. For example one can get evidence-based suggestion for cancer screening based on a set of questions such as patient's demographics, comorbidities, functional status, and mental health. (23) Also, the prediction of mortality impacts the type and intensity of treatments to be offered to older patients. For example, the use of screening tests such as colonoscopy and mammography, or intensity of therapy for diabetes mellitus in older adults can be completely different depending on the prognosis. (24–29) Likewise, intensive treatment for diabetes for the prevention of long-term complications may not have any benefit or may even cause harm in a patient with <12 months life expectancy. Patients with limited life expectancy may benefit more from palliative care to ease their symptoms and to improve their end of life experience when time to benefit from the screening test or intensive treatment exceeds their life expectancy.

The overall goal of risk stratification for clinical populations like those served by USMM is to accurately predict adverse outcomes such as mortality and medical service utilization which then allows delivering more appropriate levels of clinical services to patients with different risk levels and to align services with the patients' needs and priorities. These services can include a change in medications, nutritional support, additional home visit, offering palliative care and advanced care planning, or hospice referral. Different prognostic tools have been developed to identify high risk patients for palliative care, for example Palliative Performance Scale (30,31) and Palliative Prognostic Score (32). The palliative care tools were summarized in a publication by the National Hospice and Palliative Care Organization. (33) Hospice eligibility criteria was developed by CMS. Additionally there are different disease-specific

hospice admission criteria and guidelines for cancer, cardiac disease, pulmonary disease, dementia, etc. (34,35) Another example of the risk stratification tool for clinical population is the PRIME REGISTRY which is a platform for clinical data registry with tools for risk stratification and care planning for family medicine physicians in addition to evaluation of practice performance. (36,37)

The purpose of this dissertation is to develop a risk prediction model specific to the USMM population, which is characterized by large numbers of patients with advanced age, multi-morbidity, and functional impairment. The ultimate goal of this risk stratification approach is to improve the quality and efficiency of home-based medical services, which also can include the appropriate and timely referral to other care settings, including nursing home and hospice.

- ***Overview of population-based disease management***

Health care organizations are working to change their cost structure and to improve their outcomes. The fact that 20% of the patients with chronic conditions are responsible for about 80% of health care expenditure has brought to attention the need for improvements in the disease management of such populations. Disease management is a system of coordinated healthcare interventions for populations with specific conditions. It emphasizes prevention of exacerbations and complications of the condition, evaluation of the clinical and economic outcomes, and developing a case management plan. (38) Population-based disease management is becoming today's optimal practice pattern and is replacing the former approach of episode-based disease management. It means that instead of managing patients who are seeking treatment at a given time, all people in a target population (e.g., insurance enrollee with a particular disease) are considered targets for case management interventions designed to prevent the complications of the disease and unnecessary medical utilization. By identifying high-risk, high-cost patients, this approach can result in more timely delivery of appropriate interventions in a cost-effective manner that has the potential to save money. (33) The fundamental step in population diseases management is risk stratification in order to identify the high-risk, high-cost patients. (40) For

example, Lavery et al., evaluated a risk stratification approach for population based disease management of diabetes mellitus. (41) Also, Haas et al. in a study used several risk stratification instruments in predicting health care utilization among all adult patients in a primary care practice. (42) Although population health management for older adults who often have multiple chronic conditions is more complex than the disease-specific population health management. Tkatch et al. reviewed the literature for the population health management for older adults and found that interventions to promote health among target populations tend to be disease specific rather than based on a global concept of older adults' health. (43)

- ***Current guideline for care management of geriatric population with multi-morbidity***

The American Geriatrics Society (AGS) Expert Panel has developed guidelines for the care of older adults with multi-morbidity.(44) Boyd et al. then developed a framework of actions to translate the AGS guideline to actions steps for decision making. (45) They provided three decisional actions and then action steps for each one. Table 1.2 contains the action steps. The first action requires the estimation of life expectancy and patients' health trajectory. The RS model in this study is going to serve as an instrument for estimation of the patient's life expectancy. Risk stratification is one of the many action steps needed for improvement in the health care for older adults with multi-morbidity. The risk stratification should be used and aligned in accordance with other action steps such as identifying patient's priorities and communicating the information between clinicians, caregivers and patient.

Table 1. 2. Conceptual framework for the care of older adults with multiple chronic conditions

MCC ACTION: IDENTIFY AND COMMUNICATE PATIENTS' HEALTH PRIORITIES AND HEALTH TRAJECTORY
• Identify and communicate patients' health priorities
    o Use a validated approach to identifying patients' health priorities
    o Transmit patients' health priorities
• Assess and communicate patients' health trajectory
    o Estimate life expectancy, trajectory, and lag time (time horizon) to benefit
    o Determine patients' readiness to discuss their trajectory or prognosis
    o Assess patients' perceptions of their prognosis and trajectory

MCC ACTION: STOP, START, OR CONTINUE CARE BASED ON HEALTH PRIORITIES, POTENTIAL BENEFIT VS HARM AND BURDEN, AND HEALTH TRAJECTORY
• Acknowledge uncertainty and variable health priorities in decision making and communication
• Stop or do not start medications for which harm or burden may outweigh benefit
    o Stop medications deemed inappropriate in older adults
    o Avoid medication cascades
    o Perform serial trials if treatments may be contributing to bothersome symptoms
    o Discontinue treatments no longer indicated or needed
    o Review and adjust self-management tasks
• Consider whether the patient has advanced illness or limited life expectancy that affects benefits and harms of treatments
    o Consider health trajectory and time to benefit for preventive interventions
    o Explain cessation of screening and prevention as a shift in priorities and use positive messaging

MCC ACTION: ALIGN DECISIONS AND CARE AMONG PATIENTS, CAREGIVERS, AND OTHER CLINICIANS WITH PATIENTS' HEALTH PRIORITIES AND HEALTH TRAJECTORY
• Affirm shared understanding of patients' health priorities and the information that informs decision making
    o Agree on the factors and information that will inform decision making and care
    o Encourage patients and family/caregivers to participate in decision making
• Align decisions when patient and clinician have different perspectives
    o Link decision to something meaningful to the patient
    o Ensure that patients' health outcome goals are consistent with their healthcare preferences
    o Identify and change bothersome aspects of treatment
    o Accept patients' decisions
• Align decisions when clinicians have different perspectives or recommendations
    o Focus discussion on patients' health priorities, not only on diseases
    o Acknowledge absence of one "right answer" for patients with MCCs
    o Use collaborative negotiation to arrive at shared recommendations

MCC, multiple chronic condition. Table adapted from the 'Decision Making for Older Adults with Multiple Chronic Conditions: Executive Summary for the American Geriatrics Society Guiding Principles on the Care of Older Adults With Multi-morbidity' (45).

This conceptual framework is to illustrate how RS is used to inform clinical care and in turn disease

management at the population level.

- ***Overview of literature relevant to the RS in community living older adults***

To find the previous studies related to the subject of this thesis, we searched Pubmed and google scholar for risk stratification in older adults for mortality and also for hospice. The results were reviewed for the study population and the outcomes; the relevant studies were also reviewed for forward and backward reference searching. The results of literature review are summarized in this section. Previous studies have developed prognostic models in different geriatric populations, such as hospitalized older adults(46,47), or nursing home patients.(13) Other studies have investigated the effect of specific comorbidities or functional status on mortality in elders.(49–51) These studies are different from our homebound study population since their populations are not community-living, or are limited to those with a specific condition, or evaluate only a specific predictor in association with the outcomes. There are fewer studies that developed a prognostic index for the community-living population regardless of a specific disease or chronic condition. Table 1.3 summarizes nine relevant studies that utilized prediction models to develop a prognostic index for mortality outcome in the community-living older adults. These study populations are mostly similar to our data population; although differ in one critical feature; none of them were described as homebound, while USMM patients are all confined to home.

Yourman et al., in a systematic review for prognostic models in older adults. Their study was the main paper that contributed to the Table 1.3 on the literature review. Yourman et al. reported on 16 studies, of which only six were performed in community-living older adults. (13) The other ten prognostic indices were developed in institutionalized populations, often nursing home residents. These six models from Yourman's review, along with other applicable studies, are summarized in Table 1.3. These models are discussed in more detail in the next three chapters when the results are compared to my findings.

Overall these previous models have been built in community-living older adults with different levels of multi-morbidity and functional impairment. The investigators used different databases for their study,

11

including Medicare administrative data (52) , population health surveys (53–55) , retrospective chart review from VA hospital patients (56) , or epidemiologic cohorts (49,57,58) . Therefore they might include the oldest-old adults (49) or a much younger cohort of elderly (50 years or older). (16) They can be nursing home eligible population that are living at home, like the study population in Carey et al. research.(57)

The mortality rates among these study populations ranged from 7.5% a year in Gagne study conducted in a cohort of Medicare beneficiaries who enrolled for the drug coverage programs, to 26% a year in Fischer study conducted through retrospective chart review for all patients who were admitted and discharged from the Denver Veteran's Administration Medical Center (DVAMC). (52,56) Han et al. reported a 6-month mortality rate of 15% (a grossly estimated one-year mortality of 30%) in their study population which is higher than the 2% mortality in the total Medicare Health Outcome Survey population (MHOS).(53) The reason is that they only included MHOS participants with declining health (i.e. patients who reported their health "much worse" compared to their last year health). The investigators constructed a prognostic index using regression coefficients of the multivariable models (Table 1.3). Except for the two studies by Carey and Fried (57,58), which modeled their data using Cox proportional hazard model, other studies used logistic regression models to develop the prognostic models. The study conducted by Carey in 2008 used the members of the Program of All-inclusive Care for the Elderly (PACE) which is probably the most similar study population to the USMM patient population. They were older frail adults, eligible for nursing home, but still living in their homes. PACE is a Medicare program for adults aged 55 and older who are living with disabilities and need a nursing home level of care but can safely continue to live in the community. PACE services can include home care if needed but there is not necessarily a home-based physician visit and health services. The USMM population was also vulnerable and includes frail older adults with underlying conditions that made them homebound. However, these two population had a critical difference, which is their mortality

rates. The USMM population had a much higher mortality rate (one-year 32%) than Carey's study population (one-year 13%). Indeed, the mortality rate in the Carey's study is lower than the expected mortality rate in a population of older adults who are eligible for nursing home. In the literature, the one-year mortality rate of nursing home residents has been reported between 17.4% and 35.0%. (59–62) The lower rate of mortality in the PACE patient population studied in Carey's paper may be explained by the definition of the PACE eligibility criteria. Adults with age 55 and older who are eligible for nursing home care are participants of PACE programs; therefore patients at relatively younger age (i.e. 55 to 65) who need long-term care (e.g. due to a disability) may have a longer life expectancy which contributes to the lower overall mortality rate in the study cohort.

In summary, comparing the studies in Table 1.3, the most important observation is that although all of the studies are among the community-living older adults, but the study populations are extensively heterogeneous. The heterogeneity can be best seen in the mortality rate of different studies. Consequently, these studies are not really comparable to the USMM patient population.

Table 1. 3. Summary of previous studies that developed a prognostic index for use in community-living older adult populations

Study First author Date of publication	Study population	Country and Time interval	Outcome and Predictors
Development of a Prognostic Model for Six-Month Mortality in Older Adults With Declining Health (PROMPT) **Paul K.J. Han** (53) 2012	-N=21,870 -Medicare beneficiaries from the Medicare Health Outcome Survey (MHOS), an annual nationwide survey by CMS -Medicare beneficiaries randomly sampled each year, aged over 65, with self-reported declining health in the past year -Institutionalized and disable beneficiaries are included -6-month mortality of 15%	-US -MHOS surveys from 1998-2000, 1999-2001, 2000-2002, 2001-2003	Outcome: 6-month mortality Predictors (11): age, gender, cancer, CHF, COPD, Smoking status, proxy status, ADLs, General health perceptions, social functioning, energy/fatigue
A combined comorbidity score predicted mortality in elderly patients better than existing scores **Joshua J. Gagne** (52) 2011	-N=120679 derivation -Medicare beneficiaries who had complete drug coverage through the Pharmacy Assistance Contract for Elderly (PACE) that provides medications at minimal expense to low-income elderly -1-year mortality of 8.9% -N= 123855 validation -Medicare enrollees who had complete drug coverage through the Pharmacy Assistance for the Aged and Disabled (PAAD) -1-year mortality of 7.5%	-US -Jan 2004- Dec 2005	Outcome: 1-year mortality Predictors (20): metastatic cancer, CHF, dementia, renal failure, weight loss, hemiplegia, alcohol abuse, any tumor, cardiac arrhythmias, chronic pulmonary diseases, coagulopathy, complicated diabetes, anemias, fluid and electrolyte disorders, liver disease, peripheral vascular disorder, psychosis, pulmonary circulation disorders, HIV/AIDS, hypertension
Index to Predict 5-Year Mortality of Community-Dwelling Adults Aged 65 and Older Using Data from the National Health Interview Survey **Schonberg** (55) 2009	-N=24115 -Non-institutionalized adults aged >65 who responded to the 1997-2000 National Health Interview Survey (NHIS) with follow up from the National Death Index (NDI) -5-year mortality of 17% (estimated 1-year mortality of 3.4%)	-US -1997-2002	Outcome: 5-year mortality Predictors (11): age, gender, BMI, perceived health, emphysema, cancer, diabetes, dependency in IADLs, difficulty walking, smoking, past year hospitalization

Table 1. 3. (cont'd)			
Prediction of Mortality in Community-Living Frail Elderly People with Long-Term Care Needs **Elise C. Carey** (57) 2008	-N=3,899 -Community-based, frail, chronically ill older adults who are eligible for nursing home -A cohort of community-living participants enrolled in the Program of All-Inclusive Care for the elderly (PACE), (63)the program operates under Medicare and Medicaid waiver to deliver services to the elderly who are certified by the state's Medicaid staff as eligible for nursing home -1-year mortality of 13% and 3-year mortality of 36%	-US -1988- 1996 (participants enrolled in PACE)	Outcome: Time-to-death from the time of initial enrollment in PACE (3-year follow up) predictors (8): age, gender, dependence in the 2 ADL (toileting and dressing), CHF, COPD, Cancer, Renal failure
Screening of Older Community-Dwelling People at Risk for Death and Hospitalization: The Assistenza Socio-Sanitaria in Italia Project **Giampiero Mazzaglia** (64) 2007	-N=5396 -Community-dwelling, aged ≥65, randomly sampled from the roster of 98 Primary Care Physicians -15-month mortality of 4.7% in derivation and 3.9% in validation cohorts	-Italy, Florence -Jan 2003- Mar 2004	Outcome: 15-months mortality Predictors (5): age, gender, hospitalization in the past 6 months, use of ≥5 medications, score from a 7-item questionnaire (need help for ADLs, need help for IADLs, poor vision, poor hearing, self-perceived inadequacy of income, absence of home care services, weight loss>3 kg)
A Practical Tool to Identify Patients Who May Benefit from a Palliative Approach: The CARING Criteria **Stacy M. Fischer** (56) 2006	-N=895 -All patients admitted to general medical wards or medical ICU of the Denver Veterans' Administration Medical Center (DVAMC) -1-year mortality of 26% (from the index hospitalization)	-US (Colorado, Denver) -Feb- Jun 1999 (retrospective chart review)	Outcome: one-year mortality Predictors (5): cancer, admissions≥2, residence in a nursing home, intensive care unit admit with multi-organ failure, ≥two non-cancer hospice guidelines

15

Table 1. 3. (cont'd)			
Development and validation of a prognostic index for 4-year mortality in older adults **Sei J. Lee** (54) 2006	-N=19710 -community-dwelling adults aged >50, participants of the 1998 wave of the Health and Retirement Survey(HRS); data primarily collected through a telephone interview with a participation rate of 81% -4-year mortality of 12% in derivation and 13% in validation cohorts	-US -1998-2002	Outcome: 4-year mortality Predictors (12): age, gender, diabetes, cancer, lung disease, heart failure, smoking, body mass index, difficulty with- bathing, walking several blocks, managing money, pushing large objects
Development and Validation of a Functional Morbidity Index to Predict Mortality in Community-dwelling Elders **Elise C. Carey** (49) 2004	-N=7393 -Community-dwelling, age ≥70, Participants of the Asset and Health Dynamics Among the Oldest Old (AHEAD) study, a prospective national study that sampled community-dwelling U.S. elders age ≥70 -2-year mortality of 10% in derivation and 12% in validation cohorts	-US -AHEAD Participants who were interviewed in 1993	Outcome: 2-year mortality Predictors (6): age, gender, dependence in bathing, dependence in shopping, difficulty walking several blocks, difficulty pulling or pushing heavy objects
Risk factors for 5-year mortality in older adults **Linda P. Fried** (58) 1998	-N=5886 derivation -Participants of the Cardiovascular Health Study (CHS) aged ≥65, a prospective cohort of randomly sampled from age-stratified from the Health Care Financing Administration (HCFA) Medicare enrollment lists -5-year mortality of 12%	-US (4counties: Sacramento, CA; Washington, MD; Forsyth, NC; Allegheny, PA - Derivation 1989- 1990 Validation 1992-1993	Outcome: 5-year mortality Predictors: age, gender, income, weight, exercise, smoking, systolic blood pressure, diuretic use, fasting blood sugar, albumin, creatinine, forced vital capacity, aortic stenosis, EF, ECG abnormality, carotid artery stenosis, CHF, difficulty in IADLs, low cognitive function

- ***Statistical analysis of prediction models***

Prediction models are actually problems of either estimation or hypothesis testing. For example the question "What is the risk of a patient dying in the next 30 days?" is an estimation problem that needs a prediction model to estimate the probability of death, while the question "Is gender a predictor of a certain complication after surgery?" or "What are the important predictors of hospitalization in older adults?" is a problem of hypothesis testing. Statistical models can address both types of questions. There are three main classes of statistical models used in prediction: regression, classification, and neural network. Regression models are the most commonly used models for prediction. (24) Different regression models are used in the literature; the two commonly used models are logistic regression and time-to-event (Cox regression) analysis. Logistic regression is the most commonly used model for prognostic models; it is used when the outcome is binary (Yes/No) like death, hospital admission, ER visit, or occurrence of a complication. Most of the studies in Table 1.3 utilized logistic regression analysis to build the prognostic model. Fewer studies used the Cox proportional hazard analysis to model time-to-event as the outcome. The two studies by Carey and Fried in are examples of the Cox model utilization in the development of a prediction model. (57,58)

In the past few decades, due to the increasing size and complexity of biological data, the limits of the traditional modelling approaches have begun to be reached, and there is a need for innovative statistical analysis for the ever-growing data. (39) Advanced methods such as machine learning algorithms that allow detecting pattern and making predictions in big data with complex relationships are becoming an increasingly important method for use in the development of prediction models. (40–42) Random forest is one of the machine learning algorithms that has been occasionally used in biomedical researches. (65,68,69) The machine learning algorithms have been shown to outperform the traditional statistical models in prediction, (70–77) however some studies showed no difference in the performance of the prediction models between the traditional models and machine learning methods. (78)

In this dissertation, prediction models are developed for two primary outcomes (death and hospice admission) using both traditional statistical models and machine learning algorithm. First, a logistic model is developed for each of the two binary outcomes. Second, a random forest algorithm is used for the same outcomes to have comparable models. Third, a Cox PH model is developed to account for the time-to-event for both outcomes. The results of developed models are compared to the USMM proposed RF approaches. The performance of the three models is compared to each other to find the best model in terms of predictive performance. The ultimate goal of this study is to find the best model that can be integrated into the USMM database. The RS process is a necessary step in older adult care according to the framework of actions for the care of older adults with multi-morbidity (Table 1.2).The optimal process of using the RS model in the USMM patient population would be that for each patient a predicted probability is calculated from the base model, and then a risk level will be assigned to the patients based on their probability of death (or hospice admission). The high-risk patients would be flagged and brought to the attention of the provider team for appropriate and timely intervention. This intervention can include a range of services such as a change in medications, nutritional support, additional home visit, hospice referral, or offering palliative care and advanced care planning. Lower risk patients can be targeted for other levels of services according to the USMM policies and care plans, such as screening for prevention of morbidities, rehabilitation and other programs to preserve and enhance the functional ability, and lifestyle modifications to improve physical and mental health.

To assign the risk levels based on the predicted probabilities, the threshold for different levels of risk must be decided. An arbitrary cut point of the highest 20% of predictions is used in Chapter two to calculate the performance of the model. This cut point must be determined based on the number of patients in the system and the resources that USMM can allocate for services to the different levels of risk. A more liberal threshold for identification of the high-risk patients (for example top 30% of the predicted probability) results in a larger number of high-risk patients who need to be evaluated for an

intervention. Consequently, more human and financial resources are required to take care of these additional cases.

On the other hand, more stringent threshold, while reduces the need for resources, may result in more false-negative cases (i.e., those who are truly at high risk of death or hospice are classified as low-risk), which in turn can cost in missing adverse events in the truly high-risk patient. To summarize, there is a tradeoff in determining the threshold for different risk levels. The cut points should be determined by the USMM providers based on their objectives of risk stratification and the available resources. More formal approaches for determining the optimal RS thresholds for an organization like USMM might involve cost-effectiveness analysis.

It is critical to remember that final decision making involves the patient and caregiver's priorities and preferences. Therefore the clinician's goal of care must be aligned with the patient's and caregiver's goals.(44)

o Overall Analysis plan

The USMM clinical database (APRIMA) and claims data will be used to construct a cohort of USMM patients who were FIRST registered with USMM in the calendar year 2015 and had at least one visit in that year. After data preparation and necessary recoding of variables, available potential predictor variables (including demographics, functional status, comorbidities, laboratory tests, and socioeconomic factors) that have less than 20% missing will be considered as predictors in the analyses. There will be two outcomes of interest, death, and hospice admission that will be identified based on the presence of a date of death or date of hospice in the claims data. For validation of the models, the dataset will be randomly divided into two subgroups named derivation and validation. Then three different statistical approaches will be applied to the derivation data to develop predictive models and to generate performance metrics used to compare the different models. Each model will be then validated using the

validation data set. The discrimination of the models which described the ability of the models to accurately distinguish those with and without outcome will be measured by the area under the ROC curve (AUC) will be used as the primary measure of prediction accuracy and model performance. Calibration plots as a measure of goodness of fit will be generated when is applicable. All the analysis will be done for both outcomes separately. Details of the model development and results are provided in chapters 2-4.

- ***Objectives***

The three objectives of this dissertation are:

1. To develop and validate multivariable logistic models for prediction of 12-month mortality and 12-month hospice admission among the USMM population of community-living homebound older adults. The models will be compared to the alternative risk stratification approaches used by USMM, including the surprise question (in isolation) and the existing USMM 3-level risk stratification method.

2. To develop and validate a random forest (RF) algorithm for prediction of 12-month mortality and hospice admission. The model performance will be evaluated compared to the logistic regression (LR) model from aims 1.

3. To develop and validate a multivariable failure time model (Cox proportional hazard) to model time-to-event for mortality and hospice admission separately. These models will also be compared to the logistic regression and random forest models developed in aims 1 and 2.

- ***Introduction***

The US population is aging faster than any other time in history. (1,2) Causes of mortality have shifted

from communicable infectious diseases to chronic conditions and their complications. Diseases that

used to be lethal now can be treated or managed for years. People are living longer; therefore the

prevalence of chronic diseases, cancers, and persons with multiple comorbidities has significantly

increased in the population.(11,12) Chronic diseases often require long-term health care, and frequent

utilization of services; consequently, health care costs are growing fast as the population is aging. The

combination of the aging population and higher prevalence of multi-comorbidity in older adults imposes

a considerable burden of increased health care expenditure on governments, especially in developed

countries.(7,13–15) About one-fifth of Medicare beneficiaries have five or more chronic conditions, and

two-thirds of Medicare expenditures are related to this group.(79) The increasing need for health care

services and limited resources has brought about an essential need to identify patients who have the

most need and to allocate services to those who will benefit the most. Risk stratification methods are

commonly used for this purpose. Using statistical methods, one can develop a risk stratification model

to predict the risk of an adverse event based on observed variables. The model then can be applied to

classify patients into different risk levels and to identify the most appropriate services for each level.

Risk stratification is playing an increasingly important role in public health and clinical care. Health care

organizations are working to change their cost structure and to improve their outcomes. The fact that 20

percent of the patients with chronic conditions are responsible for about 80 percent of health care

expenditure has brought to attention the need for improvements in the disease management of such

populations. Disease management is a system of coordinated healthcare interventions for populations

with specific conditions. It emphasizes prevention of exacerbation or complications of the condition,

evaluation of the clinical and economic outcomes, and developing a case management plan. (80,81,38)

Population-based disease management is becoming today's optimal practice pattern and is replacing the former approach of episode-based disease management. This means that instead of managing patients who are seeking treatment at a given time, all people in a target population (e.g., insurance enrollee with a particular disease) are considered targets for case management interventions designed to prevent the complications of the disease and unnecessary medical utilization. By identifying high-risk, high-cost patients, this approach results in more timely delivery of appropriate interventions in a cost-effective manner that has the potential to save money.(39) Risk stratification can help to identify high-risk, high-cost patients.

Prediction models are used by researchers, health care providers, and policymakers to predict patient outcomes such as mortality and health care utilizations.(82) Prognostic indices can be used to target different services appropriately to older patients. For example, prediction of mortality in a target population can identify patients at high risk of death with consideration of palliative care programs or advanced care planning. It also helps to prevent the allocation of resources to the services that are costly and not beneficial; for example, screening for slow-growing cancer in older adults with a high risk of 1-year mortality is not reasonable. Additionally hospice care can be offered to the terminally-ill patients in order to improve the quality of life for the patients and caregivers. According to Medicare criteria, a patient is eligible for hospice services, if determined to have a terminal illness (defined as having a prognosis of 6 months or less if the disease or illness runs its normal course).(35) Risk stratification can identify patients with limited life who are eligible to be evaluated for hospice services. A risk stratification approach that predicts probability of death for a group of patient can help to identify those at high risk of death in close future (e.g., 6 months) and so can help to identify the potential candidates for hospice services.

Objectives of the research are to develop alternative risk stratification models in a unique population of community-dwelling, home-bound older adults who receive home-based medical services from the United States Medical Management (USMM) Corporation. The outcomes of interest are mortality and hospice admission. Three different statistical approaches will be applied to develop predictive models:

- Chapter 2 (current chapter): Binary outcomes in a fixed time interval (i.e., one-year mortality and hospice admission) will be modeled using a logistic regression approach

- Chapter 3:Binary outcomes in a fixed time interval (i.e., one-year mortality and hospice admission) will be modeled using a random forest model

- Chapter 4: Time-to-event will be modeled using a Cox proportional hazard model

In this chapter, the first approach is presented, namely, logistic regression analysis. In the development of the prediction model, several variable selection methods including forward, backward, and stepwise selection are applied as well as more advanced variable selection methods, including Adaptive lasso and elastic net variable selection techniques. A conventional variable selection method is also used (called manual variable selection). To handle the missing data problem, a multiple imputation approach is applied and different variable selection methods are used to develop models using the imputed data. These models are compared by their discrimination ability indicated as c-statistic (AUC- area under the ROC curve). The models are also compared to the pre-existing risk stratification approaches that are already in-use by USMM providers.

The contribution of this research to the mortality risk stratification literature are: 1. the use of community-dwelling homebound older adults, 2. incorporation of advanced variable selection techniques, 3. implementation of multiple imputation technique to manage missing data, 4. prediction of hospice admission in addition to mortality.

The rest of this chapter is organized as follows: background and literature review, methods and materials, empirical results, discussion, and conclusion.

- ***Literature review***

As discussed in chapter one, most of the previous prognostic models have been developed in a specific setting such as nursing home, emergency department, or hospital. Other authors have developed models in populations of older adults with specific conditions such as cancer, chronic kidney diseases, and cardiovascular diseases. There are a limited number of studies that focuses on the risk stratification in the community-dwelling older population.

Yourman et al., in a systematic review of prognostic indices for older adults, evaluated the accuracy and generalizability of such indices.(48) They found 16 validated indices, but only six included community-dwelling patients. The rest of them used patients in a nursing home or hospitals. Table 1.3 in chapter one, summarizes these studies along with three other relevant studies. Following is a brief summary of the seven studies that used logistic regression in model development.(53)

Of the six studies in community-dwelling patients, the only model that evaluated 1-year mortality was developed by Gagne et al. The model consisted of 20 comorbidities and resulted in a c-statistic of 0.788 (95% cl, 0.786-0.791). However, their study population included both community-dwelling and nursing home residence patients (9%). Mortality rate of this population was 7.5% a year. Their index also showed better discrimination for 30-day and 180-day than 1-year mortality.(52)

Carey et al., in two separate studies developed prognostic indices for 2-year and 3-year mortality in older adults. In their first study, an index was developed for 2-year mortality in community-living frail older adults. They included variables age and sex plus 16 functional variables in the predictive model. The final index comprised of 6 variables including age, sex, dependence in bathing, dependence in shopping, difficulty walking several blocks, and difficulty pulling/pushing heavy objects. The prognostic index had discrimination (C-statistic) of 0.74. Mortality rate in this population was 12% over 2 years.(49)

The second study was to develop a prognostic model for 3-year mortality in a cohort of nursing home eligible older adults. The study population consisted of the participants in the Program of All-Inclusive

Care for the Elderly (PACE). The PACE program provides comprehensive medical and social services to frail, community-dwelling older adults, most of them are dually eligible for Medicare and Medicaid. The program enables most of the participants to remain in the community rather than receiving nursing home care.(3) The study population were chronically ill elderly who met the criteria for nursing home placement. Using a Cox model for time to death, Carey et al., defined a score made of variables age, sex, dependency on toileting and dressing, and four comorbidities. The index showed a c-statistic of 0.69 for 3-year mortality in the validation data. Mortality rate in this population was 13% over a year.(57)

Mazzaglia et al., developed a prognostic index for 15-month mortality and hospitalization in a cohort of community-dwelling older adults in Italy. This index was developed to be used mainly by primary care physicians and consisted of age, sex, previous hospitalization, dependency on basic ADLs and IADLs, poor vision, poor hearing, use of home health services, and inadequate income. This index stratified elders to 4 risk groups; the c-statistic of this model was 0.75 for 15-month mortality. Mortality rate in this cohort was 4% over a 15 months period. (64)

Lee et al., proposed a prognostic index for 4-year mortality among older adults using 12 predictor variables. Their study population was adults older than 50 years who participated in the 1998 wave of Health and Retirement Study. The Participation rate in their study was 81%. Significant predictors in this index included age, sex, six comorbidities, and four functional status indicators such as walking several blocks and managing money. The discrimination of this index was 0.82 in the validation data. Because they included patients as young as 50 years old who were generally healthy, the authors suggested that the optimal model for an older and sicker population would include other predictor variables. Mortality rate in this cohort was 12% over 4 years. (54)

Han et al., developed a prognostic model using 11 predictors of 6-month mortality among community-living older adults with a self-reported decline in health. The outcome of 6-month mortality was chosen because the 6-month prognosis is essential in hospice referral. They used the data from the Medicare

Health Outcome Survey (MHOS), and they did not exclude institutionalized and disabled Medicare

beneficiaries. Significant predictors included age, sex, smoking status, any cancer, congestive heart

failure, COPD, ADLs, proxy status, and health-related quality of life (general health perception, social

functioning, and energy/fatigue). Their model had a c-statistic of 0.75. Mortality rate in this population

was 15% over 6 months (grossly estimated one-year mortality of 30%) – much greater than these other

studies. This mortality rate was also much greater than the 2% mortality rate of the total MHOS

population. An explanation is the selection of this particular study population which included only MHOS

participants with declining health (i.e. patients who reported their health "much worse" compared to

their last year health).(53)

Schonberg et al., studied predictors of 5-year mortality among a population aged 65 and older who

participated in National Health Interview Survey and responded to annual follow up surveys for five

years from 1997 to 2002 (74% mean participation rate). The 5-year mortality rate in this population was

17% during the study period. They used a multivariable Cox proportional hazard model with 11

predictors including age, sex, smoking status, BMI (<25 kg/m^2), dependence in IADL, difficulty walking

several blocks, general health perception, past year hospitalization, and three comorbidities. The model

had a c-statistic of 0.75. Mortality rate was 17% over 5 years. (55)

Review of the studies that developed a prognostic model in community-living older adults revealed that

the USMM population is different from the other study populations presented in table 1.3. The

difference can be seen in average age, the number of comorbidities, functional status, and the setting

(institution, nursing home, community). USMM patients are homebound, also they are older and have

higher rates of comorbidities compared to the previous study populations. But, most importantly the

USMM population mortality rate (32% over 12 months) is substantially higher than these other study

populations – where the estimated annual mortality rates were in the range of 4-8% but varied from as

low as 3.0% (54,55) to as high 30% (53). Moreover, selected predictors used in the previous prognostic

models, were not found in the USMM database. For example itemized ADL and IADL information, past

year hospitalization, income and BMI are variables that were not available in this dataset. These

differences made the pre-existing prognostic models not to be appropriate for the USMM population.

This study aims to develop a model suitable for this population and other similar older cohorts.

Variable selection is the basis of developing a prediction model. Variable selection can be made by

including and excluding rules for predictor variables, or by built-in automated methods in the statistical

software. Most of the studies cited above used logistic regression to model the outcome and identify

the important predictors of it. Carey and Schonberg made use of a Cox proportional hazard model as

well. None of these studies mentioned any application of any particular variable selection method;

instead they included all significant (usually P<.05) variables in a final MV model. Statistical software

offers different options for variable selection methods as part of the model development process

including stepwise, backward, forward methods, as well as more advanced methods such as lasso,

adaptive lasso, elastic net, and ridge regression. Although the newer selection methods such as adaptive

lasso and elastic net are not directly available in SAS for binary outcomes (Logistic or HPLOGISTIC

procedures), there are methodological papers that explain the use of the GLMSELECT procedure to

make use of these methods. Lund and Cohen suggested that although GLMSELECT procedure fits an

ordinary regression model, it can be used to select a good set of predictors for a logistic model. (83–85)

Missing data is a persistent problem in epidemiological studies. Some of the common reasons for

missing data are patients' refusal to answer, lack of knowledge, loss of contact in longitudinal studies

(due to death or relocation), and failure of routine documentation in the EMR by clinical staff. We could

not find a specific approach for management of missing data in any of the nine previously mentioned

studies that developed prognostic indices in community-living older adults. When the missing data was

described, either the observations with partly missing data were excluded from the analysis, or

missingness was included as a dummy variable in the analysis.(53,57) In this analysis, missing data is

observed in about one-third of the cohort; therefore the missing data problem is addressed in this chapter. One of the commonly known approaches to the missing data problem is multiple imputation (MI). The SAS procedure, multiple imputation uses the assumption of missing at random (MAR) for the missing data, however the use of PROC MI can be extended to the MNAR conditions. (86) Although, it is impossible by definition, to distinguish between MAR and MNAR mechanisms.(87) To build a prediction model, variable selection is applied to a complete case data when there is no missing observation. However, excluding the cases with partly missing data can induce bias into the results. Wood et al., proposed four different approaches for variable selection in the imputed data. (88) The first method involves developing the model in the complete case data and using the same variables in the imputed data. The second method is to develop a single model in the first set of imputed data. The third method is to develop separate models in each imputed dataset and then combined the selected variables from all models to form the final model. The fourth method is to use stacked imputed datasets with weighted regression. The first and third methods are used in this study for variable selection in the imputed data.

The most common methods for evaluating the accuracy of a predictive model for binary outcomes are discrimination, which is measured by area under the ROC (AUC), and calibration. The AUC (also called the concordance or C-statistic) is the most commonly used measure of discrimination of a model. It indicates how good the model classifies those with and without the outcome of interest. For a binary outcome, ROC is a plot of sensitivity against 1- specificity for all the consecutive cutoffs in the probability of an outcome.(89) AUC values can be roughly interpreted as excellent (AUC above 0.80), good (between 0.70 and 0.80), and weak (between 0.50 and 0.70). Calibration compares the predicted and observed probability of the outcome in different risk groups. Calibration plots provide a qualitative visualization of the goodness of fit, while the Hosmer-Lemeshow is a statistical test of goodness of fit. The Brier score is another measure to evaluate the goodness of fit and performance of a predictive model.(90) Brier score is an equivalent of R-square when the outcome is binary. In this study, the AUC,

calibration plots, Hosmer-Lemeshow test, and Brier score are utilized to evaluate the model

performance.(90)

- ***Methods and materials***

  o <u>Data source</u>

I conducted this study utilizing the United States Medical Management (USMM) dataset. USMM is a

family of companies that provide home-based medical care to patients across 11 US states including

Michigan, Ohio, Texas, Florida, Kansas, Virginia, Illinois, Kentucky, Missouri, Washington, and Wisconsin.

USMM specializes in home-based health care for homebound elderly individuals and other patients with

complex medical issues. The USMM providers include physicians, nurse practitioners, clinical educators,

and people with other specialties. USMM also owns several health properties, including hospices and

home-health agencies. USMM maintains a database of all patients visited in their 100 offices across the

11 states. This database includes demographics, social, functional status, clinical, laboratory, and

utilization data. This database consists of the USMM electronic medical record named APRIMA in

addition to other data sources. The USMM clinical database for the calendar year 2015 was used for this

analysis. Claims data were also available through a third party corporation, E-solution, which provides

processed claims data from CMS. The processed claims data contained limited information on only 5

events including death (date of death), hospice utilization (first and last dates of hospice in 12-week

intervals), home-health utilization (first and last dates of HH services in 8-week intervals), most recent

hospitalization (admission and discharge dates), and prior hospitalization (admission and discharge

dates). Dates of death and hospice enrollment were used as the outcomes of interest in this study.

Therefore the USMM EMR data and claims data together were used to define the study population.

o   Study population

The 2015 cohort was defined as all patients who had their first ever home-based medical visit between

January 1st and December 31st, 2015. The date of the first visit was recorded in the APRIMA EMR. The

data was then linked to the claims data, and those patients who did not have claims data available were

excluded. Patients with age <65 years were excluded. Table 2.1 contains the inclusion and exclusion

criteria for the patient population in this chapter.

Table 2. 1. Inclusion and exclusion criteria in this study patient population

Inclusion criteria
-   Register in the USMM system in the calendar year 2015
-   Have at least one visit between January 1st and Dec 31st, 2015
Exclusion criteria
-   Claims data not available
-   Age <65 years old
-   Followed up for less than 1 year

Since the purpose of this chapter is to analyze 1-year mortality and hospice admission, the cohort was

limited to the patients who had been followed up for at least 365 days or had the outcomes within a

year of their first USMM visit. Follow up time was determined by counting the days between the first

visit date and the date of the outcome (i.e., death or hospice admission), or the date of the last visit if

the outcomes did not occur. Figure 2.1 displays a flow diagram of the patient population in this study.

Figure 2. 1. Flow diagram of the study cohort
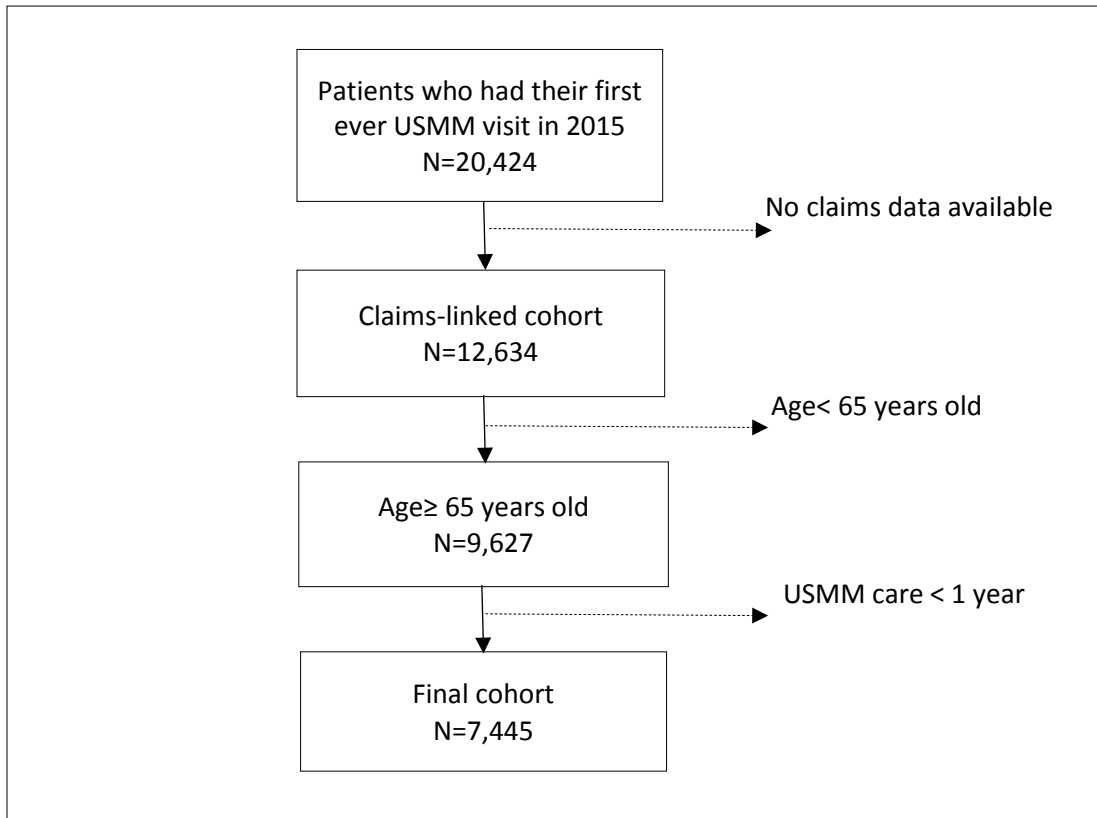
Among the 2182 patients who were excluded due to <12 months of USMM care, 88.5% (n=1932)

patients became inactive in the USMM database due to various reasons including patient opted out of

the program, relocation to nursing home, loss to follow-up (hospitalization, no response to phone call,

bad address), discharged by provider (e.g., due to not being homebound), patient moved, and other

reasons. These reasons for withdrawal are summarized in Table 2.2. Majority of these reasons were

related to patients' preference (i.e., 31% who opted out), such as patient changed the Primary Care

Physician (PCP), chose another house call program and refused the services. The rest of 11.5% (n=250)

were patients who did not have documented reasons for withdrawal from the USMM, but their total

registered time was less than 12 months (Table 2.2). Many of these patients (n=177, 71%) were visited

at the end of the year 2015 (i.e., December 2015) and their last recorded visit in the USMM database,

occurred before the December 2016. Thus the total documented time of care in the USMM system was

less than 12 months, although these patients were still active in the USMM database.

Table 2. 2. Patients with < 1-year care received from USMM (N=2182)

Inactive	N (%)
Became inactive in the USMM system	
Patient opt-out	686 (31.4%)
Nursing home admission	380 (17.4%)
Loss to follow up	206 (9.4%)
Provider excluded patient	204(9.3%)
Patient moved	202 (9.3%)
Missing reason	189 (8.9 %)
Insurance issues	65 (3.0%)
< 1-year documented care	250 (11.5%)
Total	**2182 (100%)**

o   Outcome and exposure

There are two outcomes of interest: mortality and hospice admission. One-year mortality was

determined if a date of death was recorded in the claims data within the 12 months of the first USMM

visit. Likewise, 1-year hospice admission was determined according to the recorded date of first hospice

service in the claims data. Claims data was processed data provided by E-solutions, a commercial

medical billing and claim processing company. (6) Claims data provided the dates of death, hospice,

and/or home-health services (8-weeks period), therefore the first date of earliest hospice service was

considered as the date of the outcome. If a date of death or hospice was not reported within a year

from the first visit date, then the case was counted as censored at one year. If the death occurred in

hospice, both outcomes (death and hospice admission) were analyzed as separate outcomes in each

respective analysis.

Variables with less than 20% missing observations were considered as exposure variables for the

analysis. This information was collected from the baseline visit for each patient. Table 2.3 in the results

section displays the frequency of missing data on each variable.

A total number of 41 potential predictor variables had <20% missing, including demographics and social factors: age, gender, race, insurance status representing if a patient has dual eligibility for both Medicaid and Medicare; life style factors: living alone, smoking; functional status factors: functional decline in ADLs, timed up and go (TUG), Karnofsky performance scale (KPS value); serum measures: serum albumin, cholesterol; and other factors: having a pressure ulcer, surprise question answer, number of medications, and number of lab test ordered by the provider. There are 24 medical history variables in the APRIMA EMR that documented if at the time of the current visit, the patient had an active diagnosis of the condition as defined by the CMS-Chronic Condition Warehouse (CCW).(91) These 24 variables reported as binary (yes, no) data and includes: history of hypothyroidism, asthma, atrial fibrillation, cataract, chronic kidney diseases, osteoporosis, hyperlipidemia, hypertension, anemia, breast cancer, colorectal cancer, benign prostatic hyperplasia, COPD, depression, diabetes, endometrial cancer, glaucoma, heart failure, hip/pelvic fracture, ischemic heart disease, lung cancer, prostate cancer, stroke/TIA, rheumatoid arthritis/osteoarthritis. Diagnosis count is a variable that counted the number of existed CCW variables for each patient. Another variable, cancer, was generated if a patient had any of the four different types of cancers listed in the CCW variables. History of Alzheimer's disease and acute MI were also among the CCW variables; however, the number of patients who had these conditions were too small to analyze. Thus they were dropped from consideration.

Three variables in this dataset represent the functional status of patients; functional decline in activities of daily living (ADLs), Timed Up and Go (TUG answer), and Karnofsky Performance Scale (KPS). These variables are documented by the visiting physician in the APRIMA and supplemented by Status Scope. The functional status variables and surprise question are defined in Table 2.3.

The decline in ADLs can be an indicator of developing frailty or other medical events that need attention for the timely prevention of adverse outcome. (92) Activities of daily living (ADL) include six daily activities: self-feeding, bathing, dressing, toileting, transferring, getting in/out of bed/chair. (93,94)

Instrumental Activities of Daily Living (IADLs) include activities such as shopping, housekeeping, keeping track of finances, and food preparation. Two variables in the APRIMA database indicate a decline in ADL and IADL compared to the last year. The visiting physician evaluates the functional status compared to the last visit (for ADLs) or the last Annual wellness Visit (IADLs). Unfortunately, the variable measuring decline in IADLs was excluded from the analysis due to the high rates of missing value.

Timed Up and Go test (TUG) is a simple test used to assess a person's mobility that includes both static and dynamic balance.(95) The test involves measuring the time that a person takes to rise from a standard arm chair, walk three meters at their normal pace, turn around, walk back to the chair and sit down again. It is reported in seconds and <30 seconds is considered normal. The results of this test were reported in three categories: ≤30 seconds, >30 seconds, and non-ambulatory.

Karnofsky Performance Status Scale (KPS) is another tool used to quantify patients' general well-being and functional status.(96) The score ranges from 100 to 0, where 100 is perfect health and function, and 0 is death. The score is usually reported in intervals of 10. A KPS score of 80-100 indicates the ability to carry on normal activity and to work. A score of 50-70 shows inability to work, but these patients are able to live at home and care for personal needs. A score of 40 and less indicates functional disability and inability to care for self. (36). Since only 0.4% of this population had a score of 80-100, we e re-coded KPS values into a binary variable using 40 as the threshold (i.e., ≤40 versus >= 50).

The Surprise question is a simple question answered by the provider, "Would you be surprised if this patient died in the next 6-12 months?" This question provides a valuable piece of information that has been shown in many different setting to be a strong predictor of mortality. (98,99) The predictive value of the surprise question has been evaluated explicitly in diseases such as cancer and kidney diseases. The value of the surprise question has not been well assessed in a general population of older age adults without specific diseases or conditions. A recent study has evaluated the performance of the SQ in prediction of two-year mortality in patients with serious illness from primary care clinics in Boston, MA.

The patients were screened by the primary care physicians (PCP) and enrolled in the study if they were eligible for the serious illness care program.(100) The goal of the study was to improve access to palliative care among the patients who are approaching the end of life. The performance of SQ in prediction of two-year mortality among the chronically ill, complex patients was measured by the Area under the curve and it was 0.74 when the question was asked from the primary care physicians. (100) The key features of these four measures are summarized in Table 2.3.

Table 2. 3. Definition and values of the functional status variables and surprise question

Variable	Definition	Values
ADL- decline	Functional decline in activities of daily living	Decline, improve, no change
TUG	Timed up and go is a measure of patients mobility and balance	<30 seconds, >30 seconds, non-ambulatory
KPS	Karnofsky performance scale quantifies patients' general well-being and functional status	Values range from 10-100 with lower values indicating worse functional status
Surprise question	Answer to the question "would you be surprised if this patient died in the next 6-12 months?"	Yes/ no

o   Statistical analysis

The statistical analyses for this paper was done using SAS software, version 9.4 (SAS Institute Inc., Cary, NC). The data were randomly split into two equal size cohorts, to create derivation and validation datasets. Logistic regression was applied to develop a prediction model in the derivation dataset. The model parameters then were applied to the validation cohort, and the predicted probability of the outcome was calculated for each patient.

Logistic regression model fits binary response and provides several variable selection methods to identify important predictor variables among many potential independent variables. Logistic regression is used to explain the effect of an explanatory variable x on the response Y.

$$\text{logit}\{\Pr(Y = 1|\,x)\} \; = \; \log\{\frac{\Pr(Y=1|\,x)}{1-\,\Pr(Y=1|\,x)}\} = \; \beta_0 + X\beta$$

Where Y is a binary response (i.e., 1 when death occurred and 0 when it did not), X=$(x_1, ..., x_k)$ is a vector

of explanatory variables, β_0 is the intercept parameter and β is the vector of slope parameter. (101)

Receiver operating curve (ROC) was generated for each model and the area under the curve (AUC) was

reported as an indicator of discrimination of the model in both derivation and validation data sets. AUC

(also referred to as the C-statistic) was used as the primary measure to compare the alternative

prediction models. Sensitivity and specificity of the models were also provided for comparison between

alternative RS models (i.e., multivariable logistic regression model, SQ only model, and USMM proposed

RS approach). Calibration plots were generated to show the goodness of fit for the final model

graphically. Further details are provided below in model assessment section.

- *Variable selection methods*

Several variable selection methods were applied to the derivation data set and then validated by

applying the model to the validation data set. Both outcomes, 1-year mortality and hospice admission,

were modeled using different automated variable selection methods including forward, backward and

stepwise selection. These selection methods are built-in options in PROC LOGISTIC. Stepwise selection

method with the entry level of $p < 0.2$ and stay level of $p < 0.05$ was applied to select the significant

predictors of the outcome. A total number of 41 predictor variables that had ≤20% missing observations

were included in the model building process.

There are newer variable selection methods including lasso, adaptive lasso, ridge, elastic net, and group

lasso methods. (102–105) We applied two of these methods, adaptive lasso, and elastic net variable

selection, using the SAS procedure 'Proc GLMSELECT'. These methods have advantages over the pre-

existing stepwise selection methods in specific circumstances, especially when the data set includes a

large number of predictors and a limited number of observations; and also when the predictor variables

are highly correlated with each other.(42) Adaptive lasso and elastic net allow the model to include

more than one predictor from a group of correlated predictor variables. In the adaptive lasso method, a weight vector ω is defined for parameter estimates as:

$\omega = 1/|\hat{\beta}|^{\Upsilon}$

Where, $\beta = (\beta_1, \beta_2, ..., \beta_m)$ are the adaptive lasso regression coefficients generated under the constrained optimization problem.

The parameter gamma (Υ) in the above equation is the power transformation of the parameters to form the adaptive weight. Gamma can be specified in the model statement but the default in SAS PROC GLMSELECT is 1.0 (which represents no power transformation). I applied the adaptive lasso option with seven different values of gamma between 0 and 1.

The elastic net method develops a parsimonious model by solving the least square regression problem with constraints on both the sum of the absolute coefficients as well as the sum of the squared coefficients:

Min $||y-X\beta||^2$ subject to: $\sum_{j=1}^{m}|\beta_j| \leq t_1$ and $\sum_{j=1}^{m}|\beta_j^2| \leq t_2$

Where βs are the regression coefficients, and t_1 and t_2 are the constraints applied to the sum of the absolute and sum of the squared coefficients, respectively. (105,107)

Two different options in the model statement were specified to determine the optimal model: validation data and k-fold cross-validation. For both adaptive lasso and elastic net methods, a 4-fold cross validation option was specified. The Selected variables were then included in a logistic regression model and c-statistics generated for both derivation and validation data sets.

The procedure GLMSELECT was used for variable selection only, and not for the logit model development. The underlying assumption of PROC GLMSELECT is that the outcome is continuous, however it is accepted to use the GLMSELECT procedure with the categorical outcome for variable selection only. (83) I applied the GLMSELECT procedure because SAS does not support adaptive lasso

and elastic net options in the LOGISTIC or HPLOGISTIC procedures. Then the logistic model was

developed by including the variables that were selected in GLMSELECT.

A manual variable selection approach was also used by running univariate logistic regression for all the

predictor variables and then entering those with a significance level of ≤ 0.2 into the multivariable

model. The variables with p-value ≤ 0.05 in the multivariable model were included in the final model in

addition to demographic variables (i.e., age, and sex) that were forced in regardless of significance.

- *Model performance assessment*

The most common measure of a predictive model's performance is AUC or c-statistic. Calibration plots

are another way to evaluate the performance of a predictive model. Calibration indicates the degree of

agreement between observed and predicted probabilities and is therefore a measure of model fit. By

plotting the predicted probability of the outcome against the observed probability of the event for

groups of patients (often deciles) calibration plots are diagnostic graphs that help to qualitatively

evaluate how good a model is in the prediction of the outcome. There are two methods that are

commonly used to generate calibration plots: Loess-based and decile-based.(108) In the Loess-based

method the observed and predicted probability of the event for each observation are plotted and a

loess function is used to smooth the plot over all observation. In the decile-based method, data is sorted

by the predicted probabilities and then grouped into deciles. The average observed and predicted

probabilities for each decile are calculated and plotted. A study by Austin and Steyerberg concluded that

loess-based plots have several advantages over the decile based.(109) In fact the decile calibration is

dependent on the number of groups into which data is partitioned. In this chapter a calibration plot was

generated in the validation data. A plot was also made in the derivation data for comparison.

Hosmer-Lemeshow goodness of fit test is a statistical test of GOF and is another metric used to evaluate

the prediction model. To do this test, data is sorted and divided into deciles similar to the method used

in calibration. Hosmer-Lemeshow test statistic is obtained by calculating a chi-square statistic from a 2 x

g table of observed and expected frequencies, where g is the number of groups (ten in this case):

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{(O_i - N_i \bar{\pi}_i)^2}{N_i \bar{\pi}_i (1 - \bar{\pi}_i)}$$

Where O_i is the number of event outcomes in the ith group, N_i is the number of observations in the ith group, and $\bar{\pi}_i$ is the average predicted probability of the outcome in the ith group. This statistic is compared to the χ^2 distribution with (g – 2) degrees of freedom. The large value of χ^2 and small p-value mean the lack of fit for the model.(101,110)

In the evaluation of predictive model performance, measurement of the distance between the predicted and observed outcomes is essential. R-square is the measure for this distance when the outcome is continuous and is calculated as:

$$R^2 = 1 - \frac{\sum_i (Y_i - \hat{Y})^2}{\sum_i (Y_i - \bar{Y})}$$

Where Y_i is the observed outcome, $\bar{Y}$ is the mean of the observed outcome, and $\hat{Y}$ is the predicted outcome. R^2 presents the proportion of the variation that can be explained by model, therefore larger R^2 indicates a better model. Brier score is another measures that calculates the squared difference between the actual binary outcome and the prediction.(90) It is calculated as $(Y - \hat{p})^2$ where $\hat{p}$ is the predicted probability of the binary outcome Y. Brier score is lower where the model fits better. Therefore it is 0 for the perfect fit model whereas a maximum value indicates a non-informative model. The maximum value for the non-informative model is dependent to the outcome incidence and is calculated as $P*(1 - P)^2 + P^2 *(1 - P)$, where P is the outcome incidence.(90) We calculated the maximum value for the non-informative model in this cohort and generated Brier scores for comparison between the different models.

- *Multiple imputation*

To handle the missing data on the predictor variables I used multiple-imputation procedure. To choose the sufficient number of imputations, the data set was imputed twice, once using five imputations and

then using 20. SAS procedure 'PROC MI' was used to impute missing data on categorical and continuous

variables. The parameter estimates and variances from the two MI procedures (5 or 20 imputations)

were compared. Although the parameter estimates and their standard errors were similar, 20

imputations were chosen to maximize relative efficiency. Furthermore, the increase in the computation

time was trivial when the number of imputation was increased from 5 to 20, thus computation time was

not a limitation in this dataset.

There was no missing observation on the two outcomes (i.e., death and hospice), but there was missing

on 15 independent variables. As mentioned above, 6 variables with missing observations were excluded

initially (Table 2.4) resulted in inclusion of 41 predictor variables. Nine of the 41 predictors had different

proportions (0.4-20%) of missing observations. The variables are: race, TUG answer, ADL decline, living

alone, surprise question, tobacco use, KPS, albumin, and cholesterol. All the 41 predictor variables were

used in the imputation procedure. As mentioned above, 6 variables with missing observations were

excluded initially (Table 2.4). The 28 binary variables that has no missing observations and all continuous

variables were included in the model as continuous variables. Six categorical factors (race, TUG answer,

a decline in ADLs, living alone, surprise question, and tobacco use) that had some missing observations,

were imputed using a class statement. Age, albumin, cholesterol, and KPS are recorded as continuous

variables in the data and so were included as continuous factors in the imputation model, although in

the logit model they were included as categorical variables based on their quartiles.

The multiple imputation procedure is typically followed by the MIANALYSE procedure which summarizes

the results of all imputations and provides summarized measure of effect such as relative risk, odds

ratio, or hazard ratio. Variable selection procedures for multivariable models based on data from

multiple imputations is different from available case-based methods, since the variables selected in one

imputation can be different from other imputations, thus there is not a procedure to summarize the

results of different imputations. For model selection in the imputed data, several methods were

suggested in the literature, including using the model selected in available case data, or development of the model in the first imputation and then apply it to the other imputations.(88)

I used two of the four methods proposed by Wood et al., (88) to develop a model and generate the c-statistic for it in the imputed data. The first method was using the model that was developed in the available case data. Using the same set of variables selected in manual variable selection, the model was developed in each derivation set of the imputed data and applied to the corresponding validation set. The predicted probability of outcome was generated for individual patients in each imputed validation data. Average of 20 probabilities for each patient was calculated, and then ROC and AUC were generated for both derivation and validation data sets by modeling the averaged probabilities against the observed outcomes.

As an alternative model development method, I used the following steps to select variables that were consistently selected in different imputations. In the first step, as described by Wood et al., a separate model was developed in each imputed derivation dataset. The three variable selection options (forward, backward, stepwise) that were applied in the available case analysis were also applied to each one of the 20 imputed datasets. A logistic regression model was developed in each imputed derivation data and then applied to the corresponding imputed validation data set. A predicted probability was generated for each individual in the imputed data, then the average of 20 predictions for each person was used to generate a single AUC for each selection method. Selected variables were counted over the 20 imputations for each selection method (forward, backward, and stepwise). Only variables that were selected in all 20 imputations were included in a final model for that variable selection method and was used to calculate the AUC from the validation data.

Figure 2. 2. Manual variable selection in the imputed data
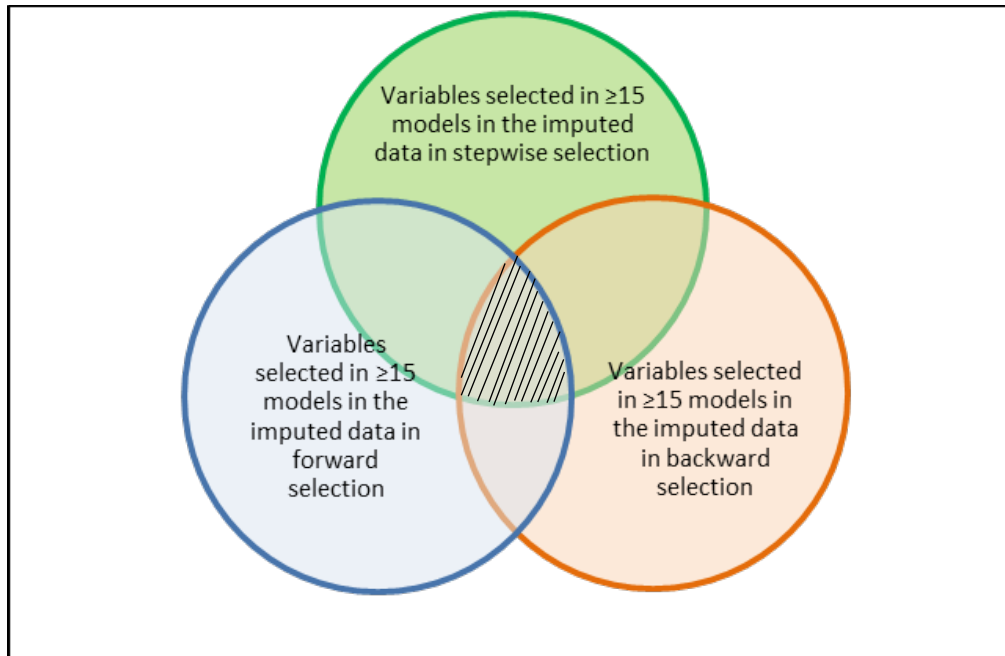
In manual variable selection in the imputed data, variables that were selected for ≥ 15 times in all three

selection methods were considered as the final model selected from the imputed data analysis (Figure

2.2). This set of variables were then applied to the original data to generate c-statistic for both

derivation and validation datasets. To compare the performance of the alternative approaches (final

model developed in this chapter, surprise question model, proposed USMM model), in addition to the

AUC, sensitivity and specificity of the different models were also calculated. In MV model, Sensitivity,

specificity and predictive values were calculated at two different thresholds. All the observations were

sorted based on their predicted probability of the outcome. Then the top 10% and top 20% threshold

were used to calculate sensitivity and specificity of the model. The two thresholds were chosen

arbitrarily for identifying the high risk and low risk groups. However the selection of optimal threshold

for risk groups depends on multiple factors including the cost of false positive cases vs. the false

negative cases. Also the services and resources that can be allocated to each risk level groups influence

the selection of threshold for RS. The selection of threshold is discussed in more details in the discussion

section of this chapter.

- o Alternative risk stratification approaches

The AUCs for the various multivariable logistic regression models were compared to each other as well

as to the alternative risk stratification approaches. The USMM researchers proposed two approaches for

risk stratification: the SQ and a 3-level risk stratification approach. The answer to the SQ is used to find

high-risk patients (answer = no). The three-level approach can be operationalized as a decision tree that

categorizes patients into three risk level (referred to as level 3, 4, 5) based on five variables: SQ,

albumin, an episode of fall, hospitalization, and ER visit since their last USMM visit. If serum albumin is

<2.5 mg/dl, the patient is considered high risk. If SQ is answered 'No' and the patient has a history of fall

or hospitalization or ER visit since their last visit, then the patient is high risk and assigned to level 5 of

risk. If SQ is answered 'No' without any fall or hospitalization/ER, then the patient is at intermediate risk

level or level 4. If Albumin is >2.5 mg/dl and SQ answer is 'Yes' then the patient is in the low-risk (level 3)

in this approach (Figure 2.3).

Figure 2. 3. The USMM proposed 3-level risk stratification approach
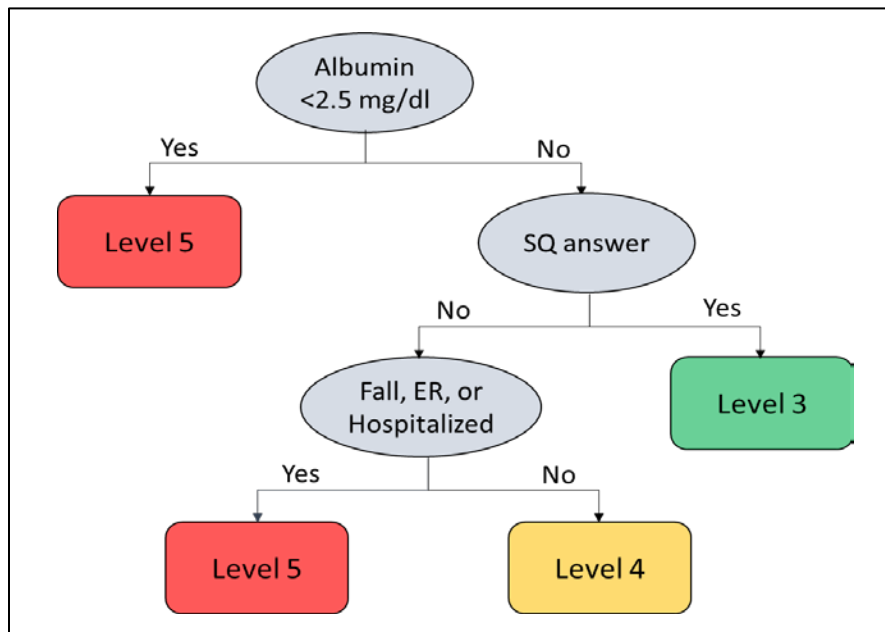
To have comparable measures between different models, univariate logistic regression analyses were performed for the SQ and for the USMM proposed risk levels. The AUC, sensitivity and specificity of these two approaches in the validation data were generated and compared to the different multivariable logit models.

- ***Results***

  o <u>Study population</u>

The final study population consisted of 7445 patients who had their first USMM visit in the calendar year of 2015, had available claims data, and were followed up for at least one year (Figure 2.1).

The minimum and maximum follow up time for this cohort were 1 and 794 days, respectively; with average (standard deviation) of 459 (239) days and median (interquartile range) of 517 (q1=246, q3=658) days (Table 2.5).

In the final cohort of 7445 patients (Table 2.4), 66% were female, 63% white, the average age was 82 years, 99% had Medicare coverage, and 27% were dual eligible (both Medicare and Medicaid); 54% of the cohort had a KPS≤40 – indicating severe disability with the need for necessary assistance and specialized care. Prevalence of hypertension, hyperlipidemia, diabetes, and cancer were 81%, 50%, 34%, and 8% respectively. Over 50% of patients had 5 or more medical conditions.

Overall, 45% (n=3345) of the cohort died, and 19% were admitted to the hospice over the total follow up time. However, the 1-year mortality and hospice admission rates within the first year of follow up were 32% (n=2408) and 10% (n=752), respectively (Table 2.5). Among hospice-admitted patients, 765 (55%) died within three months of their admission. Overall 2680 deaths (80% of all deaths) occurred outside of hospice. Table 2.4 demonstrates the population characteristics and Table 2.5 displays outcome events.

Table 2. 4. Cohort population description, by the outcome rates and unadjusted odds ratios (N=7445)

Variable	N (%)	Missing N (%)	Death %	Unadjusted OR	Hospice %	Unadjusted OR
Baseline characteristics						
Age -65 -74 -75 – 84 -85 – 94 -95+	1826 (24.5) 2249 (30.2) 2796 (37.6) 574 (7.7)	0	21.0 30.5 39.6 40.2	Ref 1.65* 2.47* 2.53*	4.5 8.5 13.9 15.5	Ref 1.95* 3.39* 3.85*
Sex -Male -Female	2513 (33.7) 4932 (66.3)	0	36.7 30.1	1.34* Ref	10.8 9.8	1.12 Ref
Race -White -Black -Other	4684 (62.9) 1148 (15.4) 201 (2.7)	1412 (19.0)	27.4 18.5 19.9	Ref 0.6* 0.66*	11.3 6.6 6.0	Ref 0.56* 0.5*
Tobacco use (current vs not) -Yes -No	645 (8.7) 6412 (86.1)	388 (5.2)	21.7 31.1	0.61* Ref	7.4 10.1	0.71* Ref
Dual-eligible -Yes -No	2024 (27.2) 5421 (72.8)	0	23.1 35.8	0.54 Ref	6.0 11.6	0.49* Ref
Lives alone -Yes -No	884 (11.9) 5511 (74.0)	1050 (14.1)	18.0 30.3	0.5* Ref	5.4 10.7	0.48* Ref
S.Q- No -No -Yes	1045 (14.0) 5381 (72.3)	1019 (13.7)	44.4 25.3	2.36* Ref	19.1 8.0	2.7* Ref
KPS -Mild /moderate (50-100) -Severe disability (10-40)	3376 (44.9) 4042 (54.3)	27 (0.4)	22.3 40.5	Ref 2.38*	5.7 13.8	Ref 2.66*
TUG -<30 sec -≥30 sec -Non ambulatory	2538 (34.1) 1377 (18.5) 2027 (27.2)	1503 (20.1)	17.5 22.9 30.3	Ref 1.4* 2.1*	7.2 10.0 10.9	Ref 1.43* 1.58*
Decline in ADLs -Decline -Improve -No change	1063 (14.3) 311 (4.2) 4889 (65.7)	1182 (15.9)	30.3 1.6 26.8	1.19* 0.05* Ref	13.2 2.3 9.6	1.43* 0.22* Ref
Pressure ulcer -Yes -No	940 (12.6) 6505 (87.4)	0	37.3 31.6	1.29* Ref	13.2 9.7	1.42* Ref

Table 2. 4. (cont'd)						
cancer						
-Yes	566 (7.6)	0	38.0	1.31*	12.9	1.35*
-No	6879 (92.4)		31.9	Ref	9.9	Ref
Cholesterol result (mg/dl) Quartiles						
-<136	1554 (20.9)	1056	38.3	2.27*	10.9	1.16
-136 - <164	1625 (21.8)	(14.2)	27.5	1.39*	9.4	0.98
-164 - <195	1589 (21.3)		24.4	1.18	9.3	0.97
- 195+	1621 (21.8)		21.5	Ref	9.6	Ref
Albumin result (g/dl) Quartiles						
-<3.2	1669 (22.4)	637 (8.6)	50.5	5.66*	13.4	2.22*
-3.2 – <3.5	1610 (21.6)		30.4	2.43*	10.9	1.77*
-3.5 - <3.8	1820 (24.5)		22.3	1.59*	9.3	1.47*
-3.8+	1709 (23.0)		15.3	Ref	6.5	Ref
Medical history (CCW variables)						
Hypothyroidism						
-Yes	2050 (27.5)	0	30.4	0.89*	9.5	0.91
-No	5395 (72.5)		33.1	Ref	10.3	Ref
Myocardial infarction						
-Yes	3 (0.04)	0	33.3	1.05	0	--
-No	7442 (99.9)		32.3	Ref	10.1	Ref
Anemia						
-Yes	2243 (30.1)	0	26.4	0.67*	10.7	1.1
-No	5202 (69.9)		34.9	Ref	9.9	Ref
Asthma						
-Yes	309 (4.2)	0	20.1	0.51*	5.8	0.54*
-No	7136 (95.9)		32.9	Ref	10.3	Ref
Atrial fibrillation						
-Yes	1233 (16.6)	0	37.4	1.31*	11.4	1.17
-No	6212 (83.4)		31.3	Ref	9.9	Ref
BPH						
-Yes	504 (6.8)	0	30.8	0.92	10.7	1.1
-No	6941 (93.2)		32.5	Ref	10.1	Ref
Breast cancer						
-Yes	224 (3.0)	0	29.9	0.89	8.9	0.87
-No	7221 (97.0)		32.4	Ref	10.1	Ref
Cataract						
-Yes	184 (2.5)	0	14.7	0.35*	3.3	0.29*
-No	7261 (97.5)		32.8	Ref	10.3	Ref
Chronic kidney diseases						
-Yes	3006 (40.4)	0	24.6	0.54*	10.3	1.03
-No	4439 (59.6)		37.6	Ref	10.0	Ref
Colorectal cancer						
-Yes	95 (1.3)	0	36.8	1.22	9.5	0.93
-No	7350 (98.7)		32.3	Ref	10.1	Ref

Table 2. 4. (cont'd)						
COPD						
-Yes	1946 (26.1)	0	29.2	0.82*	8.7	0.81*
-No	5499 (73.9)		33.5	Ref	10.6	Ref
Depression						
-Yes	1615 (21.7)	0	23.5	0.58*	9.7	0.95
-No	5830 (78.3)		34.8	Ref	10.2	Ref
Diabetes						
-Yes	2519 (33.8)	0	29.3	0.81*	8.1	0.7*
-No	4926 (66.2)		33.9	Ref	11.1	Ref
Endometrial cancer						
-Yes	27 (0.4)	0	25.9	0.73	14.8	1.55
-No	7418 (99.6)		32.4	Ref	10.1	Ref
Glaucoma						
-Yes	337 (4.5)	0	30.9	0.93	9.8	0.97
-No	7108 (95.5)		32.4	Ref	10.1	Ref
Heart failure						
-Yes	2542 (34.1)	0	29.0	0.79*	10.0	0.98
-No	4903 (65.9)		34.1	Ref	10.1	Ref
Hip fracture						
-Yes	81 (1.1)	0	35.8	1.17	9.9	0.98
-No	7364 (98.9)		32.3	Ref	10.1	Ref
Hyperlipidemia						
-Yes	3686 (49.5)	0	24.1	0.47*	8.5	0.7*
-No	3759 (50.5)		40.4	Ref	11.7	Ref
Hypertension						
-Yes	6056 (81.3)	0	29.7	0.54*	9.5	0.72*
-No	1389 (18.7)		44.1	Ref	12.7	Ref
Ischemic heart diseases						
-Yes	1270 (17.1)	0	31.6	0.96	11.3	1.16
-No	6175 (82.9)		32.5	Ref	9.9	Ref
Lung cancer						
-Yes	70 (0.9)	0	52.9	2.37*	17.1	1.86
-No	7375 (99.1)		32.2	Ref	10.0	Ref
Osteoporosis						
-Yes	819 (11.0)	0	21.1	0.53*	8.6	0.82
-No	6626 (89.0)		33.7	Ref	10.3	Ref
Prostate cancer						
-Yes	175 (2.4)	0	43.4	1.63*	17.1	1.88*
-No	7270 (97.7)		32.1	Ref	9.9	Ref
Osteoarthritis						
-Yes	2761 (37.1)	0	24.5	0.55*	9.5	0.9
-No	4684 (62.9)		37.0	Ref	10.4	Ref
TIA/stroke						
-Yes	800 (10.8)	0	29.6	0.87	12.5	1.31*
-No	6645 (89.3)		32.7	Ref	9.8	Ref

Table 2. 4. (cont'd)						
Continuous variables[†]						
Age (mean± sd)	82.2 ± 9.3	0	--	1.04	--	1.05*
Number of lab tests (Median, IQR)	0 (0 – 5)	0	--	1.02*	--	0.97*
Number of medications (Median, IQR)	9 (5 – 13)	0	--	0.98*	--	0.97*
Comorbidity count (Median, IQR)	5 (3-6)	0	--	0.81*	--	0.95*
Variables that were not included in the analysis due to >20% missing observations						
Decline IADLs -Decline -Improve -No change	730 (9.8) 524 (7.0) 984 (13.2)	5207 (69.9)	2.7 1.2 2.0	1.36 0.56 Ref	2.1 2.1 3.3	0.62 0.64 Ref
Global health compared to a year ago -Better -Worse -The same	55 (0.7) 316 (4.2) 1185 (15.9)	5889 (79.1)	21.8 54.4 28.3	0.71 3.03* Ref	10.9 15.2 7.3	1.57 2.29* Ref
Fall since last visit -Yes -No	184 (2.5) 1546 (20.8)	5715 (76.8)	35.9 34.2	1.08 Ref	8.2 9.1	0.89 Ref
Hospitalization since last visit -Yes -No	872 (11.7) 1565 (21.0)	5008 (67.3)	45.1 52.3	o.75* Ref	9.4 5.1	1.93* Ref
ER since last visit -Yes -No	790 (10.6) 1649 (22.2)	5006 (67.2)	32.2 54.4	0.4* Ref	8.2 5.5	1.55* Ref
Lost weight -Yes -No	1243 (16.7) 2431 (32.7)	3771 (50.7)	22.1 1.8	15.4* Ref	12.8 4.5	3.1* Ref

IQR: interquartile range; sd: standard deviation; S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; IADL: instrumental activities of daily living; TIA: transient ischemic attack; FU: follow-up; mg/dl: milligram per deciliter; g/dl: gram per deciliter;

[*] P-value < 0.05 in univariate analysis with the outcomes;

[†] The unadjusted OR for continuous variables were generated for 1 unit change in the independent variable; Age was included as categorical variable in the analyses;

Table 2. 5. Outcomes and follow up duration

Variable	N (%)	Missing N
FU time in days -mean ± sd* -median (q1 - q3)	459 ± 239 517 (246 - 658)	0
Death -over the total follow up time -one-year	3345 (44.9) 2408 (32.3)	0
Hospice admission -over the total follow up time -one-year	1391 (18.7) 752 (10.1)	0
* sd: standard deviation;		

Nine of the 41 patient-level independent variables that were included in the analysis, have missing data.

To explore the importance of the missing data, the association between predictor's missingness and five

key variables without missing (i.e. death, hospice, age, sex, and dual eligibility) were evaluated. A

dummy variable was generated for missing data on each of the seven predictors (1= missing and 0= non-

missing). Table 2.6 contains the p-values from the univariate regression models. Also the direction and

magnitude of the association were also shown in Table 2.6. Although nine predictor variables had

missing data, variables KPS and tobacco-use had a small percentage of missing (0.4%, and 5%,

respectively) and were not included in Table 2.6. The fact that missingness on all seven predictors was

consistently and significantly associated with a higher rate of mortality suggests that missingness was

not at random in this data. In contrast, missingness on predictors were not significantly associated with

hospice admission. Additionally older age and male gender were often associated with missingness on

the predictors. Hospice admission was not significantly associated with the missingness except for two

variables, TUG and cholesterol. A conclusion from findings shown in Table 2.6 is that missing data can be

very informative in this study and exclusion of the observations with missing data (as occurs

automatically in regression procedures) could negatively affect the validity of the model.

Table 2. 6. Association between missing observations on predictor variables and the outcomes, age, gender and Medicare/Medicaid dual-eligibility, p-values, magnitude and direction of the effect

Outcome / Variable*	Missing (N=7445)	Death (all)	Death 1-yr	Hospice (all)	Hospice 1-yr	Age	Male	Dual-eligible
Race	19%	<.0001 ↑↑	<.0001 ↑↑	0.29 =↓	0.58 =↓	0.13 =↑	0.001 ↑	0.06 =↓
SQ	14%	<.0001 ↑↑	<.0001 ↑↑	0.70 =↓	0.06 ↑	<.0001 =↑	0.05 ↑	0.17 =↓
TUG	20%	<.0001 ↑↑	<.0001 ↑↑	0.0008 ↑	<.0001 ↑	<.0001 =↑	0.01 =↑	<.0001 ↓
Lives alone	14%	<.0001 ↑↑	<.0001 ↑↑	0.26 =↓	0.23 =↑	0.02 =↑	0.03 =↑	0.43 =↓
ADL decline	16%	<.0001 ↑↑	<.0001 ↑↑	0.57 =↓	0.06 =↑	0.005 =↑	0.0006 ↑	0.4 ↓
Cholesterol	14%	<.0001 ↑↑	<.0001 ↑↑	0.46 =↑	0.03 =↑	<.0001 =↑	0.60 =↓	0.0002 ↓
Albumin	9%	<.0001 ↑↑	<.0001 ↑↑	0.46 =↑	0.23 =↑	0.23 =↑	0.66 =↓	<.0001 ↓

*Variables with ≤20% missing were evaluated for the association to outcomes and age, gender, and insurance status;
Shaded cells show the statistically significant association;
Arrows indicate the direction of the association between the missingness and outcome and number of arrows show the magnitude of the association (↑ means the outcome rate is higher when the variable is missing than when it is not missing);

o <u>Outcome: One-year mortality</u>

For each of the two outcomes (mortality and hospice admission) analyses were done in two parts, first using the available case data (original data that has missing observations), and then using the imputed data.

- *Available case analysis*

The alternative variable selection approaches (automatic and manual selection) were applied to the derivation dataset using logistic regression model. A total of 41 independent variables were included in the model building process. All variables were included in the model as categorical variable except for comorbidity count, number of medications, and number of lab tests (shown at the bottom of Table 2.4). Age, albumin and cholesterol were categorized as illustrated in Table 2.4. More than one-third of the observations were excluded from the analysis due to missing data on one or more predictors.

50

The number of observations included in the various models differed when models were based on a different set of variables (which have a different number of missing observations). For example in all three stepwise, backward and forward methods, all 41 variables were in the model statement, therefore observations with missing on any of the predictors were excluded right at the beginning. Whereas, when using adaptive lasso method, the variable selection was first made using PROC GLMSELECT and then the variables were included in the logistic regression model in both derivation and validation data sets, therefore the number of observations which are excluded from the analysis is different from the one in stepwise selection methods. The results of different variable selection methods are demonstrated in Table 2.7. The SAS built-in selection methods are reported at first, following by the adaptive lasso and elastic net selection methods (each with two selection rules), and manual selection method. At the bottom of the table, the best model that was developed in imputed data (later in this chapter) was also applied to the available data for comparison.

Brier score was generated for each model as a measure of the overall goodness of fit. As mentioned in the method section, the lower Brier score means the model fits better. However the maximum limit for the Brier score is not a constant and is calculated based on the incidence of the outcome. The incidence rate of mortality (33%) was used in the equation $P*(1-P)^2 + P^2*(1-P)$, and the maximum limit of 0.18 was calculated for the Brier score of a non-informative model.

Table 2. 7. Model development using alternative variable selection methods for 1-year mortality in available case data

AUC and 95% confidence limits for both derivation and validation data sets, Brier score in validation and final variable selected (N=3722 derivation and 3723 validation)					
Variable selection	**N analyzed* Derivation**	**AUC Derivation**	**AUC Validation**	**Brier Score Validation**	**Selected variables in the final model**
Automatic variable selection methods					
Stepwise selection	2055	0.7522 (0.7231-0.7813)	0.7697 (0.7476-0.7919)	0.1473	13 variables: age, sex, race, dual-eligible, SQ, albumin, cholesterol, KPS, ADL decline, anemia, depression, hyperlipidemia, number of meds
Forward	2055	0.7458 (0.7162-0.7754)	0.7636 (0.7411-0.7861)	0.1473	11 variables: race, dual-eligible, SQ, albumin, cholesterol, KPS, ADL-decline, anemia, depression, hyperlipidemia
Backward	2055	0.7453 (0.7166-0.7740)	0.7624 (0.7402-0.7846)	0.1479	10 variables: race, dual-eligible, SQ, albumin, cholesterol, KPS, ADL-decline, AF, IHD, dx-count
Adaptive lasso[†] (validation data, Gamma=1.0)	2089	0.7631 (0.7351-0.7911)	0.7673 (0.7427-0.7918)	0.1277	24 variables: age, sex, race, dual-eligible, SQ, albumin, cholesterol, KPS, ADL- decline, TUG, number of meds, hypothyroidism, anemia, AF, BPH, cataract, CKD, depression, diabetes, hyperlipidemia, hypertension, IHD, RA/OA, stroke/TIA
Adaptive Lasso[†] (4-fold CV Gamma=0.1)	2081	0.7645 (0.7365-0.7924)	0.7616 (0.7368-0.7863)	0.1290	27 variables: age, sex, race, dual-eligible, SQ, living-alone, albumin, cholesterol, KPS, ADL-decline, TUG, number of meds, number of labs, diagnosis-count, cancer, anemia, asthma, AF, BPH, cataract, CKD, colorectal cancer, depression, hyperlipidemia, hypertension, IHD, stroke/TIA

Table 2. 7. (cont'd)					
Elastic Net[†] (validation data)	2081	0.7644 (0.7364-0.7923)	0.7631 (0.7385-0.7876)	0.1287	32 variables: age, sex, race, dual-eligible, SQ, living-alone, albumin, cholesterol, KPS, ADL-decline, TUG, number of meds, number of labs, diagnosis-count, cancer, pressure-ulcer, hypothyroidism, anemia, asthma, AF, BPH, cataract, CKD, colorectal cancer, depression, endometrial-ca, glaucoma, hyperlipidemia, hypertension, IHD, RA/OA, stroke/TIA
Elastic Net[†] (4-fold CV)	2055	0.7653 (0.7371-0.7935)	0.7668 (0.7420-0.7916)	0.1270	33 variables: age, sex, race, dual-eligible, SQ, living-alone, smoking, albumin, cholesterol, KPS, ADL-decline, TUG, number of meds, number of labs, diagnosis-count, cancer, pressure-ulcer, hypothyroidism, anemia, asthma, AF, BPH, cataract, CKD, colorectal cancer, depression, endometrial-ca, glaucoma, hyperlipidemia, hypertension, IHD, RA/OA, stroke/TIA
Manual variable selection					
Full model	2055	0.7653 (0.7370-0.7935)	0.7664 (0.7415-0.7912)	0.1270	All 41 variables included, no selection method
Manual variable selection-final model	2290	0.7719 (0.7476-0.7962)	0.7634 (0.7410-0.7859)	0.1437	11 variables: age, race, dual-eligible, SQ, albumin, cholesterol, KPS, ADL-decline, hyperlipidemia, depression Forced to the model: sex

Table 2. 7. (cont'd)				
Model developed in imputed data and applied to the available data				
Backward variable selection- in the imputed data	2636	0.7854 (0.7648-0.8060)	0.7624 (0.7422-0.7826)	0.1564	18 variables: age, dual-eligible, SQ, albumin, cholesterol, KPS, ADL-decline, anemia, CKD, hyperlipidemia, depression, hypertension, rheumatoid arthritis, pressure-ulcer, Cataract, osteoporosis, number of meds, number of labs
S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; AF: atrial fibrillation; HF: heart failure; CKD: chronic kidney disease; RA/OA: rheumatoid arthritis/osteoarthritis; IHD: ischemic heart diseases; BPH: benign prostatic hyperplasia; TIA: transient ischemic attack; *The numbers are different because first the variable selection was done in PROC GLMSELECT and then variables included in PROC LOGISTIC to generate AUCs, so not all the variables included in the final model †Adaptive lasso and elastic net methods were conducted using two methods of validation and weighting parameters (Table 2.8)				

When applied to the validation data, the different variable selection strategies resulted in AUCs that were very similar for all models. The confidence intervals around the c-statistic are also comparable in width, so the precision of the C-statistics are also similar. The stepwise selected model (AUC=0.7697) had the highest c-statistic, although the difference between it and other models is trivial and of no practical importance. Likewise the difference in the Brier score between different models is small, although this metric indicates slightly better fit in the models that were based on adaptive lasso and elastic net selection methods. Although the advanced variable selection methods made minimal differences in the discrimination of the model, the number of selected variables was much more than with the stepwise and manual methods. Thus there was no evidence that any of the variable selection approaches has significantly better performance than other methods in terms of the discrimination ability (C-statistic); however, the manually selected model has a good performance (c-statistic of 0.7634), is parsimonious (only 11 variables) and clinically logical (it includes demographics, functional, and indicators of nutritional status including albumin and cholesterol) compared to the other models.

There are variables that were consistently selected in the different models regardless of the variable selection method, including albumin, cholesterol, ADL-decline, SQ, KPS, race, and dual eligibility for Medicare and Medicaid. This emphasizes the central importance of these variables in the prediction of mortality in this cohort of older adults. Functional status variables are also shown to be important in the prediction of the adverse outcomes. Other variables including age, sex, and TUG were frequently selected but not in all models. The most variation between different models was observed for the medical history variables. Hyperlipidemia and depression were often selected but other CCW variables such as endometrial and colorectal cancer were only occasionally selected. The number of medications was also selected in multiple variable selection methods; it can represent the general health of the patient as well as the frequency and severity of different conditions.

Table 2.8 displays the results of specifying different gamma values and different selection rules in the adaptive lasso variable selection method for 1-year mortality. Gamma is a parameter in the adaptive weight calculation, and alternative selection rules in the adaptive lasso method are k-fold cross-validation or use of validation data. Table 2.8 was generated to help determine the appropriate gamma to be used in adaptive lasso variable selection. The number of effects is the total number of variables that selected including each level of classification variables as a dummy variable. It means the number of variables is often less than the number of effects illustrated in table 2.8. This table indicates that the optimal gamma for the different selection method (cross-validation or using separate validation dataset) are different, although the difference in the optimal model criteria (ASE and CV Press) between the different gammas is minimal. Average square error (ASE) and cross-validation predicted the residual sum of squares statistic (CV Press) are the model fit summary statistics that used for variable selection. A lower score in both criteria means a better fit of the model.

55

Table 2. 8. Different gamma- adaptive lasso variable selection for 1-year mortality

Selection	Selected variables	Optimal model criteria[†]
Validation	*Number of predictors*	*ASE + 0.001*
Gamma=0	38 effects	0.1298
Gamma=0.1	36 effects	0.1298
Gamma=0.3	34 effects	0.1297
Gamma=0.5	31 effects	0.1296
Gamma=0.7	31 effects	0.1295
Gamma=0.9	31 effects	0.1294
Gamma=1.0*	29 effects	0.1294
4-fold CV	*Number of predictors*	*CV PRESS*
Gamma=0	34 effects	0.1117
Gamma=0.1*	33 effects	0.1117
Gamma=0.3	34 effects	0.1118
Gamma=0.5	32 effects	0.1118
Gamma=0.7	31 effects	0.1119
Gamma=0.9	31 effects	0.1119
Gamma=1.0	30 effects	0.1120
*Selected gamma based on the criteria and the number of variables; [†]Average square error (ASE) and CV PRESS are error measures that represent the goodness of model fit.		

Figures 2.4 and 2.5 demonstrate the process of adding and removing variables using adaptive lasso and elastic net variable selection methods, respectively. The bottom panel in each figure shows the average squared error (ASE) of each model. It illustrates the lowest ASE of the selected model that can be correlated to the predictors in the model in the top panel. Both figures show that a few steps before step 40, the minimum ASE was achieved and after that it is a plateau with no more gain from adding or removing variables. SAS output provides a table of details of the variable selection process at each step.

Figure 2. 4. Adaptive lasso variable selection process using GLMSELECT for the mortality outcome (gamma=1.0 and validation dataset)
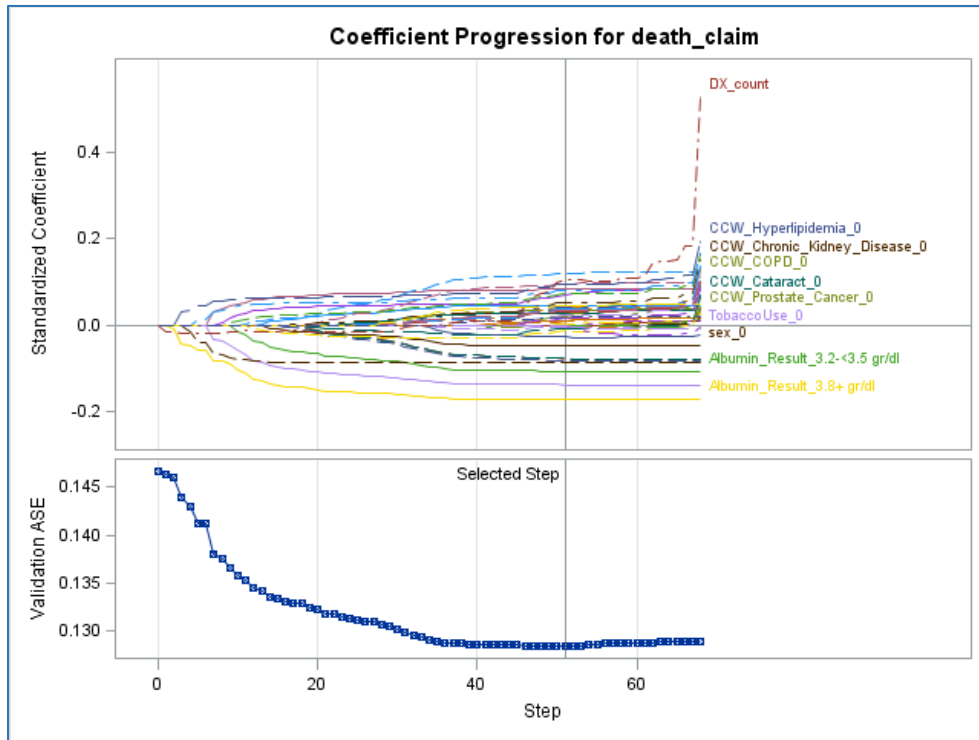
Figure 2. 5. Elastic net variable selection process using GLMSELECT for the mortality outcome (validation dataset)
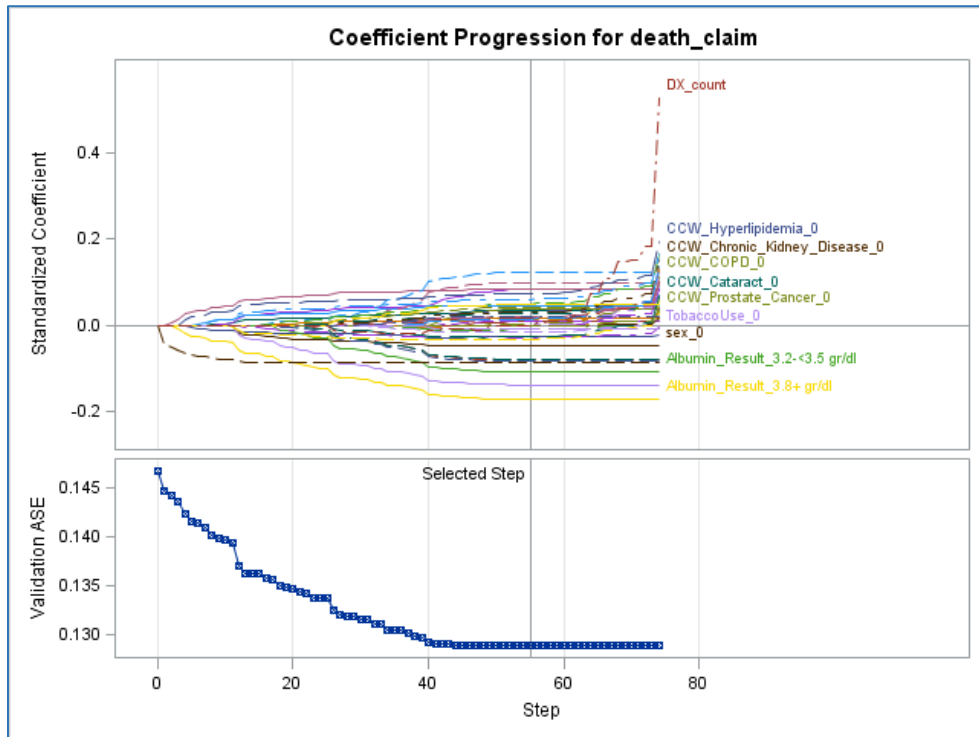
- *Imputed data analysis*

To compare and choose the optimal number of imputations multiple-imputation was performed with 5

and 20 imputations. Tables 2.8 and 2.9 show the parameter estimates and variances from the multiple

imputation procedure. These tables contain information on the continuous variables only, despite the

fact that both continuous and classification variables were included in the model and imputed. PROC MI

(SAS version 9.4) does not report the summary statistics for classification variables. Therefore the

summary tables in SAS results (Tables 2.8 and 2.9) include information on only continuous variables that

have missing data, although the model includes all the variables with and without missing data. As

described in methods, parameter estimates, variances, and confidence intervals for the variables were

similar between the 5 and 20 imputations; the number 20 was selected for imputation to maximize the

relative efficiency.

Table 2. 9. Parameter estimates for the continuous variables from multiple imputation procedure-comparison of 20 and five imputations

Variable	Mean	Std Error	95% Confidence Limits		DF	Min	Max	Mu0	t for H0: Mean= Mu0	Pr > |t|
Parameter Estimates (20 Imputations)										
Albumin Result	3.41	0.01	3.40	3.42	3260.5	3.41	3.41	0	585.70	<.0001
Cholesterol Result	167.72	0.54	166.66	168.77	1547.9	167.38	167.98	0	312.15	<.0001
KPS	44.43	0.12	44.19	44.66	7407.8	44.42	44.44	0	371.49	<.0001
Parameter Estimates (5 Imputations)										
Albumin Result	3.41	0.01	3.40	3.42	892.26	3.41	3.41	0	584.03	<.0001
Cholesterol Result	167.70	0.53	166.66	168.74	674.25	167.50	167.85	0	315.79	<.0001
KPS	44.43	0.12	44.19	44.66	7352.4	44.42	44.44	0	371.48	<.0001
Only continuous variables that includes missing values are outputs of the multiple imputation procedure										

Table 2. 10. Variance information for the continuous from multiple imputation procedure- comparison of 20 and 5 imputations

Variable	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Variance Information (20 Imputations)							
Albumin Result	0.000002	0.00003	0.00003	3260.5	0.062	0.056194	0.997
Cholesterol Result	0.03	0.26	0.29	1547.9	0.11	0.098090	0.995
KPS	0.00003	0.01	0.01	7407.8	0.002	0.002386	0.999
Variance Information (5 Imputations)							
Albumin Result	0.000002	0.00003	0.00003	892.26	0.07	0.06	0.993
Cholesterol Result	0.02	0.26	0.28	674.25	0.08	0.08	0.999
KPS	0.00003	0.01	0.01	7352.4	0.002	0.002	0.999
Only continuous variables that includes missing values are outputs of the multiple imputation procedure							

The original data set with 7445 observations were used in multiple imputation, with 20 imputations the imputed data consisted of 148900 (7445*20) observations. An indicator of the subgroups (derivation or validation) for each patient was added to the dataset before imputations, thus the derivation and validation subgroups are fixed across the 20 imputations. The alternative variable selection methods were applied to the imputed data following the same steps described previously in the method section. AUCs were generated for derivation and validation data by analyzing the individual predicted probabilities of outcome (average of predictions in 20 imputations) against the observed outcome. Table 2.11 displays the results of variable selection methods for the 1-year mortality in the imputed data.

Table 2. 11. Model development using alternative variable selection methods for 1-year mortality using imputed data, AUCs for both derivation and validation data sets

Variable selection	Derivation AUC (N=3722)	Validation AUC N=(3723)	Variables*
Automatic selection methods			
Stepwise	0.7880	0.7730	15 variables: age, dual-eligibility, SQ, albumin, cholesterol, KPS, ADL-decline, anemia, CKD, hyperlipidemia, pressure-ulcer, cancer, number of meds, number of labs, diagnosis-count
Forward	0.7879	0.7728	15 variables: age, dual-eligibility, SQ, albumin, cholesterol, KPS, ADL-decline, anemia, CKD, hyperlipidemia, pressure-ulcer, cancer, number of meds, number of labs, diagnosis-count
Backward	0.7877	0.7756	18 variables: age, dual-eligibility, SQ, albumin, cholesterol, KPS, ADL decline, anemia, cataract, CKD, depression, hyperlipidemia, hypertension, osteoporosis, rheumatoid arthritis, pressure ulcer, number of meds, number of labs
Manual selection			
Manual variable selection from Imputed data[&]	0.7812	0.7663	15 variables: age, dual-eligibility, SQ, albumin, cholesterol, KPS, ADL decline, anemia, CKD, hyperlipidemia, rheumatoid arthritis, pressure ulcer, cancer, number of meds, number of labs
Manual variable selection- from available case data, applied to the imputed data[#]	0.7634	0.7541	11 variables: age, race, dual-eligible, SQ, albumin, cholesterol, KPS, ADL decline, hyperlipidemia, depression Forced to the model: sex
S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; AF: atrial fibrillation; HF: heart failure; CKD: chronic kidney disease; AUCs in the imputed data are based on the average of 20 predictions for each individual from the 20 imputations; *Variables that are selected in all 20 imputations built the final model; [&]Variables that are selected >15 times in all three methods (forward, backward, stepwise); [#] From Table 2.7;			

- *Comparison of the risk stratification models*

Comparing the models developed in the available data shown in Table 2.7, the best model was the

manually selected model, because it is a parsimonious model while its discrimination is similar to the

other models that are much more complex (i.e., have twice as many variables). The models developed in

the imputed data did not improve the AUC compared to the available data. Backward selection had the

best AUC among the variable selection methods in the imputed data (Table 2.11). These two models

(Manual selection in available data and backward selection in imputed data) were both applied to the

available case data and were compared to the two alternative approaches proposed by USMM

providers: SQ, and 3-level risk stratification. Table 2.12 describes the prevalence of each risk level in the

cohort using these two approaches.

Table 2. 12. Prevalence of the risk levels determined by the USMM risk-stratification approaches
(N=7445)

Risk stratification approach	Risk level	N (%)[†] Total (N=7445)
SQ*	High risk (answer=No)	1045 (14.0)
	Low risk (answer=Yes)	5381 (72.3)
3-level risk approach	High risk (level 5)	532 (7.2)
	Intermediate risk (level 4)	678 (9.1)
	Low risk (level 3)	4817 (64.7)
*Surprise question; [†]There are missing values on the SQ and other variables (as reported in Table 2.4) that are used in the 3-level risk approach, hence the totals do not add to 100%; USMM risk stratification approaches were proposed by USMM providers;		

Table 2.13 displays the AUCs of the four alternative risk stratification models in this cohort (i.e., SQ, 3-

level, manual selection logistic model, backward selection in imputed data). Sensitivity and specificity of

the models also provided by defining the high risk and low-risk group and comparing the observed

events in each group. The high-risk group in the manually selected model were identified as the top 10%

and 20% of the predicted probabilities in the model. These cutoff points are selected arbitrarily to show

the impact of different cutoffs on the number of patients who are falsely categorized. The final decision

about the appropriate cutoff value for risk stratification must be made considering the resources that

the company can allocate to the interventions for different risk levels. For example if planned

interventions for the high-risk group are costly and resources (including money, facilities, and human

resources) are very limited, then a more stringent cutoff such as top 10% seems practical. Whereas if

the cost of interventions for the high-risk group is relatively low and resources can support them for a

larger number of patients, then cutoffs can be more relaxed (top 20%). Another approach can be using a

cutoff at the predicted probability of 0.50, this approach is not suitable for our data since the predicted

probability of the outcome in this study has an average and median about 0.15 which makes the

probability of 0.50 a very high bar for high-risk definition in this cohort.

The high-risk definition in the two approaches that are currently in use by USMM providers is given in

the methods section. In the surprise question approach, high-risk patients are those with the answer

"No" to the surprise question. For the 3-level USMM risk stratification approach (Figure 2.3), the high-

risk group was defined twice, first as level 5 and then level 4 and 5 as shown in Table 2.13. Sensitivity

and specificity of each model can help policymakers in the corporation to make a better decision for the

risk cutoff point based on the cost of false positive vs. false negative cases.

Table 2. 13. Comparison of the alternative risk stratification approaches for 1-year mortality (N=3723 validation)

Model	AUC validation*	N analyzed	High-risk group (prevalence)	Sensitivity	Specificity	PVP[‡]	NPV[§]
SQ only	0.5552 (0.5400-0.5705)	3227	Answer "No" (14%)	24.1%	86.9%	43.6%	73.2%
USMM 3-level risk stratification	0.5994[†] (0.5814-0.6173)	3043	Level 5 (7%)	18.1%	95.0%	58.3%	75.0%
			Level 4 and 5 (17%)	33.9%	84.9%	46.5%	76.9%
Manual selection (from available)	0.7634 (0.7410-0.7859)	2312	Top 10%	25.1%	94.0%	52.8%	82.6%
			Top 20%	44.5%	86.5%	46.7%	85.5%

Table 2. 13. (cont'd)							
Backward selection (from imputed)	0.7624 (0.7422-0.7826)	2694	Top 10%	24.7%	94.8%	60.7%	79.4%
			Top 20%	43.2%	87.6%	53.3%	82.5%

*To make the results comparable, the AUC for SQ model and USMM model was also generated from the validation data
†USMM risk level included in the model as a 3-level predictor for AUC calculation
Sensitivity is the proportion of deaths that are classified as high risk by the model; Specificity is the proportion of non-deaths that are classified as low risk by the model;
‡ PVP: predictive value positive is the proportion of model-identified high risk cases who are truly high risk
§ PVN: predictive value negative is the proportion of model-identified low risk cases who are truly low risk

Both multivariable models have much higher AUCs than the current USMM approaches. The fact that the additional variables in the multivariable model are being routinely collected and recorded in the USMM database, makes this model an excellent approach for risk stratification. As mentioned above the choice of a cutoff point is dependent on multiple factors, including the company resources, the cost of interventions for each risk level, and the cost of misclassification of patients. Sensitivity and specificities show the proportion of high and low-risk patients that are correctly classified by each model. For example, when we apply the model to a group of 1000 patients. The mortality rate is 32% in the USMM population, which means 320 of the 1000 patients died within 12 months. The manual selection model's sensitivity of 45% (at 20% cut off) means that 144 of the 320 patient who died were be classified as high risk by the model. Also from the 680 who did not die within a year, 592 are classified as low risk by the model (specificity) and 88 are classified as high risk. Predictive value positive and negative (PVP and PVN) present the percentage of the high-risk group who actually died (PVP), and the percentage of the low-risk group who survived (PVN). Predictive values change depending on the prevalence of the outcome. Mortality rate was 32% in this cohort; PVP and PVN of the model were 47% and 86%. It means that when the model applied to the 1000 patients, using the 20% cutoff, from the 200 patients classified high risk 106 died and 94 actually survived after a year. Also from 800 patients that are classified as low risk, 688 actually survived and 112 died.

Although the sensitivity and specificity of the multivariable models are better than the current

approaches, the overall sensitivity is still low which means that none of these models are very good

when used for screening older adults to identify high risk patients. However, altering the cutoff value to

classify more patients in the high-risk group increases the sensitivity. The predictive values of different

models do not differ vastly, although the PVN of the manual selection model is slightly better than the

other models.

Finally when the appropriate cutoff was determined, the model could be programmed and integrated

into the USMM database system. The high-risk patients that are identified based on the model, can be

flagged and brought to the attention of the providers for reevaluation and interventions if indicated.

- *Final model selection*

When comparing the alternative approaches in predicting 1-year mortality among this cohort, the

manually selected multivariable model has the highest c-statistic. Sensitivity and specificity of this model

can be optimized by changing the cutoff point that divides population to high and low-risk levels. Table

2.14 contains odds ratios and parameter estimates of the final predictive model for 1-year mortality

among this cohort of older adults. Examining the odds ratios and confidence intervals of the variables,

the strongest predictors of mortality are ADL-decline and low albumin. Both of these variables are

clinically essential indicators of the patient's global health. Albumin level can serve as a surrogate for

inflammatory status and also the nutritional status of a patient; decline in ADLs indicates functional

impairment. Low cholesterol was also associated with higher odds of death. Surprisingly, having a

history of hyperlipidemia showed a protective effect for mortality.

Increasing KPS, being dual eligible, and black race were all associated with a lower risk of death in this

population. Dual eligibility and black race both are more common in age groups <75 years old than

older. Thus the residual confounding can be the main reason for this relationship. The most important

predictors of death in this model are functional and nutritional indicators which are known clinically

relevant and robust components of general health status.

Table 2. 14. Final model parameter estimates and odds ratios for 1-year mortality using derivation dataset (N=3722)

Odds Ratio Estimates				Parameter estimates	
Predictor variables	Point Estimate	95% Wald Confidence Limits		Parameter estimate	P-value
ADL-decline, Decline vs. No-change	0.790	0.577	1.081	-0.2356	0.1407
ADL-decline, Improve vs. No-change	0.096	0.023	0.397	-2.3422	0.0012
Albumin, <3.2 vs 3.8+ g/dl	3.750	2.613	5.382	1.3218	<.0001
Albumin, 3.2-<3.5 vs 3.8+ g/dl	1.884	1.303	2.725	0.6336	0.0008
Albumin, 3.5-<3.8 vs 3.8+ g/dl	1.486	1.015	2.175	0.3959	0.0417
Race, Black vs. White	0.588	0.415	0.833	-0.5306	0.0028
Race, Other vs. White	0.442	0.197	0.991	-0.8156	0.0475
Surprise question, No vs. Yes	2.073	1.533	2.803	0.7289	<.0001
Cholesterol, <136 vs 195+ mg/dl	1.959	1.384	2.772	0.6724	0.0001
Cholesterol, 136-<164 vs 195+ mg/dl	1.191	0.839	1.690	0.1747	0.3285
Cholesterol, 164-<195 vs 195+ mg/dl	1.304	0.923	1.843	0.2658	0.1317
CCW-Hyperlipidemia Yes vs. No	0.531	0.417	0.676	-0.6334	<.0001
Age, 75-84 years vs. 65-74 years	1.711	1.180	2.481	0.5372	0.0046
Age, 85-94 years vs. 65-74 years	1.804	1.259	2.584	0.5898	0.0013
Age, 95+ years vs. 65-74 years	1.602	0.953	2.693	0.4712	0.0755
KPS, Severe vs. Moderate disability*	1.543	1.199	1.986	0.4340	0.0007
CCW-Depression, Yes vs. No	0.654	0.478	0.896	-0.4244	0.0082
Dual-eligibility, Yes vs. No	0.687	0.509	0.929	-0.3751	0.0146
Sex, Male vs. Female	1.151	0.886	1.497	0.1411	0.2917[†]

IQR: interquartile range; sd: standard deviation; S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; IADL: instrumental activities of daily living; TIA: transient ischemic attack; FU: follow-up; mg/dl: milligram per deciliter; g/dl: gram per deciliter; *KPS was included in the final model as a categorical variable based on the clinical application of KPS value;
[†]Sex was included in the final logistic model, although the Wald test for its coefficient was not statistically significant;

- *Calibration plots*

Calibration plots were generated for the final multivariable model applied to the validation dataset using

the two methods, loess-based and decile-based methods. Figures 2.6 and 2.7 display the loess-based

and decile-based calibration plots. Both plots show a small deviation from the 45 degree (diagonal) line.

The diagonal line indicates the prediction models. The maximum deviation is around the predicted

probability of 0.25, while in the lower and higher probabilities the model shows a better fit to the data.

The calibration plot in the derivation data was also generated as a reference. Figure 2.8 shows the

decile-based calibration plot generated from the derivation data where the calibration curve aligns very

well with the diagonal line (almost perfect prediction). The prediction model in the validation data

slightly underestimates the probability of 1-year mortality, especially in the deciles representing an

intermediate range of death (observed risk of 0.2-0.4).

Figure 2. 6. Loess-based calibration plot for the multivariable logistic model in the validation data for the outcome of 1-year mortality
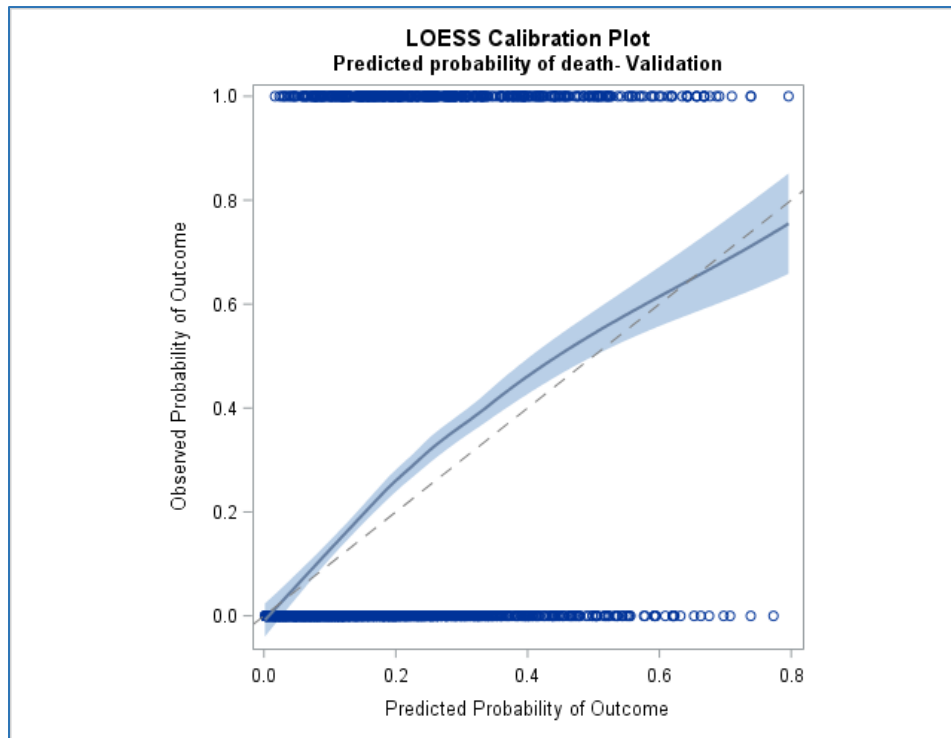
Figure 2. 7. Decile-based calibration plot for the multivariable logistic model in the validation data for the outcome of 1-year mortality
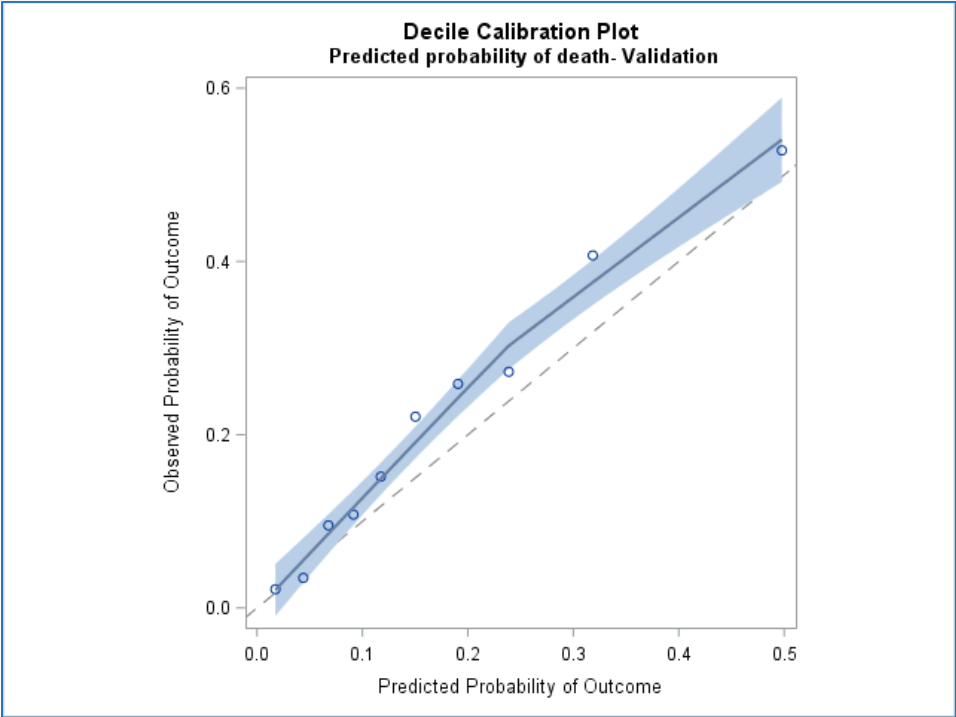
Figure 2. 8. Decile-based calibration plot for the multivariable logistic model in the derivation data for the outcome of 1-year mortality
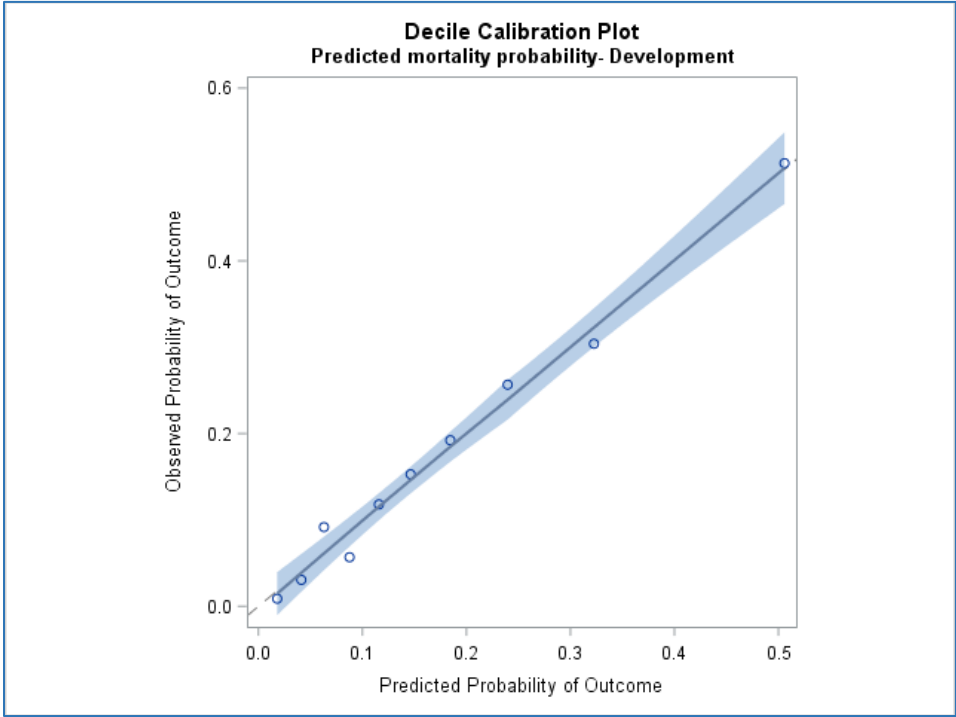
I also generated Hosmer-Lemeshow goodness of fit test statistics. The test evaluates the lack of fit of the model; a small p-value indicates a lack of fit. The Hosmer-Lemeshow goodness of fit test for the final multivariable logistic model resulted in p-values of 0.372 and <0.0001 for the derivation and validation datasets, respectively. This means a lack of fit cannot be rejected for the model in the validation data. This result is consistent with the calibration plots where the prediction model underestimated the probability of events compared to the observed events, especially for the lower probabilities of death.

o   <u>Outcome: Hospice admission</u>

In the following section I show the results for the outcome of hospice using the same modeling strategy used above for mortality.

Hospice admission in this cohort was defined according to the date of the first hospice service documented in the claims data. A total of 1391 (18.7%) patients were admitted to the hospice over the follow-up time. The Hospice admission rate within a year from the first visit was 10% among this cohort. The death occurred within six months of admission in 492 (65%) of those admitted to hospice within 12 months of their first visit. From all 1124 hospice deaths, 68% happened in the first three months of admission. Overall 2221 deaths (66% of all deaths) in this cohort occurred without hospice.

-   *Available data analysis*

The same modeling approaches were utilized as for mortality outcome. Independent variables were also the same as for mortality. Automatic variable selection methods (stepwise, forward, backward, adaptive lasso, and elastic net), and manual selection methods were applied. Using these selection methods, each model was developed in the derivation and applied to the validation datasets. The area under the ROC and Brier score were generated for each model. The results are provided in Table 2.15.

Table 2. 15. Model development using alternative variable selection methods for hospice admission using available case data

AUC and 95% confidence limits for both derivation and validation data sets (N=3722 derivation and 3723 validation)					
Variable selection	**N analyzed* Derivation**	**AUC Derivation**	**AUC Validation**	**Brier Score Validation**	**Selected variables in the final model**
Automatic variable selection methods					
Stepwise	2055	0.7819 (0.7502 - 0.8137)	0.6981 (0.6699- 0.7262)	0.0886	4 variables: age, dual-eligible, SQ, KPS
Forward	2055	0.8091 (0.7795 - 0.8387)	0.7272 (0.6976- 0.7568)	0.0874	9 variables: age, race, dual-eligible, SQ, lives alone, KPS, ADL decline, cataract, Heart failure
Backward	2055	0.7962 (0.7648 - 0.8276)	0.7295 (0.7006- 0.7585)	0.0858	8 variables: age, race, dual-eligible, SQ, lives alone, KPS, Heart failure, number of lab tests
Adaptive lasso[†] (validation data, Gamma=1.0)	2199	0.8173 (0.7881- 0.8465)	0.7440 (0.7101- 0.7779)	0.0764	18 variables: age, race, dual-eligible, SQ, TUG, ADL-decline, KPS, albumin, number of meds, number of labs, pressure-ulcer, cataract, osteoporosis, RA/OA, Hyperlipidemia, hypertension, hip fracture, diagnosis-count
Adaptive Lasso[†] (4-fold CV Gamma=0.1)	2229	0.8036 (0.7737- 0.8335)	0.7276 (0.6939- 0.7614)	0.0776	17 variables: age, race, dual-eligible, SQ, TUG, KPS, living-alone, albumin, number of meds, number of labs, pressure-ulcer, cataract, hip fracture, hyperlipidemia, hypertension, HF, AF
Elastic Net[†] (validation data)	2191	0.8181 (0.7891- 0.8471)	0.7339 (0.7003- 0.7674)	0.0779	18 variables: age, race, dual-eligible, SQ, lives alone, TUG, ADL-decline, KPS, albumin, number of meds, number of labs, pressure ulcer, AF, HF, cataract, Hyperlipidemia, hypertension, hip fracture

Table 2. 15. (cont'd)					
Elastic Net[†] (4-fold CV)	2081	0.8241 (0.7947-0.8536)	0.7313 (0.6956-0.7671)	0.0772	20 variables: age, race, dual-eligible, SQ, TUG, ADL-decline, KPS, lives-alone, albumin, cholesterol, number of meds, number of labs, pressure-ulcer, AF, HF, cataract, Hyperlipidemia, hypertension, hip fracture, colorectal cancer
Manual variable selection					
Full model	2055	0.8276 (0.7983-0.8569)	0.7090 (0.6709-0.7471)	0.0783	All 41 variables No variable selection
Manual selection-Available data	2601	0.7749 (0.7473-0.8026)	0.7351 (0.7055-0.7646)	0.0864	7 variables: age, race, dual-eligible, SQ, KPS, ADL decline Forced to the model: sex
Model developed in imputed data and applied to the available data					
Manual selection-Imputed data	3051	0.7602 (0.7335-0.7868)	0.7090 (0.6803-0.7376)	0.0877	6 variables: age, race, dual-eligible, SQ, KPS, ADL decline
S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; AF: atrial fibrillation; HF: heart failure; RA/OA: rheumatoid arthritis/osteoarthritis; *The numbers are different because first the variable selection was done in and then variables included in Proc LOGISTIC to generate AUCs, so not all the variables included in the final model and not all the missing observations were excluded. [†]Adaptive lasso and elastic net methods were conducted using two methods of validation and weighting parameters (Table 2.16)					

Overall, the alternative variable selection methods resulted in comparable c-statistic and Brier score.

Larger c-statistic and smaller Brier score both indicate a better fit of the model. To set the maximum

Brier score for a non-informative model, the incidence rate of the outcome in the validation data (10%)

was used and the maximum value of 0.25 was calculated as the Brier score for the non-informative

model. The largest AUC was seen in adaptive lasso (0.7440) and manual (0.7351) selection methods.

However, the manually-selected model included fewer variables than the adaptive lasso model (7 vs. 18

variables). Brier score was slightly better in the adaptive lasso method indicating that the adaptive lasso

model fits the data better than the other models, although the difference between the models is small.

Similarly the gain in the AUC when the adaptive lasso was used for variable selection was tiny and of no practical importance, however the model is larger with twice the number of predictors compared to the manual selection model. There are a few variables that were consistently selected regardless of the selection method, including age, dual eligibility, SQ, and KPS. Increasing age and answer 'No' to SQ were associated with higher hospice admission, whereas being dual-eligible and higher KPS decreased the risk of hospice admission, as it was for the mortality outcome. It is related to the residual confounding of age, but it also can be due to selection bias. The dual-eligible group might be somehow different from the other patients, so the unobserved variables in this group can cause a lower rate of outcomes. Variables race and ADL decline were also often selected. Being of the race black was associated with lower outcome compared to the white race, which in part is related to the residual confounding of age. As expected, the functional status of patients is an important predictor of hospice admission. Surprise question that indicates the physician's assessment of the patient's prognosis was also a good predictor of the hospice referral. Interestingly, sex was not a predictor factor of hospice admission. Nutritional status indicators (albumin and cholesterol) were not significant predictors of hospice admission, although they were essential predictors of mortality.

Similar to what was done for the mortality analysis, weighting parameter gamma was selected for the adaptive lasso method by testing different levels of gamma and two validation options (i.e. k-fold cross-validation or validation data) in the PROC GLMSELECT. The results are reported in Table 2.16. The number of effects is the total number of variables that selected including each level of classification variables as a dummy variable.

Table 2. 16. Using different gamma in adaptive lasso variable selection for 1-year hospice admission

Selection	# of selected variables	Optimal model criteria[†]
Validation	*Number of predictors*	*ASE*
Gamma=0	21 effects	0.0759
Gamma=0.1	21 effects	0.0758
Gamma=0.3	20 effects	0.0757
Gamma=0.5	24 effects	0.0755
Gamma=0.7	23 effects	0.0754
Gamma=0.9	22 effects	0.0754
Gamma=1.0*	21 effects	0.0754
4-fold CV	*Number of predictors*	*CV PRESS*
Gamma=0	18 effects	0.0717
Gamma=0.1*	19 effects	0.0717
Gamma=0.3	15 effects	0.0718
Gamma=0.5	20 effects	0.0718
Gamma=0.7	9 effects	0.0719
Gamma=0.9	8 effects	0.0720
Gamma=1.0	8 effects	0.0720
*Selected gamma based on the criteria and the number of variables; [†]Average square error (ASE) and CV PRESS are error measures that represent the goodness of model fit;		

The optimal gamma for each one of the two methods was chosen in a way to minimize the deviation from the true outcome (i.e. ASE or CV PRESS). Selected gammas were used to generate the adaptive lasso results that are shown in Table 2.16.

Figures 2.9 and 2.10 demonstrate the process of adding and removing variables using adaptive lasso and elastic net variable selection methods, respectively. The lower panel in each figure shows the average squared error of each model. It visualizes the lowest ASE of a model that can be correlated to the predictors of the model in the top panel. Figure 2.9 displays that with using the adaptive lasso selection, the optimal ASE was obtained at step 21 which was associated with the lowest ASE value of 0.075. The corresponding model consists of 18 variables. Similarly Figure 2.10 shows the optimal criteria for the elastic net selection method. The same graphical results can be generated for both methods when using 4-fold cross-validation in the selection option for the model statement.

Figure 2. 9. Adaptive lasso variable selection process using GLMSELECT for the hospice admission outcome (gamma=1.0 and validation dataset)
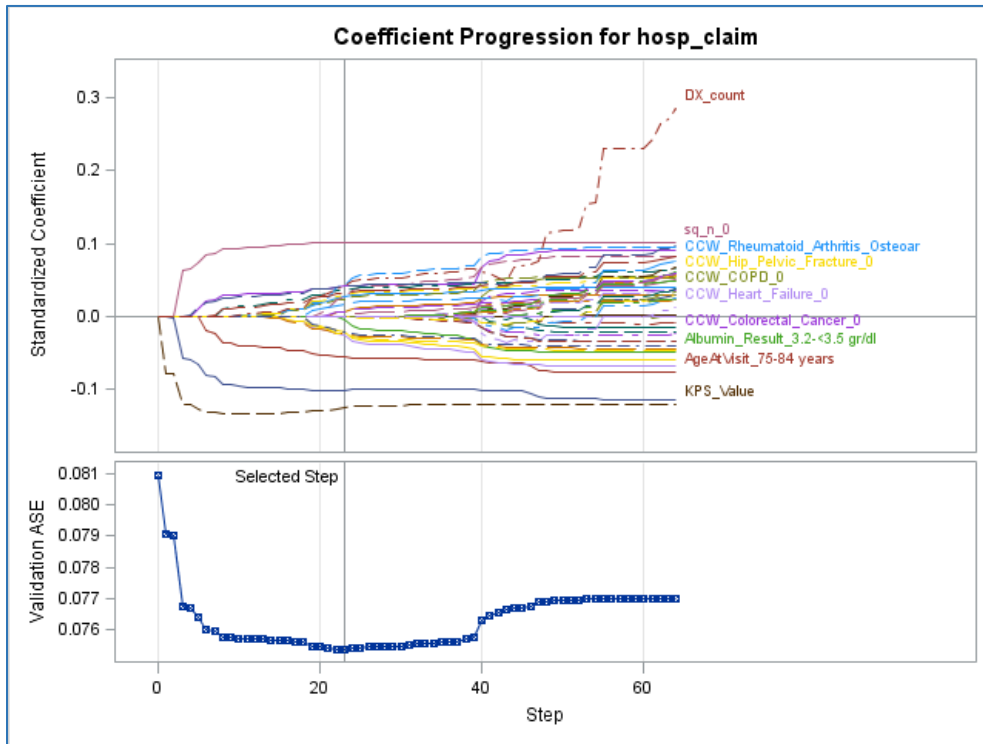
Figure 2. 10. Elastic net variable selection process using GLMSELECT for the hospice admission outcome (validation dataset)
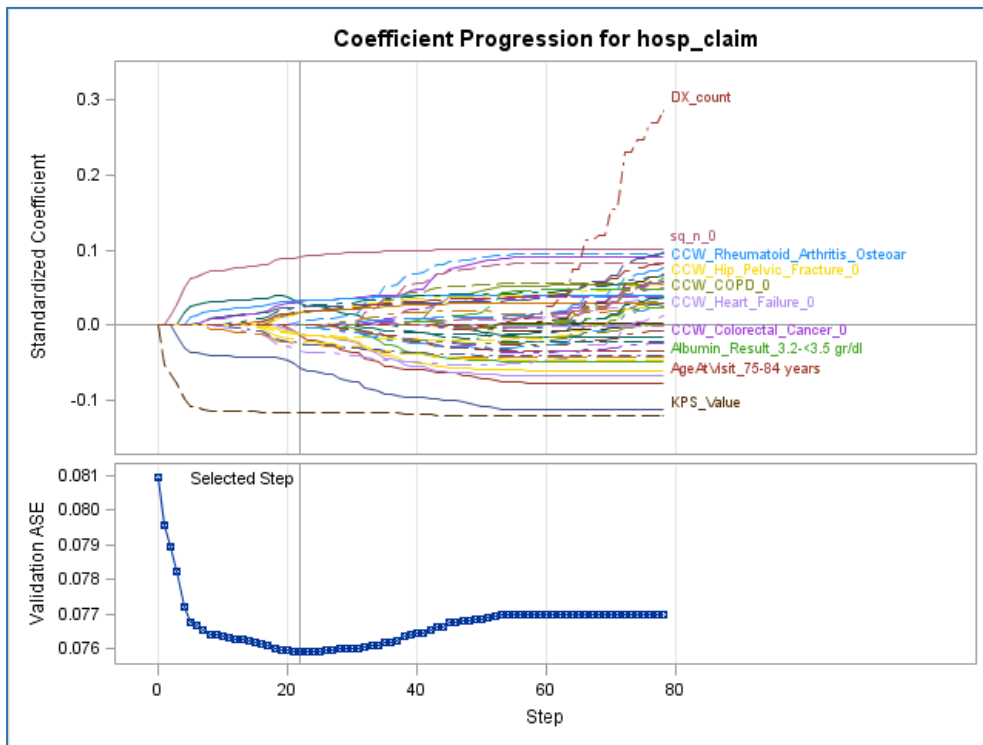
- *Imputed data analysis*

With the same considerations as in the mortality analysis, the multiple imputation procedure with 20 imputations was performed for the hospice admission analysis. The same modeling approach was applied as used for 1-year mortality. The c-statistic from different models was generated for derivation and validation data and provided in Table 2.17. Also, the variables that were selected in manual selection- available case analysis (Table 2.15) was applied to the imputed data set and AUC was generated for the imputed validation data.

Table 2. 17. Model development using alternative variable selection methods for 1-year hospice admission using imputed data, AUC and 95% confidence limits for both derivation and validation data sets

Variable selection	Derivation AUC (N=3722)	Validation AUC (N=3723)	Selected variables *
Automatic selection			
Stepwise	0.7359	0.7001	15 variables: Age, dual-eligibility, SQ, KPS, ADL decline, albumin, cholesterol, anemia, CKD, hyperlipidemia, pressure ulcer, cancer, number of meds, number lab tests, dx-count
Forward	0.7373	0.6992	6 variables: Age, dual-eligibility, SQ, KPS, AF, number lab tests
Backward	0.7339	0.6885	9 variables: Age, dual-eligibility, SQ, KPS, AF, depression, heart failure, number lab tests, dx-count
Manual selection			
Manual variable selection- Imputed data[†]	0.7227	0.7027	6 variables: Age, dual-eligibility, SQ, KPS, ADL decline, number of lab tests
Manual variable selection- (from available case analysis) [‡]	0.7204	0.6934	7 variables: Age, race, dual-eligible, SQ, KPS, ADL decline Forced to the model: sex

S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; AF: atrial fibrillation; HF: heart failure; RA/OA: rheumatoid arthritis/osteoarthritis;
*Variables that are selected in all 20 imputations;
[†] Includes variables that were selected ≥ 15 times in all three methods (forward, backward, stepwise);
[‡] The model is presented in Table 2.15;

Overall the performance of the models developed in the imputed data did not show any improvement compared to the available case data models. In fact, the AUCs of different models in the imputed data were generally smaller than those in the available case data.

Because the focus of this study is to develop prediction models, the primary performance measure for comparing the models is discrimination as measured by AUC. Sensitivity and specificity were also reported in Table 2.17 as another measure for comparing the alternative models.

- *Comparison of the risk stratification models*

The models that were developed manually from available case data, and imputed data were compared with the two alternative approaches: SQ only, and USMM 3-level risk stratification. Considering the AUCs for each model, the manually selected model in the available case data has the best discrimination. Also, sensitivity and specificity of the model at the cutoff point of top 20% high probability, are the best combination. However, the selection of the cutoff point depends on the cost and saving of false positive and false negative cases for the provider system.

Table 2. 18. Comparison of the alternative risk stratification approaches for 1-year hospice admission

Model	N validation (total=3723)	AUC validation*	High-risk group and prevalence	Sensitivity	Specificity
SQ only	3227	0.5895 (0.5633- 0.6157)	Answer "No" 14%	32.4%	85.5%
Current USMM model	3043	0.5875[†] (0.5591- 0.6158)	Level 5 7%	12.4%	91.7%
			Level 4 and 5 17%	35.5%	81.4%
Manual selection	2590	0.7351 (0.7055- 0.7646)	Top 10%	25.9%	91.8%
			Top 20%	45.9%	81.4%
Manual selection- Imputed data[‡]	3070	0.7090 (0.6803- 0.7376)	Top 10%	25.6%	91.8%
			Top 20%	42.4%	79.9%
*To make comparable results, the AUC for SQ model and USMM model was also generated from the validation data; [†]USMM risk level included in the model as a 3-level predictor for AUC calculation; [‡] The model was developed in the imputed, applied to the available case;					

- *Final model selection*

Comparing the alternative variable selection methods in this study, the manually selected model shows the best results in the discrimination ability of the model while being a parsimonious model. Penalized selection methods such as adaptive lasso and elastic net, had comparable c-statistics, although the number of variables in these models is much larger than in manually selected models. Table 2.19 shows the parameter estimates and odds ratios resulted from the manually-selected model.

Considering the significance and magnitude of the Odds ratios of the different variables in the final model, ADL decline is the most informative predictor of the hospice admission. Similar to the mortality prediction, functional status is the main predictor of need to hospice care. Age and dual-eligibility for Medicare and Medicaid are the following important variables. Surprisingly, having dual eligibility decreases the probability of hospice admission. Unlike the results of mortality outcome, nutritional status indicators (albumin and cholesterol) did not play a role in the prediction of hospice admission.

Table 2. 19. Final model parameter estimates and odds ratios for 1-year hospice admission using derivation data set (N=3722)

Odds Ratio Estimates				Parameter estimates	
Predictor variables	Point Estimate	95% Wald Confidence Limits		Parameter estimate	P-value
ADL-decline, Decline vs. No change	1.034	0.735	1.454	0.0335	0.8473
ADL-decline, Improve vs. No change	0.086	0.012	0.628	-2.4477	0.0155
Age, 75-84 years vs. 65-74 years	2.392	1.386	4.127	0.8720	0.0017
Age, 85-94 years vs. 65-74 years	3.345	1.986	5.633	1.2073	<.0001
Age, 95+ years vs. 65-74 years	3.870	2.055	7.286	1.3531	<.0001
KPS, Severe vs. moderate disability*	3.125	2.239	4.361	1.1393	<.0001
Surprise question, No vs. Yes	2.131	1.547	2.934	0.7566	<.0001
Dual-eligibility, No vs. Yes	2.023	1.336	3.064	0.7045	0.0009
Race, Black vs. White	0.654	0.426	1.004	-0.4242	0.0522
Race, Other vs. White	0.547	0.213	1.404	-0.6035	0.2095
Sex, Male vs. Female	1.165	0.866	1.567	0.1525	0.3137[†]

*KPS was included in the final model as categories based on the clinical application of KPS value;
[†]Sex was included in the final logistic model, although the Wald test for its coefficient was not statistically significant;

- *Discussion*

The study population- The USMM patient population is a unique population in terms of demographics and functional status. They are older and sicker and have more comorbidity and disability than most other similar study population that referred to in the background section. At the same time these patients remain community-dwelling and so are different from institutionalized, or hospitalized patients. They are home-bound by the CMS definition (5) which means the patient needs the help of another person or medical equipment such as a walker, or a wheelchair to leave their home, or their doctor believes that the patient's health or illness could get worse if they leave their home.(111) Therefore many of the previously developed prognostic indices are not applicable to this population. (17–19, 21–24) For example indices that include physical activities such as walking several blocks are probably not as relevant in this population as in a healthier older population.(49) Also because this population is different from those who are institutionalized, the prognostic models developed in hospitalized or nursing home setting may not be as accurate in this population.(62) The most similar population to the USMM population is the PACE participants who are nursing home, community living older adults. However the mortality rate in PACE cohort studied by Carey et al., (1-year mortality 13%) is much lower than the USMM population. They have a high frequency of comorbidities and multi-comorbidity compared to the similar studies in community-dwelling older patients. (7,49,52,57) These older adults Functional status measures including TUG (45% non-ambulatory or >30 seconds) and KPS (54% with a severe need to assistance) indicate a high prevalence of impaired function which makes this group of patients especially vulnerable and prone to the adverse outcomes. The one-year mortality rate of 32% in this cohort is much higher than the rate of comparable studies which ranged between 9% and 13% one-year mortality.(52,57) The high mortality rate in this cohort is comparable to the mortality rate of nursing home population which has been reported between 17-35% in different studies. (59–61) It indicates that the USMM population are similar to the nursing home residents in terms of mortality and

other adverse outcomes. The study reported by Carey et al., used the PACE patient population which by definition are nursing home eligible, however the one-year mortality rate was 13%.(57) As discussed in chapter one, the lower mortality rate of this population may be explained by the fact that PACE patients are adults aged 55 years and older who need the nursing-home level of care. So the PACE population might include younger adults with disabilities that made them eligible for long term care but did not necessarily increase their risk of 1-year mortality. These unique characteristics of the USMM population make them prone to adverse outcomes and denote a need for a risk stratification approach to be developed specifically for this population. Additionally, other RS indices often involved variables that are not available in this population. For example level of income, detailed information on dependency in functional status (e.g., bathing, dressing, etc.) are not available in the USMM data. This model has been developed to be used in the USMM patient population, however it could be tested for use in other similar populations.

Important predictor variables- We tested all available predictor variables including demographics, socioeconomic status, comorbidity, functional, laboratory tests and other variables such as surprise question, smoking status, number of lab tests ordered, and the number of medications. However, we should note that some of the previously known predictors of the adverse outcomes (such as number of hospitalization in the past year, decline in IADLs, recent fall) were not available for the analysis in this cohort of the USMM data.

ADL-decline, age, race, SQ, KPS, and dual-eligibility were important predictors of both outcomes- mortality and hospice admission. Functional impairment in ADLs have been shown to be predictors of adverse outcomes in hospitalized older adults. (112,113) Our findings is consistent with previous studies in showing that ADL-decline is an important predictor of death and hospice admission in the USMM patient population.

Examining the parameter estimates and p-values of final model, ADL-decline, serum albumin and cholesterol were the strongest predictors of mortality outcome in this cohort. This information can be useful in designing the interventions that focus on the nutrition, inflammatory status, and functional empowerment of older adults when they are at lower risk of mortality. Also, these findings can help USMM modify their policies for the timing and frequency of lab tests and assessments of patient function. Interestingly, serum albumin and cholesterol values were not selected in the final model for the hospice outcome, but ADL-decline was an important predictor of hospice outcome. This observation suggests that impaired functional status is more likely to be a reason for hospice referral than the biochemical laboratory tests.

Although low levels of albumin and cholesterol were associated with higher mortality rate, surprisingly, having a history of hyperlipidemia correlated with lower mortality rate in this cohort. It can say that the low cholesterol level in this population of older and/or frail patients represents a worse global health status than a history of hyperlipidemia which might be mild or treated appropriately. Specifically, statins as a lipid-lowering class of medication are proved to increase survival in CVD patients.(114,115) Low cholesterol level can show a poor nutritional status either due to poor general health or due to an underlying disease. (116)

Surprisingly, history of depression in this cohort also had a protective effect on the mortality outcome. Unobserved confounders may be an explanation for this observation. Also dual eligibility is associated with lower rates of death and hospice admission.

Variable selection methods- We tested and compared different methods of variable selection. We applied commonly used automated methods such as stepwise, backward and forward selection as well as more advanced methods including adaptive lasso and elastic net; however the advanced methods did not show superiority over the conventional selection methods in this dataset. One of the main benefits of the advanced penalized selection methods (like adaptive lasso and elastic net selection) is when faced

with high dimensional data sets with numerous predictors and a relatively small number of observations. Another advantage of these methods is in the datasets with highly correlated predictor variables.(102) However, none of the two conditions were present in our dataset.

Importance of missing data- We evaluated the association of missingness on several variables with the outcomes. Variables that had less than 20% missing observation were analyzed in the univariate and multivariable analysis for model development. We found a strong association between missingness in these variables (race, TUG, SQ, ADL, living alone, albumin, and cholesterol) and mortality (Table 2.6). A possible reason is that the patients who died were too sick at the time of the visit to be interviewed and evaluated thoroughly and so some of the variables were left missing. To assess the impact of missingness on the variable selection, we applied a multiple-imputation approach and repeated model development. We used different variable selection methods in the imputed data, however the models' performance in the imputed data were not as good as in the available case data. As mentioned before, the assumption in the multiple imputation is missing at random while in the USMM data there are evidences that suggest missing is not at random. This can explain why the model's performance is worse in the imputed data than available data. Variable selection in the imputed data resulted in a larger number of variables selected (15 vs. 9) but only a slight improvement in discrimination compared to the available case data. Also, when the selected model developed from the imputed data was applied to the available case data, the AUC was actually slightly lower relative to the AUC of the final manually selected model developed in the available case dataset (0.755 vs. 0.763).

Application of the developed model- Comparing the results of this study to the approaches that are currently practiced by the USMM providers, showed the superiority of our multivariable models – regardless of their exact specification. Using the two multivariable logistic models, the AUCs for mortality and hospice admission were substantially higher (0.763 and 0.735) compared to the two alternative approaches, USMM proposed 3-level risk model (0.599 and 0.588), and surprise question

only model (0.555 and 0.589). Comparing different cutoff points for our model and the USMM proposed

3-level model, sensitivity and specificity estimates of our model is similar or higher than the current

model for both outcomes. Consequently for any cutoff point this model can help providers to better

manage the cost of services and patients benefits by reducing the number of false positives or false

negatives. The optimal cutoff point can be selected by providers and policymakers based on the cost of

misclassification of cases. If the cost of misclassification of a high-risk patient is higher than the

misclassification of a low-risk patient then the more relaxed cutoff point (top 20% instead of top 10% of

probability) is appropriate and vice versa. The cost and benefit of services are dependent to different

factors, such as patient's benefit from receiving or harm from neglecting a given service, cost of that

service for a provider and the insurance companies, alternative options for that service, and available

resources. Thus cutoff point should be explicitly selected by the USMM providers.

Furthermore, the population can be grouped in more than two risk levels when needed; especially if

different levels of services are available for example palliative care, hospice referral, home health care,

and use of preventive services. This approach can also be useful in a clinical setting, for example

prediction of 1-year mortality risk can be beneficial to make the decision for offering a screening

procedure to an older adult patient. Our risk stratification model can be especially advantageous for

advanced care planning by identifying patients who are at high risk of mortality or hospice admission.

Considering the patients and caregivers goals, different services can be offered to patients at different

levels of risk.

To use the developed model in the practice, a statistical software (e.g., SAS, R) will be used to integrate

the final model into the USMM database. By programming the model into APRIMA, the model can be

run on all observation at each new data entry time; then a prediction probability of death and hospice

will be made for each patient. The patients will be stratified in different risk groups based on the

threshold that will be determined by the USMM providers. Finally high risk patients will be flagged and

brought to the attention of their physician for further assessment. Clinicians can discuss different

services with patient and caregiver to make the decision that is aligned with the patient's goal of

treatment.

      o   <u>Strengths</u>

This study has been conducted in a unique population of USMM patients. The database richness allows

us to use a broad array of potential predictor variables (e.g., demographics, clinical, functional status,

medical history, and lab tests) to develop the model, whereas in the studies that use billing data to build

a prognostic model these data are accessible. Moreover, in this study we used and compared several

alternative variable selection methods including new methods such as adaptive lasso and elastic net

variable selection. We also used multiple imputation to manage the missing data and evaluate the

impact of missing observations on the model selection.

      o   <u>Limitations</u>

Although the USMM database is relatively very rich data set in both quality and quantity of the

information collected from the patient population, there are still variables that were not usable due to

missing observations. Valuable information was lost on functional status including a decline in IADL

function since the last visit, a decline in global health since last year, falls, hospitalizations and ER events.

Another limitation of this research was the problem of missing data. One assumption of MI PROCEDURE

is that missing is at random. We confirmed that the missing mechanism is not MCAR (missing completely

at random). Although it is not possible to statically distinguish between MAR and MNAR, the strong

association between the missingness on predictors and the outcomes suggests the MNAR mechanism.

We used multiple-imputation in this data regardless of the MAR assumption. There were two

comorbidity variables excluded from the analysis because the number of patients with the comorbidity

was too small or zero. Finally, we validated our model using the validation data that is originated from

the USMM database. We did not use any external population to evaluate the accuracy of the model.

- *Conclusion*

We developed prognostic models for prediction of two adverse outcomes, mortality and hospice admission among the population of community-living home-bound older adults. Both models showed a significantly better performance than the current risk stratification approaches used in this population. These models also demonstrated comparable or better discrimination compared to the similar prognostic models published in the literature. These two models can be used for risk stratification among older adults in different settings (i.e., community-living, nursing home, rehabilitation centers, and hospice), also can be useful in other epidemiological studies to adjust for baseline risk factors among such population of older adults. Future studies are required to validate the models in external population. Furthermore other risk stratification models can be developed in this population trying to improve the prognostic models. Survival analysis of time to event and using machine learning techniques to reveal the possible nonlinear relationships between the outcome and predictors will be included in subsequent chapters.

- *Introduction*

The population is aging faster than any other time in history.(9,10) Increasing age is associated with a high prevalence of chronic diseases and multiple comorbid conditions, which often require long term care and frequent utilization of health care.(11,12) Health care expenditures are disproportionately higher in the older population than working-age patients.(14,15) The cost of health care for older adults imposes a considerable burden on the health systems and government through the Medicare program.(13) The increasing number of older adults and growing need for services among them along with limited resources necessitate the allocation of services to those who benefit the most. To align the appropriate levels of services with patients' needs, risk stratification methods are required. Using statistical methods, one can develop a prediction model for an outcome - such as 5-year mortality - and stratify patients based on their probability of that outcome. Then appropriate services can be allocated to each level of risk. Different risk stratification approaches have been developed for the prediction of different outcomes, including mortality, readmission, relapse, or complications of specific diseases. There are also risk stratification models developed among older populations with a specific condition (e.g., diabetes, cancer, cardiovascular diseases) or in a specific setting (e.g., emergency room, surgery, nursing home) often to predict mortality, readmission or complications.(19,46,117,118) There are also a few risk stratification models developed in the community living older populations regardless of any specific condition.(48,49,52,57,64) Some of these models were reviewed in the background section of chapter two. In chapter two, a logistic regression analysis was applied to develop a risk stratification model among a subset (derivation data) of a cohort of community living homebound older adults to predict the risk of mortality and hospice admission using a set of explanatory variables. The accuracy of the model was evaluated by using the area under the ROC curve or C-statistic, and the model was

validated using a validation subset (derived from the same database). In this chapter, I will use a

Machine Learning (ML) technique to develop a risk stratification model among the same population and

compare the performance of this model to the previous logistic model. Derivation data is often called

derivation data in ML algorithms, however to be consistent with other chapters, we used the same

terminology ('derivation' data) for the subset of data in which the model is developed.

In the next section, I provide a brief introduction to *the random forest method*, which is the machine

learning (ML) algorithm used in this chapter. It is followed by a literature review, methods and materials,

results, and discussion sections.

- ***Main concepts and definitions***

The development and use of big data are rapidly growing in medicine and public health, like many other

industries.(119) Traditional statistical methods may not be sufficient for the analysis of these big data.

The enormous sample size and high dimensionality of big data bring new statistical challenges, including

noise accumulation (i.e., too much unexplained variability within a data sample), spurious correlation,

incidental endogeneity (i.e., when the predictor variable is correlated with the error term), and

measurement errors. (120) Likewise, big medical data introduce problems such as multicollinearity (i.e.,

multiple correlation between predictors or independent variables), model complexity, the

computational cost to fit models, and model overfitting (i.e., decreased generalizability of the model).

(121) Machine learning algorithms are becoming popular in big data analysis and are increasingly used in

biomedical research as well. In the following sections some examples will be described.

- o Machine learning

Machine learning (ML) is an application of artificial intelligence that allows computers to automatically

learn and improve the algorithm without being explicitly programmed. The process of learning begins

with the input data; the algorithm searches patterns and makes predictions using an iterative approach

in order to improve future decisions. ML techniques include a wide range of statistical methods that can

be used to describe associations, search for patterns, and make predictions. ML algorithms are being

increasingly used in biomedical research. There are two main methods in ML: supervised and

unsupervised learning. Predicting an outcome based on a set of explanatory variables that are specified

by data scientists is referred to as supervised learning. Whereas unsupervised learning refers to the

exploration of associations or detection of patterns among variables regardless of a specific outcome.

There are numerous different ML algorithms; neural networks, random forests, Bayes net, and support

vector machines are a few examples. Random forests have been previously used in biomedical studies

for the development of prediction models.(65,119,122)

     o   <u>Machine learning in prediction models</u>

Prediction models are central in medicine and are utilized in everyday decision-making by physicians for

prediction of diagnostic or clinical outcomes.(22) These models are often used in medical research to

predict the outcome of a disease, result of a diagnostic test, the outcome of a new treatment,

complications of an illness, or survival of the patients. Risk prediction usually relies on parametric

regression methods, such as logistic regression or generalized linear model. However, new approaches

such as ML techniques have been introduced in epidemiologic studies as well as in many other medical

and non-medical disciplines.(74) Machine learning is often used without having any specific hypothesis

regarding the association of the variables or the pattern of the associations; thus it is an excellent

approach to explore the important predictors of an outcome in scenarios where they are many

explanatory variables. ML algorithms are specifically preferred when the number of explanatory

variables in the data is considerably larger compared to the number of observations- also referred to as

big data. (65,123) Several studies have compared different machine learning algorithms to develop

prediction models and determined that the random forest algorithms have better performance than

other machine learning approaches such as support vector machine (SVM) and Bayes net. (73,124–126)

- *Decision tree*

Decision trees (also known as classification and regression trees) are recognized as powerful tools for prediction models.(22,127) Recursive partitioning is the core idea in constructing a decision tree, and it involves dividing data set into subsets based on several independent variables or rules in order to correctly classify members of the dataset. Each tree is made of nodes, branches, and leaves. The structure of an example decision tree is shown in Figure 3.1.

Figure 3. 1. The schematic structure of a decision tree
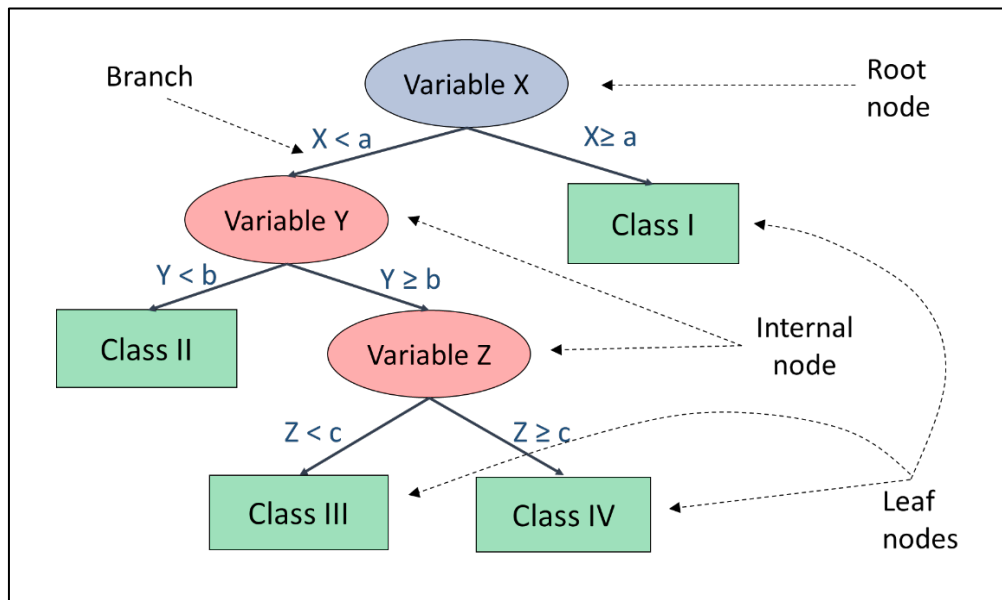
Variable X is the first variable that splits the study population the best; Variable Y is the next best variable;
Classis I-IV are the leaves and represents different classes or groups of risk predicted for the outcome of interest;
Branches are rules that connect a node to its child nodes (internal nodes and leaves) based on the value of a predictor variable;

Nodes are points of decision in a decision tree where a predictor variable splits the study population into subgroups (named child nodes) based on their observed data for that predictor. In other words, each node tests the data on an attribute (predictor). Branches are the outcomes of that attribute. The first node is called root node and uses or identifies?? the best predictor to split the cohort into two or more child nodes based on the optimal separation (maximum separation of the subgroups and minimum variability within subgroups in respect to the outcome).(22) Internal nodes (child nodes) then are split

again using the next best predictor at each node. The split will be repeated for each child node until the

algorithm reaches the final decisions or classification nodes were obtained, also called leaves.

- o Random forest

Random forest is a data mining algorithm first proposed by Leo Breiman in 2001. (68,69) It combines

several (to potentially many) decision trees (ranges from 10 to thousands) and generates predictions by

averaging over all the trees in the forest.(124) The term forest represents the numerous replications of

decision trees. Each decision tree is developed in a randomly selected subsample (bootstrap samples) of

the derivation population. Figure 3.2 displays the random forest algorithm, as presented in machine

learning literature.(68)

Bagging or bootstrap aggregation is a ML techniques for reducing the variance of an estimated

prediction. In general, when several bootstrapped samples of the original data are constructed, and

separate decision trees are trained in each subsample, averaging over these trees is referred to as

bagging or bootstrap aggregation. Random forest is a fundamental modification of bagging; in the

bagging technique, when constructing the decision trees, all predictors are searched at each split-point,

and the best splitting variable is selected. However in random forest method, building the tress involves

an additional step of randomly sampling predictors at each node. In other words the difference between

random forest and bagging is that in a random forest at each split-point of each tree, the optimal

splitting variable will be selected from a random subset of all predictors. This method minimizes the

correlation between trees and increases bagging accuracy. The number of predictors that are tested at

each split-point can be specified as a parameter in the model and is often calculated as m= Square root

(p) where m is the number of randomly selected predictors, and p is the number of all predictors. For

each bootstrap sample from the derivation dataset, there are samples left behind and not included in

the model construction. These samples are called out-of-bag (OOB) samples. The performance of each

model, when averaged over its OOB samples, is a good estimation of the model accuracy. The OOB

average square error is a good estimate of test error rate and is generated for each tree by default in SAS output.

Figure 3. 2. Random forest algorithm for regression and classification

---

Algorithm 15.1 *Random Forest for Regression or Classification.*

1. For $b = 1$ to B:

   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.

   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

      i. Select m variables at random from the p variables.

      ii. Pick the best variable/split-point among the m.

      iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

Regression: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{rf}^B(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B$.

Source: figure adapted from *'The Elements of Statistical Learning'* by Hastie, Tibshirani, and Friedman (68)

The bootstrap samples (Z) are randomly sampled with replacement in the derivation dataset. A decision tree is developed in each Z subsample using a randomly selected set of predictors. When random forest is built, to make a prediction for a new observation (x), the observation (e.g., a patient in our study) passes through all trees and the predictions from all trees are aggregated. When the outcome is interval (continuous), the prediction is the average of all trees. When the outcome is a classification variable, the average is determined by the majority vote, which means among all the trees, the class that has been predicted most often for observation (x), will be the RF prediction for this observation.

Random forests became popular during the past decade as a statistical method in many scientific fields.(69,128) RF-based methods are used for risk stratification and for identifying important variables among a large number of potential predictors. Parametric linear regression models are powerful statistical methods to explore the relationship between explanatory variables and the response in a linear fashion. These approaches generate a single model to fit the full data set. However, when the data has many explanatory variables with complex interactions, building a best single linear model can be very difficult. Random forests are proposed as an excellent alternative approach for datasets with a large number of predictors and the potential for complex interactions between them.(68,128) The structure of the decision tree permits a variable to be selected in multiple splitting nodes at different depth of the tree, also a single variable can split the nodes using different rules for different nodes. These specific structure in the RF gives it the strength to manage complex interactions. Also non linearity in the data, which requires the use of polynomial terms in parametric models, can be also handled in random forest.

Overfitting of a model often happens in a single decision tree when the tree grows deep. It means the model is too specified to the derivation data so that its generalizability to external data is weakened. Overfitted models often have poor performance when applied to validation data. Random forest overcomes this problem by averaging over hundreds of different de-correlated trees which prevents overfitting. Averaging over the trees also diminishes the sensitivity of the trees to the noise (meaningless data) so long as the trees are not correlated, and the use of bagging (bootstrap aggregation) in the random forest algorithm prevents correlation between the trees.(68)

Unlike the regression models, in construction of a random forest model, observations that are missing data on one or more independent variables are not excluded from the analysis. There are different ways to manage the missing data in random forest development. The default is that missing observations are

included as a separate category in the model building process. Therefore no observations are excluded from the analysis.

Machine learning methods, including the random forest approach, are designed to make the most accurate predictions possible, and they have demonstrated high predictive accuracy.(128) However, to gain this accuracy, random forest models do not output the same metrics as regression models. (129) A logistic regression model for example, provides beta coefficients (and odds ratios) that indicate the magnitude and direction of the predictor impact on the outcome. On the other hand, random forests do not provide conceptual equivalents to regression coefficients or measures of effect for each predictor. In fact, being non-linear, the sensitivity of a random forest model's output to the independent variable is not straight-forward to formulate. Instead, random forest models output ranked importance tables for the predictors. Importance of the predictors are ranked using different methods according to the frequency and order that each predictor was selected. Therefore, to compare random forest models with linear prediction models such as logistic regressions, discrimination metrics (i.e. AUC) and misclassification rate (defined as the fraction of study population that are misclassified– [false positive+ false negative]/total) are used as standard measures of model performance.

Compared to a single decision tree, random forests are more generalizable to new data, because the most influential predictors are selected by growing a large number of trees. However, unlike a single decision tree, the results of random forests are not interpretable as a defined set of decision nodes, rather as ranked importance of predictor variables. In other word, RF does not output a single tree that can be used manually to classify an observations based on the splitting nodes. Random forests are used to rank the variable importance; thus, one can identify the essential predictors but not the relationship between them.

The main strengths of the random forest approach can be summarized as below:

1. Suitable for nonlinear data (where not a single linear model appropriately fits to the data)

2. Avoids overfitting of the data, which is a drawback of a single decision tree

3. Provides the rank of importance of explanatory variables

4. Is robust to noise

5. Avoids exclusion of observations with partly missing data

6. Excellent for complex interaction and highly correlated data (69,128)

- *Random forest construction parameters*

In random forest models, both the outcome and explanatory variables can be categorical or quantitative (continuous). A random forest model can also handle missing observations on explanatory variables as legitimate values, so unlike logistic regression, the observations with partly missing data on explanatory variables are not excluded from the analysis. However, observations with the outcome missing will still be excluded in the RF method.

The number of trees in each forest and the depth that each tree grows can be specified at the model building phase. In addition to these two parameters, there are other optional parameters that can be specified in the model to control the characteristics of the RF; for example the minimum number of observations in the leaves can be specified. However the two main determinant of the model are number of trees and the depth of trees. An increasing number of trees increases the accuracy of the estimated probability because it represents the averaging over all the trees. The depth of a tree determines how deep the model can develop, which means how many times a tree can be split sequentially in classification until reaching the final nodes (leaves). In other words, the depth of a tree indicates how much the model fits the data. As the depth of the forest increases, the leaves will have a smaller number of observations, and the model risks overfitting the derivation data. A common mistake is to develop shallow trees (i.e., with very few splitting levels such as 3 or 4 consecutive splits) in order to avoiding overfitting the model. In fact deep trees with some degrees of overfitting are preferred to

the shallow ones, because averaging over all trees prevents the problem of overfitting of the forest to

the derivation data.

- *Variable importance*

As explained previously, random forest models do not provide coefficient estimates for the explanatory

variables, consequently there is no P-values in the RF output to measure the significance of a predictor.

Therefore, there is a need for alternative measures to evaluate the significance of independent variables

in a RF model; this alternative measure is the ranked importance table of independent variables.

Importance of a variable is the contribution of it to the model success – where success is defined as the

accuracy of the predictions. Generally, prediction models rely mostly on a few predictor variables,

although they may include many independent variables; a good measure of importance is the one that

identifies those few essential predictors among all the predictors. Identifying the importance of

variables helps to understand the relationships between the predictors and the outcome better. Also,

random forest can serve as an initial step to select relatively influential predictors from a list of all

possible predictors. Then other model development strategies can use the predictors that are selected

in RF. (23) There are different measures to rank the importance of variables, for instance, Breiman's

method, loss reduction, Stroble's method, and random branch assignment (RBA). Breiman and Stroble

methods are computationally intensive and so often have a long running time. Loss reduction method is

less intensive, however it is biased towards the correlated predictors and so inflates the importance of

correlated variables at the expense of other independent variables. The importance of a variable using

the loss reduction method is proportional to the sum of the impurity measures, summed over all the

nodes that the variable is splitting. Impurity is often measured by Gini splitting criterion, so this method

is also named Gini impurity or Gini increase. Impurity represents how well the tree split the data. Gini

Impurity measures how often a randomly chosen subject from the derivation data set will be incorrectly

labeled (regarding the outcome) if it was randomly labeled according to the distribution of labels in the

93

dataset. '*Proc HPforest'* replaced the word '*impurity*' with '*loss*', to show the reduction in loss from using the model.

The random branch assignment (RBA) is the most recently developed method to rank the variables importance and it has advantages over other methods without their drawbacks. It was introduced in 2014 by Neville and Tan (131) and was claimed to satisfy the objectives of the previously developed methods (Breiman and Strobl) (128,132) while avoiding the problems of inflating the importance of correlated variables and intensive computation. Compared to loss reduction method of ranking variable importance, the RBA method diminishes the inflation of correlated variables and so results in the most accurate ranking of the predictors. The RBA method measures the importance of each variable by replacing the splitting rule with a randomized rule in the nodes that involve the variable. When the model is structured in the derivation data, the proportion of observations in each node are saved. When evaluating the variable importance in RBA, the observations that reach the node will be randomized to the branches with a probability proportional to the observed proportions in the derivation data. Then the model fit is compared to the fit statistics of the model with the variable included. The importance measure is proportional to the randomized fit minus the model fit without randomization. In this study, the table of ranked importance of predictors was generated using both loss reduction and RBA methods.

- *Literature review*

Machine learning literature has been expanding in the past few decades. Also, biomedical researches have increasingly used ML techniques. There are studies that compared the use of machine learning methods and traditional parametric regressions and often have found that the performance of machine learning methods in risk prediction was superior to the parametric regression methods. Among Machine learning techniques, random forest has been used frequently in biomedical researches (65) because of its strengths that were summarized above. The overall goal of literature search in this chapter was to

find examples of studies that utilized ML techniques in the development of prediction models for adverse outcomes (specifically mortality) among a population of older adults.

Search Methods- Searching Pubmed for the word 'random forest' resulted in about 8500 hits (when using no other limits). Searching the words 'machine learning' and 'prediction' also resulted in the similar number of hits. Limiting the search words to 'random forest' and 'risk stratification' in the title or abstract of the paper, reduced the results to 50 hits. Searching for 'risk stratification', 'random forest', and 'mortality' found 7 results, none of them were relevant to the community living older adults. Adding 'older adult' or 'elderly' to the search resulted in no findings. Searching the three words 'random forest', 'risk stratification', and 'elderly' in all fields resulted in 34 findings. Searching the Google Scholar database with the words 'random forest', 'risk stratification', 'mortality', and 'elderly' resulted in more than 20000 hits. Reviewing the first 50 hits in Google scholar and the 34 findings in Pubmed, along with forward and backward reference searching of any relevant article, I found four studies that fit broadly into my original goal for the literature review i.e., to identify studies that utilized a random forest algorithm to develop a prediction model for an adverse outcome. I did not find any studies that are exactly comparable to this thesis topic in terms of prediction of mortality in community-living older adults.

Khalilia et al. (75) utilized the Healthcare Cost and Utilization Project (HCUP) data set to develop prediction models. They compared the performance of random forest and three other ML methods (Specifically SVM, bagging, and boosting) in predicting the risk of the following eight disease categories: breast cancer, diabetes with no complication, diabetes with complication, hypertension, coronary atherosclerosis, peripheral atherosclerosis, other circulatory diseases, and osteoporosis. Diseases categories are developed by HCUP and are based on a combination of diagnosis and procedure ICD codes. Random forest outperformed the other three models for prediction of seven out of eight disease categories when comparing AUCs.(75) Schneider et al. (133) studied mortality risk in acute cholangitis

95

patients. They developed eleven different risk prediction models, including logistic regression with

stepwise variable selection, generalized linear models with lasso penalties, and random forest model.

They found that the random forest model had the best predictive performance (AUC=0.92). Weng et al.

(77) studied the performance of 4 Machine learning techniques to predict cardiovascular risk and

compared the performance to the American College of Cardiology guidelines for prediction of the 10-

year cardiovascular event. (134,135) They concluded that the performance of machine learning

algorithm was better than the established approach which was based on the prediction of the risk of

future CVD based on the well-known risk factors such as hypertension, cholesterol, diabetes, and

smoking (coefficients from a proportional hazard model). Chong et al. (76) compared the performance

of the machine learning approach and multivariable logistic regression in prediction of the diagnosis for

pediatric traumatic brain injury in the emergency room patients. Their results demonstrated that the

machine learning model had better AUC (AUC=0.98 vs. 0.93), sensitivity, specificity, and predictive

values than the logistic regression model. Rose et al. (74) developed a super learner algorithm to predict

mortality in a population of adults 54 years and older. The super learner is an ensembling machine

learning approach that combines multiple machine learning algorithms into a single algorithm and gives

the prediction with the best (lowest) mean square error. They demonstrated that this super learner

algorithm had better performance than every single algorithm. Peng at al. (136) studied the

performance of random forest in prediction of 30-day mortality in patients diagnosed with spontaneous

intracerebral hemorrhage (ICH). They found that RF model (AUC=0.87) outperformed other models

including logistic regression model (AUC=0.78), artificial neural network algorithm (AUC=0.81), support

vector machine (AUC=0.79), and ICH score (AUC=0.72). (136)

In chapter two of this dissertation, a logistic regression model was developed for risk stratification in a

cohort of USMM patients. About one-third of the observations were excluded from the logistic

regression model due to missing data on one or more explanatory variables. Multiple imputation

approach was applied to overcome the missing data issue. However, the model developed in the

imputed data did not improve the predictive performance (AUC); in fact, it was a slight decrease in the

AUC of the model for 1-year mortality when applied to the imputed data (AUC in imputed data=0.75 vs.

AUC in available data= 0.76). In this chapter, a random forest algorithm is used to develop a risk

stratification model with the intent of improvement in the model performance. My hypothesis is that a

random forest model will have a better performance than the logistic regression model because unlike

the logistic model, random forest:

1. handles the missing data and so uses many more observations,

2. accounts for the potential non-linearity in the relationships between the explanatory and/or

outcome variables,

3. able to fit data with complex interactions (which is common in the biomedical data).

- ***Methods and materials***

Data source- we developed the random forest model utilizing the same dataset and the same study

population that were used in Chapter two.

Study population- the 2015 cohort was defined as all patients who had their first ever medical visit by a

USMM provider between January 1st and December 31st, 2015. The USMM data was linked to the claims

data, and those patients who had claims data were included. To have comparable results to the logistic

regression model (Chapter 2) the cohort was limited to the patients who have been followed up for at

least 365 days or who had an outcome (death or hospice admission) within a year of their first USMM

visit. Figure 2.1 displays the flow diagram of the study population.

Outcome- 1-year mortality was determined if a date of death was recorded in the claims data within 12

months of the first USMM visit. Likewise, 1-year hospice admission was determined according to the

recorded date of first hospice service in the claims data. Claims data was processed data and included

the intervals of hospice or home-health services (2 weeks period). Therefore the first date of earliest

hospice service was considered as the date of the outcome. If death happened in hospice, both

outcomes (death and hospice admission) were analyzed in the respective analysis.

Exposure- Variables with less than 20% missing observations were considered for the analysis. These

data were collected from the baseline visit for each patient. Random forest model can handle missing

values on explanatory variables; however, I limited the independent variables to those with less than

20% missing to have a comparable data set for both random forest and logistic regression models.

A total number of 41 variables (the same as Chapter two) were analyzed as predictors, including

demographics: age, gender, race; socioeconomic status: insurance status representing if a patient has

dual eligibility for both Medicaid and Medicare, living alone, smoking; functional status: functional

decline in ADLs, timed up and go (TUG), Karnofsky performance scale (KPS value); lab tests: serum

albumin, cholesterol; and other variables: having a pressure ulcer, surprise question answer, number of

medications, and number of lab test ordered by the provider. There are 24 medical history variables

(CCW variables) as listed in Chapter two. These 41 variables are the same predictor variables that were

used in chapter 2.

     o   Statistical analysis

The analyses for this study was done using SAS software (SAS Institute Inc., Cary, NC, version 9.4). The

data were randomly split into two equal size cohorts, derivation and validation. The two main SAS

procedures used for random forest modeling are '*Proc HPforest'* and *'Proc HP4score'*. '*HPforest'*

procedure generates random forest in the derivation data. The number of trees and the maximum

depth are specified in this procedure. The*'HP4score'* procedure is used for both scoring the validation

dataset (*'score'* statement) and ranking the variable's importance (*'importance'* statement). Scoring the

validation dataset means applying the model that was developed in the derivation dataset to the

validation dataset. The model is applied to all observations (even those with missing data), and the

prediction is generated for each. The two statements, *'score'* and *'importance'* cannot be specified at

the same time in '*proc* HP4score'. Therefore the *'HP4score'* procedure is separately specified for each of

the two statements.

In this analysis, the random forest model was developed in the derivation data using '*proc HPforest'*. The

developed model was then applied to the validation data using '*Proc HP4score'*. The predicted

probability of the outcome is computed individually for each patient. Receiver operating curve (ROC)

and area under the curve (AUC) were generated as an indicator of discrimination of the model in both

derivation and validation data sets.

A random forest model has two main parameters that can be specified in the model development

phase, the number of trees (MAXTREES) and the depth of them (MAXDEPTH). The number of trees

determines how many decision trees at maximum are developed in the forest. The default number is

100 trees. The MAXDEPTH option specifies the maximum depth of a node in each tree of the forest. It is

the number of splitting rules needed to define a node. Therefore the root node has a depth of 0, and the

children of the root node have a depth of 1 and so on. The default depth is 20. (130) To find the optimal

number of trees and depth of them in random forest analysis, the developed model was repeated with

different numbers (1, 10, 50, 100, and 200) and depths (2, 10, 20, and 30) of trees and the ROC and

respecting AUC for each model were generated. The number of trees and depth that result in the

highest AUC were selected as the model parameters. In derivation data set, increasing the depth and

the number of trees always increases the AUC, however, when the model is applied to the validation

data, there is a reflection point after which the AUC will not increase anymore. At this point, the model

begins to overfit the derivation data; thus, the discrimination decreases in the validation data. Table 3.1

and Figure 3.2 in the results section demonstrate the AUCs of the random forest model with different

numbers and depths of trees.

Additionally, to demonstrate the effect of an increasing number of trees on the accuracy of the model, a

fit statistic (average square error) was also plotted against the different numbers of trees for the full

data and out-of-bag observations (Figure 3.5). Out of bag average square error is computed among the observations that were not used to train the decision tree.

To validate the trained predictive model, in addition to generating a validation AUC, I evaluated the model accuracy by applying it to the validation data and calculating the misclassification rate (test error rate) in predicted outcomes. A 2x2 table is generated from the predicted and observed outcomes. Misclassification rate is calculated by adding the number of false positive and false negatives and divide the sum by the total number of cases. The tables of ranked importance were generated for all predictors using both methods: loss reduction, and random branch assignment (RBA).

Similar to what was done in the logistic regression model (second chapter), calibration plots and Hosmer-Lemeshow goodness of fit test were performed using the predicted probabilities generated by the random forest model in the validation dataset.

The random forest model was compared to the logistic regression model that has been developed in the same cohort, previously. The model performances were compared using AUC and misclassification rate between the two models. ROC for the two models was drawn in a single plot to make the comparison easier. Additionally, to evaluate the performance of the RF model compared to the logistic regression model regardless of the number of observations included in the analysis, the RF model was also applied to the imputed data. In the second chapter of this dissertation, multiple imputation was used to impute the missing data, and then the logistic regression model was applied to the imputed data. In this chapter, the RF model was applied to the same imputed data as used in the logistic regression chapter. The AUC of the two models in the imputed data was then compared.

Moreover, to include a fair comparison between the RF and logistic models, two more analysis steps were done. First the RF model that developed in the derivation data was applied to the exact same number of patients in the validation cohort that have been used in the logistic model (i.e., those with no missing observation on any of the model predictor variables). Second, the logistic model was applied to

100

the validation cohort while missing observations on the predictor variables were recoded as a legitimate category. The AUC of the models were then compared to assess whether the inclusion of missing observations induce the difference between the two models' performance.

- ***Results***

  o <u>Study population</u>

The study cohort consisted of 7445 patients who had their first medical visit by a USMM provider in the calendar year of 2015 and were followed up for at least 12 months. Figure 2.1 displays the flow diagram of the study population.

The patients were 66% female, 63% white, the average age was 82± 9 years, 99% had Medicare coverage, and 27% were dual eligible (both Medicare and Medicaid). Functional status of the patients measured by KPS demonstrated severe disability (KPS≤40 – defined as the need for essential assistance and specialized care) in 54% of the patients. Prevalence of hypertension, hyperlipidemia, diabetes, and cancer were 81%, 50%, 34%, and 8% respectively. Over 50% of patients had 5 or more medical conditions. The study population is the same as used in chapter two (logistic regression). Table 2.4 demonstrates the population characteristics.

The minimum and maximum follow up time for this cohort were 1 and 865 days, respectively; with average (standard deviation) of 413 (210) days and median (interquartile range) of 444 (q1=244, q3=581) days (Table 2.5).

Overall, during the total follow up time, 45% of the cohort died, and 19% were admitted to the hospice. However, the 1-year mortality and hospice admission rates within the first year of follow up were 32% and 10%, respectively. Among hospice-admitted patients, 765 (55%) died within three months of their admission. Overall 2680 deaths (80% of all deaths) occurred outside of hospice.

o Outcome: one-year mortality

- *Random forest development*

Figure 3.4 and Table 3.1 demonstrate the sensitivity of the model's AUC to the two random forest hyper-parameters, the number of trees and the depth of trees. In the derivation data set, increasing the depth and the number of trees always increases the AUC (Table 3.1).

Figure 3. 3. Impact of RF hyper-parameters on the AUCs of the random forest model applied to the validation dataset– 1-year mortality
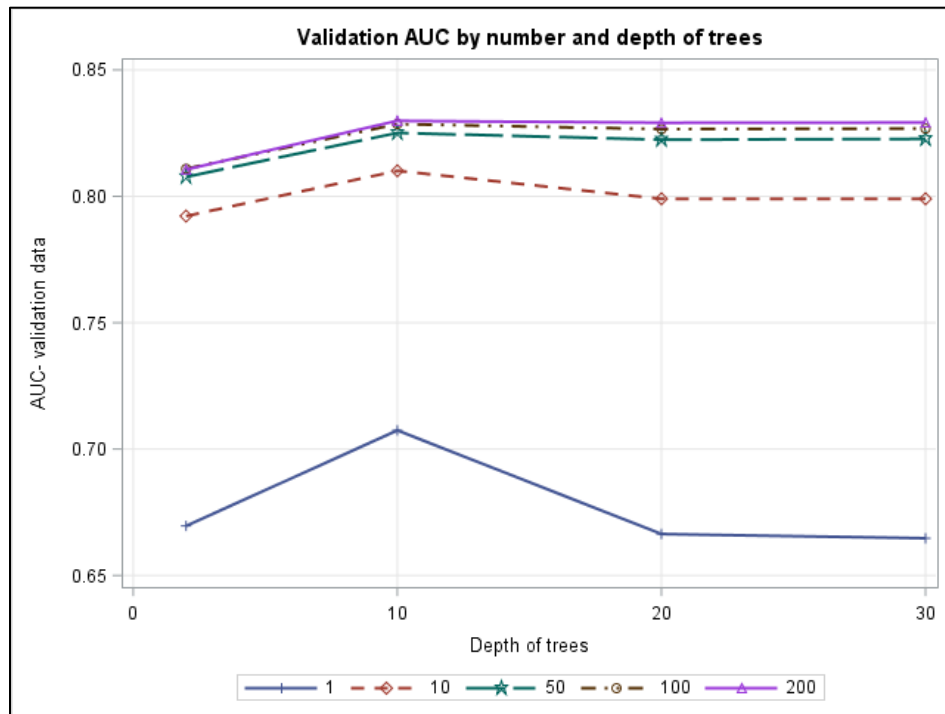
Colors and patterns indicates different number of trees
The number of trees varies between 1 and 200 and is indicated by colored lines
The depth of trees varies between 1 and 50 and is indicated on the X axis

Applying the model to the validation data reveals the point of reflection which corresponds to the optimal number of trees and depth of the forest. Figure 3.4 shows the AUCs for the different number of trees (1, 10, 50, 100, and 200) and depth (2, 10, 20, and 30). The vertical axis is the AUC, and the horizontal axis is the depth of trees. Patterns represent the different number of trees. It shows the reflection point at a depth of 10, however when the number of trees is greater than 100, there is no significant change in AUC after the depth of 10.

As shown in Table 3.1, in this data set the optimal number of trees and depth are 200 and

10respectively, because they resulted in the highest AUC, although the difference in the AUC between

depth 10 and 20 and between numbers of trees 100 to 200 is small. Therefore the selection of depth

and number of trees depends on computational aspects, i.e., for the large data set and low capacity

machine, 50 trees with the depth of 10 will be as sufficient as a model with 100 trees with the depth of

20.

Table 3. 1. AUC from random forest model in derivation and validation data sets using different depth and number of trees- mortality outcome

AUC	Max-Depth=2	Max-Depth=10	Max-Depth=20	Max-Depth=30
Derivation				
N-Trees=1	0.6883 (0.67- 0.70)	0.8132 (0.80- 0.83)	0.8519 (0.84- 0.87)	0.8524 (0.84- 0.87)
N-Trees=10	0.8214 (0.81- 0.84)	0.9226 (0.91- 0.93)	0.9809 (0.978- 0.984)	0.9820 (0.979- 0.985)
N-Trees=50	0.8340 (0.82- 0.85)	0.9409 (0.93- 0.95)	0.9924 (0.991- 0.994)	0.9933 (0.992- 0.995)
N-Trees=100	0.8352 (0.82- 0.85)	0.9453 (0.94- 0.95)	0.9943 (0.993- 0.996)	0.9949 (0.994- 0.996)
N-Trees=200	0.8342 (0.82- 0.85)	0.9453 (0.94- 0.95)	0.9951 (0.994- 0.996)	0.9957 (0.995- 0.997)
Validation				
N-Trees=1	0.6696 (0.65- 0.68)	0.7074 (0.69- 0.73)	0.6664 (0.65- 0.68)	0.6647 (0.65- 0.68)
N-Trees=10	0.7922 (0.78- 0.81)	0.8101 (0.80- 0.82)	0.7990 (0.78- 0.81)	0.7990 (0.78- 0.81)
N-Trees=50	0.8077 (0.79- 0.82)	0.8251 (0.81- 0.84)	0.8224 (0.81- 0.84)	0.8227 (0.81- 0.84)
N-Trees=100	0.8109 (0.80- 0.83)	0.8286 (0.81- 0.84)	0.8266 (0.81- 0.84)	0.8268 (0.81- 0.84)
N-Trees=200	0.8106 (0.80- 0.83)	0.8299 (0.82- 0.84)	0.8291 (0.82- 0.84)	0.8292 (0.82- 0.84)
AUC: area under the ROC curve; Max-Depth: maximum depth of decision trees in the random forest model; N-Trees: number of decision trees in the random forest model;				

Given the size of this data set, this analysis was not computationally intensive; thus, I chose to perform

the analysis with the number of trees of 200 and the depth of 30.

The validation AUC of about 83% in the model with N-trees=200 and depth≥ 10 indicates that the model

has good discrimination ability. Compared to the validation AUC from the logistic regression model,

random forest shows a 7% increase in the AUC (AUC 83% in random forest vs. 76% in the logistic

regression model).

To demonstrate the effect of an increasing number of trees on the accuracy of the model, a fit statistic

(average square error) was also plotted against the different numbers of trees for the full data and out-

of-bag observations (Figure 3.5). As expected, the average square error for out of bag sample is higher

than the one for the full data. Out of bag average square error is computed among the observations that

were not used to train the decision tree. The average squared error turns out to be stable at the number

of trees equal to 40-50 for OOB sample. After this point, increasing the trees in the forest does not

decrease the prediction error anymore.

The conclusion from Table 3.1 and Figures 4.2 and 4.3 is that the number of trees equal to 200 is more

than enough for building the forest in this data. Also, the depth of 10 can be enough, although by

increasing the depth to 30, there is no evidence of overfitting the model to the derivation data, i.e., the

AUC in the validation data did not decrease meaningfully. Thus the parameters for the construction of

RF were selected as the number of trees=200 and depth=30.

Figure 3. 4. The average squared error of the RF model by the number of trees for both OOB (top line) and full data (lower line)
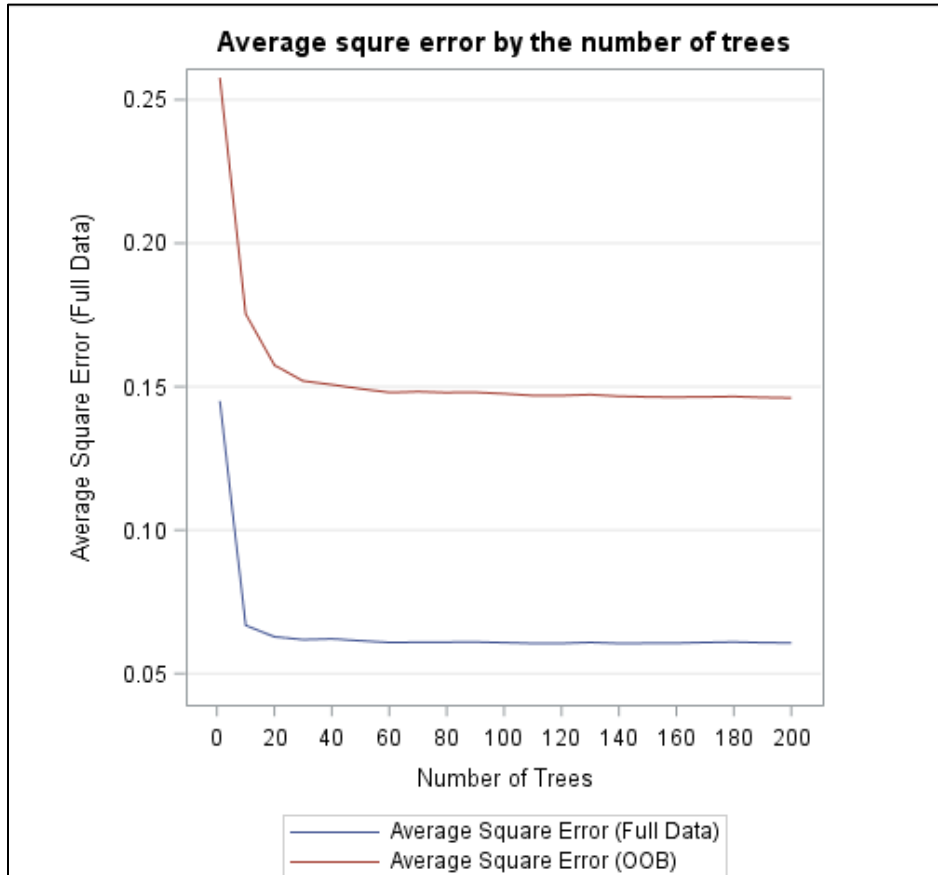
- *Variable importance*

All explanatory variables were evaluated for the importance, and the ranked importance table was

generated using both the random branch assignment (RBA) and loss reduction methods. Table 3.2

contains the first ten ranked important variables in the random forest model based on the RBA and loss

reduction methods. Complete tables of importance measures for all variables using both methods are

found in the appendix. The RBA results are provided for both derivation and validation group, whereas

the loss reduction importance is reported only for derivation data. The reason is that loss reduction is a

product of the *'Proc HPForest'* which develops the model in derivation data, whereas RBA importance is

computed in *'Proc HP4score'* which applies the developed RF algorithm to any given data including

validation and derivation. Among the highest ranked variables, TUG, Albumin, race, age, KPS, and

cholesterol are consistently among the first ten important variables when looking at the RBA importance

table. Variables ADL decline and SQ ranked high in loss reduction method; however, as discussed in the

background section, the loss reduction method can be biased towards the importance of correlated

variables. The medical history variables (CCW variables) were ranked among the first ten variables,

inconsistently. The number ten was selected arbitrarily to compare the results of the random forest

'Importance' statement with the final logistic regression model results.

Table 3. 2. The first ten ranked important variables in the random forest model- Mortality outcome

Ranked Importance	RBA- Validation data	RBA- Derivation data	Loss reduction- Derivation data
1	TUG Answer	Albumin result	TUG Answer
2	Albumin result	Cholesterol result	Race
3	Race	TUG Answer	ADL decline
4	ADL decline	Age	Surprise question
5	KPS value	Diagnosis-count	CCW-Hyperlipidemia
6	Age	Race	Lives alone
7	Cholesterol result	KPS value	Tobacco use
8	Dual eligible	CCW-Hyperlipidemia	Albumin result
9	CCW-Depression	Number of Medications	CCW-Cataract
10	CCW-Chronic Kidney Disease	Surprise question	CCW-Alzheimer's

- *Comparison to the logistic regression model*

There are different methods to evaluate the importance of variables in the logistic regression methods;

standardization of the coefficients, odds ratio, and Wald test results are a few examples. (137,138) None

of the methods is agreed upon by data scientists. I used the odds ratio and Wald test p-values to

evaluate the importance of predictors in the logistic regression model developed in the second chapter.

Considering odds ratios and parameter estimates in the logistic model, ADL-decline had the largest

effect among all the variables, following by albumin, race, SQ and cholesterol (Table 3.3). TUG was not

selected in the final logistic model since it was not significantly associated with the outcome. It is

noteworthy that in logistic regression analysis, more than 20% of the observations had missing value on

TUG variable and therefore were excluded from the analysis. On the other hand, the random forest model uses all observations in the analysis regardless of missing value on the explanatory variables. It can be a reason for the fact that TUG was not a significant predictor of outcome in the logistic regression model where it is first ranked important variable in RF.

Table 3. 3. The variable importance in the logistic regression model, (by estimates and significance) - Mortality outcome

effect	Odds Ratio Estimates			Parameter estimates	
	Point Estimate	95% Wald Confidence Limits		Parameter estimate	P-value
ADL-decline, Decline vs. No-change	0.790	0.577	1.081	-0.2356	0.1407
ADL-decline, Improve vs. No-change	0.096	0.023	0.397	-2.3422	0.0012
Albumin, 3.2-<3.5 vs 3.8+ gr/dl	1.884	1.303	2.725	0.6336	0.0008
Albumin, 3.5-<3.8 vs 3.8+ gr/dl	1.486	1.015	2.175	0.3959	0.0417
Albumin, <3.2 vs 3.8+ gr/dl	3.750	2.613	5.382	1.3218	<.0001
Race, Black vs. White	0.588	0.415	0.833	-0.5306	0.0028
Race, Other vs. White	0.442	0.197	0.991	-0.8156	0.0475
Surprise question, No vs. Yes	2.073	1.533	2.803	0.7289	<.0001
Cholesterol, 136-<164 vs 195+ gr/dl	1.191	0.839	1.690	0.1747	0.3285
Cholesterol, 164-<195 vs 195+ gr/dl	1.304	0.923	1.843	0.2658	0.1317
Cholesterol, <136 vs 195+ gr/dl	1.959	1.384	2.772	0.6724	0.0001
CCW-Hyperlipidemia Yes vs. No	0.531	0.417	0.676	-0.6334	<.0001
Age, 75-84 years vs. 65-74 years	1.711	1.180	2.481	0.5372	0.0046
Age, 85-94 years vs. 65-74 years	1.804	1.259	2.584	0.5898	0.0013
Age, 95+ years vs. 65-74 years	1.602	0.953	2.693	0.4712	0.0755
KPS, Severe vs. Moderate disability	1.543	1.199	1.986	0.4340	0.0007
CCW-Depression Yes vs. No	0.654	0.478	0.896	-0.4244	0.0082
Dual-eligibility, Yes vs. No	0.687	0.509	0.929	-0.3751	0.0146
Sex, Male vs. Female	1.151	0.886	1.497	0.1411	0.2917*

*Sex was included in the final logistic model, although the Wald test for its coefficient was not statistically significant

A correlation plot was generated to evaluate the correlation of the predicted probabilities between the two models, logistic regression and RF (Figure 3.6). The correlation was strongly positive with the coefficient of 0.6512 and p-value=<0.0001 when comparing the two models' predicted values among the 2312 patients that were included in both analyses. Notably, the best correlation is in the lowest values of the predictions (i.e., probabilities ≤0.4). It means the two models agreement on the risk of

outcome is better in lower probabilities. In other words, the two models are consistent in identifying the

low-risk group, but they are inconsistent in assigning the patients to the high-risk category.

Figure 3. 5. Correlation of the predicted probability of death between the two models (N=2312)
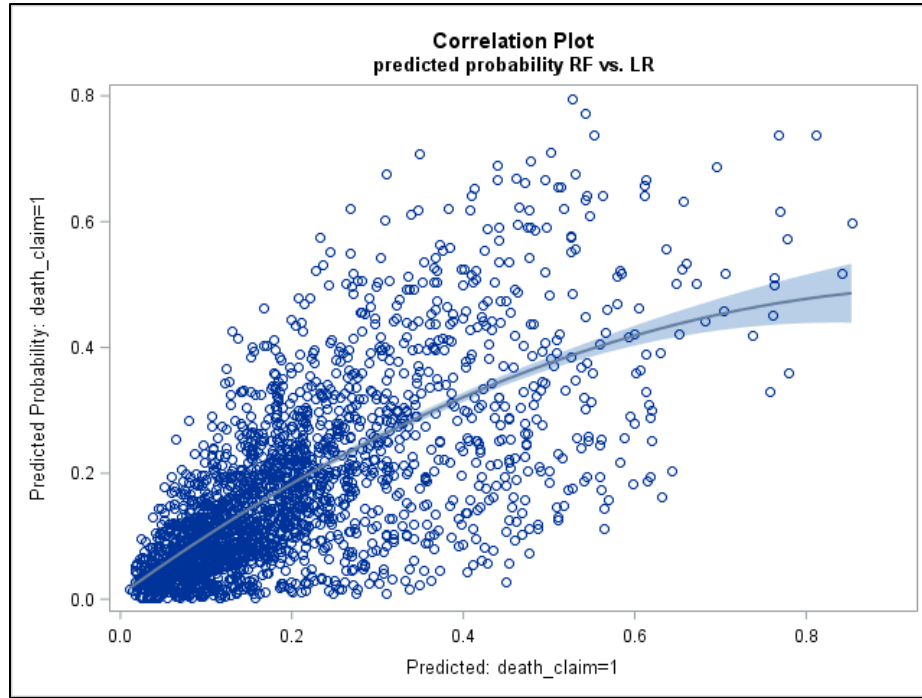
LR predictions are on the vertical axis, and RF predictions are on the horizontal axis

AUC, sensitivity, specificity, and misclassification (test error) rate were calculated for each model (LR and

RF) at two different cut points, top 10% and top 20% of predicted probability.

Table 3. 4. Comparison of the model performance for prediction of 1-year mortality, logistic regression
and random forest models (validation N=3723)

Model	AUC validation	N analyzed	High-risk group	sensitivity	specificity	PVP	PVN	Test error rate[&]
Logistic regression model	0.7634 (0.74-0.79)	2312* Death=485 (21%)	Top 10%	25.1%	94.0%	52.8%	82.6%	20.4%
			Top 20%	44.5%	86.5%	46.7%	85.5%	22.3%
Random forest model	0.8292 (0.82-0.84)	3723 Death=1241 (33.3%)	Top 10%	24.9%	97.4%	82.8%	72.2%	26.8%
			Top 20%	46.7%	93.3%	77.7%	77.8%	22.2%
PVP: predictive value positive; PVN: predictive value negative; *The number of observations that were analyzed in the logistic regression is less than total due to missing data [&]Test error rate or misclassification rate calculated as the number of misclassified predictions divided by total observations								

The AUC of the prediction model increased by 9% from the LR to the RF model (Table 3.4). Figure 3.7 shows the two ROC curves from RF and logistic regression models. Comparison of the 95% confidence intervals (CI) of the AUC between the two models demonstrated a better precision (i.e., narrower CI) of the AUC for the random forest model than logistic regression. Sensitivity of the two models was similar whereas specificity of the random forest was better than LR. Thus both models had similar ability to identify patients that died but RF model had better identification of patients who lived. Comparison of predictive values is complicated by the higher prevalence of death in the RF population which will increase PVP and decrease PVN. So despite similar sensitivity values, the RF model had lower PVN because of the higher mortality rate. The higher specificity in the RF model will results in a higher PVP but the difference in PVP between the LR and RF models are much greater because of the combination of lower false positive rates and higher prevalence. In summary, the RF results in substantially better PVP but slightly lower PVN compared to the LR model. But neither model has sufficient PVP or PVN to rule in or rule out mortality with confidence.

Misclassification rate in the RF was higher than LR (27% vs.20%) when the 10% threshold was used, whereas, with the 20% threshold, both models had similar misclassification rates (22%). A possible explanation is that the population analyzed in the LR model is smaller than those analyzed in the RF model because patients with partly missing data were excluded in the LR model. These excluded patients had higher rate of mortality (Table 3.4). Again the selection of the threshold is dependent on different factors, including the cost of the interventions and services for different risk groups and the resources that the company can allocate to them.

Figure 3. 6. Comparison of ROCs between the two models- RF and LR, logistic regression (N=2312) and random forest model (N=3723)
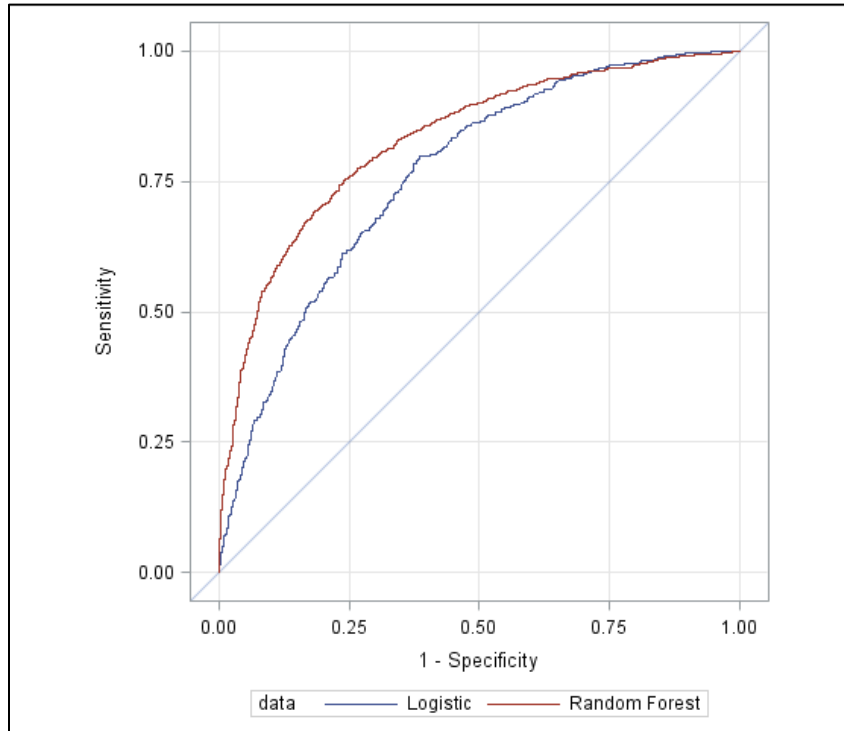
As shown in Table 3.4, the total number of observations in the validation data is different between the

two models due to missing observations. In the logistic regression procedure, observations with missing

on any of the variables are excluded, while random forest can handle observations with missing in

explanatory variables. Therefore there are 1411 fewer observations in logistic regression than the

random forest analysis. For a better understanding of the two models, the random forest was also

applied to the same 2312 patients in the logistic regression model, and the difference between AUCs

was tested using *ROC Contrast* statement.

Figure 3.8 shows the small difference between the two AUCs when the same observations are used in

the generation of the ROCs (AUC=0.77 for random forest vs. 0.76 for logistic regression). The chi-square

test showed that the difference in AUC between the two models was not statistically significant (Chi-

square=0.563 and p-value=0.45). This finding suggests that the gain in random forest's AUC compared to

the logistic regression model is mainly due to the inclusion of all patients. In fact, including the

observations with partly missing data causes the increase in AUC of the random forest model in this

data. Nevertheless, it is an essential advantage of the random forest algorithm that all observations with

and without missing data can be included in the analysis.

Figure 3. 7. Comparison of ROCs between the logistic regression and random forest models when using
the same validation cohort in both models (N=2312)
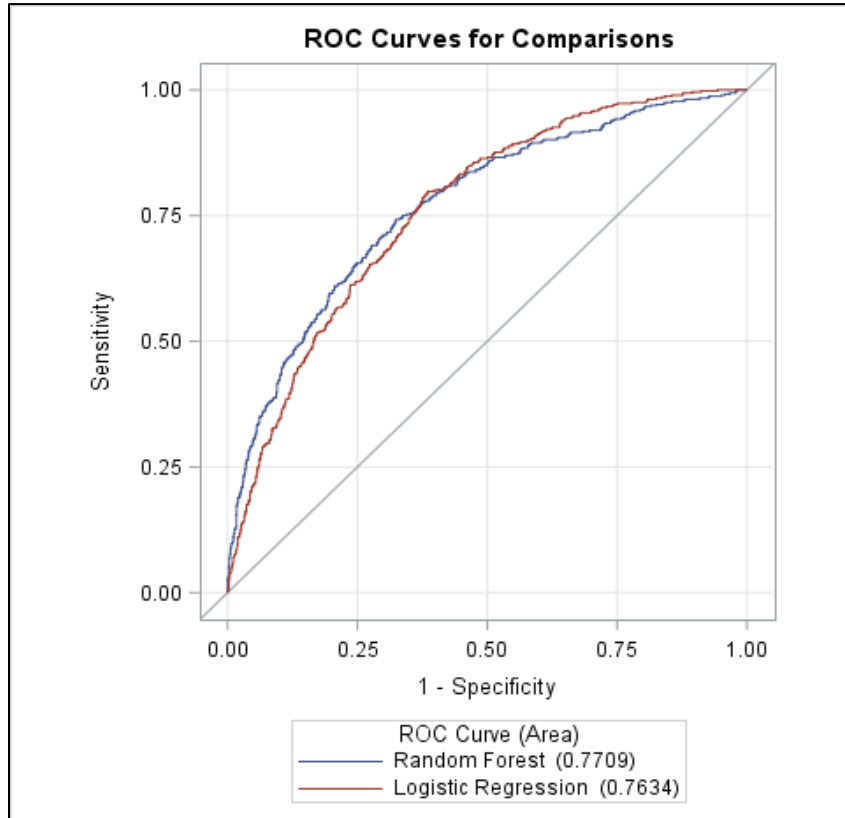
Table 3. 5. ROC and 95% confidence intervals from the RF and LR models (N=2312)

Model	Mann-Whitney			
	Area	Standard Error	95% Wald Confidence Limits	
Random Forest	0.7709	0.0122	0.7469	0.7948
Logistic Regression	0.7634	0.0114	0.7410	0.7859

Table 3. 6. ROC contrast between the two models, RF and LR

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Logistic Regression	1	0.5632	0.4530

In another attempt to confirm the fact that the main advantage of RF model in this data is due to

inclusion of missing data, the missing observations were replaced and included in the analysis as a new

category. For example a binary variable with levels 0=No and 1=Yes, now has another level 2= missing.

This way no any observations is excluded from the analysis. The LR model was applied to this population

and the AUC was much higher (N=3723, AUC=0.8379) than the LR model in the available data (N=2312,

AUC=0.7634). This sensitivity analysis confirms again that in this cohort missing data carry valuable

information in prediction of adverse outcomes.

- *Applying the RF model to imputed data*

Additionally, the RF model was applied to the imputed dataset. The imputed data was the same data as

was used in chapter two– logistic regression. Multiple imputation was used to generate the imputed

data with 20 imputations. RF model was applied to all 20 datasets, and predictions were generated for

all observations. An average of the predicted probabilities for each observation was calculated across

the 20 imputations. Then the ROC for the validation cohort was generated (Figure 3.9). The AUC for the

RF model in the imputed data was slightly lower than the LR model in imputed data (AUC=0.7605 vs.

0.7756) and was notably lower than AUC of the RF model in available data (0.8292). These results

indicate that observations with partially missing data are informative, and they cannot be excluded from

the analysis. Also, imputation of these missing observations did not result in better discrimination in the

RF or LR models. The most probable reason is that the mechanism of missing in this data is not random.

The results of modeling in the imputed data confirms again that missingness on the predictor variables

per se is important in the prediction of the outcomes.

Figure 3. 8. ROC from the random forest model applied to the imputed validation data (average of 20 predictions for each individual was generated from 20 imputed dataset)
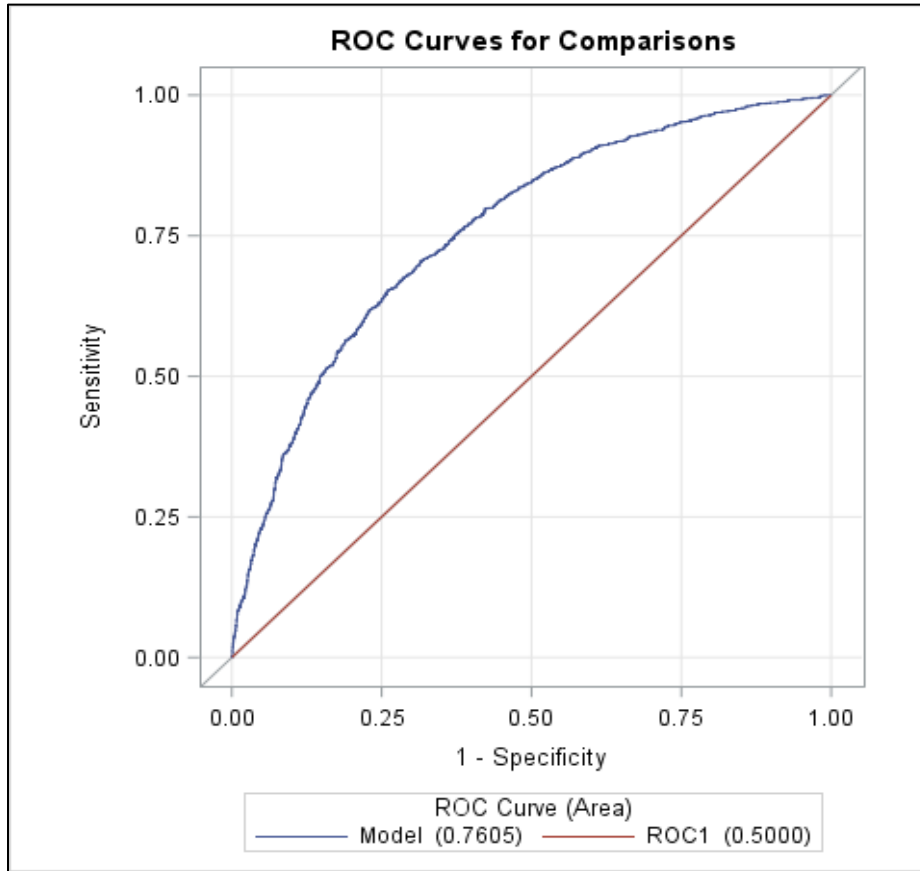
The model's ROC is compared to the null model (red line)

Table 3. 7. AUC and the 95% confidence intervals from the RF model in the imputed data

ROC Association Statistics				
ROC Model	**Mann-Whitney**			
	Area	**Standard Error**	**95% Wald Confidence Limits**	
Model	0.7605	0.00822	0.7444	0.7766

- *Model's goodness-of-fit*

To evaluate the model's goodness of fit, calibration plots and Hosmer-Lemeshow test were generated.

Similar to the logistic regression chapter, two methods were used to generate the calibration plots-

loess based and decile based. As observed in Figures 3.10 and 3.11, the model fit is the best in the lower

113

predicted probabilities. The deviation of the perfect fit model begins at the predicted probabilities of

0.50 where the model underestimate the mortality risk.

Figure 3. 9. Loess-based calibration plot for RF model- mortality outcome- validation cohort (N=3723)
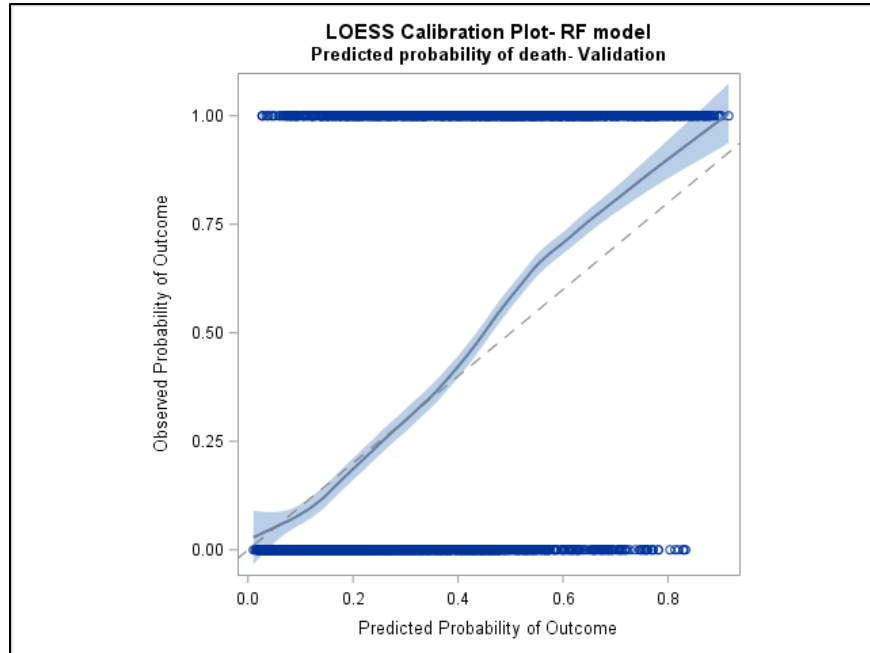
The vertical axis is the observed outcome, and the horizontal axis is the predicted probability

Figure 3. 10. Decile based calibration plot for RF model- mortality outcome-validation cohort (N=3723)
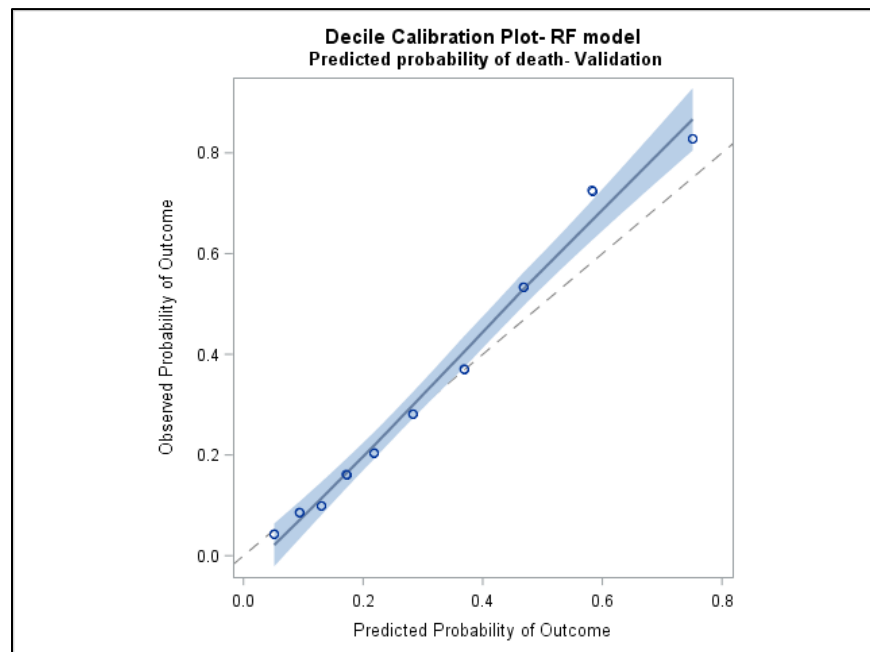
The vertical axis is the observed outcome, and the horizontal axis is the predicted probability

Hosmer-Lemeshow test for the random forest model was performed using the observed and predicted outcomes in the validation data set and resulted in a test statistic= 54.32 and p-value <0.0001 which means the model's lack of fit in this data cannot be rejected. Again it is consistent with the calibration plots, which shows loss of fit at probabilities higher than 0.50.

    o   <u>Outcome: one-year hospice admission</u>

To develop a random forest model for prediction of 1-year hospice admission and to evaluate the importance of predictors, the same methods that were applied for the mortality outcome, also were used here.

    -   *Random forest development*

To find the optimal parameters for the random forest in the prediction of hospice admission, different numbers (1, 10, 50, 100, and 200) and depths (2, 10, 20, and 30) of trees were tested, and the AUC of the models were provided in Table 3.8. Developed RF was applied to both derivation and validation datasets.

Table 3. 8. AUC from random forest model in derivation and validation data sets using different depth and number of trees- hospice outcome

AUC	Max-Depth=2	Max-Depth=10	Max-Depth=20	Max-Depth=30
Derivation				
N-Trees=1	0.5907 (0.56- 0.62)	0.7884 (0.76- 0.81)	0.8096 (0.78- 0.83)	0.8095 (0.78- 0.83)
N-Trees=10	0.7078 (0.68- 0.73)	0.9503 (0.94- 0.96)	0.9909 (0.98- 0.99)	0.9913 (0.98- 0.99)
N-Trees=50	0.7431 (0.72- 0.77)	0.9866 (0.98- 0.99)	0.9999 (0.99- 1.0)	0.9999 (0.99- 1.0)
N-Trees=100	0.7500 (0.73- 0.77)	0.9893 (0.98- 0.99)	1.00 (0.99- 1.00)	1.00 (0.99- 1.00)
N-Trees=200	0.7503 (0.73 0.77)	0.9907 (0.98- 0.99)	1.00 (1.00- 1.00)	1.00 (1.00- 1.00)
Validation				
N-Trees=1	0.5415 (0.51- 0.57)	0.5648 (0.54- 0.59)	0.5392 (0.51- 0.57)	0.5394 (0.51- 0.57)
N-Trees=10	0.6641 (0.64- 0.69)	0.6556 (0.63- 0.68)	0.6371 (0.61- 0.67)	0.6378 (0.61- 0.67)
N-Trees=50	0.6953 (0.67- 0.72)	0.6885 (0.66- 0.71)	0.6727 (0.65- 0.70)	0.6721 (0.64- 0.70)
N-Trees=100	0.6997 (0.68- 0.72)	0.6986 (0.67- 0.72)	0.6919 (0.67- 0.72)	0.6915 (0.66- 0.72)
N-Trees=200	0.7022 (0.68- 0.73)	0.7028 (0.68- 0.73)	0.6973 (0.67- 0.72)	0.6971 (0.67- 0.72)
AUC: area under the ROC curve; Max-Depth: maximum depth of decision trees in the random forest model; N-Trees: number of decision trees in the random forest model;				

The validation AUC of about 70% in the model with N-trees=200 and depth≥ 10 indicates a moderate to good discrimination ability. Compared to the validation AUC from the logistic regression model, random forest shows slightly lower AUC for the hospice admission outcome (AUC=0.7251 for logistic regression vs. 7028 for random forest).

Figure 3.12 shows the effect of increasing the number of trees on model performance. Average squares error- a fit statistics- was plotted against the different numbers of trees for the full data and out-of-bag observations. The conclusion from Table 3.8 and Figure 3.12 is that the number of trees equal to 200 is more than enough for building the forest in this data. Also, the depth of 10 can be enough, although by increasing the depth to 30, there is no evidence of overfitting the model to the derivation data, i.e., the

AUC in the validation data did not decrease meaningfully. Thus the parameters for construction of the

RF were selected as number of trees=200 and depth=30.

Figure 3. 11. The average squared error of the RF model by the number of trees- Hospice outcome
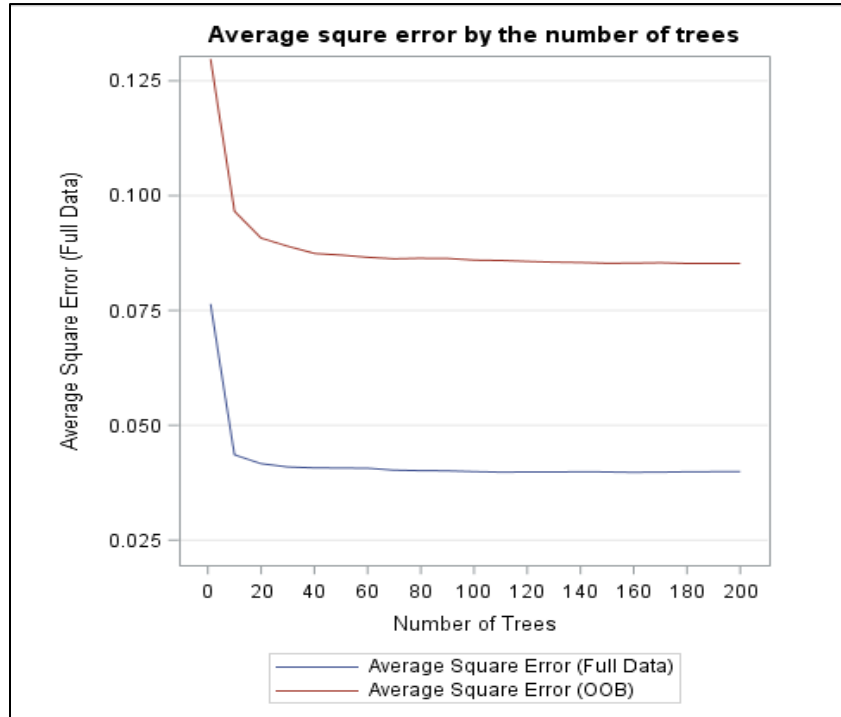
- *Variable importance*

To evaluate the importance of predictors for 1-year hospice admission, the same methods that were

applied for the mortality outcome, also were used here.

Table 3. 9. The first ten ranked important variables in the random forest model- Hospice outcome

Ranked Importance	RBA-Validation data	RBA-Derivation data	Loss reduction-Derivation data
1	Surprise question	Age	KPS (category)
2	Age	Albumin result	Surprise question
3	KPS (category)	Cholesterol result	CCW-Hip fracture
4	Number of lab tests	Number of Medications Age	CCW-Cataract
5	Albumin result	Diagnosis-count	CCW-endometrial Ca
6	Living alone	KPS (category)	CCW-Lung Ca
7	Race	Number of lab tests	CCW-Asthma
8	Dual eligible	Surprise question	Dual eligible
9	TUG Answer	TUG Answer	CCW-Prostate Ca
10	Cholesterol result	CCW-Hyperlipidemia	CCW-Glaucoma

Since RBA is the most commonly recommended method for evaluation of the variable importance in the

RF model, the predictors selected in this method are compared to the LR model. Interestingly SQ is the

first ranked important variable in the prediction of hospice admission in this model. It is consistent with

the literature on the importance of SQ (100) and with the findings in chapter two. Also, age and KPS

functional indicator are the next ranked predictors. Living alone, worse functional status, and poor

nutritional condition (indicated by albumin and cholesterol levels) are all variables that indicate the

possible need for hospice care. The main difference in predictors of the mortality and hospice admission

is that hospice admission must be ordered by a physician to be eligible for insurance reimbursement,

Medicare in this population. Therefore factors that point to patients' short life expectancy and their

inability to live at home are flags for physicians and subsequently the predictors of hospice order. It is

noteworthy that in Medicare criteria for hospice eligibility, the first requirement is a physician confirms

the patient's life expectancy of less than 6 months, but then for a patient to be covered for hospice

services, they have to give up traditional Medicare coverage (e.g., curative treatment for cancer

patients). (35,139)

It is notable that loss reduction method ranked predictors differently from the RBA method. Since the

first 10 ranked variables in RBQ methods are more similar to the variables in the logistic regression

model for hospice, this suggest that the RBA method was better in identification of the important

predictors of hospice.

- *Comparison to the logistic regression*

The logistic regression model that was developed for the hospice admission in chapter 2 included 7

variables. The results of the model are presented in Table 3.10. Odds ratios and Wald test p-values were

used to evaluate the importance of predictors in the logistic regression model. Similar to the LR model

for mortality, ADL-decline was the most influential predictor of 1-year hospice admission, followed by

age and KPS level. Five of the seven variables (surprise question, age, KPS, Race, and dual eligibility) in

118

the final logistic model are among the 10 first important variables ranked in the RF model by RBA

method in the validation data. Surprisingly the ADL-decline was not among the first 10 variables in RF. It

might be explained by the fact that ADL-decline was missing in 16% of observations and the missingness

on ADL was not associated with the hospice admission (as shown in the second chapter, Table 2.6).

Therefore including all of the observations in RF model caused ADL-decline importance to be ranked

much lower (number 11), whereas it ranked the first according to the odds ratio and p-value from the LR

model. Age and KPS are ranked second and third in both RF and LR models. Overall the importance of

predictors of hospice admission in both LR and RF models were consistent.

Table 3. 10. The variable importance in the logistic regression model, (by estimates and significance) - Hospice outcome

Odds Ratio Estimates				Parameter estimates	
effect	Point Estimate	95% Wald Confidence Limits		Parameter estimate	P-value
ADL-decline, Decline vs. No change	1.034	0.735	1.454	0.0335	0.8473
ADL-decline, Improve vs. No change	0.086	0.012	0.628	-2.4477	0.0155
Age, 75-84 years vs. 65-74 years	2.392	1.386	4.127	0.8720	0.0017
Age, 85-94 years vs. 65-74 years	3.345	1.986	5.633	1.2073	<.0001
Age, 95+ years vs. 65-74 years	3.870	2.055	7.286	1.3531	<.0001
KPS, Severe vs. moderate disability	3.125	2.239	4.361	1.1393	<.0001
Surprise question, No vs. Yes	2.131	1.547	2.934	0.7566	<.0001
Dual-eligibility, No vs. Yes	2.023	1.336	3.064	0.7045	0.0009
Race, Black vs. White	0.654	0.426	1.004	-0.4242	0.0522
Race, Other vs. White	0.547	0.213	1.404	-0.6035	0.2095
Sex, Male vs. Female	1.165	0.866	1.567	0.1525	0.3137*
*Sex was included in the final logistic model, although the Wald test for its coefficient was not statistically significant					

A correlation plot was generated to show the correlation of the predicted probabilities of outcome

between the two models (Figure 3.13). The correlation coefficient was 0.7424 and p-value was <0.0001,

using the total of 2590 observations that were analyzed in both LR and RF. The correlation was strong

and positive; however, similar to the mortality models, the best correlation is seen where the predicted

probabilities are low in both models. The two models (Rf and LR) showed higher correlation in

predictions for hospice admission compared to mortality (correlation 0.74 vs. 0.65).

Figure 3. 12. Correlation of the predicted probability of hospice admission between the two models, logistic regression and random forest, (N=2590)

Correlation Plot- Predicted probability
Logistic regression vs. random forest
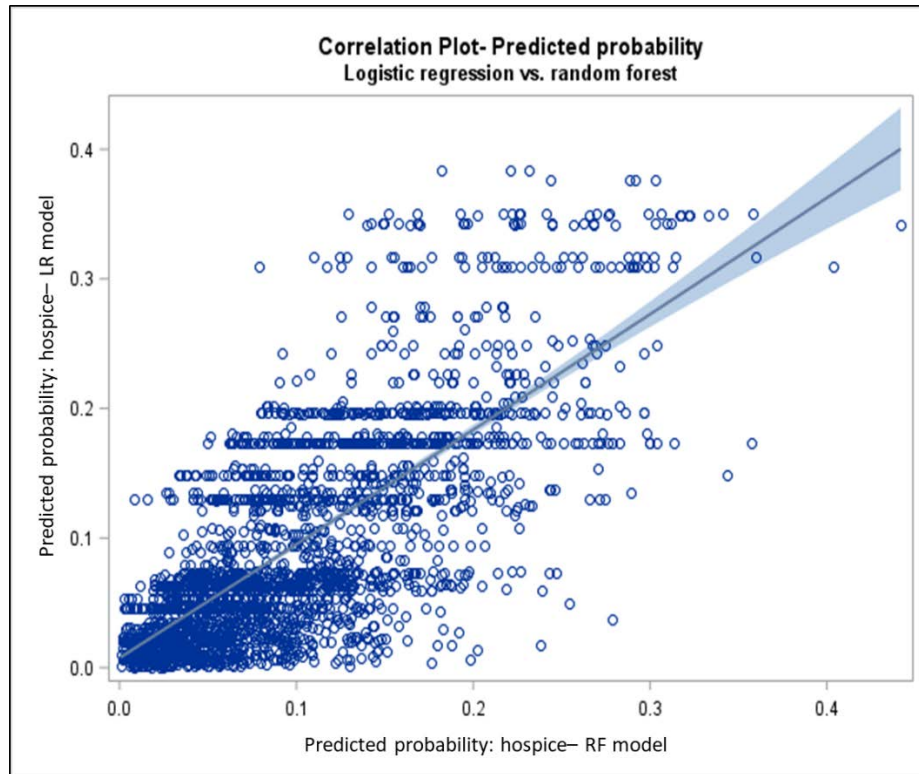
LR predictions are on the vertical axis, and RF predictions are on the horizontal axis

To compare the two prediction models, their discrimination ability was compared by generating AUC for

each model. Sensitivity, specificity, predictive values, and misclassification rates were also calculated

with two thresholds (top 10% and 20%), and the results are provided in Table 3.11.

Table 3. 11. Comparison of the model performance for prediction of 1-year hospice admission, logistic regression and random forest models (validation cohort)

Model	AUC validation	N analyzed	High-risk group	sensitivity	specificity	PVP	PVN	Test error rate[&]
Logistic regression model	0.72 (0.70-0.75)	2590* Hospice=266 (10.3%)	Top 10%	30.1%	90.5%	26.5%	91.9%	15.8%
			Top 20%	34.6%	87.2%	23.7%	92.1%	23.3%
Random forest model	0.70 (0.67-0.72)	3723 Hospice=384 (10.3%)	Top 10%	25.3%	91.8%	26.1%	91.4%	15.1 %
			Top 20%	41.2%	82.5%	21.2%	92.4%	21.8%
PVP: predictive value positive; PVN: predictive value negative; *The number of observations that were analyzed in the logistic regression is less than total due to missing data [&]Test error rate or misclassification rate calculated as the number of misclassified predictions divided by total observations								

Unlike the analysis of mortality outcome, the RF model did not have better discrimination in the prediction of hospice admission compared to the logistic regression model. Although the difference is small, at the 20% threshold, RF has higher sensitivity, also lower sensitivity and misclassification rate than the LR model. Predictive values were roughly similar between the two models. There is no significant superiority between the two models in prediction of the hospice outcome. Both models are good in the classification of the low-risk groups, but poor in identifying the high-risk patients. The most practical point for clinicians is related to the predictive values. Looking at the PVP and PVN, one can say when the models classify a patient as low risk, there is more than 90% probability that the patient will not go to hospice in the next year. Whereas among the patients who are classified as high risk for hospice, about 25% will actually go to hospice. So the models have utility in ruling out the need for hospice care (with 90% certainty) but are not useful in identifying patients that will be referred to hospice (ruling in).

Figure 3.14 displays the two ROCs from the LR and RF models in prediction of hospice admission. The two ROCs line up closely, although the AUC of the LR model is slightly larger than the RF model (0.7251

vs. 0.6971), and confidence intervals implies no significant difference between the two ROCs (Table

3.12).

Figure 3. 13. Comparison of ROCs between the two models- Hospice outcome, logistic regression
(N=2590) and random forest model (N=3723)
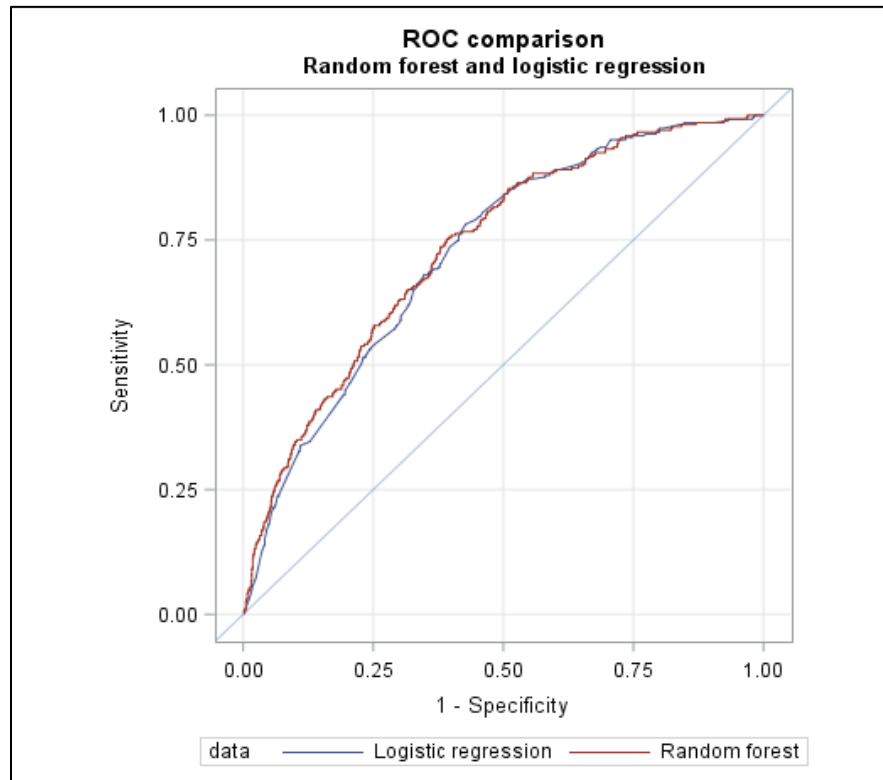
The two ROCs in Figure 3.14 were generated from different number of observations in the validation

data, i.e., 2590 observations were analyzed in the LR model, and 3723 in RF model. To assess whether

the application of the models to the same number of observations can make any difference in the model

discrimination, both models were applied to the 2590 observations that were included in the LR

analysis. Figure 3.15 shows the ROCs from the two models in the same population. The two AUCs are

very close (Table 3.12) and the statistical test for ROC contrast is non-significant for the difference

between the two model AUCs with a P-value of 0.43.

Figure 3. 14. Comparison of ROCs between the two models, logistic regression and random forest when using the same validation cohort in both models (N=2590)
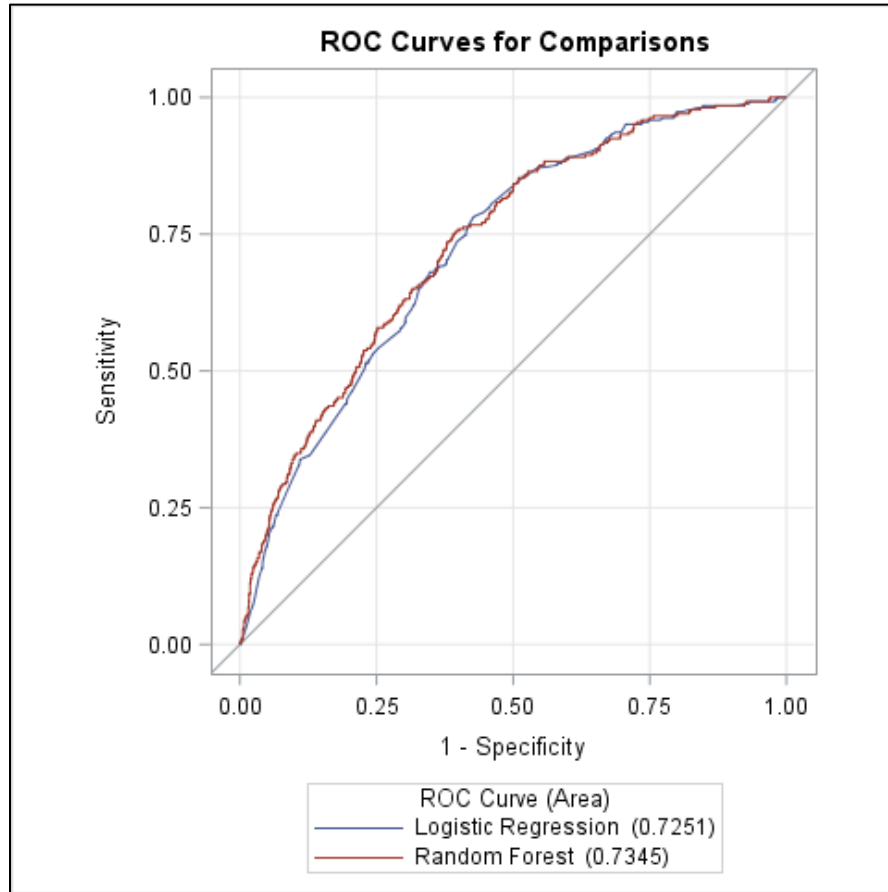
Table 3. 12. AUC and 95% confidence intervals from the two models, LR and RF applied to the same population (N=2590) - Hospice outcome

ROC Association Statistics				
ROC Model	**Mann-Whitney**			
	Area	**Standard Error**	**95% Wald Confidence Limits**	
Logistic Regression	0.7251	0.0151	0.6955	0.7547
Random Forest	0.7345	0.0154	0.7044	0.7646

- *Applying the RF model to imputed data*

The RF model that was developed in the available data was also applied to the imputed dataset. The imputed data was the same data as was used in logistic regression. Using the same methods that were described for the mortality outcome in the imputed data, a ROC was generated for the RF model in the imputed validation data (Figure 3.16 and Table 3.15). The AUC was similar to the one from the RF model in available data (AUC=0.6936 in imputed data vs. 0.6971 in available data). Indeed, imputation of the missing data did not improve discrimination in the RF model compared to when missing observations were included as missing. As discussed in chapter 2, imputation did not improve the discrimination in the logistic regression model either. It confirms that the missing observations on the predictors in these models were not associated with the hospice outcome (as was already shown in chapter 2). Unlike the mortality outcome, missingness on predictors did not predict the hospice outcome. Thus, although the random forest model allows the inclusion of all observations in the analysis, it did not improve the model discrimination in the prediction of hospice outcome in this data, mostly because missing observations were not predictive of outcome.

Figure 3. 15. ROC from the random forest model applied to the imputed validation data- Hospice outcome (N=3723)
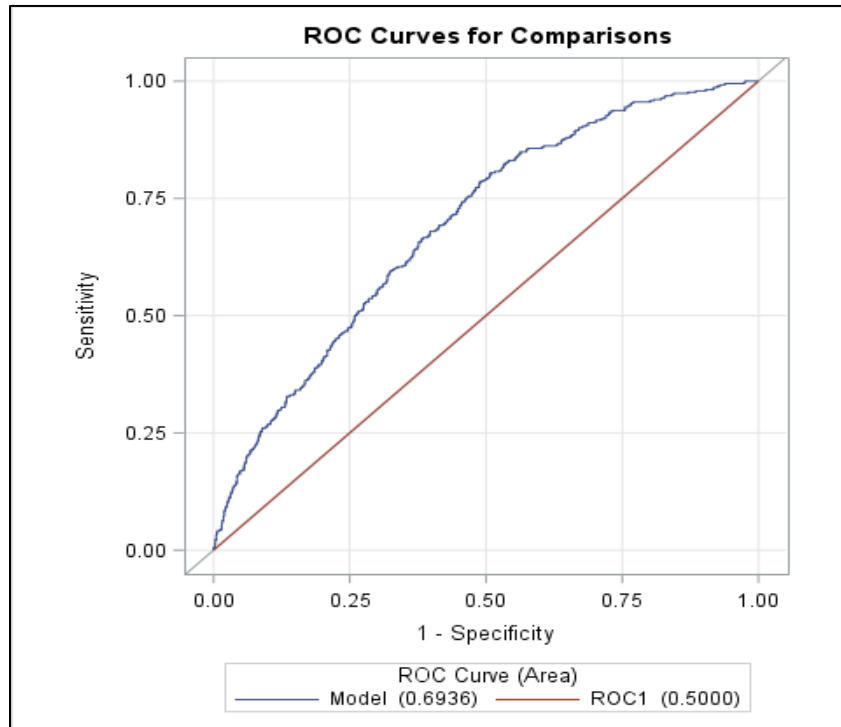
Table 3. 13. AUC and 95% confidence intervals from the RF model applied to the imputed data (20 replications) - Hospice outcome

ROC Association Statistics							
ROC Model	**Mann-Whitney**			**Somers' D**	**Gamma**	**Tau-a**	
	Area	**Standard Error**	**95% Wald Confidence Limits**				
Model	0.6936	0.0133	0.6675	0.7197	0.3872	0.3872	0.0717

Additionally, to confirm that using the RF model that was developed in the available data (with missing data included as a legitimate category) in the imputed data, a random forest model was developed in the imputed derivation data and applied to the imputed validation data. The AUC of this model was almost the same as the AUC from the previous model in the same data (AUC=0.6939 in the imputed data vs. AUC=0.6971 in the available data).

Calibration plots and Hosmer-Lemeshow test were generated for the RF model. Similar to the logistic

regression chapter, two methods were used to generate the calibration plots- loess based (Figure 3.17)

and decile based (Figure 3.18). Although the overall fit does not seem good, the model fits better in the

lower predicted probabilities than higher probabilities, like what was observed in the mortality model. In

other words, models under predict the mortality risk at higher risk levels.

Figure 3. 16. Loess-based calibration plot for the RF model- Hospice outcome
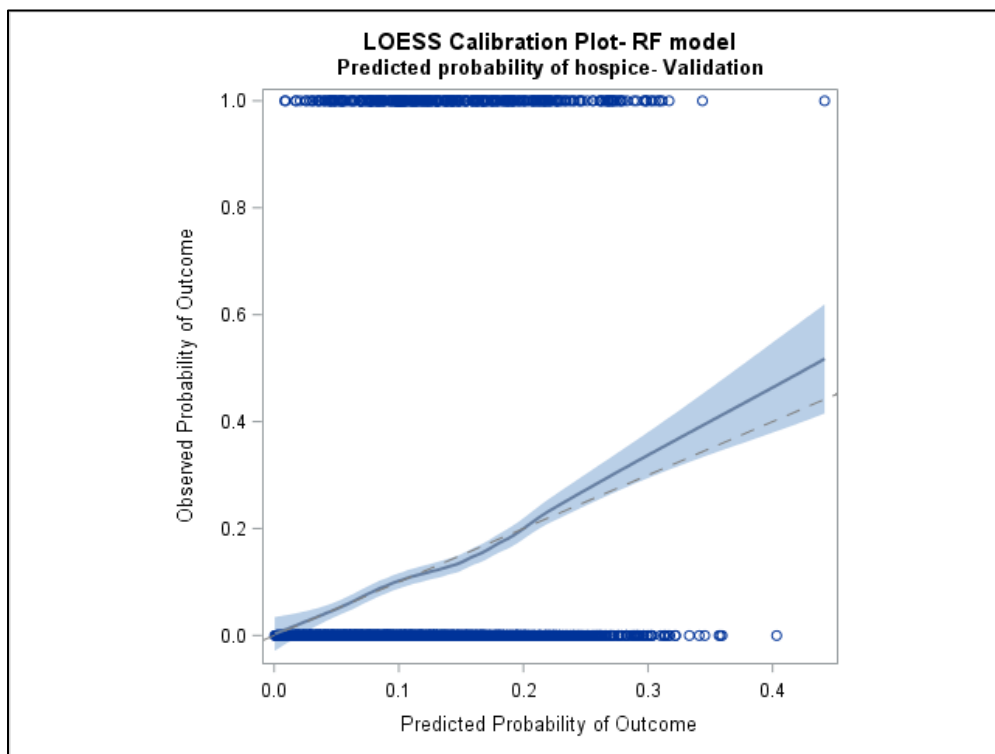
The vertical axis is the observed outcome and the horizontal is the predicted probability

Figure 3. 17. Decile based calibration plot for the RF model- Hospice outcome
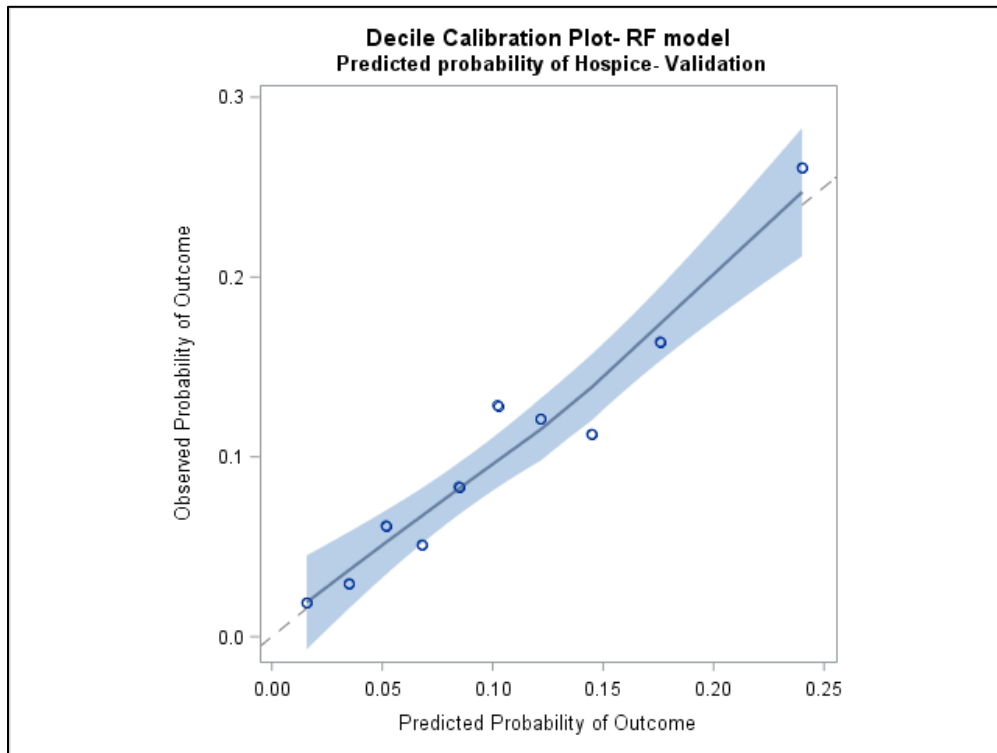
The vertical axis is the observed outcome and the horizontal axis is the predicted probability

Hosmer-Lemeshow goodness of fit test was also done and resulted in a test statistic=10.17 and p-value= 0.2532 which means there is no statistical evidence for lack of fit in this model. Although as discussed before, the interpretation of this test is limited because it is dependent to the number of groups that one selects for dividing the population.

- **Discussion**

In the validation data the random forest mortality model had a much higher AUC compared to the logistic regression mortality model ; the model's AUC was 9% higher compared to the logistic regression model (Table 3.4). The better discrimination ability of the RF in this data could be explained by the presence of underlying complex interactions and non-linear relationship between predictors and outcomes in this data, or by the inclusion of about 30% of observations that were excluded from the LR

model because of missing data. Our sensitivity analysis showed that almost all the improvement in the performance of RF compared to the LR model was due to the latter reason, i.e., inclusion of the observations with partly missing data. Misclassification rate in the RF model was higher than LR when using the top 10% of predicted probabilities as high-risk group (27% vs. 20%), however, at top 20% cut point, the misclassification rates were the same (22%) for both models. These results implies that RF estimation of the mortality risk in patients with the highest-risk (i.e., 10% P-hat) was not as good as the LR estimation. This fact was confirmed in the correlation plot as well. The random forest model shows better sensitivity and specificity than the logistic regression model except for the sensitivity at the 10% threshold in which sensitivity of both models are almost equal. In chapter two, I evaluated the performance of the logistic regression model and demonstrated its superiority compared to the alternative clinical risk prediction models currently used in this population. The random forest model for the mortality outcome is performing even better than the logistic regression model, although when the high threshold is selected for classification (i.e., top 10% of predicted probabilities), it has worse misclassification rate. In other words, to take advantage of the better discrimination, sensitivity, and specificity of the RF model without adding misclassification rate, the threshold of top 20% of predictions should be used for classification in this cohort. Predictive value positive of the RF model is substantially higher in RF than in LR, which indicates in this population with the mortality rate of 32%, about 80% of those who were identified by the RF as high risk, actually died within the next year. Whereas the PVP for LR model is about 50% (for both cutoff points). The key point from the performance measures is that both models (RF and LR) have PVPs higher than PVN which implies that the negative results of the model (i.e., predicted low risk by the model) are more reliable than the positive results (i.e., predicted high risk by the model). So the models have utility in ruling out the need for hospice care (with 90% certainty) but are not useful in identifying patients that will be referred to hospice (ruling in).

Unlike the mortality outcome, RF model for the hospice outcome did not show any improvement in the model's discrimination. In fact, the LR model shows slightly better AUC. Other performance measures including sensitivity, specificity, and predictive values were similar between the two models. Again, there is no evidence of superiority of the RF model in prediction of the hospice outcome in this data.

Although RF has many advantages especially when the data has complex interactions and non-linear relationship between the variables, in this study the main benefit of the RF model was due to its ability to include all observations including those with missing data on some predictors. We confirmed the gain in the AUC of the RF model for mortality, was mostly due to including of missing data, because when the RF applied to the complete case data (i.e., the same population as was used in LR model), the improvement in the AUC vanished; both LR and RF model had similar AUCs when applied to the complete case data. It is notable that almost one third of the observations were excluded from the LR because of missing data, and as observed in chapter two, missingness on the predictors were associated with the mortality outcome but not with the hospice admission – thus for mortality outcome we have evidence that the data is MNAR. The association between missing data on predictors (e.g., TUG, ADL-decline) and the outcome can be explained by survival bias. For instance a terminally ill patient is more likely to die before the doctor can complete the medical records. Also, in case of a very sick patient, the doctor may decide not to perform and record a given test or procedures particularly those that requires a patient's ability to move and cooperate (such as TUG).

The association between missingness on the seven key variables and the two outcomes in chapter two, revealed that there was trivial to no association between missingness and hospice admission, whereas a consistent and strong association with mortality was seen for all 7 predictors. These findings are in agreement with the results of RF model, where a substantial improvement in AUC was observed for mortality outcome, but not for hospice admission. Since missingness was not associated with the

hospice admission, inclusion of the missing data in the RF analysis, did not improve the AUC of the RF compared to the LR model for the hospice outcome.

Interestingly, when the RF model was applied to the imputed data, the AUCs for both outcomes were similar to the values from LR model (and notably lower than the RF in the available data). These observations can again be explained by the fact that missingness in this data is related to the mortality outcome, so including observations with missing variables (as in random forest) results in better performance when compared to effect of imputing the missing data. As a conclusion, the RF model had a better discrimination than logistic regression in this population for the mortality. In fact, taking account of the observations with partly missing data is the most important advantage of the RF model in this data. It seems that in this data complex interactions and non-linearity of the relationship among predictors and outcome is not a problem, because exclusion of the missing observations from the cohort resulted in the similar AUC for both RF and LR models.

The developed RF model may be incorporated into the routine data management programs in healthcare systems to facilitate identification of patients with different needs based on their risk levels and to support the provider's decision making. The developed random forest algorithm can be programmed into the USMM data system and the results would be embedded in the EMR to make a predicted probability for each new observation in the dataset. Then according to a predetermined cutoff value, different risk level patients are flagged for further attention. The model can be updated when further research results in a better model (e.g., having more complete data can result in including more predictors in the model development and it can change the final model, the changes then can be programmed into the database).

o <u>Strengths</u>

The first and most important strength of the random forest analysis was inclusion of all the observation in the analysis (i.e., observations with missing values on the predictor variables were allowed in the

130

random forest analysis whereas they were excluded by default in logistic regression model). We compared the RF and LR model and showed that the performance of both models are the same when applied to the same cohort. Another strength is our use of MI methods to account for missing data. We compared the results of RF models in imputed data and in the data with missing observations.

     o   Limitations

The most important limitation of this data was missing information. We excluded predictors with more than 20% missing values in the model development, so important variables such as hospitalization in the past year, or weight loss were excluded. There were still 9 variables with <20% missing included in the model development and which resulted in the subsequent exclusion of observations with partly missing data from the logistic regression model, however random forest overcame the problem of exclusion of missing observations. The strong association of missing values and mortality, in addition to disappearing of the AUC gain in the imputed data indicates the MNAR mechanism for missing data. Another limitation is that the RF output does not provide a tangible single model with familiar parameters that can be applied manually to any new observation. The algorithm saves hundreds of decision trees in the developed RF model and put any new data into all of the trees to make prediction and average them.

- ***Conclusion***

The use of machine learning techniques can improve the discrimination of predictive models compared to standard parametric regression models. In this study, random forest model demonstrated a better accuracy (i.e., higher AUC and lower test error rate) than the logistic regression model for mortality but not for hospice admission. The gain in the RF model discrimination in this data is mostly due to the inclusion of observations with partly missing data and the fact that missingness in this data was related to the mortality outcome (MNAR). Further analysis is needed to evaluate the performance of this model in an external database. Use of more complex machine learning methods such as an ensemble of

different algorithms (e.g., random forest, support vector machine, neural networks) could improve the

risk stratification performance of models even more.

APPENDIX

Figure 3A. 1. Correlation between predicted probability in LR and random forest

Table 3A. 1. Ranked importance of predictor variables in the random forest model, RBA method-mortality outcome

Random Branch Assignments Variable Importance		
Variable	**Margin**	**MSE**
TUG Answer	0.02621	0.00725
Albumin Result	0.02361	0.00688
Race	0.01657	0.00491
ADL decline	0.01471	0.00319
Cholesterol Result	0.00973	0.00177
KPS value	0.00809	0.00344
Surprise question	0.00769	0.00060
Age	0.00542	0.00209
CCW-Hyperlipidemia	0.00518	0.00044
Diagnosis count	0.00509	0.00010
Tobacco Use	0.00437	0.00073
Lives alone	0.00435	0.00061

Table 3A. 1. (cont'd)		
CCW-Rheumatoid Arthritis, Osteoarthritis	0.00371	0.00132
Dual eligible	0.00366	0.00143
Pressure ulcer	0.00338	0.00066
Number of lab orders	0.00302	0.00050
CCW-Chronic Kidney Disease	0.00302	0.00085
CCW-Anemia	0.00289	0.00106
CCW-Depression	0.00281	0.00110
Number of Medications	0.00278	0.00078
CCW-Atrial Fibrillation	0.00264	0.00045
Sex	0.00264	0.00060
Cancer	0.00253	0.00058
CCW-Heart Failure	0.00220	0.00039
CCW-Hypertension	0.00216	0.00027
CCW-Ischemic heart disease	0.00215	0.00049
CCW-COPD	0.00210	0.00045
CCW-Diabetes	0.00201	0.00046
CCW-Benign prostatic hyperplasia	0.00197	0.00054
CCW-Asthma	0.00190	0.00028
CCW-Colorectal Cancer	0.00187	0.00046
CCW-Osteoporosis	0.00183	0.00021
CCW-Cataract	0.00181	0.00038
CCW-Hip/pelvic fracture	0.00180	0.00042
CCW-Breast cancer	0.00178	0.00037
CCW-Glaucoma	0.00174	0.00033
CCW-Stroke/TIA	0.00172	0.00018
CCW-Lung cancer	0.00165	0.00018
CCW-Prostate cancer	0.00165	0.00032

Table 3A. 2 shows the ranked importance based on the loss reduction method. In this table, the number

of rules presents the times that each variable has been used in the forest in the decision nodes.

Table 3A. 2. Ranked importance of the explanatory variables in the random forest model, the loss
reduction method- Mortality outcome

Loss Reduction Variable Importance					
Variable	Number of Rules	Gini	OOB Gini	Margin	OOB Margin
TUG Answer	1727	0.030819	0.02574	0.061639	0.05698
Race	1061	0.017981	0.01423	0.035962	0.03208
ADL decline	858	0.014258	0.01134	0.028517	0.02581
Surprise question	894	0.008971	0.00569	0.017942	0.01468
CCW-Hyperlipidemia	1471	0.006207	0.00241	0.012414	0.00834
Lives alone	586	0.004102	0.00212	0.008204	0.00618
Tobacco use	592	0.003850	0.00200	0.007699	0.00582
Albumin Result	8229	0.044973	0.00063	0.089947	0.04620
CCW-Cataract	60	0.000158	0.00000	0.000316	0.00012
Alzheimer	0	0.000000	0.00000	0.000000	0.00000
MI	0	0.000000	0.00000	0.000000	0.00000
CCW-Endometrial cancer	0	0.000000	0.00000	0.000000	0.00000
CCW-Lung cancer	11	0.000045	-0.00003	0.000090	0.00000
CCW-Hip/pelvic fracture	19	0.000027	-0.00004	0.000055	-0.00002
CCW-Colorectal cancer	37	0.000163	-0.00006	0.000325	0.00010
CCW-Breast cancer	72	0.000133	-0.00008	0.000266	0.00006
CCW-Prostate cancer	47	0.000134	-0.00012	0.000269	-0.00003
CCW-Asthma	135	0.000253	-0.00016	0.000507	0.00005
CCW-Glaucoma	166	0.000346	-0.00021	0.000692	0.00013
CCW-Benign prostatic hyperplasia	225	0.000470	-0.00035	0.000940	0.00005
Cancer	297	0.000573	-0.00043	0.001146	0.00003
Dual eligible	1213	0.002908	-0.00046	0.005815	0.00225
CCW-Osteoporosis	557	0.000908	-0.00051	0.001817	0.00027

Table 3A. 2. (cont'd)					
CCW-Depression	923	0.001987	-0.00069	0.003974	0.00124
CCW-Stroke/TIA	469	0.000692	-0.00069	0.001385	-0.00006
CCW-Atrial fibrillation	658	0.001462	-0.00089	0.002923	0.00057
Pressure ulcer	570	0.001271	-0.00095	0.002543	0.00043
CCW-Ischemic heart disease	761	0.001176	-0.00109	0.002351	0.00002
CCW-Hypertension	1094	0.001912	-0.00129	0.003824	0.00076
CCW-Chronic kidney disease	1375	0.002707	-0.00140	0.005414	0.00129
CCW-Rheumatoid arthritis/Osteoarthritis	1760	0.003110	-0.00140	0.006220	0.00163
CCW-Anemia	1123	0.002231	-0.00146	0.004461	0.00056
Sex	1260	0.001983	-0.00159	0.003966	0.00050
CCW-COPD	1172	0.001756	-0.00162	0.003512	0.00028
CCW-Acquired hypothyroidism	1001	0.001578	-0.00165	0.003156	-0.00006
CCW-Diabetes	1303	0.001823	-0.00192	0.003645	-0.00020
CCW-Heart failure	1435	0.002131	-0.00206	0.004263	-0.00002
KPS value	6385	0.017562	-0.00672	0.035123	0.01103
Number of lab orders	6468	0.015026	-0.01314	0.030051	0.00174
Diagnosis count	9035	0.021521	-0.01899	0.043041	0.00340
Age	8496	0.025198	-0.02146	0.050396	0.00374
Cholesterol result	9707	0.036575	-0.02203	0.073149	0.01478
Number of Medications	9857	0.024768	-0.02569	0.049536	-0.00013

Figure 3A. 2. Correlation between predicted probability in LR and random forest- Hospice admission
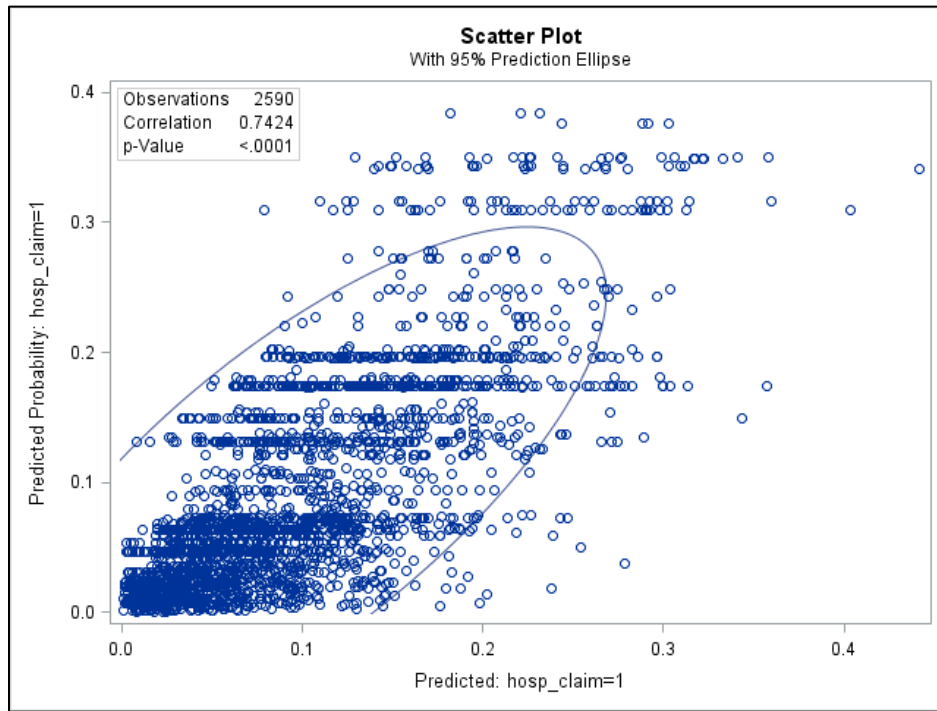
Table 3A. 3. Ranked importance of predictor variables in the random forest model, RBA method- Hospice outcome

Random Branch Assignments Variable Importance		
Variable	Margin	MSE
sq_n	0.00557	0.00232
AgeAtVisit	0.00585	0.00180
KPS_Cat	0.00215	0.00162
NumLabOrders	0.00026	0.00111
Albumin_Result	0.00113	0.00055
Lives_Alone	0.00023	0.00053
race_n	0.00044	0.00051
dual_eligible	0.00042	0.00048
TUG_Answer	0.00211	0.00025
Chol_Result	-0.00037	0.00024

Table 3A. 3. (cont'd)		
adl_decline	0.00084	0.00023
CCW_Hyperlipidemia	0.00235	0.00016
CCW_Stroke_TIA	0.00012	0.00014
CCW_Depression	0.00097	0.00011
CCW_Ischemic_Heart_Disease	0.00006	0.00009
CCW_Diabetes	0.00106	0.00006
CCW_Asthma	-0.00001	0.00005
CCW_Osteoporosis	-0.00007	0.00005
Number of meds	0.00010	0.00004
tobacco use	0.00021	0.00004
CCW_Prostate_Cancer	-0.00004	0.00003
CCW_Lung_Cancer	0.00004	0.00002
CCW_Acquired_Hypothyroidism	0.00026	0.00001
CCW_Hip_Pelvic_Fracture	-0.00003	0.00001
CCW_Hypertension	0.00165	0.00001
CCW_Cataract	-0.00003	0.00000
CCW_Breast_Cancer	0.00005	-0.00000
CCW_Endometrial_Cancer	-0.00000	-0.00001
CCW_Rheumatoid_Arthritis_Osteoar	0.00011	-0.00002
CCW_Glaucoma	0.00001	-0.00003
CCW_Colorectal_Cancer	-0.00005	-0.00004
Cancer	-0.00025	-0.00004
CCW_Anemia	0.00017	-0.00007
CCW_COPD	0.00040	-0.00007
CCW_Atrial_Fibrillation	-0.00024	-0.00009
CCW_Benign_Prostatic_Hyperplasia	-0.00010	-0.00011
CCW_Chronic_Kidney_Disease	0.00024	-0.00012
sex	0.00032	-0.00014
CCW_Heart_Failure	0.00042	-0.00017

Table 3A. 3. (cont'd)		
Pressure_Ulcer	-0.00009	-0.00017
DX_count	0.00043	-0.00054

Table 3A. 4. Ranked importance of predictor variables in the random forest model, Loss reduction method- Hospice outcome

Loss Reduction Variable Importance					
Variable	Number of Rules	Gini	OOB Gini	Margin	OOB Margin
KPS_Cat	741	0.002112	0.00109	0.004223	0.00317
sq_n	771	0.002094	0.00024	0.004189	0.00273
CCW_Hip_Pelvic_Fracture	40	0.000039	0.00001	0.000077	0.00002
CCW_Cataract	37	0.000029	0.00001	0.000059	0.00002
CCW_Endometrial_Cancer	2	0.000002	-0.00000	0.000004	-0.00000
CCW_Lung_Cancer	17	0.000027	-0.00002	0.000054	0.00001
CCW_Breast_Cancer	63	0.000067	-0.00006	0.000135	0.00002
CCW_Colorectal_Cancer	33	0.000042	-0.00006	0.000084	-0.00002
CCW_Asthma	77	0.000074	-0.00007	0.000148	-0.00006
dual_eligible	701	0.001024	-0.00007	0.002048	0.00084
CCW_Prostate_Cancer	69	0.000128	-0.00011	0.000256	0.00004
CCW_Glaucoma	103	0.000105	-0.00011	0.000209	0.00001
CCW_Osteoporosis	320	0.000298	-0.00022	0.000595	-0.00004
CCW_Stroke_TIA	321	0.000344	-0.00029	0.000689	0.00006
CCW_Benign_Prostatic_Hyperplasia	225	0.000291	-0.00031	0.000582	-0.00003
TobaccoUse	377	0.000413	-0.00033	0.000826	0.00017
Cancer	241	0.000303	-0.00034	0.000606	-0.00009
Lives_Alone	528	0.000623	-0.00035	0.001247	0.00020
race_n	691	0.001134	-0.00054	0.002269	0.00047
Pressure_Ulcer	548	0.000629	-0.00060	0.001258	0.00004
CCW_Hypertension	744	0.000853	-0.00062	0.001705	0.00021

Table 3A. 4. (cont'd)					
CCW_Ischemic_Heart_Disease	530	0.000639	-0.00063	0.001278	-0.00000
CCW_Depression	701	0.001008	-0.00065	0.002016	0.00035
CCW_Diabetes	837	0.000844	-0.00070	0.001688	0.00002
CCW_Atrial_Fibrillation	643	0.000818	-0.00073	0.001636	0.00014
adl_decline	728	0.001109	-0.00079	0.002219	0.00002
CCW_Acquired_Hypothyroidism	769	0.000868	-0.00083	0.001736	0.00015
sex	964	0.000955	-0.00085	0.001911	0.00000
CCW_COPD	789	0.000795	-0.00085	0.001590	-0.00021
CCW_Anemia	796	0.000856	-0.00090	0.001712	-0.00005
CCW_Chronic_Kidney_Disease	983	0.000898	-0.00093	0.001797	-0.00004
CCW_Hyperlipidemia	1163	0.001262	-0.00095	0.002523	0.00034
CCW_Heart_Failure	1082	0.001013	-0.00104	0.002027	-0.00009
CCW_Rheumatoid_Arthritis_Osteoar	1120	0.000966	-0.00115	0.001932	-0.00031
TUG_Answer	1302	0.002177	-0.00123	0.004355	0.00093
NumLabOrders	3484	0.006394	-0.00601	0.012788	0.00019
DX_count	5715	0.010744	-0.01101	0.021487	0.00039
AgeAtVisit	5506	0.013971	-0.01138	0.027941	0.00247
Albumin_Result	5381	0.011902	-0.01254	0.023805	-0.00022
NumMeds	6729	0.014074	-0.01565	0.028148	-0.00057
Chol_Result	6148	0.015780	-0.01700	0.031561	-0.00090

Table 3A. 5. Sample of fit statistics from the RF model for hospice outcome

				Fit Statistics			
Number of Trees	Number of Leaves	MSE (Train)	MSE (OOB)	Misclassification Rate (Train)	Misclassification Rate (OOB)	Log Loss (Train)	Log Loss (OOB)
1	265	0.0764	0.1297	0.0881	0.1336	0.972	2.241
2	513	0.0600	0.1267	0.0760	0.1317	0.368	2.029
3	738	0.0543	0.1204	0.0750	0.1260	0.240	1.787
4	988	0.0504	0.1187	0.0690	0.1250	0.183	1.644
5	1250	0.0482	0.1123	0.0690	0.1216	0.166	1.380
6	1497	0.0475	0.1074	0.0731	0.1145	0.166	1.191
7	1775	0.0460	0.1048	0.0720	0.1137	0.163	1.107
8	2016	0.0454	0.1017	0.0723	0.1106	0.161	0.989
9	2293	0.0443	0.0994	0.0717	0.1087	0.159	0.881
10	2553	0.0436	0.0966	0.0715	0.1059	0.157	0.771
100	26097	0.0399	0.0860	0.0905	0.0989	0.150	0.308
101	26390	0.0398	0.0859	0.0905	0.0989	0.150	0.308
102	26664	0.0398	0.0860	0.0903	0.0989	0.150	0.308
103	26938	0.0398	0.0859	0.0905	0.0989	0.150	0.308
104	27190	0.0398	0.0860	0.0911	0.0989	0.150	0.308
195	50953	0.0399	0.0852	0.0943	0.0989	0.150	0.304
196	51181	0.0399	0.0853	0.0940	0.0989	0.150	0.304
197	51439	0.0399	0.0853	0.0940	0.0989	0.150	0.304
198	51706	0.0399	0.0852	0.0940	0.0989	0.150	0.304
199	51939	0.0399	0.0852	0.0946	0.0989	0.150	0.304
200	52219	0.0399	0.0853	0.0948	0.0989	0.150	0.304

- *Introduction*

The need for accurate risk stratification approaches in the population of community-living older adults was discussed in previous chapters. In the second and third chapters, respectively, we developed prediction models by applying the logistic regression (LR) and random forest (RF) modeling approaches to the USMM database. These two models were developed to predict 1-year mortality and the risk of hospice admission over a similar 1-year period. However, the maximum possible follow-up time for the 2015 USMM cohort exceeds two years (maximum follow up is 794 days between Jan 1st, 2015 and Mar 6th, 2017 which is the date of the claims data inquiry). When follow-up was restricted to 12 months there were 2408 (32%) deaths, and 752 (10%) hospice admissions, whereas using all of the available follow-up time (max 794 days), the number of events and associated cumulative incidence rates increased to 3341 (45%) deaths and 1389 (19%) hospice admissions.

To capture the experience of patients who had the outcomes beyond the first year of their USMM care, and to take into account the time-to-event, we used a Cox proportional hazard model to analyze the two outcomes as time-to-death and time-to-hospice. The same cohort of USMM patients, as used in the LR (chapters two) and RF (chapter three) models, were used in this analysis as well. The Cox model's performance metrics were compared to the two alternative approaches, i.e., LR and RF models.

The objective of this chapter is, therefore, to develop and validate multivariable time-to-event (also known as failure time) Cox models for the two outcomes, death and hospice admission; and to compare their performance to the alternative models.

- ***Main concepts and definitions***

  - o <u>Survival analysis methods and the COX PH model</u>

Survival analysis is a branch of statistics that involves the modeling of time to an event. It attempts to answer questions such as what proportion of a population will survive past a certain time point, or what is the failure rate (hazard rate) in the subjects who survived up to a certain time. Analysis of survival time needs special techniques because of the nature of the follow-up time. (140) In survival data, subjects are followed until the outcome happened, but almost always the data is incomplete; some subjects withdraw the study before an event happens; some others do not experience the outcome before the end of the study. These partially observed subjects are called censored observations. There are different methods developed for survival analysis, but the two most popular models are accelerated failure time model (141) and Cox proportional hazard model.(142) Both models have the assumption of a parametric form for the effect of independent predictor variables. The difference between the two models is an assumption for the underlying survival function, where accelerated failure time assumes a parametric distribution for the underlying survival function, the Cox PH model has an unspecified survival function. The parametric form of the predictor variables enters the two models in different ways. Because of these assumptions, the Cox PH model is considered a semi-parametric model. Survival analysis can be used in several ways. For example, Kaplan-Meier (KM) curves estimate the survival function from censored data and provides a summary of the survival experience overall, and in subgroups. The log-rank test can be used to compare KM curves across subgroups. Regression analyses of time-to-event based on the accelerated failure time model or the Cox proportional hazard model serve to quantify the effect of one or more variables on survival time.

  - o <u>Definitions</u>

Let *T* denote a non-negative random variable representing the failure time for an individual in the study population. The survival function is defined as:

$$S(t) = P[T > t]$$

The probability of being event-free at time *t.* The corresponding hazard function, denoted by λ (*t*), is

defined as the instantaneous probability that the event of interest happens soon after time *t.* The

formal definition of λ (*t*), for a continuous time *T* is:

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

where *f (t)* is the probability density function of *T*. (142,143)

Cox PH model is widely used in survival data analysis to evaluate the effect of explanatory variables **Z** on

the hazard function. For each subject (index *i*) in the population his or her hazard function is expressed

as:

$$\lambda_i(t) = \lambda(t; \mathbf{Z}_i) = \lambda_0(t) \exp(\mathbf{Z}_i' \boldsymbol{\beta})$$

Where λ_0 (*t*) is unspecified, called the baseline hazard, $\mathbf{Z}_i$ is the vector of fixed (i.e., not time-

dependent) covariates for the *i*th subject, and $\boldsymbol{\beta}$ are coefficients associated with $\mathbf{Z}_i$ and assumed to be

the same for all subjects. The terminology "proportional hazards" come from the fact that for any two

subjects the hazard ratio $\lambda_i(t) / \lambda_j(t)$ is constant in time.

Hazard ratio is the ratio of two hazard rates corresponding to the two levels of an explanatory variables.

For example, patients with severe functional impairment may die at twice the rate per unit time as the

patients with no functional impairment.

The Cox model allows both continuous and categorical explanatory variables in the model; also, it

supports multivariable models while the KM method is inconvenient when faced with continuous

explanatory variables. (143,144).

Consider the following Cox model with three explanatory variables: $\lambda(t) = \lambda_0(t) \exp(x_1 \beta_1 + x_2 \beta_2 + x_3 \beta_3)$.

Holding x_2, x_3 constant and increasing x_1 by *d* units gives the hazard ratio (HR) $\exp(\beta_1 d)$. This gives us

the interpretation of $\exp(\beta_1)$ as the hazard ratio associated with a unit increase in x_1 with the other two covariates held fixed. If x_1 were binary, then $\exp(\beta_1)$ is the HR comparing group coded x_1 =1 with the group coded x_1 =0.

o <u>Performance evaluation</u>

Survival analysis models such as Cox PH are a useful approach to develop prediction models. In the LR model, the performance of a prediction model is assessed using discrimination measures such as AUC, and also calibration plots to evaluate the accuracy of the model's predictions. Equivalent discrimination measures have been developed for the Cox PH analysis. Three summary statistics can be generated for the Cox model: concordance (also known as C-index), AUC (at a specific time), and integrated AUC (iAUC). Concordance is defined as the proportion of all usable pairs of subjects for which the greater event risk was predicted for the one that experienced the event earlier. The concordance statistic in the Cox model is called C-index and can be calculated in PROC PHREG by option 'concordance'. (1,4) With some modification equivalent measures to the ROC and AUC generated for LR and RF models can also be produced from the Cox model if the measures are generated for a specific time point during the follow-up period. The definition of AUC at each time point is the same as the concordance definition, but it is limited to the events that occurred up to a specific time point at which the AUC was generated (e.g., 6-month, 1-year). It is called time-dependent AUC, and it changes at each event time point. The changes in AUC over study time can also be plotted and integrated. The integrated AUC (iAUC) is an average of the AUC at all possible time points in the study period. The C-index, time-dependent AUC, and iAUC are generated in the Cox model using PROC PHREG. (146)

Unlike the LR model, calibration for the Cox model is sparsely discussed in the literature. Calibration is a way to validate a predictive model by evaluation of its predictive accuracy, however, assessing the calibration of the Cox model is not straightforward because the predictions have to be made relative to an unspecified baseline function. (147)

o   <u>Proportional hazard assumption</u>

Proportionality assumption is the central assumption of the Cox model; hence, the model is often called

the Cox Proportional Hazard (PH) model. The PH assumption means that for any two individuals, the

hazard ratio is proportional over the follow-up time and does not depend on time. Comparing two

covariate profiles, x1, x2, the ratio of the hazards, h(t|x1) to h(t|x2) is constant in time t.

To evaluate the coefficient $\beta1$, given the x2 and x3 are the same for both subjects, the hazard ratio is

exp($\beta1$). To test the PH assumption in the Cox model, different methods have been proposed. (7) The

most common way is to generate the KM survival function for the predictor variables and observe if the

curves cross each other; non-parallel curves suggest a violation of the PH assumption that can then be

tested using a more formal statistical test.

 If x1 and x2 represent two groups, say, a treatment and a control group, the Kaplan-Meier plot of their

survival functions should not cross, if the PH holds. This is NOT a formal test of the PH assumption, but

gives a quick graphical check to see if the assumption is plausible.

A formal test of the PH assumption can be approached in two ways: [1] as suggested by Cox in his 1972

article, proportionality in a covariate x is tested by include a time-dependent term in the Cox model. This

term is x times g(t) where g(t) is often the function g(t)=log(t/c), where c is a constant—e.g, the median

follow-up time. If the regression coefficient of the interaction of x with g(t) is significant (P-value <0.05),

a violation of the PH assumption for x is indicated. There is also an overall Wald test for testing all the

interaction coefficients together. The statement TEST in PROC PHREG conducts this test. (8)

[2] a more sophisticated test of the PH assumption is the supremum test. It involves generating a few

simulations of the score process for a covariate x, and comparing it to the observed process. The

supremum test is similar in spirit to the Kolmogorov-Smirnov test comparing the maximal departure

between the observed and expected, SAS PHREG implements the procedure via the ASSESS statement.

This method takes a long time to generate the plots for each predictor levels. Therefore to test the PH

assumption of the Cox model, I used the KM survival plots and tested the 2-level interaction terms in this chapter. If the proportionality assumption is violated, then there are alternative approaches for handling non-proportionality in the Cox model; for example including a time interaction term in the model, or using accelerated failure time modelling approach.(149)

- *Literature review*

Two comparable studies used the Cox PH model to develop a prognostic model in a population of community-living older adults.

In 1998, Fried et al., developed a prognostic score in a cohort of 5201 adults aged 65 years and older to predict the 5-year mortality using a Cox PH model. (58) These adults participated in the cardiovascular health study (CHS) in 4 states – California, Maryland, North Carolina, and Pennsylvania. The 5-year mortality rate in this population was 12%. There is a difference between this cohort and USMM cohort in exclusion criteria; in the CHS cohort patients were excluded if they were wheelchair-bound, or were unable to participate in the examination at the field center, or were under cancer treatment; none of these groups were excluded from the USMM cohort. In fact, USMM patients were all home-bound based on the CMS definition (Chapter 1). The major difference between these two cohorts was their mortality rate (32% a year in USMM population vs. 2.5% a year in CHS cohort). Fried et al., assessed 78 characteristics including demographics, social, functional, physical examination, and comorbidity variables and found 20 variables to be predictors of 5-year mortality including demographics (age, gender, income), lifestyle (physical activity, smoking), comorbidities (heart failure), physical examination (systolic blood pressure, body weight), lab tests (albumin, creatinine, fasting blood sugar), respiratory test, ECG abnormalities, and echocardiography findings (Table 1.1). They included missing data on predictors as a legitimate level of the variable. They validated the model in a separate cohort of the same study by computing a risk score for each individual and then comparing the mortality rate between quantiles of the prognostic score in both the derivation and validation data. This study found a

significant difference in the mortality rates in quantiles of the prognostic score in the validation data; however, it did not provide a discrimination measure or any other performance metric for the model. (58) The proportional hazard assumption was assessed by testing the interaction between time and covariates.

In 2008, Carey et al. conducted a multi-State US-based study and developed a prognostic index to predict mortality in a cohort of community-based, chronically-ill, frail older adults. (57) This cohort had 1-year and 3-year mortality rates of 13% and 37% respectively. Carey et al. used a cox model to develop the prognostic index in the derivation cohort (n=2232) and then validate the index by applying it to the validation cohort (n=1667). They found eight variables (two demographics [age, gender], two functional [dependence in bathing, and dressing], and four comorbidities [cancer, heart failure, COPD, and chronic kidney insufficiency]) in the Cox model as significant predictors of mortality. They then developed a risk score by assigning different points to the predictor variables based on the coefficient from the Cox model. The risk score ranged from 0-14, and they assigned a 3-level risk value to each patient based on their score (i.e., 0-3 low risk, 4-5 intermediate-risk, and >5 high-risk). They compared the 1- and 3- year mortality rates between the different risk levels. They reported a good calibration based on the similarity of the mortality rates between the derivation and validation cohorts. They also reported the AUC of 0.66 and 0.69 for derivation and validation data.

Both study populations described above have differences with the USMM cohort. The first population was generally healthier, and younger (mean age=73 years) than the USMM cohort (mean age=82), and unlike USMM patients, they were not homebound. The 5-year mortality rate in this cohort was 12% (Grossly estimated one-year mortality of 2.5%) vs. 32% one year mortality rate in the USMM cohort. The second study population is more similar to the USMM cohort in terms of the age (mean age=79 years) and overall vulnerability, however, the mortality rate in this population is 13% a year, whereas it is 32% in USMM cohort. The PACE study population are also eligible for nursing-home by confirmation from the

State's Medicare staff. Both data developed the Cox model to select the predictors and then generate a risk score from those variables. To validate the model, both studies applied the risk score to the validation data and then reported the mortality rate in order to evaluate the accuracy of the model. Fried et al. did not provide discrimination measures from the Cox model, but they evaluated the proportionality assumption by testing the interaction between time and each predictor.

- ***Methods and materials***

Data source- A Cox proportional hazard model was developed utilizing the same USMM dataset as used in the LR and RF model chapters. The main difference is that all of the available follow-up time was used to identify outcomes in this analysis, whereas in the LR and RF, only outcomes that occurred within 12 months of the first visit were analyzed. Therefore outcome events (deaths and hospice admissions) that occurred after the first year of patients' registration up until the end of follow-up (median=1.4 years, max = 2.2 years) were included in this analysis. The USMM claims data was again used to determine if an outcome (death or hospice admission) occurred and if so, to identify its date. The final date of Claims data inquiry (Mar 6th, 2017) was used as the administrative end date of the study period; all subjects still in follow-up were censored at this point. Hospice coverages were reported in 3-month intervals; therefore, the date of first hospice admission was used to calculate the outcome of time-to-hospice.

Study population- As with the prior analyses, the 2015 cohort was defined as all patients who had their first-ever medical visit by a USMM provider between January 1st and December 31st, 2015. Since the outcomes of interest were recorded in the claims data, the USMM EMR data was linked to the claims data, and those patients who had claims data available were included. Patients with age<65 years were excluded. Like the previous chapters of this dissertation (LR and RF models), the cohort was limited to those who received care from the USMM for at least 12 months. In other words, if a patient was withdrawn from USMM care within the first year of registration, they were excluded. Also, four patients

were excluded due to time-to-event of zero or negative (incorrect date of event). The study population

in this chapter is the same as the chapters 2 and 3, except for the four patients that were excluded in

this chapter due to time to event less than 1 day. Table 4.1 contains the inclusion and exclusion criteria

for this population, and Figure 4.1 shows the flow diagram of the study population.

Table 4. 1. Inclusion and exclusion criteria for the Cox cohort

Inclusion criteria
- Register in the USMM system in the calendar year 2015
- Had at least one USMM visit January 1st and Dec 31st, 2015
Exclusion criteria
- Claims data not available
- Age <65 years old
- Less than 12 months care by USMM (withdrawal in the first year)
- Time-to-event <1 day

Figure 4. 1. Flow diagram of the study population

```
        ┌─────────────────────────┐
        │  Patients who had their │
        │   first-ever USMM visit │
        │          in             │
        │         2015            │
        │       N=20.424          │ ⋯⋯⋯⋯⋯⋯⋯➤  No claims data available
        └────────────┬────────────┘
                     │
                     ▼
        ┌─────────────────────────┐
        │   Claims-linked cohort  │
        │       N=12,634          │ ⋯⋯⋯⋯⋯⋯⋯➤  Age< 65 years old
        └────────────┬────────────┘
                     │
                     ▼
        ┌─────────────────────────┐
        │    Age≥ 65 years old    │
        │        N=9,627          │ ⋯⋯⋯⋯⋯⋯⋯➤  USMM care < 1 year
        └────────────┬────────────┘
                     │
                     ▼
        ┌─────────────────────────┐
        │  Received ≥ 1-year care │
        │        N=7,445          │ ⋯⋯⋯⋯⋯⋯⋯➤  Follow up time < 1 day
        └────────────┬────────────┘
                     │
                     ▼
        ┌─────────────────────────┐
        │      Final cohort       │
        │        N=7,441          │
        └─────────────────────────┘
```
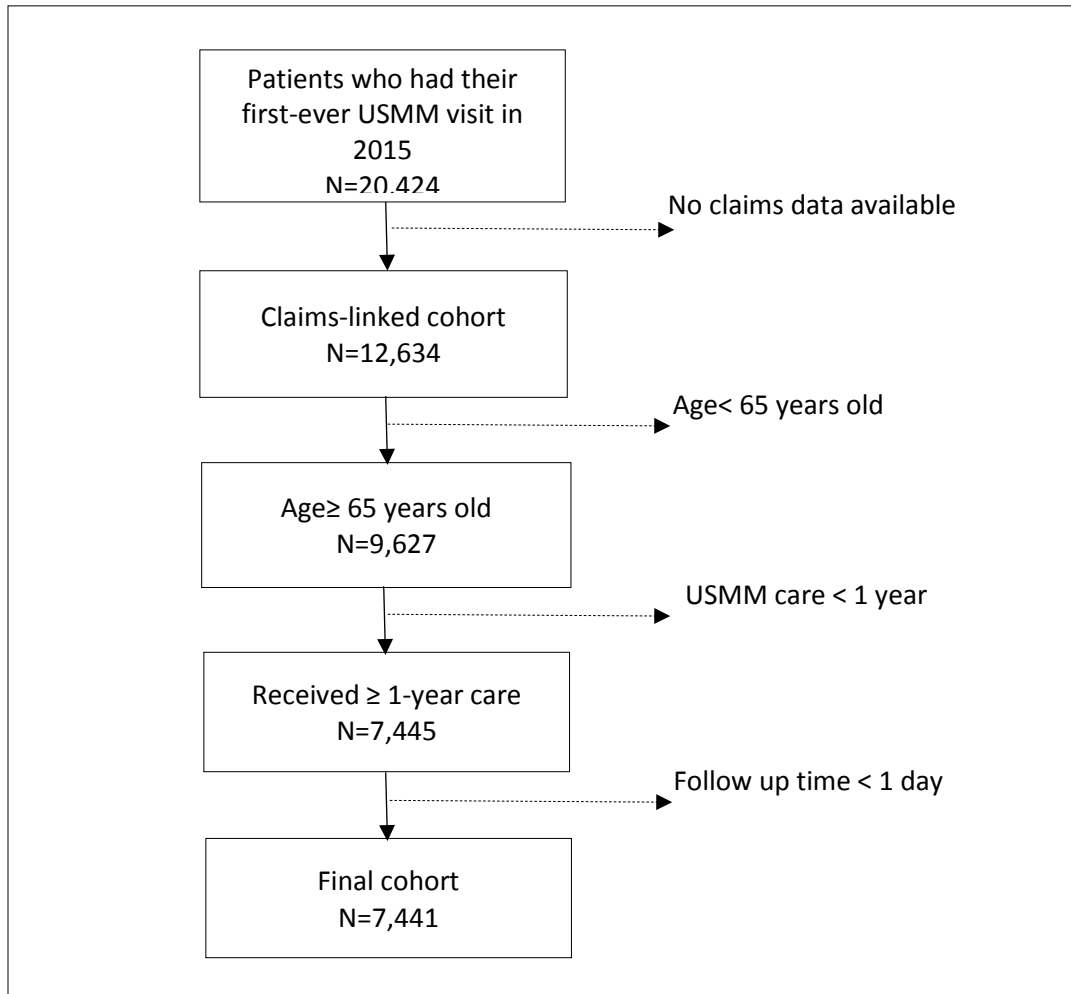
A total of 2182 patients were excluded because they had been followed up for less than 12 months; the

reasons for their withdrawal have been summarized in chapter two (Table 2.2). There were additional

four patients excluded due to time-to-event of zero (the event occurred at the same date as the first-

ever visit) or negative (incorrect date of event).

Outcomes- Time-to-death for the patients who deceased, was calculated as the number of days

between the date of the first visit (recorded in the USMM 2015 data) and the date of death (recorded in

the claims data). For those who survived, the follow-up time was calculated as the number of days

between the date of their first visit and the end of follow up defined as the date of the claims data

inquiry (March, 6th, 2017). If a patient had no outcome reported before the end of the study, then the follow-up time was the days between the first visit date and 03/06/2017, which is referred to as administrative censoring. Therefore the longest possible follow up time was 794 days or 2.2 years (the number of days between Jan 1st, 2015 and Mar 6th, 2017). Time-to-hospice was calculated for the patients who were admitted to hospice as the number of days between the date of their first visit and the date of first hospice admission (in claims data). For those who did not have a hospice admission, the follow-up time was again calculated as the number of days between the first visit and the end of follow up (i.e., Mar 6th, 2017).

Exposure variables- As with our previous approach, only variables with less than 20% missing observations were considered for the analysis. The same 41 variables as used in LR and RF models, were also included in this analysis. These were: Demographics: age, gender, race; socioeconomic status: insurance status representing if a patient has dual eligibility for both Medicaid and Medicare, living alone, smoking; functional status: functional decline in ADLs, timed up and go (TUG), Karnofsky performance scale (KPS value); lab tests: serum albumin, cholesterol; and other variables: having a pressure ulcer, surprise question answer, number of medications, and number of lab test ordered by the provider. There are 24 medical history variables as listed in the Chronic Condition Warehouse (CCW) variables: history of hypothyroidism, asthma, atrial fibrillation, cataract, chronic kidney diseases, osteoporosis, hyperlipidemia, hypertension, anemia, breast cancer, colorectal cancer, benign prostatic hyperplasia, COPD, depression, diabetes, endometrial cancer, glaucoma, heart failure, hip/pelvic fracture, ischemic heart disease, lung cancer, prostate cancer, stroke/TIA, rheumatoid arthritis/osteoarthritis. Diagnosis count is a variable that counted the number of existing CCW conditions for each patient. Another variable, cancer, was generated if a patient had one or more of the four cancers listed in the CCW variables.

o   <u>Statistical analysis</u>

The statistical analyses for this study was done using SAS software (SAS Institute Inc., Cary, NC, version 9.4). Dataset of 7441 was split into two equal-size datasets, termed derivation (n=3721) and validation (n=3720), using SAS procedure SURVEYSELECT. These derivation and validation groups are the same as were used in chapters two and three.

The Kaplan Meier (KM) survival plots were generated for the total population for both outcomes. PROC LIFETEST was used to generate the KM plots.

The Cox model was developed for each outcome using the derivation data and then applied to the validation data. Time-to-death and time-to-hospice were analyzed using PROC PHREG to develop a Cox regression model that examined all the 41 predictor variables. Different variable selection methods were examined, including automatic and manual selection methods. Automatic (SAS built-in) variable selection methods including stepwise, forward, and backward were specified in separate models and the model's performance, and the number of selected variables were compared. Also, the same manual selection method that was described for LR methods in chapter two was utilized; briefly variables that were significant in univariate analysis (p-value< 0.2), were put in a multivariable Cox model and those with p-value< 0.05 were included in the final manual selected model.

The performance measures were generated for each model for a comparison between the different variable selection methods. To compare these Cox models to the predictive models developed in the previous two chapters, I generated AUC statistics. To have comparable metrics from the Cox model, three summary statistics were generated for the final models: concordance (also known as C-index), AUC (at day 365), and integrated AUC (iAUC). (146) Concordance in the Cox model has an equivalent interpretation as the C-statistic in the LR model only the Cox model considers the timing of the event. For the Cox model, concordance is the proportion of all usable subject pairs in which the case with the higher risk predictor had an event before the case with the lower risk predictor. (146,147) In other

words, the concordance is the fraction of all pairs where the predictor score is higher for the individual

with the earlier event. Usable pairs are pairs that one or both subjects had an event. There are different

methods to generate Concordance statistic in survival analysis, naming UNO and Harrell. I used Harrell's

method, which is the default method in SAS. (140) Harrell's option in PROC PHREG also provides the

standard error for the concordance which can be used to calculate the confidence limits.

The ROC for the Cox model is sensitive to the time and can be generated at any time point in the study

period, hence called time-dependent ROC. (146) The time-dependent ROC and its respective time-

dependent AUC varies only at the event times; i.e., AUCs at the time points between the two event

times are the same. For comparison to the results from 1-year mortality in LR and RF models, I

generated an AUC and the 95% confidence limits at day 365 for Cox models using PROC PHREG options.

(150) The AUC (365) has the same definition as concordance except in AUC (365) only events that

happened between day 0 and 365 are counted. Changes in the time-dependent AUCs generate a plot

that shows AUCs and the confidence limits at all possible time points. The integrated AUC (iAUC) is an

average of all AUCs over all time points. (146) To generate ROC at day 365 and integrated AUC (iAUC) for

the model, ROCOPTION was specified in the PROC PHREG statement.

Proportionality assumption is the central assumption of the Cox model; hence, this assumption must be

satisfied for the Cox model to be an appropriate model. The proportional hazard assumption means the

hazard ratio between two individuals is independent of the time. The statistical details of the

assumption were presented in the background section. The PH assumption in this model was tested for

all variables of the final model using two methods, KM survival plots (with examining on non-parallel

curves) and testing the 2-way interaction between each covariate and time. The KM survival curve was

generated in the derivation data, stratified by the key covariates. Also, interaction terms were generated

in the PROC PHREG by multiplication of the predictors by the log function of the follow-up time divided

by the median of it. For example, the interaction of age by time was made by this formula

Age*Time=Age* log (Time/median of Time). This method of making interaction terms between the

covariates and a log function of time was first introduced in the original paper by David Cox.(142) The

time variable is divided by a constant (often its median) in order to stabilize the estimates in the model.

Then a log function of this product will be used in interaction terms. The interaction terms were added

to the model one by one and evaluated for significance at the $p < 0.05$ level. Additionally, an overall PH

test was done for all the interaction terms together in PROC PHREG. The PH test is testing the

hypothesis that all interaction's coefficients are zero. If the PH assumption is violated, the analysis will

be performed stratified by the predictor that has significant interaction term.

The model performance metrics (i.e., AUC and C-index) were then compared to the results of LR and RF

models using the validation data. The importance of explanatory variables in the final model was

assessed by their coefficient estimates and then were compared with the alternative models.

Although calibration plots were generated for the LR and RF model to validate the accuracy of

predictions, they are not as useful in the Cox model as in the previous approaches. Generation and

interpretation of the calibration plots in the Cox model is not straightforward because the predictions

are time-dependent. Therefore, a comparison of the calibration plots between the Cox and other two

models is not conducted.

- ***Results***

  o <u>Study population</u>

The starting population for this analysis consisted of 20424 patients who joined the USMM in 2015 and

had at least one visit in 2015. Since the outcomes of interest were reported in the claims data, those

with no claims data available (n=7790) were excluded. Also, 3007 patients with age <65 years old were

excluded from the analysis, because the objective of this study was to develop a risk stratification model

for the older adults who live in the community.

Additionally, 2182 patients were excluded because they were under USMM care for less than 12 months. These patients did not have an outcome, and their last documented visit with the USMM was less than 1-year from their first visit. Finally, the four patients who had time to event < 1 day were removed. Exclusion of these four patients are the only difference between the study population in this chapter and in chapters two and three. Two of these four patients died at the same date as the first visit, and the other two had negative follow up time (which is assumed to be due to a mistake in data entry). Figure 4.1 shows the flow diagram of the study cohort. The final cohort consisted of 7441 subjects. Table 4.2 displays the baseline characteristics of the patient population, as well as the unadjusted hazard ratios for both outcomes (time-to-death and time-to-hospice). In this cohort of 7441 patients, the average age was 82 years with a standard deviation of 9, 66% were female, 63% white, 99% had Medicare coverage, and 27% were dual-eligible. Prevalence of comorbidities in this population included 81% hypertension, 51% hyperlipidemia, 34% diabetes, 26% COPD, and 7% cancer. Impaired functional status was documented in this population by three variables: KPS (54% severe need for assistance), TUG (45% abnormal test or non-ambulatory), and ADL (14% decline in ADL). In the univariate analysis of the CCW comorbidities, 13 variables (for mortality outcome) and nine variables (for hospice outcome) had the significant unadjusted hazard ratio less than 1.0, which means that a positive history of the disease was associated with a lower hazard of the outcomes (death or hospice admission). Overall the characteristics of this cohort were the same as the cohort of 7445 patients analyzed in the prior two models, LR and RF (see Tables 2.3 and 3.1).

Table 4. 2. Study population characteristics and association of predictors with the outcomes (N=7441) over an average of 459 days of follow-up

Variable	N (%)	Missing N (%)	Death %	Unadjusted HR	Hospice %	Unadjusted HR
Baseline characteristics						
Age						
-65 -74	1826 (24.5)		29.4	Ref	9.0	Ref
-75 – 84	2247 (30.2)	0	42.7	1.60*	16.8	2.17*
-85 – 94	2794 (37.6)		54.2	2.22*	24.3	3.69*
-95+	574 (7.7)		58.0	2.38*	29.3	4.47*
Sex						
-Male	2512 (33.8)	0	49.8	1.25*	18.9	1.11
-Female	4929 (66.2)		42.4	Ref	18.6	Ref
Race						
-White	4681 (62.9)	1411	40.4	Ref	21.0	Ref
-Black	1148 (15.4)	(19.0)	29.3	0.67*	11.8	0.50*
-Other	201 (2.7)		28.4	0.65*	11.0	0.46*
Tobacco use (current vs not)						
-Yes	644 (8.7)	387 (5.2)	32.0	0.67*	13.8	0.65*
-No	6410 (86.1)		44.0	Ref	19.0	Ref
Dual-eligible						
-Yes	2024 (27.2)	0	34.2	0.62*	13.4	0.54*
-No	5417 (72.8)		48.9	Ref	20.6	Ref
Lives alone						
-Yes	884 (11.9)	1049	27.7	0.56*	11.2	0.46*
-No	5508 (74.0)	(14.1)	43.7	Ref	20.1	Ref
S.Q- No						
-No	1044 (14.0)	1017	61.6	2.01*	29.7	2.6*
-Yes	5380 (72.3)	(13.7)	37.3	Ref	16.6	Ref
KPS						
-Mild /moderate (50-100)	3376 (45.4)	27 (0.4)	32.2	Ref	12.8	Ref
-Severe disability (10-40)	4038 (54.3)		55.3	2.08*	23.7	2.54*
TUG						
-<30 sec	2538 (34.1)	1499	28.6	Ref	15.2	Ref
-≥30 sec	1377 (18.5)	(20.1)	35.3	1.28*	19.2	1.35*
-Non-ambulatory	2027 (27.2)		44.9	1.80*	20.4	1.61*
Decline in ADLs						
-Decline	1062 (14.3)	1181	46.4	1.22*	24.3	1.43*
-Improve	311 (4.2)	(15.9)	10.0	0.21*	9.3	0.39*
-No change	4887 (65.7)		39.4	Ref	18.2	Ref
Pressure ulcer						
-Yes	940 (12.6)	0	53.1	1.28*	22.1	1.35*
-No	6501 (87.4)		43.7	Ref	18.2	Ref

Table 4. 2. (cont'd)						
cancer						
-Yes	565 (7.6)	0	49.6	1.18*	19.3	1.13
-No	6876 (92.4)		44.5	Ref	18.6	Ref
Cholesterol result (mg/dl) Quartiles						
-<136	1554 (20.9)	1054	53.0	1.92*	19.8	1.40*
-136 - <164	1623 (21.8)	(14.2)	38.9	1.25*	18.1	1.08
-164 - <195	1589 (21.3)		37.4	1.17*	18.3	1.05
- 195+	1621 (21.8)		33.0	Ref	18.0	Ref
Albumin result (g/dl) Quartiles						
-<3.2	1667 (22.4)	636 (8.6)	67.1	4.19*	23.5	3.33*
-3.2 – <3.5	1609 (21.6)		44.0	2.15*	20.0	2.01*
-3.5 – <3.8	1820 (24.5)		34.2	1.55*	18.5	1.66*
-3.8+	1709 (23.0)		23.7	Ref	12.5	Ref
Medical history						
Hypothyroidism						
-Yes	2050 (27.5)	0	43.1	0.90*	18.4	0.92
-No	53915 (72.5)		45.6	Ref	18.8	Ref
Myocardial infarction						
-Yes	3 (0.04)	0	66.7	1.65	33.3	1.87
-No	7438 (99.9)		44.9	Ref	18.7	Ref
Anemia						
-Yes	2243 (30.1)	0	40.2	0.78*	19.6	0.96
-No	5198 (69.9)		46.9	Ref	18.3	Ref
Asthma						
-Yes	309 (4.2)	0	33.3	0.65*	14.9	0.65*
-No	7132 (95.9)		45.4	Ref	18.8	Ref
Atrial fibrillation						
-Yes	1231 (16.5)	0	53.0	1.29*	21.9	1.31*
-No	6210 (83.5)		43.3	Ref	18.0	Ref
Benign prostatic hyperplasia						
-Yes	504 (6.8)	0	45.2	0.99	20.0	1.05
-No	6937 (93.2)		44.9	Ref	18.6	Ref
Breast cancer						
-Yes	224 (3.0)	0	39.7	0.86	16.5	0.84
-No	7217 (97.0)		45.1	Ref	18.7	Ref
Cataract						
-Yes	184 (2.5)	0	21.2	0.39*	7.6	0.31*
-No	7257 (97.5)		45.5	Ref	19.0	Ref
Chronic kidney diseases						
-Yes	3005 (40.4)	0	38.0	0.67*	19.3	0.88*
-No	4436 (59.6)		49.6	Ref	18.2	Ref

Table 4. 2. (cont'd)						
Colorectal cancer						
-Yes	94 (1.3)	0	51.1	1.22	20.2	1.20
-No	7347 (98.7)		44.8	Ref	18.7	Ref
COPD						
-Yes	1945 (26.1)	0	42.1	0.88*	17.0	0.82*
-No	5496 (73.9)		45.9	Ref	19.3	Ref
Depression						
-Yes	1615 (21.7)	0	36.6	0.69*	19.0	0.86*
-No	5826 (78.3)		47.2	Ref	18.6	Ref
Diabetes						
-Yes	2518 (33.8)	0	40.8	0.82*	16.3	0.74*
-No	4923 (66.2)		47.0	Ref	20.0	Ref
Endometrial cancer						
-Yes	27 (0.4)	0	40.7	0.84	22.2	1.11
-No	7414 (99.6)		44.9	Ref	18.7	Ref
Glaucoma						
-Yes	337 (4.5)	0	41.0	0.87	16.3	0.81
-No	7104 (95.5)		45.1	Ref	18.8	Ref
Heart failure						
-Yes	2541 (34.1)	0	41.5	0.84*	18.5	0.90
-No	4900 (65.9)		46.7	Ref	18.8	Ref
Hip fracture						
-Yes	81 (1.1)	0	48.2	1.06	22.2	1.16
-No	7360 (98.9)		44.9	Ref	18.6	Ref
Hyperlipidemia						
-Yes	3686 (49.5)	0	35.9	0.56*	17.0	0.64*
-No	3755 (50.5)		53.7	Ref	20.3	Ref
Hypertension						
-Yes	6055 (81.4)	0	42.2	0.63*	18.3	0.69*
-No	1386 (18.6)		56.8	Ref	20.5	Ref
Ischemic heart diseases						
-Yes	1269 (17.1)	0	45.2	0.99	21.3	1.15*
-No	6172 (82.9)		44.9	Ref	18.1	Ref
Lung cancer						
-Yes	70 (0.9)	0	66.7	1.88*	21.4	1.59
-No	7371 (99.1)		44.7	Ref	18.6	Ref
Osteoporosis						
-Yes	818 (11.0)	0	33.3	0.63*	19.2	0.86
-No	6623 (89.0)		46.3	Ref	18.6	Ref
Prostate cancer						
-Yes	175 (2.4)	0	56.0	1.38*	18.6	1.34
-No	7266 (97.7)		44.6	Ref	21.1	Ref
Osteoarthritis						
-Yes	2760 (37.1)	0	37.0	0.65*	18.7	0.82*
-No	4681 (62.9)		49.5	Ref	18.7	Ref

Table 4. 2. (cont'd)							
TIA/stroke							
-Yes	799 (10.7)	0	45.3	0.97	23.7	1.28*	
-No	6642 (89.3)		44.9	Ref	18.1	Ref	
Continuous variables[†]							
Age (mean± sd)	82.2 ± 9.3	0	--	1.03*	--	1.06*	
Albumin g/dl (mean± sd)	3.4 ± 0.5	636 (8.6)		0.32*		0.41*	
Cholesterol mg/dl (mean± sd)	167.7 ± 44.0	1054 (14.2)		0.99*		0.99*	
Number of lab tests (Median, IQR)	0 (0 – 5)	0	--	1.01*	--	0.97*	
Number of medications (Median, IQR)	9 (5 – 13)	0	--	0.98*	--	0.98*	
Diagnosis count (Median, IQR)	5 (3-6)	0	--	0.86*	--	0.92*	
Variables that were not included in the analysis due to >20% missing observations							
Decline IADLs							
-Decline	730 (9.8)	5203	10.7	0.90	7.7	0.62*	
-Improve	524 (7.0)	(69.9)	7.3	0.60*	7.6	0.61*	
-No change	984 (13.2)		12.0	Ref	12.3	Ref	
Global health compared to a year ago							
-Better	55 (0.7)	5886	30.9	0.74	14.6	0.86	
-Worse	315 (4.2)	(79.1)	72.4	2.50*	27.6	2.96*	
-The same	1185 (15.9)		38.5	Ref	15.1	Ref	
Fall since last visit							
-Yes	184 (2.5)	5712	49.5	1.08	15.8	0.86	
-No	1545 (20.8)	(76.8)	46.1	Ref	17.8	Ref	
Hospitalization since last visit							
-Yes	870 (11.7)	5007	57.9	o.84*	18.2	1.18	
-No	1564 (21.0)	(67.3)	65.8	Ref	14.5	Ref	
ER since last visit							
-Yes	788 (10.6)	5005	45.1	0.54*	17.1	0.83	
-No	1648 (22.2)	(67.3)	67.4	Ref	14.8	Ref	
Lost weight							
-Yes	1243 (16.7)	3767	37.4	3.53*	23.7	2.35*	
-No	2431 (32.7)	(50.6)	13.1	Ref	12.9	Ref	

IQR: interquartile range; sd: standard deviation; S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; IADL: instrumental activities of daily living; TIA: transient ischemic attack; FU: follow-up; mg/dl: milligram per deciliter; g/dl: gram per deciliter;

[*] P-value < 0.05 in univariate analysis with the outcomes;

[†] The unadjusted HR for continuous variables were generated for 1 unit change in the independent variable; however, only three variables were included as continuous in the analyses: Number of meds, number of labs, and diagnosis count;

The maximum and minimum follow up time for the mortality outcome were 1 and 794 days,

respectively; with a median of 517 days (q1=246, and q3=658) and mean of 459 days or 1.25 years. The

median and mean follow-up time for hospice outcome were 497 and 440 days, respectively. From the

7441 patients, 45% (n=3341) died over the FU period, and 19% (n= 1389) were admitted to hospice. Of

those admitted to hospice 1122 (81%) died in hospice by the end of follow-up in March 2017. Overall

2219 deaths (66% of all deaths) occurred outside of hospice. A total of 3833 patients were censored at

the end of the study without experiencing the outcomes. Table 4.3 displays the follow-up time and the

frequency of outcomes.

Table 4. 3. Follow-up time and outcomes in the Cox study cohort (N=7441)

Variable	N (%)
Outcome: death	
Number of deaths over the total follow-up time	3341 (44.9)
Follow-up time in days -mean ± sd -median (q1 - q3)	459 ± 239 517 (246 - 658)
Outcome: hospice admission	
Number of hospice admissions over the follow-up time	1389 (18.7)
Follow-up time in days -mean ± sd -median (q1 - q3)	440 ± 242 497 (207 - 647)
Hospice admitted patients (n=1389)	
Number of deaths in the hospice over the follow-up time	1122 (80.8)
Follow up time from hospice admission (Time to death or censoring) (days) -mean ± sd -median (q1 - q3)	104 ± 116 58 (10 - 169)

o   Outcome: one-year mortality

Figure 4.2 illustrates the Kaplan Meier (KM) survival curve for the total cohort (n=7441). In this cohort, 3341 (45%) events (deaths) happened over the follow-up time, and 4100 (55%) observations were censored at the end of follow-up which means they were alive at the administrative end date.

Figure 4. 2. KM survival plot for the whole data (N=7441)
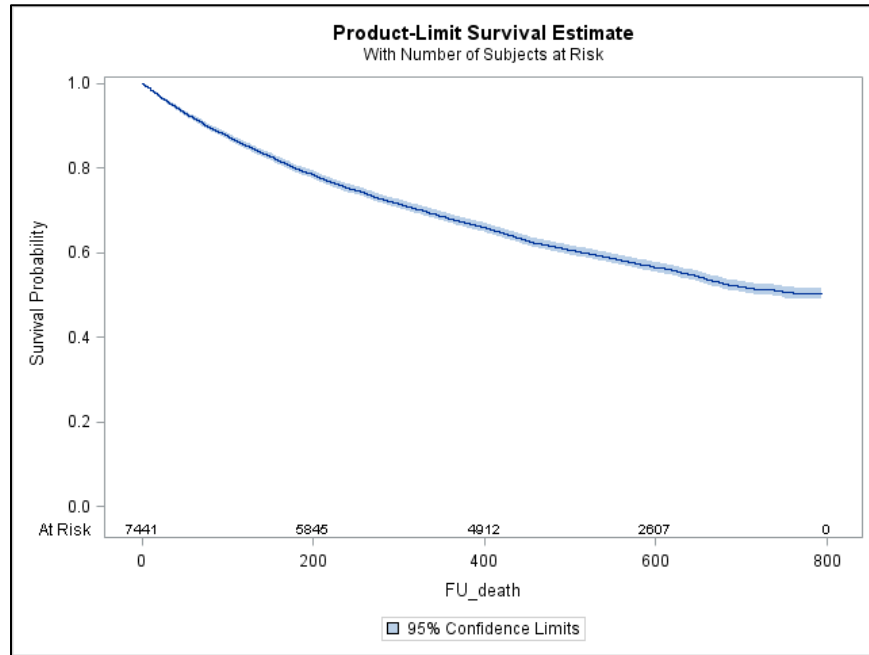
The number of at-risk patients is shown inside the plot over the time axis

Figure 4.3 is the estimated hazard rate over the follow-up time for the whole population. The time unit in hazard rates analyses, is day. The hazard rate is the highest at the beginning of the follow-up and decreases over time with two spikes of increase at about 450 and 650 days. Overall the hazard rates are not constant over time; however, the difference between the maximum and minimum hazard rates are relatively small (0.04% vs. 0.15%).

Figure 4. 3. Hazard rate estimates for the whole data (N=7441)

Epanechnikov Kernel-Smoothed Hazard Function
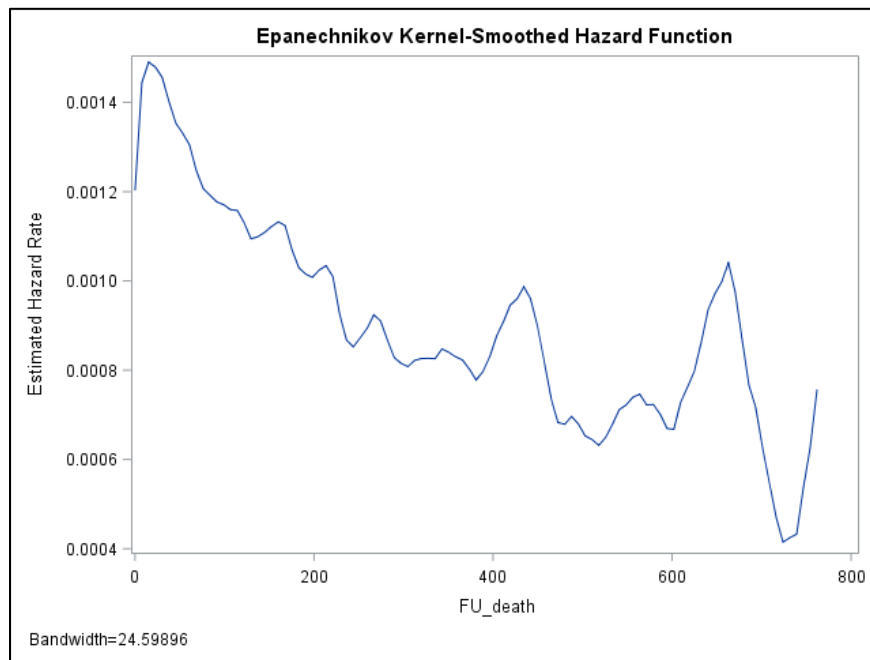
- *Model development*

To develop the cox model in the derivation data, four methods of variable selection were used, including stepwise, forward, backward, and manual selection. A full model that included all the predictors was also presented. The developed models were then applied to the validation data, and performance metrics were generated for comparison between the alternative variable selection models. The C-index and AUC (365) from each method were reported for both derivation and validation datasets (Table 4.4). The number of observations and predictors for each model are shown in Table 4.4.

Table 4. 4. Comparison of alternative variable selection methods in the derivation data (N= 3721)

Model selection	Derivation		Validation		N analyzed validation	Variables
	C-index *	AUC at 365 days[†]	C-index	AUC at 365 days		
Full model (all)	0.7168 (0.70-0.74)	0.7480	0.7035 (0.69-0.72)	0.7475 (0.58-0.92)	2073	41 variables
Stepwise	0.7004 (0.68-0.72)	0.7347	0.6961 (0.68-0.71)	0.7404 (0.71-0.77)	2312	9 variables: age, sex, race, SQ, albumin, cholesterol, KPS, ADL-decline, hyperlipidemia
Forward	0.7004 (0.68-0.72)	0.7347	0.6961 (0.68-0.71)	0.7404 (0.71-0.77)	2312	9 variables: age, sex, race, SQ, albumin, cholesterol, KPS, ADL-decline, hyperlipidemia
Backward	0.7107 (0.69-0.73)	0.7389	0.7059 (0.69-0.72)	0.7504 (0.68-0.82)	2312	32 variables: age, race, SQ, albumin, cholesterol, KPS, ADL-decline, hypothyroidism, anemia, asthma, AF, BPH, breast ca, cataract, CKD, Colorectal ca, COPD, depression, DM, endometrial ca, glaucoma, HF, hip fx, hyperlipidemia, hypertension, IHD, lung ca, osteoporosis, prostate ca, RA/OA, stroke/TIA, diagnosis-count
Manual selection	0.7163 (0.70-0.73)	0.7570	0.6924 (0.68-0.71)	0.7346 (0.71-0.76)	2312	8 variables: age, race, SQ, albumin, cholesterol, KPS, ADL-decline, hyperlipidemia

S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; AF: atrial fibrillation; BPH: benign prostatic hyperplasia; ca: cancer; fx: fracture; CKD: chronic kidney diseases; COPD: chronic obstructive pulmonary diseases; DM: diabetes mellitus; HF: heart failure; IHD: ischemic heart diseases; RA/OA: rheumatoid arthritis/osteoarthritis; TIA: transient ischemic attack;
*Confidence intervals for the C-index was calculated by using the standard error for Harrell's estimate of the concordance;
[†]Confidence intervals for the AUC (365) in the derivation cohort are not provided because using variable selection methods and multiple iterations of the model cause a very wide CL for the AUC;

The performance of the models developed with different variable selection methods was similar, so the model that was selected through stepwise variable selection was chosen as the final best model because it has slightly better AUC (365) and C-index based on the validation data, while also being parsimonious with only nine variables. The backward selection model resulted in a tiny increase in AUC (365) compared to the stepwise model; however, the confidence interval for this statistic is not attained, and

the number of variables is much higher than the stepwise model (32 vs. 9). Manual variable selection

resulted in a model with eight variables, but performance measures were slightly lower than stepwise

selection. Unlike the results of LR variable selection methods, the variables selected in the Cox model

with different selection method are also very consistent. Indeed, selected variables are precisely the

same in stepwise, forward, and manual selection methods, except for the variable sex that was not

selected in the manual selection model. The eight variables included in all the models (including

backward selection), were demographics (age, race), SQ, nutritional status indicators (albumin and

cholesterol), history of hyperlipidemia and functional status indicators (KPS, ADL-decline). These results

are consistent with the important variables selected from the other approaches developed in chapters

two and three (Table 4.9, comparison of the three approaches).

Final Selected model- The model developed with stepwise variable selection was the best model. The C-

index and AUC (365) of the model in the derivation data were 0.7004 and 0.7140, respectively.

To evaluate the importance of variables in the Cox multivariable model, the parameter estimates,

hazard ratios, and 95% confidence limits from the stepwise model are presented in Table 4.5. Albumin,

age and surprise questions have the largest hazard ratios. Similar to what was observed in the LR model,

the lowest levels of albumin and cholesterol resulted in the highest hazard ratios for mortality and

hospice. As expected, male sex and older ages are also associated with increased HRs, although age does

not show a dose-response relationship. It means that the HR for age 95+ years was not higher than the

HR for age 85-95. The direction and magnitude of the hazard ratios in the Cox model is similar to the

odds ratios from the LR model. ADL (improve vs. no change) also has a relatively large HR; however, the

prevalence of this value (ADL=improve) is very low (4%) and so of little clinical importance in this

population. For the ADL variable, "no change" had a higher hazard compared to the "decline", although

it is not statistically significant. The relationship between ADL and mortality is also similar between the

Cox and LR models.

Table 4. 5. Parameter estimates, hazard ratios, and 95% CL for predictors of the MV Cox model for mortality outcome- derivation data (N=2289)

| Variable	Parameter Estimate	P-value	Hazard Ratio	95% HR Confidence Limits	
Age, 75-84 years vs. 65-74 years	0.37319	0.0070	1.452	1.107	1.905
Age, 85-94 years vs. 65-74 years	0.61698	<.0001	1.853	1.438	2.388
Age, 95+ years vs. 65-74 years	0.40875	0.0311	1.505	1.038	2.182
Age, 65-74 years	Ref				
Sex, Male vs. Female	0.23499	0.0137	1.265	1.049	1.525
Race, Black vs. White	-0.35441	0.0042	0.702	0.550	0.894
Race, Other vs. White	-0.37761	0.1660	0.685	0.402	1.170
Race, White	Ref				
Surprise question, No vs. Yes	0.51093	<.0001	1.667	1.344	2.067
Albumin, <3.2 vs 3.8+ gr/dl	1.14039	<.0001	3.128	2.372	4.124
Albumin, 3.2-<3.5 vs 3.8+ gr/dl	0.66669	<.0001	1.948	1.474	2.573
Albumin, 3.5-<3.8 vs 3.8+ gr/dl	0.43304	0.0032	1.542	1.156	2.056
Albumin, 3.8+ gr/dl	Ref				
Cholesterol, <136 vs 195+ gr/dl	0.44608	0.0005	1.562	1.218	2.004
Cholesterol, 136-<164 vs 195+ gr/dl	-0.07845	0.5551	0.925	0.712	1.200
Cholesterol, 164-<195 vs 195+ gr/dl	0.16740	0.1857	1.182	0.923	1.515
Cholesterol, 195+ gr/dl	Ref				
KPS, Severe vs. Moderate disability	0.40052	<.0001	1.493	1.239	1.799
ADL-decline, Decline vs. No-change	-0.05349	0.6268	0.948	0.764	1.176
ADL-decline, Improve vs. No-change	-0.82404	0.0050	0.439	0.247	0.780
ADL-decline, No-change	Ref				
CCW-Hyperlipidemia, No vs. Yes	0.39815	<.0001	1.489	1.251	1.772
*KPS values 0-40 indicate severe disability, while values 50-100 shows moderate/mild and no disability;					

- *Model performance*

To validate the model, it was applied to the validation dataset, and the three performance metrics were

generated. Table 4.6 shows the C-index of the model in the validation data, and Figure 4.4 presents the

AUC at 365 days.

Table 4. 6. Concordance (C-index) of the Cox MV model for mortality in the validation data (N=2312)

Harrell's Concordance Statistic						
Source	Estimate	Standard Error	Comparable Pairs			
			Concordance	Discordance	Tied in Predictor	Tied in Time
Model	0.6961	0.0086	1053996	459642	1441	554

Figure 4. 4. ROC for the mortality outcome at time=365 days and AUC (365) from Cox MV model-validation data (N=2312)
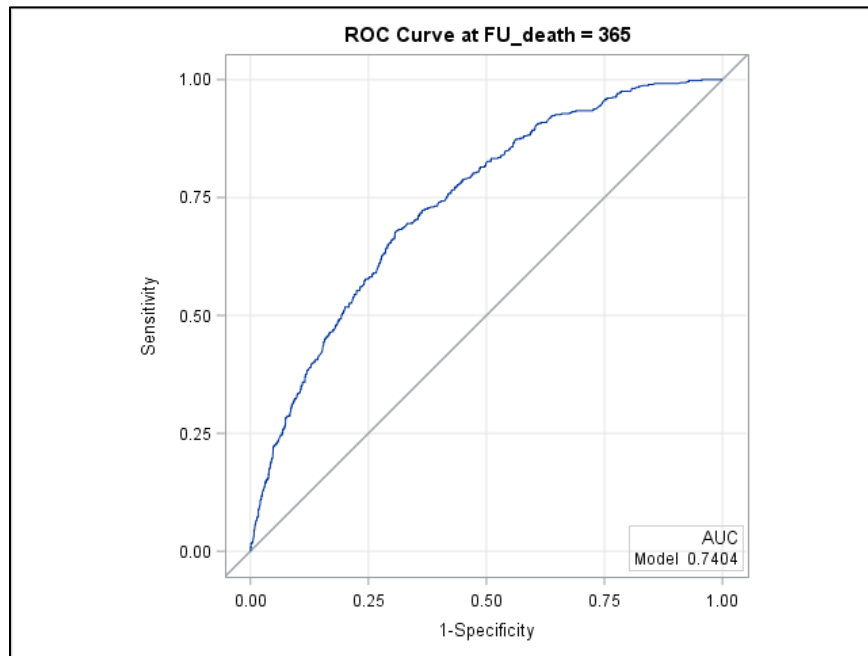
Time-dependent AUC was generated for the Cox MV model in the validation data and resulted in an

iAUC of 0.7318. This drop in the AUC at the end of the study period is related to the censored subjects at

the end of the study for which no event has been reported.

Figure 4. 5. Time dependent AUC (stepwise selection, validation data) (N=2312)
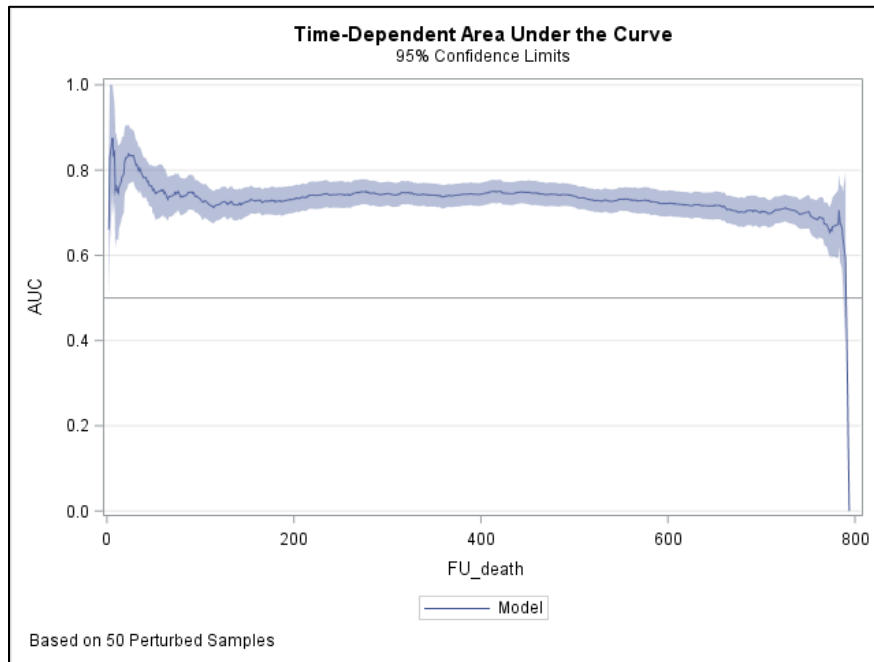
Integrated Time-Dependent AUC		
Source	Estimate	Tau
Model	0.7318	750

- *Proportionality assumption*

To test the proportionality assumption in this data, KM survival plots were generated in the derivation cohort and stratified by all nine predictors. Also, the interaction terms for these variables by the time were included in the PHREG procedure to evaluate the effect of each level of predictors over time. Figures 4.6 – 4.14 illustrate the KM survival curves stratified by the five key predictors. None of the KM survival plots show curves that cross each other over time. That is to say there is no graphical evidence of non-proportionality.

Figure 4. 6. KM survival curve stratified by age- derivation data
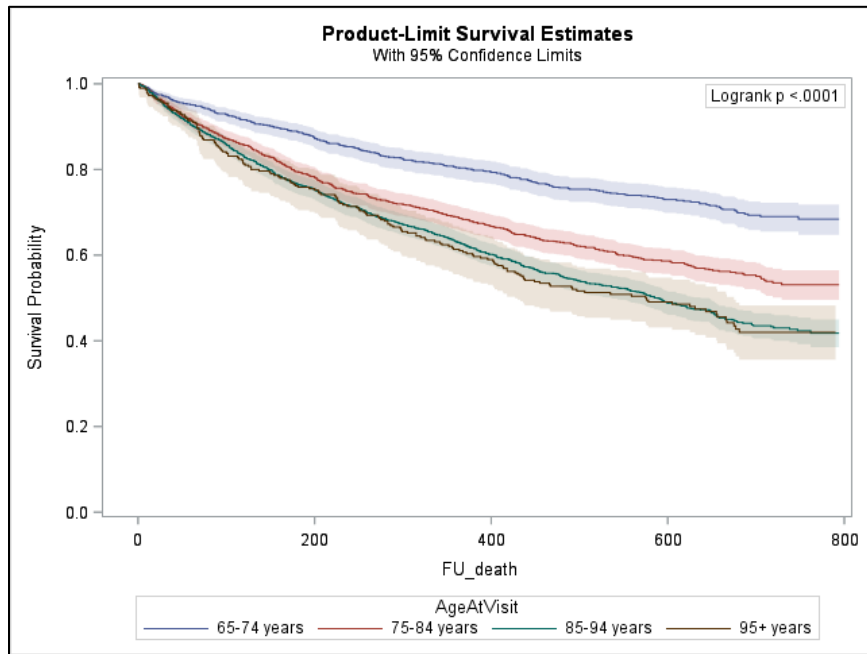

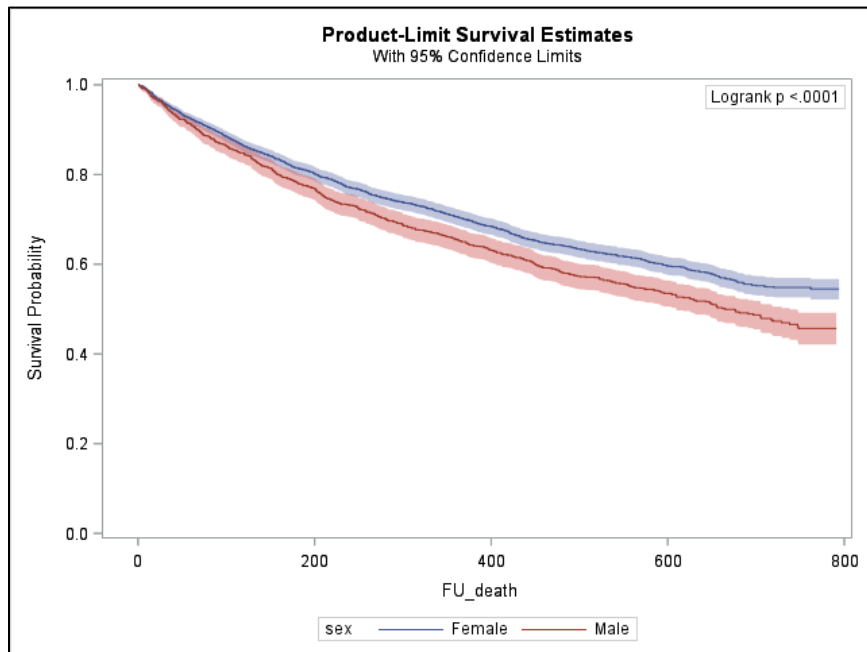Figure 4. 7. KM survival curve stratified by sex- derivation data

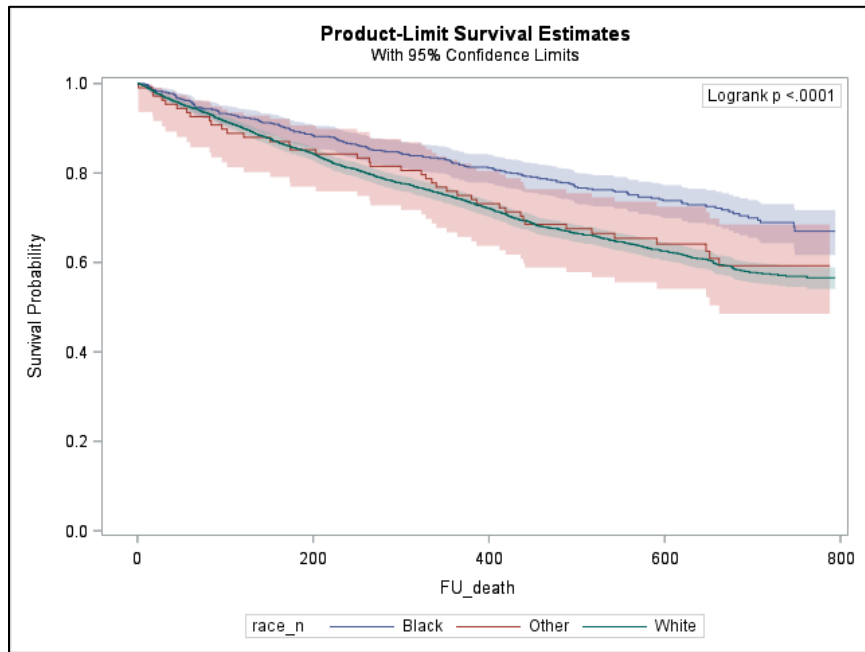Figure 4. 8. KM survival curve stratified by race- derivation data

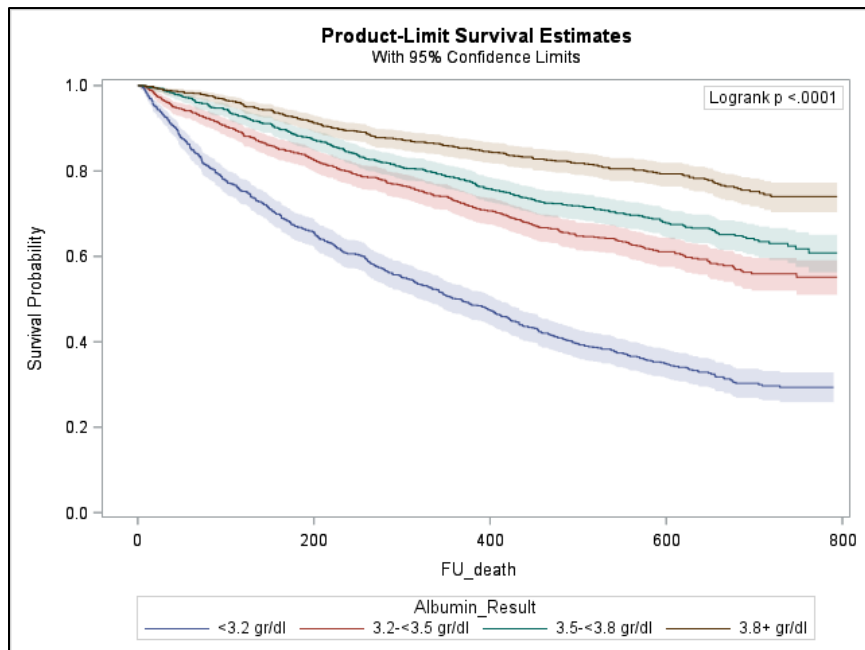Figure 4. 9. KM survival curve stratified by albumin- derivation data

Figure 4. 10. KM survival curve stratified by cholesterol- derivation data
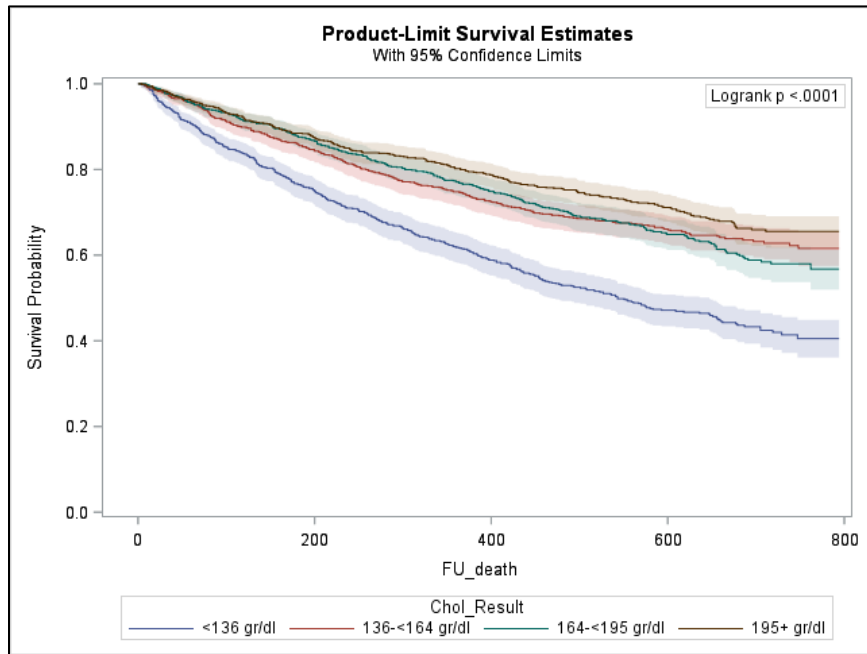

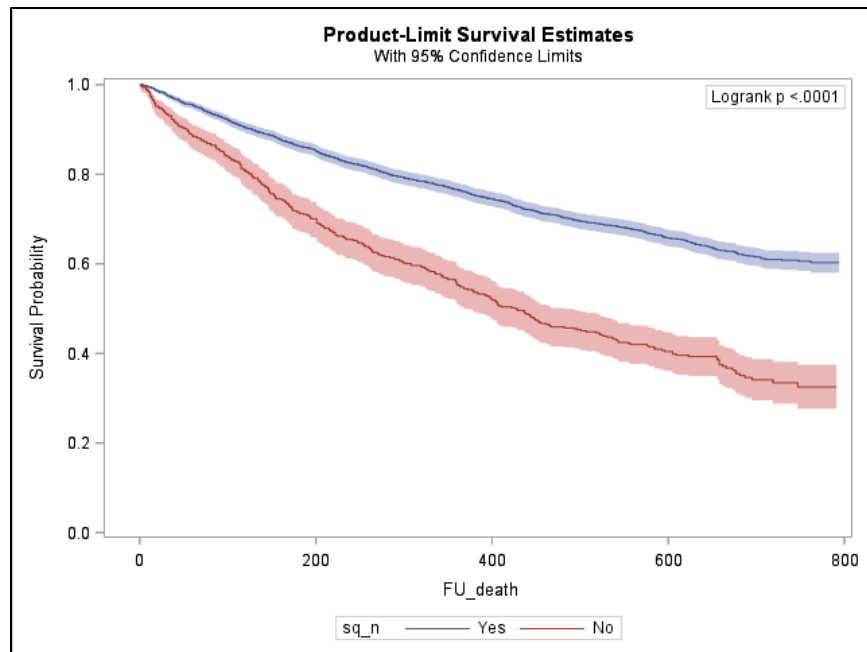Figure 4. 11. KM survival curve stratified by SQ- derivation data

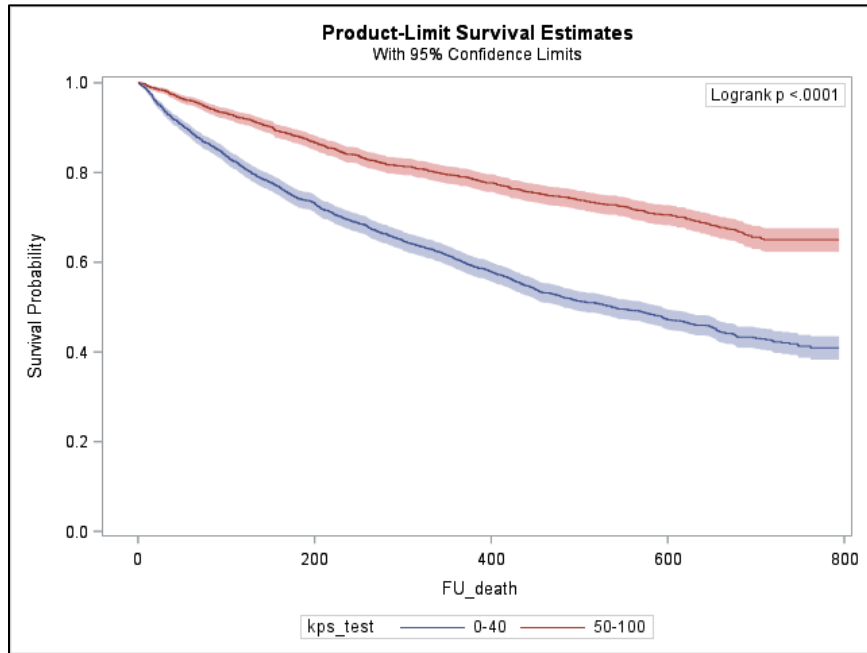Figure 4. 12. KM survival curve stratified by KPS- derivation data

Figure 4. 13. KM survival curve stratified by ADL decline- derivation data
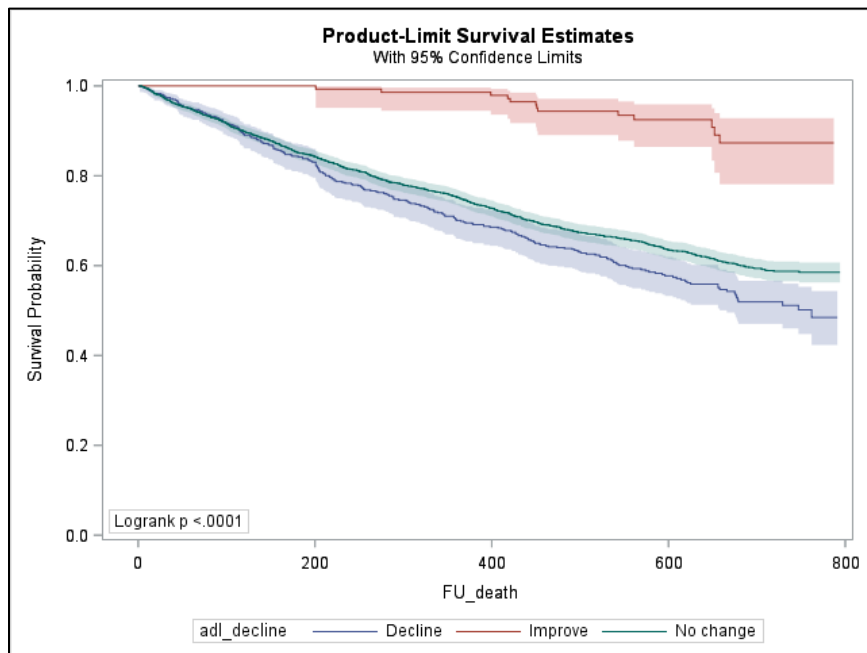
Figure 4. 14. KM survival curve stratified by hyperlipidemia- derivation data
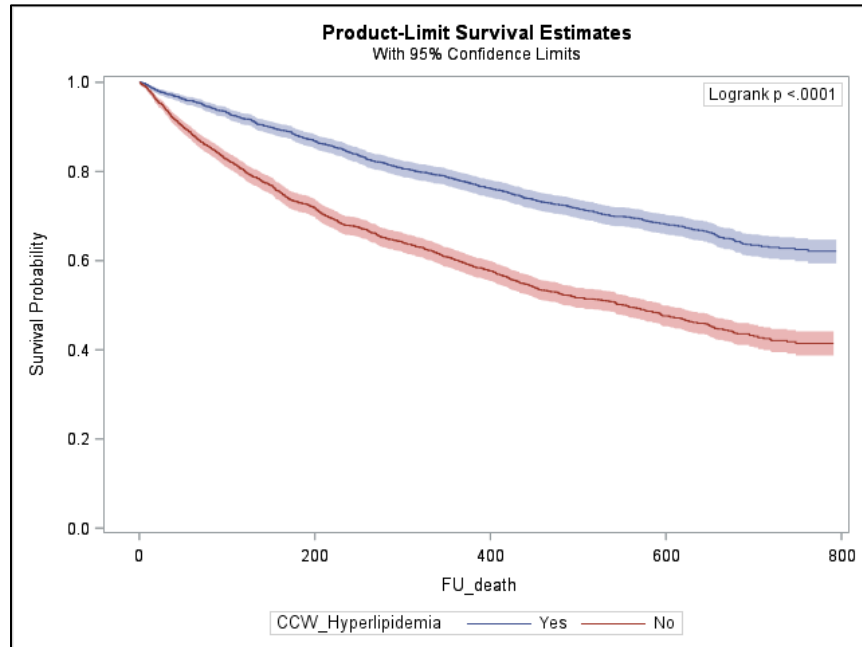
Additionally, the 2-way interactions between time and predictors in the final model were also tested by adding them into the final main effects Cox model. The significance of the coefficient of interaction terms indicates the violation of the proportionality assumption for that predictor. Table 4.7 contains the estimates and p-values for the nine interaction terms. The main effects in the model, are not shown in this table. Three of the interaction terms (Cholesterol, ADL, and hyperlipidemia) were statistically significant at the P<0.05 level; however, stratified KM survival curves did not show any evidence of the significant violation of proportionality assumption.

Table 4. 7. Parameter estimates and p-values for the interaction terms between time and key predictors-derivation data

Parameter	DF	Parameter Estimate	P-value for interaction	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Age*Time	1	0.02164	0.6850	1.022	0.920	1.135
Sex*Time	1	0.13337	0.1647	1.143	0.947	1.379
Race*Time	1	0.13021	0.2119	1.139	0.928	1.397
Albumin*Time	1	0.06108	0.1745	1.063	0.973	1.161
Cholesterol*Time	1	0.08470	**0.0419**	1.088	1.003	1.181
SQ*Time	1	0.00529	0.9592	1.005	0.821	1.231
KPS*Time	1	0.04411	0.6497	1.045	0.864	1.264
ADL-decline*Time	1	0.29245	**0.0091**	1.340	1.075	1.669
Hyperlipidemia*Time	1	0.20372	**0.0311**	1.226	1.019	1.475

An overall test for proportionality was performed using the statement TEST in PROC PHREG when all the interaction terms included in the model and the test statement. The result of this test is consistent with the results in Table 4.7 and rejected the null hypothesis that overall none of the interaction coefficients are statistically different from zero (Table 4.8). Considering the KM survival curves (Figures 4.6-4.14) which do not show a significant violation of the PH assumption, the Cox model can be appropriately used to model the mortality in this cohort.

Table 4. 8. Overall test for proportionality assumption for all interaction terms together

Linear Hypotheses Testing Results			
Label	Wald Chi-Square	DF	P-value
PH- test	20.8861	9	0.0132

- *Comparison between the alternative approaches (Cox, LR, and RF)*

To compare the performance of this model to the previous approaches, the AUC of the best model in

each of the LR and RF approaches were compared to the Cox model results (Table 4.9). Because the

outcome of interest for LR and RF models was 1-year mortality, the AUC at day 365 for Cox MV model

was reported for comparison. Also, C-index as an overall measure of discrimination over the study time

was displayed in Table 4.9.

Table 4. 9. Comparison of the model performance between the three models, Cox, LR, and RF models
using validation dataset

Model	N analyzed, Validation	AUC at 1-year Validation	Variables
Cox Model	2312*	0.7404 (0.71- 0.77) 0.6961[†] (0.68- 0.71)	9 variables: age, sex, race, SQ, albumin, cholesterol, KPS, ADL-decline, hyperlipidemia,
Logistic regression	2312*	0.7634 (0.74- 0.79)	11 variables: age, sex, race, dual-eligible, SQ, albumin, cholesterol, KPS, ADL-decline, hyperlipidemia, depression
Random forest	3723	0.8292 (0.82- 0.84)	15 first ranked important variables: TUG, albumin, race, ADL-decline, cholesterol, KPS, SQ, age, hyperlipidemia, diagnosis-count, tobacco, living-alone, RA/OA, dual-eligible, pressure-ulcer
S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; RA/OA: rheumatoid arthritis/ osteoarthritis; *Observations with partly missing data were excluded by default in LR and Cox procedures; [†]C-index is the concordance obtained by applying the developed Cox model to the validation data;			

Compared to the LR and RF models, Cox MV model had the lowest discrimination ability in this data.

Therefore, the analysis of time-to-event instead of the fixed time event analysis (1-year mortality) does

not seem to improve the accuracy of the predictions. Similar to the variable selection in the LR model,

TUG was not selected in the Cox model; however, it is high-ranked in the variable importance in the RF

model. Missing data on TUG resulted in the exclusion of 20% of observation from the LR and Cox

analysis.

o   Outcome: one-year hospice admission

The second outcome, time-to-hospice, was analyzed following the same methods used for the mortality

outcome. Figure 4.15 is the KM survival curve for time to hospice in the total cohort (n=7441). During

the study time, 1389 (19%) events (hospice admissions) occurred and 6052 (81%) of patients were

censored, 30% were censored due to death and the rest of 51% were censored at the administrative end

date of the study.

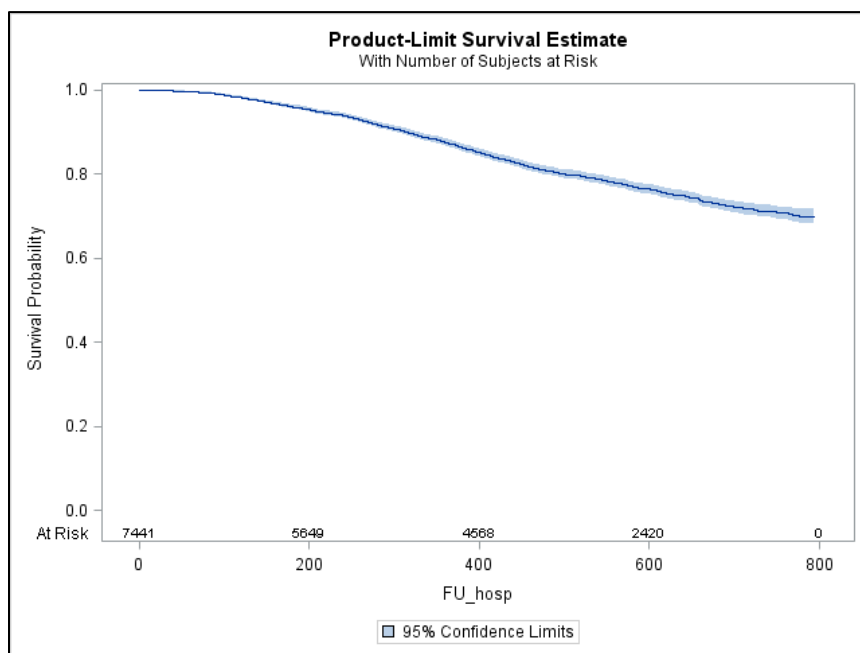Figure 4. 15. KM plot for time-to-hospice admission in the whole cohort (N=7441)

Figure 4.16 illustrates the estimated hazard rates for hospice admission over follow-up time. Unlike the

hazard rate for mortality, the hazard rate for hospice admission is low at the beginning but increases

until around a year from the first USMM visit, then decreases over time.

Figure 4. 16. Hazard rate for hospice admission from the first USMM visit- whole cohort (N=7441)
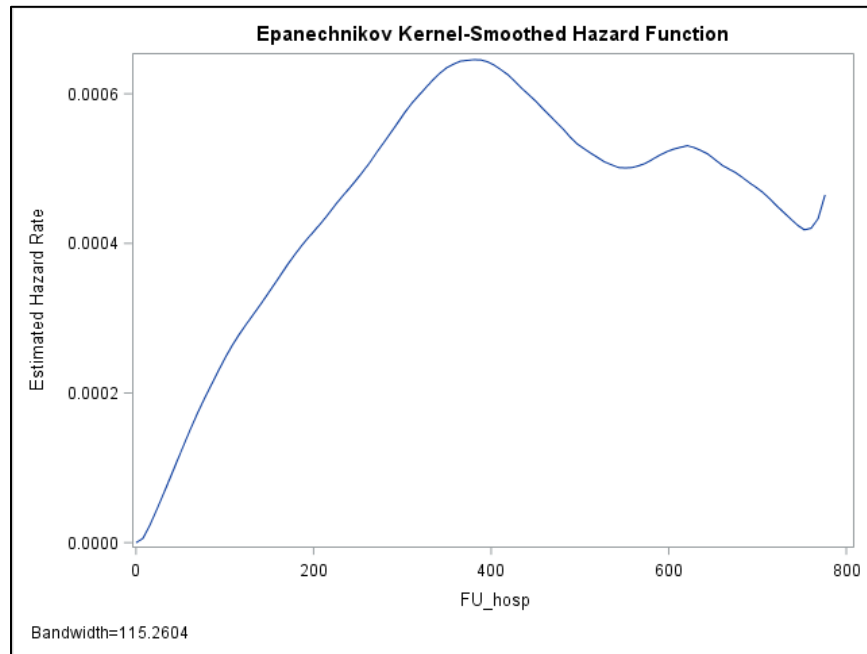

Figure 4.17 shows the KM survival plots for mortality stratified by the hospice status. The red color

showed the hospice admitted group, and it shows very few events at the beginning of the curve. It

seems that patients who were admitted to hospice had their first few months (about 180 days) free of

death, which is not correct. In fact, 59 % of patients who were admitted to hospice died within the first

three months of their admission. Note that the time-to-death in Figure 4.17 represents the number of

days between the first-ever visit and the date of death. Therefore the slow slope at the left end of the

hospice curve does not show the number of deaths at the beginning of hospice stay. It shows the

patients who finally admitted to hospice had a lower number of deaths within the first six months of

their joining USMM services.

Likewise, Figure 4.18 displays the hazard rates for mortality in the total cohort stratified by hospice

admission. Interestingly the hazard rate is increasing in the hospice admitted group, whereas it is slowly

decreasing in those without hospice admission.

Figure 4. 17. KM plot for time-to-death from the first USMM visit stratified by hospice admission status (N=7441)
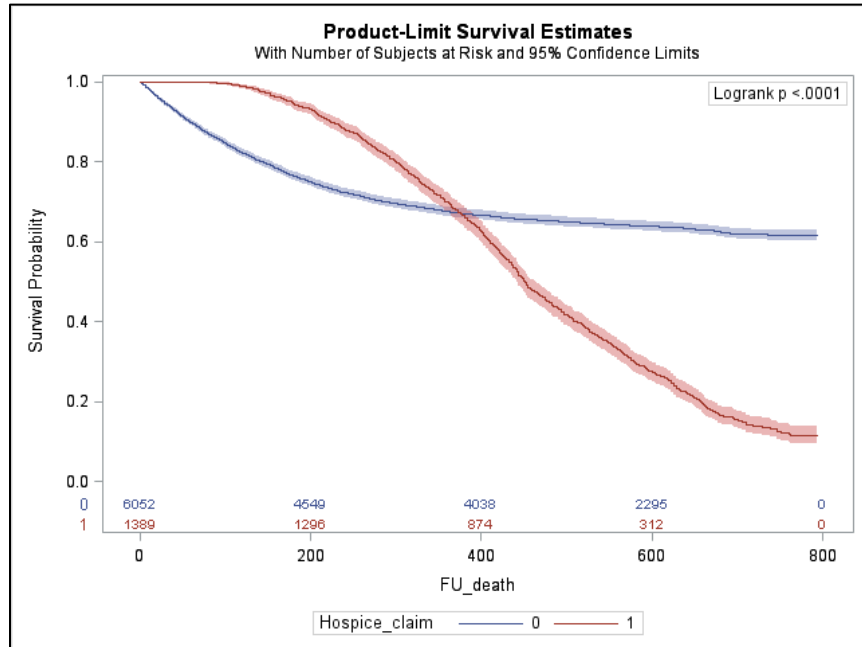

Figure 4. 18. Estimated hazard rates for time-to-death stratified by hospice admission status (N=7441)
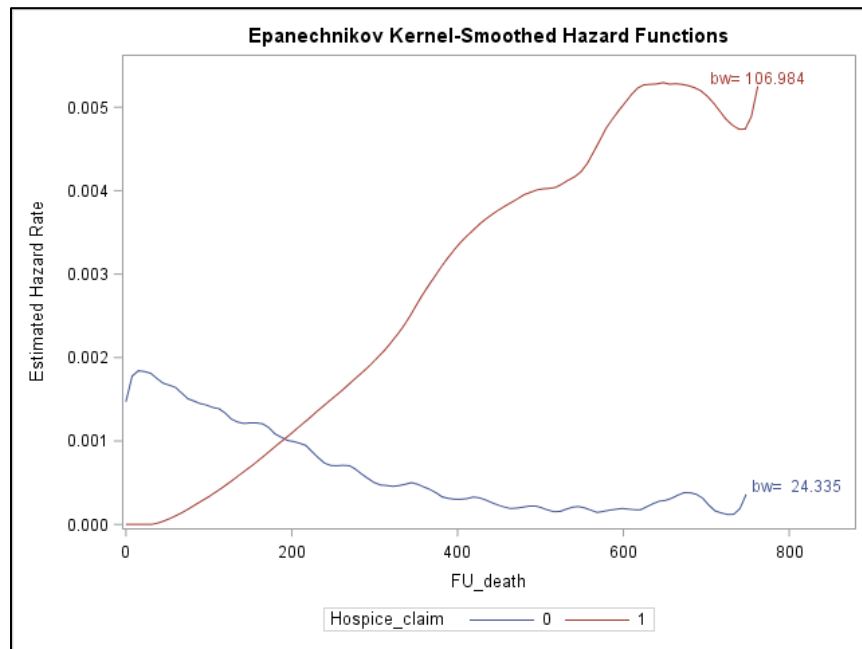
Figure 4.19 illustrates the KM survival curve for the time from hospice admission to death among the 1389 hospice admitted patients. In this group, 1122 (81%) deaths occurred during the follow-up time, and 267 (19%) subjects were censored at the end of the study.

Figure 4. 19. KM survival among hospice admitted patients (N=1389)
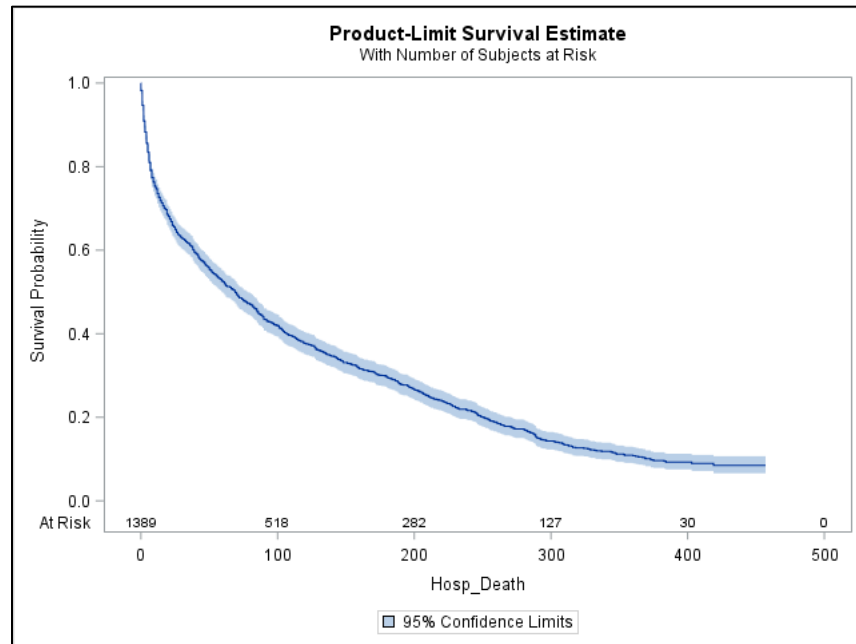
As mentioned above, 59% of patients who were admitted to hospice died within the first three months of their admission, and 76% died within the first six months. The steep slope of the KM curve in Figure 4.19 confirms the high rate of death within the first 100 days in the hospice admitted population. Figure 4.20 displays the estimated hazard rate for mortality among the hospice admitted patients over time; it is another illustration of the high rate of death at the beginning of hospice admission in this cohort. The life expectancy <6months in the 76% of hospice admitted group means that the screening for hospice based on the CMS eligibility criteria was done with a reasonable estimation of patients' prognosis.

Figure 4. 20. Estimated hazard rate for mortality among hospice admitted patients (N=1389)
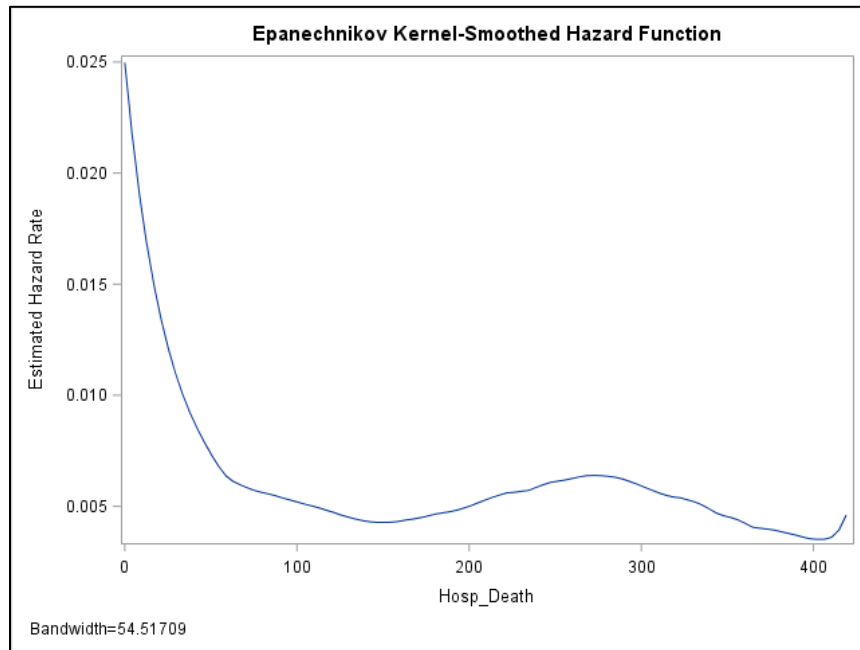
- *Model development*

Similar to the methods for mortality outcome, four variable selection methods were applied to develop

the models in the PHREG procedure. Table 4.10 compares the performance results of these methods.

Table 4. 10. Alternative variable selection methods for hospice outcome- derivation data (N=3721)

Model selection	Derivation		Validation		N analyzed validation	Variables
	C-index*	AUC at 365 days[†]	C-index	AUC at 365 days		
Full model (all)	0.7075 (0.69-0.73)	0.7502	0.6837 (0.67-0.70)	0.7207 (0.49-0.95)	2073	41 variables
Stepwise	0.6947 (0.67-0.71)	0.7396	0.6750 (0.66-0.69)	0.7199 (0.68-0.76)	2498	9 variables: age, race, SQ, living-alone, albumin, KPS, Hip fx, hyperlipidemia, number of labs,
Forward	0.6947 (0.67-0.71)	0.7396	0.6750 (0.66-0.69)	0.7199 (0.68-0.76)	2498	9 variables: age, race, SQ, living-alone, albumin, KPS, Hip fx, hyperlipidemia, number of labs,

Table 4. 10. (cont'd)					
Backward	0.7006 (0.68-0.72)	0.7426	0.6730 (0.66-0.69)	0.7152 (--)	2498	29 variables: age, race, dual-eligible, SQ, living-alone, albumin, KPS, cancer, hypothyroidism, anemia, asthma, AF, BPH, cataract, CKD, COPD, depression, DM, glaucoma, HF, hip fx, hyperlipidemia, hypertension, IHD, osteoporosis, RA/OA, stroke/TIA, number of labs, diagnosis-count
Manual selection	0.6866 (0.67-0.71)	0. 7330	0.6827 (0.67-0.70)	0.7212 (0.66-0.78)	2227	10 variables: age, race, dual-eligible, SQ, living-alone, albumin, KPS, TUG, hyperlipidemia, number of labs,

S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; AF: atrial fibrillation; BPH: benign prostatic hyperplasia; ca: cancer; fx: fracture; CKD: chronic kidney diseases; COPD: chronic obstructive pulmonary diseases; DM: diabetes mellitus; HF: heart failure; IHD: ischemic heart diseases; RA/OA: rheumatoid arthritis/osteoarthritis; TIA: transient ischemic attack;
*Confidence intervals for the C-index was calculated by using the standard error for Harrell's estimate of the concordance;
†Confidence intervals for the AUC (365) in the derivation cohort are not provided because using variable selection methods and multiple iterations of the model cause a very wide CL for the AUC;

The performance measures of the models were very similar when using different variable selection methods. However, the stepwise selection method is the most parsimonious model with nine variables. The manually selected model had one additional variable (TUG) that did not make a significant change in the AUC and C-index compared to the stepwise model. However, 271 more observations were excluded due to missing on TUG. Therefore the best Cox MV model for the outcome hospice admission is the one selected through stepwise variable selection method.

The variables that were consistently selected in all four selection methods are age, race, SQ, living alone, KPS, albumin, hyperlipidemia, and the number of lab tests.

Final selected model- The stepwise selected model with nine variables had the AUC (365) of 0.7291 and C-index of 0.6947 in the derivation data. The parameter estimates, p-value, and hazard ratios for predictors in this model are shown in Table 4.11.

Table 4. 11. Parameter estimates and hazard ratios from the Cox model for hospice outcome, derivation data (N=2055)

Variable	Parameter Estimate	P-value	Hazard Ratio	95% HR Confidence Limits	
Age, 75-84 years vs. 65-74 years	0.96481	<.0001	2.624	1.703	4.043
Age, 85-94 years vs. 65-74 years	1.33531	<.0001	3.801	2.527	5.719
Age, 95+ years vs. 65-74 years	1.27587	<.0001	3.582	2.167	5.921
Age, 65-74 years	Ref				
Race, Black vs. white	-0.72244	<.0001	0.486	0.345	0.684
Race, Other vs. White	-0.69036	0.0720	0.501	0.236	1.064
Race, White	Ref				
Surprise question, No vs. Yes	0.77141	<.0001	2.163	1.677	2.789
Living-alone, Yes vs. No	-0.63107	0.0032	0.532	0.350	0.809
Albumin, <3.2 vs 3.8+ gr/dl	0.90295	<.0001	2.467	1.759	3.459
Albumin, 3.2-<3.5 vs 3.8+ gr/dl	0.51051	0.0033	1.666	1.185	2.342
Albumin, 3.5-<3.8 vs 3.8+ gr/dl	0.34028	0.0560	1.405	0.991	1.992
Albumin, 3.8+ gr/dl	Ref				
KPS, Severe vs. Moderate disability*	0.83124	<.0001	2.296	1.788	2.949
CCW-Hip/pelvic fracture, No vs. Yes	1.00489	0.0472	2.732	1.013	7.369
CCW-Hyperlipidemia, No vs. Yes	0.33907	0.0023	1.404	1.129	1.745
Number of Labs (continuous)	-0.03736	0.0022	0.963	0.941	0.987
*KPS values 0-40 indicate severe disability, while values 50-100 shows moderate/mild and no disability;					

Based on the parameter estimates, age, hip-fracture, albumin, and KPS had the strongest impact on the time-to-hospice. Hip- fracture has a large coefficient estimate, but the prevalence of it in this population is very low (1%). Therefore the effect of this variable would not be clinically meaningful. Surprisingly, the direction of the association between hip fracture and hospice admission shows that patient with a history of hip fracture had less hospice admission than those without hip fracture. A possible

explanation might be the old patients with hip fracture are more likely to die before hospice referral or before making a decision about hospice admission (e.g., death in the hospital while hospitalized for the index fracture. Mortality rate of patients with a hip fracture is 15-30% in the first year after the fracture and a small proportion of patients are discharged to hospice. (151,152) Additionally, we observed in this cohort that many comorbidities had a reverse association with both outcomes, death, and hospice admission. The possible explanations are discussed in Chapter 5.

Age, albumin, and KPS were also among the most important predictors for mortality in all different variable selection methods.

- *Model performance*

The predictive performance of the Cox MV model for the hospice outcome is evaluated by applying the model to the validation data. C-index and AUC at day 365 from the model in the validation data were shown in Table 4.12 and Figure 4.21, respectively.

Table 4. 12. Concordance of the Cox MV model for hospice outcome- validation data (N=2498)

Harrell's Concordance Statistic						
Source	Estimate	Standard Error	Comparable Pairs			
			Concordance	Discordance	Tied in Predictor	Tied in Time
Model	0.6750	0.0080	1337345	642072	6604	881

Figure 4. 21. ROC at day 365 from the Cox MV model for the hospice outcome- validation data (N=2498)

ROC Curve at FU_hosp = 365

	AUC
Model	0.7199
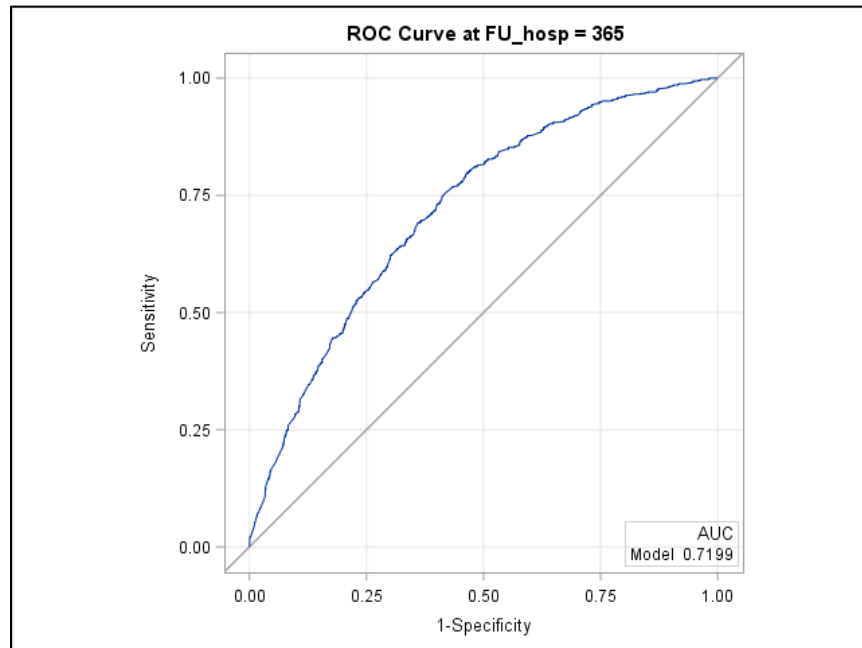
A time-dependent AUC and its summary measure (iAUC) was also generated for the model in the

validation data. Similar to the results of mortality outcome AUC was around 0.70 for most of the follow-

up time, except for the end of follow-up time that shows a steep drop in the AUC.

Figure 4. 22. Integrated AUC from the Cox MV model for hospice outcome- validation data (N=2498)
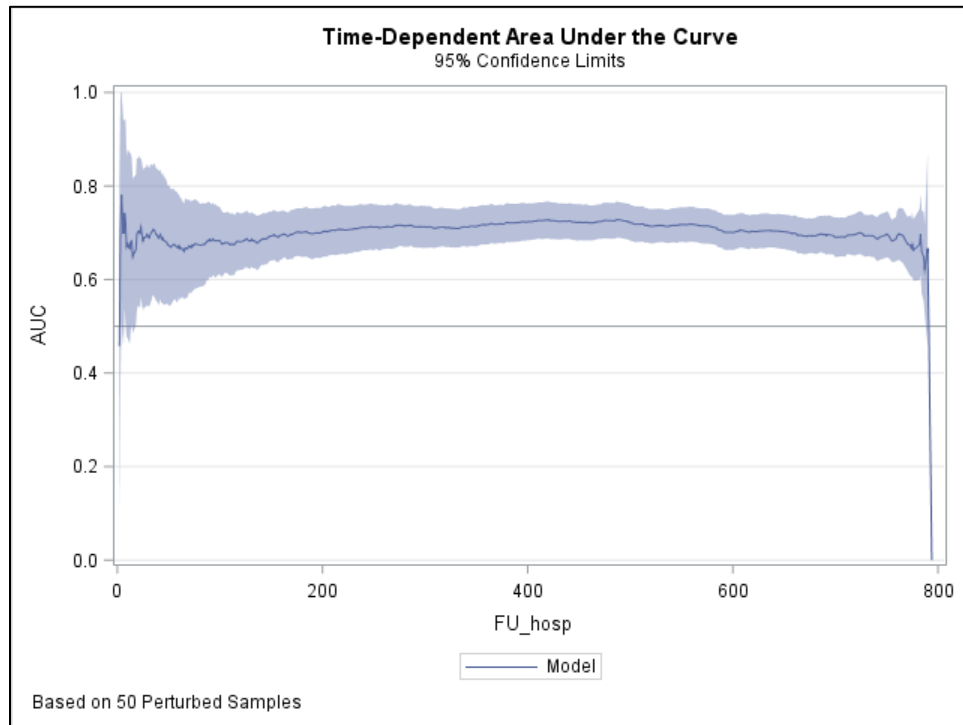
Integrated Time-Dependent AUC		
Source	Estimate	Tau
Model	0.7028	750

- *Proportionality assumption test*

Proportionality assumption was tested for all predictors in the final Cox MV model for the hospice admission, naming age, race, SQ, albumin, and KPS. KM survival curves stratified by the five predictors were generated for the patients in the derivation dataset (Figures 4.23- 4.30). None of the KM curves showed a significant violation of the proportionality assumption. There was not a clear crossing between lines for different levels of predictor, except for albumin, in which the curves for the two middle categories (i.e., 3.2-<3.5 and 3.5-<3.8) were crossing; however, they line up closely, and therefore the crossing does not necessarily mean the violation of the proportionality assumption.

Figure 4. 23. KM survival curve stratified by age- derivation data

Figure 4. 24. KM survival curve stratified by race- derivation data

Figure 4. 25. KM survival curve stratified by SQ- derivation data

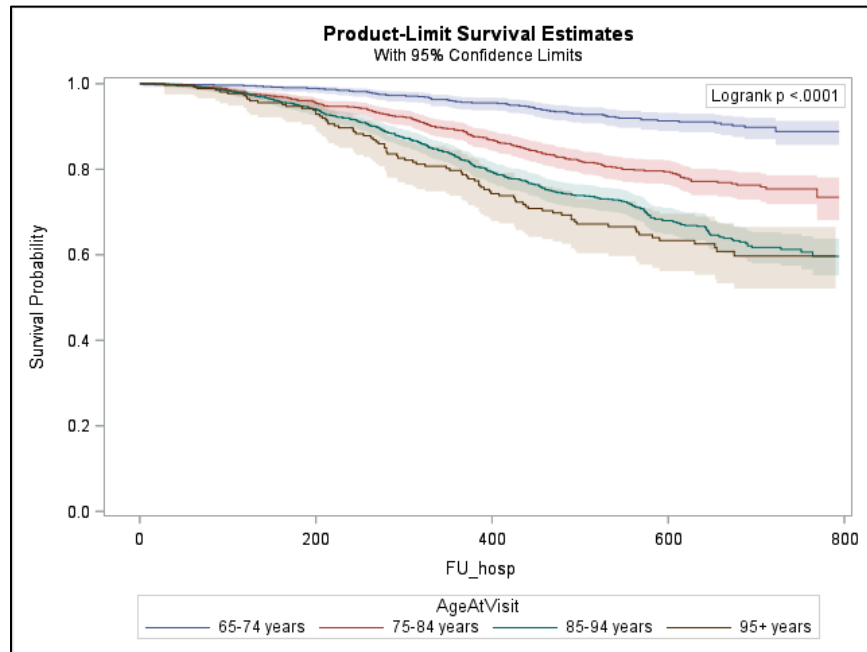Figure 4. 26. KM survival curve stratified by living alone- derivation data

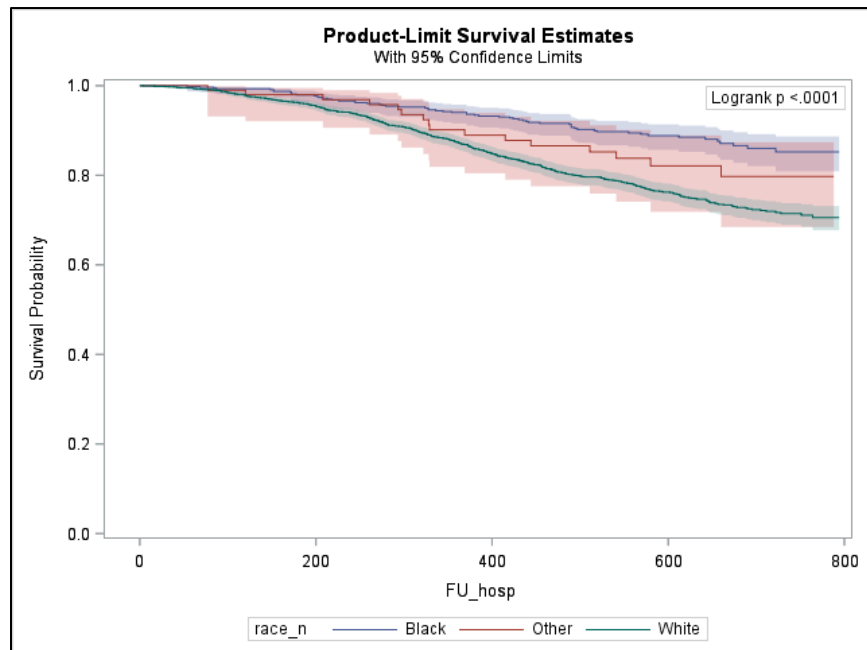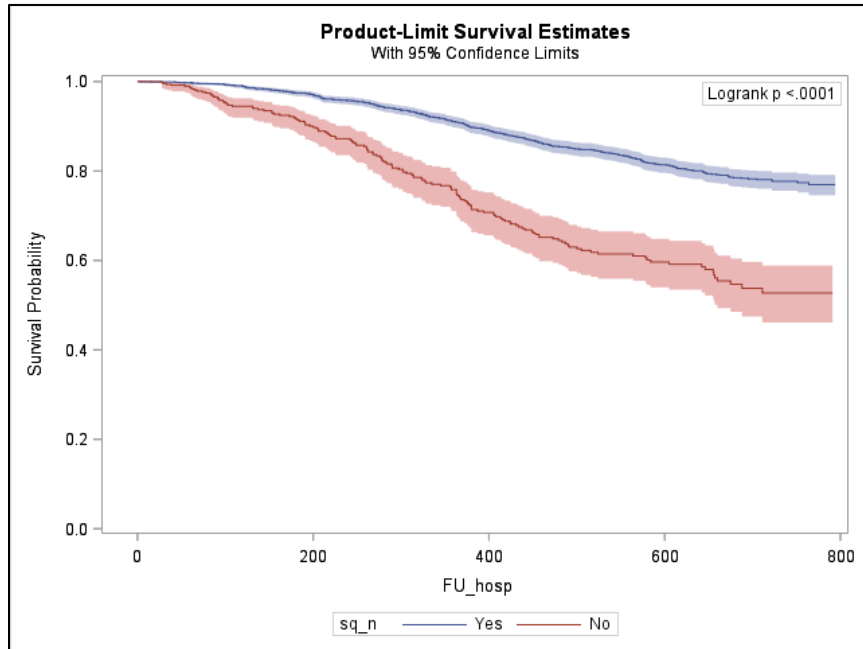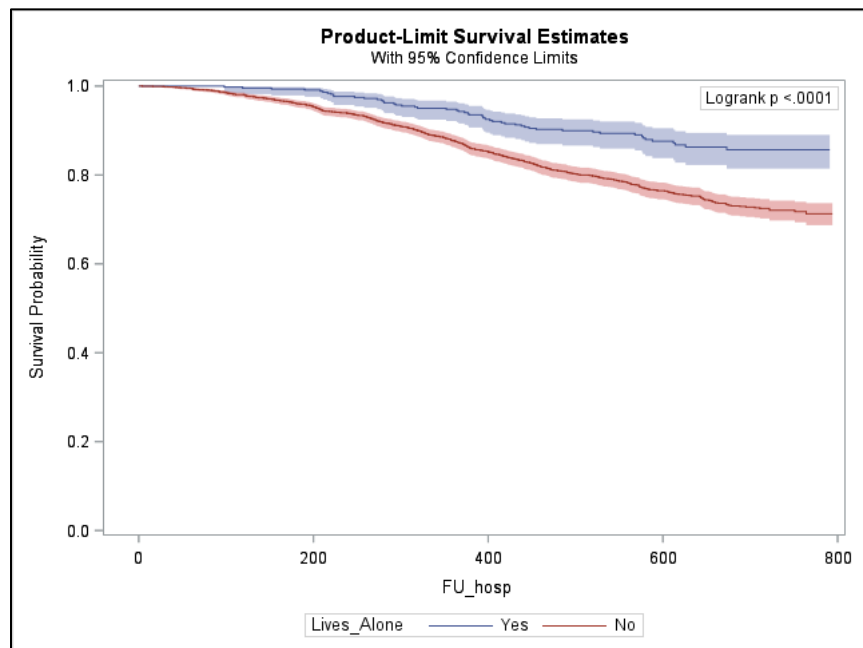Figure 4. 27. KM survival curve stratified by albumin- derivation data
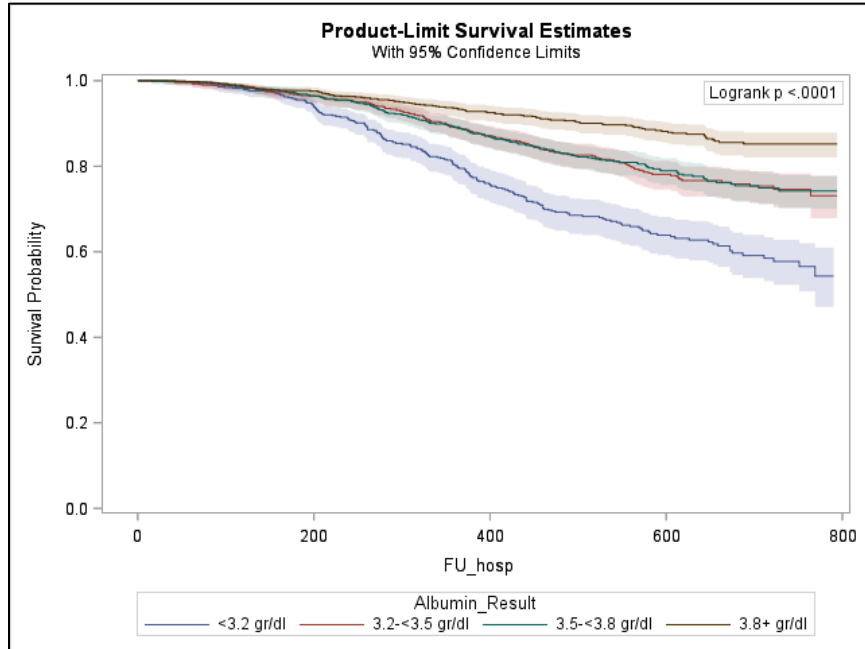
Figure 4. 28. KM survival curve stratified by KPS- derivation data
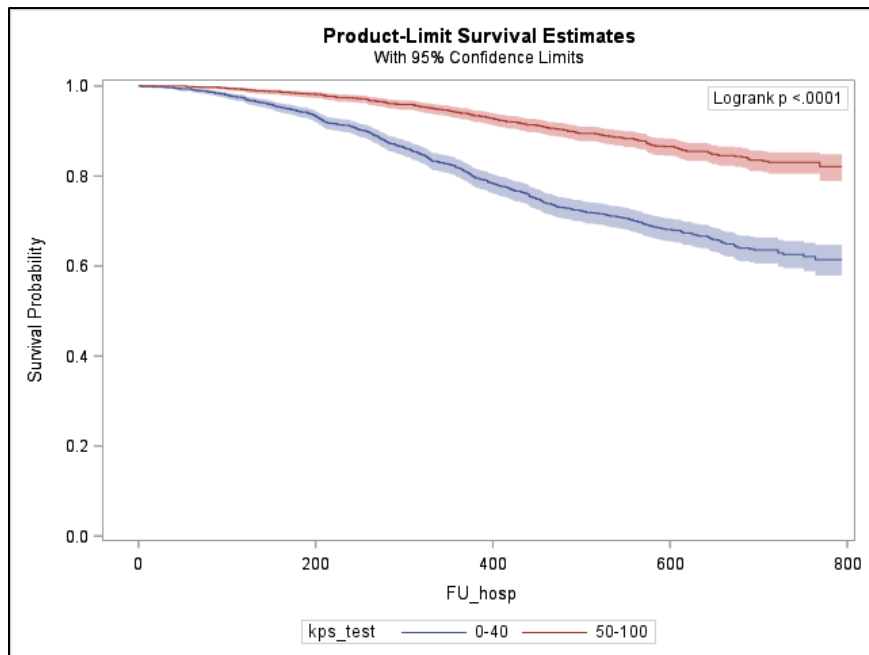
Figure 4. 29. KM survival curve stratified by hip fracture- derivation data
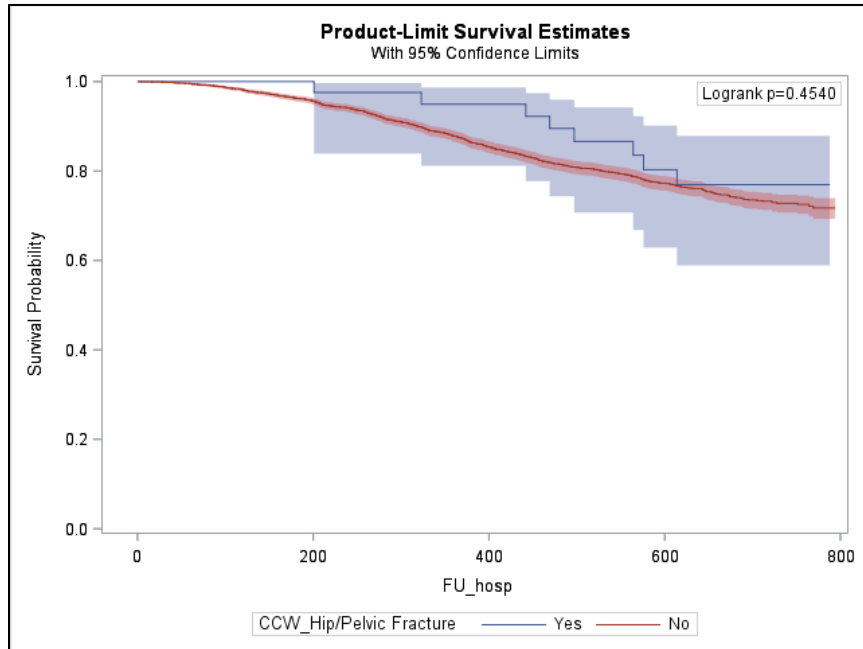
Figure 4. 30. KM survival curve stratified by hyperlipidemia- derivation data
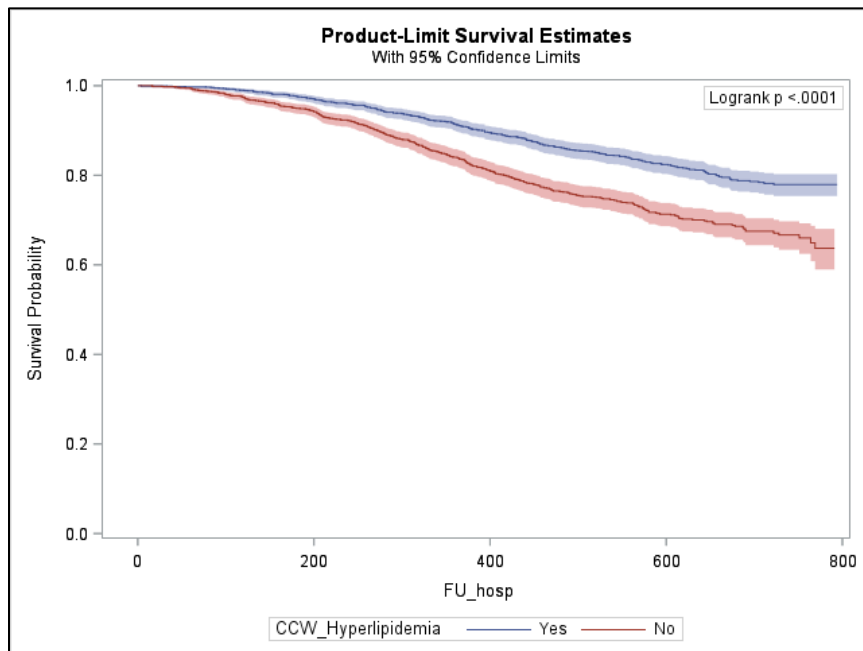
Additionally, the 2-way interactions between time and predictors in the final model were also tested by adding them into the final main effects Cox model. Table 4.13 contains the estimates and p-values for the nine interaction terms. All the interaction terms were non-significant at the significance level of 0.05, which implies that none of the interaction coefficients are statistically different from zero.

Table 4. 13. Parameter estimates and p-values for the interaction terms between time and key predictors- derivation data

Parameter	DF	Parameter Estimate	P-value For interaction	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Age*Time	1	0.07382	0.4732	1.077	0.880	1.317
Race*Time	1	0.06076	0.7513	1.063	0.730	1.547
SQ*Time	1	0.26372	0.1374	1.302	0.919	1.843
Albumin*Time	1	-0.14643	0.0523	0.864	0.745	1.001
KPS*Time		-0.24868	0.2212	0.780	0.524	1.162
Lives-alone*Time		0.05288	0.8620	1.054	0.581	1.914
Hip-fracture*Time		0.64143	0.5364	1.899	0.249	14.509
Hyperlipidemia*Time		0.28888	0.0917	1.335	0.954	1.867

An overall test for proportionality was performed using the statement TEST in PROC PHREG when all the interaction terms included in the model and in the test statement (Table 4.14). The overall PH test result was also statistically non-significant, which means there is no statistical evidence to reject the hypothesis that all of the interaction coefficients are different from zero. In other words, there is no evidence of a violation of the proportionality assumption in this data. Therefore the Cox model can be appropriately applied in this data to model the outcomes of interest.

Table 4. 14. Overall test for proportionality assumption for all interaction terms together

Linear Hypotheses Testing Results			
Label	Wald Chi-Square	DF	P-value
PH-test	10.3797	8	0.2394

- *Comparison between the alternative approaches (LR, RF, and Cox)*

To compare the performance of this model to the previous approaches (LR and RF models), the AUC of

the best model in each of the LR and RF approaches were compared to the Cox model results (Table

4.15). The performance measures of the Cox model for prediction of hospice admission was comparable

to the previous approaches, LR and RF. Interestingly, the LR model with only seven predictors has the

best discrimination among the three approaches for hospice outcome. Four of the selected predictors

were in common between LR and Cox model, naming age, race, SQ, and KPS. Age, SQ and KPS were also

the three highest-ranked important variables in RF model and race was the seventh important one.

Table 4. 15. Comparison of the Cox model performance with the LR and RF models- Hospice outcome

Model	N analyzed, Validation	AUC at 1-year Validation	Variables
Cox Model	2498	0.7199 (0.68- 0.76) 0.6750*	9 variables: age, race, SQ, living-alone, albumin, KPS, hip-fx, hyperlipidemia, number of labs,
Logistic regression	2590	0.7251 (0.70- 0.75)	7 variables: age, sex, race, dual-eligible, SQ, KPS, ADL-decline
Random forest	3723	0.6971 (0.67- 0.72)	15 first ranked important variables: SQ, age, KPS, number of labs, albumin, living-alone, race, dual-eligible, TUG, cholesterol, ADL-decline, hyperlipidemia, stroke/TIA, depression, IHD
S.Q: surprise question; KPS: Karnofsky performance scale; TUG: timed up and go; ADL: activities of daily living; RA/OA: rheumatoid arthritis/ osteoarthritis; TIA: transient ischemic attack; IHD: ischemic heart diseases; hip-fx: Hip fracture; *C-index is the concordance measure for the cox model over the study time;			

- *Discussion*

Overall the Cox PH models that were developed in this data for the two outcomes had a good performance in terms of prediction accuracy (AUC at 365=0.74 for mortality and 0.72 for hospice outcome) when using the rule of thumb for interpretation for the AUC. AUC values can be roughly interpreted as excellent (AUC above 0.80), good (between 0.70 and 0.80), and weak (between 0.50 and 0.70). However, comparing the Cox model to the LR (AUC of 0.76 for mortality and 0.73 for hospice outcome), and RF(AUC of 0.83 for mortality and 0.70 for hospice outcome) models that were developed in the previous two chapters, its performance was worse than the other two models for the mortality outcome and was comparable for the hospice outcome (Tables 4.9 and 4.15). The RF model outperformed the other two models for mortality, but not for hospice admission. A possible explanation for the poor performance of the RF in the prediction of hospice admission is the fact that missingness on predictors was significantly associated with mortality, but not with hospice admission (Table 2.6). In Chapter three, we discussed that the gain in AUC of the RF model for mortality was mainly due to including the incomplete cases. Since there was no association between the missingness and hospice outcome, including the missing observations, did not increase the accuracy of the RF model for hospice.

Cox model in this data did not show any improvement in the prediction accuracy compared to the other two models. One reason might be the fact that the maximum follow-up time in this study was about two years, and the mean was only 1.25 years. So there was not long enough follow-up time to make a difference in the model performance when including the time component in the analysis.

The mortality and hospice admission are similar outcomes, and so the predictors of them are expected to be similar when developing the models. There were six variables that predict both outcomes in different models (i.e., age, race, SQ, albumin, KPS, hyperlipidemia). However, the performance of the Cox model in prediction of mortality was a little better than the prediction of hospice admission. This difference can be in part due to the fact that unlike death, hospice admission is not completely a result

of patient's risk level, rather it depends on other factors, including patient and family preferences. For example, a patient who is at high risk for hospice admission (identified by the model) can refuse the hospice admission and therefore die at home or hospital. This scenario results in a false positive case (patients is high risk by the model but was not actually admitted). In fact, all the three approaches (LR, RF, and Cox models) had higher accuracy in prediction of mortality outcome than hospice admission. This fact supports the hypothesis that the nature of hospice admission outcome, and not the model specification itself, is the reason for this poorer performance of the models for hospice admission.

The importance of variables in the Cox model was appraised by the value of coefficient parameter estimates and the corresponding hazard ratios in the MV model. Comparing the importance of variables between the Cox model and the other two models, revealed a few predictors that are consistently selected in all approaches. These variables include age, race, albumin, SQ, and KPS, which were selected in all three approaches and for both outcomes (death, hospice admission). Older age was associated with an increasing rate of adverse outcomes. However, the hazard ratio for the oldest old (95+ years) was lower compared to the age group 85-95 t=years (1.5 vs. 1.8). This paradoxical effect might be due to survival bias(153), meaning those who survived up to ages >95 years had better health than the other group. Surprisingly, black patients had a lower rate of adverse outcomes compared to white. On average black patients are younger than whites in this cohort (mean 79 vs. 83 years). However, the association of race and outcomes persisted after adjustment for age. There might, therefore, be unobserved characteristics of the population that made the black patients overall healthier than the whites. These five variables are strongly predicting the adverse outcomes in this population of older adults. Albumin is a surrogate of the patient nutritional status (154) and low albumin is associated with impaired functional status and disability. (155) Low albumin and low cholesterol both have been shown to be associated with an increased rate of death in older adults. (156–159) Different factors can explain the effect of low albumin on the mortality rate. For example, poor nutrition and low albumin

concentration can be indicators of an underlying disease or the patient's inability to take care of

themselves. Albumin level decreases with increasing age independent of health status. (160)

Additionally, low levels of albumin have been shown to be indicators of inflammation and inadequate

nutrition in patients with chronic conditions. (161) As discussed in Chapter two, importance of the

answer 'No' to SQ in the prediction of mortality has been shown frequently in cancer and chronic kidney

diseases (98,99,162) however, its prognostic value in older adults was not well evaluated.

KPS is an indicator of the patient's functional status and disability and lower KPS that indicates severe

disability had a higher rate of adverse outcomes. KPS score has been often used to determine the

prognosis of cancer patients.(97,163,164) Its performance in the prediction of adverse outcomes among

the older adults was better or equally well as the ADL and IADL measures. (165) In this cohort of

community-living older adults, lower KPS was an essential predictor of the adverse outcomes (Tables

4.10 and 4.16).

Other variables such as ADL-decline, and cholesterol, were consistently selected in all three models (Cox,

LR, and RF) only for the mortality outcome. Whereas for the hospice outcome, living alone, dual-

eligibility and number of lab tests ordered were also important in the prediction of outcome.

Interestingly, a history of hyperlipidemia had usually a protective effect against the adverse outcomes. It

might partly be due to the known protective effect of lipid-lowering medications, particularly statins,

which increase survival in cardiovascular diseases. (12–14) Many of the chronic conditions in this data

had an inverse association with the outcomes, including diabetes, hyperlipidemia, hypertension,

depression, cataract, and chronic kidney diseases. In univariate analysis of the CCW variables and the

two outcomes (Table 4.2), all of these variables are significantly associated with a lower hazard of the

outcomes. However, in the adjusted analysis, nearly all of these effects became statistically non-

significant.

Among the 24 CCW comorbidities, often the presence of the comorbidity had a protective effect against the outcomes; in univariate analysis, 13 and nine comorbidities had statistically significant hazard ratios <1.0 (protective effect) for mortality and hospice outcome, respectively. However, almost all of these associations became non-significant after adjustment, except for hyperlipidemia, which was consistently significant and entered in almost all final models of the three approaches and for both outcomes. A possible explanation for the protective effect of hyperlipidemia might be the treatments that can affect the survival of patients. For example, lipid-lowering medications, especially statins, have been shown to decrease mortality in cardiovascular diseases.(166,167) Their effect in many other conditions and diseases has also been studied. (169–171)

Another potential explanation is the inconsistency in the documentation of the comorbidities. For instance, it is expected that the provider cannot complete the documentation of all comorbidities in a very sick patient. Since these CCW variables are recorded as binary (yes/no) variables in the APRIMA, it is likely that the default value is 'No' unless otherwise documented. Therefore if for some reason the information was not attainable, the comorbidity is recorded as absent. The same reasoning is valid for explaining the strong association between the missingness on predictors and the mortality. In this scenario, the EMR for sicker patients with a poorer prognosis is more likely to be incomplete on comorbidities compared to healthier patients with better prognosis. However, the prevalence of most chronic conditions in this population is higher than the US population of age ≥ 65 years old. The prevalence of chronic conditions in the US population was evaluated in a study using administrative claims data for a population-based cohort of over 31 million Medicare Fee-for-service beneficiaries. (7) For example prevalence of following conditions in this cohort vs. the general elderly are: hypertension (81% vs. 60%), hyperlipidemia (50% vs. 45%), heart failure (34% vs. 18%), COPD (26% vs. 11%), chronic kidney disease (40% vs. 13%), and cancer (8% vs. 7%). These findings weaken the previous assumption that there is a lack of documentation for chronic conditions in the APRIMA data; however, it still is

plausible that for those at highest risk of mortality the documentation of comorbidities is less than optimal. Lastly, it is likely that comorbidities such as cataract that were not lethal can be identified and treated more in older patients who survived because of the good general health. Anyhow it is not possible to confirm any of these potential explanations for some of the paradoxical associations between CCW comorbidities and outcomes.

In the analysis of time from hospice admission to death, the average survival time in hospice was 104 days, and the median was 58 days (Table 4.3). According to the Medicare criteria, a patient is eligible for hospice services, if determined to have a terminal illness (defined as having a prognosis of 6 months or less if the disease or illness runs its normal course). (35) In this cohort, 76% of patients who have been admitted to hospice died within the first six months. It indicates that the screening and referral process for the hospice-eligible patients accurately identify and refer these patients, so the criteria of life expectancy <6 months is met for 75% of the patients who were admitted to hospice. Mortality after hospice admission in this data was very high soon after admission, i.e., 21% died within seven days of admission, 59% died within three months of their admission, and 24% lived beyond six months of their admission. A large hospice study on Medicare beneficiaries who were enrolled in the hospice program in 5 US states showed that 15% of patients died in 7 days and 15% lived beyond six months of their enrollment date. The median survival in hospice was 36 days. (172) The higher rate of early death (<7 days) in the USMM cohort compared to the previous study (21% vs. 15%) implies that hospice referral was delayed until the very end of life for about one-fifth of those who were ultimately admitted to hospice. On the other hand, the higher rate of long stay (>6 months) in the USMM cohort compared to the previous study (24% vs. 15%) indicates that screening and referral process requires improvement, to avoid the potential for over-use of hospice facilities by the patients for whom the life expectancy was underestimated.

o   <u>Limitations</u>

Patient turnover in the USMM system is high; therefore about 23% of the 9627 patients who met the primary inclusion criteria for this study, were excluded because the total time they were under USMM care was < 1 year. Another limitation of this study was the missing data. Some key variables, such as a decline in IADLs, recent hospitalization, and recent fall were left out from the analysis due to a large number of missing observations. Moreover, some other variables that were included in the analysis had some missing observations. The missingness on those variables was strongly associated with the mortality outcome. On the other hand, in the SAS procedure PHREG (and LOGISTIC) the observations with missing on any predictor are excluded by default from the analysis at the beginning of the model development. It means some valuable information from observations with partly missing data was lost in this analysis.

Another limitation of this analysis was that the advanced variable selection methods that were used in the logistic regression model development, such as adaptive lasso and elastic net, are not available options for survival model development. However, I applied the commonly used variable selection methods, namely stepwise, backward, forward, in addition to a manual selection method. Additionally, there was no evidence of an improvement in the model performance using these advanced variable selection methods in LR models in this data.

In this analysis, the dataset was made by linking the USMM EMR database (APRIMA) to the processed claims data provided by a third party Company named eSolution. The claims data contained information on 62% of the 2015 cohort, which means 7790 patients were not linked to the claims data and excluded from the analyses. It is not clear whether these patients did not appear in claims data because they did not have any event, or because for some reason, their information was not obtained by eSolution. If the first hypothesis is correct, the event rates in this cohort, and consequently, the analyses results will change dramatically from what it is now.

Lastly, the lack of information about the patients' enrollment in the USMM programs, confined our ability in understanding and interpretation for the paradoxical findings. Answer to questions such as "how and when does a patient enroll in the USMM care?", "how long did they meet the definition of homebound?", "what are the motivations for joining the USMM program?", and "where did the patients receive care before?" can help to better understand the models and explain the findings.

o Future direction

Having the USMM data for a longer follow up time and more complete data can help to improve the prediction model for survival analysis. The maximum follow-up time in this cohort was about two years, with an average of 1.25 years. In this study, separate Cox models were developed for each outcome, but since death and hospice admission can be competing risk, a future analysis accounting for competing risks (173,174) might be useful to assess the joint effect of the two outcomes on survival. Finally, in this research, only the baseline values of the independent variables were considered. Most of the independent variables did not change over the study period, however, if the data were available for a longer follow-up time, and documentation was improved to reducing missing data, predictors that may change over time especially functional measures (KPS, ADL, TUG), lab tests (albumin, cholesterol), and body weight could be evaluated as time-varying covariates. This trajectory-based analysis (with time-varying predictors) can be one of the future analysis when the required data is available.

The Cox model developed in this chapter can be later used to develop a prognostic index using the same methods applied by Fried and Carey. (57,58) A prognostic index is generated by assigning different points to the predictors (based on their Cox regression coefficients) and is easily usable for risk determination in different settings.

The RF model was the best among the three models (LR and Cox) for the mortality outcome, although its accuracy for hospice outcome was poor compared with the other two models. Survival tree is a similar concept to the decision tree, only with survival time as the outcome. Additionally, survival random

forest is an alternative method to the survival tree. (175) It develops multiple survival tree using

randomly selected subsamples of the data. The survival time was estimated by averaging over all the

survival trees. The software packages for survival tree and survival random forest is available R statistical

software and can be used in a future study to evaluate whether using this approach can improve the

survival analysis in this cohort.

- ***Conclusion***

The survival analysis of these data for the two outcomes of mortality and hospice admission did not

indicate any essential superiority to the LR or RF model. Despite the inclusion of additional outcome

events and taking account of the time to the event, the Cox model performance measures (C-index and

AUC at 365) were worse than the other two models for the mortality outcome and was comparable for

the hospice admission outcome. However, the most important predictors of both outcomes in this

analysis were consistent with the selected variables in the other two models. Variables age, race, KPS,

albumin, and SQ were among the most important predictors in all three approaches. This is to say that

collecting data on these variables is essential in the prediction of mortality and hospice admission

among homebound older adults.

This research aimed to develop, validate, and compare three different prediction models to be used for risk stratification in the USMM patient population. The USMM database was used to construct a cohort of community-living homebound older adults for this study. The three objectives of the study were:

4. To develop and validate multivariable logistic models for prediction of 12-month mortality and hospice admission among the USMM population of community-living homebound older adults.

5. To develop and validate a random forest (RF) algorithm for prediction of 12-month mortality and hospice admission among the USMM population. The model performance will be evaluated compared to the logistic regression (LR) model from aim 1 and Cox model from aim 3.

6. To develop and validate a multivariable failure time model (Cox proportional hazard) to model time-to-event for mortality and hospice admission separately. These models will also be compared to the logistic regression and random forest models developed in aims 1 and 2.

The prediction models developed for the three aims were compared primarily by their discrimination ability. The area under the receiver operating curve (AUC) and its equivalents for the Cox model were generated for the models. Calibration methods were also applied to evaluate and compare the model's goodness of fit for LR and RF models. Additionally, the specific variables that were selected in the final model for each approach were compared to evaluate the importance of individual predictors in the different models.

The important aspects of this study include:

1. Using a unique clinical population of community-living homebound older adults

2. Using a rich database that includes a wide range of different types of EMR-based information including demographics, socioeconomic variables, comorbidities, functional status, and

laboratory test results, that was linked to claims data to obtain information on outcomes events

and utilization

3. Using multiple imputation for missing data and applying the models developed in the available

   date to the imputed data in addition to developing models in the imputed data

4. Applying different variable selection methods including advanced methods such as adaptive

   lasso and elastic net to build multivariable models

5. Utilizing a machine learning algorithm -random forest- for model development in order to

   handle both missing data and account for potential non-linear relationships in the data

6. Comparing the different models by generating discrimination metrics for all three models

- ***Population***

This cohort of USMM patients is a group of older adults that is different from most of the other

comparable study populations that are summarized in chapter one. (48,49,53,54,56,57) Unlike

institutionalized older patients, USMM patients live in the community. However, importantly, they were

homebound based on the definition from the CMS. (111) These patients needed to receive health

services at home because they were unable to leave home to seek medical services or because,

according to a physician's judgment, leaving home would be associated with an unacceptably high level

of risk them. These characteristics made this cohort different from other populations that are commonly

studied in the literature, such as nursing home patients, (48) hospitalized older adults, (46,47) older

patients who visit the ER, (176,177) and community-living non-homebound elderly.(53,54) More

importantly, the one-year mortality rate in the USMM cohort (32%) was much higher than the mortality

rates reported in these other populations (Table 1.3) but was more comparable to the mortality rates

reported in nursing home populations which range from 17 to 34%.(59–61,178) Therefore because of

the uniqueness of the USMM population and the fact that existing RS models are likely not applicable to

this population there was a knowledge gap regarding the most appropriate RS models for the USMM

patient population. This dissertation aimed to develop alternative risk stratification models for this cohort.

- ***Data source***

The USMM database is a rich data source including a wide variety of different variables for each patient. The data was obtained from 2 different sources: USMM electronic medical record named APRIMA and the claims data processed by a third party called eSolution. (6) The USMM dataset includes many variables; however, a drawback of such a large dataset is the missing data. There are other problems in data collection (e.g., data on some variables were not collected on every visit, rather a previously collected value, was repeated for the next few visits), documentation (information on a single variable were documented in different datasets and therefore there was not a unique source of data for a single variable) and storage in each one of the different sources that may cause inaccurate inferences. For example, event rate in this population were calculated based on the reported event from claims data, where the data is not available for about 1/3 of the cohort. Therefore the accuracy of the analysis results which depends to the event rate, cannot be confirmed.

One of the main issues affecting the source data in this study is the uncertainty about those patients who were not linked to the claims data. There were 7790 patients (38%) in the USMM 2015 cohort that did not have any claims data reported, and so these subjects were excluded from all analyses. Each patient in the USMM database has a unique ID number and this ID links the APRIMA and claims databases together. The reason for a patient ID not being found in the claims data is unclear. There are two possible scenarios. The first possibility is that claims data for the 7790 IDs were missed for some reason; for example there was a delay between an event and reporting it in the claims data or claims data from patients with only private insurance are not reported to Center for Medicare and Medicaid Services (CMS). However, only 15% of the excluded group had commercial insurances so this

explanation seems unlikely. It means that coverage by a private insurance is not likely to be the reason

for absence of any claims data for the 7790 IDs.

The second possible explanation for the absence of claims data is that the 7790 subjects did not

experience any outcome event. If this scenario is correct, then exclusion of these observations from the

analysis substantially inflates the event rates in the remainder of the cohort, because all of the excluded

subjects were actually event-free (i.e., were alive and were not admitted to hospice) during the study

period. There is some evidence that the second scenario is false. There are a substantial number of

observations (n=6201, 49%) in the claims data which did not have any outcome event. However because

the claims data we received was processed data, it is possible that there was some other type of claims

information other than the events reported in the processed claims data that we received for the 7790

IDs who did not experience death or hospice. If this scenario is true then they should have been retained

in the analysis and assumed to be still alive and not in hospice. This uncertainty about the origin of the

claims data and the reason for missing claims for 7790 patients remains a major limitation of this

dataset.

- ***The importance of the missing data***

Missing data is a persistent problem in biomedical studies.(179,180) In this dataset, a different number

of missing observations were found for different variables. Variables with more than 20% missing

observations were not included in the model development phase of this study, so some valuable

information has certainly been overlooked. Variables such as 'decline in IADLs', 'general health reported

by patient', 'fall' and 'hospitalization' were some of the variables that were excluded. The high missing

rates of data likely reflect USMM's approach to data collection and documentation. For example, some

variables listed in APRIMA require medical examination and documentation by USMM staff at each

home visit, while others are documented only annually at the annual wellness visit. In the latter

category are variables such as 'IADL decline' and 'general health' which are part of the annual wellness

visit. However, other variables such as 'hospitalization since last visit' and 'fall since last visit' should

have been recorded at each medical visit In other words, when a variable is evaluated annually (e.g.,

change in general? health compared to the prior year) it will be recorded as missing in the routine visits

that are conducted every 4 weeks or so between the annual wellness visits. H however missing data on

variables that explicitly indicate an incident events since last visit (i.e., fall, hospital visit) cannot be

explained by this same mechanism.

Furthermore, there are nine variables that were included in the analysis that had missing rates between

0.4% and 20%. These variables were race, surprise question, Timed Up and Go (TUG), living-alone,

decline in activities of daily living (ADL), albumin level, cholesterol level, smoking status, and KPS (Table

2.5). In univariate analysis of these variables with the outcomes, where missing values counted as a

legitimate category in the analysis, missingness was significantly associated with at least one of the

outcomes which suggests that the data is missing not at random (MNAR).

Typically, in most SAS statistical procedures, observations with any missing values are excluded from the

analysis by default. (86) Therefore in this study, about one third of patients with partly missing

observations were excluded automatically at the beginning of the model development in LR and Cox

models. Given the strong association found between the missingness of some predictors and the

outcomes (Table 2.5), the exclusion of these data can induce bias into the results. In other words, the

missingness in this data is informative and ignoring it can potentially undermine the validity of the

results. (181)

In order to further evaluate the influence of missing observations, a multiple imputation procedure was

applied to this data. The assumption was that with inclusion of all observations the risk of bias due to

missing data is reduced. Also, with increased number of observations that are included in the analysis,

the model does not lose its power and precision.(88,181) Surprisingly, using the imputed data did not

improve the prediction model performance in this data. With the LR model, the model developed in the

imputed data did not have a better discrimination than the original model developed in the available

data. Application of the Cox model to the imputed data is not straightforward and the results from the

imputed sets cannot be summarized with a single measure of performance or discrimination. (182)

Therefore the Cox model was not applied in the imputed data.

In the RF model, missing data are not excluded from the analysis. As explained in chapter 3, the random

forest procedure lets the missing values be included in the analysis as a legitimate category. The

performance of random forest model was remarkably better than the LR model for the mortality

outcome. The RF model that was developed in available data, was also applied to the imputed data; the

discriminative performance of the model in the imputed data was similar to the LR model and was

notably worse than the RF in the available data which means that imputation for missing values in this

data cannot capture the information that missing values represent. The reason is that the basic

assumption of the multiple imputation method used in SAS was that the data were MAR (missing at

random), whereas the associations between the missingness of independent variables and the

outcomes suggested that MNAR is the more likely explanation of the mechanism of missing data.

To conclude, missing data in this study was an important predictor of the outcomes. In the RF model,

exclusion of the missing observations (i.e. limiting the cohort to the same members as in LR model) or

imputation of the missing values (applying the RF model to the imputed data) diminishes the

performance of the model equally. These results are therefore supportive of the hypothesis that missing

data in this dataset is missing not at random (MNAR). The multiple imputation procedure uses the

assumption of missing at random (MAR) for imputation, (86,88) which is probably the main reason why

the model performance was worse in the imputed data than the available data. The better performance

of the RF model for mortality outcome when including missing observations, suggests the clear

advantage of the RF when data are MNAR.

- ***Using multiple imputation method in management of missing data***

Missing data in this analysis resulted in the exclusion of almost one-third of the observations from the LR and Cox analyses. Multiple imputation (MI) is a commonly used method in dealing with missing data. (181) The SAS procedure PROC MI builds a specified number of imputed datasets; PROC MIANALYZE applies a specified model to each data and summarizes the results from all imputations to generate measures of interest such as regression coefficients and effect size (e.g., odds ratio or relative risk). The MIANALYZE procedure reads and combines the coefficients and standard errors that were generated from the model in each imputed set. These statistics are stored in tables and covariance matrices produced by the regression model in each imputation. There are two sources of variance when multiple imputation is used: 'within imputation variance' which is the result of the variation between observations in each imputed dataset, and 'between imputation variance' which is the result of the variation in the data between the different imputed dataset. Using the between and within covariance matrices, PROC MIANALYZE derives valid multivariable inferences based on Wald tests. (183)

The MIANALYZE procedure does not support AUC option or its equivalent, so to summarize the AUCs from the imputed datasets in the LR model we applied a manual method that was described in chapter two. The method involved taking the average of all predicted probabilities from the 20 imputed datasets for each patient and then generating an estimate of the AUC from these average probabilities. This method cannot be applied to the Cox model results, because generating the AUC from the averaged survival is not an option in PROC PHREG. Lastly, as discussed above, the underlying assumption of missing at random for the multiple imputation procedure was not satisfied in this data. The exact mechanism of missing cannot be identified in this data. Although we can reject that the data missing is completely at random (MCAR), it is not possible to distinguish between missing at random (MAR) and missing not at random (MNAR). However, as evidenced by the significant association of missingness on predictors and the mortality outcome, MNAR is likely the primary mechanism of missing in this data.

Thus multiple imputation may not be an appropriate method to handle the missing data if it is MNAR. A

sensitivity analysis can test the appropriateness of the MI procedure in this data by using a pattern-

mixture model approach which models the distribution of a response as the mixture of a distribution of

the observed responses and a distribution of the missing responses. (86)

As a conclusion, the multiple imputation procedure used in this analysis did not improve the model

performance compared to the model based on only available data. The most likely explanation is that

that MI uses the assumption of MAR, whereas evidences suggests that MNAR is the mechanism of

missing in the USMM dataset.

- ***Variable selection methods***

In this analysis different variable selection methods including stepwise, backward, and forward selection

were applied to develop the LR and Cox models. In LR models, the more advanced methods of adaptive

lasso and elastic net were also used in variable selection. Although SAS does not support these methods

in the logistic procedure, these selection methods are supported in GLMSELECT procedure and were

used to select the variables for the LR model. However, using these variable selection methods did not

improve the performance of the models in this data. As explained in chapter two, adaptive lasso and

elastic net are useful methods in big data analysis where the number of predictors is very large and the

number of observations is relatively small (high dimensional data) such as genetic analysis data.

(102,184) It was not the case in this study where the number of observations was almost 200 times the

number of predictors.

- ***Using random forest method***

Use of machine learning (ML) algorithms has been increasing in many disciplines including biomedical

research. (65,122) Some studies have found that ML-based analyses outperformed the traditional

methods in finding risk predictors and improving the predictive model accuracy. (185) However accuracy

of any predictive models, ML-based or not, depends on the quality of the data. Thus common problems

with the EMR data (such as missing data, timeliness of the available data, and poor quality data) affect

the ML-based methods same as other traditional methods.(186) Random forest is the machine learning

algorithm that has been used in this dissertation. Two key advantages of random forest are its ability to

include incomplete (partly missing) observations and to capture non-linear relationships and complex

interactions. (69) Using random forest to develop a prediction model in this data let the missing values

to be included as legitimate values in the analyses. In other words, all observations can contribute in

model development without the need for imputation of missing values. The random forest resulted in

substantially improved discrimination compared to the LR model for the mortality outcome. Although

the primary reason of using the RF model in this data was to explore and capture any non-linear

relationships (higher degree) and complex interactions in the data, the improvement in the RF model

performance for mortality outcome was mainly due to the inclusion of missing data; we concluded this

because when the RF model was applied to the subjects with no missing (that were analyzed in the LR

model), the model's AUC was very similar to the LR model AUC. Also when the RF was applied to the

imputed data, the AUC was again similar to the LR model. Therefore RF improved the discrimination

only when missing observations are included as missing. Additionally, when the missing values were

recoded as a legitimate category and included in the LR model, the LR model performance was

comparable (and slightly better) than the random forest model. It confirms again that the gain in the

AUC of RF is almost completely due to inclusion of missing data.

In contrast to the mortality outcome, the RF performance in the analysis for hospice outcome was not

notably different from the LR model. It can be concluded that missingness on the predictors is not

associated with hospice admission as it has been presented in Table 2.5. Missingness was itself a

predictor of the mortality and the possibility of MNAR mechanism for missing data was reinforced again.

For example when a patients is very sick and at the end of life, it is more likely that physicians or other

health professionals do not evaluate all of the predictors and complete the EMR.

The conclusion from the comparison of the RF and LR models was that RF model had substantially improved AUC for the mortality outcome compared to the LR model. The main advantage of RF model in this data was due to the inclusion of missing observations. The same AUC gain was also observed when the missing values were recoded and included in the LR model.

- ***Important predictors of mortality and hospice***

The importance of variables in the prediction of outcomes were evaluated by the magnitude of their effect in LR and Cox model (adjusted odds ratio and adjusted hazard ratio, respectively). Unlike LR and Cox models, RF does not provide coefficients for the predictors, rather it generates a table for ranked importance of the variables. A few variables were among the most important variables in all three approaches and for both outcomes. These variables were age, race, SQ, albumin, KPS, and hyperlipidemia. ADL-decline and cholesterol level were also selected in multiple models.

Older age and male sex are associated with higher rate of both outcomes. African American patients in our study had lower risk of mortality (adjusted OR=0.59, 95% CL=0.42 – 0.83) and hospice admission (adjusted OR=0.65, 95% CL=0.43 – 1.0) than whites. A study that evaluated the racial difference in mortality among Medicare beneficiaries demonstrated a substantially higher mortality among Black older adults. (187) The lower mortality rate in African American compared to the whites in this study could have been explained by the age difference between the two race groups, black patients were younger than the whites (mean 79 vs 83 years). However, the association of race and mortality persisted after adjustment for age. There might be therefore unobserved factors (such as socio economic status, or education) that caused the black patient in this cohort less susceptible to death and hospice admission.

As expected, a 'No' answer to the surprise question was also strongly related to both outcomes. Validity of the answer No to SQ in prediction of mortality has been shown frequently in cancer and chronic kidney diseases (98,99,162) however, its prognostic value in older adults was not well evaluated. A study

in Spain showed the value of SQ as a screening tool to identify older patients who may require palliative care.(188)

Albumin has been used as a surrogate of the patient nutritional status.(154) Low albumin is associated with inflammation (161), impaired functional status and disability.(155) Low albumin and low cholesterol both have been shown to be associated with an increased rate of death in older adults.(156–160) Different factors can explain the effect of low albumin on the mortality rate. For example poor nutrition and low albumin concentration can be indicators of an inflammatory status, an underlying disease, or patient's inability to take care of themselves. Albumin level decreases with increasing age independent of health status. (160) In our study, lower level of albumin and cholesterol were associated with higher risk of death and hospice admission. Albumin is consistently was among the most important variables in different models for both outcomes (Tables 4.8 and 4.14).

KPS is an indicator of patient functional status and disability. Lower values of KPS indicate more severe disability and is associated with higher rate of adverse outcomes. KPS score has been often used to determine the prognosis of cancer patients,(97,163,164) and is used as part of hospice eligibility criteria in some diseases such as cancer.(189) Its performance in prediction of adverse outcomes among a population of elderly veterans (who were referred to geriatric care clinic) was better or equally good as the use of ADL and IADL measures. (165) In the USMM cohort of community-living older adults, lower KPS was an essential predictor of adverse outcomes (Tables 4.8 and 4.14). Routine documentation of KPS is valuable approach for health care programs in older adults.

With respect to changes in ADL compared to the prior assessment, 66% of the USMM had no-change whereas 14% declined and 4% improved (data were missing in 16%). Improvement in ADL compared to 'no-change' was, as expected, associated with lower risk of death. However, unexpectedly, a decline in ADLs also had slightly lower risk compared to 'no-change'; the latter association was not statistically significant. We do not have an explanation for this paradoxical finding.

Timed up and go (TUG) variable had 20% missing values but was included in the model development.

This variable was the first ranked important variable in RF analysis for the mortality and the 9th for the

hospice outcome. As mentioned before, in LR and Cox models, observations with any missing value are

left out from the analysis by default, whereas in RF the partly-missing observations are also included.

One can conclude that missing on TUG is again an important predictor of the outcomes in this data. To

measure TUG, the patient needs to understand the test and have motivation and ability to do the test. It

is very likely that the doctor or other health professionals who visited the patients overlook testing TUG

for terminally ill patients, bed-bound patients, or when patient's safety can be a concern. Consequently,

missing on TUG will be a strong predictor of mortality or hospice admission.

In this dataset, 24 variables representing the CCW comorbidities were evaluated as the predictors of the

outcomes. Interestingly, often the presence of the comorbidity had a protective effect against the

outcomes. However, almost all of these associations became non-significant after adjustment, except

for hyperlipidemia which remained consistently significantly associated with better outcomes (and was

included in most of the final models). As discussed in chapter four, the reason for this protective effect

of hyperlipidemia may be in part due to the treatments that can affect the survival of patients. For

example, lipid-lowering medications, especially statins, have been shown to decrease mortality in

cardiovascular diseases. (166,167) The protective effect of statins have been reported in many other

conditions and diseases. (169–171)

Another potential explanation is the inconsistency in the documentation of the comorbidities. For

instance, it is expected that the provider does not complete the documentation of all comorbidities in a

very sick patient.(190) Since these CCW variables are recorded as binary (yes/no) variables in the

APRIMA, it is likely that the default value is 'No' unless otherwise documented. In this scenario, the EMR

for sicker patients with a poorer prognosis, are more likely to be incomplete on comorbidities than the

healthier patients with better prognosis. However, the prevalence of most chronic conditions in this population is higher than the US population of age ≥ 65 years old.

The prevalence of chronic conditions in the US population was evaluated in a study using administrative claims data for a population-based cohort of over 31 million Medicare Fee-for-service beneficiaries. (7) For example prevalence of following conditions in this cohort vs. the general elderly are: hypertension (81% vs. 60%), hyperlipidemia (50% vs. 45%), heart failure (34% vs. 18%), TIA/stroke (11% vs. 5%), diabetes (34% vs. 27%), atrial fibrillation (17% vs. 9%), COPD (26% vs. 11%), chronic kidney disease (40% vs. 13%), and cancer (8% vs. 7%). On the other hand, the comorbidity rates for some variables are lower in the USMM cohort: ischemic heart diseases (17% vs. 35%), osteoporosis (11% vs. 14%), and Alzheimer's disease (0% in this cohort vs. 13% in the US population). These results weaken the previous assumption that there is a lack of documentation for chronic conditions in the APRIMA data, however it still is plausible that for those at highest risk of mortality the documentation of comorbidities is less than optimal. Unfortunately it is not possible to confirm any of these potential explanations for some of the paradoxical associations between CCW comorbidities and outcomes.

Difference between the two outcomes - The two outcomes of interest, death, and hospice admission are clearly related variables. According to the Medicare criteria, a patient is eligible for hospice services, if determined to have a terminal illness (defined as having a prognosis of 6 months or less if the disease or illness runs its normal course). (35) Therefore the models for the two outcomes are expected to be similar in terms of selected predictors and the performance of the models. However, in this data using the same set of potential predictors, the model performance for the two outcomes was different in terms of the AUC and selected variables in the final models.

Mortality after hospice admission in this data was very high soon after admission i.e., 21% died within seven days of admission, 59% died within three months of their admission and 25% lived beyond 6 months of their admission. Median survival after hospice admission was 58 days. A large hospice study

of Medicare beneficiaries who were enrolled in hospice program in 5 US states showed that 15% of patients died in 7 days and 15% lived beyond 6 months of their enrollment date. The median survival in hospice was 36 days. (172) In the USMM cohort, early death (≤ 7 days) occurred in higher proportion of the patients than the previous study which implies the patients were referred to hospice late. However, it is noticeable that unlike death, hospice admission is dependent on factors other than the patient's clinical condition. For instance, unobserved variables such as patient or family preferences can influence the admission and its timing. So it is probable that some caregivers preferred to take care of the patient at home until the very end of life. On the other hand, racial and ethnic disparity in end-of-life care has been shown in literature. Black patients are more likely to receive higher intensity (e.g., intensive care unit) and higher cost care (frequent hospitalizations and ER visits) instead of hospice enrollment at the end of life.(191–193) Therefore the late admission to hospice does not necessarily mean that the original hospice referral by USMM providers was late. On the other hand, about 25% of the hospice admitted patients lived beyond six months compared to the 15% in the national study which implies that screening and referral process require improvement, to avoid the potential for over-use of hospice facilities by the patients for whom the life expectancy was underestimated.

- *Limitations*

This study had limitations. The details of limitations was provided in each chapters 2-4, here is a summary of the limitations of this dissertation.

1. Although the USMM database is a rich data set with a wide range of information collected, but missing data is a serious problem. There are potential predictors that are not included in the analysis because of the high missing rate: decline in IADL function since the last visit, a decline in global health since last year, falls, hospitalizations and ER events.

2. We used the independent variables data that were collected at baseline (first visit in the USMM system) because the change of variables over time were not well documented.

3. Another limitation of this analysis is the assumption about the mechanism of missing data. The Multiple Imputation procedure has a basic assumption of missing at random. We used multiple-imputation in this data although there were evidences that the mechanism of missing is this data is MNAR (missing not at random).

4. There were two comorbidity variables excluded from the analysis because the number of patients with the comorbidity was too small or zero.

5. To evaluate the accuracy of the model, we used the validation data which is originated from the same database as derivation cohort. Using an external validation data was useful to confirm external validity of the model.

- ***Future direction***

The models developed in this study were validated using the data from the same origin as the derivation data. To assess the external validity of the models, future application of the model to cohorts of community-living homebound older adults is needed. It is interesting to also evaluate the validity of the model among older adults who are not homebound. Inclusion of variables which indicates functional status in all possible ranges from normal to severely impaired (i.e., KPS, TUG, and ADL) allow to use these models in prediction of outcomes in general older population. In the RF model I used a single machine learning algorithm to predict the mortality outcome and it resulted in a remarkably improved discrimination ability of the model. Researchers commonly use an ensemble of different machine learning algorithms to obtain a better model. (74) Future studies can incorporate different machine learning algorithms to attain a prediction model with higher discrimination. In this study, separate Cox models were developed for each outcome, but since death and hospice admission can be competing risk, a future analysis accounting for competing risks (173,174) might be useful to assess the joint effect of the two outcomes on survival. Finally in this research only the baseline values of the independent

variables were considered. Most of the independent variables did not change over the study period, however if the data were available for a longer follow up time, and documentation was improved to reducing missing data, predictors that may change over time especially functional measures (KPS, ADL, TUG), lab tests (albumin, cholesterol), and body weight could be evaluated as time-varying covariates. This trajectory based analysis (with time-varying predictors) can be one of the future analysis when the required data is available. Learning about the quality of the USMM data was an important result of this dissertation. The quality of the USMM database needs substantial improvement; development of a protocol that regulate the data collection process and significantly improve the quality of database.

- ***Potential implementation of new RS approach for USMM***

Our models can be programmed and integrated into the electronic medical databases to stratify patients and provide them with targeted care. These developed model can be received by the USMM computer programmers using different statistical software that support our models (e.g., SAS and R support all three models: logistic, random forest, and Cox PH). Then the programmer can code the model into the data system in order to run the model on all observations at the time of each new data entry. Logistic and Cox regression models can be programmed into the USMM database by knowing the regression coefficients for each predictor, however for RF model, a statistical software is required. The fact that in ML-based algorithms, the user cannot directly see how exactly the predictions are generated, remains as a limitation in utilizing ML-based algorithms.(185) Ultimately a predicted probability for each patient is calculated from the model, and then a risk level will be assigned to them based on their probability of death (or hospice admission). The high-risk patients would be flagged and brought to the attention of the provider team for appropriate and timely intervention. The intervention can include a range of services such as a change in medications, nutritional support, additional home visit, hospice referral, or offering palliative care and advanced care planning. Lower risk patients can be targeted for other levels of services according to the USMM policies and care plans. For example,

preventive services such as providing medical equipment to reduce the risk of fall, screening tests, more

intensive treatment regimen for prevention of complications of diabetes, and rehabilitation referral are

services that may be offered to the patients with estimated long survival and low risk of adverse

outcomes.

As the conceptual framework from American Geriatrics Society Guiding Principles indicates (Table 1.2),

estimation of life expectancy and health trajectory is a part of the suggested care for older adults with

multiple comorbidity. It is important that all decisions and care options must be aligned with patient's

priorities and health trajectory; and must be communicated with patient, caregiver, and other clinicians.

Finally, the results of this research can be used by the USMM to improve the quality of the database in

terms of data collection and documentation.

- ***Conclusion***

As a conclusion, the different statistical approaches for the development of a prediction model in this

data resulted in similar model discrimination, except for the random forest model for the mortality

outcome which had remarkably better discrimination than other models. A few variables such as SQ,

KPS, and albumin were consistently associated with both outcomes. We think that these variables

should be considered by researchers who are working on prognostic indices for older populations. SQ

and KPS are simple but valuable pieces of information that can be quickly evaluated and documented by

physicians or other health providers.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Sonnega A, Robinson K, Levy H. Home and community-based service and other senior service use: Prevalence and characteristics in a national sample. *Home Health Care Serv Q*. 2017;36(1):16–28.

2. Siegler EL, Lama SD, Knight MG, et al. Community-Based Supports and Services for Older Adults: A Primer for Clinicians. *J Geriatr* [electronic article]. 2015. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4339950/). (Accessed October 29, 2019)

3. Program of All-Inclusive Care for the Elderly. *Centers for Medicare and Medicaid Services*. (https://www.medicaid.gov/medicaid/ltss/pace/index.html). (Accessed November 5, 2019)

4. Anderson KA, Dabelko-Schoeny HI, Fields NL. Home- and Community-Based Services for Older Adults: Aging in Context. Columbia University Press; 2018.

5. Medicare Bulletin- January 2019. (https://www.cgsmedicare.com/hhh/pubs/mb_hhh/2019/j15_hhh_01-19.pdf). (Accessed July 25, 2019)

6. Medical Billing and Claim Processing Software | Medicare Eligibility Software. *eSolutions*. (https://www.esolutionsinc.com/). (Accessed July 14, 2019)

7. Salive ME. Multimorbidity in Older Adults. *Epidemiol Rev*. 2013;35(1):75–83.

8. White N, Kupeli N, Vickerstaff V, et al. How accurate is the 'Surprise Question' at identifying patients at the end of life? A systematic review and meta-analysis. *BMC Medicine*. 2017;15(1):139.

9. Werner CA. The Older Population: 2010. 2010;(https://digitalcommons.unomaha.edu/cparpublications/60). (Accessed May 2, 2019)

10. Public Health and Aging: Trends in Aging --- United States and Worldwide. (https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5206a2.htm). (Accessed May 2, 2019)

11. Barnett K, Mercer SW, Norbury M, et al. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*. 2012;380(9836):37–43.

12. Guralnik JM. Assessing the impact of comorbidity in the older population. *Annals of Epidemiology*. 1996;6(5):376–380.

13. Alemayehu B, Warner KE. The Lifetime Distribution of Health Care Costs. *Health Services Research*. 2004;39(3):627–642.

14. de Meijer C, Wouterse B, Polder J, et al. The effect of population aging on health expenditure growth: a critical review. *Eur J Ageing*. 2013;10(4):353–361.

15.   Payne G, Laporte A, Deber R, et al. Counting Backward to Health Care's Future: Using Time-to-Death Modeling to Identify Changes in End-of-Life Morbidity and the Impact of Aging on Health Care Expenditures. *The Milbank Quarterly*. 2007;85(2):213–257.

16.   Gage Brian F., van Walraven Carl, Pearce Lesly, et al. Selecting Patients With Atrial Fibrillation for Anticoagulation. *Circulation*. 2004;110(16):2287–2292.

17.   Hustey FM, Mion LC, Connor JT, et al. A Brief Risk Stratification Tool to Predict Functional Decline in Older Adults Discharged from Emergency Departments. *Journal of the American Geriatrics Society*. 2007;55(8):1269–1274.

18.   Martin TP, Hanusa BH, Kapoor WN. Risk Stratification of Patients With Syncope. *Annals of Emergency Medicine*. 1997;29(4):459–466.

19.   Meldon SW, Mion LC, Palmer RM, et al. A Brief Risk-stratification Tool to Predict Repeat Emergency Department Visits and Hospitalizationsin Older Patients Discharged from the Emergency Department. *Academic Emergency Medicine*. 2003;10(3):224–232.

20.   Levy D, Wilson PWF, Anderson KM, et al. Stratifying the patient at risk from coronary disease: New insights from the framingham heart study. *American Heart Journal*. 1990;119(3, Part 2):712–717.

21.   Sanchis J, Bonanad C, Ruiz V, et al. Frailty and other geriatric conditions for risk stratification of older patients with acute coronary syndrome. *American Heart Journal*. 2014;168(5):784-791.e2.

22.   Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer Science & Business Media; 2008 508 p.

23.   ePrognosis. (https://eprognosis.ucsf.edu/). (Accessed October 19, 2019)

24.   Huang ES, Zhang Q, Gandra N, et al. The effect of comorbid illness and functional status on the expected benefits of intensive glucose control in older patients with type 2 diabetes: a decision analysis. *Ann Intern Med*. 2008;149(1):11–19.

25.   Mor V, Pacala JT, Rakowski W. Mammography for older women: Who uses, who benefits? *Journal of Gerontology*. 1992;47(Spec Issue):43–49.

26.   Schonberg MA, McCarthy EP, Davis RB, et al. Breast Cancer Screening in Women Aged 80 and Older: Results from a National Survey. *Journal of the American Geriatrics Society*. 2004;52(10):1688–1695.

27.   Meissner HI, Tiro JA, Haggstrom D, et al. Does Patient Health and Hysterectomy Status Influence Cervical Cancer Screening in Older Women? *J GEN INTERN MED*. 2008;23(11):1822.

28.   Wee CC, McCarthy EP, Phillips RS. Factors associated with colon cancer screening: the role of patient factors and physician counseling. *Preventive Medicine*. 2005;41(1):23–29.

29.   Goldberg TH, Chavin SI. Preventive Medicine and Screening in Older Adults. *Journal of the American Geriatrics Society*. 1997;45(3):344–354.

30.   Harrold J, Rickerson E, Carroll JT, et al. Is the Palliative Performance Scale a Useful Predictor of Mortality in a Heterogeneous Hospice Population? *Journal of Palliative Medicine*. 2005;8(3):503–509.

31.   Lau F, Downing GM, Lesperance M, et al. Use of Palliative Performance Scale in End-of-Life Prognostication. *Journal of Palliative Medicine*. 2006;9(5):1066–1075.

32.   Glare P, Eychmueller S, Virik K. The use of the palliative prognostic score in patients with diagnoses other than cancer. *Journal of Pain and Symptom Management*. 2003;26(4):883–885.

33.   Arenella C. The Importance of Risk Stratification for Referrals to Palliative Care Programs. *National Hospice and Palliative Care Organization*. 2016;

34.   Hospice_Card__JSR_SSR_JMH_20.pdf. (https://cdn.ymaws.com/www.nmnpc.org/resource/resmgr/2018_annual_conf-_presentations-handouts/6_johnson/Hospice_Card__JSR_SSR_JMH_20.pdf). (Accessed October 22, 2019)

35.   Casarett DJ. Rethinking Hospice Eligibility Criteria. *JAMA*. 2011;305(10):1031–1032.

36.   Todd. PRIME Registry |Family medicine's only specialty outpatient quality registry. *PRIME Registry*. (https://primeregistry.org/). (Accessed October 21, 2019)

37.   Risk_Stratification_Care_QuickStart_Guide.pdf. (https://primeregistry.org/wp-content/uploads/2019/08/Risk_Stratification_Care_QuickStart_Guide.pdf). (Accessed October 21, 2019)

38.   Steenkamer BM, Drewes HW, Heijink R, et al. Defining Population Health Management: A Scoping Review of the Literature. *Population Health Management*. 2016;20(1):74–85.

39.   Sprague L. Disease Management to Population-Based Health: Steps in the Right Direction? *NHPF Issue Brief*. 2003;(791):16.

40.   Action-Guide_Pop-Health_Models-of-Care-Sept-2017.pdf. (http://www.nachc.org/wp-content/uploads/2017/09/Action-Guide_Pop-Health_Models-of-Care-Sept-2017.pdf). (Accessed October 21, 2019)

41.   Lavery LA, Armstrong DG, Wunderlich RP, et al. Predictive Value of Foot Pressure Assessment as Part of a Population-Based Diabetes Disease Management Program. *Diabetes Care*. 2003;26(4):1069–1073.

42.   Haas LR, Takahashi PY, Shah ND, et al. Risk-stratification methods for identifying patients for care coordination. *Am J Manag Care*. 2013;19(9):725–732.

43.   Tkatch R, Musich S, MacLeod S, et al. Population Health Management for Older Adults. *Gerontol Geriatr Med* [electronic article]. 2016;2. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5486489/). (Accessed October 21, 2019)

44.   Guiding Principles for the Care of Older Adults with Multimorbidity: An Approach for Clinicians. *Journal of the American Geriatrics Society*. 2012;60(10):E1–E25.

45. Boyd C, Smith CD, Masoudi FA, et al. Decision Making for Older Adults With Multiple Chronic Conditions: Executive Summary for the American Geriatrics Society Guiding Principles on the Care of Older Adults With Multimorbidity. *Journal of the American Geriatrics Society*. 2019;67(4):665–673.

46. Inouye SK, Peduzzi PN, Robison JT, et al. Importance of Functional Measures in Predicting Mortality Among Older Hospitalized Patients. *JAMA*. 1998;279(15):1187–1193.

47. Pilotto A, Ferrucci L, Franceschi M, et al. Development and Validation of a Multidimensional Prognostic Index for One-Year Mortality from Comprehensive Geriatric Assessment in Hospitalized Older Patients. *Rejuvenation Research*. 2008;11(1):151–161.

48. Yourman LC, Lee SJ, Schonberg MA, et al. Prognostic Indices for Older Adults: A Systematic Review. *JAMA*. 2012;307(2):182–192.

49. Carey EC, Walter LC, Lindquist K, et al. Development and Validation of a Functional Morbidity Index to Predict Mortality in Community-dwelling Elders. *Journal of General Internal Medicine*. 2004;19(10):1027–1033.

50. Cappola AR, Fried LP, Arnold AM, et al. Thyroid Status, Cardiovascular Risk, and Mortality in Older Adults. *JAMA*. 2006;295(9):1033–1041.

51. Studenski S, Perera S, Patel K, et al. Gait Speed and Survival in Older Adults. *JAMA*. 2011;305(1):50–58.

52. Gagne JJ, Glynn RJ, Avorn J, et al. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *Journal of Clinical Epidemiology*. 2011;64(7):749–759.

53. Han PKJ, Lee M, Reeve BB, et al. Development of a Prognostic Model for Six-Month Mortality in Older Adults With Declining Health. *Journal of Pain and Symptom Management*. 2012;43(3):527–539.

54. Lee SJ, Lindquist K, Segal MR, et al. Development and Validation of a Prognostic Index for 4-Year Mortality in Older Adults. *JAMA*. 2006;295(7):801–808.

55. Schonberg MA, Davis RB, McCarthy EP, et al. Index to Predict 5-Year Mortality of Community-Dwelling Adults Aged 65 and Older Using Data from the National Health Interview Survey. *J GEN INTERN MED*. 2009;24(10):1115.

56. Fischer SM, Gozansky WS, Sauaia A, et al. A Practical Tool to Identify Patients Who May Benefit from a Palliative Approach: The CARING Criteria. *Journal of Pain and Symptom Management*. 2006;31(4):285–292.

57. Carey EC, Covinsky KE, Lui L-Y, et al. Prediction of Mortality in Community-Living Frail Elderly People with Long-Term Care Needs. *Journal of the American Geriatrics Society*. 2008;56(1):68–75.

58. Fried LP, Kronmal RA, Newman AB, et al. Risk Factors for 5-Year Mortality in Older Adults: The Cardiovascular Health Study. *JAMA*. 1998;279(8):585–592.

59.   Tabue-Teguo M, Kelaiditi E, Demougeot L, et al. Frailty Index and Mortality in Nursing Home Residents in France: Results From the INCUR Study. *Journal of the American Medical Directors Association*. 2015;16(7):603–606.

60.   Li S, Middleton A, Ottenbacher KJ, et al. Trajectories Over the First Year of Long-Term Care Nursing Home Residence. *Journal of the American Medical Directors Association*. 2018;19(4):333–341.

61.   Vossius C, Selbæk G, Šaltytė Benth J, et al. Mortality in nursing home residents: A longitudinal study over three years. *PLoS One* [electronic article]. 2018;13(9). (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6143238/). (Accessed October 22, 2019)

62.   Flacker JM, Kiely DK. Mortality-Related Factors and 1-Year Survival in Nursing Home Residents. *Journal of the American Geriatrics Society*. 2003;51(2):213–221.

63.   Eng C, Pedulla J, Eleazer GP, et al. Program of All-inclusive Care for the Elderly (PACE): An Innovative Model of Integrated Geriatric Care and Financing. *Journal of the American Geriatrics Society*. 1997;45(2):223–232.

64.   Mazzaglia G, Roti L, Corsini G, et al. Screening of Older Community-Dwelling People at Risk for Death and Hospitalization: The Assistenza Socio-Sanitaria in Italia Project. *Journal of the American Geriatrics Society*. 2007;55(12):1955–1960.

65.   Koohy H. The rise and fall of machine learning methods in biomedical research. *F1000Res* [electronic article]. 2018;6. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5760972/). (Accessed May 28, 2019)

66.   Kording KP, Benjamin AS, Farhoodi R, et al. The Roles of Machine Learning in Biomedical Science. National Academies Press (US); 2018 (Accessed April 18, 2019).(https://www.ncbi.nlm.nih.gov/books/NBK481619/). (Accessed April 18, 2019)

67.   Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res* [electronic article]. 2016;18(12). (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5238707/). (Accessed February 28, 2019)

68.   Hastie T, Tibshirani R, Friedman J. Random Forests. In: *The Elements of Statistical Learning*. New York, NY: Springer New York; 2009 (Accessed February 28, 2019):1–18.(http://www.springerlink.com/index/10.1007/b94608_15). (Accessed February 28, 2019)

69.   Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5–32.

70.   Maniruzzaman Md, Rahman MdJ, Al-MehediHasan Md, et al. Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. *J Med Syst*. 2018;42(5):92.

71.   Xu W, Zhang J, Zhang Q, et al. Risk prediction of type II diabetes based on random forest model. In: *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*. 2017:382–386.

72. Ion Titapiccolo J, Ferrario M, Cerutti S, et al. Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients. *Expert Systems with Applications*. 2013;40(11):4679–4686.

73. Chen Y, Cao W, Gao X, et al. Predicting postoperative complications of head and neck squamous cell carcinoma in elderly patients using random forest algorithm model. *BMC Medical Informatics and Decision Making*. 2015;15(1):44.

74. Rose S. Mortality Risk Score Prediction in an Elderly Population Using Machine Learning. *Am J Epidemiol*. 2013;177(5):443–452.

75. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform Decis Mak*. 2011;11:51.

76. Chong S-L, Liu N, Barbier S, et al. Predictive modeling in pediatric traumatic brain injury using machine learning. *BMC Med Res Methodol* [electronic article]. 2015;15. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4374377/). (Accessed March 31, 2019)

77. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*. 2017;12(4):e0174944.

78. Kattan MW. Comparison of Cox Regression With Other Methods for Determining Prediction Models and Nomograms. *The Journal of Urology*. 2003;170(6, Supplement):S6–S10.

79. Horvath J, Berenson R. Developing the Right Approaches to Chronic Care in Medicare. *Medicare Policy Brief*. 2004;3.

80. Wagner EH, Austin BT, Von Korff M. Organizing Care for Patients with Chronic Illness. *The Milbank Quarterly*. 1996;74(4):511–544.

81. Ellrodt G, Cook DJ, Lee J, et al. Evidence-Based Disease Management. *JAMA*. 1997;278(20):1687–1692.

82. DeSalvo KB, Fan VS, McDonell MB, et al. Predicting Mortality and Healthcare Utilization with a Single Question. *Health Serv Res*. 2005;40(4):1234–1246.

83. Robert A. Cohen. SAS Global Forum 2009 Statistics and Data Analysis.

84. Cohen RA. Introducing the GLMSELECT procedure for model selection. In: *Proceedings of the Thirty-First Annual SAS Users Group International Conference*. SAS Institute Inc.; 2006:Paper 207.

85. Lund B. Logistic Model Selection with SAS® PROC's LOGISTIC, HPLOGISTIC,. *MWSUG 2017 - Paper AA02*. :18.

86. SAS® Help Center: PROC MI Statement. (https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_mi_syntax01.htm&docsetVersion=14.3&locale=en). (Accessed February 24, 2019)

87. Perkins NJ, Cole SR, Harel O, et al. Principled Approaches to Missing Data in Epidemiologic Studies. *Am J Epidemiol*. 2018;187(3):568–575.

88. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Statistics in Medicine*. 2008;27(17):3227–3246.

89. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its Associated Cutoff Point. *Biometrical Journal*. 2005;47(4):458–472.

90. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.

91. Chronic Conditions Data Warehouse. (https://www2.ccwdata.org/web/guest/home/). (Accessed November 12, 2019)

92. Lunney JR, Lynn J, Foley DJ, et al. PAtterns of functional decline at the end of life. *JAMA*. 2003;289(18):2387–2392.

93. Centers for Medicare and Medicaid Services Releases 2012 MCBS Access to Care Research Files. (https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/Downloads/Data_Brief_002.pdf). (Accessed September 13, 2019)

94. Millán-Calenti JC, Tubío J, Pita-Fernández S, et al. Prevalence of functional disability in activities of daily living (ADL), instrumental activities of daily living (IADL) and associated factors, as predictors of morbidity and mortality. *Archives of Gerontology and Geriatrics*. 2010;50(3):306–310.

95. Shumway-Cook A, Brauer S, Woollacott M. Predicting the Probability for Falls in Community-Dwelling Older Adults Using the Timed Up & Go Test. *Physical Therapy*. 2000;80(9):896–903.

96. Friendlander AH, Ettinger RL. Karnofsky performance status scale. *Special Care in Dentistry*. 2009;29(4):147–148.

97. Schag CC, Heinrich RL, Ganz PA. Karnofsky performance status revisited: reliability, validity, and guidelines. *JCO*. 1984;2(3):187–193.

98. Moss AH, Ganjoo J, Sharma S, et al. Utility of the "Surprise" Question to Identify Dialysis Patients with High Mortality. *CJASN*. 2008;3(5):1379–1384.

99. Moss AH, Lunney JR, Culp S, et al. Prognostic Significance of the "Surprise" Question in Cancer Patients. *Journal of Palliative Medicine*. 2010;13(7):837–840.

100. Lakin JR, Robinson MG, Obermeyer Z, et al. Prioritizing Primary Care Patients for a Communication Intervention Using the "Surprise Question": a Prospective Cohort Study. *J Gen Intern Med*. 2019;34(8):1467–1474.

101. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley; 1989.

102. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 2005;67(2):301–320.

225

103. Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*. 2006;101(476):1418–1429.

104. Jacob L, Obozinski G, Vert J-P. Group lasso with overlap and graph lasso. In: *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. Montreal, Quebec, Canada: ACM Press; 2009 (Accessed September 16, 2019):1–8.(http://portal.acm.org/citation.cfm?doid=1553374.1553431). (Accessed September 16, 2019)

105. Model-Selection Methods :: SAS/STAT(R) 14.1 User's Guide. (http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_glmselect_details01.htm). (Accessed September 16, 2019)

106. Elastic Net Selection (ELASTICNET) :: SAS/STAT(R) 13.2 User's Guide. (http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/viewer.htm#statug_glmselect_details12.htm). (Accessed May 21, 2019)

107. Elastic Net Selection (ELASTICNET) :: SAS/STAT(R) 14.1 User's Guide. (http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_glmselect_details12.htm). (Accessed September 16, 2019)

108. Calibration plots in SAS. *The DO Loop*. (https://blogs.sas.com/content/iml/2018/05/14/calibration-plots-in-sas.html). (Accessed May 22, 2019)

109. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*. 2014;33(3):517–535.

110. PROC LOGISTIC: The Hosmer-Lemeshow Goodness-of-Fit Test :: SAS/STAT(R) 9.3 User's Guide. (https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_logistic_sect046.htm). (Accessed April 12, 2019)

111. The homebound requirement. *Medicare Interactive*. (https://www.medicareinteractive.org/get-answers/medicare-covered-services/home-health-services/the-homebound-requirement). (Accessed June 15, 2019)

112. Covinsky KE, Justice AC, Rosenthal GE, et al. Measuring Prognosis and Case Mix in Hospitalized Elders: The Importance of Functional Status. *J GEN INTERN MED*. 1997;12(4):203–208.

113. DePalma G, Xu H, Covinsky KE, et al. Hospital Readmission Among Older Adults Who Return Home With Unmet Need for ADL Disability. *Gerontologist*. 2013;53(3):454–461.

114. Hebert PR, Gaziano JM, Chan KS, et al. Cholesterol Lowering With Statin Drugs, Risk of Stroke, and Total Mortality: An Overview of Randomized Trials. *JAMA*. 1997;278(4):313–321.

115. Mills EJ, Rachlis B, Wu P, et al. Primary Prevention of Cardiovascular Mortality and Events With Statin Treatments: A Network Meta-Analysis Involving More Than 65,000 Patients. *J Am Coll Cardiol*. 2008;52(22):1769–1781.

116. Omran ML, Morley JE. Assessment of protein energy malnutrition in older persons, part II: laboratory evaluation. *Nutrition*. 2000;16(2):131–140.

117. Whellan DJ, Cox M, Hernandez AF, et al. Utilization of Hospice and Predicted Mortality Risk Among Older Patients Hospitalized With Heart Failure: Findings From GWTG-HF. *Journal of Cardiac Failure*. 2012;18(6):471–477.

118. O'Hare AM, Bertenthal D, Covinsky KE, et al. Mortality Risk Stratification in Chronic Kidney Disease: One Size for All Ages? *JASN*. 2006;17(3):846–853.

119. Larrañaga P, Calvo B, Santana R, et al. Machine learning in bioinformatics. *Brief Bioinform*. 2006;7(1):86–112.

120. Fan J, Han F, Liu H. Challenges of Big Data analysis. *Natl Sci Rev*. 2014;1(2):293–314.

121. Lee CH, Yoon H-J. Medical big data: promise and challenges. *Kidney Res Clin Pract*. 2017;36(1):3–11.

122. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. *Pattern Recognition Letters*. 2010;31(14):2225–2236.

123. Singh SP, Jaiswal UC. Machine Learning for Big Data: A New Perspective. 2018;13(5):10.

124. Xu W, Zhang J, Zhang Q, et al. Risk prediction of type II diabetes based on random forest model. In: *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*. 2017:382–386.

125. Maniruzzaman Md, Rahman MdJ, Al-MehediHasan Md, et al. Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. *J Med Syst* [electronic article]. 2018;42(5). (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5893681/). (Accessed March 31, 2019)

126. Zhang W, Zeng F, Wu X, et al. A Comparative Study of Ensemble Learning Approaches in the Classification of Breast Cancer Metastasis. In: *2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*. Shanghai, China: IEEE; 2009 (Accessed November 8, 2019):242–245.(http://ieeexplore.ieee.org/document/5260680/). (Accessed November 8, 2019)

127. SONG Y, LU Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*. 2015;27(2):130–135.

128. Strobl C, Boulesteix A-L, Kneib T, et al. Conditional variable importance for random forests. *BMC Bioinformatics*. 2008;9(1):307.

129. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res* [electronic article]. 2016;18(12). (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5238707/). (Accessed March 31, 2019)

130. SAS Enterprise Miner 14.2: High-Performance Procedures. :273.

131. Neville, P. G., and Tan, P.-Y. A Forest Measure of Variable Importance Resistant to Correlations. Alexandria, VA: American Statistical Association; 2014

132. Breiman, L., and Cutler, A. Manual–setting up, using, and understanding random forests V4. 0. 2003;(https://www. stat. berkeley. edu/~ breiman/Using_random_forests_v4. 0. pdf.)

133. Schneider J, Hapfelmeier A, Thöres S, et al. Mortality Risk for Acute Cholangitis (MAC): a risk prediction model for in-hospital mortality in patients with acute cholangitis. *BMC Gastroenterol* [electronic article]. 2016;16. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4746925/). (Accessed March 31, 2019)

134. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology*. 2014;63(25 Part B):2935–2959.

135. D'Agostino RB, Vasan RS, Pencina MJ, et al. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation*. 2008;117(6):743–753.

136. Peng S-Y, Chuang Y-C, Kang T-W, et al. Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination. *European Journal of Neurology*. 2010;17(7):945–950.

137. Koslowsky S, Consultant SA, Hanks H. On Variable Importance in Logistic Regression - Predictive Analytics Times - machine learning & data science news. *Predictive Analytics Times*. 2018;(https://www.predictiveanalyticsworld.com/patimes/on-variable-importance-in-logistic-regression/9649/). (Accessed May 3, 2019)

138. Thompson D. Ranking predictors in logistic regression. *Paper D10-2009. Online available at http://www. mwsug. org/proceedings/2009/stats/MWSUG-2009-D10. pdf.(visited 2015, June 25)*. 2009;

139. Casarett DJ, Fishman JM, Lu HL, et al. The Terrible Choice: Re-Evaluating Hospice Eligibility Criteria for Cancer. *J Clin Oncol*. 2009;27(6):953–959.

140. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361–387.

141. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. John Wiley & Sons; 2011 464 p.

142. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1972;34(2):187–202.

143. PROC PHREG: Overview :: SAS/STAT(R) 9.3 User's Guide. (https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_phreg_sect001.htm). (Accessed July 8, 2019)

144. Cox DR. Analysis of Survival Data. Chapman and Hall/CRC; 2018 (Accessed July 15, 2019).(https://www.taylorfrancis.com/books/9781315137438). (Accessed July 15, 2019)

145. Liu L, Forman S, Barton B. 236-2009: Fitting Cox Model Using PROC PHREG and Beyond in SAS®. 2009;10.

146. Guo C. Evaluating Predictive Accuracy of Survival Models with PROC PHREG. :16.

147. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*. 2013;13(1):33.

148. Testing the proportional hazard assumption in Cox models. (https://stats.idre.ucla.edu/other/examples/asa2/testing-the-proportional-hazard-assumption-in-cox-models/). (Accessed July 8, 2019)

149. Patel K, Kay R, Rowell L. Comparing proportional hazards and accelerated failure time models: an application in influenza. *Pharmaceutical Statistics*. 2006;5(3):213–224.

150. Gardiner JC. Evaluating the accuracy of clinical prediction models for binary and survival outcomes. 2018

151. Hannan EL, Magaziner J, Wang JJ, et al. Mortality and Locomotion 6 Months After Hospitalization for Hip Fracture: Risk Factors and Risk-Adjusted Hospital Outcomes. *JAMA*. 2001;285(21):2736–2742.

152. Nguyen-Oghalai TU, Kuo Y, Zhang DD, et al. Discharge Setting for Patients with Hip Fracture: Trends from 2001 to 2005. *Journal of the American Geriatrics Society*. 2008;56(6):1063–1068.

153. Glesby MJM. Survivor Treatment Selection Bias in Observational Studies: Examples from the AIDS Literature. *Ann Intern Med*. 1996;124(11):999.

154. Ritchie CS, Burgio KL, Locher JL, et al. Nutritional status of urban homebound older adults. *Am J Clin Nutr*. 1997;66(4):815–818.

155. Salive ME, Cornoni-Huntley J, Phillips CL, et al. Serum albumin in older persons: Relationship with age and health status. *Journal of Clinical Epidemiology*. 1992;45(3):213–221.

156. Manolio T A, Ettinger W H, Tracy R P, et al. Epidemiology of low cholesterol levels in older adults. The Cardiovascular Health Study. *Circulation*. 1993;87(3):728–737.

157. Forette B, Tortrat D, Wolmark Y. Cholesterol as Risk Factor for Mortality in Elderly Women. *The Lancet*. 1989;333(8643):868–870.

158. Goldwasser P, Feldman J. Association of serum albumin and mortality risk. *Journal of Clinical Epidemiology*. 1997;50(6):693–703.

159. Sahyoun NR, Jacques PF, Dallal G, et al. Use of albumin as a predictor of mortality in community-dwelling and institutionalized elderly populations. *Journal of Clinical Epidemiology*. 1996;49(9):981–988.

160. Klonoff-Cohen H, Barrett-Connor EL, Edelstein SL. Albumin levels as a predictor of mortality in the healthy elderly. *Journal of Clinical Epidemiology*. 1992;45(3):207–212.

161. Don BR, Kaysen G. Poor Nutritional Status and Inflamation: Serum Albumin: Relationship to Inflammation and Nutrition. *Seminars in Dialysis*. 2004;17(6):432–437.

162. Ouchi K, Jambaulikar G, George NR, et al. The "Surprise Question" Asked of Emergency Physicians May Predict 12-Month Mortality among Older Emergency Department Patients. *Journal of Palliative Medicine*. 2017;21(2):236–240.

163. Weizer Alon Z., Joshi Daya, Daignault Stephanie, et al. Performance Status is a Predictor of Overall Survival of Elderly Patients With Muscle Invasive Bladder Cancer. *Journal of Urology*. 2007;177(4):1287–1293.

164. Yates JW, Chalmer B, McKegney FP. Evaluation of patients with advanced cancer using the karnofsky performance status. *Cancer*. 1980;45(8):2220–2224.

165. Crooks V, Waller S, Smith T, et al. The Use of the Karnofsky Performance Scale in Determining Outcomes and Risk in Geriatric Outpatients. *J Gerontol*. 1991;46(4):M139–M144.

166. Brugts JJ, Yetgin T, Hoeks SE, et al. The benefits of statins in people without established cardiovascular disease but with cardiovascular risk factors: meta-analysis of randomised controlled trials. *BMJ*. 2009;338:b2376.

167. Taylor F, Ward K, Moore TH, et al. Statins for the primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews* [electronic article]. 2011;(1). (https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD004816.pub4/abstract). (Accessed July 7, 2019)

168. Mann D, Reynolds K, Smith D, et al. Trends in Statin Use and Low-Density Lipoprotein Cholesterol Levels Among US Adults: Impact of the 2001 National Cholesterol Education Program Guidelines. *Ann Pharmacother*. 2008;42(9):1208–1215.

169. Vaughan CJ, Murphy MB, Buckley BM. Statins do more than just lower cholesterol. *The Lancet*. 1996;348(9034):1079–1082.

170. Almog Yaniv, Shefer Alexander, Novack Victor, et al. Prior Statin Therapy Is Associated With a Decreased Rate of Severe Sepsis. *Circulation*. 2004;110(7):880–885.

171. Søyseth V, Brekke PH, Smith P, et al. Statin use is associated with reduced mortality in COPD. *European Respiratory Journal*. 2007;29(2):279–283.

172. Christakis NA, Escarce JJ. Survival of Medicare Patients after Enrollment in Hospice Programs. *New England Journal of Medicine*. 1996;335(3):172–178.

173. So Y. Using the PHREG Procedure to Analyze Competing-Risks Data. :9.

174. Satagopan JM, Ben-Porat L, Berwick M, et al. A note on competing risks in survival data analysis. *British Journal of Cancer*. 2004;91(7):1229.

175. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. *Ann. Appl. Stat.* 2008;2(3):841–860.

176. Carpenter CR, Shelton E, Fowler S, et al. Risk Factors and Screening Instruments to Predict Adverse Outcomes for Undifferentiated Older Emergency Department Patients: A Systematic Review and Meta-analysis. *Academic Emergency Medicine*. 2015;22(1):1–21.

177. Hastings SN, Purser JL, Johnson KS, et al. Frailty Predicts Some but Not All Adverse Outcomes in Older Adults Discharged from the Emergency Department. *Journal of the American Geriatrics Society*. 2008;56(9):1651–1657.

178. Flacker JM, Kiely DK. Mortality-Related Factors and 1-Year Survival in Nursing Home Residents. *Journal of the American Geriatrics Society*. 2003;51(2):213–221.

179. Berglund PA. 265-2010: An Introduction to Multiple Imputation of Complex Sample Data Using SAS® 9.2. 2010;12.

180. Rubin DB. Inference and missing data. :12.

181. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.

182. Moscovici JL. Combining Survival Analysis Results after Multiple Imputation of Censored Event Times. :11.

183. SAS/STAT MIANALYZE Procedure. (https://support.sas.com/rnd/app/stat/procedures/mianalyze.html). (Accessed July 22, 2019)

184. PROC GLMSELECT Statement :: SAS/STAT(R) 13.1 User's Guide. (http://support.sas.com/documentation/cdl/en/statug/66859/HTML/default/viewer.htm#statug_glmselect_syntax01.htm). (Accessed February 24, 2019)

185. Peterson ED. Machine Learning, Predictive Analytics, and Clinical Practice: Can the Past Inform the Present? *JAMA* [electronic article]. 2019;(https://jamanetwork.com/journals/jama/fullarticle/2756195). (Accessed November 24, 2019)

186. Goldstein BA, Navar AM, Pencina MJ. Risk Prediction With Electronic Health Records: The Importance of Model Validation and Clinical Context. *JAMA Cardiol*. 2016;1(9):976–977.

187. Gornick ME, Eggers PW, Reilly TW, et al. Effects of Race and Income on Mortality and Use of Services among Medicare Beneficiaries. *New England Journal of Medicine*. 1996;335(11):791–799.

188.	Gómez-Batiste X, Martínez-Muñoz M, Blay C, et al. Utility of the NECPAL CCOMS-ICO© tool and the Surprise Question as screening tools for early palliative care and to predict mortality in patients with advanced chronic conditions: A cohort study. *Palliat Med*. 2017;31(8):754–763.

189.	Hospice Eligibility Criteria & Requirements: Crossroads. *Crossroad Hospice and Palliative Care*. (https://www.crossroadshospice.com/hospice-care/hospice-eligibility-criteria/##targetText=Hospice%20eligibility%20requirements%3A,taking%20into%20consideratio n%20edema%20weight)&targetText=Specific%20decline%20in%20condition). (Accessed November 5, 2019)

190.	Sharafoddini A, Dubin JA, Maslove DM, et al. A New Insight Into Missing Data in Intensive Care Unit Patient Profiles: Observational Study. *JMIR Med Inform* [electronic article]. 2019;7(1). (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6329436/). (Accessed November 15, 2019)

191.	Byhoff E, Harris JA, Langa KM, et al. An Examination of Racial and Ethnic Differences in End-of-Life Medicare Expenditures. *J Am Geriatr Soc*. 2016;64(9):1789–1797.

192.	Hanchate A, Kronman AC, Young-Xu Y, et al. Racial and Ethnic Differences in End-Of-Life Costs: Why Do Minorities Cost More Than Whites? *Arch Intern Med*. 2009;169(5):493–501.

193.	Rizzuto J, Aldridge MD. Racial Disparities in Hospice Outcomes: A Race or Hospice-Level Effect? *J Am Geriatr Soc*. 2018;66(2):407–413.