2020

Applications of Machine Learning Methods in Health Outcomes Research: Heart Failure in Women

Khalid Abdullah Alhussain
khalhussain@mix.wvu.edu

Recommended Citation

Alhussain, Khalid Abdullah, "Applications of Machine Learning Methods in Health Outcomes Research: Heart Failure in Women" (2020). *Graduate Theses, Dissertations, and Problem Reports*. 7979.
https://researchrepository.wvu.edu/etd/7979

2020

Applications of Machine Learning Methods in Health Outcomes Research: Heart Failure in Women

Khalid Abdullah Alhussain

Applications of Machine Learning Methods in Health Outcomes Research: Heart Failure in Women

Khalid Alhussain

Dissertation submitted
to the School of Pharmacy
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in
Health Services and Outcomes Research

Usha Sambamoorthi, PhD, Co-Chair
Nilanjana Dwibedi, PhD, Co-Chair
Traci J. LeMasters, Ph.D.
Ranjita Misra, PhD
Danielle E. Rose, PhD

Department of Pharmaceutical Systems and Policy

Morgantown, West Virginia
2020

Keywords: heart failure, women, postmenopausal women, unsupervised machine learning, supervised machine learning

ABSTRACT

Applications of Machine Learning Methods in Health Outcomes Research: Heart Failure in Women

Khalid Alhussain

There is robust evidence that heart failure (HF) is associated with substantial mortality, morbidity, poor health-related quality of life, healthcare utilization, and economic burden. Previous research has revealed that there are sex differences in the epidemiology, etiology, and disease burden of HF. However, research on HF among women, especially postmenopausal women, is limited. To fill the knowledge gap, the three related aims of this dissertation were to: (1) identify knowledge gaps in HF research among women, especially postmenopausal women, using unsupervised machine learning methods and big data (i.e., articles published in PubMed); (2) identify emerging predictors (i.e., polypharmacy and some prescription medications) of incident HF among postmenopausal women using supervised machine learning methods; (3) identify leading predictors of HF-related emergency room use among postmenopausal women using supervised machine learning methods with data from a large commercial insurance claims database in the United States. This study utilized machine learning methods. In the first aim, non-negative matrix factorization algorithms were used to cluster HF articles based on the primary topic. Clusters were independently validated and labeled by three investigators familiar with HF research. The most understudied area among women was atrial fibrillation. Among postmenopausal women, the most understudied topic was stress-induced cardiomyopathy. For the second and third aims, a retrospective cohort design and Optum's de-identified Clinformatics® Data Mart Database (Optum, Eden Prairie, MN), de-identified health insurance claims data, were used. In the second aim, multivariable logistic regression and three classification machine learning algorithms (cross-validated logistic regression (CVLR), random forest (RF), and eXtreme Gradient Boosting (XGBoost) algorithms) were used to identify predictors of incident HF among postmenopausal women. The associations of the leading predictors to incident HF were explored with an interpretable machine learning SHapley Additive exPlanations (SHAP) technique. The eight leading predictors of incident HF consistent across all models were: older age, arrhythmia, polypharmacy, Medicare, chronic obstructive pulmonary disease (COPD), coronary artery disease, hypertension, and chronic kidney disease. Some prescription medications such as sulfonylureas and antibiotics other than fluoroquinolones predicted incident HF in some machine learning algorithms. In the third aim, a random forest algorithm was used to identify predictors of HF-related emergency room use among postmenopausal women. Interpretable machine learning techniques were used to explain the association of leading predictors to HF-related emergency room use. Random forest algorithm had high predictive accuracy in the test dataset (Area Under the Curve: 94%, sensitivity: 93%, specificity: 77%, and accuracy: 0.81). We found that the number of HF-related emergency room visits at baseline, fragmented care, age, insurance type (Health Maintenance Organization), and coronary artery disease were the top five predictors of HF-related emergency room use among postmenopausal women. Partial dependence plots suggested positive associations of the top predictors with HF-related emergency room use. However, insurance type was found to be negatively associated with HF-related emergency room use. Findings from this dissertation suggest that machine learning algorithms can achieve comparable and better predictive accuracy compared to traditional statistical models.

ACKNOWLEDGMENTS

I would like to express the deepest appreciation and gratitude to my mentors Dr. Usha Sambamoorthi and Dr. Nilanjana Dwibedi for their continuous guidance, endless help, and continued support throughout the process of producing this dissertation. I am very grateful to my committee members, Dr. Traci J. LeMasters, Dr. Ranjita Misra, and Dr. Danielle E. Rose, for sharing their expertise, wisdom, and their valuable time to mentor my research work.

I would also like to express my sincere thanks to my colleagues and alumni for their continued help and valuable memories. They have been always there for me with their suggestions and assistance. I would also like to thank my friends for their support and encouragement, especially Abdulaziz Al Mouslem and Fahad Alzahrani.

I would like to acknowledge my funding sources, King Faisal University and the Saudi Arabian Cultural Mission (SACM), which made all of this work possible.

I wholeheartedly appreciate my family's role in achieving this goal. I would like to thank my parents and my siblings for the unconditional support and endless love that they provided to me during my graduate studies.

Table of Contents

CHAPTER 1

1. Introduction

1.1 Background and Significance

Heart failure and its epidemiology

Heart failure (HF) is a complex clinical condition that impairs the ability of the heart to eject or fill enough blood to meet the body's needs[1,2]. This condition affects about 64 million people globally[3], and it is growing in prevalence. The prevalence of HF varies across countries. For example, the prevalence of HF ranges from 1% to 6.7% in Asian countries[4], 1% to 2.2% in European countries[4], and 2.2% in the United States (US)[5]. The epidemiology of HF varies by sex and age. American men have a higher overall prevalence of HF than American women (2.4% vs. 2.1%)[5]. However, the incidence of HF is higher among older American women than their men counterparts[5].

Disease burden of HF

Although its prevalence seems to be relatively low compared to other cardiovascular diseases[5], HF is considered a major public health problem. This is because it is associated with substantial mortality, morbidity, poor health-related quality of life (HRQoL), healthcare utilization, and economic burden[5–12]. These negative consequences of HF affect men and women differently[13–15]. For example, women have higher HF mortality rates than men[5]. In the US, there were 78,356 deaths due to HF in 2016; about 55% of those deaths were among women[5]. In terms of HRQoL, a study by Dewan et al. revealed that women with HF reported lower scores on almost all domains of HRQoL compared to men with HF[13]. Furthermore, patients with HF have high healthcare utilization. HF hospitalizations are still high even after the slight decrease that has been observed over recent years[10,11]. In 2014, there were 978,135 hospital admissions and over a million emergency room (ER) visits due to HF in the US[7]. Most of those hospitalizations and ER visits were made by older patients (aged $\geq$ 65 years), specifically older women. About 38% (N= 367,779) of hospital admissions and 37% (N= 394,244) of ER visits were made by older women[7]. Because of this, the costs of HF management are high and will remain a significant concern for the US healthcare system. In 2012, total healthcare expenditures associated with HF were $20.9 billion[8]. These costs are projected to rise to $53.1 billion in 2030[8].

Etiology of HF in women

There is evidence that there are differences in HF etiology between men and women[14,15]. Women tend to develop HF at an older age compared to men[14,15] because young women are protected against the development of HF through the protective effect of female sex hormone, estrogen[16]. However, estrogen levels decrease after menopause. The decline in the level of endogenous estrogen can increase the risk of HF in postmenopausal women[17,18]. This may explain why older women (i.e., postmenopausal women) have a higher incidence of HF than older men. In addition to the estrogen effect, women and men differ in risk factors for HF. Although women and men share some risk factors for HF, these factors may affect them differently. For example, a systematic review and meta-analysis of cohort studies found that atrial fibrillation conferred higher risk for HF in women than men[19]. Another study indicated that hypertension confers higher HF risk in women, whereas the effect of myocardial infarction as a risk factor for HF is higher in men[15].

Research on HF among women, especially postmenopausal women

Despite the sex differences in HF etiology and disease burden, research on HF among women is limited. Women are often underrepresented in clinical trials for HF and their participation has not changed over time[20,21]. A systematic review examining the enrollment of women and other minorities in 118 HF clinical trials revealed that women represented only 27% of participants in clinical trials for HF, and the participation of older population was low[21]. Considering this information, we speculate that the representation of postmenopausal women in HF clinical trials is even lower. With such inclusion disparities, there may be significant knowledge gaps in HF research among women, especially postmenopausal women.

Modifiable risk factors for HF among postmenopausal women

Given the high incidence of HF in older women (i.e., postmenopausal women)[5], identification of risk factors for primary prevention of HF is crucial. This can reduce the disease burden and improve health outcomes in this population. Several studies have investigated risk factors for HF in postmenopausal women[22–28], but few included modifiable factors[23,28]. A study by LaMonte et al. examined the association between physical activity and HF incidence in postmenopausal women and found that levels of recreational physical activity, including walking, are inversely associated with HF risk[28]. Such finding is helpful for prevention of HF. Studies identifying modifiable risk factors for HF in postmenopausal women are needed.

Emerging risk factors for HF

There is emerging evidence that polypharmacy may increase the risk of HF[29]. A study by Chen et al. found that polypharmacy was associated with an increased risk for HF among older individuals with atrial fibrillation[29]. This increased risk can occur due to adverse drug reactions, drug-drug interactions, or both. Polypharmacy is common among postmenopausal women because of their high prevalence of multimorbidity[30,31]. Postmenopausal women are more likely to develop some health conditions such as vasomotor symptoms[32,33], diabetes mellitus[34,35], mental health conditions[36,37], bacterial infections[38], and pain[39,40]. These health conditions are treated with prescription medications such as oral antidiabetics, antiepileptics, and antibiotics. Prescription medication use can be effective to treat conditions that are prescribed for; however, they may increase the risk for HF in postmenopausal women[41].

Oral antidiabetic medications

Previous studies have suggested that some oral antidiabetic medications may increase the risk for HF[41,42]. For example, sulfonylureas, an antidiabetic class that exerts their hypoglycemic effects by stimulating insulin secretion from the pancreatic beta cells, have been found to be associated with a higher risk for HF compared to metformin[42,43]. This association was dose-response; higher doses of sulfonylureas were associated with a higher risk for incident HF[43]. Moreover, thiazolidinediones, an antidiabetic class that acts by improving insulin sensitivity, have been shown to increase the risk for HF in several meta-analyses included randomized controlled trials and observational studies[44–46]. Another oral antidiabetic class is dipeptidyl peptidase-4 (DPP-4) inhibitors including sitagliptin, saxagliptin, alogliptin, and linagliptin. These medications exert their hypoglycemic effects by increasing insulin secretion and decreasing glucagon levels through the prevention of the degradation of incretin hormones and glucagon-like peptide-1[47]. DPP-4 inhibitors have also been linked to HF risk. Results from a meta-analysis of all randomized trials of DPP-4 inhibitors indicated that patients using any DPP-4 inhibitor had a higher overall risk of acute HF compared to placebo or other classes[48]. This suggests a possible negative effect of this class; however, the mechanism of this effect is unclear. Unlike the above-mentioned oral antidiabetic medications, metformin may have cardiovascular benefits[42].

Antiepileptic medications

Pregabalin and gabapentin, structural analogues of the inhibitory neurotransmitter γ-Aminobutyric Acid (GABA), are widely used antiepileptic medications[49]. They are also used as analgesics in patients with neuropathic pain[49]. In a case report study, a 54-year-old woman with no cardiac history developed HF after a normal dose of pregabalin use[50]. The mechanism of the possible effect of pregabalin on incident HF is not well-understood. This may be because of the inhibition of the L-type calcium channels[50], which means gabapentin use could lead to the same effect[51]. In a Canadian population-based study, pregabalin was compared to gabapentin in terms of HF risk and no statistically significant differences were observed between both medications[52].

Antibiotics

Recently, concerns regarding the cardiovascular safety of antibiotics have been raised. In 2019, a study examined the association between antibiotic use and cardiovascular events in women[53]. After a follow-up of 7.6 years, 2.9% developed cardiovascular events. It was found that women who took antibiotics for 2 months or longer during late adulthood (age 60 and older) were 32% more likely to develop cardiovascular disease, and those used antibiotics for 2 months or longer in their middle age were 28% more likely to develop cardiovascular disease compared to those who did not use antibiotics in the same life-stage. The increased risk associated with antibiotic use could be explained by the alterations in the gut microbiota. In other words, antibiotics destroy probiotic bacteria (beneficial bacteria), which may increase the colonization of viruses, pathogenic bacteria, or other micro-organisms[54]. Prior research has linked the imbalance in the gut microbiota with inflammation and narrowing of the blood vessels, stroke, and heart disease[55–58]. Furthermore, a case-control study tied fluoroquinolones to the risk of aortic and mitral regurgitation, conditions in which the blood backflows into the heart[59]. This increased risk can occur due to the potential adverse effect of fluoroquinolones. The US Food and Drug Administration (FDA) has added a warning to the labeling of all fluoroquinolones stating that these drugs can increase the risk of rupture or dissection of aortic aneurysms[60]. The development of these heart valve disorders can lead to HF.

Emergency room use among postmenopausal women with HF

Even though HF is considered a chronic disease, those with HF require emergency care for acute symptoms, resulting in a high utilization ER[61]. A previous study has revealed that about one-third of patients with HF use the ER frequently[62]. Data from 2014 showed that American older women have higher HF-related ER visits than their men counterparts[7]. Such high utilization of ER imposes burden on the US healthcare system (i.e., high hospitalization and expenditures)[62,63]. In a study using data from more than 113,000 patients with HF in California and Florida hospitals, it was found that in one year $3.08 billion were spent on the ER and inpatient services for HF in Florida alone[62]. This burden can be reduced since the majority of HF-related ER use are avoidable[64].

Factors contributing to the emergency room use

Prior research refuted the common misperception that the uninsured individuals use the ER more than the insured individuals[65–67]. For example, a study using 2013 nationally representative survey data from the US found that 14.3% of insured adults (aged 19-64 years) had at least one ER visit, whereas 9.6% of uninsured adults used the ER at least once after adjusting for demographics and self-reported health status[65]. This emphasizes that health insurance does not guarantee access to primary care; even insured individuals may use ER because of the lack of access to primary care. Other patient-level factors associated with ER use have been identified in the previous studies[68–76]. For example, chronic physical conditions[71,75], mental illness[72,73], polypharmacy[71,74], and substance abuse[71] were found to be associated with ER use. However, those studies have been conducted among all adults, older individuals, and those with specific chronic conditions (e.g., diabetes and COPD).

Special needs for postmenopausal women that may increase ER use

Due to the hormonal changes, postmenopausal women may experience vasomotor symptoms such as hot flashes and night sweats. A study by Williams et al. indicated that 65% of American postmenopausal women experience vasomotor symptoms[32]. These symptoms can increase the probability of ER use[77]. In addition, postmenopausal women have a high prevalence of other factors contributing to ER use (i.e., mental illness)[36,37].

In summary, our literature review suggests the lack of 1) comprehensive review of the literature of HF among women, especially postmenopausal women; 2) real-world evidence on the effect of polypharmacy and some prescription medications used to treat co-existing health conditions among postmenopausal women (i.e., oral antidiabetics, antiepileptics, and antibiotics) on incident HF; 3) real-world evidence on predictors of HF-related ER use among postmenopausal women. It is imperative to fill these gaps in the literature. Identification of knowledge gaps in the literature of HF can provide an overall picture of HF research among women, particularly postmenopausal women. Such information can help researchers and funding agencies to address research gaps in this population. Furthermore, identification of modifiable predictors of HF including emerging risk factors (i.e., polypharmacy and prescription medication use) in real-world settings using diverse and representative population-based data can provide essential information for clinicians, payers, patients, and other stakeholders to weigh the harms and benefits of medications and personalize treatment plans. Moreover, an examination of leading predictors of HF-related ER use by utilizing real-world health insurance data can assist payers and policymakers to identify subgroups of postmenopausal women at high risk for ER use and develop specific interventions that could decrease ER utilization and improve health outcomes.

1.2 Innovation

1. There has been a transformational shift in population health landscape in terms of the availability of payer data for research and the requirement of electronic health records (EHR) to track patient's health, emphasis on patient outcomes and value-based care. The availability of big data due to this transformation has made health analytics an integral part of improving population health. The present study uses novel approaches such as topic modeling and predictive modeling.

2. This study represents a series of "firsts". It is the first study using big data (PubMed) and unsupervised machine learning methods to identify research topics in the literature of HF among women; the first study includes emerging risk factors (i.e., polypharmacy and prescription medication use) to identify predictors of incident HF among postmenopausal women. This can help identify those patients at risk for developing HF so that they can benefit from preventive care. It is the first study to identify predictors of HF-related ER use among postmenopausal women.

3. Use of natural language processing (NLP) and text mining techniques to screen and identify relevant articles and extract the objective(s) of each study from PubMed abstracts. This allowed us to provide less time-consuming methods.

1.3 Specific Aims

Aim 1: Identify knowledge gaps in heart failure research among women, especially postmenopausal women, using unsupervised machine learning methods and big data (i.e., articles published in PubMed).

Aim 2: Identify emerging predictors (i.e., polypharmacy and some prescription medications) of incident heart failure among postmenopausal women using supervised machine learning methods.

Hypothesis: Polypharmacy and use of fluoroquinolones, sulfonylureas, thiazolidinediones (TZDs), dipeptidyl peptidase-4 (DPP-4) inhibitors, gabapentin, and pregabalin will be positively associated with incident heart failure.

Aim 3: Identify leading predictors of heart failure-related emergency room use among postmenopausal women using supervised machine learning methods with data from a large commercial insurance claims database in the United States.

Hypothesis: Polypharmacy and the use of fluoroquinolones, sulfonylureas, dipeptidyl peptidase-4 (DPP-4) inhibitors, and gabapentin will be positively associated with heart failure-related emergency room use.

1.4 Approach

Machine learning techniques in health services and outcomes research

Machine learning (ML) methods have been in existence since 1950; however, the use of alternative, non-parametric ML approaches has risen significantly following the pioneering work by Breiman[78]. Numerous studies in health services and outcomes research have used ML methods and have found them to outperform traditional statistical approaches in some cases[79–81].

Unlike traditional parametric statistical models, ML methods are assumption-free and robust to outliers, multicollinearity issues, and high-level interaction terms[82].

Although multivariable logistic regression can be used to create predictive models, the predictive ability of logistic regression that uses only statistical significance may not be the best compared to ML algorithms. Therefore, we used supervised ML classification algorithms: 1) cross-validated logistic regression (CVLR), 2) random forests (RF), and 3) eXtreme Gradient Boosting (XGBoost). These algorithms were selected because of their growing popularity in clinical settings for prediction of binary outcomes and their ability to detect complex associations between the outcome and predictors and interactions between covariates[83,84].

The main advantage of CVLR is its ability to provide meaningful and easy-to-interpret results such as odds ratios (ORs), which can provide clinical information on the impact of predictors on the occurrence of the event of interest. RF algorithm, a tree-based technique, is becoming popular and has been shown to perform very well in medical settings[83,84]. RF algorithm has several advantages including its ability to handle missing data, run efficiently on large datasets, handle non-linearity and a large number of independent variables, and produce highly accurate and precise estimates[88].

In addition to their predictive abilities, ML methods provide more efficient and less time-consuming methods for text analysis. Unsupervised ML algorithms enabled us to cluster a large number of PubMed articles studying HF among women; this would not be feasible without ML methods.

Conceptual framework

We used the modified determinants of health outcome and chronic disease model, which was originally proposed by Wilkinson and Marmot[89]. This model was used to guide the selection of the study features (Figure 1), Based on this model, a disease incidence (i.e., HF) can be influenced by five domains. These domains include: *(1) biological factors* (e.g., age), *(2) access to care factors* (e.g., type of insurance), *(3) community resources* (e.g., geographical region), *(4) medication-related factors* (e.g., cardiovascular disease treatment such as polypharmacy and prescription medication use), and *(5) health-related risk factors*, which consist of two sub-domain: (a) chronic health conditions such as diabetes, asthma, and chronic obstructive pulmonary disease, and (b) lifestyle factors such as substance abuse and obesity)

Data sources

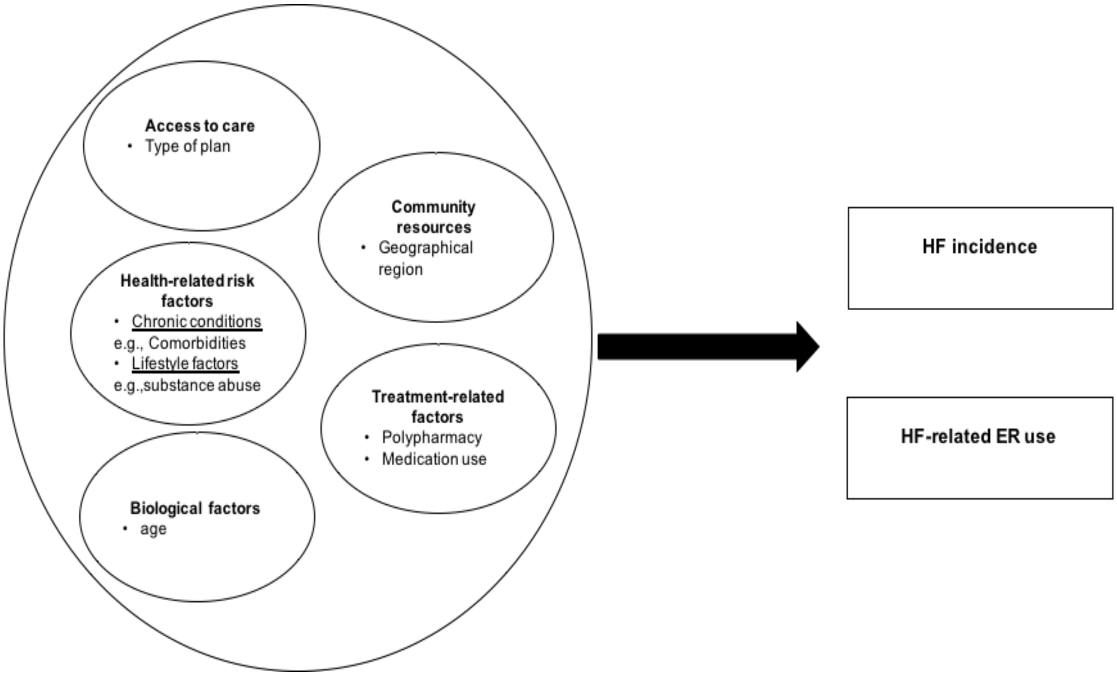
Chapter 2: PubMed

PubMed is a free resource supporting the search and retrieval of biomedical and life sciences literature and has been available since 1996. The PubMed database comprises more than 30 million citations and abstracts of biomedical literature from MEDLINE, life science journals, and online books. PubMed was developed and is maintained by the National Center for Biotechnology Information (NCBI), at the US National Library of Medicine (NLM), located at the National Institutes of Health (NIH)[90].

Chapter 3 & 4: Optum's de-identified Clinformatics® Data Mart Database (Optum, Eden Prairie, MN)

Data were derived from Optum's de-identified Clinformatics® Data Mart Database (Optum, Eden Prairie, MN). This geographically diverse database contains healthcare claims from a 10% sample of 47 million individuals. Of whom, about 80% purchased insurance through their employers. The data contain inpatient, outpatient and pharmacy claims, lab results, and certain demographic characteristics that are routinely collected during health insurance enrollment[91].

Figure 1: Adapted determinants of health outcomes and chronic disease model



1.5 References

1. Heart Failure. National Heart, Lung, and Blood Institute website. Accessed August 02, 2020. https://www.nhlbi.nih.gov/health-topics/heart-failure

2. Heart Failure. American Heart Association website. Accessed August 02, 2020. https://www.heart.org/en/health-topics/heart-failure

3. Lippi G, Sanchis-Gomar F. Global epidemiology and future trends of heart failure.

4. Savarese G, Lund LH. Global Public Health Burden of Heart Failure. *Card Fail Rev*. 2017;3(1):7-11. doi:10.15420/cfr.2016:25:2

5. Benjamin EJ, Muntner P, Alonso A, et al. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation*. 2019;139(10):e56-e528. doi:10.1161/CIR.0000000000000659

6. Hospitalization for Congestive Heart Failure: United States, 2000–2010. https://www.cdc.gov/nchs/products/databriefs/db108.htm. Accessed February 24, 2020.

7. Jackson SL, Tong X, King RJ, Loustalot F, Hong Y, Ritchey MD. National Burden of Heart Failure Events in the United States, 2006 to 2014. *Circ Heart Fail*. 2018;11(12):e004873. doi:10.1161/CIRCHEARTFAILURE.117.004873

8. Heidenreich PA, Albert NM, Allen LA, et al. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. *Circ Heart Fail*. 2013;6(3):606-619. doi:10.1161/HHF.0b013e318291329a

9. Bash LD, Weitzman D, Blaustein RO, Sharon O, Shalev V, Chodick G. Comprehensive healthcare resource use among newly diagnosed congestive heart failure. *Isr J Health Policy Res*. 2017;6:26. doi:10.1186/s13584-017-0149-0

10. Chen J, Normand S-LT, Wang Y, Krumholz HM. National and Regional Trends in Heart

Failure Hospitalization and Mortality Rates for Medicare Beneficiaries, 1998-2008. *JAMA*. 2011;306(15):1669-1678. doi:10.1001/jama.2011.1474

11. Akintoye E, Briasoulis A, Egbe A, et al. National Trends in Admission and In-Hospital Mortality of Patients With Heart Failure in the United States (2001-2014). *J Am Heart Assoc*. 2017;6(12). doi:10.1161/JAHA.117.006955

12. Juenger J, Schellberg D, Kraemer S, et al. Health related quality of life in patients with congestive heart failure: comparison with other chronic diseases and relation to functional variables. *Heart*. 2002;87(3):235-241. doi:10.1136/heart.87.3.235

13. Dewan P, Rorth R, Jhund PS, et al. Differential Impact of Heart Failure With Reduced Ejection Fraction on Men and Women. *J Am Coll Cardiol*. 2019;73(1):29-40. doi:10.1016/j.jacc.2018.09.081

14. Bozkurt B, Khalaf S. Heart Failure in Women. *Methodist Debakey Cardiovasc J*. 2017;13(4):216-223. doi:10.14797/mdcj-13-4-216

15. Azad N, Kathiravelu A, Minoosepeher S, Hebert P, Fergusson D. Gender differences in the etiology of heart failure: A systematic review. *J Geriatr Cardiol*. 2011;8(1):15-23. doi:10.3724/SP.J.1263.2011.00015

16. Iorga A, Cunningham CM, Moazeni S, Ruffenach G, Umar S, Eghbali M. The protective role of estrogen and estrogen receptors in cardiovascular disease and the controversial use of estrogen therapy. *Biol Sex Differ*. 2017;8(1):33. doi:10.1186/s13293-017-0152-8

17. Menopause and Heart Disease. American Heart Association website. Accessed August 3, 2020. www.heart.org. https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/menopause-and-heart-disease

18. Pardhe BD, Ghimire S, Shakya J, et al. Elevated Cardiovascular Risks among

Postmenopausal Women: A Community Based Case Control Study from Nepal. *Biochem Res Int*. 2017;2017:3824903. doi:10.1155/2017/3824903

19. Emdin CA, Wong CX, Hsiao AJ, et al. Atrial fibrillation as risk factor for cardiovascular disease and death in women compared with men: systematic review and meta-analysis of cohort studies. *bmj*. 2016;352:h7013.

20. Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women, and minorities in heart failure clinical trials. *Arch Intern Med*. 2002;162(15):1682-1688. doi:10.1001/archinte.162.15.1682

21. Tahhan AS, Vaduganathan M, Greene SJ, et al. Enrollment of Older Patients, Women, and Racial and Ethnic Minorities in Contemporary Heart Failure Clinical Trials: A Systematic Review. *JAMA Cardiol*. 2018;3(10):1011-1019. doi:10.1001/jamacardio.2018.2559

22. Appiah D, Schreiner PJ, Demerath EW, Loehr LR, Chang PP, Folsom AR. Association of Age at Menopause With Incident Heart Failure: A Prospective Cohort Study and Meta-Analysis. *J Am Heart Assoc*. 2016;5(8). doi:10.1161/JAHA.116.003769

23. Bibbins-Domingo K, Lin F, Vittinghoff E, et al. Predictors of heart failure among women with coronary disease. *Circulation*. 2004;110(11):1424-1430. doi:10.1161/01.CIR.0000141726.01302.83

24. Eaton CB, Abdulbaki AM, Margolis KL, et al. Racial and ethnic differences in incident hospitalized heart failure in postmenopausal women: the Women’s Health Initiative. *Circulation*. 2012;126(6):688-696.

25. Ebong IA, Watson KE, Goff Jr DC, et al. Age at menopause and incident heart failure: the Multi-Ethnic Study of Atherosclerosis. *Menopause (New York, NY)*. 2014;21(6):585.

26. Hall PS, Nah G, Howard B V, et al. Reproductive Factors and Incidence of Heart Failure Hospitalization in the Women’s Health Initiative. *J Am Coll Cardiol*. 2017;69(20):2517-2526. doi:10.1016/j.jacc.2017.03.557

27. Rahman I, Åkesson A, Wolk A. Relationship between age at natural menopause and risk of heart failure. *Menopause*. 2015;22(1):12-16.

28. LaMonte MJ, Manson JE, Chomistek AK, et al. Physical Activity and Incidence of Heart Failure in Postmenopausal Women. *JACC Heart Fail*. 2018;6(12):983-995. doi:10.1016/j.jchf.2018.06.020

29. Chen N, Alam AB, Lutsey PL, et al. Polypharmacy, Adverse Outcomes, and Treatment Effectiveness in Patients≥ 75 With Atrial Fibrillation. *J Am Heart Assoc*. 2020;9:e015089.

30. Percent of U.S. Adults 55 and Over with Chronic Conditions. Centers for Disease Control and Prevention (CDC) website. Published November 6, 2015. Accessed March 5, 2020. https://www.cdc.gov/nchs/health_policy/adult_chronic_conditions.htm.

31. Buttorff C, Ruder T, Bauman M. *Multiple Chronic Conditions in the United States*. Rand Santa Monica, CA; 2017.

32. Williams RE, Kalilani L, DiBenedetti DB, et al. Frequency and severity of vasomotor symptoms among peri- and postmenopausal women in the United States. *Climacteric*. 2008;11(1):32-43. doi:10.1080/13697130701744696

33. Thurston RC, Joffe H. Vasomotor symptoms and menopause: findings from the Study of Women’s Health across the Nation. *Obstet Gynecol Clin North Am*. 2011;38(3):489-501. doi:10.1016/j.ogc.2011.05.006

34. Slopien R, Wender-Ozegowska E, Rogowicz-Frontczak A, et al. Menopause and diabetes: EMAS clinical guide. *Maturitas*. 2018;117:6-10. doi:10.1016/j.maturitas.2018.08.009

35. Szmuilowicz ED, Stuenkel CA, Seely EW. Influence of menopause on diabetes and diabetes risk. *Nat Rev Endocrinol*. 2009;5(10):553-558. doi:10.1038/nrendo.2009.166

36. Searles S, Makarewicz JA, Dumas JA. The role of estradiol in schizophrenia diagnosis and symptoms in postmenopausal women. *Schizophr Res*. 2018;196:35-38. doi:10.1016/j.schres.2017.05.024

37. Bromberger JT, Kravitz HM, Chang Y-F, Cyranowski JM, Brown C, Matthews KA. Major depression during and after the menopausal transition: Study of Women's Health Across the Nation (SWAN). *Psychol Med*. 2011;41(9):1879-1888. doi:10.1017/S003329171100016X

38. Raz R. Urinary tract infection in postmenopausal women. *Korean J Urol*. 2011;52(12):801-808. doi:10.4111/kju.2011.52.12.801

39. Brown WJ, Mishra GD, Dobson A. Changes in physical symptoms during the menopause transition. *Int J Behav Med*. 2002;9(1):53-67. doi:10.1207/s15327558ijbm0901_04

40. Dugan SA, Powell LH, Kravitz HM, Everson Rose SA, Karavolos K, Luborsky J. Musculoskeletal pain and menopausal status. *Clin J Pain*. 2006;22(4):325-331. doi:10.1097/01.ajp.0000208249.07949.d5

41. Page RL, O'Bryant CL, Cheng D, et al. Drugs that may cause or exacerbate heart failure: a scientific statement from the American Heart Association. *Circulation*. 2016;134(6):e32--e69.

42. Azimova K, San Juan Z, Mukherjee D. Cardiovascular safety profile of currently available diabetic drugs. *Ochsner J*. 2014;14(4):616-632.

43. McAlister FA, Eurich DT, Majumdar SR, Johnson JA. The risk of heart failure in patients with type 2 diabetes treated with oral agent monotherapy. *Eur J Heart Fail*.

2008;10(7):703-708.

44. Loke YK, Kwok CS, Singh S. Comparative cardiovascular effects of thiazolidinediones: systematic review and meta-analysis of observational studies. *BMJ*. 2011;342:d1309. doi:10.1136/bmj.d1309

45. Lago RM, Singh PP, Nesto RW. Congestive heart failure and cardiovascular death in patients with prediabetes and type 2 diabetes given thiazolidinediones: a meta-analysis of randomised clinical trials. *Lancet (London, England)*. 2007;370(9593):1129-1136. doi:10.1016/S0140-6736(07)61514-1

46. Hernandez A V, Usmani A, Rajamanickam A, Moheet A. Thiazolidinediones and risk of heart failure in patients with or at high risk of type 2 diabetes mellitus. *Am J Cardiovasc Drugs*. 2011;11(2):115-128.

47. Deacon CF, Holst JJ. Dipeptidyl peptidase-4 inhibitors for the treatment of type 2 diabetes: comparison, efficacy and safety. *Expert Opin Pharmacother*. 2013;14(15):2047-2058.

48. Monami M, Dicembrini I, Mannucci E. Dipeptidyl peptidase-4 inhibitors and heart failure: a meta-analysis of randomized clinical trials. *Nutr Metab Cardiovasc Dis*. 2014;24(7):689-697. doi:10.1016/j.numecd.2014.01.017

49. Sills GJ. The mechanisms of action of gabapentin and pregabalin. *Curr Opin Pharmacol*. 2006;6(1):108-113.

50. Erdougan G, Ceyhan D, Güleç S. Possible heart failure associated with pregabalin use: case report. *Agri*. 2011;23(2):80-83.

51. Stefani A, Spadoni F, Giacomini P, Lavaroni F, Bernardi G. The effects of gabapentin on different ligand-and voltage-gated currents in isolated cortical neurons. *Epilepsy Res*.

2001;43(3):239-248.

52. Ho JM-W, Macdonald EM, Luo J, et al. Pregabalin and heart failure: A population-based study. *Pharmacoepidemiol Drug Saf*. 2017;26(9):1087-1092. doi:10.1002/pds.4239

53. Heianza Y, Zheng Y, Ma W, et al. Duration and life-stage of antibiotic use and risk of cardiovascular events in women. *Eur Heart J*. 2019;40(47):3838-3845. doi:10.1093/eurheartj/ehz231

54. Durack J, Lynch S V. The gut microbiome: Relationships with disease and opportunities for therapy. *J Exp Med*. 2019;216(1):20-40. doi:10.1084/jem.20180448

55. Jin M, Qian Z, Yin J, Xu W, Zhou X. The role of intestinal microbiota in cardiovascular disease. *J Cell Mol Med*. 2019;23(4):2343-2350. doi:10.1111/jcmm.14195

56. Jie Z, Xia H, Zhong S-L, et al. The gut microbiome in atherosclerotic cardiovascular disease. *Nat Commun*. 2017;8(1):845. doi:10.1038/s41467-017-00900-1

57. Emoto T, Yamashita T, Sasaki N, et al. Analysis of Gut Microbiota in Coronary Artery Disease Patients: a Possible Link between Gut Microbiota and Coronary Artery Disease. *J Atheroscler Thromb*. 2016;23(8):908-921. doi:10.5551/jat.32672

58. Yoshida N, Yamashita T, Hirata K. Gut microbiome and cardiovascular diseases. *Diseases*. 2018;6(3):56.

59. Etminan M, Sodhi M, Ganjizadeh-Zavareh S, Carleton B, Kezouh A, Brophy JM. Oral Fluoroquinolones and Risk of Mitral and Aortic Regurgitation. *J Am Coll Cardiol*. 2019;74(11):1444 LP - 1450. doi:10.1016/j.jacc.2019.07.035

60. Aschenbrenner DS. New warning for fluoroquinolone antibiotics. *AJN Am J Nurs*. 2019;119(4):20.

61. Weintraub NL, Collins SP, Pang PS, et al. Acute heart failure syndromes: emergency

department presentation, treatment, and disposition: current approaches and future aims: a scientific statement from the American Heart Association. *Circulation*. 2010;122(19):1975-1996.

62. Hasegawa K, Tsugawa Y, Camargo Jr CA, Brown DFM. Frequent utilization of the emergency department for acute heart failure syndrome: a population-based study. *Circ Cardiovasc Qual Outcomes*. 2014;7(5):735-742.

63. Blecker S, Ladapo JA, Doran KM, Goldfeld KS, Katz S. Emergency department visits for heart failure and subsequent hospitalization or observation unit admission. *Am Heart J*. 2014;168(6):901-908.

64. *Ready, Risk, Reward: Improving Care for Patients with Chronic Conditions*.; 2019.

65. Zhou RA, Baicker K, Taubman S, Finkelstein AN. The uninsured do not use the emergency department more—they use other care less. *Health Aff*. 2017;36(12):2115-2122.

66. Mann C. Targeting Medicaid super-utilizers to decrease costs and improve quality. *Centers Medicare Medicaid Serv*. 2013.

67. Billings J, Parikh N, Mijanovich T. Emergency department use in New York City: a survey of Bronx patients. *Issue Brief (Commonw Fund)*. 2000;(435):1-5.

68. Tsai C-L, Griswold SK, Clark S, Camargo CA. Factors associated with frequency of emergency department visits for chronic obstructive pulmonary disease exacerbation. *J Gen Intern Med*. 2007;22(6):799-804.

69. Krieg C, Hudon C, Chouinard M-C, Dufour I. Individual predictors of frequent emergency department use: a scoping review. *BMC Health Serv Res*. 2016;16(1):594.

70. Egede LE. Patterns and correlates of emergency department use by individuals with

diabetes. *Diabetes Care*. 2004;27(7):1748-1750.

71. Agarwal P, Bias TK, Madhavan S, Sambamoorthi N, Frisbee S, Sambamoorthi U. Factors associated with emergency department visits: A multistate analysis of adult fee-for-service Medicaid beneficiaries. *Heal Serv Res Manag Epidemiol*. 2016;3:2333392816648549.

72. Alhussain K, Meraya AM, Sambamoorthi U. Serious psychological distress and emergency room use among adults with multimorbidity in the United States. *Psychiatry J*. 2017;2017.

73. Niedzwiecki MJ, Sharma PJ, Kanzaria HK, McConville S, Hsia RY. Factors associated with emergency department use by patients with and without mental health diagnoses. *JAMA Netw open*. 2018;1(6):e183528--e183528.

74. Dufour I, Chouinard M-C, Dubuc N, Beaudin J, Lafontaine S, Hudon C. Factors associated with frequent use of emergency-department services in a geriatric population: a systematic review. *BMC Geriatr*. 2019;19(1):185.

75. Fan L, Shah MN, Veazie PJ, Friedman B. Factors associated with emergency department use among the rural elderly. *J Rural Heal*. 2011;27(1):39-49.

76. Liu CW, Einstadter D, Cebul RD. Care fragmentation and emergency department use among complex patients with diabetes. *Am J Manag Care*. 2010;16(6):413-420.

77. Sarrel P, Portman D, Lefebvre P, et al. Incremental direct and indirect costs of untreated vasomotor symptoms. *Menopause*. 2015;22(3):260-266.

78. Breiman L, others. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat Sci*. 2001;16(3):199-231.

79. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges.

Eur Heart J. 2016;38(23):1805-1814. doi:10.1093/eurheartj/ehw302

80. Couronné R, Probst P, Boulesteix A-L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*. 2018;19(1):270.

81. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One*. 2018;13(8).

82. Boulesteix A-L, Schmid M. Machine learning versus statistical modeling. *Biometrical J*. 2014;56(4):588-593.

83. Sakr S, Elshawi R, Ahmed A, et al. Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project. *PLoS One*. 2018;13(4).

84. Andrews PJD, Sleeman DH, Statham PFX, et al. Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. *J Neurosurg*. 2002;97(2):326-336.

85. Jain S. Applications of Logistic Model to Medical Research. *Biometrical J*. 1987;29(3):369-374.

86. Kruppa J, Liu Y, Diener H-C, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications. *Biometrical J*. 2014;56(4):564-583.

87. Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. *Biometrical J*. 2014;56(4):601-606.

88. Crown WH. Potential application of machine learning in health outcomes research and

some statistical cautions. *Value Heal*. 2015;18(2):137-140.

89. Marmot M, Wilkinson R. *Social Determinants of Health*. OUP Oxford; 2005.

90. PubMed Overview. PubMed website. Accessed March 4, 2020. https://pubmed.ncbi.nlm.nih.gov/about/

91. Clinformatics® Data Mart. Optum websit. Accessed March 5, 2020. https://www.optum.com/content/dam/optum/resources/productSheets/Clinformatics_for_Data_Mart.pdf

CHAPTER 2

2. Identifying Knowledge Gaps in Heart Failure Research among Women Using Unsupervised Machine Learning Methods

2.1 Abstract

Objective: To identify knowledge gaps in heart failure (HF) research among women, especially postmenopausal women.

Materials & Methods: We retrieved HF articles from PubMed. Natural language processing and text mining techniques were used to screen relevant articles and identify study objective(s) from abstracts. After text pre-processing, we performed topic modeling with non-negative matrix factorization to cluster articles based on the primary topic. Clusters were independently validated and labeled by three investigators familiar with HF research.

Results: Our model yielded 15 topic clusters from articles on HF among women. The smallest cluster was about atrial fibrillation. From articles specific to postmenopausal women, 5 clusters were identified. The smallest cluster was about stress-induced cardiomyopathy.

Conclusion: Topic modeling can help identify understudied areas in medical research.

2.2 Introduction

Heart failure (HF) affects at least 26 million people worldwide, and its prevalence has been increasing over the past decades[1]. For example, HF prevalence is expected to rise from 2.42% in 2012 to 2.97% in 2030 in the United States (US)[2]. The grown prevalence of HF, along with its high mortality and morbidity[3] as well as poor health-related quality of life (HRQoL)[4] make HF a major global health problem. HF mortality has been assessed in several countries[1]. In a registry-based study enrolling 12,440 patients with acute or chronic HF from 21 European and/or Mediterranean countries, the 1-year mortality rates varied across countries; it ranged from 21.6% to 36.5% in patients with acute HF, and from 6.9% to 15.6% in those with chronic HF[5]. In the US, the 1-year mortality in patients with HF ranged from 35.1% to 37.5%[6]. Even if they survive, patients with HF have poor HRQoL, both physical and mental components, compared to the general population[4]. In addition, HF has a high economic burden. Healthcare spending on HF constitutes 1-2% of the global healthcare budget, mainly due to hospitalization costs[7]. Cost estimates varied from a country to another. For instance, total annual costs per patient with HF ranged from $868 for South Korea to $25,532 for Germany[7]. Regardless of the differences across countries, in general, HF has a significant health and economic burden worldwide.

With that being said, there is a need to study HF. A major consideration that should be taken into account in future studies is the sex differences in HF burden and risk factors. For example, women with HF have poorer HRQoL compared to their men counterparts[8]. Furthermore, women tend to develop HF at an older age than men[3,9], which can be explained by the female sex hormone, estrogen. Estrogen has anti-atherosclerotic and anti-inflammatory properties, which positively affects the inner layer of artery wall[10,11]. However, estrogen levels decrease after menopause. The decline in the level of endogenous estrogen increases the risk of

HF in postmenopausal women[12,13]. In terms of risk factors for HF, hypertension is more common in women, whereas myocardial infarction is more prevalent in men[9].

Despite these differences between women and men with HF, women are underrepresented in clinical trials for HF[14,15]. A recent systematic review examined the enrollment of women and other minorities in 118 HF clinical trials[15]. This study revealed that women represented only 27% of participants in clinical trials for HF, and women's participation has not significantly changed over time.

With such underrepresentation of women in HF clinical trials, significant knowledge gaps in HF research among women may exist. These knowledge gaps need to be identified and addressed. To date, no study has reviewed all published HF research among women, specifically among postmenopausal women. Systematic reviews and meta-analysis focus on a single topic (example: mortality, treatment, biological markers)[16,17]. However, conducting a broad search of "heart failure" and women in the PubMed database yields over 100,000 articles. Manually reading all these articles and summarizing the topics will not be feasible.

With the wide-spread digital transformation and ability of processing and understanding of the text by machine through natural language processing (NLP), it is now possible to use digital technology to cluster all HF research among women based on their primary objectives. Such approach cannot only save the researchers' time by substituting computer time[18] but also discovers knowledge gaps in HF research among women. Therefore, the objective of the current study is to identify knowledge gaps in HF research among women, especially postmenopausal women using unsupervised machine learning methods and articles published in the PubMed database.

2.3 Methods

Data source, search strategies, and procedures

Our data source was PubMed, a free database comprises more than 30 million citations and abstracts of biomedical literature from MEDLINE, life science journals, and online books[19]. We only searched PubMed (i.e., no other databases) because we wanted to assess the feasibility of using unsupervised machine learning methods for identifying knowledge gaps. We identified articles on HF research in women from the inception (1959) until 3 December 2019. We conducted two search strategies: (1) broad, where we focused on all women, and (2) specific, where the focus was on postmenopausal women. For search #1, we used the following keywords and medical subject headings (MeSH): ("heart failure" OR "congestive heart failure" OR "cardiac failure" OR "heart failure therapy" OR "ejection fraction"). For search #2, we used the following strategy: ("heart failure" OR "congestive heart failure" OR "cardiac failure" OR "heart failure therapy" OR "ejection fraction" AND "postmenopause" OR "menopause"). We included "ejection fraction" as one of the search terms because ejection fraction plays a key role in HF diagnosis and outcomes[20]. For both searches, we used PubMed search filters on sex (female), species (humans), and text availability (abstract) to enhance our search strategies. For the purpose of this study, no restrictions (e.g., study design or country) were used.

Procedures

Articles retrieved from the PubMed searches were stored in Comma-separated Values files. We removed duplicates based on article titles. We identified relevant articles based on "study objectives" because the objectives of an article can provide a clear and exact intent of the study. We only included studies having at least one of the HF terms (i.e., "heart failure" and "cardiac failure") in their objectives.

As our main interest was in summarizing the HF research in women and postmenopausal women, we used “topic models”, a type of statistical model for identifying a set of “topics” that best describes a given document (in this case, given PubMed article). Topic modeling is an unsupervised machine learning method that automatically clusters a set of documents according to “semantic structures” or topics that are similar. It has to be noted that topic modeling can group words within the same context as well as distinguish the use of the same words in a different context. Furthermore, topic modeling does not require pre-existing knowledge of the categories of the articles[18]. Topic modeling has been applied on different medical datasets including lung cancer, breast cancer, and Salmonella PFGE genotyping datasets[21]. Following the framework for smart literature review of big data, we used three key steps: pre-processing, topic modeling, and post-processing of outcomes[18]. All procedures and modeling were conducted with Python 3.7.

Text pre-processing

Text pre-processing is a crucial step in the process of building any model. Typically, text pre-processing helps machine learning algorithms by removing or filtering less useful parts of the text through various methods such as punctuation and stop word removal. In the current study, we restricted NLP and text mining techniques to the objective(s) of the study rather than the full text or abstract of the article. The reason behind this is that a study objective provides specific information about the study, while the full text and abstract have information that may not be directly related to the primary topic of the study (e.g., literature review and statistical analysis). We pre-processed the text using the Natural Language Toolkit (NLTK), one of the most powerful platforms for processing human language in Python software. We first removed common words (e.g., a, is, the, and) that carry less important meaning (stop words) than keywords. Examples of such words are “introduction”, “background”, “methods”, “results”, and “conclusions” that are used in almost all structured abstracts. After removing unnecessary words, we conducted two more steps (i.e., tokenization and lemmatization). Tokenization is the process of splitting text into a list of tokens, and lemmatization is a morphological analysis of the words (e.g., using the lemma “study” for studies, study, studied, studying).

Topic modeling with non-negative matrix factorization

As topic modeling involves grouping similar word patterns to identify topics, there are several algorithms such as Non-Negative Matrix Factorization (NMF) based on linear algebra are available. We selected NMF to identify topics and classify the documents according to these topics at the same time. NMF computes term frequency-inverse document frequency (TF-IDF), a weighting scheme that assigns each word in our dataset (i.e., PubMed abstracts) a weight. The higher the weight, the more important the word is. To compute the TF-IDF weighting, we used TfidfVectorizer with n-gram range from 1 to 2 from the scikit-learn Python module.

We performed topic modeling on all studies of women with HF (search#1) and studies specific to postmenopausal women (search#2). To identify the optimal number of clusters, we ran the algorithm with a different number of topics (n); for example, we specified the value of n as 5, 10, 15, 20, and 25. Then, we manually evaluated the outputs from all models and selected the most interpretable model. All analyses were performed using Python 3.7.

Post-processing

Validation of topic modeling: human intelligence

During the post-processing, we reviewed the clusters identified to ensure that they are interpretable. Moreover, we used an expert evaluation to validate the topic models. Clusters yielded from our model were independently labeled and validated by three investigators familiar with HF research. In case of a disagreement on the cluster label, discussion among the investigators was a sensible first step. Disagreements among investigators were resolved by consensus. If a disagreement could not be resolved, investigators reviewed that cluster in depth; they randomly reviewed the titles and abstracts of 40 articles within that cluster. Finally, we reported the frequency and percentage of agreements and disagreements.

2.4 Results

Study retrieval and selection

Automated extraction using search strategy #1 yielded 69,558 articles related to HF in women. Of these, 6 articles with no abstract and 53 duplicates were removed. The remaining, 69,499 articles, were electronically screened for relevance (i.e. study objective(s) must have at least one of the HF terms). This process yielded 32,946 eligible HF articles for topic modeling.

Using a separate search strategy #2, where the focus on postmenopausal women, there were 41,519 articles with abstract after 150 duplicates were removed. After electronically screening, 41,442 articles were excluded because they were not relevant based on the study objective(s) (i.e. absence of all HF terms in the study objective). A final list of 77 articles were included in the topic modeling. Flow charts illustrating each step of this process are shown in Figure 1.

Topic clusters

A description of the topic clusters is shown in Table 1. For search strategy #1, the topic model with 15 topic clusters was selected because it was the most interpretable model for HF articles in women. In terms of size, the largest topic cluster consisted of 4,578 articles (%13.9), whereas the smallest topic cluster consisted of 808 articles (%2.5) (Figure 2). The most studied topic in HF among women was epidemiology and disease burden of HF. For search strategy #2, the most interpretable topic model yielded 5 clusters out of 77 articles on HF in postmenopausal women. The largest cluster size was 34 articles (44.2%) while the smallest cluster size was 6 articles (7.8%) (Figure 3). The most studied topic in postmenopausal women was cardiovascular risk. (e.g., effects of lipid accumulation product and blood pressure on cardiovascular risk in postmenopausal women).

Understudied research topics in the literature of HF among women

Based on the cluster size, the three most understudied topics are (1) atrial fibrillation, (2) systolic and diastolic dysfunction, and (3) left ventricular ejection fraction phenotypes. The knowledge gaps are even greater in the literature of HF among postmenopausal women. Only 6 articles studied stress-induced cardiomyopathy. The effect of breast cancer and chemotherapy on HF was discussed in 12 articles. Also, the incidence of HF in postmenopausal women was studied in 12 articles.

Cluster validation and labeling

Topic clusters were independently validated and labeled by the first, second, and seventh authors. The percentage of agreement among authors on topic labels is presented in Table 2. For search strategy #1, the agreement percentage was 80%, which means authors agreed on 12 out of 15 topic labels. Regarding the other three clusters, disagreements were resolved by reviewing those clusters in depth. For search strategy #2, there were no disagreements on the topic labels.

2.5 Discussion

The main objective of this study was to explore knowledge gaps in HF research among all women and postmenopausal women. We achieved this objective by using topic modeling, an unsupervised machine learning method. Our approach saved researchers’ time once the program was developed. Our program took only 1 minute and 4 seconds to cluster 32,946 articles into 15 topics. This hybrid approach was more comprehensive and less time-consuming than the expert-based manual literature review method. For example, a study by Myers et al. was conducted to assess the progress of CVD research output between 2002 and 2011 using the expert-based manual literature review method[22]. In that study, a physician read the abstracts and decided whether a study was relevant. Although there were 47,897 articles related to CVD in 2002 and 54,488 articles in 2011, only 3,000 articles randomly selected each year were reviewed. This is mainly because it was difficult to manually review more than 100,000 abstracts.

Our current study has revealed that atrial fibrillation is the most understudied area in the literature of HF among women. Prior research in this area has discussed the epidemiology of atrial fibrillation, role of natriuretic peptide, and risk of stroke in patients with atrial fibrillation and heart failure. Nevertheless, this research area should be further explored for several reasons. First, there is a positive association between AF and HF [23,24], and this association can be explained by shared risk factors and pathophysiology[25]. Thus, these two diseases can be regularly encountered concomitantly in clinical practice. Patients with concomitant HF and AF may have even worse symptoms and poorer prognosis, which means they may respond to treatment differently than those with HF or AF alone[24,25]. Furthermore, the co-occurrence of HF and AF may increase the risk of HF hospitalization and all-cause mortality, as previous studies shown[26,27]. With that being said, future research focusing on the comorbidity of HF and AF in women is needed. This can improve the health outcomes of women affected by these two conditions and the cost-effectiveness of their care.

Another important finding was that the volume of research on HF in postmenopausal women is small. In this study, we only identified 77 articles on HF in postmenopausal women compared to 32,946 in women in general. Based on the content of those articles, the most understudied topic is stress-induced cardiomyopathy. This may be because this condition is rare. In the US, stress-induced cardiomyopathy was diagnosed in about 0.02% of all nationwide hospitalizations[28]. Of those, 90.6% were women. It is well-known that this condition is more common in women than men[29–33]. Therefore, future studies should investigate this topic and address knowledge gaps in this area.

Another major understudied area is the incidence of HF in postmenopausal women. For instance, few studies examined risk factors for the incidence of HF in postmenopausal women. There is a critical need to identify factors associated with HF incidence in this population and address the modifiable risk factors. There may be emerging risk factors such as medication use. Medications that may increase the risk of HF should be identified. For example, one of the clusters yielded from our model was related to cardio-oncology in advanced breast cancer.

Identification of research gaps is the first step towards reducing HF risk and improving health outcomes in women. Our findings provide an overview of HF research among women. Such information can help researchers and funding agencies to prioritize and address research gaps. Using data from this study along with the insights of the professional community may contribute to the development of a research roadmap for HF in women.

Potential limitations and strengths of this study should be noted. First, no evaluation metrics were used to assess the accuracy of clusters yielded from unsupervised machine learning. However, this limitation was addressed by independently validating and labeling clusters yielded from our model by three investigators familiar with HF research. Second, we were not able to extract the study objective(s) from unstructured abstracts. In that case, we analyzed the full abstract. Finally, we only searched one database (i.e., PubMed) to retrieve HF articles, which might impact on the number of articles included in this study. Despite these limitations, this study had several strengths. To our knowledge, this was the first study to use big data (PubMed) and unsupervised machine learning methods to identify research topics in the literature of HF among women. In addition, we used NLP and text mining techniques to screen and identify relevant articles and extract the objective(s) of each study from PubMed abstracts.

2.6 Conclusion

The present study was able to identify gaps in the literature of HF among women, particularly postmenopausal women, using unsupervised machine learning methods. This approach is promising and effective for the discovery of knowledge gaps in medical research. Once unsupervised machine learning procedures are established, clustering a large number of research articles can be performed within a short time. However, human intelligence is required to interpret and validate the results.

Table 1: Description of the topic clusters

Clusters Yielded from Search #1 (n=15)

Topic/Cluster Label	**Key Words - Examples**	**No. of Articles**
Epidemiology/disease burden of HF	“prevalence”, “hf risk”, “factor”, “obesity”, “incident”, “chronic hf”, “acute hf”, “systolic hf”, “advanced hf”, “hf preserved”, “hf outcome”, “hf hospitalization”, “outpatient”, “mortality”, “population”	4,578
Heart procedures - mainly valvular	“surgery”, “operation”, “valve replacement”, “mitral regurgitation”, “aortic valve”, “tricuspid”, “bypass”, “coronary artery”, “echocardiography”, “dilated cardiomyopathy”, “stenosis”, “treatment”, “cardiac failure”, “congestive heart”, “severe”, “underwent”, “complication”	4,515
Clinical markers in chronic HF*	'inflammation', 'tnfalpha', 'endothelial', 'cytokine', 'cell', 'marker', 'activation', 'oxidative', 'sympathetic', 'muscle', 'serum', 'breathing', 'sdb', 'sleep', 'severity', 'renal', 'copd', 'anemia', 'prognosis', 'elderly', 'congestive heart', 'chronic heart'	3,243
Myocardial infarction	“myocardial infarction”, “acute myocardial infarction”, “coronary artery”, “cardiac index”, “incidence”, “age”, “diabetes”, “stroke”, “outcome”, “hospitalization”, “survival”, “all-cause”, “sudden”, “death”, “mortality”	2,990
Health-related quality of life (HRQoL)	“quality of life”, “health-related quality”, “depressive symptom”, “depression”, “physical”, “symptom”, “status”, ”selfcare”, “questionnaire”, “program”, “intervention”, “education”, “social”, “service”, “caregiver”	2,909
Hemodynamic effects*	“hemodynamic”, “pulmonary artery”, “pulmonary capillary”, “systemic vascular”, “vascular resistance”, “arterial pressure”, “heart rate”, “wedge pressure”, “blood pressure”, “cardiac index”, “stroke”	2,562
Pharmacotherapy	“ace inhibitor”, “beta-blockers”, “diuretic”, “receptor blocker”, “arb”, “captopril”, “digoxin”, “enalapril”, “angiotensin receptor”, “antagonist”, “inhibition”, “dose”, “mg”, “placebo”, “drug”, “class”	1,713
Cardiac biomarkers	“brain natriuretic”, “bnp level”, “anp”, “b-type natriuretic”, “nt-pro-bnp”, “natriuretic peptide”, “‘n-terminal pro-brain”, “serum”, “plasma”, “pgml”, “marker”, “concentration”, “measurement”, “prognostic value”, “diagnosis”	1,654

Acute decompensated heart failure	“acute decompensated”, “adhf”, “worsening renal”, “wrf”, “renal dysfunction”, “aki”, “emergency department”, “inhospital”, “admission”, “hospitalized”, “nesiritide”, “diuretic”	1,545
Exercise	“aerobic”, “cardiopulmonary exercise”, “peak exercise”, “training”, “ventilation”, “exercise test”, “exercise tolerance”, “functional capacity”, “oxygen uptake”, “oxygen consumption”, “peak vo”, “vevco”	1,469
Cardiac resynchronization therapy	“cardiac resynchronization”, “crt”, “crt-d”, “icd”, “defibrillator”, “implantation”, “dyssynchrony”, “pacing”, “bundle branch”, “branch block”, “lbbb, 'delay', 'remodeling', 'biventricular', 'lead', 'qrs duration'	1,295
Left ventricular assist device & heart transplantation	“lvad implantation”, “pump”, “bridge”, “mechanical circulatory”, “assist device”, “heartmate”, “cardiac transplantation”, “recovery”, “experience”, “survival”, “advanced heart”, “end-stage heart”	1,255
Left ventricular ejection fraction phenotypes	“hfpef”, “hfmref”, “hfref”, “reduced ef”, “preserved ef”, “midrange”, “pathophysiology”, “hypertension”, “prognostic”, “outcome”, “ejection fraction”	1,209
Systolic & diastolic dysfunction*	'systolic dysfunction', 'diastolic dysfunction', 'lv systolic', 'lv dysfunction', 'lv diastolic', 'velocity', 'right ventricular', 'myocardial', 'doppler', 'pacing', 'filling', 'volume', 'echocardiography', 'diastolic function', 'ejection fraction'	1,201
Atrial fibrillation	“atrial fibrillation”, “af”, “af sinus”, “paroxysmal”, “sinus rhythm”, “permanent atrial”, “persistent atrial”, “af hf”, “incidence”, “new-onset af”, “rate control”, “cardioversion”, “pacing”, “catheter ablation”, “digoxin”	808
Clusters Yielded from Search #2 (n=5)		
Cardiovascular disease risk	“cardiovascular risk”, “risk factor”, “myocardial infarction”, “coronary artery”, “sex”, “estrogen”, “hrt”, “postmenopausal woman”, “blood pressure”, “hypertension”, “stroke”, “obesity”, “diabetes”, “morbidity”, “mortality”, “death”	34
Role of female sex hormone in HF	“sex hormone”, “female”, “estrogen”, “menopause”, “age”, “protective”, “endothelial”, “risk marker”, “lvdd”, “diastolic dysfunction”, “preserved ejection”, “ejection fraction”, “hfpef”, “microvascular”, “role”, “mechanism”	13
Effect of breast cancer and chemotherapy on HF	“breast cancer”, “advanced breast”, “chemotherapy”, “cyclophosphamide”, “tamoxifen”, “methotrexate”, “doxorubicin”, “mitoxantrone”, “cmf”, “combination”, “course”, “regimen”,	12

	“drug”, “agent”, “dose”, “toxicity”, “progression”, “remission”, “alopecia”, “response”	
HF incidence	“hf incidence”, “incident hf”, “incident heart”, “risk incident”, “risk heart”, “age”, “early'”, “age menopause”, “effect cardiac”, “cvd”, “hf postmenopausal”, “sex hormone”, “hrt”, “deficit”, “vitamin”, “supplementation”	12
Stress-induced cardiomyopathy	“stress”, “takotsubo syndrome”, “takotsubo cardiomyopathy”, “tt”, “acute”, “syndrome”, “condition”, “reversible”, “rare”, “segment”, “pathophysiology”, “coronary artery”, “left ventricle”, “activation”, “diagnosis”, “imaging”, “admitted”, “morbidity”, “mortality”	6

Note: VO_2 is the rate of oxygen consumption measured during incremental exercise, and vevco refers to minute ventilation-to-carbon dioxide output (VE/VCO2). * indicates a cluster that was reviewed in depth.
Abbreviations: hf: heart failure; tnfalpha: tumor necrosis factor alpha; sdb: sleep disordered breathing; copd: chronic obstructive pulmonary disease; arb: angiotensin II receptor blocker; mg: milligram; bnp: brain or B-type natriuretic peptide; anp: atrial natriuretic peptide; adhf: acute decompensated heart failure; wrf: worsening renal function; aki: acute kidney injury; crt: cardiac resynchronization therapy; crt-d: cardiac resynchronization therapy defibrillator; icd: implantable cardioverter defibrillator; lbbb: left bundle branch block; lvad: left ventricular assist device; lv: left ventricular; hfpef: heart failure with preserved ejection fraction; hfmref: heart failure with mid-range ejection fraction; hfref: heart failure with reduced ejection fraction; ef: ejection fraction; af: atrial fibrillation; hrt: hormone replacement therapy; lvdd: left ventricular diastolic dysfunction; cmf: cyclophosphamide, methotrexate, fluorouracil; cvd: cardiovascular disease; tt: takotsubo.

Table 2: Percentage of agreement and disagreement among authors on topic labels

Search strategy	No. of topic clusters	Agreements n, (%)	Disagreements n, (%)
Search #1	15	12 (80%)	3 (20%)
Search #2	5	5 (100%)	0 (0%)

Note: Disagreements were on the following topic labels: systolic & diastolic dysfunction, clinical markers in chronic HF, and hemodynamic effects.

Figure 1.A: Flow chart for the selection of studies

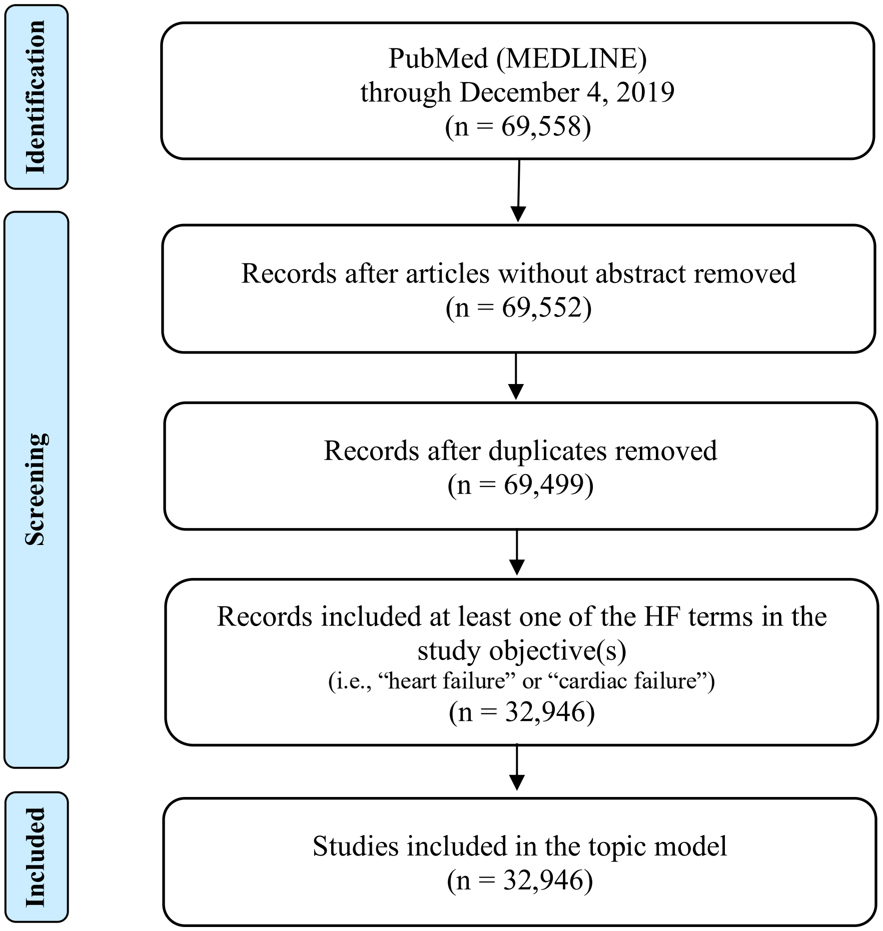

Identification
Screening
Included
PubMed (MEDLINE)
through December 4, 2019
(n = 69,558)
Records after articles without abstract removed
(n = 69,552)
Records after duplicates removed
(n = 69,499)
Records included at least one of the HF terms in the study objective(s)
(i.e., "heart failure" or "cardiac failure")
(n = 32,946)
Studies included in the topic model
(n = 32,946)

Figure 1.B: Flow chart for the selection of studies

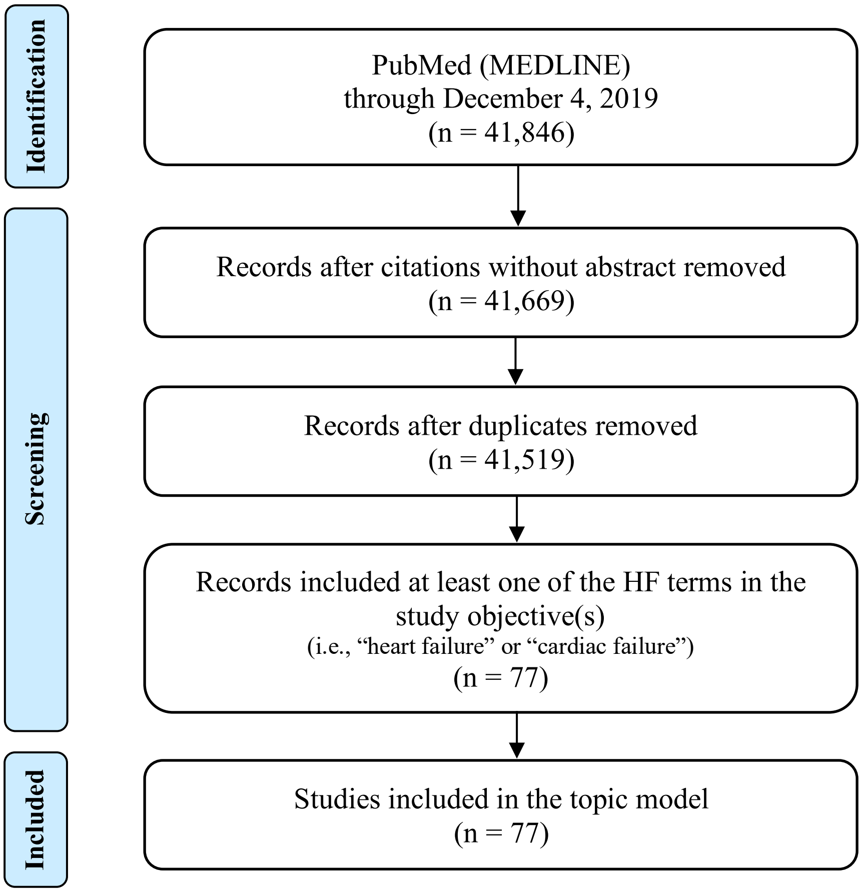

Identification
Screening
Included
PubMed (MEDLINE)
through December 4, 2019
(n = 41,846)
Records after citations without abstract removed
(n = 41,669)
Records after duplicates removed
(n = 41,519)
Records included at least one of the HF terms in the study objective(s)
(i.e., "heart failure" or "cardiac failure")
(n = 77)
Studies included in the topic model
(n = 77)

Figure 2. Distribution of HF article clusters yielded from search strategy # 1 based on main topics

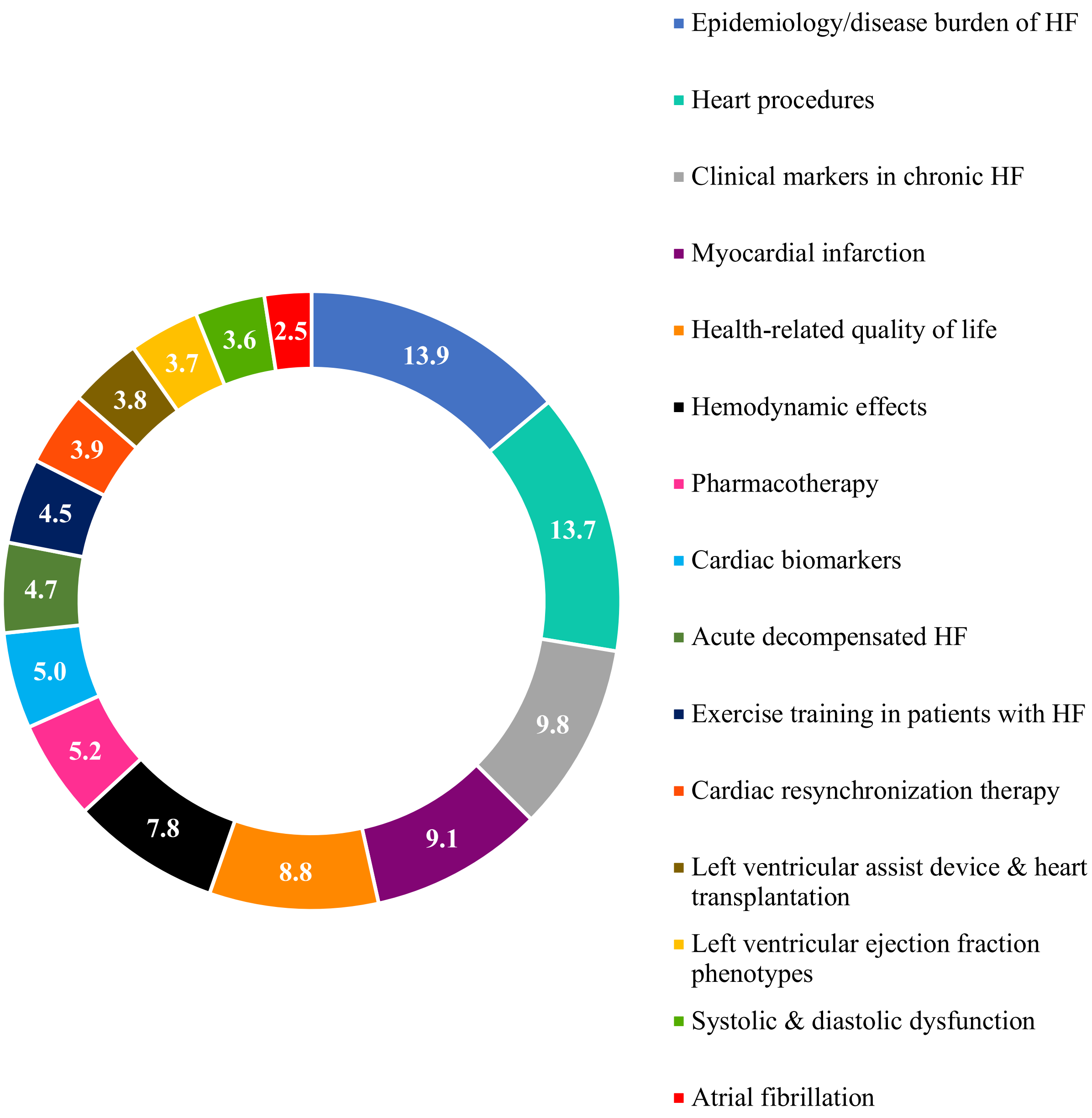
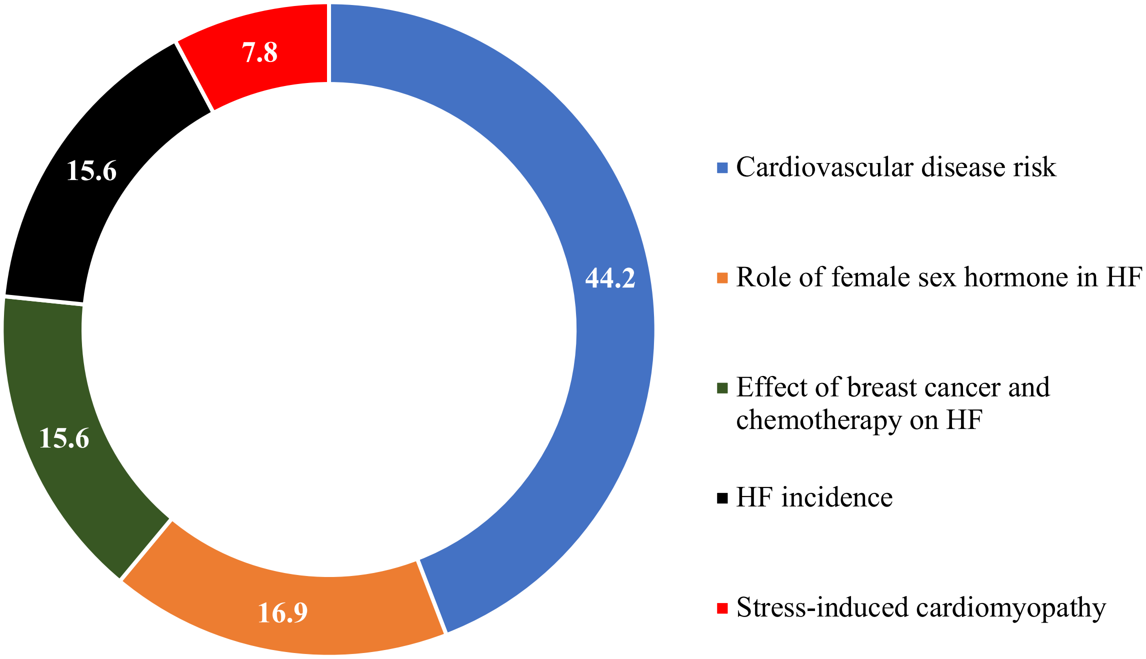


Figure 3. Distribution of HF article clusters yielded from search strategy # 2 based on main topics



2.7 References

1. Savarese G, Lund LH. Global Public Health Burden of Heart Failure. *Card Fail Rev*. 2017;3(1):7-11. doi:10.15420/cfr.2016:25:2

2. Heidenreich PA, Albert NM, Allen LA, et al. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. *Circ Heart Fail*. 2013;6(3):606-619. doi:10.1161/HHF.0b013e318291329a

3. Benjamin EJ, Muntner P, Alonso A, et al. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation*. 2019;139(10):e56-e528. doi:10.1161/CIR.0000000000000659

4. Juenger J, Schellberg D, Kraemer S, et al. Health related quality of life in patients with congestive heart failure: comparison with other chronic diseases and relation to functional variables. *Heart*. 2002;87(3):235-241. doi:10.1136/heart.87.3.235

5. Crespo-Leiro MG, Anker SD, Maggioni AP, et al. European Society of Cardiology Heart Failure Long-Term Registry (ESC-HF-LT): 1-year follow-up outcomes and differences across regions. *Eur J Heart Fail*. 2016;18(6):613-625.

6. Cheng RK, Cox M, Neely ML, et al. Outcomes in patients with heart failure with preserved, borderline, and reduced ejection fraction in the Medicare population. *Am Heart J*. 2014;168(5):721-730.

7. Lesyuk W, Kriza C, Kolominsky-Rabas P. Cost-of-illness studies in heart failure: a systematic review 2004--2016. *BMC Cardiovasc Disord*. 2018;18(1):74.

8. Dewan P, Rorth R, Jhund PS, et al. Differential Impact of Heart Failure With Reduced Ejection Fraction on Men and Women. *J Am Coll Cardiol*. 2019;73(1):29-40. doi:10.1016/j.jacc.2018.09.081

9. Azad N, Kathiravelu A, Minoosepeher S, Hebert P, Fergusson D. Gender differences in the etiology of heart failure: A systematic review. *J Geriatr Cardiol*. 2011;8(1):15-23. doi:10.3724/SP.J.1263.2011.00015

10. Williams JK, Adams MR, Klopfenstein HS. Estrogen modulates responses of atherosclerotic coronary arteries. *Circulation*. 1990;81(5):1680-1687. doi:10.1161/01.cir.81.5.1680

11. Nofer J-R. Estrogens and atherosclerosis: insights from animal models and cell systems. *J Mol Endocrinol*. 2012;48(2):R13-29. doi:10.1530/JME-11-0145

12. Menopause and Heart Disease. American Heart Association website. Accessed August 3, 2020. www.heart.org. https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/menopause-and-heart-disease

13. Pardhe BD, Ghimire S, Shakya J, et al. Elevated Cardiovascular Risks among Postmenopausal Women: A Community Based Case Control Study from Nepal. *Biochem Res Int*. 2017;2017:3824903. doi:10.1155/2017/3824903

14. Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women, and minorities in heart failure clinical trials. *Arch Intern Med*. 2002;162(15):1682-1688. doi:10.1001/archinte.162.15.1682

15. Tahhan AS, Vaduganathan M, Greene SJ, et al. Enrollment of Older Patients, Women, and Racial and Ethnic Minorities in Contemporary Heart Failure Clinical Trials: A Systematic Review. *JAMA Cardiol*. 2018;3(10):1011-1019. doi:10.1001/jamacardio.2018.2559

16. Gwadry-Sridhar FH, Flintoft V, Lee DS, Lee H, Guyatt GH. A systematic review and meta-analysis of studies comparing readmission rates and mortality rates in patients with

heart failure. *Arch Intern Med*. 2004;164(21):2315-2320.

17. Pufulete M, Maishman R, Dabner L, et al. B-type natriuretic peptide-guided therapy for heart failure (HF): a systematic review and meta-analysis of individual participant data (IPD) and aggregate data. *Syst Rev*. 2018;7(1):112.

18. Asmussen CB, Møller C. Smart literature review: a practical topic modelling approach to exploratory literature. 2019.

19. PubMed Overview. PubMed website. Accessed March 4, 2020. https://pubmed.ncbi.nlm.nih.gov/about/

20. Solomon SD, Anavekar N, Skali H, et al. Influence of Ejection Fraction on Cardiovascular Outcomes in a Broad Spectrum of Heart Failure Patients. *Circulation*. 2005;112(24):3738-3744. doi:10.1161/CIRCULATIONAHA.105.561423

21. Zhao W, Zou W, Chen JJ. Topic modeling for cluster analysis of large biological and medical datasets. In: *BMC Bioinformatics*. Vol 15. ; 2014:S11.

22. Myers L, Mendis S. Cardiovascular disease research output in WHO priority areas between 2002 and 2011. *J Epidemiol Glob Health*. 2014;4(1):23-28.

23. Anter E, Jessup M, Callans DJ. Atrial fibrillation and heart failure: treatment considerations for a dual epidemic. *Circulation*. 2009;119(18):2516-2525.

24. Wang TJ, Larson MG, Levy D, et al. Temporal relations of atrial fibrillation and congestive heart failure and their joint influence on mortality: the Framingham Heart Study. *Circulation*. 2003;107(23):2920-2925.

25. Kotecha D, Piccini JP. Atrial fibrillation in heart failure: what should we do? *Eur Heart J*. 2015;36(46):3250-3257. doi:10.1093/eurheartj/ehv513

26. Chamberlain AM, Redfield MM, Alonso A, Weston SA, Roger VL. Atrial fibrillation and

mortality in heart failure: a community study. *Circ Heart Fail*. 2011;4(6):740-746. doi:10.1161/CIRCHEARTFAILURE.111.962688

27. Zareba W, Steinberg JS, McNitt S, et al. Implantable cardioverter-defibrillator therapy and risk of congestive heart failure or death in MADIT II patients with atrial fibrillation. *Hear Rhythm*. 2006;3(6):631-637.

28. Deshmukh A, Kumar G, Pant S, Rihal C, Murugiah K, Mehta JL. Prevalence of Takotsubo cardiomyopathy in the United States. *Am Heart J*. 2012;164(1):66-71.

29. Akashi YJ, Goldstein DS, Barbaro G, Ueyama T. Takotsubo cardiomyopathy: a new form of acute, reversible heart failure. *Circulation*. 2008;118(25):2754-2762.

30. Bybee KA, Kara T, Prasad A, et al. Systematic Review: Transient Left Ventricular Apical Ballooning: A Syndrome That Mimics ST-Segment Elevation Myocardial Infarction. *Ann Intern Med*. 2004;141(11):858-865. doi:10.7326/0003-4819-141-11-200412070-00010

31. Kurowski V, Kaiser A, von Hof K, et al. Apical and midventricular transient left ventricular dysfunction syndrome (tako-tsubo cardiomyopathy) frequency, mechanisms, and prognosis. *Chest*. 2007;132(3):809-816.

32. Sharkey SW, Lesser JR, Zenovich AG, et al. Acute and reversible cardiomyopathy provoked by stress in women from the United States. *Circulation*. 2005;111(4):472-479.

33. Tsuchihashi K, Ueshima K, Uchida T, et al. Transient left ventricular apical ballooning without coronary artery stenosis: a novel heart syndrome mimicking acute myocardial infarction. *J Am Coll Cardiol*. 2001;38(1):11-18.

CHAPTER 3

3. Emerging Predictors of Incident Heart Failure (HF) among Commercially Insured Postmenopausal Women

3.1 Abstract

Objective: To identify emerging predictors (polypharmacy and some prescription medications) of incident HF among postmenopausal women using supervised machine learning methods.

Methods: The current study used a retrospective cohort design with a baseline and follow-up period. The baseline period was used to identify risk factors for HF among postmenopausal women without HF (N = 152,592). Data were obtained from Optum's de-identified Clinformatics® Data Mart Database (Optum, Eden Prairie, MN), de-identified health insurance claims data, for the period (2007 – 2016). The study cohort consisted of postmenopausal women (age ≥ 50 years) who were free of HF during the baseline period. The target variable was incident HF identified during the two-year follow-up period. Features (i.e., independent variables) were selected based on a conceptual framework and published literature. Multivariable logistic regression and three classification machine learning algorithms (cross-validated logistic regression (CVLR), random forest (RF), and eXtreme Gradient Boosting (XGBoost) algorithms) were used to identify predictors of HF. All models were compared in terms of their predictive abilities (accuracy, sensitivity, specificity, and Area Under the Curve (AUC)). The associations of the leading predictors to incident HF were explored with an interpretable machine learning SHapley Additive exPlanations (SHAP) technique.

Results: About 2.1% of postmenopausal women (N = 3,213) developed HF during the 2-year follow-up period. The predictive accuracy was highest in the random forest algorithm with AUC of 0.87, sensitivity of 0.87, and specificity of 0.71. The eight leading predictors of incident HF consistent across all models were: older age, arrhythmia, polypharmacy, Medicare, COPD, coronary artery disease, hypertension, and chronic kidney disease. Individual medications such as sulfonylureas and antibiotics other than fluoroquinolones also predicted incident HF, but only in CVLR and RF for sulfonylureas, and only antibiotic use other than fluoroquinolones predicted HF when using XGBoost.

Conclusion: Machine learning methods identified some novel risk factors for incident HF in postmenopausal women. Further research with prospective cohorts is needed to confirm the effects of specific prescription medications on HF.

3.2 Introduction

Numerous studies have used statistical or machine learning methods to identify risk factors for heart failure (HF) among both men and women, older individuals, and those with specific chronic conditions (e.g., diabetes, coronary artery disease)[1–6]. Although there are sex differences in the etiology of HF and late-age onset of HF in women[7,8], only 7 studies have exclusively focused on incident HF among postmenopausal women[9–15]. Of those, three used data from Women's Health Initiative (WHI)[9–11]. These previous studies have shed light on several risk factors including medical conditions, lifestyle behaviors such as physical activity, race, sex-specific risk factors such as number of live births, age at first pregnancy, and age from menarche to menopause. However, those studies have several limitations such as not examining polypharmacy and prescription medication use[9–13,15], not US-based[12], specific to certain US geographical areas[13], or specific to postmenopausal women with coronary artery disease[14]. Although a study by Bibbins-Domingo et al. included medication use, it only examined the effect of medications for coronary artery disease (i.e., aspirin, angiotensin-converting-enzyme (ACE) inhibitors, beta-blockers, digoxin, diuretics, calcium channel blockers, and statin) on incident HF among postmenopausal women with coronary artery disease[14].

There is emerging evidence that polypharmacy can increase incident HF[16]. A recent study using a large healthcare claims database has indicated that polypharmacy is associated with a high risk of HF among older individuals with atrial fibrillation[16]. In addition, some prescription medications used to treat the risk factors for HF can increase the risk of HF in addition to their risk for adverse drug reactions and drug-drug interactions[17–20]. For example, a published systematic review suggests that among those with diabetes, a risk factor for HF, except metformin all other oral antidiabetics were associated with increased risk of HF[20]. Recently, fluoroquinolones, antibiotics used to treat infections, have been tied to an increased risk of aortic and mitral regurgitation, conditions in which the blood backflows into the heart and may lead to HF development[21]. The significant risk associated with fluoroquinolones can mainly occur due to its potential adverse effect of increasing the risk of aortic dissections[22]. Case reports have also suggested that analgesic, antiepileptic, and anxiolytic medications can lead to significant HF[23,24].

Therefore, an examination of the risk of polypharmacy and specific prescription medications on incident HF risk after controlling for established risk factors among postmenopausal women is needed. In this study, we focused on postmenopausal women for many reasons: 1) hormonal changes that may place them at higher risk for HF[25]; 2) high prevalence of established risk factors for HF[26,27]; and 3) postmenopausal women are more likely to use prescription medications for treating prevalent conditions such as diabetes and bacterial infections[28–30].

However, to date, no study has included oral antidiabetics, antibiotics, and antiepileptics as predictors of incident HF among postmenopausal women. Identification of prescription medications that predict incident HF among postmenopausal women can help clinicians, payers, patients, and other stakeholders to weigh the harms and benefits of commonly used medications and personalize treatment plans. Therefore, this present study used real-world data of commercially insured postmenopausal women to examine whether oral antidiabetics, antiepileptics, and antibiotics are leading predictors of incident HF using supervised machine learning methods. In this study, women aged 50 or older were considered to be postmenopausal based on the average age of postmenopausal women in the US, as well as, previous research[31,32].

3.3 Methods

Study design

We used a retrospective cohort study design with a 2-year baseline period and a 2-year follow-up period. Baseline and follow-up periods were defined using a calendar year approach. The HF free cohort was identified using both years of the baseline period and incident HF was identified using the 2-year follow-up period. HF risk factors were measured during the 2nd year of the baseline period.

Study cohort

The cohort consisted of postmenopausal women (age ≥ 50 years) who were free of HF during the baseline period. To identify and exclude those with HF during the cohort identification period, postmenopausal women who had at least one inpatient claim or two outpatient claims (30 days apart) for HF were considered as having HF[33]. We also excluded postmenopausal women with the following heart valvular disorders: mitral valve disease or insufficiency, aortic valve disease or insufficiency, and aortic valve or mitral valve regurgitation due to their family history. These valve disorders were identified based on ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) or ICD-10 CM (International Classification of Diseases, Tenth Revision, Clinical Modification) diagnosis codes (Appendix 6.2). Finally, all postmenopausal women had to be continuously enrolled in a commercial insurance plan with both medical and pharmaceutical benefits throughout the observation period. We pooled 6 cohorts (2008-2011; 2009-2012; 2010-2013; 2011-2014; 2012-2015; and 2013-2016) to gain adequate sample. After applying the inclusion/exclusion criteria, the final analytical cohort consisted of 152,592 postmenopausal without HF during the baseline period.

Data source

For this study, we used de-identified health insurance claims data from Optum's de-identified Clinformatics® Data Mart Database (Optum, Eden Prairie, MN) for the period from January 2007 to December 2016. This geographically diverse database contains healthcare claims from a 10% sample of 47 million individuals. Of whom, about 80% purchased insurance through their employers; individuals insured in Medicare Advantage plans were also included in this dataset. The data contain inpatient, outpatient, and pharmacy claims, as well as, certain demographic characteristics that are routinely collected during health insurance enrollment. Use of prescription medications was obtained from pharmacy claims. We used National Drug Codes (NDCs) and American Hospital Formulary Service (AHFS) classification system codes to identify oral antidiabetics, antibiotics, and antiepileptics (Appendix 6.3).

Outcome

Incident HF (yes/no)

The primary outcome was the development of HF (incident HF) during the follow-up period, and this was measured as a binary variable to indicate if incident HF occurred during the follow-up period (yes or no). Incident HF was identified using ICD-9 and ICD-10 codes (see Appendix 6.2). Postmenopausal women who had at least one inpatient claim or two outpatient claims (30 days apart) for HF during the follow-up period were classified as having incident HF.

Risk factors (i.e., features)

Risk factors for HF, also known as features, were selected based on prior published literature and our conceptual framework. We used the modified determinants of health outcome and chronic disease model, which was originally proposed by Wilkinson and Marmot[34], to create an initial list of features (N=37) (see Table 1). Based on this model, a disease incidence (in our case, HF) can be influenced by *(1) biological factors* (e.g., age), *(2) access to care factors* (e.g., type of insurance plan), *(3) community resources* (e.g., geographical region), *(4) medication-related factors* (e.g., polypharmacy, defined as ≥ 6 medications excluding antibiotics and antidiabetic medications), and *(5) health status* measured by chronic health conditions such as diabetes, asthma, and chronic obstructive pulmonary disease (COPD), coronary artery disease, acute myocardial infarction, and hypertension, and *6) lifestyle factors* such as substance abuse and obesity. Medication use was derived from prescription drugs file using NDCs or AFHS classification codes. Three classes of oral antidiabetics were selected (thiazolidines, sulfonylureas, and dipeptidyl peptidase-4 (DPP-4) inhibitors) because they have been linked to negative cardiovascular diseases, including HF[18-24]. We also include metformin because it has been shown to have protective effects. For antibiotic use, we created a 3-level variable with the following categories: 1) any fluoroquinolone use, 2) other antibiotics, and 3) no antibiotics use.

Analytic approach: machine learning algorithms

We used three different supervised machine learning algorithms to identify the leading predictors of incident HF among postmenopausal women. First, we used a cross-validated logistic regression (CVLR), which is widely used to predict the occurrence of an event in clinical research. For the CVLR model, we used a 10-fold cross validation approach. The second method is random forest (RF) classification. The third algorithm used in this study is eXtreme Gradient Boosting (XGBoost) algorithm.

Model evaluation

The predictive abilities of all machine learning algorithms were evaluated by obtaining the following measures: accuracy, sensitivity, specificity, and area under the ROC curve (AUC) using a test dataset. In addition, we built a multivariable logistic regression model using the same features (i.e., independent variables). This statistical model serves as a base model to compare the performance of our machine learning models.

Model development

The first step on model development is the random split of training (70%) and test datasets (30%). Our dataset was highly imbalanced with only 2.1% (N = 3,213) of postmenopausal women with incident HF; such severe imbalance is difficult to model and requires specialized techniques (example: under and over sampling). We used an undersampling technique by randomly selecting women without HF until we reached a 1:1 ratio of those with and without incident HF. The balanced dataset (N= 2,233 with HF and 2,265 without HF) was used to train our machine learning models. We used the original test dataset (that did not undersample women without HF) to evaluate model performance.

Tuning of hyperparameters

An important step in building a machine learning algorithm is the tuning of the hyperparameters of the algorithm (e.g., the number of trees in the forest and depth of the decision tree). This process can reduce overfitting to training data and improve the predictive ability of the algorithm. We used automated methods to adjust the parameters of our machine learning algorithms (e.g., grid search)

Feature importance

In the CVLR algorithm, the importance of the baseline features was obtained based on feature importance. For the RF algorithm, feature importance was obtained using the mean decrease in prediction accuracy without the variable in the model and mean decrease in the Gini index, a measure of impurity of the dataset, by including the variable. Similar to RF, feature importance in XGBoost is measured by each feature's gain. In other words, feature importance is determined based on the contribution of each feature to the final prediction.

Interpretable feature associations to incident HF

To explain the association of leading predictors to incident HF, an interpretable machine learning technique called SHapley Additive exPlanations (SHAP) was used. SHAP values derive the direction of association and importance of features by using the marginal contribution of each of the features with all combinations of other features included in the model. Dataset construction was performed using SAS 9.4 (Cary, NC) and all predictive models were built in R software (R Development Core Team, Vienna, Austria).

3.4 Results

Baseline characteristics of the study cohort

The characteristics of the study cohort by incident HF in the original dataset (N=152,592) are presented in Table 2. In the original dataset, only 2.1% (N = 3,213) of postmenopausal women developed HF during the 2-year follow-up period. Women aged 80 years and older had a higher percentage (4.4%) of incident HF compared to those aged 50-64 years (0.3%). We found that 5.1% of those with polypharmacy had incident HF, whereas only 1.5% of those without polypharmacy developed HF during the follow-up period. In terms of prescription medication use, a higher proportion of postmenopausal women with fluoroquinolones had incident HF compared to those with no fluoroquinolone use (3.2% vs. 1.9%). We observed that 4.8% of those with sulfonylurea use developed incident HF compared to 2.0% of those with no sulfonylurea use. Among those with DPP-4 inhibitor use, 4.0% developed HF during the follow-up period, while 2.1% of those with no DDP-4 inhibitor use had incident HF. We also found that those with gabapentin use had a higher percentage of incident HF than those with no gabapentin use (4.4%

vs. 2.0%). Regarding chronic conditions, a higher percentage of those with acute myocardial infarction (11.6%) developed HF during the follow-up period compared to 2.1% of those without acute myocardial infarction. Also, postmenopausal women with coronary artery disease had a higher percentage (8.1%) of incident HF than those with no coronary artery disease (1.7%). In regard to other factors, we found that obese women had a higher proportion of incident HF compared to non-obese women (3.3% vs. 2.0%).

Performance of machine learning algorithms using test data

Table 3 summarizes the performance metrics of all models obtained by testing the models with the test dataset. Based on the AUC score, the RF model was the best model for predicting incident HF in postmenopausal women. It has an AUC of 87%. The sensitivity was 87%, and the specificity was 71%. The sensitivity ranged from 0.78 in the multivariable logistic regression model, 0.78 in CVLR, and 0.82 in XGBoost. Specificity values for these models were: 0.71, 0.74, and 0.69, respectively.

Feature importance in machine learning algorithms

Common leading predictors of incident HF across all machine learning algorithms were old age (≥ 80 years), arrhythmia, polypharmacy, Medicare, Chronic Obstructive Pulmonary Disease (COPD), coronary artery disease, hypertension, chronic kidney disease, and diabetes. Table 4 summarizes leading predictors of incident HF from all machine learning algorithms. Regarding prescription medications, sulfonylurea use was identified as a predictor of incident HF in the CVLR and RF models. Adjusted odds ratios and 95% confidence intervals of top significant predictors of incident HF yielded from CVLR are presented in Figure 1. Antibiotic use (other than fluoroquinolones) ranked 12th in the XGBoost model. Figure 2 shows the top 10 predictors of incident HF from the XGBoost.

Feature association

SHAP summary plot explains the feature effect on the prediction and the direction of association of study features to incident HF (Figure 3). In this plot, each observation is represented with a single dot, and each dot is presented with a color, either yellow or purple, depending on its value. Yellow indicates that the feature value is ” No”, while purple indicates that the feature value is “Yes”1. The x-axis of the SHAP summary plot expresses the marginal contribution of the feature to the change in the predicted probability of incident HF, and the y-axis represents leading predictors based on their SHAP values. Our SHAP summary plot suggested positive associations of old age, Medicare, polypharmacy, arrhythmia, hypertension, COPD, coronary artery disease, and diabetes to incident HF. In contrast, it showed that postmenopausal women with hyperlipidemia were less likely to develop incident HF (i.e., negative associations).

3.5 Discussion

Using machine learning algorithms, this study identified modifiable and non-modifiable leading predictors of incident HF among postmenopausal women. Our study confirmed that older age is a strong predictor of incident HF, which is an established risk factor[4,6,14]. Prior research has shown that aging is associated with some structural and functional changes (e.g., myocardial thickness and a decline in physiological processes) that negatively affect the heart and arterial system. These changes increase the risk of cardiovascular disease, including HF[35–37].

In our study, polypharmacy, defined as taking 6 or more medications, was a leading predictor in all algorithms (ranked 5 in RF, ranked 3 in XGBoost, and ranked 3 in CVLR). In our study cohort, among those with incident HF, 41.1% had polypharmacy compared to 16.4% in those without incident HF. The presence of polypharmacy in this population could be attributed to the high prevalence of multiple chronic conditions among postmenopausal women[38,39]. To manage these conditions, they may seek healthcare from multiple specialists and providers[40]. This can increase the number of prescriptions medication and duplicate therapies[41]. Our findings have implications for promoting evidence-based methods to reduce polypharmacy; for example, engaging pharmacists and incorporating their recommendations, reviewing patients' medications regularly, and educating patients[42].

Furthermore, the presence of chronic health conditions can predict incident HF in this population. Our models identified chronic conditions (i.e., arrhythmia, coronary artery disease, hypertension, chronic kidney disease, and diabetes) as the leading predictors of HF risk, consistent with the literature[1,3,14]. We also identified COPD and stroke as predictors of incident HF among postmenopausal women. This is in line with a previous study showing that COPD patients have a higher risk to develop HF compared to those without COPD[43]. The relationship between COPD and cardiovascular diseases, including HF, is complex and includes several biological mechanisms[44]. It has been suggested that severe COPD may lead to HF through pulmonary hypertension[45]. Early identification and good management of COPD may decrease the risk of HF in postmenopausal women. For example, screening for COPD may help to reduce the risk of HF in this population.

The prediction of incident HF by specific medications was not consistent. Although they were not leading predictors, sulfonylureas predicted incident HF in the CVLR and RF algorithms, and antibiotics other than fluoroquinolones were found to predict incident HF in the XGBoost algorithm. Further research with prospective cohorts is needed to confirm the effect of sulfonylureas and antibiotics on incident HF. Fluoroquinolones were not identified as a predictor of incident HF even though they were found to be associated with heart valve disorders that may increase the risk of HF, as shown by a previous case-control study[21]. These conflicting findings may be due to the differences in study designs, analytical approaches, and study populations. Other study medications did not predict incident HF. Future research evaluating the cardiovascular safety of prescription medications among postmenopausal women may need to consider using a prospective design and the cumulative use of medications.

Our study had both strengths and limitations. We used a representative real-world sample of commercially insured postmenopausal women to predict incident HF. This allowed us to generate real-world evidence on predictors of incident HF and cardiovascular safety of polypharmacy in this understudied population. We examined a comprehensive set of risk factors including established and some novel risk factors (e.g., polypharmacy and specific prescription medications). We also utilized three classification machine learning methods to increase the rigor, robustness, and precision of our investigation. However, these study findings should be interpreted in the context of its limitations. Our data lacked some important clinical variables (e.g., type and severity of HF, laboratory findings, and severity of chronic conditions), socioeconomic characteristics (e.g., income and education), and race. Not including these variables might influence the performance of our models.

3.6 Conclusion

Findings from this study confirmed established risk factors of incident HF as well as some novel risk factors using supervised machine learning algorithms. Among the modifiable factors, the negative effect of polypharmacy was highlighted, suggesting that medication utilization review may be an important element of HF prevention among postmenopausal women. Future studies need to incorporate biological factors to identify the contribution of medication-related factors on incident HF and to increase predictive accuracy.

Table 1
List of Baseline Study Features (N = 37) Considered
Postmenopausal Women (Age ≥ 50 Years)
Optum Clinformatics Data Mart 10% Sample (2007 – 2016)

Feature	Measurement Levels	Data Source	Basis of Measurement
Age group	A 3-level variable: 1) 50-64 years; 2) 65-79 years; 3) ≥ 80 years)	Enrollment file	
Medicare insurance	Yes/No	Enrollment file	
HMO	Yes/No	Enrollment file	
ER use during the baseline period	Yes/No	Outpatient claims	Revenue Center Codes
Polypharmacy	(>6 drugs for consecutive 90 days) excluding oral antidiabetics and antibiotics	Prescription Drug Claims	Generic Name
Fluoroquinolone use Other antibiotic use	Yes/No	Prescription Drug Claims	AHFS
Metformin use Sulfonylurea use DPP4 inhibitor use Thiazolidines use	Yes/No	Prescription Drug Claims	National Drug Codes
Pregabalin Gabapentin	Yes/No	Prescription Drug Claims	National Drug Codes
Acute myocardial infarction, arrhythmia, arthritis, asthma, cancer, chronic kidney disease, COPD, coronary artery disease, dementia, diabetes, hepatitis, hyperlipidemia, hypertension, osteoporosis, stroke, sleep disorders	Yes/No	Inpatient and outpatient claims	ICD-9/ICD-10 Codes
Anxiety, bipolar, depression, psycho, schizophrenia	Yes/No	Inpatient and outpatient claims	ICD-9/ICD-10 Codes
Obesity	Yes/No	Inpatient and outpatient claims	ICD-9/ICD-10 Codes
Any substance abuse	Yes/No	Inpatient and outpatient claims	ICD-9/ICD-10 Codes
Region of residence	A 4-level variable (Northeast, Midwest, South, West)	Enrollment File	

Abbreviations: HMO: Health maintenance organization; ER: emergency room; AHFS: American Hospital Formulary Service; DPP-4 inhibitors: Dipeptidyl Peptidase-4 inhibitors.

Table 2
Baseline Characteristics of Study Cohort
By Incident Heart Failure
Postmenopausal Women (Age > 50 Years)
Optum Clinformatics Data Mart 10% Sample (2007-2016)
Row Percentages

	Incident HF		**No Incident HF**		**P-value**
	N	**%**	**N**	**%**	
ALL	**3,213**	**2.1**	**149,379**	**97.9**	
Biological Factors					
Age in years					<0.001
50-64 years	78	0.3	25,284	99.7	
65-79 years	770	1.0	73,083	99.0	
80 years and older	2,365	4.4	51,012	95.6	
Access to Care Factors					
Medicare insurance					<0.001
Yes	69	11.6	528	88.4	
No	3,144	2.1	148,851	97.9	
Medication-related Factors					
Polypharmacy					<0.001
Yes	1,324	5.1	24,519	94.9	
No	1,889	1.5	124,860	98.5	
Fluoroquinolone use					<0.001
Yes	661	3.2	19,896	96.8	
No	2,552	1.9	129,483	98.1	
Other antibiotic use					0.884
Yes	994	2.1	46,033	97.9	
No	2,219	2.1	103,346	97.9	
Metformin use					<0.001
Yes	449	3.4	12,717	96.6	
No	2,764	2.0	136,662	98.0	
Sulfonylurea use					<0.001
Yes	324	4.8	6,358	95.2	
No	2,889	2.0	143,021	98.0	
Thiazolidines use					<0.001
Yes	74	3.7	1,952	96.3	
No	3,139	2.1	147,427	97.9	
DPP4 inhibitor use					<0.001
Yes	99	4.0	2,386	96.0	
No	3,114	2.1	146,993	97.9	
Pregabalin					<0.001
Yes	67	3.8	1,685	96.2	
No	3,146	2.1	147,694	97.9	
Gabapentin					<0.001
Yes	294	4.4	6,346	95.6	
No	2,919	2.0	143,033	98.0	
Health-related Risk Factors					
Hypertension					<0.001
Yes	2,536	3.3	73,634	96.7	
No	677	0.9	75,745	99.1	
Coronary artery disease					<0.001
Yes	824	8.1	9,288	91.9	
No	2,389	1.7	140,091	98.3	

Continued

Acute myocardial infarction					<0.001
Yes	69	11.6	528	88.4	
No	3,144	2.1	148,851	97.9	
Arrhythmia					<0.001
Yes	926	6.4	13,624	93.6	
No	2,287	1.7	135,755	98.3	
Stroke					<0.001
Yes	411	5.5	7,048	94.5	
No	2,802	1.9	142,331	98.1	
Hyperlipidemia					<0.001
Yes	1,942	2.6	72,232	97.4	
No	1,271	1.6	77,147	98.4	
Diabetes					<0.001
Yes	1,330	3.8	33,476	96.2	
No	1,883	1.6	115,903	98.4	
COPD					<0.001
Yes	736	5.9	11,636	94.1	
No	2,477	1.8	137,743	98.2	
Chronic kidney disease					<0.001
Yes	614	6.7	8,614	93.3	
No	2,599	1.8	140,765	98.2	
Obesity					<0.001
Yes	309	3.3	9,132	96.7	
No	2,904	2.0	140,247	98.0	
Any substance abuse					0.001
Yes	242	3.3	7,047	96.7	
No	2,971	2.0	142,332	98.0	

Note: Based on 152,592 postmenopausal women aged 50 years and older. P-values were obtained from Chi-square test.

Abbreviations: HF: Heart failure; HMO: Health maintenance organization; DPP-4 inhibitors: dipeptidyl peptidase-4 inhibitors; COPD: chronic obstructive pulmonary disease.

Table 3
Performance of Cross-validated Logistic Regression, Random Forest, XGBoost, and Multivariable Logistic Regression Models on Incident Heart Failure
Postmenopausal Women (Age ≥ 50 Years)
Optum Clinformatics Data Mart 10% Sample (2007-2016)

Method	Accuracy	Sensitivity	Specificity	AUC
CVLR	0.74	0.78	0.74	0.76
RF	0.71	0.87	0.71	0.87
XGBoost	0.70	0.82	0.69	0.84
Multivariable logistic regression	0.72	0.78	0.71	0.82

Note: Performance metrics of the multivariable logistic regression were based on 152,592 postmenopausal women aged 50 years and older. For machine learning models, performance metrics were obtained using the original test dataset consisting of 45,778 postmenopausal women aged 50 years and older.

Abbreviations: CVLR: Cross-validated logistic regression; RF: Random forest; XGBoost: eXtreme Gradient Boosting; AUC: area under the curve.

Table 4
Consistent Predictors Out of 15 Leading Predictors of Incident HF
Postmenopausal Women (Age ≥ 50 Years)
Optum Clinformatics Data Mart 10% Sample (2007-2016)

Predictor	CVLR	RF	XGBoost
Old age (≥ 80 years)	1	1	1
Arrhythmia	2	2	4
Polypharmacy	3	5	3
Medicare	5	4	2
COPD	4	3	7
CAD	6	6	8
Hypertension	7	8	5
CKD	9	7	11
Diabetes	11	15	10
Hyperlipidemia	8	x	6
Middle age (65-79 years)	10	9	x
HMO	x	12	9
Stroke	12	10	x
Sulfonylureas	14	11	x
Midwest	13	x	15
South	x	14	14
Antibiotic (other than fluoroquinolones)	x	x	12
Dementia	x	13	x
Arthritis	x	x	13
Obesity	15	x	x

Note: Based on 2,233 postmenopausal women aged 50 years and older (training dataset).
Abbreviations: CVLR: Cross-validated logistic regression; RF: Random forest; XGBoost: extreme gradient boosting; COPD: Chronic obstructive pulmonary disease; CAD: Coronary artery disease; CKD: Chronic kidney disease; HMO: Health maintenance organization.

Figure 1: Adjusted odds ratios and 95% confidence intervals of top predictors from cross-validated logistic regression on incident heart failure
Postmenopausal Women (Age ≥ 50 Years)
Optum Clinformatics Data Mart 10% Sample (2007-2016)

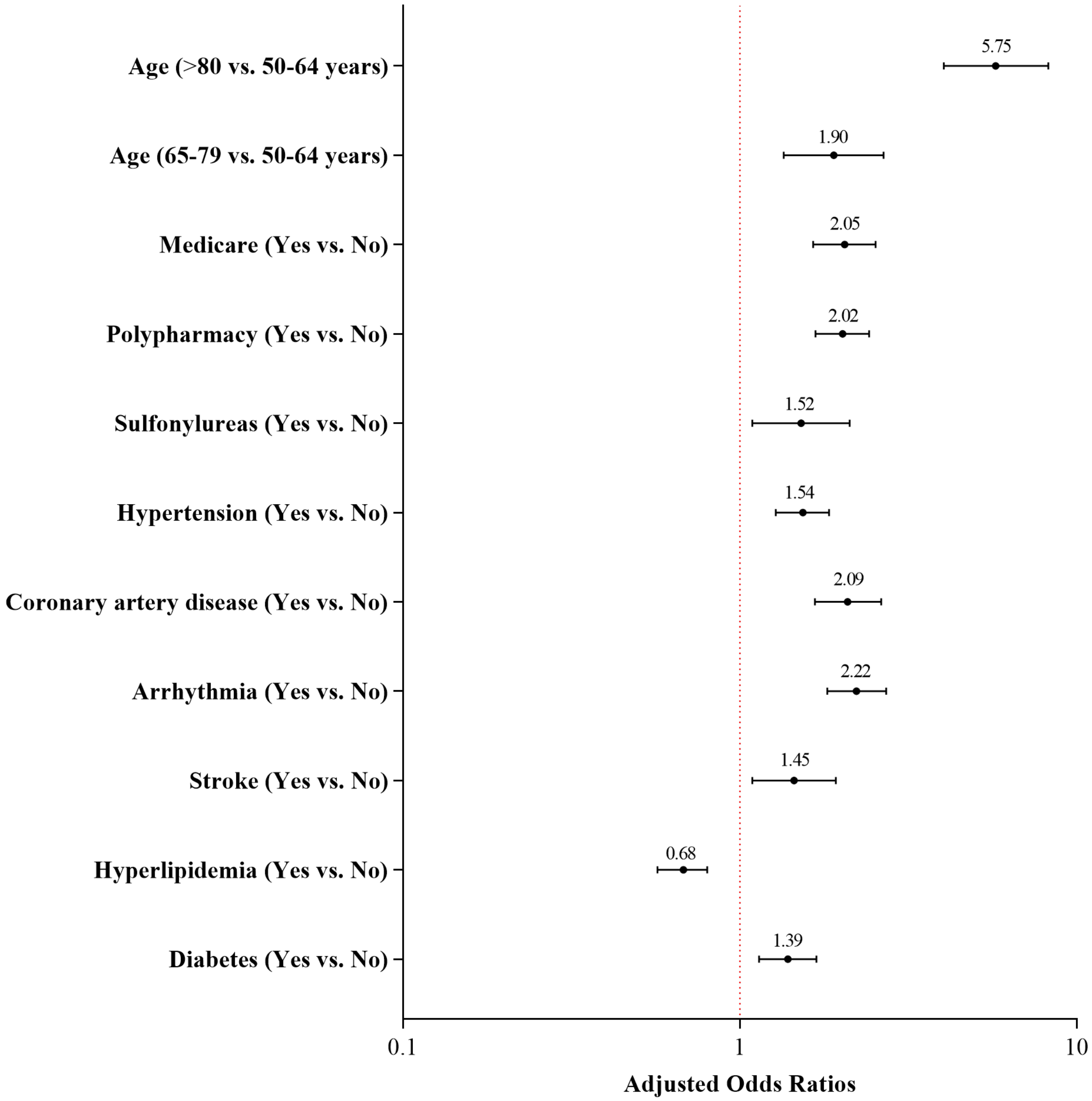
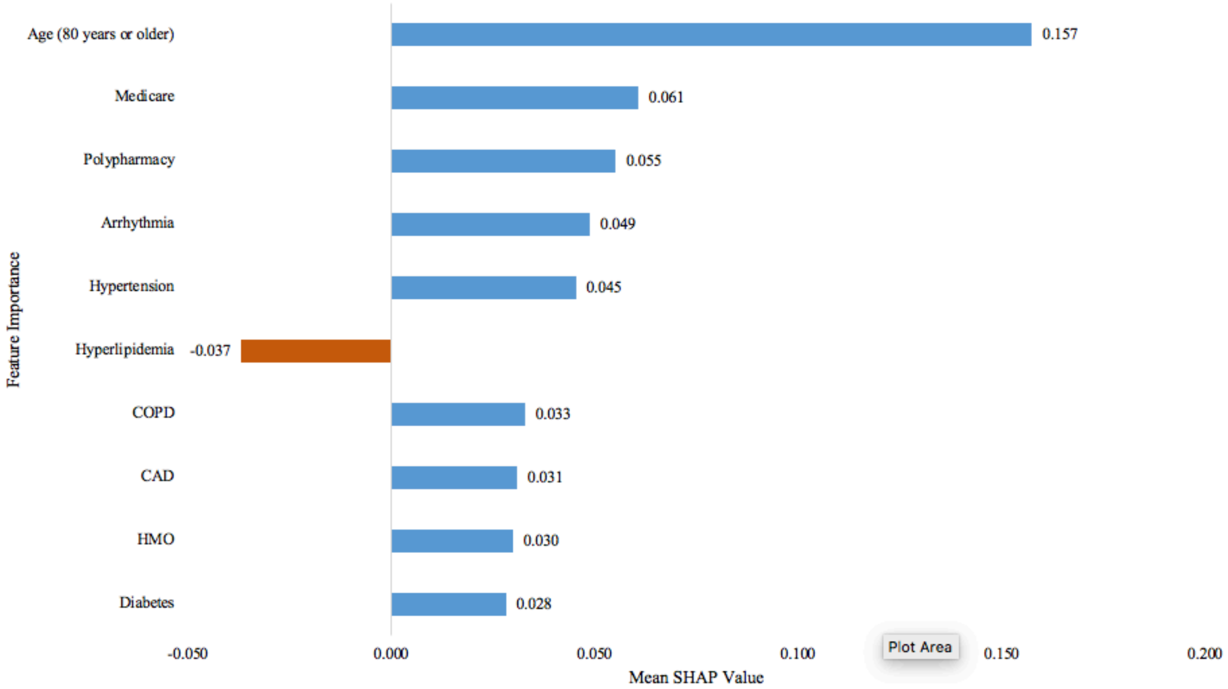
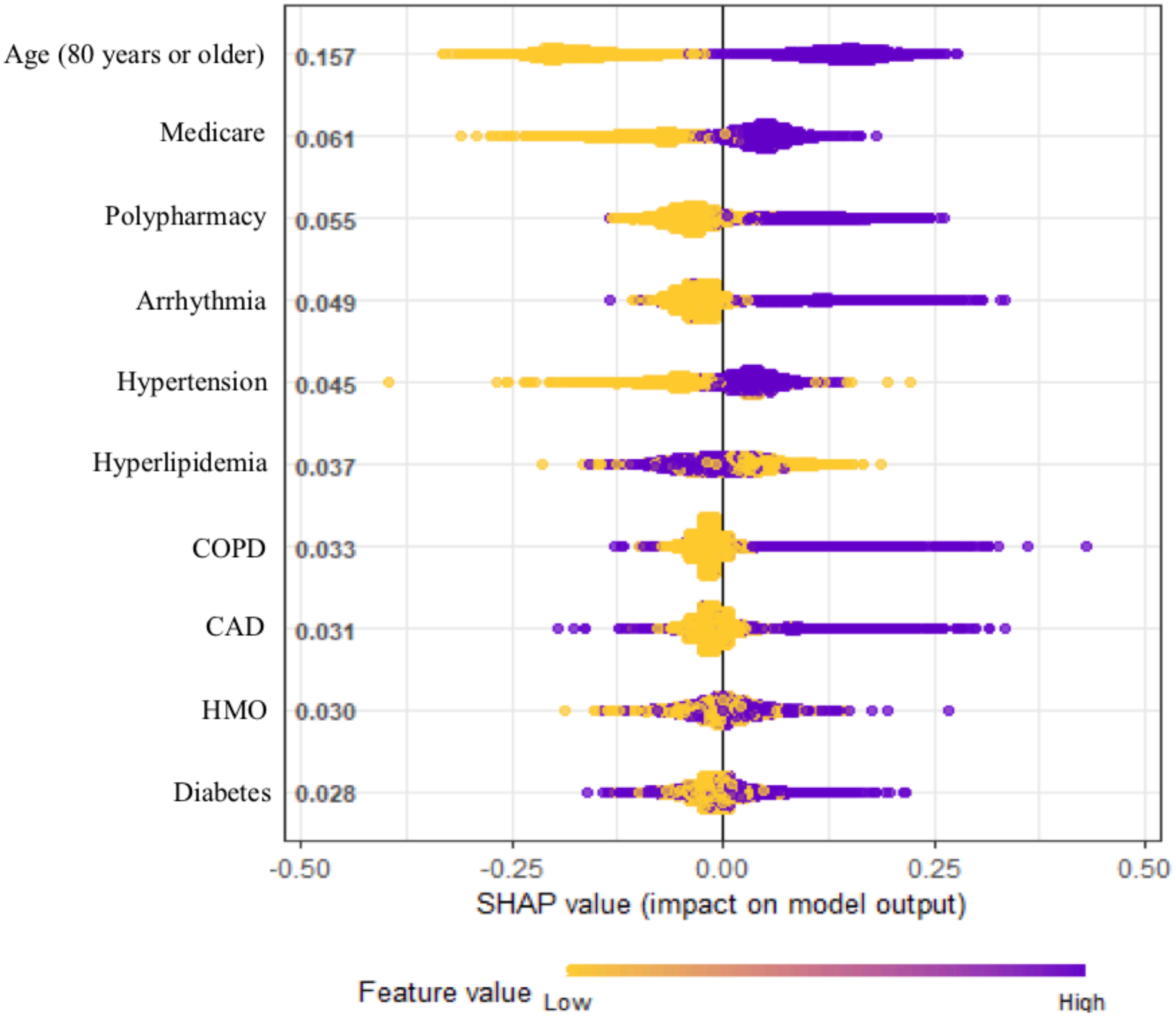


Figure 2: Top predictors of incident heart failure from XGboost algorithm and SHAP values Postmenopausal Women (Age > 50 Years) Optum Clinformatics Data Mart 10% Sample (2007-2016)



Note: Based on 2,233 postmenopausal women aged 50 years and older (training dataset).
Abbreviations: COPD: Chronic obstructive pulmonary disease; CAD: Coronary artery disease; HMO: Health maintenance organization.

Figure 3: SHAP value summary plot for top predictors of incident heart failure
Postmenopausal Women (Age ≥ 50 Years)
Optum Clinformatics Data Mart 10% Sample (2007-2016)



Note: Based on 2,233 postmenopausal women aged 50 years and older (training dataset). Features in this plot are categorical (Yes/No). Yellow dots indicate "No" (i.e., absence) and purple dots indicate "Yes" (i.e., presence).
Abbreviations: COPD: Chronic obstructive pulmonary disease; CAD: Coronary artery disease; HMO: Health maintenance organization.

3.7 References

1. Butler J, Kalogeropoulos A, Georgiopoulou V, et al. Incident heart failure prediction in the elderly: the health ABC heart failure score. *Circ Hear Fail*. 2008;1(2):125-133.

2. Segar MW, Vaduganathan M, Patel K V, et al. Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score. *Diabetes Care*. 2019;42(12):2298-2306.

3. Driscoll A, Barnes EH, Blankenberg S, et al. Predictors of incident heart failure in patients after an acute coronary syndrome: the LIPID heart failure risk-prediction model. *Int J Cardiol*. 2017;248:361-368.

4. Yang H, Negishi K, Otahal P, Marwick TH. Clinical prediction of incident heart failure risk: a systematic review and meta-analysis. *Open Hear*. 2015;2(1).

5. Goyal A, Norton CR, Thomas TN, et al. Predictors of incident heart failure in a large insured population: a one million person-year follow-up study. *Circ Hear Fail*. 2010;3(6):698-705.

6. Abdissa SG. Predictors of incident heart failure in a cohort of patients with ischemic heart disease. *Pan Afr Med J*. 2020;35.

7. Azad N, Kathiravelu A, Minoosepeher S, Hebert P, Fergusson D. Gender differences in the etiology of heart failure: A systematic review. *J Geriatr Cardiol*. 2011;8(1):15-23. doi:10.3724/SP.J.1263.2011.00015

8. Bozkurt B, Khalaf S. Heart Failure in Women. *Methodist Debakey Cardiovasc J*. 2017;13(4):216-223. doi:10.14797/mdcj-13-4-216

9. Hall PS, Nah G, Howard B V, et al. Reproductive Factors and Incidence of Heart Failure

Hospitalization in the Women's Health Initiative. *J Am Coll Cardiol*. 2017;69(20):2517-2526. doi:10.1016/j.jacc.2017.03.557

10. LaMonte MJ, Manson JE, Chomistek AK, et al. Physical Activity and Incidence of Heart Failure in Postmenopausal Women. *JACC Heart Fail*. 2018;6(12):983-995. doi:10.1016/j.jchf.2018.06.020

11. Eaton CB, Abdulbaki AM, Margolis KL, et al. Racial and ethnic differences in incident hospitalized heart failure in postmenopausal women: the Women's Health Initiative. *Circulation*. 2012;126(6):688-696.

12. Rahman I, Åkesson A, Wolk A. Relationship between age at natural menopause and risk of heart failure. *Menopause*. 2015;22(1):12-16.

13. Ebong IA, Watson KE, Goff Jr DC, et al. Age at menopause and incident heart failure: the Multi-Ethnic Study of Atherosclerosis. *Menopause (New York, NY)*. 2014;21(6):585.

14. Bibbins-Domingo K, Lin F, Vittinghoff E, et al. Predictors of heart failure among women with coronary disease. *Circulation*. 2004;110(11):1424-1430. doi:10.1161/01.CIR.0000141726.01302.83

15. Appiah D, Schreiner PJ, Demerath EW, Loehr LR, Chang PP, Folsom AR. Association of Age at Menopause With Incident Heart Failure: A Prospective Cohort Study and Meta-Analysis. *J Am Heart Assoc*. 2016;5(8). doi:10.1161/JAHA.116.003769

16. Chen N, Alam AB, Lutsey PL, et al. Polypharmacy, Adverse Outcomes, and Treatment Effectiveness in Patients≥ 75 With Atrial Fibrillation. *J Am Heart Assoc*. 2020;9:e015089.

17. Vrettos I, Voukelatou P, Katsoras A, Theotoka D, Kalliakmanis A. Diseases linked to polypharmacy in elderly patients. *Curr Gerontol Geriatr Res*. 2017;2017.

18. Page RL, O'Bryant CL, Cheng D, et al. Drugs that may cause or exacerbate heart failure: a scientific statement from the American Heart Association. *Circulation*. 2016;134(6):e32--e69.

19. Azimova K, San Juan Z, Mukherjee D. Cardiovascular safety profile of currently available diabetic drugs. *Ochsner J*. 2014;14(4):616-632.

20. Eurich DT, McAlister FA, Blackburn DF, et al. Benefits and harms of antidiabetic agents in patients with diabetes and heart failure: systematic review. *BMJ*. 2007;335(7618):497. doi:10.1136/bmj.39314.620174.80

21. Etminan M, Sodhi M, Ganjizadeh-Zavareh S, Carleton B, Kezouh A, Brophy JM. Oral Fluoroquinolones and Risk of Mitral and Aortic Regurgitation. *J Am Coll Cardiol*. 2019;74(11):1444 LP - 1450. doi:10.1016/j.jacc.2019.07.035

22. Aschenbrenner DS. New warning for fluoroquinolone antibiotics. *AJN Am J Nurs*. 2019;119(4):20.

23. Murphy N, Mockler M, Ryder M, Ledwidge M, McDonald K. Decompensation of chronic heart failure associated with pregabalin in patients with neuropathic pain. *J Card Fail*. 2007;13(3):227-229.

24. De Smedt RHE, Jaarsma T, Van Den Broek SAJ, Haaijer-Ruskamp FM. Decompensation of chronic heart failure associated with pregabalin in a 73-year-old patient with postherpetic neuralgia: a case report. *Br J Clin Pharmacol*. 2008;66(2):327.

25. Menopause and Heart Disease. American Heart Association website. Accessed August 3, 2020. www.heart.org. https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/menopause-and-heart-disease

26. Centers for Disease Control and Prevention (CDC). National Diabetes Statistics Report 2020. Estimates of Diabetes and Its Burden in the United States.; 2020.

27. Li Q, Wang X, Ni Y, et al. Epidemiological characteristics and risk factors of T2DM in Chinese premenopausal and postmenopausal women. *Lipids Health Dis*. 2019;18(1):155. doi:10.1186/s12944-019-1091-7

28. Small R, Friedman GD, Ettinger B. Concomitant medication use in postmenopausal women using estrogen therapy. *Menopause*. 2001;8(2):120-126. doi:10.1097/00042192-200103000-00007

29. Payne J, Neutel I, Cho R, DesMeules M. Factors Associated with Women's Medication Use. *BMC Womens Health*. 2004;4 Suppl 1:S29. doi:10.1186/1472-6874-4-S1-S29

30. Manteuffel M, Williams S, Chen W, Verbrugge RR, Pittman DG, Steinkellner A. Influence of patient sex and gender on medication use, adherence, and prescribing alignment with guidelines. *J Womens Health (Larchmt)*. 2014;23(2):112-119. doi:10.1089/jwh.2012.3972

31. Phipps AI, Ichikawa L, Bowles EJA, et al. Defining menopausal status in epidemiologic studies: a comparison of multiple approaches and their effects on breast cancer rates. *Maturitas*. 2010;67(1):60-66.

32. Menopause. Medscape website. Accessed March 5, 2020. https://emedicine.medscape.com/article/264088-overview#a2

33. Clinformatics® Data Mart. Optum websit. Accessed March 5, 2020. https://www.optum.com/content/dam/optum/resources/productSheets/Clinformatics_for_Data_Mart.pdf

34. Marmot M, Wilkinson R. *Social Determinants of Health*. OUP Oxford; 2005.

35. North BJ, Sinclair DA. The intersection between aging and cardiovascular disease. *Circ Res*. 2012;110(8):1097-1108.

36. Strait JB, Lakatta EG. Aging-associated cardiovascular changes and their relationship to heart failure. *Heart Fail Clin*. 2012;8(1):143-164.

37. Gerstenblith G, Frederiksen J, Yin FC, Fortuin NJ, Lakatta EG, Weisfeldt ML. Echocardiographic assessment of a normal adult aging population. *Circulation*. 1977;56(2):273-278.

38. Percent of U.S. Adults 55 and Over with Chronic Conditions. Centers for Disease Control and Prevention (CDC) website. Published November 6, 2015. Accessed March 5, 2020. https://www.cdc.gov/nchs/health_policy/adult_chronic_conditions.htm.

39. Buttorff C, Ruder T, Bauman M. *Multiple Chronic Conditions in the United States*.; 2017. doi:10.7249/tl221

40. Vogeli C, Shields AE, Lee TA, et al. Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. *J Gen Intern Med*. 2007;22(3):391-395.

41. Sherman JJ, Davis L, Daniels K. Addressing the polypharmacy conundrum. *US Pharm*. 2017;42(6).

42. Wang KA, Camargo M, Veluswamy RR. Evidence-based strategies to reduce polypharmacy: a review. *OA Elder Med*. 2013;1(1):6.

43. Curkendall SM, deluise C, Jones JK, et al. Cardiovascular disease in patients with chronic obstructive pulmonary disease, Saskatchewan Canada: cardiovascular disease in COPD

patients. *Ann Epidemiol*. 2006;16(1):63-70.

44. André S, Conde B, Fragoso E, et al. COPD and cardiovascular disease. *Pulmonology*. 2019;25(3):168-176.

45. de Miguel Díez J, Chancafe Morgan J, Jiménez García R. The association between COPD and heart failure risk: a review. Int J Chron Obstruct Pulmon Dis. 2013;8:305-312. doi:10.2147/COPD.S31236.

CHAPTER 4

4. Predictors of Heart Failure-Related Emergency Room (ER) Use with Random Forest Classification Algorithm among Commercially Insured Postmenopausal Women

4.1 Abstract

Objective: To identify leading predictors of heart failure-related emergency room (HF-related ER) use among postmenopausal women using supervised machine learning methods with data from a large commercial insurance claims database in the United States.

Methods: This is a retrospective cohort study with a 1-year baseline and 1-year follow-up period. We used de-identified health insurance claims data from Optum's de-identified Clinformatics® Data Mart Database (Optum, Eden Prairie, MN) for the period (2015 – 2016). The study cohort consisted of postmenopausal women (age ≥ 50 years) with HF during the baseline period. HF-related ER use was derived from the outpatient claims using revenue and ICD-9/ICD-10 codes. We used random forest algorithm for the primary analysis. We used interpretable machine learning techniques to explain the association of leading predictors to HF-related ER use.

Results: The study cohort consisted of 6,182 postmenopausal women with HF (mean age: 76.1 years). During the follow-up period, 27.4% (N = 1,692) had HF-related ER use. Random forest algorithm had high predictive accuracy in the test dataset (Area Under the Curve 94%, sensitivity 93%, 77% specificity, and accuracy 0.81). We found that the number of HF-related ER visits at baseline, fragmented care, age, insurance type (Health Maintenance Organization), and coronary artery disease were the top 5 predictors of HF-related ER use among postmenopausal women. Partial dependence plots suggested positive associations of the top predictors with HF-related ER use. However, insurance type was found to be negatively associated with HF-related ER use.

Conclusion: The random forest classification algorithm showed very high predictive accuracy of HR-related ER use and identified subgroups of HF patients who are at high risk for HF-related ER use.

4.2 Introduction

Nearly 50% of medical care is delivered in emergency rooms (ERs)[1]. However, ER visits are an important measure of the quality of care[2], as many of these ER visits are preventable[3]. On the other hand, providers in the ER make decisions about the hospitalization of a patient and ER utilization may present opportunities to reduce hospital utilization[4]. Notably, as heart failure (HF) is an ambulatory care sensitive condition that can be managed with primary care, hospital admissions for HF are considered preventable[5]. Beginning October 1, 2012, the Centers for Medicare and Medicaid Services (CMS) instituted the Hospital Readmissions Reduction Program (HRRP) that imposes fiscal penalties for excessive HF-related 30-day readmissions[6]. Therefore, ERs may be used to successfully managing HF exacerbations. However, there is some evidence that in the first few years following the implementation of HRRP, there was an increase in post-discharge ER visits and observation stays[7].

Although HF may be initially diagnosed in ERs[8], a large population-based study found that nearly one-third of patients with HF used the ER frequently[9]. Despite the emergence of urgent care centers as an alternative for care when primary care physicians are not available, HF patients may get treatment from ERs due to their perceptions and seriousness of symptoms[8]. In 2014, there were more than a million ER visits due to HF in the United States (US)[10]. Of those, about 37% were made by older women. In an analysis of 2017 discharge data from approximately 750 hospitals, it was reported that of the 70,092 ER visits for HF, nearly 57,534 visits were avoidable[11].

Previous studies have identified factors contributing to ER use in general. For example, chronic physical conditions[12,13], fragmented care[14], mental illness[15,16], polypharmacy[12,17], and substance abuse[12,18] were found to be associated with ER use. Several studies have examined

HF-related ER use. Yet, these studies are limited by use of older data (1992 – 2001)[19], examining the combined use of ER and hospitalization[20], and a narrow focus on specific states – California and Florida[9].

A review of ER use in the US and UK, not specific to HF, elucidated that the reasons for ER are associated with the availability of primary care, perceptions of urgency, convenience, health system factors, and cost[21]. However, in this review, studies focusing on emerging risk factors such as polypharmacy and medications that can exacerbate HF symptoms leading to ER use were not available. Therefore, a study examining predictors of HF-related ER use is needed. In this study, we focused on postmenopausal women for several reasons. Unlike other ER visits[22], HF-related ER visits are higher among older women than older men[10]. Further, women with HF have higher rates of readmission for HF mostly through ERs[23]. In addition to the HF-related reasons, women have other risk factors that may increase the probability of ER use such as women's special healthcare needs (e.g., vasomotor symptoms)[24] and higher prevalence of mental illness compared to men[25].

To date, no study has evaluated the leading predictors of ER use among postmenopausal women. Examining leading predictors from available data during a clinical encounter may assist payers and policymakers to identify subgroups of women who may be at high risk for ER use and tailor interventions that could reduce ER utilization and enhance health outcomes as ER use is associated with poor health outcomes among HF patients[20]. Therefore, the primary objective of this study is to identify the leading predictors of HF-related ER use among postmenopausal women using supervised machine learning methods with data from large commercial insurance claims. We also used interpretable machine learning techniques to evaluate the associations of leading predictors to HF-related ER use.

4.3 Methods

Study design

This study used a retrospective cohort design with a 1-year baseline period (calendar year 2015) and a 1-year follow-up period (calendar year 2016).

Data source

Data were obtained from de-identified health insurance claims data from Optum's de-identified Clinformatics® Data Mart Database (Optum, Eden Prairie, MN) for the period from January 2015 to December 2016. This database is geographically diverse and contains healthcare claims from a 10% sample of 47 million individuals; the majority of those individuals purchased insurance through their employers. In addition, this dataset includes individuals insured in Medicare Advantage plans. Some demographic characteristics, inpatient, outpatient, and pharmacy claims are available in the dataset[26].

Study Cohort

The cohort was comprised of 6,182 postmenopausal women (age ≥ 50 years) with HF. In this study, we used age of 50 years at baseline as a cut-off to define postmenopausal women. This is based on the average age of postmenopausal women in the US and prior research[27,28]. HF was identified using on ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) or ICD-10 CM (International Classification of Diseases, Tenth Revision, Clinical Modification) codes (i.e., ICD-9: 428; ICD-10: I50). Women who had at least one inpatient claim or two outpatient claims (30 days apart) for HF during the baseline period (calendar year 2015) were considered as having HF. Women had to be continuously enrolled in a commercial insurance plan with both medical and pharmaceutical benefits throughout the observation period. The final cohort size was 6,182 postmenopausal women with prevalent HF.

Outcome

HF-related ER use (yes/no)

We created a dichotomous variable with "1" indicating at least one HF-related ER visit and "0" indicating no HF-related ER visit during the follow-up period. HF-related ER use was identified from outpatient claims using the revenue codes of 0450 – 0459 and the HF diagnosis based on the ICD-9 and ICD-10 codes.

Predictors of ER use

A total of 37 predictors were selected based on the modified determinants of health outcome and chronic disease model, which was originally proposed by Wilkinson and Marmot[29], and prior literature. These features include age, access to care, healthcare utilization, community resources, health status, health behavior, and treatment-related factors (see Table 1). To assess healthcare utilization during the baseline period, we used two features: 1) the number of HF-related ER visits and 2) fragmented care. Fragmented care was measured using the Fragmentation of Care Index (FCI)[14,30]. This index measures the fragmented care of patients based on their total number of healthcare visits, the number of different providers visited, and the number of visits to each provider. The FCI score ranges from 0 to 1, where a higher FCI score indicates higher levels of fragmented care. Regarding prescription medication use, we selected medications among postmenopausal women that have been linked to HF in prior research[31–35]. These were oral antidiabetic medications (sulfonylureas and dipeptidyl peptidase-4 (DPP-4) inhibitors), antibiotics (fluoroquinolones and other antibiotics), antiepileptic medications (gabapentin). Although metformin was not among oral antidiabetics that tied to HF, we included it to examine if it has a protective effect and leads to lower HF-related ER use. We also included medications that were used to treat HF (beta-blockers, angiotensin-converting–enzyme (ACE) inhibitors, angiotensin-receptor blockers (ARBs), and diuretics) as well as antihyperlipidemic medications) because these may reduce HF exacerbations and reduce the risk of HF-related ER use. These medications were derived from prescription drugs file using National Drug Codes (NDCs) and American Hospital Formulary Service (AHFS) classification system codes (Appendix 6.3).

Analytical approach

Prediction of ER use with Random Forest (RF) Algorithm

Several Machine learning algorithms have been used to predict ER utlization[36–38]. RF is a decision-tree based ensemble algorithm with many decision trees. These decisions trees are constructed using random sampling of training data points and random subsets of features when making the decision nodes. In the case of RF for binary target variables, each tree provides a prediction for each observation. At test time, the final prediction class (e.g., "Yes HF-related ER use" or "No HF-related ER use") for an observation is obtained using the maximum number of times the test subject belonged to the class (e.g., "Yes HF-related ER use" vs"No HF-related ER use"). Feature importance was assessed using two measures: 1) the mean decrease in prediction accuracy without the variable in the model and 2) mean decrease in the Gini index, a measure of impurity of the dataset, by including the variable. For both measures, the higher the score, the more important the variable is.

In machine learning, prediction, rather than the predictor-outcome relationships, is the main focus. As we are also interested in the direction of associations, we "unboxed" a random forest classifier to enhance interpretation by using "model-agnostic" partial dependence plots (PDPs). PDPs explain the marginal effect of each study feature (i.e., predictors) on the predicted outcome (i.e., “Yes HF-related ER use” vs. “No HF-related ER use”). These plots do not only assess linear relationships, but also non-linear relationships[39].

Our dataset was randomly split into a 70% training dataset, which was internally validated (Out-of-bag –OOB sample), and a 30% test sample. OOB sample, was used to estimate the performance of RF models. For many classification machine learning algorithms, having a balanced outcome (i.e., 50% “Yes HF-related ER use” and 50% “No HF-related ER use”) is ideal. If one class has a much higher prevalence than another, the model will have better predictive accuracy only for the majority class. Our dataset was imbalanced with 27.4% of postmenopausal women having HF-related ER use during the follow-up period. Such imbalance can negatively affect the training of the RF classifier. To train the RF classifier on a balanced dataset, we used a down-sampling method to achieve 1:1 ratio of “Yes HF-related ER use” (N = 1,185) and “No HF-related ER use” (N = 1,183) of the trained dataset. RF algorithm was trained using the down-sampled data set.

All supervised machine learning algorithms require adjustments of “hyperparameters” for better predictive accuracy. In the RF algorithms, they are the number of trees and the number of variables used to make the decision nodes. We varied these hyperparameters while training. The final trained model consisted of 4 variables that were randomly split and 500 trees. However, the prediction was evaluated on the original test dataset. The predictive abilities of the RF algorithm were evaluated by obtaining the following measures using the test dataset: accuracy, sensitivity, specificity, and area under the ROC curve (AUC) using a test dataset.

Our model included 37 features (Table 1). Dataset construction was performed using SAS 9.4 (Cary, NC) and the RF model was built in R software (R Development Core Team, Vienna, Austria).

Use of multivariable logistic regression as comparator with random forest algorithm

A multivariable logistic regression model was built in SAS 9.4. This model served as a base model to compare the performance of our RF model. The comparison was based on their predictive abilities. We also reported the significant predictors of HF-related ER use from the multivariable logistic regression.

4.4 Results

Description of the study cohort by HF-related ER Use

In our study cohort, 27.4% (N = 1,692) had at least one HF-related ER visit during the follow-up period. The characteristics of the study cohort by HF-related ER use during the follow-up period are described in Table 2. The mean age of postmenopausal women with HF-related ER use was 75.8 years, and it was 76.2 years for those without HF-related ER use. On average, those with HF-related ER use during the follow-up period had an average of 3 HF-related visits during the baseline period. On the other hand, those without HF-related ER use during the follow-up period had an average of 1 HF-related ER visit during the baseline period. The average score of FCI was 0.68 among those with HF-related ER use, whereas it was 0.62 among those without HF-related ER use. With regard to the type of health insurance, 20.5% of postmenopausal women with HMO had at least one HF-related ER visit, while 32.5% of those with no HMO had HF-related ER use during the follow-up period. As compared to those with no chronic kidney disease, postmenopausal women with chronic kidney disease had a higher proportion of HF-related ER use (30.9% vs. 24.9%). We also observed that those with COPD had a higher percentage of HF-related ER use than those without COPD. Postmenopausal women with coronary artery disease had a higher proportion (32.1%) of HF-related ER use as compared to those with no coronary artery disease (22.3%). We also found that a higher percentage of postmenopausal women with diabetes had HF-related ER use during the follow-up period compared to those without diabetes (30.6% vs. 23.8%).

Performance of random forest and multivariable logistic regression

For our RF model, the accuracy was 81%; the sensitivity was 93%; the specificity was 77%. The AUC of the RF model was 94%. Using multivariable logistic regression on the same dataset, we obtained the following results: the accuracy was 66%; the sensitivity was 65%; the specificity was 67%, and the AUC was 73%.

Leading predictors of HF-related ER use

Based on feature importance from the RF model, we observed that the number of HF-related ER visits during the baseline period and fragmented care were the top 2 predictors of HF-related ER use during the follow-up period. In addition, age and HMO were identified as leading predictors of HF-related ER use. In terms of chronic conditions, coronary artery disease, arrhythmia, chronic kidney disease, arthritis, COPD, diabetes, and cancer were found to predict HF-related ER use. With regard to prescription medications, diuretics were among the top 15 predictors of HF-related ER use (Figure 1).

Significant predictors of HF-related ER use were also obtained from multivariable logistic regression. Based on this model, fragmented care, region, Medicare insurance, number of HF-related ER visits during the baseline period, acute myocardial infarction, coronary artery disease, arrhythmia, chronic kidney disease, diabetes, hypertension, and diuretics were positively associated with HF-related ER use among postmenopausal women. Figure 2 summarizes the adjusted odds ratios and 95% confidence intervals for the significant predictors of HF-related ER use yielded from the multivariable logistic regression model.

Associations of features to HF-related ER use

Partial dependence plots (PDPs) generated by RF showed the non-linear relationships between the number of HF-related ER visits during the baseline period and fragmented care, and age with HF-related ER use during the follow-up period (Figure 3). In these plots, the Y-axis expresses the log of the fraction of the votes that indicate the presence of HF-related ER use. The X-axis expresses the value of the predictor, which is 0 or 1 for categorical features. For example, the PDP shows that the presence of chronic kidney disease was associated with a higher likelihood of being classified as having HF-related ER use. PDPs also suggested positive associations of coronary artery disease, arrhythmia, chronic kidney disease, arthritis, COPD, diabetes, and cancer, regions (i.e., Midwest and South) to HF-related ER use. However, HMO was found to be negatively associated with HF-related ER use. PDPs of prescription medications indicated that sulfonylureas and DPP-4 were positively associated with HF-related ER use (Figure 4).

4.5 Discussion

In our large population-based cohort of postmenopausal women, 27.4% had HF-related ER use in 2016. We identified the number of HF-related ER visits during the baseline period as the leading predictor of HF-related ER use in the subsequent year. Although we did not explore the reasons for HF-related ER use, prior studies indicate that a majority of HF patients report frailty, and those with frailty are more likely to use ERs even a year after diagnosis[40]. It is also possible that HF patients may perceive that their condition required the resources and facilities offered by the ER[21]. As concluded by a review of reviews, multimodal interventions (support for self-management practices, education, and strong primary care) may be needed to reduce the risk of ER use among HF patients[41]. Additionally, “screening-in-triage” with telehealth may be an option to reduce the risk of ER use. In a matched cohort study, there were no differences in care received by patients with chest pain between telehealth and in-person screening[42].

Another leading predictor of HF-related ER use was fragmented care. The relationship of fragmented care and ER use has been observed in prior research[14,43]. In our cohort study, HF patients had multiple chronic conditions consistent with the published research[44]. Multimorbidity often leads to receiving care from multiple providers. About half of older individuals with Medicare receive care from two to five different providers with 12% receiving care from ten or more different providers[45]. Without effective collaboration between providers (e.g., a cardiologist and mental health provider), the quality of care decreases, and HF-related ER use may increase[46,47]. To overcome this, prior research suggested implementing transition of care interventions including patient education, telephone follow-up, medication reconciliation, and home visits[48,49]. Our study findings have implications for predictive analytics in identifying high-risk ER use patients and the opportunity to implement targeted care-coordination interventions to reduce the risk of ER use[50].

The PDPs revealed non-linear relationships of age, care fragmentation, and baseline ER visits. For example, the likelihood of ER visits increased with increased levels of fragmentation of care and leveled off at very high levels of care fragmentation. These findings suggest that very high levels of fragmented care may reflect the high clinical need and “heightened surveillance” that may have reduced the risk of ER use. The likelihood of HF-related ER use was high from the age of 50 to 55 years, and then it decreased from age of 56 to 64 years. After that, age was positively associated with HF-related ER use. Given the fact that HF is a progressive disease, those aged 65 years and older might use the ER due to the severity of HF.

In this study of HF-related ER use among postmenopausal women with HF, The RF algorithm outperformed the multivariable logistic regression. This better performance of the RF algorithm can be due to its ability to detect non-linear relationships between study features and HF-related ER use.

Potential limitations and strengths of this study should be noted. One limitation of our study was that we did not include the type and severity of HF in our models. This study also did not include socioeconomic characteristics, which have been found to predict ER use. This study had several strengths. This was the first study to use a comprehensive list of factors including prescription medications to predict HF-related ER use among postmenopausal women. RF classifier model was able to detect non-linear relationships. In statistical learning methods, each additional test run on the data (e.g., stratification, interaction) increases the statistical error. However, as RF is based on an algorithmic approach, we were able to detect non-linear relationships without loss of power.

4.6 Conclusion

Using the RF classification algorithm, we were able to predict HF-related ER use among postmenopausal women with high accuracy. Our findings show the complex relationships between predictors of HF-related ER use, suggesting there is a need to identify high-risk patients with predictive algorithms and developing targeted interventions to reduce the risk of ER visits among postmenopausal women with HF.

Table 1
Baseline Study Features
Postmenopausal Women with Heart Failure (Age ≥ 50 Years)
Optum Clinformatics Data Mart 10% Sample (2015 – 2016)

	Mean	SD
Age	76.1	9.0
Number of HF-related ER visits	1.6	3.3
Fragmented care (FCI)	0.64	0.18
	N	**%**
Medicare insurance		
Yes	5,794	93.7
HMO		
Yes	2,655	42.9
Polypharmacy		
Yes	2,403	38.9
Antihyperlipidemic use		
Yes	3,317	53.7
Beta blocker use		
Yes	3,843	62.2
ACE inhibitor use		
Yes	2,060	33.3
ARB use		
Yes	1,492	24.1
Diuretic use		
Yes	3,916	63.3
Fluoroquinolone use		
Yes	1,767	28.6
Other antibiotic use		
Yes	1,932	31.3
Gabapentin use		
Yes	960	15.5
Metformin use		
Yes	802	13.0
Sulfonylurea use		
Yes	586	9.5
DPP4 inhibitor use		
Yes	294	4.8
Hypertension		
Yes	5,711	92.4
Coronary artery disease		
Yes	3,182	51.5
Acute myocardial infarction		
Yes	467	7.6
Arrhythmia		
Yes	3,984	64.4
Stroke		
Yes	1,374	22.2
Hyperlipidemia		
Yes	4,538	73.4
Diabetes		
Yes	3,271	52.9
Cancer		
Yes	1,407	22.8

Continued

Asthma		
Yes	1,038	16.8
COPD		
Yes	2,449	39.6
Arthritis		
Yes	2,619	42.4
Osteoporosis		
Yes	1,108	17.9
Chronic kidney disease		
Yes	2,539	41.1
Anxiety		
Yes	1,229	19.9
Depression		
Yes	1,635	26.4
Dementia		
Yes	1,029	16.6
Sleep disorders		
Yes	1,444	23.4
Obesity		
Yes	1,559	25.2
Any substance abuse		
Yes	648	10.5
Region of residence		
Northeast	846	13.7
Midwest	1,491	24.1
South	2,199	35.6
West	1,646	26.6

Note: Based on 6,182 postmenopausal women (age ≥ 50 years) with heart failure enrolled in commercial insurance plans, alive, with continuous enrollment in pharmacy and medical benefits in 2015 and 2016.

Abbreviations: FCI: Fragmentation of Care Index; HMO: Health maintenance organization; ACE inhibitors: angiotensin-converting-enzyme inhibitors; ARBs: Angiotensin II receptor blockers; DPP-4 inhibitors: Dipeptidyl Peptidase-4 inhibitors; COPD: chronic obstructive pulmonary disease.

Table 2
Baseline Characteristics of Study Cohort
By Heart Failure-related Emergency Room Use During the Follow-up Period
Postmenopausal Women with HF (age ≥ 50 years)
Optum Clinformatics Data Mart 10% Sample (2015-2016)
Row Percentages

	HF-related ER Use (N=1,692) 27.4%		**No HF-related ER Use** (N=4,490) 72.6%		
	Continuous Features				
	Mean	**SD**	**Mean**	**SD**	**P-value**
Age in years	75.80	9.27	76.21	8.84	0.115
Number of HF-related ER visits	3.02	4.35	1.04	2.58	<0.001
Care fragmentation (FCI)	0.68	0.14	0.62	0.19	<0.001
	Categorical Features				
	N	**%**	**N**	**%**	**P-value**
Medicare insurance					<0.001
Yes	1,621	28.0	4,173	72.0	
No	71	18.3	317	81.7	
HMO					<0.001
Yes	544	20.5	2,111	79.5	
No	1,148	32.5	2,379	67.5	
Polypharmacy					<0.001
Yes	780	32.5	1,623	67.5	
No	912	24.1	2,867	75.9	
Antihyperlipidemic					0.249
Yes	928	28.0	2,389	72.0	
No	764	26.7	2,101	73.3	
Beta blockers					0.008
Yes	1,097	28.5	2,746	71.5	
No	595	25.4	1,744	74.6	
ACE inhibitors					0.680
Yes	557	27.0	1,503	73.0	
No	1,135	27.5	2,987	72.5	
ARBs					0.221
Yes	390	26.1	1,102	73.9	
No	1,302	27.8	3,388	72.2	
Diuretics					<0.001
Yes	1,151	29.4	2,765	70.6	
No	541	23.9	1,725	76.1	
Fluoroquinolone use					<0.001
Yes	548	31.0	1,219	69.0	
No	1,144	25.9	3,271	74.1	
Other antibiotics					0.431
Yes	516	26.7	1,416	73.3	
No	1,176	27.7	3,074	72.3	
Gabapentin use					0.001
Yes	307	32.0	653	68.0	
No	1,385	26.5	3,837	73.5	
Metformin use					0.252
Yes	233	29.1	569	70.9	
No	1,459	27.1	3,921	72.9	

Continued

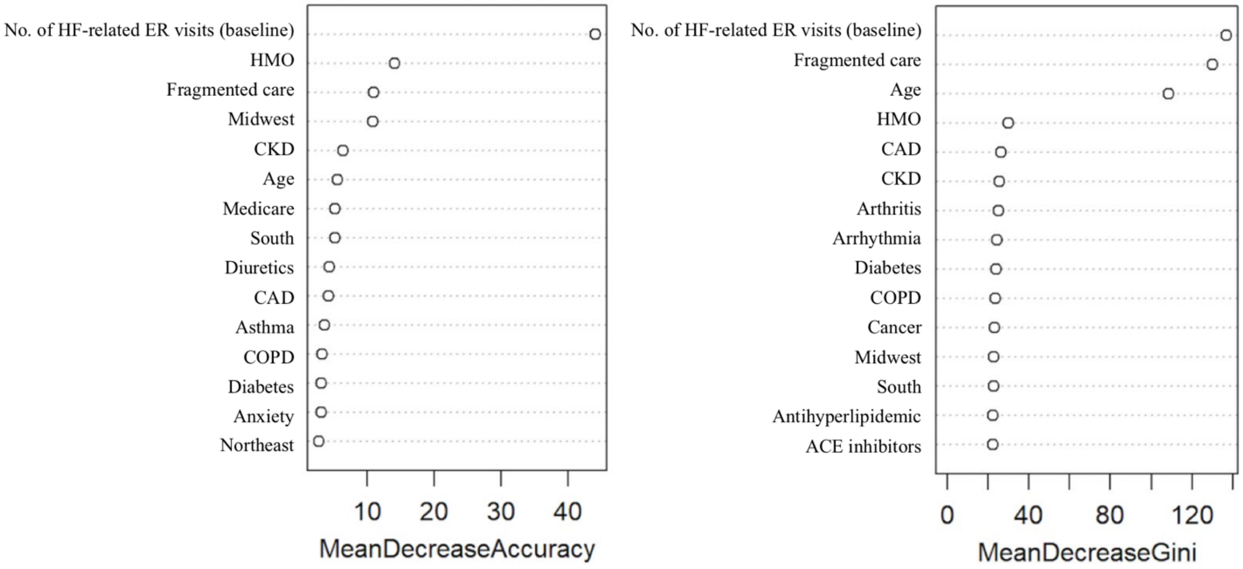
Sulfonylurea use					<0.001
Yes	197	33.6	389	66.4	
No	1,495	26.7	4,101	73.3	
DPP4 inhibitor use					0.070
Yes	94	32.0	200	68.0	
No	1,598	27.1	4,290	72.9	
Hypertension					<0.001
Yes	1,619	28.3	4,092	71.7	
No	73	15.5	398	84.5	
Coronary artery disease					<0.001
Yes	1,023	32.1	2,159	67.9	
No	669	22.3	2,331	77.7	
Acute myocardial infarction					0.126
Yes	142	30.4	325	69.6	
No	1,550	27.1	4,165	72.9	
Arrhythmia					<0.001
Yes	1,197	30.0	2,787	70.0	
No	495	22.5	1,703	77.5	
Stroke					0.006
Yes	416	30.3	958	69.7	
No	1,276	26.5	3,532	73.5	
Hyperlipidemia					<0.001
Yes	1,337	29.5	3,201	70.5	
No	355	21.6	1,289	78.4	
Diabetes					<0.001
Yes	1,000	30.6	2,271	69.4	
No	692	23.8	2,219	76.2	
Cancer					0.004
Yes	427	30.3	980	69.7	
No	1,265	26.5	3,510	73.5	
Asthma					<0.001
Yes	371	35.7	667	64.3	
No	1,321	25.7	3,823	74.3	
COPD					<0.001
Yes	821	33.5	1,628	66.5	
No	871	23.3	2,862	76.7	
Arthritis					<0.001
Yes	788	30.1	1,831	69.9	
No	904	25.4	2,659	74.6	
Osteoporosis					0.926
Yes	302	27.3	806	72.7	
No	1,390	27.4	3,684	72.6	
Chronic kidney disease					<0.001
Yes	784	30.9	1,755	69.1	
No	908	24.9	2,735	75.1	
Anxiety					<0.001
Yes	423	34.4	806	65.6	
No	1,269	25.6	3,684	74.4	
Depression					<0.001
Yes	515	31.5	1,120	68.5	
No	1,177	25.9	3,370	74.1	

Continued

Dementia					0.461
Yes	272	26.4	757	73.6	
No	1,420	27.6	3,733	72.4	
Sleep disorders					<0.001
Yes	489	33.9	955	66.1	
No	1,203	25.4	3,535	74.6	
Obesity					
Yes	480	30.8	1,079	69.2	0.001
No	1,212	26.2	3,411	73.8	
Any substance abuse					<0.001
Yes	227	35.0	421	65.0	
No	1,465	26.5	4,069	73.5	
Region of residence					<0.001
Northeast	242	28.6	604	71.4	
Midwest	521	34.9	970	65.1	
South	668	30.4	1,531	69.6	
West	261	15.9	1,385	84.1	

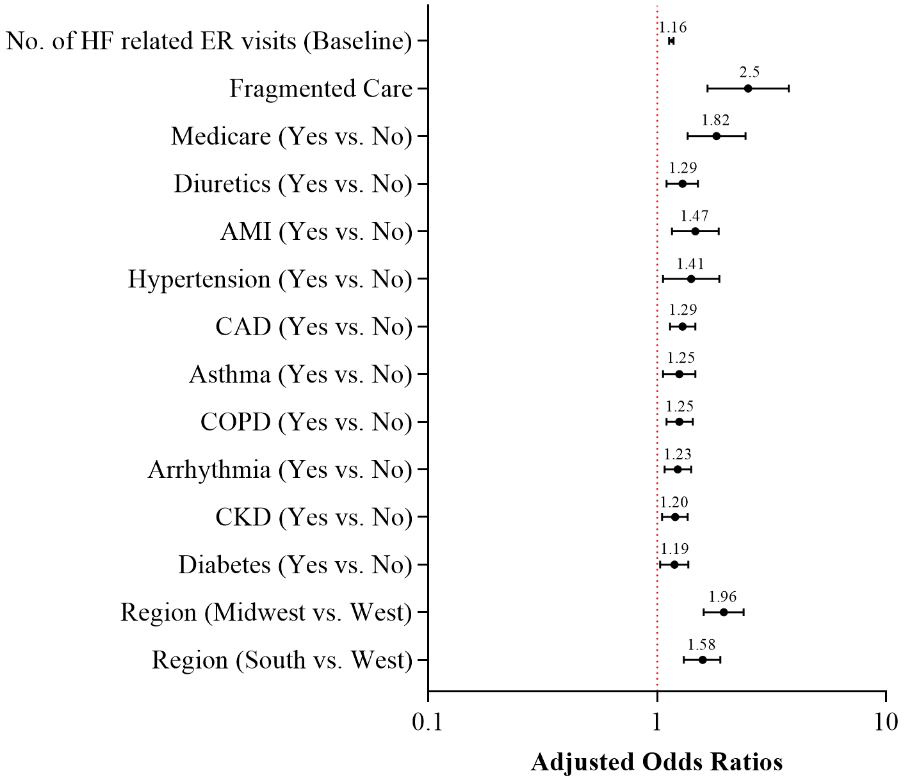
Note: Based on 6,182 postmenopausal women (age ≥50 years) with heart failure enrolled in commercial insurance plans, alive, with continuous enrollment in pharmacy and medical benefits in 2015 and 2016.
-values were obtained from t-test foe continuous features and Chi-square test for categorical features.
Abbreviations: FCI: Fragmentation of Care Index; HMO: Health maintenance organization; ACE inhibitors: angiotensin-converting-enzyme inhibitors; ARBs: Angiotensin II receptor blockers; DPP-4 inhibitors: Dipeptidyl Peptidase-4 inhibitors; COPD: chronic obstructive pulmonary disease.

Figure 1: Top predictors of heart failure-related emergency room use from the random forest
Postmenopausal Women with heart failure (Age ≥ 50 Years)
Optum Clinformatics Data Mart 10% Sample (2015-2016)



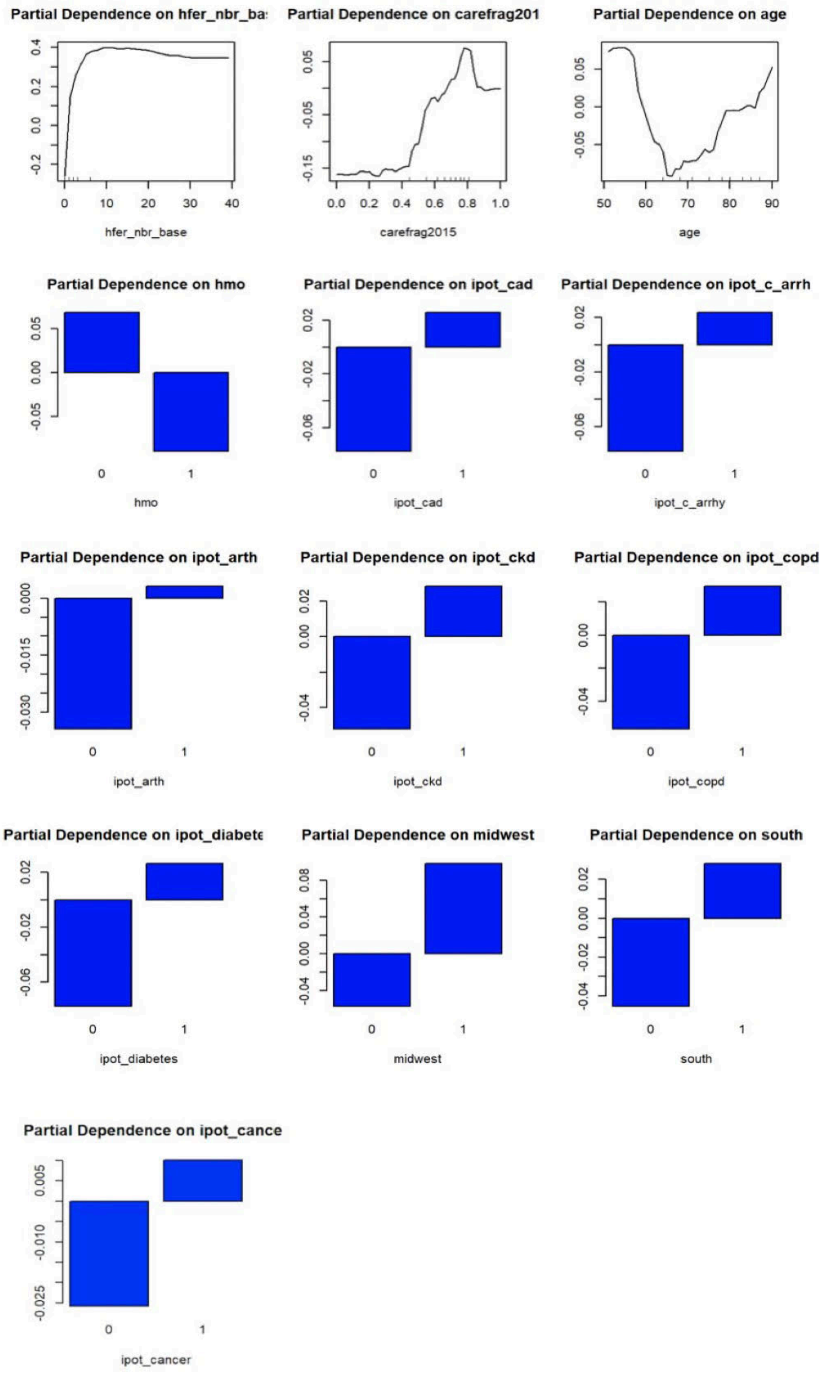
Note: Based on postmenopausal women (age > 50 years) with heart failure enrolled in commercial insurance plans, alive, with continuous enrollment in pharmacy and medical benefits in 2015 and 2016 using the training dataset (N = 2,368).
Abbreviations: HMO: Health maintenance organization; CKD: Chronic kidney disease; CAD: Coronary artery disease; COPD: Chronic obstructive pulmonary disease; ACE inhibitors: Angiotensin-converting-enzyme inhibitors.

Figure 2: Adjusted odds ratios and 95% confidence intervals of top predictors from multivariable logistic regression on heart failure-related emergency room use
Postmenopausal Women with heart failure (Age ≥ 50 Years)
Optum Clinformatics Data Mart 10% Sample (2015-2016)



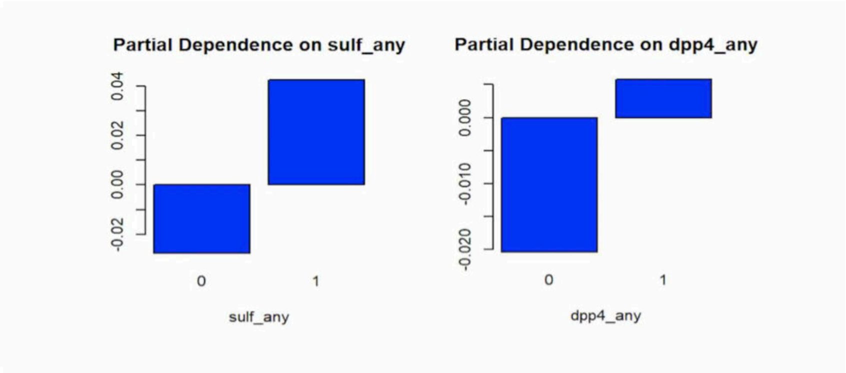
Abbreviations: AMI: Acute myocardial infarction; CAD: Coronary artery disease; COPD: Chronic obstructive pulmonary disease; CKD: Chronic kidney disease.

Figure 3: Partial dependence plots of predictors of heart failure-related emergency room use Postmenopausal Women with heart failure (Age > 50 Years) Optum Clinformatics Data Mart 10% Sample (2015-2016)



Note: Based on postmenopausal women (age > 50 years) with heart failure enrolled in commercial insurance plans, alive, with continuous enrollment in pharmacy and medical benefits in 2015 and 2016 using the training dataset (N = 2,368).

Figure 4: Partial dependence plots of prescription medications associated with HF-related ER use Postmenopausal Women with heart failure (Age ≥ 50 Years) Optum Clinformatics Data Mart 10% Sample (2015-2016)



Note: Based on postmenopausal women (age > 50 years) with heart failure enrolled in commercial insurance plans, alive, with continuous enrollment in pharmacy and medical benefits in 2015 and 2016 using the training dataset (N = 2,368).

4.7 References

1. Marcozzi D, Carr B, Liferidge A, Baehr N, Browne B. Trends in the contribution of emergency departments to the provision of hospital-associated health care in the USA. *Int J Heal Serv*. 2018;48(2):267-288.

2. Dowd B, Karmarker M, Swenson T, et al. Emergency department utilization as a measure of physician performance. *Am J Med Qual*. 2014;29(2):135-143.

3. Uscher-Pines L, Pines J, Kellermann A, Gillen E, Mehrotra A. Emergency department visits for nonurgent conditions: systematic literature review. *Am J Manag Care*. 2013;19(1):47-59.

4. Kocher KE, Dimick JB, Nallamothu BK. Changes in the source of unscheduled hospitalizations in the United States. *Med Care*. 2013:689-698.

5. AHRQ quality indicators—guide to prevention quality indicators: hospital admission for ambulatory care sensitive conditions. Publication No: 02-R0203. 2001. Agency for Healthcare Research and Quality website. Accessed August 02, 2020. https://www.ahrq.gov/downloads/pub/ahrqqi/pqiguide.pdf

6. Hospital Readmissions Reduction Program (HRRP). Centers for Medicare & Medicaid Services website. Accessed August 2, 2020. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.

7. Wadhera RK, Maddox KEJ, Kazi DS, Shen C, Yeh RW. Hospital revisits within 30 days after discharge for medical conditions targeted by the Hospital Readmissions Reduction Program in the United States: national retrospective analysis. *bmj*. 2019;366:l4563.

8. Weintraub NL, Collins SP, Pang PS, et al. Acute heart failure syndromes: emergency department presentation, treatment, and disposition: current approaches and future aims: a

scientific statement from the American Heart Association. *Circulation*. 2010;122(19):1975-1996.

9. Hasegawa K, Tsugawa Y, Camargo Jr CA, Brown DFM. Frequent utilization of the emergency department for acute heart failure syndrome: a population-based study. *Circ Cardiovasc Qual Outcomes*. 2014;7(5):735-742.

10. Jackson SL, Tong X, King RJ, Loustalot F, Hong Y, Ritchey MD. National Burden of Heart Failure Events in the United States, 2006 to 2014. *Circ Heart Fail*. 2018;11(12):e004873. doi:10.1161/CIRCHEARTFAILURE.117.004873

11. Ready, Risk, Reward: Improving Care for Patients with Chronic Conditions. Premier Inc. website. 2019. Accessed August 02, 2020. http://offers.premierinc.com/rs/381-NBB-525/images/Improving Care for Chronic Conditions%2C Premier.pdf

12. Agarwal P, Bias TK, Madhavan S, Sambamoorthi N, Frisbee S, Sambamoorthi U. Factors associated with emergency department visits: A multistate analysis of adult fee-for-service Medicaid beneficiaries. *Heal Serv Res Manag Epidemiol*. 2016;3:2333392816648549.

13. Fan L, Shah MN, Veazie PJ, Friedman B. Factors associated with emergency department use among the rural elderly. *J Rural Heal*. 2011;27(1):39-49.

14. Liu CW, Einstadter D, Cebul RD. Care fragmentation and emergency department use among complex patients with diabetes. *Am J Manag Care*. 2010;16(6):413-420.

15. Niedzwiecki MJ, Sharma PJ, Kanzaria HK, McConville S, Hsia RY. Factors associated with emergency department use by patients with and without mental health diagnoses. *JAMA Netw open*. 2018;1(6):e183528--e183528.

16. Alhussain K, Meraya AM, Sambamoorthi U. Serious psychological distress and emergency room use among adults with multimorbidity in the United States. *Psychiatry J*.

2017;2017.

17. Dufour I, Chouinard M-C, Dubuc N, Beaudin J, Lafontaine S, Hudon C. Factors associated with frequent use of emergency-department services in a geriatric population: a systematic review. *BMC Geriatr*. 2019;19(1):185.

18. Drug-Related Hospital Emergency Room Visits. National Institute on Drug Abuse (NIDA) website. May 2011. Accessed August 02, 2020. https://www.drugabuse.gov/sites/default/files/hospitalvisits.pdf

19. Hugli O, Braun JE, Kim S, Pelletier AJ, Camargo Jr CA. United States emergency department visits for acute decompensated heart failure, 1992 to 2001. *Am J Cardiol*. 2005;96(11):1537-1542.

20. Rame JE, Sheffield MA, Dries DL, et al. Outcomes after emergency department discharge with a primary diagnosis of heart failure. *Am Heart J*. 2001;142(4):714-719.

21. Coster JE, Turner JK, Bradbury D, Cantrell A. Why do people choose emergency and urgent care services? A rapid review utilizing a systematic literature search and narrative synthesis. *Acad Emerg Med*. 2017;24(9):1137-1149.

22. Milani SA, Crooke H, Cottler LB, Striley CW. Sex differences in frequent ED use among those with multimorbid chronic diseases. *Am J Emerg Med*. 2016;34(11):2127-2131.

23. Hoang-Kim A, Parpia C, Freitas C, et al. Readmission rates following heart failure: a scoping review of sex and gender based considerations. *BMC Cardiovasc Disord*. 2020;20:1-19.

24. Sarrel P, Portman D, Lefebvre P, et al. Incremental direct and indirect costs of untreated vasomotor symptoms. *Menopause*. 2015;22(3):260-266.

25. Brody DJ, Pratt LA, Hughes JP. *Prevalence of Depression among Adults Aged 20 and*

over: United States, 2013-2016. US Department of Health and Human Services, Centers for Disease Control and~…; 2018.

26. Clinformatics® Data Mart. Optum websit. Accessed August 2, 2020. https://www.optum.com/content/dam/optum/resources/productSheets/Clinformatics_for_Data_Mart.pdf

27. Phipps AI, Ichikawa L, Bowles EJA, et al. Defining menopausal status in epidemiologic studies: a comparison of multiple approaches and their effects on breast cancer rates. *Maturitas*. 2010;67(1):60-66.

28. Menopause. Medscape website. Accessed August 2, 2020. https://emedicine.medscape.com/article/264088-overview#a2

29. Marmot M, Wilkinson R. *Social Determinants of Health*. OUP Oxford; 2005.

30. Liu S, Yeung PC. Measuring fragmentation of ambulatory care in a tripartite healthcare system. *BMC Health Serv Res*. 2013;13(1):176.

31. Eurich DT. McAlister FA, Blackburn DF, Majumdar SR, Tsuyuki RT, Varney J, Johnson JA. *Benefits harms antidiabetic agents patients with diabetes Hear Fail Syst Rev BMJ*. 2007;335:497-501.

32. Durack J, Lynch S V. The gut microbiome: Relationships with disease and opportunities for therapy. *J Exp Med*. 2019;216(1):20-40. doi:10.1084/jem.20180448

33. Etminan M, Sodhi M, Ganjizadeh-Zavareh S, Carleton B, Kezouh A, Brophy JM. Oral Fluoroquinolones and Risk of Mitral and Aortic Regurgitation. *J Am Coll Cardiol*. 2019;74(11):1444 LP - 1450. doi:10.1016/j.jacc.2019.07.035

34. Heianza Y, Zheng Y, Ma W, et al. Duration and life-stage of antibiotic use and risk of cardiovascular events in women. *Eur Heart J*. 2019;40(47):3838-3845.

doi:10.1093/eurheartj/ehz231

35. Page RL, O’Bryant CL, Cheng D, et al. Drugs that may cause or exacerbate heart failure: a scientific statement from the American Heart Association. *Circulation*. 2016;134(6):e32--e69.

36. Jones A, Costa AP, Pesevski A, McNicholas PD. Predicting hospital and emergency department utilization among community-dwelling older adults: Statistical and machine learning approaches. *PLoS One*. 2018;13(11):e0206662.

37. Qiao Z, Sun N, Li X, Xia E, Zhao S, Qin Y. Using Machine Learning Approaches for Emergency Room Visit Prediction Based on Electronic Health Record Data. In: *MIE*. ; 2018:111-115.

38. Vest JR, Ben-Assuli O. Prediction of emergency department revisits using area-level social determinants of health measures and health information exchange information. *Int J Med Inform*. 2019;129:205-210.

39. Cafri G, Bailey BA. Understanding variable effects from black box prediction: Quantifying effects in tree ensembles using partial dependence. *J Data Sci*. 2016;14(1):67-95.

40. McNallan SM, Singh M, Chamberlain AM, et al. Frailty and healthcare utilization among patients with heart failure in the community. *JACC Hear Fail*. 2013;1(2):135-141.

41. den Heede K, de Voorde C. Interventions to reduce emergency department utilisation: a review of reviews. *Health Policy (New York)*. 2016;120(12):1337-1349.

42. Rademacher NJ, Cole G, Psoter KJ, et al. Use of telemedicine to screen patients in the emergency department: matched cohort study evaluating efficiency and patient safety of telemedicine. *JMIR Med informatics*. 2019;7(2):e11233.

43. Frandsen BR, Joynt KE, Rebitzer JB, Jha AK. Care fragmentation, quality, and costs among chronically ill patients. *Am J Manag Care*. 2015;21(5):355-362.

44. Manemann SM, Chamberlain AM, Roger VL, et al. Multimorbidity and functional limitation in individuals with heart failure: a prospective community study. *J Am Geriatr Soc*. 2018;66(6):1101-1107.

45. Report to the Congress: Increasing the Value of Medicare. Medicare Payment Advisory Commission (MedPAC) website. June 2006. Accessed August 02, 2020. http://www.medpac.gov/docs/default-source/reports/Jun06_EntireReport.pdf

46. Maeng DD, Martsolf GR, Scanlon DP, Christianson JB. Care coordination for the chronically ill: understanding the patient's perspective. *Health Serv Res*. 2012;47(5):1960-1979.

47. Schoen C, Osborn R, How SKH, Doty MM, Peugh J. In Chronic Condition: Experiences Of Patients With Complex Health Care Needs, In Eight Countries, 2008: Chronically ill US patients have the most negative access, coordination, and safety experiences. *Health Aff*. 2008;27(Suppl1):w1--w16.

48. Mansukhani RP, Bridgeman MB, Candelario D, Eckert LJ. Exploring transitional care: evidence-based strategies for improving provider communication and reducing readmissions. *Pharm Ther*. 2015;40(10):690.

49. Albert NM, Barnason S, Deswal A, et al. Transitions of care in heart failure: a scientific statement from the American Heart Association. *Circ Hear Fail*. 2015;8(2):384-409.

50. David G, Smith-McLallen A, Ukert B. The effect of predictive analytics-driven interventions on healthcare utilization. *J Health Econ*. 2019;64:68-79.

CHAPTER 5

5. Summary and Conclusion

5.1 Summary of Findings and Discussion

Machine learning approaches to modeling of epidemiologic and healthcare data are becoming very common. In this dissertation, we applied natural language processing, unsupervised machine learning algorithms, specifically topic modeling, to identify research gaps in the published literature, supervised machine learning algorithms to accurately predict the diagnosis of incident HF and healthcare utilization among postmenopausal women. Although the purpose of machine learning algorithms is “prediction” rather than predictor-outcome relationships, we “unboxed” the algorithms with interpretable machine learning techniques.

Women over 50 years, about the age of natural menopause, are at increased risk for cardiovascular disease including HF due to a decline in the natural hormone estrogen, which has been shown to be cardio-protective in women[1]. In 2017, 1 in every 5 female deaths were due to CVD[2]. Specifically, heart failure (HF) is a chronic, progressive condition accounts for 35% of all CVD deaths among women[3].

Understudied research topics in the literature of heart failure (HF) among women

Although studies report significant sex differences in HF etiology, risk factors, and HF disease burden, women are underrepresented in HF-related clinical trials[4,5] and observational studies, which may result in significant knowledge gaps in women-specific HF research. Utilizing unsupervised machine learning methods, our study identified knowledge gaps in the literature of heart failure (HF) among women. Based on the published HF studies in PubMed between 1959 until 3 December 2019, the top three most understudied topics were (1) atrial fibrillation, (2) systolic and diastolic dysfunction, and (3) left ventricular ejection fraction phenotypes. The co-occurrence of atrial fibrillation and HF is common in clinical practice[6] and may lead to worse symptoms, poorer prognosis, high healthcare utilization, and all-cause mortality[7–10]. Nevertheless, our analysis revealed that treatments and interventions specific to those with HF and atrial fibrillation have not been well-studied in the literature as the prior research in this area focused on the epidemiology of atrial fibrillation, role of natriuretic peptide, and risk of stroke in patients with atrial fibrillation and HF.

Substantial knowledge gaps in the literature of HF among postmenopausal women

In our study, we only identified 77 articles on HF in postmenopausal women compared to 32,946 in women in general. Among the 77 articles, the most understudied topic was stress-induced cardiomyopathy, which can be due to the rarity of this condition. However, stress-induced cardiomyopathy is more common in women than men[11–15]. Other understudied areas were about the effect of breast cancer and chemotherapy on HF and the incidence of HF in postmenopausal women.

Leading predictors of incident HF among postmenopausal women

Our review also identified only 7 studies that have exclusively focused on incident HF among postmenopausal women[3,16–21] with 3 studies using data from Women's Health Initiative (WHI)[16,18]. While these studies have shed light on modifiable and non-modifiable risk factors, emerging evidence from case reports and observational studies suggest that some prescription medications (e.g., oral antidiabetics, antibiotics, and antiepileptic medications) may confer a high risk for HF. Therefore, we examined the risk of incident HF among postmenopausal women with a comprehensive list of risk factors and several machine learning approaches (cross-validated logistic regression, random forest, extreme Gradient Boosting (XGBoost)) using a commercial insurance claims database.

In our cohort study, 2.1% of postmenopausal women developed HF during the 2-year follow-up period consistent with published studies[22]. Polypharmacy, older age, and arrhythmia were consistent predictors of incident HF across all machine learning algorithms. In addition to established risk factors, we identified some novel predictors of incident HF among postmenopausal women. For example, polypharmacy ranked 3rd, after older age and arrhythmia, as a leading predictor of incident HF. Although not a leading predictor, sulfonylurea use predicted incident HF. Antibiotic use other than fluoroquinolones was identified as a predictor in one of the three machine learning models.

Identification of HF patients at high risk for heart failure-related emergency room use (HF-related ER use)

ER use is associated with negative health outcomes[23]. Specifically, HF is considered as an ambulatory sensitive condition and some ER visits may be preventable[24]. Therefore, we analyzed predictors of HF-related ER use among postmenopausal women using a large commercial insurance claims database and random forest for classification, a machine learning algorithm. Findings from our study have indicated that the number of HF-related ER visits at baseline, fragmented care, age, insurance type (Health Maintenance Organization)), and coronary artery disease were the key predictors of HF-related ER use among postmenopausal women. These predictors, except HMO, were found to be positively associated with HF-related ER use.

5.2 Implications and Suggestions for Future Research

Our study findings unveiled the gaps in HF research among women and highlight the need for research focusing on the treatment and management of women who concomitantly have atrial fibrillation and HF. Given the small proportion of articles published on HF among postmenopausal women and unique characteristics of this population, future research should study postmenopausal women and leverage big data and electronic health records. Conducting studies focusing on postmenopausal women can enhance our understanding of the needs of this population and improve their health outcomes.

Furthermore, results of this study underscore the importance of medication management among postmenopausal women. Given the high prevalence of polypharmacy and its negative effects on HF risk among postmenopausal women, our results have implications for promoting evidence-based methods to reduce polypharmacy such as medication utilization review and patient education. Although this study identified some prescription medications (i.e., sulfonylureas and antibiotics other than fluoroquinolones) as predictors of incident HF, these findings need to be confirmed in future research.

Our machine learning models were able to identify HF patients at high risk for ER use with high predictive accuracy. This suggests the use of predictive analytics in identifying high-risk ER use patients. Identification of those at high risk for ER use can assist payers and policymakers to tailor interventions that could decrease ER use and improve health outcomes. As the top two predictors of HF-related ER use were healthcare utilization features (i.e., number of HF-related ER visits at baseline and fragmented care), our findings have implications for implementing interventions that can reduce fragmented care and ER utilization (e.g., telehealth).

A novel and unique contribution of our study is the application of machine learning methods. Findings from all three studies suggest that machine learning algorithms can achieve comparable and, in some cases, better predictive accuracy compared to traditional statistical models. Our study on research gaps in women with HF confirmed the feasibility of using unsupervised machine learning methods (i.e., topic modeling). Our hybrid method was not only more comprehensive but less time-consuming than the expert-based manual literature review method. Even though when only used one database (i.e., PubMed), our approach is promising and effective for the discovery of knowledge gaps in medical research. Future research should collect data from multiple databases to capture all published articles in the literature. In terms of supervised machine learning methods, our findings have shown better predictive abilities of machine learning methods compared to traditional methods.

Moreover, this study used interpretable machine learning techniques (i.e., SHapley Additive exPlanations (SHAP) and partial dependence plots) to explain the association between study features and the target feature. With such high predictive abilities and enhancement in the interpretability of machine learning algorithms, the use of machine learning methods may continue to expand in the HF area.

5.3 Strengths and Limitations

This present study has several strengths: 1) it is the first study to identify knowledge gaps in HF research among women, especially postmenopausal women, using unsupervised machine learning methods and articles published in PubMed database; 2) use of NLP and text mining techniques to screen and identify relevant articles and extract the objective(s) of each study from PubMed abstract; 3) use of nationally representative real-world data of commercially insured postmenopausal women (aged ≥ 50 years); 4) use of a retrospective cohort study design to track postmenopausal women over time; 4) including a comprehensive set of risk factors (e.g., polypharmacy and specific prescription medications); and 5) use of several machine learning classifiers to increase the rigor, robustness, and precision of our investigation.

In contrast, this study has some potential limitations. First, no evaluation metrics were used to assess the accuracy of clusters yielded from the unsupervised machine learning model. To overcome this limitation, three investigators familiar with HF research independently validating and labeling clusters yielded from our model. In addition, we only searched one database (i.e., PubMed) to retrieve HF articles, which might impact on the number of articles included in this study. Our data lacked some important variables including clinical factors (e.g., type and severity of HF, laboratory findings, and severity of chronic conditions), socioeconomic characteristics (e.g., income and education), and race. Not including these variables might influence the performance of our models.

5.4 Conclusion

In the HF research area, women, specifically postmenopausal women, are understudied. The co-occurrence of atrial fibrillation with HF in women and stress-induced cardiomyopathy in postmenopausal women are the most understudied topics in the literature. Among postmenopausal women, polypharmacy was identified as a major risk factor for incident HF; Among postmenopausal women with HF, the number of HF-related ER use at baseline and fragmented care were the top two predictors of the HF-related ER use in the subsequent year.

Collectively, our study findings identified risk factors that can be modified to reduce the risk of incident HF and suboptimal utilization (ER visits) of healthcare resources. Furthermore, our studies highlighted the usefulness of machine learning methods as promising tools in health outcomes research. These methods outperform traditional methods (e.g., expert-based manual literature review and statistical methods). With the ongoing enhancement in the interpretability of machine learning methods, the adoption of these methods may increase in future HF research.

5.5 References

1. Iorga A, Cunningham CM, Moazeni S, Ruffenach G, Umar S, Eghbali M. The protective role of estrogen and estrogen receptors in cardiovascular disease and the controversial use of estrogen therapy. *Biol Sex Differ*. 2017;8(1):33. doi:10.1186/s13293-017-0152-8
2. Women and Heart Disease. Centers for Disease Control and Prevention (CDC) website. Accessed August 03, 2020. https://www.cdc.gov/heartdisease/women.htm
3. Appiah D, Schreiner PJ, Demerath EW, Loehr LR, Chang PP, Folsom AR. Association of Age at Menopause With Incident Heart Failure: A Prospective Cohort Study and Meta-Analysis. *J Am Heart Assoc*. 2016;5(8). doi:10.1161/JAHA.116.003769
4. Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women, and minorities in heart failure clinical trials. *Arch Intern Med*. 2002;162(15):1682-1688. doi:10.1001/archinte.162.15.1682
5. Tahhan AS, Vaduganathan M, Greene SJ, et al. Enrollment of Older Patients, Women, and Racial and Ethnic Minorities in Contemporary Heart Failure Clinical Trials: A Systematic Review. *JAMA Cardiol*. 2018;3(10):1011-1019. doi:10.1001/jamacardio.2018.2559
6. Anter E, Jessup M, Callans DJ. Atrial fibrillation and heart failure: treatment considerations for a dual epidemic. *Circulation*. 2009;119(18):2516-2525.
7. Kotecha D, Piccini JP. Atrial fibrillation in heart failure: what should we do? *Eur Heart J*. 2015;36(46):3250-3257. doi:10.1093/eurheartj/ehv513
8. Chamberlain AM, Redfield MM, Alonso A, Weston SA, Roger VL. Atrial fibrillation and mortality in heart failure: a community study. *Circ Heart Fail*. 2011;4(6):740-746. doi:10.1161/CIRCHEARTFAILURE.111.962688

9. Wang TJ, Larson MG, Levy D, et al. Temporal relations of atrial fibrillation and congestive heart failure and their joint influence on mortality: the Framingham Heart Study. *Circulation*. 2003;107(23):2920-2925.

10. Zareba W, Steinberg JS, McNitt S, et al. Implantable cardioverter-defibrillator therapy and risk of congestive heart failure or death in MADIT II patients with atrial fibrillation. *Hear Rhythm*. 2006;3(6):631-637.

11. Kurowski V, Kaiser A, von Hof K, et al. Apical and midventricular transient left ventricular dysfunction syndrome (tako-tsubo cardiomyopathy) frequency, mechanisms, and prognosis. *Chest*. 2007;132(3):809-816.

12. Deshmukh A, Kumar G, Pant S, Rihal C, Murugiah K, Mehta JL. Prevalence of Takotsubo cardiomyopathy in the United States. *Am Heart J*. 2012;164(1):66-71.

13. Bybee KA, Kara T, Prasad A, et al. Systematic Review: Transient Left Ventricular Apical Ballooning: A Syndrome That Mimics ST-Segment Elevation Myocardial Infarction. *Ann Intern Med*. 2004;141(11):858-865. doi:10.7326/0003-4819-141-11-200412070-00010

14. Akashi YJ, Goldstein DS, Barbaro G, Ueyama T. Takotsubo cardiomyopathy: a new form of acute, reversible heart failure. *Circulation*. 2008;118(25):2754-2762.

15. Sharkey SW, Lesser JR, Zenovich AG, et al. Acute and reversible cardiomyopathy provoked by stress in women from the United States. *Circulation*. 2005;111(4):472-479.

16. Hall PS, Nah G, Howard B V, et al. Reproductive Factors and Incidence of Heart Failure Hospitalization in the Women’s Health Initiative. *J Am Coll Cardiol*. 2017;69(20):2517-2526. doi:10.1016/j.jacc.2017.03.557

17. LaMonte MJ, Manson JE, Chomistek AK, et al. Physical Activity and Incidence of Heart Failure in Postmenopausal Women. *JACC Heart Fail*. 2018;6(12):983-995.

doi:10.1016/j.jchf.2018.06.020

18. Eaton CB, Abdulbaki AM, Margolis KL, et al. Racial and ethnic differences in incident hospitalized heart failure in postmenopausal women: the Women's Health Initiative. *Circulation*. 2012;126(6):688-696.

19. Rahman I, Åkesson A, Wolk A. Relationship between age at natural menopause and risk of heart failure. *Menopause*. 2015;22(1):12-16.

20. Ebong IA, Watson KE, Goff Jr DC, et al. Age at menopause and incident heart failure: the Multi-Ethnic Study of Atherosclerosis. *Menopause (New York, NY)*. 2014;21(6):585.

21. Bibbins-Domingo K, Lin F, Vittinghoff E, et al. Predictors of heart failure among women with coronary disease. *Circulation*. 2004;110(11):1424-1430. doi:10.1161/01.CIR.0000141726.01302.83

22. Ho KKL, Pinsky JL, Kannel WB, Levy D. The epidemiology of heart failure: the Framingham Study. *J Am Coll Cardiol*. 1993;22(4 Supplement 1):A6--A13.

23. Schnitker L, Martin-Khan M, Beattie E, Gray L. Negative health outcomes and adverse events in older people attending emergency departments: a systematic review. *Australas Emerg Nurs J*. 2011;14(3):141-162.

24. AHRQ quality indicators—guide to prevention quality indicators: hospital admission for ambulatory care sensitive conditions. Publication No: 02-R0203. 2001. Agency for Healthcare Research and Quality website. Accessed August 02, 2020. https://www.ahrq.gov/downloads/pub/ahrqqi/pqiguide.pdf

6. Appendices

Appendix 6.1
Python Codes for Topic Modeling on the literature of HF among Women

```
#import libraries
import re
import string
import pandas as pd
import numpy as np
import datetime

#search pubmed and get the count of articles via python using Biopython
from Bio import Entrez
Entrez.email = 'khaled_al-hussain@hotmail.com'
handle = Entrez.egquery(term="(heart failure[MeSH Terms] OR congestive heart failure[MeSH Terms] OR cardiac
failure[MeSH Terms] OR ejection fraction AND hasabstract[text] AND Humans[Mesh] AND Female[MeSH
Terms])")
record =Entrez.read(handle)
for row in record ['eGQueryResult']:
   if row['DbName']=='pubmed':
      record_count = (row["Count"])
      print(record_count) #we can compare this count to the count we get from the website

#search pubmed and get the count of articles via python using Biopython
from Bio import Entrez
Entrez.email = 'khaled_al-hussain@hotmail.com'
handle = Entrez.egquery(term="(heart failure[MeSH Terms] OR congestive heart failure[MeSH Terms] OR cardiac
failure[MeSH Terms] OR ejection fraction AND hasabstract[text] AND Humans[Mesh] AND Female[MeSH
Terms])")
record =Entrez.read(handle)
for row in record ['eGQueryResult']:
   if row['DbName']=='pubmed':
      record_count = (row["Count"])
      print(record_count) #we can compare this count to the count we get from the website

#retrieve IDs of all articles
handle = Entrez.esearch(db='pubmed', term="(heart failure[MeSH Terms] OR congestive heart failure[MeSH
Terms] OR cardiac failure[MeSH Terms] OR ejection fraction AND hasabstract[text] AND Humans[Mesh] AND
Female[MeSH Terms])", retmax = 300000)
record =Entrez.read(handle)
handle.close()
idlist = record["IdList"]
print(idlist)
len(idlist) #double check

#divde id lists into multiple files
record_count = int(record_count)
file_count = (record_count/10000)

idlist_1 = idlist[:10000]
idlist_2 = idlist[10000:20000]
idlist_3 = idlist[20000:30000]
idlist_4 = idlist[30000:40000]
idlist_5 = idlist[40000:50000]
idlist_6 = idlist[50000:60000]
idlist_7 = idlist[60000:70000]
```

```
#import necessary packages for retrieving the content of articles
from Bio import Medline

#retrieve the 1st 10000 articles
handle_1 = Entrez.efetch(db="pubmed", id = idlist_1, rettype = "medline", retmode = "text")
records_1 = Medline.parse(handle_1)
records_1 = list(records_1)
#save titles & abstracts in a txt file - 1st 10000 articles
with open("//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations1_fem.txt", "w") as file1:
    for record in records_1:
        x = record.get("AB", "?")
        x = x.lower().split(".")
        str1 = ''.join(x)
        y = re.split('methods: |methods:: |material and methods: |material and methods |materials and methods: 
|materials and methods |methods and materials: | methods and materials |methods & materials: |methods & materials 
|patients and materials |patients and methods |methods and results: |study design: |design: |patients: |participants: 
settings:| setting',str1)
        obj = y[0]
        obj_1 = obj.translate(str.maketrans('', '', string.punctuation))
        file_1 = record.get("PMID", "?") + " &&& " + record.get("TI", "?") + " &&& "  + record.get("AB", "?") + " 
&&& " + obj_1 + "\n"
        file1.write(str(file_1))

#retrieve the 2nd 10000 articles
handle_2 = Entrez.efetch(db="pubmed", id = idlist_2, rettype = "medline", retmode = "text")
records_2 = Medline.parse(handle_2)
records_2 = list(records_2)
#save titles & abstracts in a txt file - 2nd 10000 articles
with open("//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations2_fem.txt", "w") as file2:
    for record in records_2:
        x = record.get("AB", "?")
        x = x.lower().split(".")
        str1 = ''.join(x)
        y = re.split('methods: |methods:: |material and methods: |material and methods |materials and methods: 
|materials and methods |methods and materials: | methods and materials |methods & materials: |methods & materials 
|patients and materials |patients and methods |methods and results: |study design: |design: |patients: |participants: 
settings:| setting',str1)
        obj = y[0]
        obj_2 = obj.translate(str.maketrans('', '', string.punctuation))
        file_2 = record.get("PMID", "?") + " &&& " + record.get("TI", "?") + " &&& "  + record.get("AB", "?") + " 
&&& " + obj_2 + "\n"
        file2.write(str(file_2))

#retrieve the 3rd 10000 articles
handle_3 = Entrez.efetch(db="pubmed", id = idlist_3, rettype = "medline", retmode = "text")
records_3 = Medline.parse(handle_3)
records_3 = list(records_3)
#save titles & abstracts in a txt file - 3rd 10000 articles
with open("//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations3_fem.txt", "w") as file3:
    for record in records_3:
        x = record.get("AB", "?")
        x = x.lower().split(".")
        str1 = ''.join(x)
        y = re.split('methods: |methods:: |material and methods: |material and methods |materials and methods: 
|materials and methods |methods and materials: | methods and materials |methods & materials: |methods & materials
|patients and materials |patients and methods |methods and results: |study design: |design: |patients: |participants:
settings:| setting',str1)
        obj = y[0]
        obj_3 = obj.translate(str.maketrans(", ", string.punctuation))
        file_3 = record.get("PMID", "?") + " &&& " + record.get("TI", "?") + " &&& "  + record.get("AB", "?") + "
&&& " + obj_3 + "\n"
        file3.write(str(file_3))

#retrieve the 4th 10000 articles
handle_4 = Entrez.efetch(db="pubmed", id = idlist_4, rettype = "medline", retmode = "text")
records_4 = Medline.parse(handle_4)
records_4 = list(records_4)
#save titles & abstracts in a txt file - 4th 10000 articles
with open("//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations4_fem.txt", "w") as file4:
    for record in records_4:
        x = record.get("AB", "?")
        x = x.lower().split(".")
        str1 = ''.join(x)
        y = re.split('methods: |methods:: |material and methods: |material and methods |materials and methods:
|materials and methods |methods and materials: | methods and materials |methods & materials: |methods & materials
|patients and materials |patients and methods |methods and results: |study design: |design: |patients: |participants:
settings:| setting',str1)
        obj = y[0]
        obj_4 = obj.translate(str.maketrans(", ", string.punctuation))
        file_4 = record.get("PMID", "?") + " &&& " + record.get("TI", "?") + " &&& "  + record.get("AB", "?") + "
&&& " + obj_4 + "\n"
        file4.write(str(file_4))

#retrieve the 5th 10000 articles
handle_5 = Entrez.efetch(db="pubmed", id = idlist_5, rettype = "medline", retmode = "text")
records_5 = Medline.parse(handle_5)
records_5 = list(records_5)
#save titles & abstracts in a txt file - 5th 10000 articles
with open("//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations5_fem.txt", "w") as file5:
    for record in records_5:
        x = record.get("AB", "?")
        x = x.lower().split(".")
        str1 = ''.join(x)
        y = re.split('methods: |methods:: |material and methods: |material and methods |materials and methods:
|materials and methods |methods and materials: | methods and materials |methods & materials: |methods & materials
|patients and materials |patients and methods |methods and results: |study design: |design: |patients: |participants:
settings:| setting',str1)
        obj = y[0]
        obj_5 = obj.translate(str.maketrans(", ", string.punctuation))
        file_5 = record.get("PMID", "?") + " &&& " + record.get("TI", "?") + " &&& "  + record.get("AB", "?") + "
&&& " + obj_5 + "\n"
        file5.write(str(file_5))

#retrieve the 6th 10000 articles
handle_6 = Entrez.efetch(db="pubmed", id = idlist_6, rettype = "medline", retmode = "text")
records_6 = Medline.parse(handle_6)
records_6 = list(records_6)
#save titles & abstracts in a txt file - 6th 10000 articles
with open("//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations6_fem.txt", "w") as file6:
    for record in records_6:
        x = record.get("AB", "?")
        x = x.lower().split(".")
        str1 = ''.join(x)
        y = re.split('methods: |methods:: |material and methods: |material and methods |materials and methods:
|materials and methods |methods and materials: | methods and materials |methods & materials: |methods & materials
|patients and materials |patients and methods |methods and results: |study design: |design: |patients: |participants:
settings:| setting',str1)
        obj = y[0]
        obj_6 = obj.translate(str.maketrans('', '', string.punctuation))
        file_6 = record.get("PMID", "?") + " &&& " + record.get("TI", "?") + " &&& "  + record.get("AB", "?") + "
&&& " + obj_6 + "\n"
        file6.write(str(file_6))

#retrieve the 7th 10000 articles
handle_7 = Entrez.efetch(db="pubmed", id = idlist_7, rettype = "medline", retmode = "text")
records_7 = Medline.parse(handle_7)
records_7 = list(records_7)
#save titles & abstracts in a txt file - 7th 10000 articles
with open("//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations7_fem.txt", "w") as file7:
    for record in records_7:
        x = record.get("AB", "?")
        x = x.lower().split(".")
        str1 = ''.join(x)
        y = re.split('methods: |methods:: |material and methods: |material and methods |materials and methods:
|materials and methods |methods and materials: | methods and materials |methods & materials: |methods & materials
|patients and materials |patients and methods |methods and results: |study design: |design: |patients: |participants:
settings:| setting',str1)
        obj = y[0]
        obj_7 = obj.translate(str.maketrans('', '', string.punctuation))
        file_7 = record.get("PMID", "?") + " &&& " + record.get("TI", "?") + " &&& "  + record.get("AB", "?") + "
&&& " + obj_7 + "\n"
        file7.write(str(file_7))

#combine all content files
filenames1 = ["//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations1_fem.txt",
"//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations2_fem.txt", "//Users/khalidabdullah
1/Desktop/Health Outcomes Research/SLR/citations3_fem.txt"]
with open("//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations1-3_fem.txt", "w") as outfile:
    for fname in filenames1:
        with open(fname) as infile:
            for line in infile:
                outfile.write(line)

#combine all content files
filenames2 = ["//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations4_fem.txt",
"//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations5_fem.txt", "//Users/khalidabdullah
1/Desktop/Health Outcomes Research/SLR/citations6_fem.txt", "//Users/khalidabdullah 1/Desktop/Health
Outcomes Research/SLR/citations7_fem.txt"]
with open("//Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/citations4-7_fem.txt", "w") as outfile:
    for fname in filenames2:
        with open(fname) as infile:
            for line in infile:
                outfile.write(line)


#convert text into a dataframe
from io import StringIO
```

```
import pandas as pd
content_data = StringIO("""PMID&&&title&&&abstract&&&objective
""")

df = pd.read_csv(content_data, sep="&&&")
path="/Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/df1_fem.csv"
df_csv=df.to_csv(path)

#import all csv datasets
import pandas as pd
df1 = pd.read_csv("/Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/df1_fem.csv")
df2 = pd.read_csv("/Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/df2_fem.csv")

#merge all dataframs
df_all = pd.concat([df1, df2])

#check # of articles
df_all

#remove whitespaces
df_all['abstract'] = df_all['abstract'].apply(lambda x : x.strip().lower())
#exclude citations without abstract
df_all = df_all[df_all['abstract'] != "?"]
#check # of articles after removing citations with no abstracts
df_all

# drop duplicate values by title
df_all.drop_duplicates(subset="title", keep='first', inplace=True)
#check no. of articles after removing duplicates
df_all

#Detect missing values
missing_data = df_all.isnull()

for column in missing_data. columns.values.tolist():
            print (column)
            print (missing_data[column].value_counts())
            print (" ")

#replace missing values in objective with abstract
df_all['objective'].fillna(df_all.abstract, inplace = True)

#remove punctuations from titles
import string
df_all['title'] = df_all['title'].apply(lambda x : x.strip().capitalize().translate(str.maketrans('', '', string.punctuation)))

#add new columns for preprocessed titles, abstracts & objectives
preprocessed_title = df_all['title']
df_all['preprocessed_title'] = preprocessed_title
preprocessed_abstract = df_all['abstract']
df_all['preprocessed_abstract'] = preprocessed_abstract
preprocessed_objective = df_all['objective']
df_all['preprocessed_objective'] = preprocessed_objective

#text preprocessing
#remove whitespaces and do lowercase
```

```
df_all['preprocessed_title'] = df_all['preprocessed_title'].apply(lambda x : x.strip().lower())
df_all['preprocessed_abstract'] = df_all['preprocessed_abstract'].apply(lambda x : x.strip().lower())
df_all['preprocessed_objective'] = df_all['preprocessed_objective'].apply(lambda x : x.strip().lower())

#remove punctuations
df_all['preprocessed_abstract'] = df_all['preprocessed_abstract'].apply(lambda x : x.translate(str.maketrans('', '',
string.punctuation)))
#remove numbers
df_all['preprocessed_title'] = df_all['preprocessed_title'].str.replace('\d+', '')
df_all['preprocessed_abstract'] = df_all['preprocessed_abstract'].str.replace('\d+', '')
df_all['preprocessed_objective'] = df_all['preprocessed_objective'].str.replace('\d+', '')

#remove stopwords
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
#add more stopwords
stop_words = set(stopwords.words('english'))
stop_words.add('background')
stop_words.add('introduction')
stop_words.add('aims')
stop_words.add('aim')
stop_words.add('aimed')
stop_words.add('purpose')
stop_words.add('objectives')
stop_words.add('objective')
stop_words.add('methods')
stop_words.add('analysis')
stop_words.add('analyses')
stop_words.add('results')
stop_words.add('finding')
stop_words.add('findings')
stop_words.add('discussion')
stop_words.add('discussions')
stop_words.add('conclusion')
stop_words.add('conclusions')
stop_words.add('case')
stop_words.add('cases')
stop_words.add('study')
stop_words.add('studies')
stop_words.add('patient')
stop_words.add('patients')
stop_words.add('subject')
stop_words.add('subjects')
stop_words.add('disease')
stop_words.add('diseases')
stop_words.add('report')
stop_words.add('reports')
stop_words.add('group')
stop_words.add('groups')
stop_words.add('use')
stop_words.add('uses')
stop_words.add('using')
stop_words.add('used')
stop_words.add('analyze')
stop_words.add('analyzes')
stop_words.add('analyzed')
stop_words.add('clinical')
stop_words.add('show')
stop_words.add('shows')
stop_words.add('showed')
stop_words.add('shown')
stop_words.add('examine')
stop_words.add('examines')
stop_words.add('examined')
stop_words.add('investigate')
stop_words.add('investigates')
stop_words.add('investigated')
stop_words.add('determine')
stop_words.add('determines')
stop_words.add('determined')
stop_words.add('assess')
stop_words.add('assesses')
stop_words.add('assessed')
stop_words.add('evaluate')
stop_words.add('evaluates')
stop_words.add('evaluated')
stop_words.add('measure')
stop_words.add('measures')
stop_words.add('measured')
stop_words.add('sought')
stop_words.add('compare')
stop_words.add('compares')
stop_words.add('compared')
stop_words.add('observe')
stop_words.add('observes')
stop_words.add('observed')
stop_words.add('reveal')
stop_words.add('reveals')
stop_words.add('revealed')
stop_words.add('day')
stop_words.add('days')
stop_words.add('week')
stop_words.add('weeks')
stop_words.add('month')
stop_words.add('months')
stop_words.add('year')
stop_words.add('years')
stop_words.add('yearold')
stop_words.add('significantly')
stop_words.add('significant')
stop_words.add('review')
stop_words.add('data')
stop_words.add('normal')
stop_words.add('confidence')
stop_words.add('interval')
stop_words.add('increase')
stop_words.add('increases')
stop_words.add('increased')
stop_words.add('high')
stop_words.add('higher')
stop_words.add('highest')
stop_words.add('low')
stop_words.add('lower')
stop_words.add('lowest')
stop_words.add('decrease')
stop_words.add('decreases')
stop_words.add('decreased')
stop_words.add('change')
stop_words.add('changes')
stop_words.add('changed')
stop_words.add('plus')
stop_words.add('publication')
stop_words.add('publications')
stop_words.add('need')
stop_words.add('needs')
stop_words.add('different')
stop_words.add('differences')
stop_words.add('difference')
stop_words.add('association')
stop_words.add('associations')
stop_words.add('associated')
stop_words.add('relationship')
stop_words.add('relationships')
stop_words.add('related')
stop_words.add('known')
stop_words.add('unknown')
stop_words.add('clear')
stop_words.add('unclear')

df_all['preprocessed_title'] = df_all['preprocessed_title'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop_words)]))
df_all['preprocessed_abstract'] = df_all['preprocessed_abstract'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop_words)]))
df_all['preprocessed_objective'] = df_all['preprocessed_objective'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop_words)]))

#tokenize
from nltk.tokenize import word_tokenize
df_all['preprocessed_title'] = df_all['preprocessed_title'].apply(word_tokenize)
df_all['preprocessed_abstract'] = df_all['preprocessed_abstract'].apply(word_tokenize)
df_all['preprocessed_objective'] = df_all['preprocessed_objective'].apply(word_tokenize)

#lemmatize
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()

def word_lemmatizer(text):
    lem_text = [lemmatizer.lemmatize(i) for i in text]
    return lem_text

df_all['preprocessed_title'] = df_all['preprocessed_title'].apply(lambda x: word_lemmatizer(x))
df_all['preprocessed_abstract'] = df_all['preprocessed_abstract'].apply(lambda x: word_lemmatizer(x))
df_all['preprocessed_objective'] = df_all['preprocessed_objective'].apply(lambda x: word_lemmatizer(x))


#drop index and delete unncessary column from the dataframe
df_all=df_all.reset_index(drop=True)
```

```
df_all = df_all.drop("Unnamed: 0", axis=1)
df_all

#export the merged and preprocessed dataframe
path = "/Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/df_nodup_fem_clean.csv"
df_all_csv = df_all.to_csv(path)


#import all csv datasets
df_all = pd.read_csv("/Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/df_nodup_fem_clean.csv")

#filter by keywords
df_all['HF'] = np.where(df_all.objective.str.contains('heart failure'), 1,
                    np.where(df_all.objective.str.contains('HF'), 1,
                    np.where(df_all.objective.str.contains('cardiac failure'), 1,
                    np.where(df_all.objective.str.contains('congestive heart failure'), 1,
                    np.where(df_all.objective.str.contains('CHF'), 1,
                        0)))))

df_HF_all = df_all[df_all.HF == 1]

#export the merged and preprocessed dataframe
path = "/Users/khalidabdullah 1/Desktop/Health Outcomes Research/SLR/df_nodup_hf_all_hfilter.csv"
df_all_csv = df_HF_all.to_csv(path)

#import
df_HF_all = pd.read_csv("/Users/khalidabdullah 1/Desktop/Health Outcomes
Research/SLR/df_nodup_hf_all_hfilter.csv")

#topic modeling
#import
df_HF_all = pd.read_csv("/Users/khalidabdullah 1/Desktop/Health Outcomes
Research/SLR/df_nodup_hf_all_hfilter.csv")

begin_time = datetime.datetime.now()
#NMF model
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_vect = TfidfVectorizer(ngram_range=(1,2), max_df=0.8, min_df=2, stop_words='english')
doc_term_matrix_1 = tfidf_vect.fit_transform(df_HF_all['preprocessed_objective'].values.astype('U'))

from sklearn.decomposition import NMF
nmf = NMF(n_components=15, random_state=42)
nmf.fit(doc_term_matrix_1)

for i,topic in enumerate(nmf.components_):
    print(f'Top 20 words for topic #{i}:')
    print([tfidf_vect.get_feature_names()[i] for i in topic.argsort()[-20:]])
    print('\n')

#add the topics to the dataset and displays the first five rows:
topic_values = nmf.transform(doc_term_matrix_1)
df_HF_all['Topic'] = topic_values.argmax(axis=1)
df_HF_all.head()

print(datetime.datetime.now() - begin_time)
```

```
#random articles from each group
tp_0 = df_HF_all[df_HF_all.Topic == 0]
tp_0.sample(40)
tp_0 = tp_0.sample(40)
#export
path = "/Users/khalidabdullah 1/Desktop/Health Outcomes Research/Aim1/individual clusters/topic0_40rs.csv"
df_all_csv = tp_0.to_csv(path)

tp_2 = df_HF_all[df_HF_all.Topic == 2]
tp_2.sample(40)
tp_2 = tp_2.sample(40)
#export
path = "/Users/khalidabdullah 1/Desktop/Health Outcomes Research/Aim1/individual clusters/topic2_40rs.csv"
df_all_csv = tp_2.to_csv(path)

tp_6 = df_HF_all[df_HF_all.Topic == 6]
tp_6.sample(40)
tp_6 = tp_6.sample(40)
#export
path = "/Users/khalidabdullah 1/Desktop/Health Outcomes Research/Aim1/individual clusters/topic6_40rs.csv"
df_all_csv = tp_6.to_csv(path)
```

Appendix 6.2
ICD-9-CM and ICD-10-CM Codes for Identifying Heart Failure and Heart Valve Disorders

Diagnosis	ICD-9-CM code	ICD-10-CM code
Heart failure	428	I50
Mitral valve insufficiency and aortic valve insufficiency	396.3	I08.0
Multiple involvement of mitral and aortic valves	396.8	I08.8
Mitral and aortic valve diseases, unspecified	396.9	I08.9
Other and unspecified mitral valve diseases	394.9	I05.8
Mitral valve disorders	424.0	I34
Aortic valve disorders	424.1	I35

Appendix 6.3
Measurements of Prescription Medications

Medication name	Measurements (Code type)	Identification codes
Oral antidiabetics		
Metformin	NDCs	metformin.xlsx
Sulfonylureas (Glimepiride, Glipizide, Glyburide, Tolbutamide, Tolazamide, Chlorpropamide)	NDCs	sulfonylureas.xlsx
DPP-4 inhibitors (Sitagliptin, Saxagliptin, Alogliptin, and Linagliptin)	NDCs	DPP4_inhibitors.xlsx
Antiepileptics		
Pregabalin	NDCs	pregabalin.xlsx
Gabapentin	NDCs	gabapentin.xlsx
Antibiotics		
Fluoroquinolones	AHFS classification	‘081218’
Other antibiotics	AHFS classification	'520404', '081202', '081206', '081207', '081208', '081212', '081216', '081220', '081224', '081228', '082400'
Heart Failure Medications		
ACE inhibitors	AHFS classification	‘243204’
Beta-blockers	AHFS classification	‘242400’
ARBs	AHFS classification	‘243208’
Diuretics	AHFS classification	‘402800’, ‘402808’, ‘402810’, ‘402812’, ‘402816’, ‘402820’, ‘402824’, ‘402892’
Other medications		
Antihyperlipidemic medications	AHFS classification	‘240600’,’ 240604’, 240605’,’ 240606’,’ 240608’, ‘240692’

Abbreviations: NDCs: National Drug Codes; AHFS: American Hospital Formulary Service; DPP-4 inhibitors: Dipeptidyl Peptidase-4 inhibitors; ACE inhibitors: Angiotensin-converting-enzyme inhibitors; ARBs: Angiotensin II receptor blockers.

Appendix 6.4

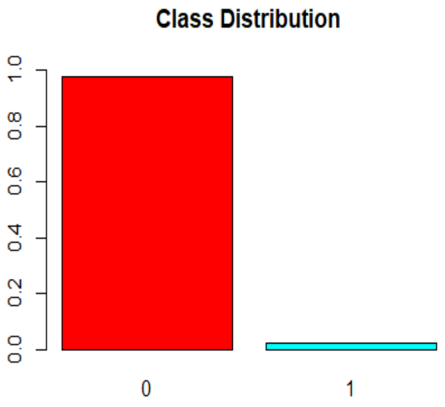
R Codes for Machine Learning Algorithms to Identify Predictors of Incident Heart Failure among Postmenopausal Women

A. CVLR Algorithm: Predictors of Incident HF

```
#read data
library(haven)
#read sas data---must install package haven and load library haven#
df <-
read_sas("Z:/OPTUM_10pct/projects/Khalid_phd/Aim_2/sasdata/hfree_2007_2016_n.sas7bdat
", NULL)

hf <-
df[c('hf_fu12','abrx_3grp','antiep_grp','metrx_any','tzd_any','dpp4_any','sulf_any','
age_3grp','polyrx_gn_ge6','anyabuse','ins_mcare','hmo','region_grp4','er_nbr',
'anx_any','bipolar','psycho','deprn','schiz','ipot_arth','ipot_asth','ipot_cancer','i
pot_cad','ipot_mi','sleep','obesity','ipot_c_arrhy','ipot_ckd','ipot_copd','ipot_deme
ntia','ipot_hepatitis','ipot_hilipid','ipot_htn','ipot_diabetes','ipot_stroke','ipot_
osteop')]

#convert to factor variable---for logistic regression code the dv as 0(no) and 1
(yes)#
library(plyr)
hf$hf_fu12  <-factor(hf$hf_fu12)
hf$hf_fu12 <- revalue(hf$hf_fu12, c("1"= "1", "2"= "0")) #changing label 2 to 0
hf$hf_fu12 <- relevel(hf$hf_fu12, ref = "0") #changing reference category for log reg
summary(hf$hf_fu12)
##      0      1
## 149379   3213
table(hf$hf_fu12)
##
##      0      1
## 149379   3213
#recode indep variables to indicate categorical status to R#
hf$abrx_3grp    <-factor(hf$abrx_3grp)
hf$antiep_grp   <-factor(hf$antiep_grp)
hf$metrx_any    <-factor(hf$metrx_any)
hf$tzd_any  <-factor(hf$tzd_any)
hf$dpp4_any     <-factor(hf$dpp4_any)
hf$sulf_any <-factor(hf$sulf_any)
hf$age_3grp <-factor(hf$age_3grp)
hf$polyrx_gn_ge6 <-factor(hf$polyrx_gn_ge6)
hf$anyabuse <-factor(hf$anyabuse)
hf$ins_mcare    <-factor(hf$ins_mcare)
hf$hmo  <-factor(hf$hmo)
hf$region_grp4  <-factor(hf$region_grp4)
hf$anx_any  <-factor(hf$anx_any)
hf$bipolar  <-factor(hf$bipolar)
hf$psycho   <-factor(hf$psycho)
hf$deprn    <-factor(hf$deprn)
hf$schiz    <-factor(hf$schiz)
hf$ipot_arth    <-factor(hf$ipot_arth)
hf$ipot_asth    <-factor(hf$ipot_asth)
hf$ipot_cancer  <-factor(hf$ipot_cancer)
hf$ipot_c_arrhy <-factor(hf$ipot_c_arrhy)
```

```
hf$ipot_ckd <-factor(hf$ipot_ckd)
hf$ipot_copd    <-factor(hf$ipot_copd)
hf$ipot_dementia    <-factor(hf$ipot_dementia)
hf$ipot_hepatitis   <-factor(hf$ipot_hepatitis)
hf$ipot_hilipid <-factor(hf$ipot_hilipid)
hf$ipot_htn <-factor(hf$ipot_htn)
hf$ipot_diabetes    <-factor(hf$ipot_diabetes)
hf$ipot_stroke  <-factor(hf$ipot_stroke)
hf$ipot_osteop  <-factor(hf$ipot_osteop)
hf$ipot_cad <-factor(hf$ipot_cad)
hf$ipot_mi  <-factor(hf$ipot_mi)
hf$sleep    <-factor(hf$sleep)
hf$obesity  <-factor(hf$obesity)

#create reference grps for R#
hf$abrx_3grp    <-C(hf$abrx_3grp,contr.treatment, base = 3)
hf$antiep_grp   <-C(hf$antiep_grp,contr.treatment, base = 4)
hf$metrx_any    <-C(hf$metrx_any,contr.treatment, base = 2)
hf$tzd_any  <-C(hf$tzd_any,contr.treatment, base = 2)
hf$dpp4_any     <-C(hf$dpp4_any,contr.treatment, base = 2)
hf$sulf_any <-C(hf$sulf_any,contr.treatment, base = 2)
hf$age_3grp <-C(hf$age_3grp,contr.treatment, base = 1)
hf$polyrx_gn_ge6 <-C(hf$polyrx_gn_ge6,contr.treatment, base = 2)
hf$anyabuse <-C(hf$anyabuse,contr.treatment, base = 2)
hf$ins_mcare    <-C(hf$ins_mcare,contr.treatment, base = 2)
hf$hmo  <-C(hf$hmo,contr.treatment, base = 2)
hf$region_grp4  <-C(hf$region_grp4,contr.treatment, base = 4)
hf$anx_any  <-C(hf$anx_any,contr.treatment, base = 2)
hf$bipolar  <-C(hf$bipolar,contr.treatment, base = 2)
hf$psycho   <-C(hf$psycho,contr.treatment, base = 2)
hf$deprn    <-C(hf$deprn,contr.treatment, base = 2)
hf$schiz    <-C(hf$schiz,contr.treatment, base = 2)
hf$ipot_arth    <-C(hf$ipot_arth,contr.treatment, base = 2)
hf$ipot_asth    <-C(hf$ipot_asth,contr.treatment, base = 2)
hf$ipot_cancer  <-C(hf$ipot_cancer,contr.treatment, base = 2)
hf$ipot_c_arrhy <-C(hf$ipot_c_arrhy,contr.treatment, base = 2)
hf$ipot_cad <-C(hf$ipot_cad,contr.treatment, base = 2)
hf$ipot_mi  <-C(hf$ipot_mi,contr.treatment, base = 2)
hf$ipot_ckd <-C(hf$ipot_ckd,contr.treatment, base = 2)
hf$ipot_copd    <-C(hf$ipot_copd,contr.treatment, base = 2)
hf$ipot_dementia    <-C(hf$ipot_dementia,contr.treatment, base = 2)
hf$ipot_hepatitis   <-C(hf$ipot_hepatitis,contr.treatment, base = 2)
hf$ipot_hilipid <-C(hf$ipot_hilipid,contr.treatment, base = 2)
hf$ipot_htn <-C(hf$ipot_htn,contr.treatment, base = 2)
hf$ipot_diabetes    <-C(hf$ipot_diabetes,contr.treatment, base = 2)
hf$ipot_stroke  <-C(hf$ipot_stroke,contr.treatment, base = 2)
hf$ipot_osteop  <-C(hf$ipot_osteop,contr.treatment, base = 2)
hf$sleep    <-C(hf$sleep,contr.treatment, base = 2)
hf$obesity  <-C(hf$obesity,contr.treatment, base = 2)

#check the target feature distribution in the dataset
#check the class balance
table(hf$hf_fu12)
```

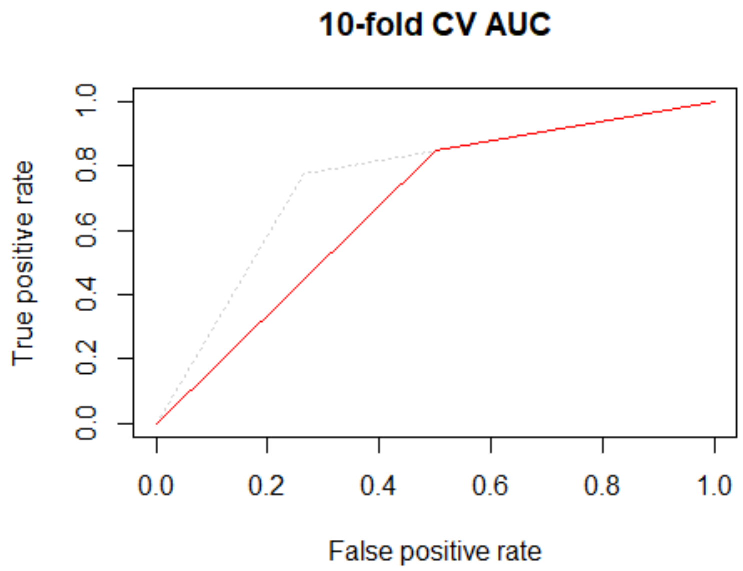
```
##
##      0      1
## 149379   3213
barplot(prop.table(table(hf$hf_fu12)),
        col = rainbow(2),
        ylim = c(0,1),
        main = "Class Distribution")
```



```
table(hf$hf_fu12)
##      0      1
## 149379   3213
prop.table(table(hf$hf_fu12))
##          0          1
## 0.97894385 0.02105615
#data partition into 70% train and 30% test (original dataset)
set.seed(123)  #set seed to make the analyses repeatable#
library(caret)
hf1 = sort(sample(nrow(hf),nrow(hf)*0.7))
hforig_train = hf[hf1,]   #training dataset
hforig_test = hf[-hf1,]   #test dataset
#fix the imbalanced dataset with undersampling
library(ROSE)
set.seed(999)
hf_us <- ovun.sample(hf_fu12~., data=hf, method="under",N=6426)$data
table(hf_us$hf_fu12)
##    0    1
## 3213 3213
#data partition into 70% train and 30% test#
set.seed(123)  #set seed to make the analyses repeatable#
library(caret)
hf1 = sort(sample(nrow(hf_us),nrow(hf_us)*0.7))
hftrain = hf_us[hf1,]   #training dataset
hftest = hf_us[-hf1,]   #test dataset

#check the target feature distribution in the training dataset
table(hftrain$hf_fu12)
##    0    1
## 2265 2233
print('distribution in the training dataset',prop.table(table(hftrain$hf_fu12)))
```

```
#10-fold cross-validation#
library(caret)
ctrl <- trainControl(method = "repeatedcv", number = 10, savePredictions = TRUE)


#Fit model
#this model was selected after comparing three models, and removing colinear
predictors (model_3)
cv_model <-
train(hf_fu12~age_3grp+ins_mcare+hmo+er_nbr+polyrx_gn_ge6+abrx_3grp+antiep_grp+metrx_
any+sulf_any+tzd_any+dpp4_any+ipot_htn+ipot_cad+ipot_mi+ipot_c_arrhy+ipot_stroke+ipot
_hilipid+ipot_diabetes+ipot_cancer+ipot_asth+ipot_copd+ipot_arth+ipot_osteop+ipot_ckd
+ipot_hepatitis+anx_any+deprn+bipolar+psycho+schiz+ipot_dementia+sleep+obesity+anyabu
se+region_grp4,
              data=hftrain,
              method = "glm",
              family = "binomial",
              trControl = ctrl)

#Summarize the CV Log Reg results
summary(cv_model)
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8612  -0.7810  -0.3026   0.8435   2.4337
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.796897   0.181712 -15.392  < 2e-16 ***
## age_3grp2         0.642593   0.173342   3.707  0.00021 ***
## age_3grp3         1.750022   0.182275   9.601  < 2e-16 ***
## ins_mcare1        0.715658   0.108248   6.611 3.81e-11 ***
## hmo1              0.142801   0.089972   1.587  0.11247
## er_nbr           -0.003399   0.026568  -0.128  0.89821
## polyrx_gn_ge61    0.701036   0.094113   7.449 9.42e-14 ***
## abrx_3grp1        0.121424   0.111663   1.087  0.27686
## abrx_3grp2       -0.081598   0.086050  -0.948  0.34300
## antiep_grp1       0.702288   0.381580   1.840  0.06570 .
## antiep_grp2       0.319651   0.156795   2.039  0.04148 *
## antiep_grp3      -0.596341   0.573773  -1.039  0.29865
## metrx_any1       -0.244137   0.146225  -1.670  0.09500 .
## sulf_any1         0.416069   0.170084   2.446  0.01443 *
## tzd_any1          0.638198   0.289101   2.208  0.02728 *
## dpp4_any1        -0.429273   0.246374  -1.742  0.08144 .
## ipot_htn1         0.428963   0.091124   4.707 2.51e-06 ***
## ipot_cad1         0.739214   0.115368   6.407 1.48e-10 ***
## ipot_mi1          0.225317   0.394528   0.571  0.56793
## ipot_c_arrhy1     0.799263   0.102256   7.816 5.44e-15 ***
## ipot_stroke1      0.370826   0.145890   2.542  0.01103 *
## ipot_hilipid1    -0.390490   0.085371  -4.574 4.78e-06 ***
## ipot_diabetes1    0.327725   0.099818   3.283  0.00103 **
## ipot_cancer1     -0.226106   0.102585  -2.204  0.02752 *
## ipot_asth1        0.230550   0.143977   1.601  0.10931
## ipot_copd1       0.824661   0.114153   7.224 5.04e-13 ***
## ipot_arth1       0.057651   0.086488   0.667  0.50504    
## ipot_osteop1    -0.181306   0.107703  -1.683  0.09230 .  
## ipot_ckd1        0.545173   0.121357   4.492 7.05e-06 ***
## ipot_hepatitis1 -0.671336   0.397109  -1.691  0.09092 .  
## anx_any1        -0.212897   0.138642  -1.536  0.12464    
## deprn1          -0.031554   0.115025  -0.274  0.78384    
## bipolar1         0.292729   0.391997   0.747  0.45521    
## psycho1         -0.629779   0.289058  -2.179  0.02935 *  
## schiz1           0.026756   0.589883   0.045  0.96382    
## ipot_dementia1   0.283585   0.168020   1.688  0.09145 .  
## sleep1           0.167869   0.126929   1.323  0.18599    
## obesity1         0.333359   0.142255   2.343  0.01911 *  
## anyabuse1        0.103839   0.158753   0.654  0.51305    
## region_grp41     0.191831   0.134815   1.423  0.15476    
## region_grp42     0.260797   0.104771   2.489  0.01280 *  
## region_grp43     0.195617   0.095451   2.049  0.04042 *  
## region_grp45    -0.336468   0.492044  -0.684  0.49409    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 6235.3  on 4497  degrees of freedom
## Residual deviance: 4600.5  on 4455  degrees of freedom
## AIC: 4686.5
## 
## Number of Fisher Scoring iterations: 4
exp(cv_model$finalModel$coefficients) #to get the ORs
##     (Intercept)       age_3grp2       age_3grp3      ins_mcare1            hmo1 
##      0.06099905      1.90140549      5.75472803      2.04553301      1.15350058 
##          er_nbr  polyrx_gn_ge61      abrx_3grp1      abrx_3grp2     antiep_grp1 
##      0.99660700      2.01583916      1.12910332      0.92164273      2.01836540 
##     antiep_grp2     antiep_grp3      metrx_any1       sulf_any1        tzd_any1 
##      1.37664661      0.55082321      0.78337995      1.51599061      1.89306720 
##       dpp4_any1       ipot_htn1       ipot_cad1        ipot_mi1   ipot_c_arrhy1 
##      0.65098204      1.53566462      2.09428912      1.25271952      2.22390132 
##    ipot_stroke1   ipot_hilipid1  ipot_diabetes1    ipot_cancer1      ipot_asth1 
##      1.44893121      0.67672541      1.38780672      0.79763335      1.25929235 
##      ipot_copd1      ipot_arth1    ipot_osteop1       ipot_ckd1 ipot_hepatitis1 
##      2.28110647      1.05934508      0.83417978      1.72490613      0.51102551 
##        anx_any1          deprn1        bipolar1         psycho1          schiz1 
##      0.80823972      0.96893890      1.34007901      0.53270928      1.02711738 
##  ipot_dementia1          sleep1        obesity1       anyabuse1    region_grp41 
##      1.32788156      1.18278189      1.39564893      1.10942207      1.21146547 
##    region_grp42    region_grp43    region_grp45 
##      1.29796440      1.21606071      0.71428873
#variable importance
#returns the absolute value of the t-statistic for each model parameter
varImp(cv_model)
## glm variable importance
## 
##   only 20 most important variables shown (out of 42)
## 
##                 Overall
## age_3grp3       100.00
## ipot_c_arrhy1    81.32
## polyrx_gn_ge61   77.48
## ipot_copd1       75.13
## ins_mcare1       68.71
## ipot_cad1        66.58
## ipot_htn1        48.79
## ipot_hilipid1    47.39
## ipot_ckd1        46.54
## age_3grp2        38.32
## ipot_diabetes1   33.88
## ipot_stroke1     26.13
## region_grp42     25.57
## sulf_any1        25.13
## obesity1         24.05
## tzd_any1         22.63
## ipot_cancer1     22.59
## psycho1          22.33
## region_grp43     20.97
## antiep_grp2      20.86
#Summarize the accuracy and kappa
cv_model
## Generalized Linear Model
##
## 4498 samples
##   35 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 4047, 4048, 4048, 4049, 4049, 4048, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7385647  0.4771807
#calculate accuracy
calc_acc = function(actual,predicted) {
  mean(actual == predicted)
}

#Make predictions on test data
head(predict(cv_model, newdata = hftest, type = "prob"))
##            0         1
## 3  0.6470469 0.3529531
## 4  0.8122266 0.1877734
## 6  0.4899400 0.5100600
## 7  0.7985798 0.2014202
## 9  0.5175976 0.4824024
## 10 0.1741712 0.8258288
#test accuracy of predictions
calc_acc(actual = hftest$hf_fu12,
         predicted = predict(cv_model, newdata = hftest))
## [1] 0.7349585
#get confusion matrix using test dataset
pred = predict(cv_model, newdata=hftest)
```

```
confusionMatrix(data=pred, hftest$hf_fu12, positive = '1')
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 683 246
##          1 265 734
##
##                Accuracy : 0.735
##                  95% CI : (0.7147, 0.7545)
##     No Information Rate : 0.5083
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.4696
##
##  Mcnemar's Test P-Value : 0.4259
##
##             Sensitivity : 0.7490
##             Specificity : 0.7205
##          Pos Pred Value : 0.7347
##          Neg Pred Value : 0.7352
##              Prevalence : 0.5083
##          Detection Rate : 0.3807
##    Detection Prevalence : 0.5182
##       Balanced Accuracy : 0.7347
##
##        'Positive' Class : 1
##
#calculate accuracy
#Make predictions on original test data
head(predict(cv_model, newdata = hforig_test, type = "prob"))
##           0          1
## 1 0.9515797 0.04842035
## 2 0.9636499 0.03635011
## 3 0.8924323 0.10756772
## 4 0.9094337 0.09056626
## 5 0.8949819 0.10501811
## 6 0.5559103 0.44408974
#test accuracy of predictions
calc_acc(actual = hforig_test$hf_fu12,
         predicted = predict(cv_model, newdata = hforig_test))
## [1] 0.7360741
#get confusion matrix using original test dataset
pred2 = predict(cv_model, newdata=hforig_test)
confusionMatrix(data=pred2, hforig_test$hf_fu12, positive = '1')
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 32914   225
##          1 11857   782
##
##                Accuracy : 0.7361
##                  95% CI : (0.732, 0.7401)
##     No Information Rate : 0.978
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.077
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.77656
##             Specificity : 0.73516
##          Pos Pred Value : 0.06187
##          Neg Pred Value : 0.99321
##              Prevalence : 0.02200
##          Detection Rate : 0.01708
##    Detection Prevalence : 0.27609
##       Balanced Accuracy : 0.75586
##
##        'Positive' Class : 1
#ROC#
library (cvAUC)
print(auc_value <-cvAUC(as.numeric(pred2), as.numeric(hforig_test$hf_fu12),
label.ordering = NULL, folds = 10))
## $perf
## A performance instance
##   'False positive rate' vs. 'True positive rate' (alpha: 'Cutoff')
##   with 3 data points
## $fold.AUC
## [1] 0.7558637
##
## $cvAUC
## [1] 0.7558637
#Plot fold AUCs
plot(auc_value$perf, col="grey82", lty=3, main="10-fold CV AUC")

#Plot CV AUC
plot(auc_value$perf, col="red", avg="vertical", add=TRUE)
```

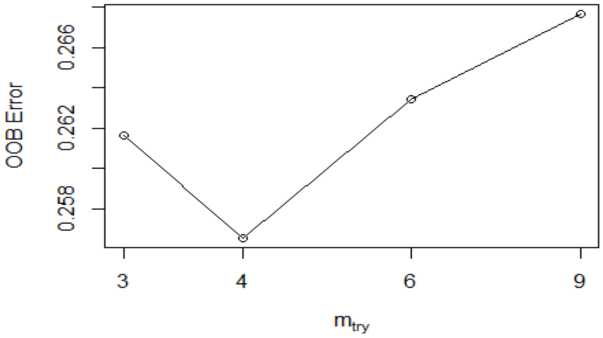


B. Random Forest Algorithm: Predictors of incident HF among Postmenopausal Women

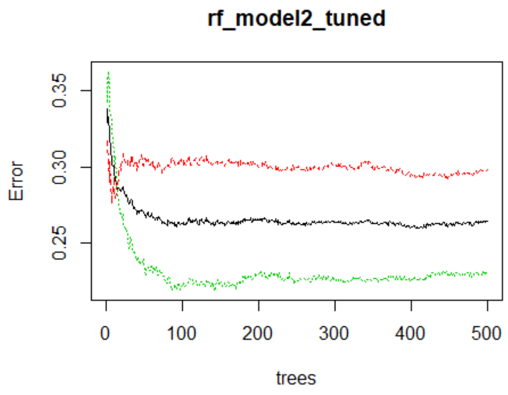
```
#read data
library(haven)
#read sas data---must install package haven and load library haven#
df <-
read_sas("Z:/OPTUM_10pct/projects/Khalid_phd/Aim_2/sasdata/hfree_2007_2016_n.sas7bdat
", NULL)

hf <- df[c('hf_fu12','pregabarx_any','gabarx_any','fqrx_any','abrx_othr',
           'metrx_any','tzd_any','dpp4_any','sulf_any',

'age_old','age_middle','polyrx_gn_ge6','anyabuse','ins_mcare','er_nbr','hmo','midwest
','northeast','south','anx_any','bipolar','psycho','deprn','schiz',

'ipot_arth','ipot_asth','ipot_cancer','ipot_cad','ipot_mi','sleep','obesity',

'ipot_c_arrhy','ipot_ckd','ipot_copd','ipot_dementia','ipot_hepatitis','ipot_hilipid'
,
           'ipot_htn','ipot_diabetes','ipot_stroke','ipot_osteop')]

#convert to factor variable
library(plyr)
hf$hf_fu12  <-factor(hf$hf_fu12)
hf$hf_fu12 <- revalue(hf$hf_fu12, c("1"= "1", "2"= "0")) #changing label 2 to 0
hf$hf_fu12 <- relevel(hf$hf_fu12, ref = "0")
table(hf$hf_fu12)
##      0      1
## 149379   3213
#recode indep variables to indicate categorical status to R#
hf$fqrx_any <-factor(hf$fqrx_any)
hf$abrx_othr    <-factor(hf$abrx_othr)
hf$gabarx_any   <-factor(hf$gabarx_any)
hf$pregabarx_any    <-factor(hf$pregabarx_any)
hf$metrx_any    <-factor(hf$metrx_any)
hf$tzd_any  <-factor(hf$tzd_any)
hf$dpp4_any     <-factor(hf$dpp4_any)
hf$sulf_any <-factor(hf$sulf_any)
hf$age_old  <-factor(hf$age_old)
hf$age_middle   <-factor(hf$age_middle)
hf$polyrx_gn_ge6 <-factor(hf$polyrx_gn_ge6)
hf$anyabuse <-factor(hf$anyabuse)
hf$hmo  <-factor(hf$hmo)
hf$ins_mcare    <-factor(hf$ins_mcare)
hf$midwest  <-factor(hf$midwest)
hf$south    <-factor(hf$south)
hf$notheast <-factor(hf$notheast)
hf$anx_any  <-factor(hf$anx_any)
hf$bipolar  <-factor(hf$bipolar)
hf$psycho   <-factor(hf$psycho)
hf$deprn    <-factor(hf$deprn)
hf$schiz    <-factor(hf$schiz)
```

```
hf$ipot_arth    <-factor(hf$ipot_arth)
hf$ipot_asth    <-factor(hf$ipot_asth)
hf$ipot_cancer  <-factor(hf$ipot_cancer)
hf$ipot_c_arrhy <-factor(hf$ipot_c_arrhy)
hf$ipot_ckd <-factor(hf$ipot_ckd)
hf$ipot_copd    <-factor(hf$ipot_copd)
hf$ipot_dementia    <-factor(hf$ipot_dementia)
hf$ipot_hepatitis   <-factor(hf$ipot_hepatitis)
hf$ipot_hilipid <-factor(hf$ipot_hilipid)
hf$ipot_htn <-factor(hf$ipot_htn)
hf$ipot_diabetes    <-factor(hf$ipot_diabetes)
hf$ipot_stroke  <-factor(hf$ipot_stroke)
hf$ipot_osteop  <-factor(hf$ipot_osteop)
hf$ipot_cad <-factor(hf$ipot_cad)
hf$ipot_mi  <-factor(hf$ipot_mi)
hf$sleep    <-factor(hf$sleep)
hf$obesity  <-factor(hf$obesity)
#fix the imbalanced dataset with undersampling
library(ROSE)
set.seed(999)
hf_us <- ovun.sample(hf_fu12~., data=hf, method="under",N=6426)$data
table(hf_us$hf_fu12)
##    0    1
## 3213 3213
#data partition into 70% train and 30% test#
set.seed(123)  #set seed to make the analyses repeatable#
library(caret)
hf1 = sort(sample(nrow(hf_us),nrow(hf_us)*0.7))
hftrain = hf_us[hf1,]   #training dataset
hftest = hf_us[-hf1,]   #test dataset

#check the target feature distribution in the training dataset
table(hftrain$hf_fu12)
##
##    0    1
## 2265 2233
print('distribution in the training dataset',prop.table(table(hftrain$hf_fu12)))
# Algorithm Tune (tuneRF)
library(randomForest)
set.seed(111)
x <- hftrain[c('abrx_othr','fqrx_any',
'gabarx_any','pregabarx_any','metrx_any','tzd_any','dpp4_any','sulf_any','age_old','a
ge_middle',

'polyrx_gn_ge6','anyabuse','ins_mcare','hmo','midwest','south','northeast','er_nbr',

'anx_any','bipolar','psycho','deprn','schiz','ipot_arth','ipot_asth','ipot_cancer','i
pot_c_arrhy','ipot_ckd',

'ipot_copd','ipot_dementia','ipot_hepatitis','ipot_hilipid','ipot_htn','ipot_diabetes
','ipot_stroke',
               'ipot_osteop','ipot_cad','ipot_mi','sleep','obesity')]
y <- hftrain$hf_fu12

bestmtry <- tuneRF(x, y, stepFactor=1.5, improve=1e-5, ntree=500)
```

```
## mtry = 6  OOB error = 26.35%
## Searching left ...
## mtry = 4     OOB error = 25.66%
## 0.02616034 1e-05
## mtry = 3     OOB error = 26.17%
## -0.01993068 1e-05
## Searching right ...
## mtry = 9     OOB error = 26.77%
## -0.04332756 1e-05
```

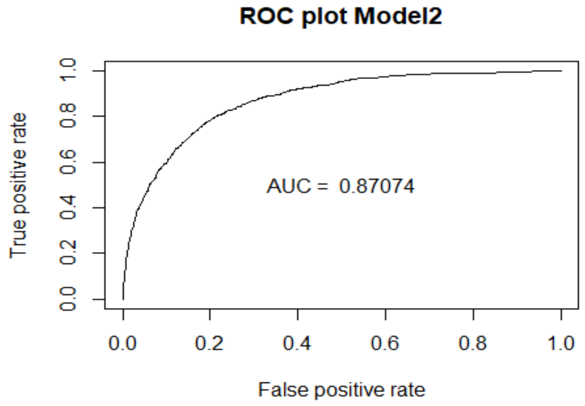


```
print(bestmtry)
##       mtry  OOBError
## 3.OOB    3 0.2616719
## 4.OOB    4 0.2565585
## 6.OOB    6 0.2634504
## 9.OOB    9 0.2676745
#random forest method
library(randomForest)
#use set seet to make it repeatable again#
set.seed(111)
rf_model2_tuned<-
randomForest(hf_fu12~age_old+age_middle+abrx_othr+fqrx_any+gabarx_any+pregabarx_any+m
etrx_any+tzd_any+dpp4_any+sulf_any+polyrx_gn_ge6

+anyabuse+ins_mcare+hmo+midwest+south+northeast+er_nbr
                                      +anx_any+bipolar+psycho+deprn
+schiz+ipot_arth+ipot_asth+ipot_cancer+ipot_c_arrhy+ipot_ckd

+ipot_copd+ipot_dementia+ipot_hepatitis+ipot_hilipid+ipot_htn+ipot_diabetes+ipot_stro
ke
                                      +ipot_osteop+ipot_cad+ipot_mi+sleep+obesity,
                                      data=hftrain,
                                      ntreeTry = 500,
                                      mtry =4,
                                      importance = TRUE)

#Print results from tuned Model
print(rf_model2_tuned)
## Call:
##  randomForest(formula = hf_fu12 ~ age_old + age_middle + abrx_othr +      fqrx_any
+ gabarx_any + pregabarx_any + metrx_any + tzd_any +      dpp4_any + sulf_any +
polyrx_gn_ge6 + anyabuse + ins_mcare +      hmo + midwest + south + northeast +
er_nbr + anx_any + bipolar +      psycho + deprn + schiz + ipot_arth + ipot_asth +
ipot_cancer +      ipot_c_arrhy + ipot_ckd + ipot_copd + ipot_dementia +
ipot_hepatitis +      ipot_hilipid + ipot_htn + ipot_diabetes + ipot_stroke +
ipot_osteop +      ipot_cad + ipot_mi + sleep + obesity, data = hftrain, ntreeTry =
500,      mtry = 4, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 26.41%
## Confusion matrix:
##      0    1 class.error
## 0 1589  676   0.2984547
## 1  512 1721   0.2292880
#error rate of random forest tuned model
plot(rf_model2_tuned)
```



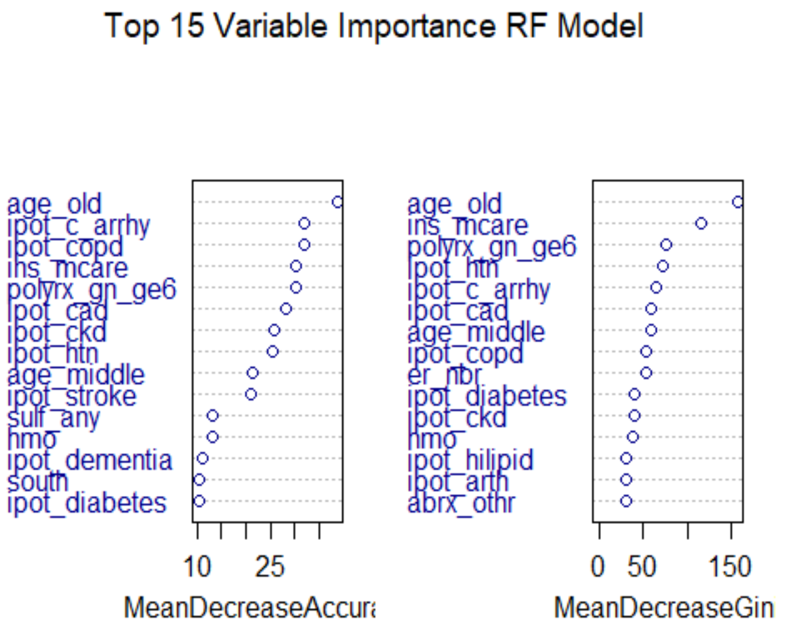
```
library(caret)
#predict and specify model we created using training data#
pred_model2 <-predict(rf_model2_tuned,hftrain)
confusionMatrix(pred_model2,hftrain$hf_fu12, positive = "1")
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1973  285
##          1  292 1948
##
##                Accuracy : 0.8717
##                  95% CI : (0.8616, 0.8814)
##     No Information Rate : 0.5036
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.7434
##
##  Mcnemar's Test P-Value : 0.8028
##
##             Sensitivity : 0.8724
##             Specificity : 0.8711
##          Pos Pred Value : 0.8696
##          Neg Pred Value : 0.8738
##              Prevalence : 0.4964
##          Detection Rate : 0.4331
##    Detection Prevalence : 0.4980
##       Balanced Accuracy : 0.8717
##
##        'Positive' Class : 1
##
#predict for test data#
pred_test2<-predict(rf_model2_tuned,hftest)
#get confusion matrix for test#
confusionMatrix(pred_test2,hftest$hf_fu12, positive = "1")
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 648 227
##          1 300 753
##
##                Accuracy : 0.7267
##                  95% CI : (0.7062, 0.7465)
##     No Information Rate : 0.5083
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4525
##
##  Mcnemar's Test P-Value : 0.001711
##
##             Sensitivity : 0.7684
##             Specificity : 0.6835
##          Pos Pred Value : 0.7151
##          Neg Pred Value : 0.7406
##              Prevalence : 0.5083
##          Detection Rate : 0.3906
##    Detection Prevalence : 0.5462
##       Balanced Accuracy : 0.7260
##
##        'Positive' Class : 1
#predict for original test data
pred_test3<-predict(rf_model2_tuned,hforig_test)
#get confusion matrix for test#
confusionMatrix(pred_test3,hforig_test$hf_fu12, positive = "1")
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 31757   135
##          1 13014   872
##
##                Accuracy : 0.7128
##                  95% CI : (0.7086, 0.7169)
##     No Information Rate : 0.978
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0793
##
##  Mcnemar's Test P-Value : <2e-16
```

```
##
##             Sensitivity : 0.86594
##             Specificity : 0.70932
##          Pos Pred Value : 0.06280
##          Neg Pred Value : 0.99577
##              Prevalence : 0.02200
##          Detection Rate : 0.01905
##    Detection Prevalence : 0.30333
##       Balanced Accuracy : 0.78763
##
##        'Positive' Class : 1
##
#install.packages("ROCR")
library(ROCR)
library(gplots)
#get predictions
oob.votes2 <- predict(rf_model2_tuned,hforig_test,type="prob")
head(oob.votes2)
##       0     1
## 1 0.974 0.026
## 2 0.950 0.050
## 3 0.996 0.004
## 4 0.900 0.100
## 5 0.898 0.102
## 6 0.452 0.548
oob.pred2<-oob.votes2[,2] #storing the prob of hf (1)
predictions2=as.vector(oob.pred2)
pred2=prediction(predictions2,hforig_test$hf_fu12)

#Calculate the AUC value
perf_AUC2=performance(pred2,"auc")
AUC2=perf_AUC2@y.values[[1]]

#plot the ROC curve
perf_ROC2=performance(pred2,"tpr","fpr")
plot(perf_ROC2, main="ROC plot Model2")
text(0.5,0.5,paste("AUC = ",format(AUC2, digits=5, scientific=FALSE)))
```



```
#get feature importance
importance(rf_model2_tuned)
```

```
##                          0           1 MeanDecreaseAccuracy MeanDecreaseGini
## age_old        19.05924285 40.65633157           38.7364207       157.114257
## age_middle     -8.73625341 23.47211445           21.2255557        58.496222
## abrx_othr       5.33158982 -1.52785455            2.7096516        30.039318
## fqrx_any        4.75607086 -0.66063327            3.1151737        23.660686
## gabarx_any      8.12408140  4.17010484            9.2107646        18.006772
## pregabarx_any   6.11173866 -1.16578795            3.7627089         6.480584
## metrx_any       8.18065924 -5.29870143            2.9358726        19.279095
## tzd_any         1.11546102  1.47035173            2.0028788         6.826976
## dpp4_any        5.36795920  1.26300673            5.2599164         9.640396
## sulf_any       15.04692278 -1.09202125           13.3134333        19.856049
## polyrx_gn_ge6  30.07587914  6.47092282           30.0666369        75.660382
## anyabuse        1.70508333  4.61712991            4.9983531        17.054730
## ins_mcare      16.90783146 23.92616409           30.1654214       116.635570
## hmo            12.14584020  1.21548703           13.0455428        38.836338
## midwest         2.36875931  1.02100583            2.3837802        24.857204
## south           9.71304487  4.48977061           10.3598085        29.556802
## northeast       0.40334598 -0.92297131           -0.4541694        18.554312
## er_nbr         14.81981339 -2.03859619            9.4621374        53.316327
## anx_any         0.03399209  4.05870640            3.4391783        18.389866
## bipolar         5.33497462 -4.08829291            1.7843846         4.164862
## psycho          6.80132383 -1.69399152            3.9590919         7.409968
## deprn           6.78041033 -2.70402681            2.7650871        23.088709
## schiz           2.12869053 -0.61966009            1.0702105         2.108506
## ipot_arth       1.10069519  4.17567945            4.1372278        30.696912
## ipot_asth       5.22647497 -1.25764437            3.2283363        19.046154
## ipot_cancer     3.70165849 -1.20016877            1.5842526        25.968135
## ipot_c_arrhy   33.35711299  9.58333354           32.0595909        64.518867
## ipot_ckd       26.01593250  2.01322700           25.7445972        39.691864
## ipot_copd      29.22903354 12.78705115           31.8242289        53.533113
## ipot_dementia  13.98423969 -2.91090089           11.2422955        16.749594
## ipot_hepatitis  1.13262081  0.06937672            0.8537319         4.176250
## ipot_hilipid   -4.21136260  7.45178545            3.5418554        30.699301
## ipot_htn       10.70859892 16.66928201           25.2759936        71.918921
## ipot_diabetes   8.85301124  2.60112764           10.2976553        39.858353
## ipot_stroke    26.39645319 -6.01529528           20.9933682        26.842238
## ipot_osteop     4.64829449 -2.11561456            1.6423567        23.459672
## ipot_cad       25.36288727  8.65254521           28.1031499        58.671870
## ipot_mi         7.32955525 -5.75995424            1.8741233         4.948871
## sleep           1.76454427  5.07874536            5.2185984        22.077589
## obesity        -0.33433954  2.69630979            1.8475211        18.686181
varImpPlot(rf_model2_tuned,sort=T, main="Top 15 Variable Importance RF
Model",n.var=15,col="blue4")
```



C. XGBoost: Predictors of Incident HF among Postmenopausal Women

```
#read data#
library(haven)     #to read the SAS file
library(tidyverse)
library(xgboost)
library(caret)
#read sas data---must install package haven and load library haven#
df <-
read_sas("Z:/OPTUM_10pct/projects/Khalid_phd/Aim_2/sasdata/hfree_2007_2016_xg.sas7bda
t", NULL) #converted all variables to 0s and 1s and made dummy variables where
necessary

hf <- df[c('hf_fu12','abrx_othr','fqrx_any',
'gabarx_any','pregabarx_any','metrx_any','tzd_any','dpp4_any','sulf_any','age_old','a
ge_middle',
          'midwest','south','northeast',
           'polyrx_gn_ge6','anyabuse','ins_mcare','hmo','er_nbr',

'anx_any','bipolar','psycho','deprn','schiz','omi','ipot_arth','ipot_asth','ipot_canc
er','ipot_c_arrhy','ipot_ckd',
           'ipot_copd','ipot_dementia','ipot_hepatitis','ipot_hilipid',

'ipot_htn','ipot_diabetes','ipot_stroke','ipot_osteop','ipot_cad','ipot_mi','sleep','
obesity')]

hf[is.na(hf)] = 0 #setting missing values to zero along with other missing values

#look at structure of data#
dim(hf)
## [1] 152592     42
head(hf) #pay attention to all potential categorical variables to ensure they are
coded as 0 and 1
```

```
## # A tibble: 6 x 42
##   hf_fu12 abrx_othr fqrx_any gabarx_any pregabarx_any metrx_any tzd_any dpp4_any
##     <dbl>     <dbl>    <dbl>      <dbl>         <dbl>     <dbl>   <dbl>    <dbl>
## 1       0         0        0          0             0         0       0        0
## 2       0         0        1          0             0         0       0        0
## 3       0         0        0          0             0         0       0        0
## 4       0         0        0          0             0         0       0        0
## 5       0         0        0          0             0         0       0        0
## 6       0         1        0          1             0         0       0        0
## # ... with 34 more variables: sulf_any <dbl>, age_old <dbl>, age_middle <dbl>,
## #   midwest <dbl>, south <dbl>, northeast <dbl>, polyrx_gn_ge6 <dbl>,
## #   anyabuse <dbl>, ins_mcare <dbl>, hmo <dbl>, er_nbr <dbl>, anx_any <dbl>,
## #   bipolar <dbl>, psycho <dbl>, deprn <dbl>, schiz <dbl>, omi <dbl>,
## #   ipot_arth <dbl>, ipot_asth <dbl>, ipot_cancer <dbl>, ipot_c_arrhy <dbl>,
## #   ipot_ckd <dbl>, ipot_copd <dbl>, ipot_dementia <dbl>, ipot_hepatitis <dbl>,
## #   ipot_hilipid <dbl>, ipot_htn <dbl>, ipot_diabetes <dbl>, ipot_stroke <dbl>,
## #   ipot_osteop <dbl>, ipot_cad <dbl>, ipot_mi <dbl>, sleep <dbl>,
## #   obesity <dbl>
         #also make sure that variables with multiple categories are converted to
dummy vars e.g. age_4grp, region
#str(hf)
```

#Keep only required vars and create a subset of the dataset #make sure all variables are numeric

```
#select only required vars for the ease of analysis
hf_select = hf[,c('hf_fu12','abrx_othr','fqrx_any',
'gabarx_any','pregabarx_any','metrx_any','tzd_any','dpp4_any','sulf_any','age_old','a
ge_middle',
           'midwest','south','northeast',
           'polyrx_gn_ge6','anyabuse','ins_mcare','hmo','er_nbr',

'anx_any','bipolar','psycho','deprn','schiz','ipot_arth','ipot_asth','ipot_cancer','i
pot_c_arrhy','ipot_ckd',
           'ipot_copd','ipot_dementia','ipot_hepatitis','ipot_hilipid',

'ipot_htn','ipot_diabetes','ipot_stroke','ipot_osteop','ipot_cad','ipot_mi','sleep','
obesity')]
dim(hf_select)
## [1] 152592     41
#dependent variable labels
#1st set of variables
hf_select$hf_fu12 <- as.factor(hf_select$hf_fu12)
levels(hf_select$hf_fu12)<- list("no" = "0" , "yes" = "1") #caret requires labels
head(hf_select$hf_fu12)
## [1] no no no no no no
## Levels: no yes
table(hf_select$hf_fu12) #make sure labels are correct
##     no    yes
## 149379   3213
#1st set of variables
set.seed(123)
hf_select1 <- as.data.frame(hf_select) #converting to a data frame for sampling;
random sampling does not work otherwise
n = nrow(hf_select1) #get total no. of rows

train.index = sample(n,floor(0.7*n)) #randomly select 70% rows from hf_select
```

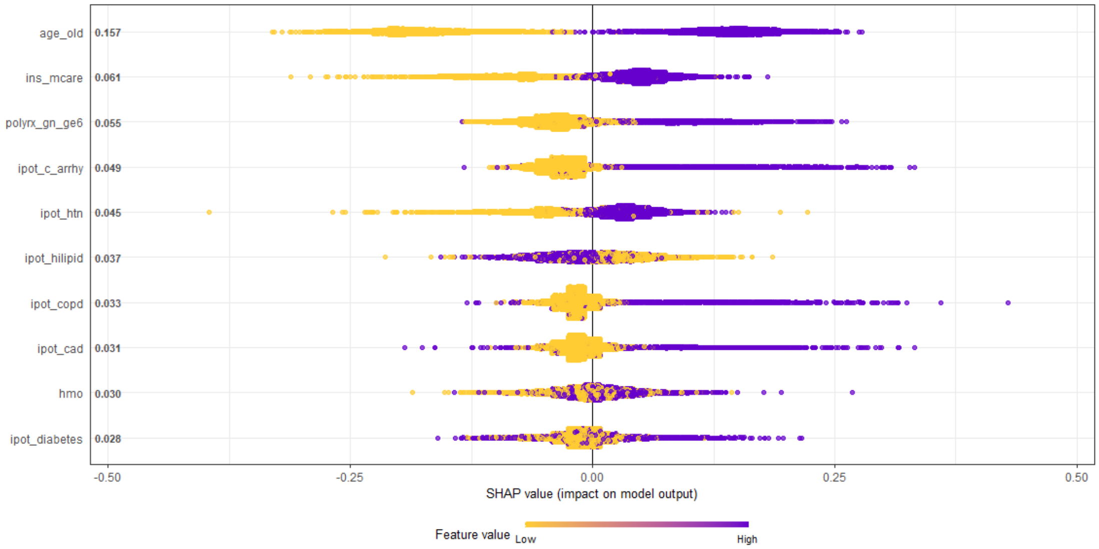
```
#original training data set
set.seed(123)
hforig_train_data <- hf_select1[train.index,] #this will select rows in train.index
head(hforig_train_data)
##        hf_fu12 abrx_othr fqrx_any gabarx_any pregabarx_any metrx_any tzd_any
## 134058     yes         1        0          1             0         0       0
## 124022      no         0        0          0             0         0       0
## 103065      no         0        0          0             0         0       0
## 124507      no         0        0          0             0         0       0
## 45404       no         1        0          0             0         1       0
## 65161       no         0        0          0             0         0       0
##        dpp4_any sulf_any age_old age_middle midwest south northeast
## 134058        0        0       1          0       0     0         0
## 124022        0        0       0          1       0     1         0
## 103065        0        0       0          1       0     0         0
## 124507        0        0       1          0       0     0         1
## 45404         1        0       0          1       0     1         0
## 65161         0        0       1          0       0     0         0
##        polyrx_gn_ge6 anyabuse ins_mcare hmo er_nbr anx_any bipolar psycho deprn
## 134058             1        0         1   1      0       0       0      0     1
## 124022             0        0         1   1      0       0       0      0     0
## 103065             0        0         0   0      0       0       0      0     0
## 124507             0        0         1   0      0       0       0      0     0
## 45404              1        0         0   0      0       0       0      0     0
## 65161              1        0         1   1      1       0       0      0     0
##        schiz ipot_arth ipot_asth ipot_cancer ipot_c_arrhy ipot_ckd ipot_copd
## 134058     0         1         0           0            1        1         1
## 124022     0         0         0           0            0        0         0
## 103065     0         0         0           0            0        0         0
## 124507     0         0         0           0            0        0         0
## 45404      0         0         0           1            0        0         1
## 65161      0         0         0           0            0        0         0
##        ipot_dementia ipot_hepatitis ipot_hilipid ipot_htn ipot_diabetes
## 134058             1              0            0        1             0
## 124022             0              0            0        0             0
## 103065             0              0            0        0             0
## 124507             0              0            0        0             0
## 45404              0              0            1        1             1
## 65161              0              0            0        1             0
##        ipot_stroke ipot_osteop ipot_cad ipot_mi sleep obesity
## 134058           0           0        0       0     0       0
## 124022           0           0        0       0     0       0
## 103065           0           0        0       0     0       0
## 124507           0           0        0       0     0       0
## 45404            0           0        0       0     1       0
## 65161            0           0        0       0     0       0
dim(hforig_train_data)
## [1] 106814     41
#original test data set
set.seed(123)
hforig_test_data <- hf_select1[-train.index,] #this will select those rows not in
train.index
head(hforig_test_data)
##    hf_fu12 abrx_othr fqrx_any gabarx_any pregabarx_any metrx_any tzd_any
## 3       no         0        0          0             0         0       0
## 4       no         0        0          0             0         0       0
## 5       no         0        0          0             0         0       0
## 6       no         1        0          1             0         0       0
## 9       no         0        1          0             0         0       0
## 11      no         1        0          0             0         0       0
##    dpp4_any sulf_any age_old age_middle midwest south northeast polyrx_gn_ge6
## 3         0        0       0          0       0     0         0             0
## 4         0        0       0          0       1     0         0             0
## 5         0        0       0          1       0     0         0             0
## 6         0        0       0          1       0     0         0             0
## 9         0        0       0          0       0     0         0             0
## 11        0        0       1          0       0     0         0             1
##    anyabuse ins_mcare hmo er_nbr anx_any bipolar psycho deprn schiz ipot_arth
## 3         0         0   0      0       0       0      0     0     0         0
## 4         0         0   1      0       1       0      0     0     0         0
## 5         0         0   0      0       0       0      0     0     0         0
## 6         0         0   0      0       0       0      0     0     0         0
## 9         1         0   0      0       0       0      0     0     0         0
## 11        0         0   0      0       0       0      0     0     0         0
##    ipot_asth ipot_cancer ipot_c_arrhy ipot_ckd ipot_copd ipot_dementia
## 3          0           0            0        0         0             0
## 4          0           0            0        0         0             0
## 5          0           0            0        0         0             0
## 6          0           0            0        0         0             0
## 9          0           0            0        0         0             0
## 11         0           1            0        0         0             0
##    ipot_hepatitis ipot_hilipid ipot_htn ipot_diabetes ipot_stroke ipot_osteop
## 3               0            0        0             0           0           1
## 4               1            0        0             0           0           0
## 5               0            1        1             0           0           0
## 6               0            1        0             0           0           0
## 9               0            0        1             0           0           0
## 11              0            0        1             0           0           0
##    ipot_cad ipot_mi sleep obesity
## 3         0       0     0       0
## 4         0       0     0       0
## 5         0       0     0       0
## 6         0       0     0       0
## 9         0       0     0       0
## 11        0       0     0       0
dim(hforig_test_data)
## [1] 45778    41
#fix the imbalanced dataset with undersampling
library(ROSE)
set.seed(999)
hf_select_us <- ovun.sample(hf_fu12~., data=hf_select, method="under",N=6426)$data
table(hf_select_us$hf_fu12)
##
##   no  yes
## 3213 3213
#1st set of variables
set.seed(123)
hf_select_us<- as.data.frame(hf_select_us) #converting to a data frame for sampling;
random sampling does not work otherwise
n = nrow(hf_select_us) #get total no. of rows
```

```
train.index = sample(n,floor(0.7*n)) #randomly select 70% rows from hf_select

#undersampled training data set
set.seed(123)
hftrain_data <- hf_select_us[train.index,] #this will select rows in train.index
head(hftrain_data)
##      hf_fu12 abrx_othr fqrx_any gabarx_any pregabarx_any metrx_any tzd_any
## 2463      no         1        0          0             0         0       0
## 2511      no         1        0          0             0         0       0
## 2227      no         0        0          0             0         0       0
## 526       no         0        1          0             0         0       0
## 4291     yes         0        0          0             0         0       0
## 2986      no         0        0          0             0         0       0
##      dpp4_any sulf_any age_old age_middle midwest south northeast polyrx_gn_ge6
## 2463        0        0       1          0       0     0         0             0
## 2511        0        0       1          0       0     0         1             0
## 2227        0        0       0          1       0     1         0             0
## 526         0        0       0          1       0     0         1             0
## 4291        0        0       1          0       1     0         0             0
## 2986        0        0       0          0       0     0         1             1
##      anyabuse ins_mcare hmo er_nbr anx_any bipolar psycho deprn schiz ipot_arth
## 2463        0         1   0      0       0       0      0     0     0         0
## 2511        0         1   0      1       0       0      0     0     0         0
## 2227        0         0   0      0       0       0      0     0     0         0
## 526         0         1   0      0       0       0      0     0     0         0
## 4291        0         1   1      0       0       0      0     0     0         0
## 2986        0         1   0      1       1       1      0     0     0         1
##      ipot_asth ipot_cancer ipot_c_arrhy ipot_ckd ipot_copd ipot_dementia
## 2463         0           0            1        0         0             0
## 2511         0           0            0        0         0             0
## 2227         0           0            0        0         0             0
## 526          0           0            0        0         0             0
## 4291         0           0            1        0         0             0
## 2986         0           0            0        0         0             0
##      ipot_hepatitis ipot_hilipid ipot_htn ipot_diabetes ipot_stroke ipot_osteop
## 2463              0            1        0             0           0           0
## 2511              0            1        1             0           1           0
## 2227              0            0        0             0           0           0
## 526               0            0        1             1           0           0
## 4291              0            1        1             0           0           0
## 2986              0            1        0             0           0           0
##      ipot_cad ipot_mi sleep obesity
## 2463        0       0     0       0
## 2511        0       0     0       0
## 2227        0       0     0       0
## 526         0       0     0       0
## 4291        0       0     0       0
## 2986        0       0     0       0
dim(hftrain_data)
## [1] 4498   41
#undersampled test data set
set.seed(123)
hftest_data <- hf_select_us[-train.index,] #this will select those rows not in
```

```
train.index
head(hftest_data)
##    hf_fu12 abrx_othr fqrx_any gabarx_any pregabarx_any metrx_any tzd_any
## 3       no         0        0          0             0         0       0
## 4       no         0        0          0             0         0       0
## 6       no         1        0          0             0         0       0
## 7       no         1        0          0             0         0       0
## 9       no         0        0          0             0         0       0
## 10      no         0        1          0             0         0       0
##    dpp4_any sulf_any age_old age_middle midwest south northeast polyrx_gn_ge6
## 3         0        0       0          1       1     0         0             0
## 4         0        0       0          1       1     0         0             0
## 6         0        0       0          1       0     1         0             0
## 7         0        0       0          1       0     0         0             0
## 9         0        0       1          0       1     0         0             0
## 10        1        1       0          1       0     1         0             1
##    anyabuse ins_mcare hmo er_nbr anx_any bipolar psycho deprn schiz ipot_arth
## 3         0         1   1      0       0       0      0     0     0         0
## 4         0         0   0      0       0       0      0     0     0         0
## 6         0         1   1      0       0       0      0     1     0         0
## 7         0         1   1      0       0       0      0     0     0         0
## 9         0         1   0      0       0       0      0     0     0         0
## 10        0         1   0      0       0       0      0     0     0         0
##    ipot_asth ipot_cancer ipot_c_arrhy ipot_ckd ipot_copd ipot_dementia
## 3          0           0            0        0         0             0
## 4          0           0            0        0         0             0
## 6          0           0            0        0         1             0
## 7          0           0            0        0         0             0
## 9          0           0            0        0         0             0
## 10         0           0            1        0         1             0
##    ipot_hepatitis ipot_hilipid ipot_htn ipot_diabetes ipot_stroke ipot_osteop
## 3               0            0        1             0           0           0
## 4               0            0        1             0           0           0
## 6               0            0        1             0           0           0
## 7               0            0        0             0           0           0
## 9               0            0        0             0           0           0
## 10              0            1        1             1           0           0
##    ipot_cad ipot_mi sleep obesity
## 3         0       0     0       0
## 4         0       0     0       0
## 6         0       0     0       0
## 7         0       0     0       0
## 9         0       0     0       0
## 10        0       0     0       0
dim(hftest_data)
## [1] 1928   41
#install.packages("SHAPforxgboost")
library(SHAPforxgboost)
#Running the same xgboost model with the following command due to to non-numeric var
error with shap.values function
library(xgboost)
library(ggplot2)
hftrain <- subset(hftrain_data, select = -c(hf_fu12)) #copy hftrain_data and drop the
DV
dim(hftrain)
```

```
## [1] 4498   40
hftrain_label <- hftrain_data[,"hf_fu12"] #capture labels of the dv
head(hftrain_label)
## [1] no  no  no  no  yes no
## Levels: no yes
hftest <- subset(hftest_data, select = -c(hf_fu12)) #copy hftest_data and drop the DV
dim(hftest)
## [1] 1928   40
hftest_label <- hftest_data[,"hf_fu12"] #capture labels of the dv
head(hftest_label)
## [1] no no no no no no
## Levels: no yes
hforig_test <- subset(hforig_test_data, select = -c(hf_fu12)) #copy hforig_test_data
and drop the DV
dim(hforig_test)
## [1] 45778    40
hforig_test_label <- hforig_test_data[,"hf_fu12"] #capture labels of the dv
head(hforig_test_label)
## [1] no no no no no no
## Levels: no yes
#hyperparameter tuning results from the final model tuned using caret package
params <- list (objective = "multi:softprob",
                nrounds = 700,
                eta = 0.01,
                max_depth = 3,
                gamma = 0,
                subsample = 0.5,
                colsample_bytree = 1,
                min_child_weight = 1,
                eval_metric = "auc"
                )

#run the xgboost model
xgb_train <- xgboost::xgboost(data = as.matrix(hftrain),
                         label = hftrain_label,
                         xgb_param = params,
                         nrounds = params$nrounds,
                         verbose = FALSE
                         )
## [22:12:29] WARNING: amalgamation/../src/learner.cc:480:
## Parameters: { xgb_param } might not be used.
##
##   This may not be accurate due to some parameters are only used in language
bindings but
##   passed down to XGBoost core.  Or some parameters are not used but slip through
this
##   verification. Please open an issue if you find above cases.
#print the model
xgb_train
## ##### xgb.Booster
## raw: 2.7 Mb
## call:
##   xgb.train(params = params, data = dtrain, nrounds = nrounds,
##     watchlist = watchlist, verbose = verbose, print_every_n = print_every_n,
##     early_stopping_rounds = early_stopping_rounds, maximize = maximize,
##     save_period = save_period, save_name = save_name, xgb_model = xgb_model,
##     callbacks = callbacks, xgb_param = ..1)
## params (as set within xgb.train):
##   xgb_param = "multi:softprob", validate_parameters = "700", xgb_param = "0.01",
validate_parameters = "3", xgb_param = "0", validate_parameters = "0.5", xgb_param =
"1", validate_parameters = "1", xgb_param = "auc", validate_parameters = "TRUE"
## xgb.attributes:
##   niter
## callbacks:
##   cb.evaluation.log()
## # of features: 40
## niter: 700
## nfeatures : 40
## evaluation_log:
##     iter train_rmse
##        1   0.835650
##        2   0.653909
## ---
##      699   0.209106
##      700   0.209073
#run the xgboost model
xgb_test <- xgboost::xgboost(data = as.matrix(hftest),
                             label = hftest_label,
                             xgb_param = params,
                             nrounds = params$nrounds,
                             verbose = FALSE
                             )
## [22:12:35] WARNING: amalgamation/../src/learner.cc:480:
## Parameters: { xgb_param } might not be used.
##
##   This may not be accurate due to some parameters are only used in language
bindings but
##   passed down to XGBoost core.  Or some parameters are not used but slip through
this
##   verification. Please open an issue if you find above cases.
#print the model
xgb_test
## ##### xgb.Booster
## raw: 2.7 Mb
## call:
##   xgb.train(params = params, data = dtrain, nrounds = nrounds,
##     watchlist = watchlist, verbose = verbose, print_every_n = print_every_n,
##     early_stopping_rounds = early_stopping_rounds, maximize = maximize,
##     save_period = save_period, save_name = save_name, xgb_model = xgb_model,
##     callbacks = callbacks, xgb_param = ..1)
## params (as set within xgb.train):
##   xgb_param = "multi:softprob", validate_parameters = "700", xgb_param = "0.01",
validate_parameters = "3", xgb_param = "0", validate_parameters = "0.5", xgb_param =
"1", validate_parameters = "1", xgb_param = "auc", validate_parameters = "TRUE"
## xgb.attributes:
##   niter
## callbacks:
##   cb.evaluation.log()
## # of features: 40
## niter: 700
## nfeatures : 40
## evaluation_log:
##     iter train_rmse
##        1   0.844361
##        2   0.659897
## ---
##      699   0.151722
##      700   0.151700
#run the xgboost model using original dataset
xgb_test_orig <- xgboost::xgboost(data = as.matrix(hforig_test),
                        label = hforig_test_label,
                        xgb_param = params,
                        nrounds = params$nrounds,
                        verbose = FALSE
                        )
## [22:12:40] WARNING: amalgamation/../src/learner.cc:480:
## Parameters: { xgb_param } might not be used.
##
##   This may not be accurate due to some parameters are only used in language
bindings but
##   passed down to XGBoost core.  Or some parameters are not used but slip through
this
##   verification. Please open an issue if you find above cases.
#print the model
xgb_test_orig
## ##### xgb.Booster
## raw: 2.8 Mb
## call:
##   xgb.train(params = params, data = dtrain, nrounds = nrounds,
##     watchlist = watchlist, verbose = verbose, print_every_n = print_every_n,
##     early_stopping_rounds = early_stopping_rounds, maximize = maximize,
##     save_period = save_period, save_name = save_name, xgb_model = xgb_model,
##     callbacks = callbacks, xgb_param = ..1)
## params (as set within xgb.train):
##   xgb_param = "multi:softprob", validate_parameters = "700", xgb_param = "0.01",
validate_parameters = "3", xgb_param = "0", validate_parameters = "0.5", xgb_param =
"1", validate_parameters = "1", xgb_param = "auc", validate_parameters = "TRUE"
## xgb.attributes:
##   niter
## callbacks:
##   cb.evaluation.log()
## # of features: 40
## niter: 700
## nfeatures : 40
## evaluation_log:
##     iter train_rmse
##        1   0.392997
##        2   0.293120
## ---
##      699   0.097959
##      700   0.097957
#Get SHAP values and ranked features by mean|SHAP| for train data
set.seed(222)
shapvalues_trn <- shap.values(xgb_train, hftrain)
```

```
meanshap_trn <- shapvalues_trn$mean_shap_score

#Prepare long form data for dependende plot
#shaplong_trn <- shap.prep(xgb_train, X_train = hftrain)

#plot the SHAP value summmary plot
shap.plot.summary.wrap1(xgb_train, as.matrix(hftrain), top_n = 10) #dilute helps when
there are a lot of data points
```



```
#Plot of meaan SHAP score vs top 10 predictors
library(ggplot2)
trainshap_names <- as.data.frame(names(meanshap_trn[1:15])) #get names of all
features sorted by mean SHAP score
trainshap_val <- as.data.frame(unname(meanshap_trn[1:15])) #get sorted mean SHAP
values
trainshap <- cbind(trainshap_names, trainshap_val)
colnames(trainshap) <- c("feature", "meanSHAP") #copied this table then to Excel to
make the graphs
```

Appendix 6.5
R Codes for Random Forest Algorithm to Identify Predictors of Heart Failure-related Emergency Room Use among Postmenopausal Women

```
#read data#
library(haven)     #to read the SAS file
library(tidyverse)
library(xgboost)
library(caret)
#read sas data---must install package haven and load library haven#
df <- read_sas("Z:/OPTUM_10pct/projects/Khalid_phd/Aim_3/sasdata/hf2015_2016_hfxg.sas
7bdat", NULL) #converted all variables to 0s and 1s and made dummy variables where ne
cessary
hf <- df[c('hfer_use','hfer_use_base','hfer_nbr_base','ip_nbr_base','carefrag2015','l
ipdrx_any','bbrx_any','acerx_any','arbrx_any','diurx_any',
         abrx_3grp','fqrx_any','abrx_othr','antiep_grp', 'pregabarx_any','gabarx_any'
,'metrx_any','tzd_any','dpp4_any','sulf_any','age','age_3grp','age_old','age_middle',
'age_young','polyrx_gn_ge6','anyabuse','ins_mcare',
'hmo','region_grp4','midwest','northeast','south','anx_any','deprn',
'ipot_arth','ipot_asth','ipot_cancer','ipot_c_arrhy','ipot_cad','ipot_mi','ipot_ckd',
'ipot_copd','ipot_dementia','ipot_hilipid','ipot_htn','ipot_diabetes','ipot_stroke','
ipot_osteop','sleep_2015','obesity_2015')]

# convert NA to 0
hf[is.na(hf)] <- 0

#convert to factor variable---for RF 1 is hfer 0 is no hfer#
# required for caret package
table(hf$hfer_use) #before changing the levels
##    0    1
## 4490 1692

hf$hfer_use<-as.factor(hf$hfer_use)

#recode indep variables to indicate categorical status to R#
hf$hfer_use_base <-factor(hf$hfer_use_base)
hf$lipdrx_any   <-factor(hf$lipdrx_any)
hf$bbrx_any <-factor(hf$bbrx_any)
hf$acerx_any    <-factor(hf$acerx_any)
hf$arbrx_any    <-factor(hf$arbrx_any)
hf$diurx_any        <-factor(hf$diurx_any)
hf$abrx_3grp    <-factor(hf$abrx_3grp)
hf$fqrx_any <-factor(hf$fqrx_any)
hf$abrx_othr    <-factor(hf$abrx_othr)
hf$abrx_3grp    <-factor(hf$abrx_3grp)
hf$antiep_grp   <-factor(hf$antiep_grp)
hf$gabarx_any   <-factor(hf$gabarx_any)
hf$metrx_any    <-factor(hf$metrx_any)
hf$tzd_any  <-factor(hf$tzd_any)
hf$dpp4_any     <-factor(hf$dpp4_any)
hf$sulf_any <-factor(hf$sulf_any)
hf$age_3grp <-factor(hf$age_3grp)
hf$age_old  <-factor(hf$age_old)
hf$age_middle   <-factor(hf$age_middle)
hf$polyrx_gn_ge6 <-factor(hf$polyrx_gn_ge6)
hf$anyabuse <-factor(hf$anyabuse)
```

```
hf$ins_mcare    <-factor(hf$ins_mcare)
hf$hmo  <-factor(hf$hmo)
hf$region_grp4  <-factor(hf$region_grp4)
hf$midwest  <-factor(hf$midwest)
hf$northeast    <-factor(hf$northeast)
hf$south    <-factor(hf$south)
hf$anx_any  <-factor(hf$anx_any)
hf$deprn    <-factor(hf$deprn)
hf$ipot_arth    <-factor(hf$ipot_arth)
hf$ipot_asth    <-factor(hf$ipot_asth)
hf$ipot_cancer  <-factor(hf$ipot_cancer)
hf$ipot_cad <-factor(hf$ipot_cad)
hf$ipot_mi  <-factor(hf$ipot_mi)
hf$ipot_c_arrhy <-factor(hf$ipot_c_arrhy)
hf$ipot_ckd <-factor(hf$ipot_ckd)
hf$ipot_copd    <-factor(hf$ipot_copd)
hf$ipot_dementia    <-factor(hf$ipot_dementia)
hf$ipot_hilipid <-factor(hf$ipot_hilipid)
hf$ipot_htn <-factor(hf$ipot_htn)
hf$ipot_diabetes    <-factor(hf$ipot_diabetes)
hf$ipot_stroke  <-factor(hf$ipot_stroke)
hf$ipot_osteop  <-factor(hf$ipot_osteop)
hf$sleep_2015   <-factor(hf$sleep_2015)
hf$obesity_2015 <-factor(hf$obesity_2015)

#numeric variables
hf$age <- as.numeric(hf$age)
hf$carefrag2015 <- as.numeric (hf$carefrag2015)
hf$hfer_nbr_base <- as.numeric (hf$hfer_nbr_base)


#look at structure of data#
dim(hf)

## [1] 6182   51

head(hf) #pay attention to all potential categorical variables to ensure they are cod
ed as 0 and 1

## # A tibble: 6 x 51
##   hfer_use hfer_use_base hfer_nbr_base ip_nbr_base carefrag2015 lipdrx_any
##   <fct>    <fct>                 <dbl>       <dbl>        <dbl> <fct>
## 1 0        0                         0           1        0.479 0
## 2 0        1                         6           4        0.86  1
## 3 0        0                         0           0        0.571 0
## 4 0        0                         0           0        0.679 1
## 5 0        0                         0           0        0.681 1
## 6 0        1                         2           1        0.627 1
## # ... with 45 more variables: bbrx_any <fct>, acerx_any <fct>, arbrx_any <fct>,
## #   diurx_any <fct>, abrx_3grp <fct>, fqrx_any <fct>, abrx_othr <fct>,
## #   antiep_grp <fct>, pregabarx_any <dbl>, gabarx_any <fct>, metrx_any <fct>,
## #   tzd_any <fct>, dpp4_any <fct>, sulf_any <fct>, age <dbl>, age_3grp <fct>,
## #   age_old <fct>, age_middle <fct>, age_young <dbl>, polyrx_gn_ge6 <fct>,
## #   anyabuse <fct>, ins_mcare <fct>, hmo <fct>, region_grp4 <fct>,
## #   midwest <fct>, northeast <fct>, south <fct>, anx_any <fct>, deprn <fct>,
## #   ipot_arth <fct>, ipot_asth <fct>, ipot_cancer <fct>, ipot_c_arrhy <fct>,
## #   ipot_cad <fct>, ipot_mi <fct>, ipot_ckd <fct>, ipot_copd <fct>,
## #   ipot_dementia <fct>, ipot_hilipid <fct>, ipot_htn <fct>,
## #   ipot_diabetes <fct>, ipot_stroke <fct>, ipot_osteop <fct>,
## #   sleep_2015 <fct>, obesity_2015 <fct>

#also make sure that variables with multiple categories are converted to dummy
#str(hf)

#select only required vars for the ease of analysis
#based on lit review (RFE with ER use)
hf_select = hf[,c('hfer_use','hfer_nbr_base','carefrag2015','lipdrx_any','bbrx_any','
acerx_any','arbrx_any','diurx_any','fqrx_any','abrx_othr','gabarx_any','metrx_any','d
pp4_any',
'sulf_any','age','age_old','age_middle','polyrx_gn_ge6','anyabuse','ins_mcare',
'hmo','midwest','northeast','south','anx_any','deprn','ipot_arth','ipot_asth','ipot_c
ancer','ipot_c_arrhy','ipot_cad','ipot_mi','ipot_ckd','ipot_copd','ipot_dementia',
'ipot_hilipid','ipot_htn','ipot_diabetes','ipot_stroke','ipot_osteop','sleep_2015','o
besity_2015')]
dim(hf_select)

## [1] 6182   42

set.seed(100)
hf_select1 <- as.data.frame(hf_select) #converting to a data frame for sampling; rand
om sampling does not work otherwise
n = nrow(hf_select1) #get total no. of rows

train.index = sample(n,floor(0.7*n)) #randomly select 70% rows from hf_select

#training data set
hforig_train <- hf_select1[train.index,] #this will select rows in train.index
head(hforig_train)

##      hfer_use hfer_nbr_base carefrag2015 lipdrx_any bbrx_any acerx_any
## 3786        0             0    0.6047431          0        1         0
## 503         0             0    0.7472527          0        1         0
## 3430        1             3    0.7574595          0        0         0
## 3696        0             1    0.6900585          1        1         1
## 6131        1             2    0.7526316          0        0         0
## 4090        1             2    0.7564103          1        1         0
##      arbrx_any diurx_any fqrx_any abrx_othr gabarx_any metrx_any dpp4_any
## 3786         0         1        0         0          0         0        0
## 503          0         1        0         1          0         0        0
## 3430         0         0        0         0          0         0        0
## 3696         0         1        1         0          0         0        0
## 6131         0         0        0         0          0         0        0
## 4090         0         1        0         0          0         0        0
##      sulf_any age age_old age_middle polyrx_gn_ge6 anyabuse ins_mcare hmo
## 3786        0  68       0          1             0        0         1   1
## 503         0  82       1          0             0        0         1   0
## 3430        0  70       0          1             0        0         1   1
## 3696        0  66       0          1             1        0         1   0
## 6131        0  71       0          1             0        0         1   1
## 4090        0  87       1          0             0        0         1   0
##      midwest northeast south anx_any deprn ipot_arth ipot_asth ipot_cancer
## 3786       0         0     1       0     0         0         0           0
## 503        0         1     0       0     0         0         0           1
## 3430       0         0     0       0     1         1         0           1
## 3696       0         0     1       1     1         0         0           0
## 6131       0         0     1       0     0         1         0           0
## 4090       1         0     0       0     0         1         0           1
##      ipot_c_arrhy ipot_cad ipot_mi ipot_ckd ipot_copd ipot_dementia
## 3786            0        0       0        0         0             0
## 503             1        0       0        0         1             0
## 3430            1        0       0        1         0             0
## 3696            0        1       0        1         0             0
## 6131            1        0       0        1         0             0
## 4090            1        1       0        0         0             0
##      ipot_hilipid ipot_htn ipot_diabetes ipot_stroke ipot_osteop sleep_2015
## 3786            1        1             1           0           0          0
## 503             0        1             0           0           1          0
## 3430            0        1             1           0           1          1
## 3696            1        1             1           0           0          0
## 6131            1        1             0           1           0          1
## 4090            1        1             1           1           0          0
##      obesity_2015
## 3786            0
## 503             0
## 3430            0
## 3696            1
## 6131            0
## 4090            0
```

```
dim(hforig_train)
```

```
## [1] 4327   42
```

```
#test data set
hforig_test <- hf_select1[-train.index,] #this will select those rows not in train.in
dex
head(hforig_test)
```

```
##    hfer_use hfer_nbr_base carefrag2015 lipdrx_any bbrx_any acerx_any arbrx_any
## 3         0             0    0.5714286          0        1         0         0
## 5         0             0    0.6810631          1        1         0         1
## 8         1             0    0.6000000          1        1         0         0
## 11        0             0    0.5416667          0        0         1         0
## 13        0             0    0.7500000          1        1         0         0
## 15        0             0    0.6719368          0        0         0         0
##    diurx_any fqrx_any abrx_othr gabarx_any metrx_any dpp4_any sulf_any age
## 3          1        0         1          0         0        0        0  88
## 5          1        1         0          1         0        0        0  70
## 8          1        0         1          0         0        0        0  90
## 11         1        1         0          1         0        0        0  83
## 13         1        0         1          0         0        0        1  83
## 15         0        0         0          0         0        0        0  85
##    age_old age_middle polyrx_gn_ge6 anyabuse ins_mcare hmo midwest northeast
## 3        1          0             0        0         1   0       1         0
## 5        0          1             0        0         1   1       0         0
## 8        1          0             0        0         1   0       1         0
## 11       1          0             0        0         1   1       0         0
## 13       1          0             0        0         1   0       1         0
## 15       1          0             0        0         1   1       0         1
##    south anx_any deprn ipot_arth ipot_asth ipot_cancer ipot_c_arrhy ipot_cad
## 3      0       0     0         1         0           1            1        1
## 5      0       1     1         1         1           0            0        0
## 8      0       0     0         0         0           0            1        0
## 11     0       0     1         0         0           0            0        0
## 13     0       0     0         0         1           1            1        0
## 15     0       1     1         0         0           0            0        0
##    ipot_mi ipot_ckd ipot_copd ipot_dementia ipot_hilipid ipot_htn ipot_diabetes
## 3        0        1         0             0            0        0             0
## 5        0        1         1             0            1        1             0
## 8        0        0         1             1            0        0             0
## 11       0        0         0             0            0        1             0
## 13       0        0         1             0            1        1             1
## 15       0        0         1             1            0        1             0
##    ipot_stroke ipot_osteop sleep_2015 obesity_2015
## 3            0           1          1            0
## 5            0           0          0            0
## 8            1           0          0            0
## 11           0           0          0            0
## 13           0           0          0            0
## 15           1           0          0            0
```

```
dim(hforig_test)
```

```
## [1] 1855   42
```

```
library(ROSE)
```

```
set.seed(999)
hf_select_us <- ovun.sample(hfer_use~., data=hf_select, method="under",N=3384)$data
table(hf_select_us$hfer_use)
```

```
##    0    1
## 1692 1692
```

```
#1st set of variables
set.seed(123)
hf_select_us<- as.data.frame(hf_select_us) #converting to a data frame for sampling;
random sampling does not work otherwise
n = nrow(hf_select_us) #get total no. of rows

train.index = sample(n,floor(0.7*n)) #randomly select 70% rows from hf_select

#undersampled training data set
hftrain <- hf_select_us[train.index,] #this will select rows in train.index
head(hftrain)
```

```
##      hfer_use hfer_nbr_base carefrag2015 lipdrx_any bbrx_any acerx_any
## 2463        1             0    0.7692308          1        1         1
## 2511        1             0    0.6666667          0        0         1
## 2227        1             6    0.6203067          0        0         1
## 526         0             0    0.3888889          0        0         1
## 195          0             0    0.7229437          1        0         0
## 2986         1             9    0.7331628          1        1         1
##      arbrx_any diurx_any fqrx_any abrx_othr gabarx_any metrx_any dpp4_any
## 2463         0         1        0         1          0         0        0
## 2511         0         1        0         1          0         0        0
## 2227         0         1        0         1          0         0        0
## 526          0         1        0         0          0         0        0
## 195          0         0        1         0          0         0        0
## 2986         0         1        0         1          0         0        0
##      sulf_any age age_old age_middle polyrx_gn_ge6 anyabuse ins_mcare hmo
## 2463        0  84       1          0             0        0         1   0
## 2511        0  87       1          0             0        0         1   0
## 2227        0  89       1          0             1        0         1   0
## 526         0  77       0          1             0        0         1   1
## 195         0  78       0          1             1        0         1   1
## 2986        0  57       0          0             1        0         1   1
##      midwest northeast south anx_any deprn ipot_arth ipot_asth ipot_cancer
## 2463       0         0     1       0     0         1         0           1
## 2511       1         0     0       0     0         1         0           1
## 2227       0         0     0       0     0         0         0           1
## 526        0         0     0       0     0         1         0           0
## 195        0         0     0       0     0         1         0           0
## 2986       1         0     0       1     0         1         0           1
##      ipot_c_arrhy ipot_cad ipot_mi ipot_ckd ipot_copd ipot_dementia
## 2463            1        0       0        0         1             0
## 2511            1        0       0        0         0             0
## 2227            1        1       0        1         1             0
## 526             0        0       0        1         0             0
## 195             1        1       0        1         1             1
## 2986            1        0       0        1         0             0
##      ipot_hilipid ipot_htn ipot_diabetes ipot_stroke ipot_osteop sleep_2015
## 2463            1        1             0           0           1          0
## 2511            1        1             0           0           0          1
## 2227            1        1             0           0           1          1
## 526             0        1             0           0           0          0
## 195             1        1             1           1           1          0
## 2986            1        1             1           0           0          1
##      obesity_2015
## 2463            0
## 2511            0
## 2227            0
## 526             1
## 195             0
## 2986            1
```

```
dim(hftrain)
```

```
## [1] 2368   42
```

```
#undersampled test data set
hftest <- hf_select_us[-train.index,] #this will select those rows not in train.index
head(hftest)
```

```
##    hfer_use hfer_nbr_base carefrag2015 lipdrx_any bbrx_any acerx_any arbrx_any
## 3         0             4    0.7820513          0        0         0         1
## 6         0             0    0.6666667          1        1         0         1
## 12        0             2    0.4000000          0        0         0         0
## 14        0             0    0.4746377          1        1         0         0
## 15        0             0    0.5454545          0        1         1         0
## 22        0             0    0.5846154          1        1         1         0
##    diurx_any fqrx_any abrx_othr gabarx_any metrx_any dpp4_any sulf_any age
## 3          1        0         0          0         0        0        0  87
## 6          0        0         0          0         0        0        0  75
## 12         0        1         0          0         0        0        0  85
## 14         0        1         0          0         0        0        1  68
## 15         0        0         0          0         0        0        0  88
## 22         1        1         0          1         0        0        0  83
##    age_old age_middle polyrx_gn_ge6 anyabuse ins_mcare hmo midwest northeast
## 3        1          0             1        0         1   0       0         0
## 6        0          1             0        0         1   0       0         1
## 12       1          0             1        0         1   0       0         0
## 14       0          1             1        0         0   0       0         0
## 15       1          0             0        0         1   1       0         1
## 22       1          0             1        0         1   1       0         1
##    south anx_any deprn ipot_arth ipot_asth ipot_cancer ipot_c_arrhy ipot_cad
## 3      1       0     1         0         1           0            0        0
## 6      0       0     0         0         0           0            0        1
## 12     1       0     0         0         0           0            0        0
## 14     1       0     0         1         0           0            0        1
## 15     0       0     0         1         0           0            0        0
## 22     0       0     1         0         1           0            0        0
##    ipot_mi ipot_ckd ipot_copd ipot_dementia ipot_hilipid ipot_htn ipot_diabetes
## 3        0        0         1             0            1        1             0
## 6        0        0         0             0            1        1             1
## 12       0        1         0             1            0        1             0
## 14       0        1         1             0            1        1             1
## 15       0        0         0             1            0        1             0
## 22       0        1         1             0            0        1             0
##    ipot_stroke ipot_osteop sleep_2015 obesity_2015
## 3            0           0          0            1
## 6            0           0          0            0
## 12           0           0          0            1
## 14           1           0          0            0
## 15           0           0          0            0
## 22           0           0          0            0
```

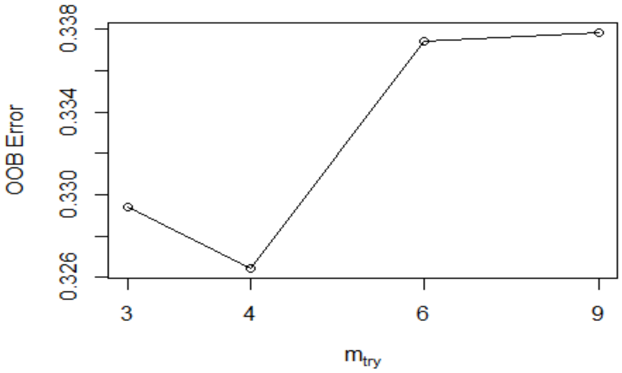
```
dim(hftest)
```

```
## [1] 1016   42
```

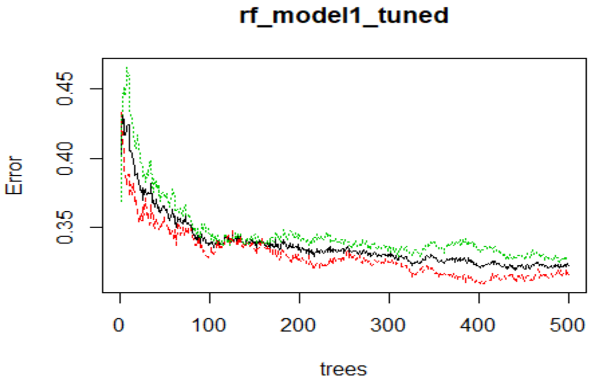
```
#random forest method
library(randomForest)
```

```
# Algorithm Tune (tuneRF)
ind_vars = hftrain[c('hfer_nbr_base','carefrag2015','lipdrx_any','bbrx_any','acerx_an
y','arbrx_ny','diurx_any','fqrx_any','abrx_othr','gabarx_any','metrx_any','sulf_any',
'dpp4_any','age','polyrx_gn_ge6','ins_mcare','hmo','anx_any','deprn','ipot_arth','ipo
t_asth','ipot_cancer','ipot_c_arrhy','ipot_cad','ipot_mi','ipot_ckd','ipot_copd','ipo
t_dementia','ipot_hilipid','ipot_htn','ipot_diabetes','ipot_stroke','ipot_osteop','sl
eep_2015','obesity_2015','northeast','midwest','south')]
```

```
set.seed(999)
bestmtry <- tuneRF(ind_vars,
                   hftrain$hfer_use,
                   stepFactor=1.5,
                   improve=1e-5,
                   ntree=500)

## mtry = 6  OOB error = 33.74%
## Searching left ...
## mtry = 4     OOB error = 32.64%
## 0.03254068 1e-05
## mtry = 3     OOB error = 32.94%
## -0.009055627 1e-05
## Searching right ...
## mtry = 9     OOB error = 33.78%
## -0.03492885 1e-05
```



```
print(bestmtry)

##       mtry  OOBError
## 3.OOB    3 0.3293919
## 4.OOB    4 0.3264358
## 6.OOB    6 0.3374155
## 9.OOB    9 0.3378378

#random forest method
library(randomForest)
#use set seet to make it repeatable again#
set.seed(111)
rf_model1_tuned<-randomForest(hfer_use~hfer_nbr_base+carefrag2015+lipdrx_any+bbrx_any
+acerx_any+arbrx_any+diurx_any
                        +fqrx_any+abrx_othr+gabarx_any+metrx_any+sulf_any+dpp4_any+ag
e+polyrx_gn_ge6
                        +ins_mcare+hmo
                        +anx_any+deprn+ipot_arth+ipot_asth+ipot_cancer+ipot_c_arrhy+i
pot_cad+ipot_mi+ipot_ckd
                        +ipot_copd+ipot_dementia+ipot_hilipid+ipot_htn+ipot_diabetes+
ipot_stroke+ipot_osteop
                        +sleep_2015+obesity_2015+northeast+midwest+south,
                        data=hftrain,
                        ntreeTry = 500,
                        mtry = 4,
                        importance = TRUE)

#Print results from Model 1
print(rf_model1_tuned)
## Call:
##  randomForest(formula = hfer_use ~ hfer_nbr_base + carefrag2015 +      lipdrx_any
+ bbrx_any + acerx_any + arbrx_any + diurx_any +      fqrx_any + abrx_othr + gabarx_a
ny + metrx_any + sulf_any +      dpp4_any + age + polyrx_gn_ge6 + ins_mcare + hmo + a
nx_any +      deprn + ipot_arth + ipot_asth + ipot_cancer + ipot_c_arrhy +      ipot_
cad + ipot_mi + ipot_ckd + ipot_copd + ipot_dementia +      ipot_hilipid + ipot_htn +
ipot_diabetes + ipot_stroke + ipot_osteop +      sleep_2015 + obesity_2015 + northeas
t + midwest + south,      data = hftrain, ntreeTry = 500, mtry = 4, importance = TRUE
)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 32.26%
## Confusion matrix:
##     0   1 class.error
## 0 809 374   0.3161454
## 1 390 795   0.3291139

#error rate of random forest model 1
plot(rf_model1_tuned)
```



```
library(caret)
#predict using training data#
pred_model1<-predict(rf_model1_tuned,hftrain)

head(pred_model1)

## 2463 2511 2227  526  195 2986
##    1    1    1    0    0    1
## Levels: 0 1

head(hftrain$hfer_use)
```

```
## [1] 1 1 1 0 0 1
## Levels: 0 1
```

```
confusionMatrix(pred_model1,hftrain$hfer_use, positive = "1")
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1183    4
##          1    0 1181
##                                           
##                Accuracy : 0.9983          
##                  95% CI : (0.9957, 0.9995)
##     No Information Rate : 0.5004          
##     P-Value [Acc > NIR] : <2e-16          
##                                           
##                   Kappa : 0.9966          
##                                           
##  Mcnemar's Test P-Value : 0.1336          
##                                           
##             Sensitivity : 0.9966          
##             Specificity : 1.0000          
##          Pos Pred Value : 1.0000          
##          Neg Pred Value : 0.9966          
##              Prevalence : 0.5004          
##          Detection Rate : 0.4987          
##    Detection Prevalence : 0.4987          
##       Balanced Accuracy : 0.9983          
##                                           
##        'Positive' Class : 1               
## 
```

```
#predict using original test data
pred_test1<-predict(rf_model1_tuned,hforig_test)
pred_test1_prob<-predict(rf_model1_tuned,hforig_test, type = "prob")
head(pred_test1_prob)
```

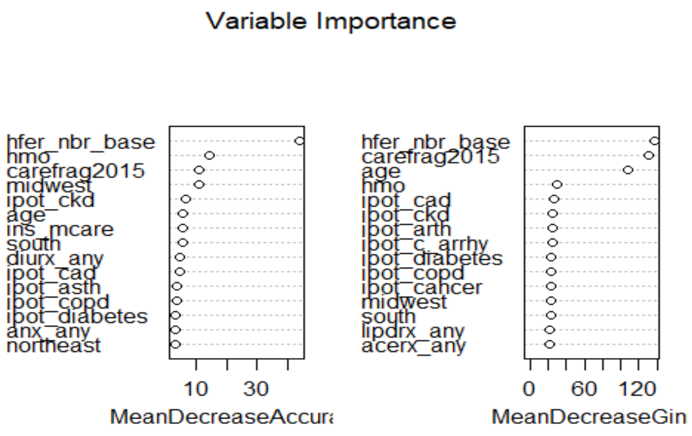
```
##        0     1
## 3  0.560 0.440
## 5  0.680 0.320
## 8  0.258 0.742
## 11 0.844 0.156
## 13 0.712 0.288
## 15 0.850 0.150
```

```
pred_test1_prob <- pred_test1_prob[,"1"]
head(pred_test1_prob)
```

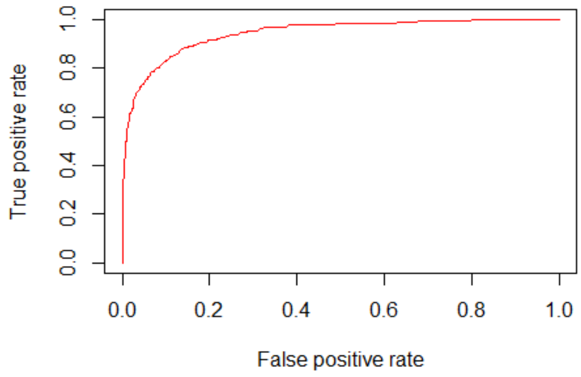
```
##     3     5     8    11    13    15 
## 0.440 0.320 0.742 0.156 0.288 0.150
```

```
#get confusion matrix for original test#
confusionMatrix(pred_test1,hforig_test$hfer_use, positive = "1")
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1034   38
##          1  311  472
##                                           
##                Accuracy : 0.8119          
##                  95% CI : (0.7933, 0.8294)
##     No Information Rate : 0.7251          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.5953          
##                                           
##  Mcnemar's Test P-Value : < 2.2e-16       
##                                           
##             Sensitivity : 0.9255          
##             Specificity : 0.7688          
##          Pos Pred Value : 0.6028          
##          Neg Pred Value : 0.9646          
##              Prevalence : 0.2749          
##          Detection Rate : 0.2544          
##    Detection Prevalence : 0.4221          
##       Balanced Accuracy : 0.8471          
##                                           
##        'Positive' Class : 1               
## 

#Top predictors
varimpplot <- varImpPlot(rf_model1_tuned, n.var = 15, sort = TRUE, main = "Variable I
mportance")
```



```
library(gplots)
library(ROCR)
library(pROC)
#get ROC
rocrpred<- prediction(pred_test1_prob,hforig_test$hfer_use)#, label.ordering = c("non
e", "any"))
rocrperf<- performance(rocrpred, 'tpr', 'fpr')
plot(rocrperf, add = F, col = 'red')
```



```
#print auc#
rocrauc<- performance(rocrpred, measure = 'auc')
print(rocrauc@y.values)

## [[1]]
## [1] 0.9430702
ci.auc <- ci.auc(hforig_test$hfer_use, pred_test1_prob)
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
print(ci.auc)
## 95% CI: 0.9316-0.9546 (DeLong)

#confusion matrix
library(caret)
cm_rf <- confusionMatrix(pred_test1, hforig_test$hfer_use, positive = "1")
cm_rf

## Confusion Matrix and Statistics
##           Reference
## Prediction    0    1
##          0 1034   38
##          1  311  472
##
##                Accuracy : 0.8119
##                  95% CI : (0.7933, 0.8294)
##     No Information Rate : 0.7251
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5953
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9255
##             Specificity : 0.7688
##          Pos Pred Value : 0.6028
##          Neg Pred Value : 0.9646
##              Prevalence : 0.2749
##          Detection Rate : 0.2544
##    Detection Prevalence : 0.4221
##       Balanced Accuracy : 0.8471
##
##        'Positive' Class : 1
##

cm_rf_pr <- confusionMatrix(pred_test1, hforig_test$hfer_use, mode = "prec_recall", p
ositive = "1")
cm_rf_pr

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1034   38
##          1  311  472
##
##                Accuracy : 0.8119
##                  95% CI : (0.7933, 0.8294)
##     No Information Rate : 0.7251
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5953
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##               Precision : 0.6028
##                  Recall : 0.9255
##                      F1 : 0.7301
##              Prevalence : 0.2749
##          Detection Rate : 0.2544
##    Detection Prevalence : 0.4221
##       Balanced Accuracy : 0.8471
##
##        'Positive' Class : 1
##
```

Partial dependence plot for RF model 1

```
library(pdp)

## Warning: package 'pdp' was built under R version 3.6.3

##
## Attaching package: 'pdp'

## The following object is masked from 'package:purrr':
##
##     partial

#Top 10 variables original dataset
set.seed(242)
imp1 <- importance(rf_model1_tuned)
imp1
```

```
##                        0           1 MeanDecreaseAccuracy MeanDecreaseGini
## hfer_nbr_base 37.9672413 31.69336327          44.08377548        137.03336
## carefrag2015  10.4608211  4.24562341          10.89968741        130.09695
## lipdrx_any    -2.5601229  4.78048639           2.04418855         22.33132
## bbrx_any      -2.3729118  5.27503586           2.23643454         21.36956
## acerx_any      0.1331093  1.71712378           1.30584286         22.31277
## arbrx_any      1.8492774  0.30319629           1.68750788         20.09725
## diurx_any     -5.0713278  9.39094258           4.32418246         21.42383
## fqrx_any      -2.0382352  1.41728811          -0.30646592         20.05076
## abrx_othr     -1.4625565  2.71553019           0.93644702         20.70491
## gabarx_any    -3.7722654  0.49641120          -2.40978381         15.97759
## metrx_any      2.2282682 -1.76665163           0.12370855         12.35641
## sulf_any      -1.4856569  3.61339175           1.51202162         11.80321
## dpp4_any       0.5362780  2.61770132           2.47009631          7.03576
## age           -1.6994047  8.57539381           5.47750677        108.46775
## polyrx_gn_ge6 -1.0989806  4.39272934           2.54858348         21.66605
## ins_mcare      5.8584641  0.91159792           5.20582688          9.75016
## hmo           15.8185951  2.59289870          14.09555785         29.86181
## anx_any       -0.5070949  4.55717651           3.08334030         19.34062
## deprn         -1.3965579  0.10010417          -0.90352708         21.11822
## ipot_arth     -1.1000132  1.66661520           0.36980653         25.01414
## ipot_asth      3.5929585  1.45577013           3.54877800         18.57081
## ipot_cancer    2.7456150 -0.13317346           1.92121773         23.25871
## ipot_c_arrhy   2.7190771 -0.08009677           1.95925267         24.38003
## ipot_cad       0.2453323  5.83131669           4.23808921         26.23646
## ipot_mi       -0.2403934  0.97937367           0.52862988         11.14757
## ipot_ckd       4.5804428  4.24242907           6.42749927         25.50455
## ipot_copd      1.3811830  3.05323679           3.24835354         23.73637
## ipot_dementia  2.3347037 -1.74374286           0.48722012         17.46825
## ipot_hilipid  -0.2331155  1.20571046           0.66311835         19.94742
## ipot_htn      -1.8232423  2.67448562           0.51214389         10.27341
## ipot_diabetes -0.8857299  5.07297520           3.08591910         23.90229
## ipot_stroke    1.5296095 -0.22298623           0.88030321         21.21619
## ipot_osteop    1.4972770 -1.35086018           0.05416766         18.87292
## sleep_2015    -2.0024586  0.72516398          -0.70714800         20.60697
## obesity_2015  -2.9784928  2.38712362          -0.42696839         20.21741
## northeast     -1.4733704  5.01404526           2.74216816         14.73760
## midwest        9.6635759  4.78474310          10.85688006         22.77794
## south          1.2319487  5.58399229           5.20433935         22.64246
```

```
impvar1 <- rownames(imp1) [order(imp1[, 1], decreasing=TRUE)]
impvar1
```

```
##  [1] "hfer_nbr_base" "hmo"           "carefrag2015"  "midwest"
##  [5] "ins_mcare"     "ipot_ckd"      "ipot_asth"     "ipot_cancer"
##  [9] "ipot_c_arrhy"  "ipot_dementia" "metrx_any"     "arbrx_any"
## [13] "ipot_stroke"   "ipot_osteop"   "ipot_copd"     "south"
## [17] "dpp4_any"      "ipot_cad"      "acerx_any"     "ipot_hilipid"
## [21] "ipot_mi"       "anx_any"       "ipot_diabetes" "polyrx_gn_ge6"
## [25] "ipot_arth"     "deprn"         "abrx_othr"     "northeast"
## [29] "sulf_any"      "age"           "ipot_htn"      "sleep_2015"
## [33] "fqrx_any"      "bbrx_any"      "lipdrx_any"    "obesity_2015"
## [37] "gabarx_any"    "diurx_any"
```

```
op <- par(mfrow=c(2,3))
for (i in seq_along(impvar1))  {
     partialPlot(rf_model1_tuned, hftrain, impvar1[i],xlab = impvar1[i],
                 main = paste("Partial Dependence on", impvar1[i]), which.class = "1"
)
}
```