9-12-2023 2:30 PM

Association of Fall-Related Injuries and Different Diagnoses in Older Adults of Ontario: A Machine Learning Approach

Sorour Rostampour, *Western University*

Supervisor: Zecevic, Aleksandra, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Health and Rehabilitation Sciences

Abstract

Falls are the leading cause of injury-related hospitalizations among older adults in Canada. This study aimed to identify the most informative diagnostic categories associated with fall-related injuries (FRIs) using three machine learning algorithms: decision tree, random forest, and extreme gradient boosting tree (XGBoost). Secondary data from two Ontario health administrative databases (NACRS, DAD) covering the period 2006-2015 were analyzed. Older adults (aged ≥ 65 years) who sought treatment for FRIs in emergency departments (ED) or hospitals, as indicated by Canadian version of the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10-CA) codes for falls and injuries, were included in the study. Accuracy, sensitivity, specificity, precision, and F1 score measures were calculated for each model. A total of 631,339 ED admissions and 304,495 hospitalizations were recorded due to FRIs. The random forest model demonstrated the highest sensitivity and accuracy in both datasets. Dyspnea and secondary malignant neoplasm of liver and intrahepatic bile duct were the most informative ICD-10-CA code and disease for FRIs among older adults admitted to ED and hospitals. These findings indicate that machine learning models can also be used to study FRIs as they are capable of handling large datasets and providing a better than 60% accuracy. Also, diagnostic categories linked to FRIs have a potential to enhance healthcare providers ‘ability to prevent FRIs in the future.

Keywords

older adults, falls, injury, machine learning, artificial intelligence, prediction, diagnosis, ICD-10-CA

Summary for Lay Audience

In Canada, falls are responsible for many emergency room visits and hospitalizations among older adults. This study explored a connection between injuries older adults experienced after a fall and other diagnoses they got while in an emergency room or a hospital. We used advanced computer calculations, also called machine learning, to determine which diagnostic categories are closely related with fall related injuries and provide the most useful information. Data from two large health databases (NACRS, DAD) covering the years 2006 to 2015 in Canadian province of Ontario were analyzed. Three machine learning algorithms were compared for accuracy and sensitivity. The results revealed that the random forest model was the most accurate and sensitive. Two diagnostic categories were identified as informative: in the emergency department, the presence of dyspnea or shortness of breath was found to be a notable factor, and in hospitals the presence of an abnormal tumor in the bile duct and liver, also known as the secondary malignant neoplasm of liver and intrahepatic bile duct, were identified as highly relevant. These findings show that machine learning models can be used in studies about fall-related injuries (FRIs). These models can handle big amounts of data and have accuracy higher than 60%. The most informative diagnostic categories associated with FRIs can help healthcare providers better understand the risks of falls in older adults and improve their ability to prevent FRIs in the future.

This thesis is dedicated to Kurdistan, land of resilience and hope.

Acknowledgments

This study was supported by IC/ES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by IC/ES or the Ontario MOHLTC is intended or should be inferred 1.[1]

I would like to begin by expressing my utmost gratitude to my supervisor, Dr. Aleksandra Zecevic, for her unwavering guidance and support throughout my MSc studies. Words cannot adequately convey the depth of my gratitude for your invaluable contributions to both my professional development and personal growth. You are an exceptional researcher and supervisor, and I am sincerely grateful for the ways in which you have encouraged, challenged, comforted, and cared for me throughout this lengthy journey. Your dedication to research, exceptional communication skills, openness to novelty and the unknown, and unwavering optimism have deeply influenced me and enriched my experience. Upon my arrival in this new country and academic environment, your generous knowledge and selfless support provided me with a sense of reassurance. As an immigrant in an unfamiliar land, meeting someone as warm, welcoming, and positive as you in my early days here reaffirmed that I had made the right decision to move to Canada. It filled me with gratitude and dispelled any worries I may have had. I deeply appreciate that you arranged regular individual meetings with me to discuss my research progress and that you promptly addressed every question I posed. Your understanding and support during times of need were truly invaluable. Furthermore, I am sincerely grateful for your assistance in connecting me with other academics and professors whenever I required additional help. Your humble support in facilitating these connections is greatly appreciated. I hope that our relationship can extend beyond the academic realm and that our mutual learning and growth continue indefinitely.

I would like to thank Dr. Richard G. Booth, one of my esteemed advisory committee members. Your strategic wisdom and professional expertise in handling big data have been

---

[1] Italicized portions of text are a required wording provided by IC/ES and obligatory for inclusion in studies using IC/ES data (ICES, 2022).

invaluable throughout the development of this thesis. Without your exceptional guidance and unwavering support, this research endeavor would not have been possible. I am immensely thankful for your instrumental role in crafting the wording of the result section of my thesis and preparing the manuscripts for submission. Your generosity in sharing your wealth of experience working with IC/ES, offering insights on data management, and providing constructive feedback and advice on my writing has been truly remarkable.

Dr. Tarun Katapally, a valued member of my advisory committee, for his invaluable expertise in quantitative studies. I greatly appreciate his direct and concise feedback and suggestions, which have been instrumental in shaping the development of my research.

I would also like to extend my heartfelt thanks to Dr. Katarina Grolinger for her invaluable guidance in utilizing statistical computing software such as R and SAS. Her insightful comments on the performance metrics of the selected machine learning models have been immensely valuable to my work.

Finally, special acknowledgement goes to my thesis examination committee members. Thank you very much for reviewing my thesis and challenging me with insightful ideas to improve.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

CDC	Centers for Disease Control and Prevention
CIHI	Canadian Institute for Health Information
DAD	Discharge Abstract Database
ED	Emergency Department
ICD-10	International Statistical Classification of Disease and Related Health Problems 10th Edition
ICD-10-CA	International Statistical Classification of Disease and Related Health Problems 10th Edition, with Canadian Enhancements
ICD-11	International Statistical Classification of Diseases and Related Health Problems, 11th Revision
IC/ES	Institute for Clinical Evaluative Sciences, in 2018, the institute formerly known as the Institute for Clinical Evaluative Sciences formally adopted the initialism IC/ES as its official name. This change acknowledges the growth and evolution of the organization's research since its inception in 1992, while retaining the familiarity of the former acronym within the scientific community and beyond.
IKN	IC/ES Key Number
NACRS	National Ambulatory Care Reporting System
PHAC	Public Health Agency of Canada
RPDB	Registered Persons Database
FRIs	Fall-Related Injuries
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting
CDC	Centers for Disease Control and Prevention
SQL	Structured Query Language
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

Chapter 1

1. Introduction and Literature Review

This thesis describes a population-based study that examined FRIs in older adults, defined as people who are 65 years or older. The first section of this thesis is a brief introduction to population aging in Canada and globally. This introduction provides the context on the impact of falls and injuries they cause at a population level. The research on falls and FRIs is vast, therefore a literature review included here focused on population-based studies. Methodology of choice for the study are machine learning algorithms applied to administrative health databases that contain diagnostic codes according to ICD-10-CA. Hence, the literature review also includes a synthesis of knowledge from the key studies on using machine learning and diagnostic codes in prediction of injuries, falls, and FRIs in older adult population. The review highlights trends, identifies gaps, critiques methodologies, and builds compelling rationale for a new research study that challenges the unresolved problems in the application of machine learning in research of FRIs. The second chapter outlines research methods and the third describes findings from this study. The final chapters, discussion and conclusions, tie in findings from the previous literature, identify strengths and limitations, and provide guidance for future research.

1.1 Population Aging

Increased longevity is considered an accomplishment of the 20th and 21st centuries, however the continuously rising average age of the human population poses a significant challenge to supporting older adults in communities around the world (Sander et al., 2015). By the middle of 21st century, the global population of older adults is projected to reach 1.5 billion (Colombo et al., 2012). Population aging is also evident in Canada, where the population of older adults is currently higher than that of youth aged 15 or below (Statistics Canada, 2017). At the provincial level, Ontario has one of the fastest-growing older adult populations. In 2016, older adults accounted for 16.4% of Ontario's population and this number is projected to increase to 25% by 2041. This is an increase from 3 million in 2016 to 4.6 million in 2041 (Government of Ontario Ministry for Seniors and Accessibility, 2017). The demand to enhance the health conditions, reduce preventable injuries and improve the quality of life of older adults is increasing

both globally and provincially (Colombo et al., 2012; Geithner & McKenney, 2010; Lagiewka, 2012).

1.2 Falls and Fall-Related Injuries

Falling is a common issue among older adults, which may lead to injuries (Peel, 2011). There are different definitions for falls. According to the World Health Organization (WHO), a fall is defined as an event that leads an individual to come to rest on the ground, floor, or lower level accidentally (WHO, 2008). Factors causing falls are categorized into intrinsic and extrinsic. Intrinsic factors originate from individuals' conditions, such as, decreased balance, vision loss, and polypharmacy (Källstrand-Ericson & Hildingh, 2009). Extrinsic factors are related to the environment and context such as poor lighting and slippery surfaces (Lord et al., 2006). A fall can be the direct or indirect result of the individuals' conditions, the tasks they perform, their environment, or a combination of the three (Erkal, 2010).

According to Centers for Disease Control and Prevention (CDC) in the United States, falls are not considered a normal part of aging (CDC, 2022). However, they happen frequently among people aged 65 years or older. Moreland et al. (2020) reported that about 36 million falls happen annually among older adults in the United States. As a result, prevention of falls has been of interest among researchers around the world, and many countries consider fall prevention a health priority (Lamb et al., 2005).

Because of their elevated risk of injury, older adults are more likely to suffer significant consequences from falls. For adults over the age of 65, falling is the most common cause of fatal and nonfatal injuries (Stevens et al., 2008), and an important challenge for healthcare professionals (Gimigliano, 2020). Falls can lead to serious personal, societal, and economic issues. Peel (2011) found that 20-23% of community-dwelling older adults living independently have reported a fall at least once a year. These increases to 60% among the oldest-old group, or individuals 85 years of age and older (Peel, 2011). Bergland and Wyller (2004) defined serious FRIs as "fractures, dislocations, head injuries resulting in loss of consciousness, and other injuries resulting in medical care". In this study, FRIs were operationally defined as any injury diagnosis caused by a fall resulting in a visit to the ED, hospital or death after admission to the ED or hospital. To the author's knowledge, there is no consensus on the number of minor,

moderate and serious injuries caused by falls in older adults. For instance, Ambrose et al. (2013) suggested that 30-50% of falls in community-dwelling older adults may cause minor injuries such as lacerations or bruises and 5-10% of falls can cause serious injuries (e.g., hip fractures or traumatic brain injuries). However, Peel (2011) has estimated that, on average, 10-15% of falls may result in serious injuries. Additionally, between 30% and 40% of older adults who experienced an injury due to fall (Stevens et al., 2008) while up to 6% of these falls resulted in severe injuries, including fractures, concussions, and head injuries that required hospitalization (Nachreiner et al., 2007). In Canada, Scott et al. (2004) reported that about 50% of older adults who fall sustain minor injuries, while 5-25% experience severe injuries, such as sprains or fractures. Requiring ED or hospital level of care. This indicates that falls cause a higher risk of hospitalization (Gillespie et al., 2012; Rubenstein, 2006).

The most common types of FRIs among older adults include cuts, bruises, fractures, sprains, and head injuries (Milat et al., 2011). The impact of FRIs on older adults, healthcare systems, and society is substantial (Rubenstein, 2006). The treatment of FRIs can be costly. In the United States, approximately $50 billion is expended annually on medical expenses associated with non-fatal fall injuries, and also $754 million is allocated for medical costs linked to fatal falls (Florence et al., 2018). A recent report by Parachute (2021) revealed that in Canada, injurious falls among older adults incurred a financial burden of $5.6 billion Canadian Dollar for the health care system. In Ontario, the direct healthcare costs attributed to "falls from the same level" were $458 million, with the total cost reaching $610 million (SMARTRISK, 2009). These figures illustrate the significant economic impact of falls on older adults and the healthcare system in Canada. Research on FRIs illustrate that relatively small portion of incidents, namely falls, leads to extreme consequences such as serious injuries and related healthcare cost (Zecevic et al., 2012). Further research that can support the development of reliable solutions for reduction of high-risk incidents that cause injuries is needed.

1.3 Risk Factors for Falls and Injuries

The risk factors for falls are very well researched. A study summarized 16 articles

in a literature review and reported the following risk factors for falls in older individuals: vision deficits or impairment, mobility limitations, balance deficits, gait deficits, and weakness (Rubenstein, 2006). Similarly, findings from Jehu et al. (2021), underscore the significance of four domains of falls-risk factors: balance and mobility, medications, psychological factors, and sensory and neuromuscular risk factors. Each of these domains was found to be linked to a higher likelihood of experiencing repeated falls.

Gait and balance problems emerged as prominent risk factors associated with falls. For instance, Deandrea et al. (2010) reported that gait and balance problems, vertigo, and use of walking aids were associated with falling among recurrent fallers. As individuals grow older, their ability to maintain balance while controlling their walking tends to diminish, leading to an increase in gait variability. This age-related decline in balance and gait poses a higher risk of falls among older adults (Osoba et al., 2019).

Polypharmacy and medication use, increasingly prevalent in older adults, were reported as strong risk factors. For example, Zaninotto et al. (2020) found that the likelihood of being hospitalized after experiencing a fall raised with the number of medications taken. The frequency of hospital admissions due to falls started at 1.5% for individuals not taking any medications, increased to 4.7% for those on 1-4 medications, further increased to 7.9% for individuals with polypharmacy (5 to 9 medications), and peaking at 14.8% for those with heightened polypharmacy (10+ medications). Shuto et al. (2010) reported that the first-time use of medications like antihypertensives, antiparkinsonians, anti-anxiety drugs, and hypnotics was closely linked to a higher likelihood of experiencing falls.

The research on risk factors for FRIs in older adults is rich but leaves space for improvement for the following reasons. Some studies are either subsidiary to fall prediction studies or conducted on a specific subsection of population that cannot be generalized. Examples of such subgroups of older adults are people with vestibular dysfunction (Marchetti, 1994), kidney problems (Kistler et al., 2018) or lower limb amputations (Wong et al., 2016). Other studies are focused on particular cultural group such as Chinese (Pi et al., 2016) Taiwanese (Li et al., 2016), Korean (Kim et al., 2018), and Indonesian (Pengpid & Peltzer, 2018). Frequently, research on risk factors for injuries due to falls is combined with risk factors for falls. Therefore,

it is important to investigate FRIs separately, to gather evidence and develop strategies for injury prevention, given their significant impact on morbidity and mortality (McClure et al., 2005). Although fall prevention is not within the scope of this research, some findings of this research have the potential to enrich fall prevention.

Association between some diseases and FRIs or falls were reported in previous research. They include osteoarthritis (Barbour et al., 2019), multiple sclerosis (Peterson et al., 2008), musculoskeletal disorders (Aoyagi et al., 1998), Parkinson's disease (Balash et al., 2005; Bloem et al., 2001), and dementia (Borges et al., 2015; Li, 2016; Pedroso et al., 2012; Van Doorn et al., 2003). Barbour et al. (2019) found out that the presence of knee osteoarthritis was linked to a higher likelihood of experiencing injurious falls in older men living in the community. However, there was no evidence to suggest that knee osteoarthritis predicted injurious falls among women. Aoyagi et al. (1998) reported that musculoskeletal disorders can increase the prevalence of falls and fractures among older adults. Frequent falls are prevalent in individuals diagnosed with Parkinson's disease (Balash et al., 2005; Bloem et al., 2001). A study conducted by Peterson et al. (2008) surveyed middle aged and older adult participants with multiple sclerosis about their medical care for injuries resulting from falls, revealing a significant number of respondents who reported injurious falls. The results indicate that addressing the fear of falling and osteoporosis management were crucial elements in a comprehensive program aimed at preventing FRIs among individuals with multiple sclerosis. Van Doorn et al. (2003) reported that dementia is an independent risk factor for falling, and the increased frequency of falls among individuals with dementia, compared to those without the condition, exposes them to a greater likelihood of experiencing injurious falls over time. Alzheimer's disease was also identified as a risk factor for falling (Borges et al., 2015; Pedroso et al., 2012) and hospitalization due to fall-related bone fractures (Li, 2016) .

According to recently published "World guidelines for falls prevention and management for older adults: A global initiative" by Montero-Odasso and colleagues (2022), older adults should be categorized into low, intermediate, or high-risk groups for falls. This categorization can be determined through opportunistic case-finding, where physicians ask about recent experiences of falls during visits, or when older adults

present with a fall or an FRI. For those at high risk, a comprehensive multifactorial fall risk assessment should be offered, encompassing mobility, sensory function, activities of daily living, cognitive function, autonomic function, nutrition history, medication history, environmental risk, and disease history. The guideline specifies several diseases associated with falls, including osteoarthritis, diabetes mellitus, neurological disorders (such as Parkinson's disease, polyneuropathy, and stroke), cardiovascular diseases, depressive disorders, cognition, delirium, anemia, thyroid disease, electrolyte disorders, frailty, sarcopenia, fracture risk (osteoporosis), and acute pneumonia (Montero-Odasso et al., 2022). The guideline provides some evidence about the associations between falls, injurious falls, and various diseases. However, it is still unclear how other diseases (identified by medical diagnoses in an emergency room or a hospital) relate to FRIs. The specific focus on studying the connection between diagnoses and their combinations would help fill the gap in understanding how other diseases relate to FRIs.

1.4 Machine Learning

The term data science has been frequently used over the past decade to describe the new 'science' created to tease out the hidden knowledge and value in the data collected from different businesses such as social media or retail (Hastie et al., 2009; Provost & Fawcett, 2013). While terms data science and machine learning sometimes may be used interchangeably, they are not the same thing. Data science is a collection of specialties that are deployed to use data for decision making, while machine learning is focused on the tools or algorithms used in data science (Provost & Fawcett, 2013). Machine learning is a field of artificial intelligence where algorithms learn from data to improve their performance on specific tasks or predictions (Hastie et al., 2009; Provost & Fawcett, 2013).

Machine learning is a subset of artificial intelligence that uses computer science algorithms that can improve automatically by learning from data. Machine learning algorithms are broadly divided into two classes: supervised and unsupervised learning (Hastie et al., 2009). In unsupervised learning, a model is trained to find patterns or structures in a dataset without being provided with explicit labels or targets. It involves learning from unlabeled data to discover underlying relationships, groupings, or representations within the data (Hastie et al., 2009). Unlike supervised learning, unsupervised learning algorithms do not have a previously

known outcome and are focused on pattern detection, such as naturally formed communities. Clustering algorithms, such as k-means, belong to unsupervised learning family (Hastie et al., 2009). An example of using unsupervised learning to predict falls could involve sensor data collected from individuals, such as accelerometers or gyroscopes. The data could include measurements related to body movement, orientation, or other relevant features. Supervised learning describes a type of algorithm in which the format of the outcome is known (Provost & Fawcett, 2013). For example, a supervised learning algorithm can be developed to predict whether an older adult that fell will sustain a certain type of injury. In this case the algorithm predicts a label or class, which, in this example is a binary variable (i.e., 0 or 1 or yes or no) for injury. Since it predicts a class, this algorithm is also called a classifier (Provost & Fawcett, 2013). Common examples of classification algorithms are logistic regression, naïve Bayes classifier, support vector machines, and decision tree, which is probably the most popular algorithm (X. Wu et al., 2008). The second category of supervised learning models is trained to predict a numeric value (e.g., the time between two consecutive falls for an older adult) rather than a class. These algorithms are known as regression algorithms (Provost & Fawcett, 2013). The well-known example of these algorithms is linear regression. Another example is neural network, when trained to predict a value (Hastie et al., 2009).

Nowadays, using machine learning is feasible because of advances in computer science theory, on one side, and the unprecedented availability of large datasets. Many new technologies such as wearable sensors, social media, relational databases, and smartphones allow automatic and continuous data collection. More importantly, the culture of data collection and management have allowed researchers in many disciplines, including gerontology and risk predictions, to collect large datasets at a rapidly growing rate (Provost & Fawcett, 2013; Speiser et al., 2021). Complex data requires a type of analysis that is sophisticated enough to capture the nuances of data and its internal patterns. That is one of the reasons behind the popularity of machine learning as it can mine multidimensional large datasets (Provost & Fawcett, 2013). Machine learning offers several distinct advantages over much simpler classification algorithms such as logistic regression. First and foremost, machine learning models are capable of automatically

learning and adapting to complex patterns and features within data. This adaptability enables them to handle high-dimensional data sets, where simpler algorithms may struggle to reach an acceptable accuracy (Hastie et al., 2009). Additionally, machine learning algorithms excel at variable extraction, reducing the need for manual variable engineering, which can be time-consuming and error-prone in traditional approaches (Hastie et al., 2009; Provost & Fawcett, 2013). Furthermore, they can continually improve their performance through iterative training and fine-tuning, making them well-suited for dynamic and evolving data environments (Hastie et al., 2009; Provost & Fawcett, 2013). Lastly, machine learning algorithms often outperform simpler classification algorithms in terms of predictive accuracy, as they can capture subtle relationships and dependencies within data that might be overlooked by less sophisticated methods, leading to more precise and robust classification results (Hastie et al., 2009).

1.5 Machine Learning in Gerontology

Researchers in gerontology have embarked on adopting data science and machine learning to address problems in their field. To identify what is already known in the literature in recent years, a review was completed in 2021 to map the application of machine learning to the problems of gerontology, with emphasis on falls in older adults. Additionally, the review aided in determining the appropriate machine learning models to employ. Moreover, it helped to identify the prevalent performance metrics suitable for the present study. The review uncovered that although the application of machine learning is not prevalent in gerontology, it is rapidly gaining popularity, due to its high performance and predictive power. A summary of studies included in the literature review is presented in Table 1-1.

Table 1-1

Summary of the Studies Identified in the Literature Review on the Use of Machine Learning in Gerontology

Author	Year	Target variable(s) and problem(s) addressed	Target population	Variables	Machine learning algorithm(s)	Performance metric(s) of machine learning model(s)
Ateeq	2018	Prediction of FRIs	Canadians aged 12 years or above	Income, alcohol use, physical activities, dwelling and household variables, consultation with health professional, fruit and vegetable consumption, height, and weight	LR, RF, and PCA	Accuracy, and sensitivity
Heo et al.	2019	Favorable and poor outcomes of stroke in people with acute stroke	Individuals with acute stroke	Patient demographics (age, and sex), NIHSS, time from onset to admission, stroke subtypes, previous diseases, blood test values, medication history, smoking status	DNN, RF, and LR	One-number accuracy
Badgujar et al.	2020	Falls prediction	Older adults	Fall and gait patterns	SVM, and DT	Confusion matrix, and related measures (accuracy, and sensitivity)
Kalatzis et al.	2020	Prediction of stress (yes or no)	Older adults	Heart rate variables	ANN	One-number accuracy
Park	2020	Early detection of mild cognitive impairment	Older adults aged 65 or older with mild cognitive impairment	Results of MoCA-K and mSTS-MCI tests for dementia	LR	Accuracy, sensitivity, and specificity
Spooner et al.	2020	Prediction of survival in dementia	Older adults	Demographic, medical and family history, psychological scores, quality of life ratings, etc.	FS, RF, BCR, and PCR	One-number accuracy

Table 1-2

Summary of the Studies Identified in the Literature Review on the Use of Machine Learning in Gerontology

Author	Year	Target variable(s) and problem(s) addressed	Target population	Variables	Machine learning algorithms	Performance metric(s) of machine learning model(s)
Tarekegn et al.	2020	Predicting frailty conditions (e.g., mortality and fracture)	Older adults with fracture	Clinical characteristics (number of hospitalizations, number of ED visits, disease history, disability history), and socioeconomic factors (age, citizenship, housing and work status, marital status, level of education, type of family)	ANN, GP, SVM, RF, LR, and DT	Accuracy, and k-fold cross validation
Wu et al.	2020	Major osteoporotic fracture risk	Male older adults diagnosed with osteoporosis	Demographic and clinical characteristics related to genetics	GBA, RF, ANN, and LR	One-number accuracy
Wu et al.	2020	Falls	Not specific age range provided	Data from wearable sensors such as acceleration, angular velocity, and attitude angle	RF, and DT	One-number accuracy
Ali et al.	2021	Early detection of Parkinson's disease	Older adults	Demographics, facial expressions (smiling face, disgusted face, surprising face)	SVM, KMC, and LR	Confusion matrix, accuracy, precision, recall, AUC, and F-1 score
Awais et al.	2021	Classification of most common activities of daily living: walking, sitting, standing, and lying	Community-dwelling older adults aged 65 or older	Body posture while using a wearable sensor	LSTM, SVM, and FS	F-score, and confusion matrix
Cuaya-Simbro et al.	2021	Fall Risk Prediction	Older adults with osteoporosis	Balance parameters with open or closed eyes	NB, SVM, AdaBoost, and RF	Accuracy, sensitivity, and specificity

Table 1-3

Summary of the Studies Identified in the Literature Review on the Use of Machine Learning in Gerontology

Author	Year	Target variable(s) and problem(s) addressed	Target population	Variables	Machine learning algorithms	Performance metric(s) of machine learning model(s)
Fricke et al.	2021	Automatic classification of EMG patterns in gait disorders	Patients aged 18 or older with different neurological diseases	Gait patterns	CNN, SVM, and KNN	One-number accuracy
Greene et al.	2021	Fall risk prediction and its association with balance	People aged 60 or older	Height, weight, and balance	LR	Accuracy, and sensitivity
Jang et al.	2021	Classification of stages of Alzheimer's disease	Individuals aged 50 years or older with Alzheimer's disease	68 variables extracted from four tasks related to language and eye movement	LR, RF, and GNB	Cross-validation, and AUC
Makino et al.	(2021)	Fall prediction	Community-dwelling older adults aged 65 years or older	Fall history, age, sex, fear of falling, prescribed medication, knee osteoarthritis, lower limb pain, gait speed, and timed up and go test	DT	Accuracy, sensitivity, PPV, and NPV
Martino et al.	2021	Malnutrition risk prediction	Frail older adults	Nutritional intake, dietary habits, and body composition	LASSO, SVM, KNN, RF, AdaBoost, and RUSB	Accuracy, precision, recall, and specificity
Rodríguez et al.	2021	Audio-based activity recognition system for reminders to take medications	Community-dwelling older adults aged 65 or older	Variables extracted from speech recognition	HMM	Three-fold cross validation, and confusion matrix
Saeed et al.	2021	Fall patterns	Community-dwelling older adults	Position of participants while falling	RF, DT, SVM, ANN, and NB	One-number accuracy
Speiser et al.	2021	Prediction of serious fall-related injuries	Older adults	Race, education, body mass index, marital status, health behaviors	RF, and DT	One-number accuracy, and AUC

Note. CHAID= Chi-square Automatic Interaction Detector. LR = Logistic Regression. RF = Random Forest. PCA = Principal Component Analysis. DNN = Deep Neural Network. SVM = Support Vector Machine. DT = Decision Tree. ANN = Artificial Neural Network. LR = Logistic Regression. FS = Feature Selection. BCR = Boosted Cox Regression.

PCR = Penalized Cox Regression. GP = Genetic Programming. GBA= Gradient Boosted Algorithm. KMC= K-Means Clustering. LSTM= Long Short-Term Memory. NB= Naïve Bayes. AdaBoost=Adaptive Boosting. CNN = Convolutional Neural Network. KNN = K Nearest Neighbour. GNB = Gaussian Naïve Bayes. PPV= Positive Predictive Value. NPV=Negative Predictive Value. LASSO = Logistic Regression Shrinkage and Selection Operator. RUSB = Random Under Sampling Boosting. HMM = Hidden Markov Models. AUC = Area Under the Curve. NIHSS = National Institutes of Health Stroke Scale (It is a standardized neurological examination tool used to assess the severity of stroke symptoms). MoCA-K = Montreal Cognitive Assessment - Korean version (a cognitive screening tool that is used to assess cognitive impairment and detect early signs of dementia. It is based on the original Montreal Cognitive Assessment but has been adapted for the Korean population). mSTS-MCI = mobile Screening Test System for Mild Cognitive Impairment.

Six studies related to falls in older adults investigated fall prediction (Cuaya-Simbro et al., 2021; Greene et al., 2021; Makino et al., 2021; Saeed et al., 2021; Y. Wu et al., 2020), and two studies addressed the application of machine learning in FRIs (Ateeq., 2018; Speiser et al., 2021). Ateeq (2018) used logistic regression and random forest to predict the FRIs in all age groups of community dwelling adults and children. They reported that random forest outperformed the logistic regression, but the accuracy of their models did not exceed 61%. Model evaluation was limited to accuracy and sensitivity and authors did not provide either a confusion matrix or the area under the curve. The other study on predicting FRIs in older adults was done by Speiser et al. (2021). The authors used random forest and decision tree on data of 1,635 community-dwelling older adults with different characteristics such as age, grip strength trial, race, body mass index, and education. They reported that decision tree could predict FRIs with an accuracy of 85 %, while the accuracy of random forest reached only 73%. Both training and testing accuracy of the decision tree was higher. The fact that the decision tree outperformed random forest in predicting injury is counterintuitive. Typically, it is anticipated that an ensemble learning algorithm will outperform a base learner (Hastie et al., 2009). If the opposite outcome arises, it is customary for authors to elucidate potential data characteristics or model tuning methods that contributed to the unexpected result. Regrettably, such explanations were absent in this study.

This literature review identified only two studies on the application of machine learning to FRIs. This limited amount of research on the topic and the abovementioned study limitations further encouraged conducting a new study on machine learning application to FRIs.

1.6 International Classification of Diseases

According to the Centers for Medicare and Medicaid Services (2023) International Classification of Diseases Tenth Revision, also known as ICD-10, codes are defined as a standardized system of alphanumeric codes used to report and categorize medical diagnoses and procedures. More than 68,000 codes make up the ICD-10 coding system, which corresponds to medical diagnoses or treatments (Centers for Medicare & Medicaid Services, 2023). These codes offer a standardized method of reporting medical data, enabling precise and consistent documentation between various healthcare practitioners and systems. ICD-10 codes are

consistently gathered in EDs and hospitals. They provide a standardized approach to reporting various diagnoses. Moreover, ICD-10 codes are globally utilized, ensuring comparability across diverse countries, cultures, and contexts (CIHI, 2022).

1.6.1 ICD-10 Codes and Injuries

A literature review was conducted at the end of 2022 to assess the current understanding of ICD-10 codes and their relationship with medical outcomes. The objective was to identify common areas of ICD-10 code utilization and whether these codes have been used in studies on injuries, or more importantly studies on FRIs. Examining utilization of ICD codes in prior literature, helped ascertain whether different diagnoses and their association with FRIs had been previously documented or not. Table 1-2 summarizes 16 most relevant research articles that utilized machine learning to investigate associations between the ICD-10 codes and clinical outcomes or injuries.

Table 1-2

Summary of the Studies Identified in the Literature Review on the Use of ICD-10 Codes and Injuries Using Machine Learning Algorithms

Author	Year	Target variable(s) and problem(s) addressed	Target population	Variables	Machine learning algorithms
Choi et al.	2018	Probability of suicide death	South Korean adults	Sex, age, type of insurance, household income, disability, and eight ICD-10 codes related to mental and behavioural disorders	Cox regression, SVM, DNN
Betts et al.	2019	Common maternal postpartum complications requiring an inpatient episode of care	Women giving birth	Maternal health data	GBT
Deschepper et al.	2019	Unplanned readmission to the hospital	Hospital patients	Age, length of stay, and pathology data	GBM, RF, PLR
McMaster et al.	2019	Adverse drug reactions	Hospital patients	Adverse drug reaction codes, primary diagnosis, primary diagnosis of adverse drug event probability grouping, and length of stay	RF. LR, SVM
Olsavszky et al.	2020	Forecasting the number of monthly hospital admissions for diagnoses of ten deadliest diseases*	All hospitalized patients in Romania from 2008 to 2018	ICD-10 codes	AutoTS
Su et al.	2020	Suicide risk	Children and adolescents	Demographic, clinical, and mental health data	LR
Weegar and Sundström	2020	Cervical cancer diagnosis	Swedish women with cervical cancer	Clinical, diagnostic, and treatment data	RF, NB, SVM
Bolourani et al.	2021	Lower extremity amputation	Trauma patients with arterial injury	Demographic, injury, laboratory, and imaging data	RF, XGBoost, LR
Cowling et al.	2021	Patient mortality	Hospital patients	Age, sex, socioeconomic status, and ICD-10 codes	LR, GBT
Edgcomb et al.	2021	Suicide attempt and self-harm	Women with mental illness hospitalized for general medical conditions	Demographics, clinical, and mental health data	RF

Table 1-2

Summary of the Studies Identified in the Literature Review on the Use of ICD-10 Codes and Injuries Using Machine Learning Algorithms

Author	Year	Target variable(s) and problem(s) addressed	Target population	Variables	Machine learning algorithms
Huda et al.	2021	Risk of developing wild-type transthyretin amyloid cardiomyopathy	Patients at risk for wild-type transthyretin amyloid cardiomyopathy	Demographics, clinical, echocardiographic, electrocardiographic, and laboratory data	LR, XGBoost
McCann-Pineo et al.	2021	Predictors of opioid administration and prescribing	ED patients receiving opioids age 18 or older	Sociodemographic variables, and ED clinical variables (chief complaint, discharge diagnosis)	RF, GBM, NB
Shah et al.	2021	Major complications and readmission	Patients undergoing lumbar spinal fusion	Demographic, clinical, and surgical data	XGBoost, AdaBoost, GBM, RF
Tran et al.	2021	In-hospital mortality following a trauma	Trauma patients	Demographic, injury codes, and physiological variables	XGBoost
Lee et al.	2022	In-hospital mortality	Physical trauma patients in Korea, no age limitation	ICD-10 codes, patient age, gender, intentionality, injury mechanism and emergent symptom, AVPU scale, KTAS, and procedure codes	DL, AdaBoost, XGBoost, LightGBM
Tran et al.	2022	In-hospital mortality following a traumatic injury	Trauma patients	Demographic, injury-related codes, and physiological variables	XGBoost

Note. SVM = Support Vector Machine. DNN = Deep Neural Network. GBT = Gradient Boosting Trees. GBM = Gradient Boosting Model. RF = Random Forest. PLR = Penalized Logistic Regression. AutoTS = Automated Time Series. NB = Naïve Bayes. XGBoost = Extreme Gradient Boosting. LR = Logistic Regression. AdaBoost = Adaptive Boosting. DL = Deep Leaning. LightGBM = Light Gradient Boosting Model. COPD = Chronic Obstructive Pulmonary Disease. AVPU = Alert/Verbal/Painful/Unresponsive (simple method used in medical settings to assess a person's level of consciousness and responsiveness). KTAS = Korean Triage and Acuity Scale (a system used in South Korea to prioritize patients in emergency departments based on the severity of their condition). *Ten deadliest diseases included: ischemic heart diseases, stroke, COPD, lower respiratory infections, Alzheimer's disease, lung cancer, diabetes mellitus, road injuries, diarrheal diseases, and tuberculosis.

None of the studies focused on predicting FRIs using ICD-10 codes in older adults. Only two studies, by the same research team (Tran et al., 2021, 2022), investigated post-injury occurrences, such as in-hospital mortality following trauma or traumatic injuries that were a result of a fall, gunshot wound, stabbing, blunt injury, motor vehicle collision, or motorcycle collision. Articles by Huda et al. (2021), Bolourani et al. (2021), Shah et al. (2021), and Weegar and Sundström (2020) predicted clinical outcomes (i.e., the risk of developing wild-type transthyretin amyloid cardiomyopathy, lower extremity amputation, major complications, and readmission after lumbar spinal fusion), while studies by Edgcomb et al. (2021), and Su et al. (2020), predicted suicide attempts, self-harm, and suicide risk. The age distribution of the population varied significantly between studies. Huda et al. (2021), Shah et al. (2021), and Betts et al. (2019) did not specify age ranges, while McCann-Pineo et al. (2021) and Choi et al. (2018) targeted patients over the age of 18 years.

Most previous studies that used ICD-10 codes as variables combined them with other factors such as socioeconomic (Cowling et al., 2021) or demographic data (Bolourani et al., 2021; Choi et al., 2018; Shah et al., 2021). Only one study solely focused on diagnostic categories, but it only examined ischemic heart diseases, stroke, chronic obstructive pulmonary disease, lower respiratory infections, Alzheimer's disease, lung cancer, diabetes mellitus, road injuries, diarrheal diseases, and tuberculosis (Olsavszky et al., 2020). Furthermore, these studies primarily focused on predicting limited outcomes such as suicide (Choi et al., 2018), cancer (Weegar & Sundström, 2020), mortality (Cowling et al., 2021; Kim et al., 2018; Tran et al., 2021), and drug reactions (McMaster et al., 2019) rather than FRIs. Therefore, examination of all diagnostic categories related to ED or hospital admission in association with FRIs is an unexplored field of research.

Based on the literature review, the machine learning models that emerged most frequently are random forest, logistic regression, and XGBoost. These models were used in multiple studies in different research contexts. Random forest was used in three studies (Deschepper et al., 2019; Edgcomb, Shaddox, et al., 2021; McMaster et al., 2019), logistic regression was used in two studies (Cowling et al., 2021; Su et al., 2020), and XGBoost was used in two studies (Bolourani et al., 2021; Huda et al., 2021). These machine learning models have

demonstrated effectiveness and popularity in various research domains, leading to their frequent utilization in the forementioned studies.

In summary, the studies included in this literature review demonstrate the use of machine learning algorithms to analyze medical data across diverse contexts is increasing rapidly in recent years. Predictions mainly focus on the risk of developing different medical conditions and forecasting the number of hospital admissions for specific diagnoses. These studies emphasize assessing performance metrics to achieve their targeted outcomes. However, our literature search did not yield specific studies on the application of machine learning algorithms for analyzing the association between FRIs and other diseases using ICD-10 codes in older adults.

1.7 Research Gap

The literature review helped identify several gaps and the need for further research. First, there is insufficient research examining the association between diagnostic codes and FRIs. Second, many studies have been conducted on falls prediction, but very few have focused on FRIs prediction using machine learning algorithms. These studies mostly used logistic regression which is a linear classifier. Some compared logistic regression with random forest. This comparison is not the most useful knowing that logistic regression is not the base learner for random forest of analysis and not more robust algorithms. They often lacked clear explanations of the data cleaning and preparation process, resulting in models that can be improved. Additionally, most studies used accuracy as the only performance metric while other performance metrics such as sensitivity, precision and F1 score are available and could offer a better understanding of the model performance. Finally, to the author's knowledge, at the present time there are no studies that have used ICD-10-CA diagnostic codes for FRIs in older adults using robust machine learning algorithms. Therefore, additional research is warranted.

Based on the research gap, this study intends to demonstrate how three different machine learning algorithms can be used to study the association between FRIs and ICD-10-CA using ED and hospital data. This study will cast light on strengths and limitations of machine learning algorithms in the study of FRIs. This approach improves interpretability and reproducibility of the study in other jurisdictions or settings. Unlike ICD-10 codes, variables like race, income and education level can have different meanings in different contexts or over time, therefore, training

a model based on such variables makes it difficult to replicate the result by other researchers. To mind the gap in previous studies, where data cleaning was not clearly described, this study will clearly document steps required for data preparation, the most time and resources demanding step in a machine learning-based project (Hastie et al., 2009; Provost & Fawcett, 2013). Having detailed documentation will enhance transparency and reproducibility of the study.

1.8 Research Questions

Based on the literature review and identified gaps, three research questions were identified for this project:

1) Which categories of ICD-10-CA diagnostic codes are most informative when associated with FRIs?

2) What is the difference between ICD-10-CA diagnostic code categories associated with FRIs reported in EDs (NACRS database) and hospitals (DAD database)?

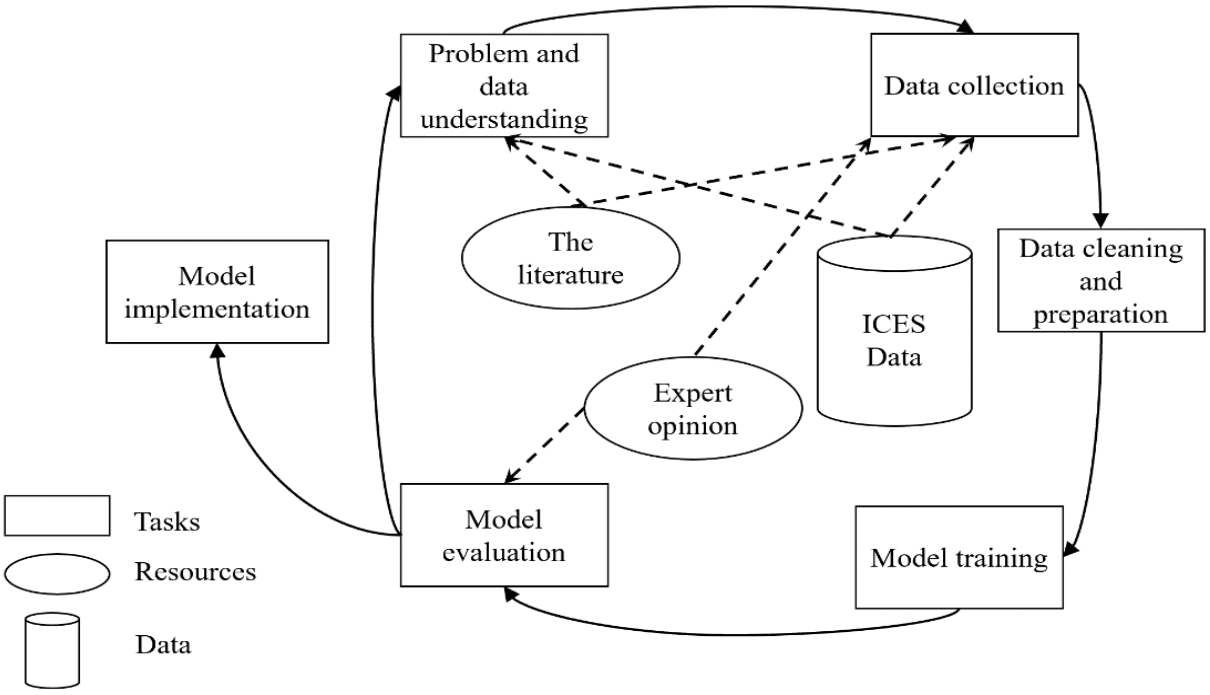
3) Which machine learning model is the most accurate and sensitive for determining associations between ICD-10-CA diagnostic codes and FRIs?

Chapter 2

2. Methods

In this chapter, the main methodological steps of the study are described. These steps follow the methodology proposed schematically in Figure 2-1. This methodology is inspired by the Cross-Industry Standard Process for Data Mining framework, which is a well-known approach in data mining (Provost & Fawcett, 2013).

Figure 2-1
Study Framework for Providing Main Methodological Steps of the Study



Secondary data was extracted from two databases, the National Ambulatory Care Reporting System (NACRS) and the Discharge Abstract Database (DAD). NACRS encompasses closed cases from community and hospital-based ambulatory care, such as day surgery, outpatient and community clinics, and ED data. Only ED data was included in this study. DAD contains hospital inpatient records that include administrative, demographic, and clinical patient information at the time of patient discharge, death, or transfer (Ontario Ministry of Health and Long-Term Care, 2012). Together, these databases provided the data for an overview of two levels of healthcare that an individual

may require in the event of an FRI.

2.1 Dataset Creation and Extraction

The origin and components of the databases used to build a dataset for this study are described in this section. The databases NACRS and DAD are managed by the Canadian Institute for Health Information (CIHI), which obtains healthcare information from regional health ministries or authorities, as well as community and hospital facilities (Chan et al., 2013; CIHI, 2011). After collecting the data, CIHI cleans and validates the data and subsequently provides IC/ES with access to the data.

The dataset creation plan (Appendix A) was submitted to IC/ES in 2019, and a data analyst extracted the data from NACRS and DAD databases. The period covered in the present study was nine years, from 2006 to 2015. Although this observation window may seem outdated, the study provides a valuable baseline data for the baby boomer generation's entry into the retirement age (Hogan et al., 2008). In addition, data from the Ontario Registered Persons Database (RPDB), which includes sociodemographic data on residents of Ontario, was added.

The research team received a "Master" dataset and a "Full" dataset. The "Master" dataset included sociodemographic data from RPDB, with data on age, sex, the nearest neighborhood income quintile, and five chronic condition codes (i.e., diabetes, hypertension, dementia, asthma, and chronic obstructive pulmonary disease). The "Full" dataset encompassed ICD-10-CA codes and contained combined codes for at least one injury (S00-S99 or T00-T14) and one fall (W00-W19) code per observation. Every Ontario resident eligible to receive healthcare is assigned a unique encoded identifier called an IC/ES key number (IKN), which enables linking between all IC/ES databases (Iron & Sykora, 2015). It is important to mention that since data was extracted for this study, the 11th revision of the International Classification of Diseases (ICD-11) codes came into effect in January 2022 as the result of a comprehensive and collaborative process led by the WHO (Harrison et al., 2021). Ethics approval for this study was obtained from the Research Ethics Board at Western University (HSREB #211336, Appendix B). This study is a continuation of a research program exploring FRIs on the population level using the same ICD-

10-CA codes and comparable timeframe done by Ming (2020) and Lappan (2021).

2.2 Population Inclusion and Exclusion Criteria

All older adult residents of Ontario who were admitted to an Ontario ED or hospital for an FRI comprised the population of interest. The inclusion criteria for this population were: (1) older adults diagnosed with FRI or who passed away as a result of FRI between January 1, 2006, and December 31, 2015; (2) older adults admitted to an ED or hospital who were 65 years of age or older at the time of admission; (3) a diagnosis of an injury, as described by ICD-10-CA codes S00-S99 or T00-T14 codes, combined with a diagnosis of a fall, as defined by ICD-10-CA codes W00-W19, to assure an FRI. Older adult residents of Ontario who were admitted to ED or hospital between January 1, 2006, and December 31, 2015, and were matched to the case group by sex and age but did not experience FRIs were included as instances without target variableto help train the machine learning models, so the models could discriminate between patients who had FRIs and those who did not. This group was matched to the FRI group by sex, age, Charlston Comorbidity Index score and LHIN, with a ratio of 1.5 to 1. The date of visiting ED due to FRI was defined as the index event date. Participants were deemed ineligible for the study if they met any of the following criteria: (1) lacked a valid IC/ES Key Number (IKN); (2) a person passed away prior to ED or hospital admission; (3) a person did not reside in Ontario; (4) a person was younger than 65 at the index date; or (5) a person had an FRI while in a hospital.

2.3 Data Cleaning and Preparation

The data cleaning and preparation process was completed using SAS 9.4 M7 (2020) and R version 3.6.3 (R Core Team, 2020). DAD and NACRS datasets underwent the same steps described below. The SAS and R codes used for the preparation of these databases can be found in Appendix C and Appendix D, respectively.

In the first step, unwanted variables were removed. For example, the dataset from NACRS originally had 81 columns, and the dataset from DAD had 87 columns. Not all variables were relevant. The relevant variables were chosen after considering the literature review findings, the study goals, and the gaps that needed to be addressed. Ten diagnostic codes (*dx10code1* to *dx10code10*), *sex*, *age* at the time of admission, and study IDs (*study_ID*) were

chosen as variables from NACRS, while 25 diagnostic codes (*dx10code1* to *dx10code25*), *sex, age* at the time of admission, and study IDs of observations (*study_ID*) were chosen from DAD (Table 2-1). The *drop* function in SAS was used to remove unnecessary variables. Ultimately, 13 columns from NACRS dataset and 28 columns from DAD dataset remained.

Table 2-1
List of Variables Selected from Databases

Databases	Variables Extracted
NACRS	Sex, age at the time of admission, main diagnosis, other diagnosis, index date, unique study ID
DAD	Sex, age at the time of admission, diagnosis, unique study ID

The original datasets included both the observations of individuals who experienced FRIs and controls who did not have FRIs. The next step in data preparation was to identify FRI records, which was done as follows. The data cleaning process began by merging the *Full* and *Master* datasets using the variable *Study_ID* as the identifier for individual patient records. This merging was performed using the *merge ()* function in R version 3.6.3 (R Core Team, 2020). The resulting merged dataset was referred to as the *Master-Full* dataset. To refine the dataset, observations that did not include FRI diagnoses were filtered out. This was accomplished by using the *filter ()* function to remove the control group data, retaining only FRI cases. Additionally, the variable *Days from index date to registration date* had to have a value of zero, indicating that a patient was registered at the ED on the same day as the *index date* (the date of injury). This step confirmed that the treatment was provided specifically for FRIs. Next, the NACRS and DAD datasets were separately merged with the *Master-Full* dataset to collate FRI records with predictive variables for both ED and hospital patients. Finally, to ensure that there was only one observation per person or IKN, any duplicate observations with the same *Study_ID* were removed. The *distinct ()* function was used to retain only the first observation per *Study_ID* in each dataset. The new FRI tables were assigned a new column called *FRI* with *Y* (Yes) values and merged with the original NACRS and DAD tables separately using SAS *merge* function. Next, the empty cells were assigned *N* (No) meaning that no FRI occurred. The full list of

commands used for finding FRIs can be found in Appendix C and Appendix D. Additionally, the full list of ICD-10-CA diagnostic categories for FRIs can be found in Appendix E.

After the number of FRIs was obtained, an initial descriptive analysis was performed on sex and age using Structured Query Language (SQL) codes in SAS. The objective of this analysis was to gain a better understanding of relative frequency of different values of these variables in the developed dataset. The specific SQL queries used for this analysis can be found in Appendix C.

Further work was conducted to categorize the diagnostic codes into 21 categories, which were used as input variables. The reason for this step was that ICD-10-CA codes were stored as alphanumeric codes in the raw datasets rather than categories, but these diagnoses are presented in 21 main chapters or categories in ICD-10-CA manual (Centers for Medicare & Medicaid Services, 2023) . In this study, the term category is used instead of chapter to emphasize that each category represents a group of interrelated diagnoses. The full list of 21 categories used in the current study can be found in Appendix F. In the raw datasets, the diagnostic codes were stored in different columns without specifying their respective category (Figure 2-2). To be able to utilize the diagnostic codes as input variables, they had to be placed in their respective category. For example, codes *A00* to *B99* were placed in *Category 1* that describes certain infectious and parasitic diseases, and codes *G00* to *G99* were assigned to *Category 6* that relates to diseases of the nervous system. The codes for injury and falls that defined FRI were part of categories 19 and 20 and were excluded to avoid having variables related to FRI among the inputs. This was done as follows. In *Category 19* (injury, poisoning and certain other consequences of external causes from *S00* to *T98*) codes *T00* to *T14* were removed. Similarly, codes *W00* to *W19* and *S00* to *S99* were removed from *Category 20* (external causes of morbidity and mortality from *V01* to *Y98*). Each category branches into numerous subcategories. However, due to the scope of this study, it was not feasible to identify subcategories for all 21 categories. Instead, the focus was placed only on select subcategories that emerged through the analysis as the most informative when associated with FRIs in EDs and in hospitals.

Figure 2-2
Raw DAD Dataset

Diagnosis code	Diagnosis code	Diagnosis code	Diagnosis code	Diagnosis code
I428	I500	I480	I472	Z136
K255	C3410	M069	I978	I480
R074				
A492				
C3411	80703			
D700	Y433	E876	C3499	I100
G459				
K318	K921	K317	N179	
I517	I420	E780		
K921	I2510	K922		
E860	R53	C911	Z933	Z751
I200	R074	I100	E11900	E785
I480				
I480	I200			
R074	R55	G309	F009	
C679				
C711	94403			
J181	R060	M4024	Z751	
J440	J209	I219	Z548	
M173	G4738	E11229	N180	E877
T822	Y832			
I2141				

To illustrate how each category was derived from the original datasets, *Category 1* will be explained here as an example. To identify observations with certain infectious and parasitic diseases (*Category 1*), SQL queries were used to extract all records for codes *A00* to *A999* and *B00* to *B99* separately. These two sets of records were then merged using the *merge* function and duplicate records were removed with the *nodup* function. The resulting records represented observations with certain infectious and parasitic diseases and were assigned a new column or variable named *Category1* that was populated with *Y* using the *set* function. Next, the *Category1* records were merged with the original table and any empty cells were assigned *N* to indicate that these observations did not have the diseases in this category. Using the *keep* function, only the columns for *Category1* and *study_ID*s were retained to create the *Subfinal_Category1* table. The same process was followed to create *Subfinal* tables for the remaining 20 categories. The final NACRS observations for FRIs and final tables of all 21 categories were merged using the *merge* function. The resulting final tables of NACRS contained all binary values of *Y* (yes) or *N* (no) and was used to train machine learning models.

Before determining which categories required further investigation, it was necessary to first identify the two most informative ones associated with FRIs in ED and hospital. These two

categories of ICD-10-CA codes were then further divided into subcategories. To explain this process, an example is provided here. For instance, *Category 2*, neoplasms, included diagnostic codes from *C00* to *D48* and encompassed 20 subcategories associated with neoplasms. One such subcategory was benign neoplasms, denoted by codes *D10* to *D36*. The process of categorizing subcategories within the first two informative categories followed the same *Y* or *N* procedure described above for the diagnostic categories. For a comprehensive list of subcategories within the first two informative categories see Appendix G.

2.4 Data Analysis

After the data was prepared, it was split into training and testing data. The models were trained using the training data. An in-depth discussion about theories behind all algorithms is beyond the scope of this study. Given the popularity of decision trees and tree-based ensemble learning algorithms, their theory overview is briefly discussed here as an example.

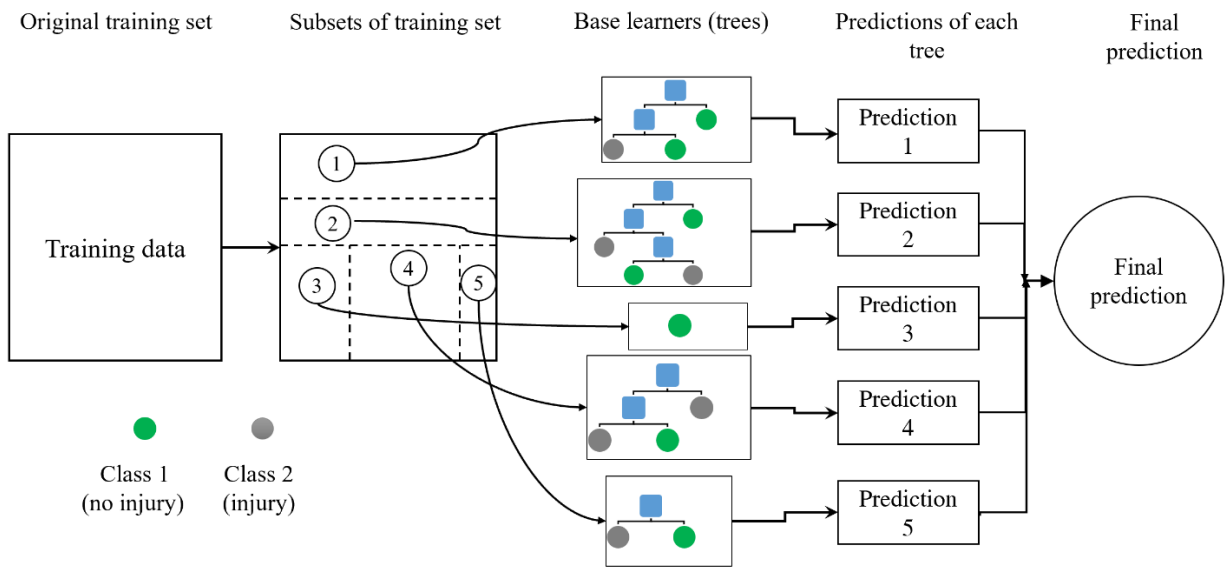
A decision tree algorithm segments the data into subsets to minimize disorder in each subset. The process of segmentation continues until the algorithms reach 'pure' leaves. In the context of a decision tree algorithm, pure leaves refer to the final subsets or nodes in the tree where the data is homogeneous or consists entirely of one class or category. These leaves represent the ultimate outcome or prediction of the decision tree for a given set of input variables (Hastie et al., 2009). A common way to define disorder is by using entropy formula as presented in Equation (1).

$$Entropy = IE\ (p1, p2, \ldots, pc) = -\sum_{i=1}^{c} p_i log(pi), \qquad (1)$$

where, p_i is the probability of belonging to class i (thus the sum of p_is is equal to 1) and c is the number of classes.

Ensemble learning algorithms (i.e., random forest and XGBoost trees) are trained based on an ensemble of trees rather than a single predictor. These algorithms split the data into several subsets and train a tree (or a single base learner) based on each subset as shown in Figure 2-3. The final prediction is an aggregation (e.g., via majority vote) of the individual predictions.

Figure 2-3
Illustration of the Theory Behind Ensemble Learning Algorithms



Note. The green and brown circles represent the leaves, and the blue squares represent the nodes (where the three divides).

Although ensemble learning algorithms are more accurate, they do not result in an explicit model like a decision tree. However, they could be interpreted using their importance factor output. These algorithms calculate variable importance based on the same formula of a decision tree (e.g., entropy). The key difference is that ensemble learning algorithms average out the importance (Piryonesi & El-Diraby, 2020a). Thus, the equation for determining the importance of attribute x_l can be formulated as follows (Hastie et al., 2009).

$$I_l^2 = \frac{1}{M}\sum_{m=1}^{M} I^2(T_m) \qquad (2)$$

where I_l^2 is the importance of variable x_l, $I^2(T_m)$ is the importance of that variable in decision tree m, and M is the number of base learners (i.e., decision trees).

2.5 Model Training

Data analysis was performed using R version 3.6.3 (R Core Team, 2020). Based on the literature review, consideration of the selected variables' nature, the availability of software and resources, and the advantages of different models in this context, three machine learning models

were chosen: the decision tree (Badgujar & Pillai, 2020; Speiser et al., 2021; Y. Wu et al., 2020) random forest (Ateeq, 2018; Cuaya-Simbro et al., 2021; Speiser et al., 2021; Y. Wu et al., 2020), and gradient boosting models (Cuaya-Simbro et al., 2021; Di Martino et al., 2021; Q. Wu et al., 2020). They were previously employed by researchers for predicting and investigating the association between selected variables and falls or injurious falls. However, earlier studies have limitations (e.g., lack of sufficient explanation of results and low accuracy) that will be addressed in the current study.

Decision tree has been one of the most popular machine learning algorithms (Wu et al., 2008) because of the "open box" approach it utilizes and its ease of interpretation and visualization (Hastie et al., 2009). This contrasts with the "black-box" nature of some machine learning algorithms, such as neural networks, that makes their interpretation difficult (Piryonesi & El-Diraby, 2020a; Provost & Fawcett, 2013). Unlike models such as K- nearest neighbor, decision tree results in an explicit model that can be implemented and used to classify new data independent of the training set, and its training requires no initial assumption such as normality of residual errors or independence of input variables (e.g., as required by naïve Bayes classifier) (Hastie et al., 2009; Provost & Fawcett, 2013). In addition to these features, decision trees are known to have a superior accuracy compared to linear models such as logistic regression or naïve Bayes classifier (Hastie et al., 2009).

The two ensemble learning algorithms, i.e., random forest and XGBoost trees, were adopted to check if they can enhance the accuracy of the decision tree's classification. The ensemble form of a decision tree would be a random forest or an XGBoost tree (Hastie et al., 2009). The selection of multiple models was motivated by the desire to compare their performance and attain greater confidence in the obtained results, addressing the third research question of the study: "Which machine learning model is the most accurate and sensitive for determining associations between ICD-10-CA diagnostic codes and FRIs?". Furthermore, having these three models can help make an apple-to-apple comparison in addressing some of the limitations discussed in Sections 1.5 and 1.7. Finally, the limited availability of machine learning software within the IC/ES environment also led to selection of these three machine learning algorithms. First, the absence of Python software led to restricted options for choosing more models and packages. Second, all analysis had to be performed on secure IC/ES servers and only

the results could be exported. Consequently, the three models chosen for this study are decision tree, random forest, and XGBoost. The codes used for each model are available in Appendix D.

In all three machine learning models used in this study, the first step was that the prepared data was initially split into two parts with a ratio of 70% for the training set and 30% for the testing set. This ratio of is a common best practice in the literature for machine learning across different disciplines (Awais et al., 2021; Hastie et al., 2009; Speiser et al., 2021). The training sets were used to develop the models, by teaching the algorithms how to classify individuals as either having FRI or not (*Y* or *N*). Meanwhile, the test sets that contain the data that was not used in the model's training, were used to assess the performance of each model. This helped to estimate the performance metrics of each model and identify the most informative variables. Machine learning libraries used in this research do the splitting randomly and therefore the training set and test set have similar distributions.

The first algorithm utilized was the decision tree. Decision trees have few weaknesses that could be addressed by their ensembles. Examples of such weaknesses are a lack of robustness and relatively low accuracy (Hastie et al., 2009; X. Wu et al., 2008). Random forests and XGBoost trees are examples of ensembles of trees that do considerably better than a single learner. Ensemble learning algorithms usually have a higher robustness given the fact the result is averaged over an ensemble of trees (Hastie et al., 2009; Wu et al., 2008). Robustness is the idea of having a more stable and reliable model (Piryonesi & El-Diraby, 2020b). In ensemble learning, the data is randomly split into multiple subsets that could be both horizontal and vertical. The number of subsets is equal to the number of base learners or weak learners. Next, a tree (base learner) is trained based on each subset. For every new incoming example, each tree will make its own prediction (in this case a Yes/*Y* or a No/*N*). The final prediction of the ensemble model is an aggregate of the predictions of base learners. While in a random forest the final prediction is the result of a majority vote, the voting is weighted in an XGBoost algorithm. It means that a larger weight is assigned to more accurate trees to amplify their prediction and diminish the less accurate base learners (Piryonesi et al., 2021). Random forest was used as the second algorithm in the study and an XGBoost tree was the final algorithm employed. How many trees are used to construct the model depends on the number of iterations. (Hastie et al., 2009). Improved performance may result from more iterations, but it can also result in overfitting

(Hastie et al., 2009). Several methods can be used to determine the ideal number of iterations for an XGBoost model. Grid search and random search are most common. Grid search is a technique for determining a model's ideal hyperparameters by examining how the model performs with various combinations of hyperparameters. The validation set is used to evaluate the model's performance after training it over several iterations. Choosing the number of iterations that produces the best performance allows one to determine the optimal number. (Hastie et al., 2009).Grid search and random search both use methods that search over sets of hyperparameters, but random search does so over a set of hyperparameters rather than a predefined grid (Hastie et al., 2009). Using a random set of iterations, the model was trained, and its effectiveness was evaluated using a validation set. At the end, 100 iterations were used as the ideal number because they produced the best results.

2.6 Model Evaluation

All three machine learning models were evaluated for accuracy, sensitivity, specificity, precision, and F1 score to assess each model's performance and to compare the models against each other. Variable informativeness was assessed in all three models. The model's predictions are summarized in the confusion matrix, which is useful for calculating performance metrics like sensitivity and accuracy (Boehmke & Greenwell, 2019; Kuhn & Johnson, 2013; Nolan & Lang, 2015).

A table that lists the conclusions drawn by a binary classification model is known as a confusion matrix. It displays how many predictions the model made were true positive (TP), true negative (TN), false positive (FP), and false negative (FN) (Boehmke & Greenwell, 2019; Kuhn & Johnson, 2013; Nolan & Lang, 2015). In the context of FRIs, a true positive occurs when the model accurately predicts an injury, and a true negative occurs when the model accurately predicts no FRIs. False positive prediction happens when the model predicts an injury when there is not one, and a false negative prediction happens when the model predicts no injury when an injury really did happen. Accuracy, sensitivity, specificity, precision, and F1 score are commonly used performance metrics that can be calculated using a confusion matrix (Kuhn & Johnson, 2013).

Accuracy is defined as the ratio of the number of correct predictions to the total number

of predictions. It measures how accurately the model foresees both favorable and unfavorable events (Boehmke & Greenwell, 2019; Kuhn & Johnson, 2013; Nolan & Lang, 2015). Accuracy can be calculated using the following formula:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

Sensitivity measures the proportion of true positives that are correctly identified by the model. It is defined as the ratio of the number of true positives to the total number of actual positive instances (Kuhn & Johnson, 2013). Sensitivity can be calculated using the following formula:

$$Sensitivity = TP / (TP + FN) \quad (4)$$

Specificity, also known as the true negative rate, is a metric used to evaluate the performance of a model in correctly identifying negative instances (Boehmke & Greenwell, 2019; Kuhn & Johnson, 2013; Nolan & Lang, 2015). Specificity can be calculated using the following formula:

$$Specificity = TN / (TN + FP) \quad (5)$$

Precision, also referred to as positive predictive value, is a metric that assesses the accuracy of a model in correctly identifying positive instances (Boehmke & Greenwell, 2019; Kuhn & Johnson, 2013; Nolan & Lang, 2015). Precision focuses on the proportion of correctly predicted positive instances out of all instances classified as positive. It provides insight into the model's ability to avoid false positives and is particularly useful in scenarios where the cost of false alarms is high, such as in medical diagnosis or fraud detection (Boehmke & Greenwell, 2019; Kuhn & Johnson, 2013; Nolan & Lang, 2015).Precision can be computed using the formula:

$$Precision = TP / (TP + FP) \quad (6)$$

The F1 score is a measure that combines precision and sensitivity into a single metric, providing a balanced evaluation of a model's performance. (Boehmke & Greenwell, 2019; Kuhn & Johnson, 2013; Nolan & Lang, 2015). The F1 score considers both false positives and false

negatives, making it a useful metric when there is an uneven class distribution or when both types of errors are equally important. It provides a comprehensive assessment of a model's effectiveness in identifying positive instances while minimizing misclassifications. In the context of FRIs, F1 score can be calculated as follows:

$$F1\ Score = 2 * (Precision * Sensitivity) / (Precision + Sensitivity) \quad (7)$$

Variable importance is a useful output of machine learning algorithms. It is a metric that measures the impact of input variables on the output of a machine learning model. Thus, it is a useful tool to understand the behavior of the model and identify the most important or informative variables. There are several methods for calculating variable importance, and they differ depending on the type of model being used (Hastie et al., 2009; Provost & Fawcett, 2013).

Decision trees create segmentations in the dataset. They start with the most informative variable and split the data based on a test. Therefore, at least two branches grow out of each node. Then, the nodes in each branch will split based on their informativeness. The terminal node of a tree is called a leaf, which is ideally pure and belongs to a particular class (Hastie et al., 2009; Opher & Ostfeld, 2011). Different trees may use different measures for defining informativeness. Most trees rely on entropy or the Gini index (Hastie et al., 2009; Provost & Fawcett, 2013; Wu et al., 2008). The information gain and the homogeneity of each leaf will be determined by the parameters of the tree (Ergen et al., 2015; Wu et al., 2008). The Gini index measures the degree of heterogeneity of a set of samples based on the target variable, and information gain measures the reduction in entropy achieved by splitting the data on a given variable. The variables that lead to the greatest reduction in impurity or entropy are considered the most important (Hastie et al., 2009; Provost & Fawcett, 2013; Wu et al., 2008). The trees developed in this study are based on entropy the most well-known implementation of decision trees (Hastie et al., 2009; Provost & Fawcett, 2013).

In random forests, variable importance is usually calculated by aggregating the importance scores of individual decision trees in the forest. The importance of a variable is calculated by measuring the reduction in classification accuracy that occurs when that variable is randomly permuted. The variables that cause the greatest decrease in accuracy are considered the most important (Hastie et al., 2009; Provost & Fawcett, 2013; Wu et al., 2008).

In XGBoost, variable importance can be calculated using the gain or cover metrics. Gain measures the reduction in loss achieved by splitting the data on a given variable and covers the number of observations that are associated with a given variable. The variables that lead to the greatest reduction in loss or have the most observations associated with them are considered the most important (Kuhn & Johnson, 2013; Lantz, 2019).

In R, the *varImp* function from the *caret* library was used to calculate variable importance for decision trees and random forests. For XGBoost, the *xgb.importance* function from the XGBoost package was used. These functions return a data frame containing the importance scores for each variable, which was visualized using a bar plot or other visualization tool (Boehmke & Greenwell, 2019; Lantz, 2019; Zaki & Meira Jr, 2020).

Chapter 3

3. Results

The Results chapter is divided into three sections. The first presents findings on the most informative diagnostic categories associated with FRIs at the emergency department level of care. The second section presents findings on the most informative diagnostic categories associated with FRIs for individuals who were admitted into the hospital level of care. Both sections are divided into four parts, starting with results from different machine learning models that answer the first research question: "Which categories of ICD-10-CA diagnostic codes are most informative when associated with FRIs?", followed by findings from the analysis of subcategories of diagnostic codes to answer the second research question: "What is the difference between ICD-10-CA diagnostic code categories associated with FRIs reported in EDs (NACRS database) and hospitals (DAD database)?", and ending with a summary of findings on accuracy, sensitivity, specificity, precision, and F1 score of the three machine learning models to answer the third research question: "Which machine learning model is the most accurate and sensitive for determining associations between ICD-10-CA diagnostic codes and FRIs?"

3.1 Emergency Level of Care

3.1.1 Demographics

The current study analyzed a total of 1,248,029 observations in the NACRS dataset (EDs) out of which 631,339 observations were identified as FRIs. Table 3-1 provides age and Table 3-2 provides sex information for the identified FRI observations in the NACRS dataset. The age group of 75 to 79 years had the highest incidence of FRIs, while the age group over 90 experienced the lowest number of FRIs. The overall mean age of patients was 77.8 years. Females experienced nearly twice as many FRIs as males.

Table 3-1
Age Distribution of FRI Observations in EDs of Ontario (NACRS Dataset)

Age group	Number of observations	Percentage of total frequency (%)
65-69	116,645	18.5
70-74	97,320	15.4
75-79	165,688	26.3
80-84	108,645	17.2
85-89	87,385	13.8
Over 90	55,656	8.8
Total	631,339	100.0

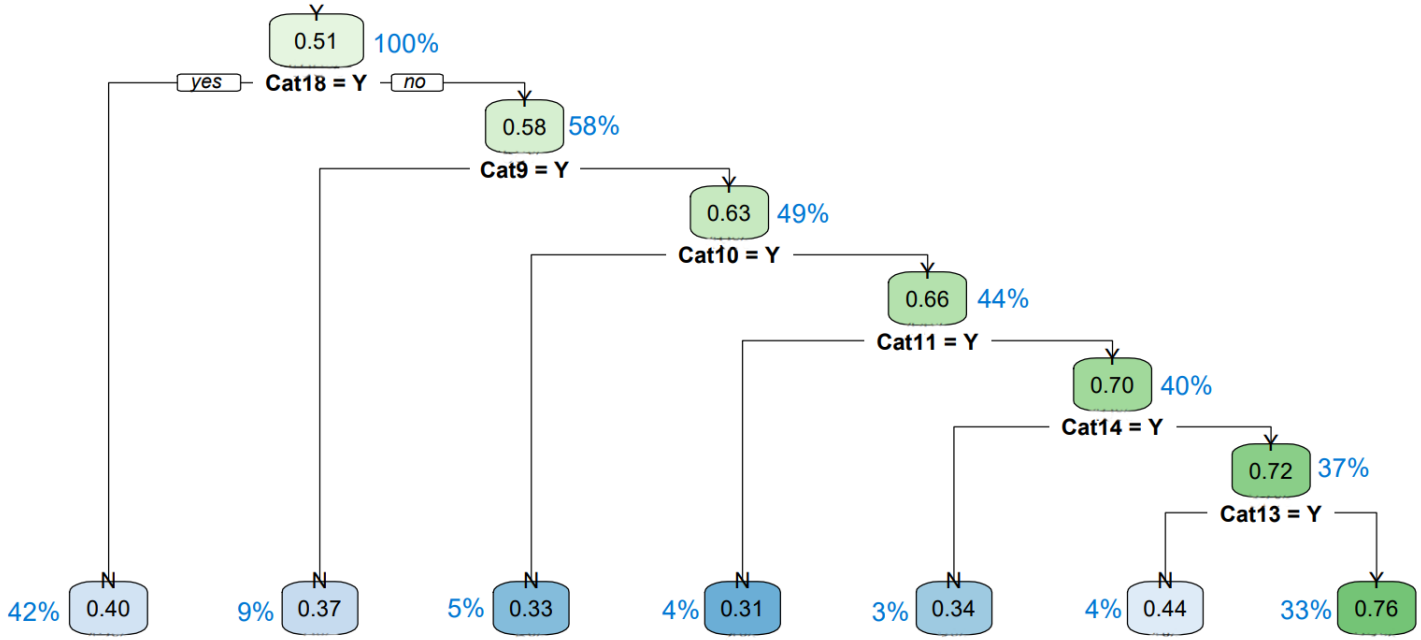
Table 3-2
Sex Distribution of FRI Observations in EDs of Ontario (NACRS Dataset)

Sex	Number of observations	Percentage of total frequency (%)
Females	420,077	66.5
Males	211,262	33.5
Total	631,339	100

3.1.2 Decision Tree

The target variable of the decision tree model was the occurrence or non-occurrence of FRIs over a 9-year period. Figure 3-1 shows the final decision tree model for input variables.

Figure 3-1
Visualization of the Decision Tree for FRIs Using Diagnostic Categories of ICD-10 Codes (Cat1 to Cat21) in NACRS Dataset



Note. The tree consists of various components: starting from the root node, each internal node represents a decision based on specific input variables and their conditions (Y=Yes or N=No). The branches emanating from each internal node depict the possible outcomes based on the input variables' conditions. The leaf nodes, or terminal nodes, signify the final predictions, representing specific combinations of input variables. Within each leaf node, the percentages indicate the proportion of observations falling into that category (Y=Yes or N=No). By examining the tree's structure, splitting criteria, branches, leaf nodes, and the percentages in each leaf, we can gain a comprehensive understanding of the relationship between the input variables and the predicted outcomes of FRIs based on the ICD-10-CA diagnostic categories. The lighter and darker shades of blue and green convey information about the certainty or probability associated with the predicted outcomes. Darker shades of blue or green indicate a higher probability or a stronger association with a particular outcome, representing a higher confidence in the prediction being made. On the other hand, lighter shades of blue or green suggest lower probabilities or weaker associations with the predicted outcomes, indicating a lower confidence in the corresponding predictions. This color scheme helps to distinguish between more reliable or significant associations depicted by darker colors and less robust relationships represented by lighter colors, allowing for a visual assessment of the model's confidence in specific conditions or categories within the decision tree; Cat18=symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified;Cat9= diseases of the circulatory system; Cat10=diseases of the respiratory system; Cat11= diseases of the digestive system; Cat14= diseases of the genitourinary system; Cat13= diseases of the musculoskeletal system and connective tissue.

The performance of the model was evaluated by generating a confusion matrix. Table 3-3 shows a confusion matrix that resulted from testing the results of the decision tree, and Table 3-4 shows the performance metric of the model.

Table 3-3
Confusion Matrix of the Decision Tree Model in NACRS Dataset

N=374,628	Actual yes	Actual no	Totals
Predicted yes	131,559	57,875	189,434
Predicted no	21,544	163,650	185,194
Totals	153,103	221,525	374,628

Table 3-4
Performance Metric of the Decision Tree Model in NACRS Dataset

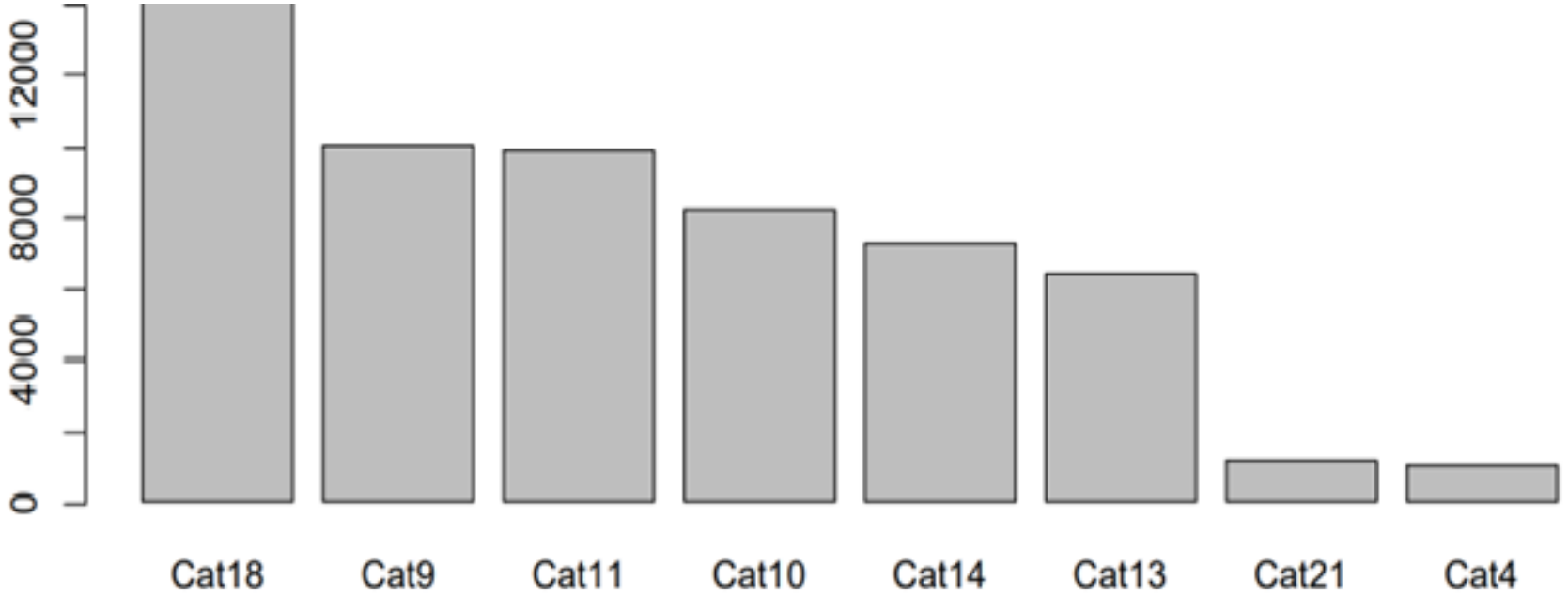
Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
78.8	86.0	73.9	69.5	76.8

Out of the total 374,628 instances, the model correctly predicted 131,559 true positive cases and 163,650 true negative cases (Table 3-3). However, it also made 21,544 false positive predictions and 57,875 false negative predictions. The overall accuracy of the model was 78.8%, reflecting the proportion of correct predictions out of all instances., The sensitivity was notable 86.0%, indicating the model's ability to correctly identify positive cases (Table 3-4). The specificity was 73.9%, representing its proficiency in identifying negative cases. The precision stood at 69.5%, showcasing the accuracy of positive predictions among all instances labeled as positive. Lastly, the F1-Score was 76.8%, which considered both precision and sensitivity, providing a balanced assessment of the model's overall performance.

The three most informative variables were identified in the following order: *Category 18*, which represents symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified; *Category 9*, denoting diseases of the circulatory system; and *Category 11*,

corresponding to diseases of the digestive system. Figure 3-2 displays the most prominent diagnostic categories, ranked according to their degree of informativeness.

Figure 3-2
The Importance of Diagnostic Categories of the Decision Tree Model in NACRS Dataset



Note. The x-axis represents ICD-10-CA categories and the y-axis represents the entropy values associated with each variable. Cat18=symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; Cat9=diseases of the circulatory system; Cat11=diseases of the digestive system; Cat10=diseases of the respiratory system; Cat14=diseases of the genitourinary system; Cat13=diseases of the musculoskeletal system and connective tissue; Cat21=factors influencing health status and contact with health services; Cat4=Endocrine, nutritional and metabolic diseases.

3.1.3 Random Forest

The target variable of the random forest model was the occurrence or non-occurrence of FRIs over a 9-year period. As discussed above, random forest is as an ensemble of multiple trees, and visualizing it is not an option. However, the variable importance can provide some information about the nature of the model and its relation to the input data. To evaluate the performance of the model, accuracy, sensitivity, specificity, precision, and F1 score were chosen among the results of the confusion matrix (Table 3-5). Table 3-6 also shows the performance metric of the random forest model.

Table 3-5
Confusion Matrix of the Random Forest Model in NACRS Dataset

N=374,455	Actual yes	Actual no	Totals
Predicted yes	131,902	57,571	189,473
Predicted no	20,876	164,106	184,982
Totals	152,778	221,677	374,455

Table 3-6
Performance Metrics of the Random Forest Model in NACRS Dataset

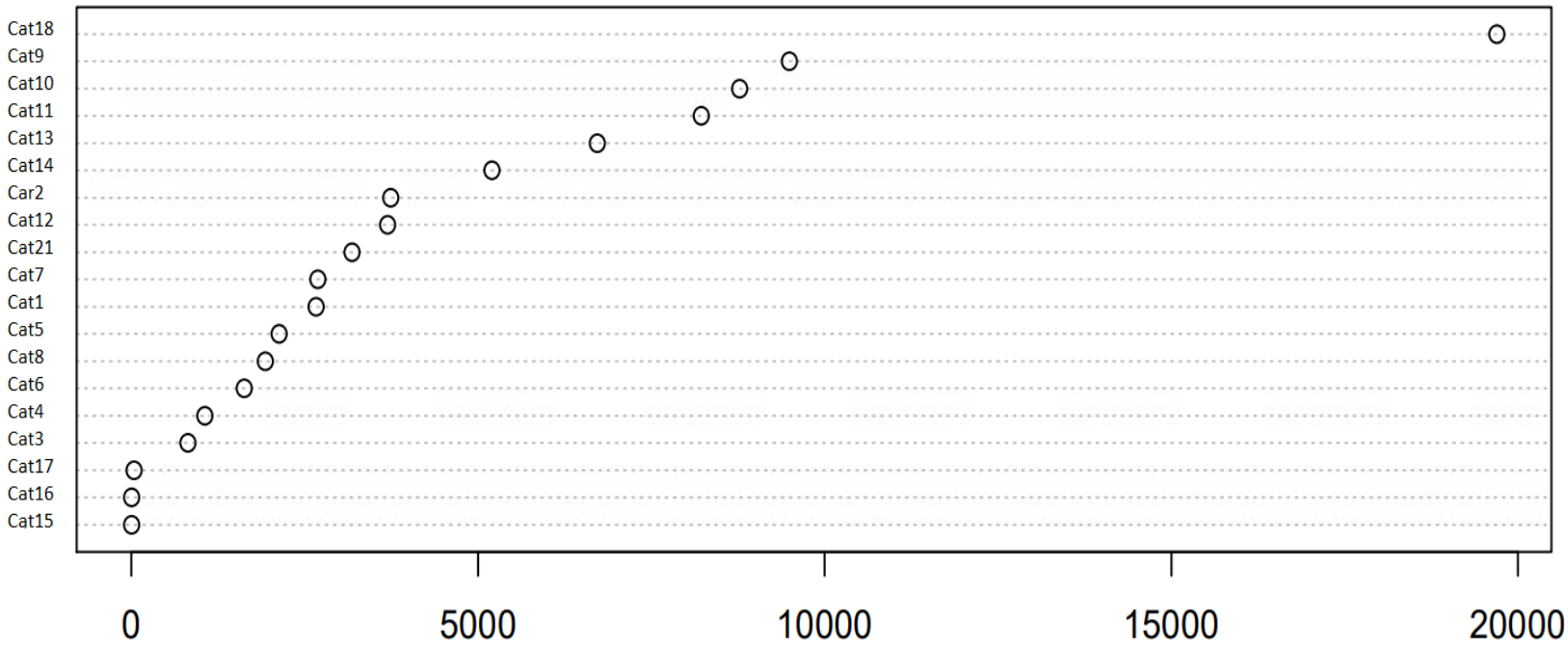
Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
78.8	86.3	73.8	69.3	76.9

Out of the total 374,455 instances, the model correctly predicted 131,902 cases as 'yes' and 164,106 cases as 'no,' while misclassifying 57,571 and 20,876 instances, respectively (Table 3-5). The model's overall accuracy was78.8%, indicating that it correctly classified approximately 79% of the cases. The model's sensitivity was 86.3%, indicating its ability to correctly identify positive cases (Table 3-6). On the other hand, the specificity, or true negative rate, stood at 73.8%, representing the model's capacity to correctly identify negative cases. The precision of the model was 69.3%, reflecting the proportion of true positive predictions among

all positive predictions. Lastly, the F1-score was 76.9%, providing an overall assessment of the model's predictive ability.

Next, the three most informative variables were identified in the following order: *Category 18*, which represents symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified; *Category 9*, denoting diseases of the circulatory system; and *Category 10*, corresponding to diseases of the respiratory system. The list of diagnostic categories in terms of their informativeness is presented in Figure 3-3.

Figure 3-3
The Informativeness of Diagnostic Categories of the Random Forest Model in NACRS Dataset



Note. The y-axis represents the category and the x-axis represents the Gini index values associated with each variable. Cat18=symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; Cat9=diseases of the circulatory system; Cat10= diseases of the respiratory system; Cat11=diseases of the digestive system; Cat13=diseases of the musculoskeletal system and connective tissue; Cat14=diseases of the genitourinary system; Cat2= neoplasms; Cat12=diseases of the skin and subcutaneous tissue; Cat21=factors influencing health status and contact with health services; Cat7=diseases of the skin and subcutaneous tissue; Cat1=certain infectious and parasitic diseases; Cat5= mental and behavioural disorders; Cat8=diseases of the ear and mastoid process; Cat6=diseases of the nervous system; Cat4=endocrine, nutritional and metabolic diseases Cat3=diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism; Cat17= congenital malformations, deformations, and chromosomal abnormalities; Cat16= certain conditions originating in the perinatal period; Cat15= pregnancy, childbirth and the puerperium.

3.1.4 Extreme Gradient Boosting Tree (XGBoost Tree)

A parameter search was conducted to determine the optimal number of iterations, and it was determined that 100 iterations yielded the highest level of model performance, as shown in Figure 3-4. To evaluate the performance of the model, a confusion matrix was generated. The results of the confusion matrix are presented in Table 3-7 and the performance metrics are also provided in Table 3-8.

Table 3-7
Confusion Matrix of the XGBoost Model in NACRS Dataset

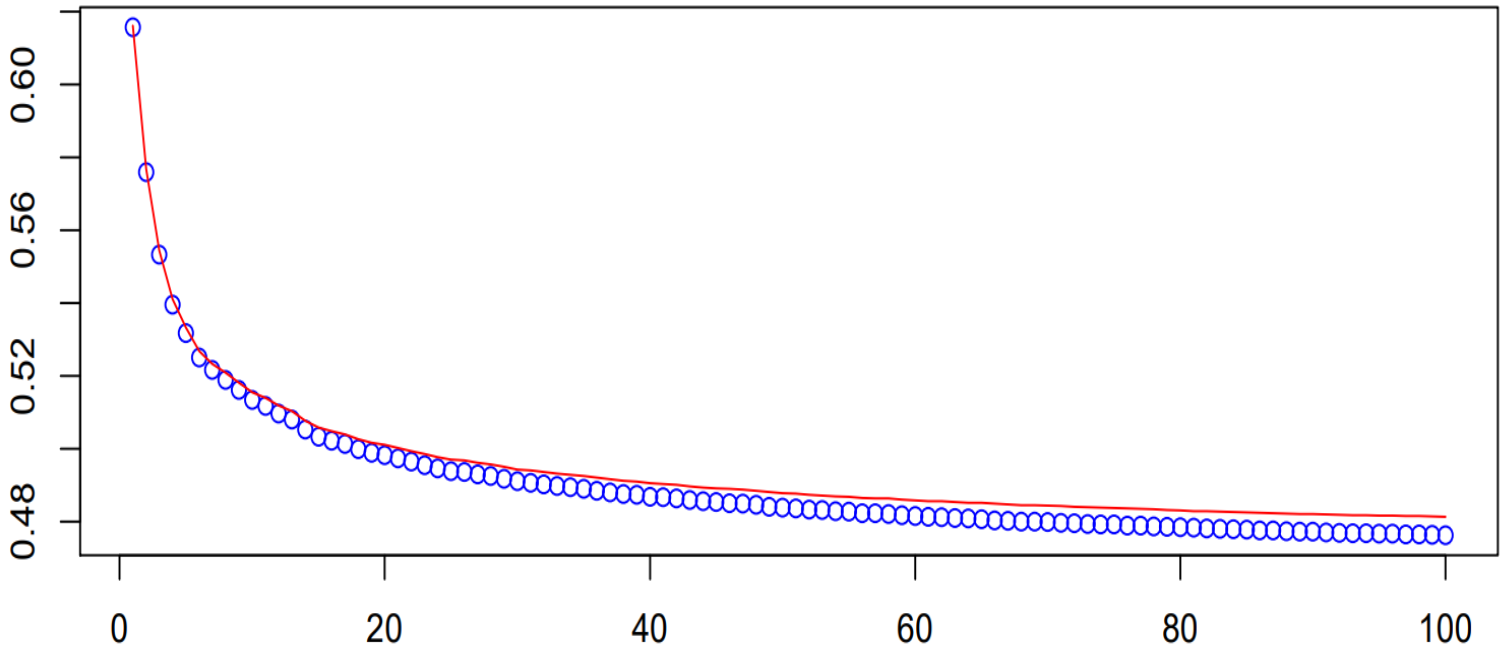
N=374,628	Actual yes	Actual no	Totals
Predicted yes	133,762	24,401	158,163
Predicted no	55,672	160,793	216,465
Totals	189,434	185,194	374,628

Table 3-8
Performance Metrics of the XGBoost Model in NACRS Dataset

Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
78.6	70.6	86.8	84.6	77.0

The confusion matrix (Table 3-7) shows that out of the total 374,628 instances, the model correctly predicted 133,762 instances as positive and 160,793 instances as negative. However, there were 24,401 false positives and 55,672 false negatives. Overall, the model achieved an accuracy of 78.6%, indicating that approximately 79% of the predictions were correct.

Figure 3-4
Parameter Search for XGBoost Model in NACRS: Number of Trees

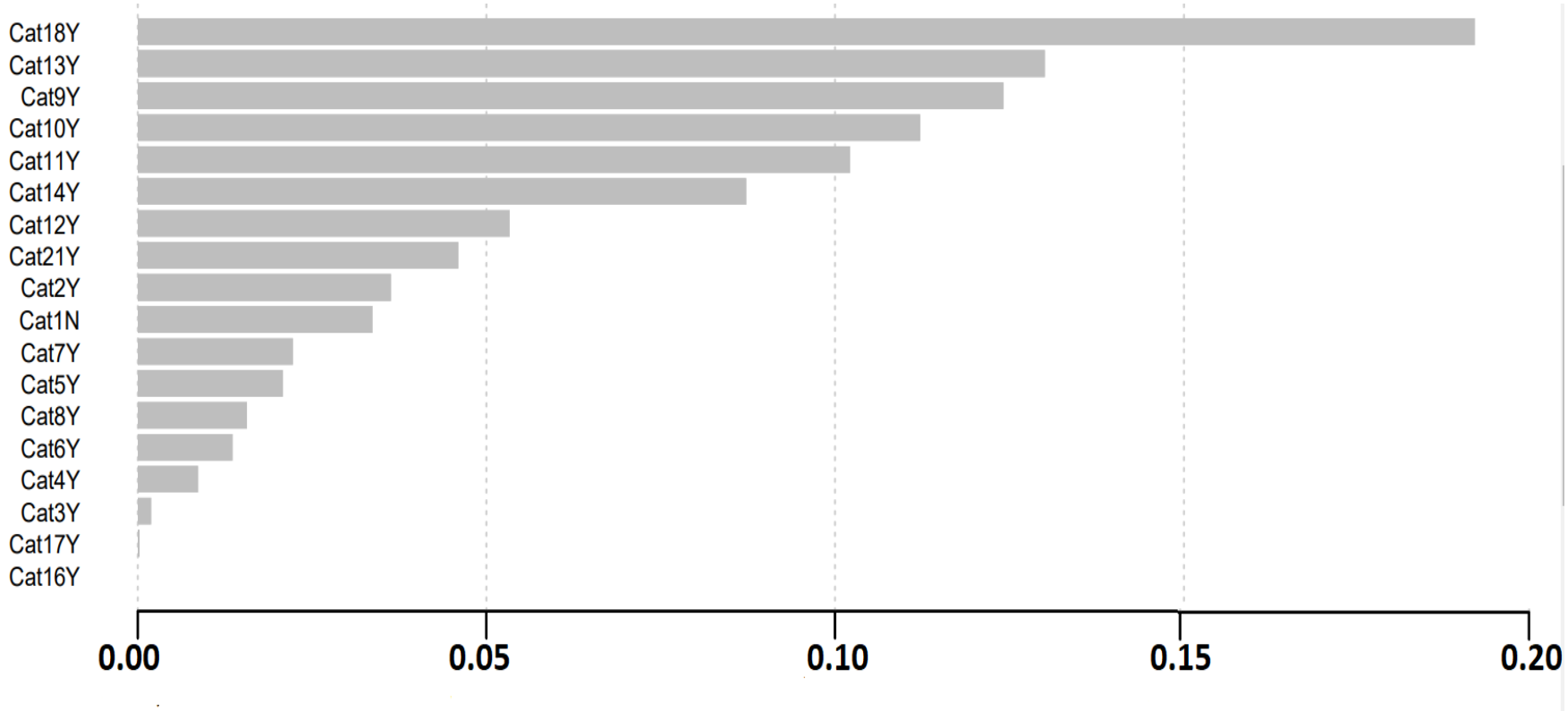


Note. The graph displays the performance of a trained XGBoost tree model during the training process. The y-axis represents the train-mlogloss, which is a measure of the model's multiclass log loss on the training data. The x-axis represents the number of iterations (in this case number of trees) or rounds during the training process.

Further analysis of the performance metrics (Table 3-8), shown that the model demonstrated a sensitivity at 70.6%, indicating its ability to correctly identify true positives relative to the actual positive cases. The model showed high specificity at 86.8%, indicating its proficiency in correctly identifying true negatives relative to the actual negative cases. The precision of 84.6% indicated the proportion of true positive predictions among all positive predictions, signifying the model's ability to avoid false positives. The F1-score was at 77.0%, reflecting a balanced trade-off between precision and sensitivity.

The three most informative variables were subsequently identified, in order, as *Category 18*, which pertains to symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified, *Category 13*, which relates to diseases of the musculoskeletal system and connective tissue, and *Category 9*, which relates to diseases of the circulatory system. The list of diagnostic categories in terms of their informativeness is presented in Figure 3-5.

Figure 3-5
The Informativeness of Diagnostic Categories of the XGBoost in NACRS Dataset



Note. The numbers on the x-axis correspond to the variable importance scores. These scores quantify the relative importance of each variable in influencing the model's predictions. Higher values indicate greater importance, while lower values indicate lesser importance. Cat18= symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; Cat13= diseases of the musculoskeletal system and connective tissue; Cat9= diseases of the circulatory system; Cat10= diseases of the respiratory system; Cat11= diseases of the digestive system; Cat14= diseases of the genitourinary system; C12= diseases of the skin and subcutaneous tissue; Cat21= factors influencing health status and contact with health services; Cat2= neoplasms; Cat1=certain infectious and parasitic diseases; Cat7=diseases of the eye and adnexa, Cat5= mental and behavioral disorders; Cat8= diseases of the ear and mastoid process; Cat6= diseases of the nervous system; Cat4= endocrine, nutritional and metabolic diseases; Cat3= diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism; Cat17= congenital malformations, deformations, and chromosomal abnormalities; Cat16= certain conditions originating in the perinatal period.

3.1.5 Summary of Outcomes for the Emergency Department

Table 3-9 shows the first three most informative variables in three different models identifying the categories of ICD-10-CA diagnostic codes that were the most informative in association with FRIs. In all three models *Category 18* emerged as the most informative. To further identify the most informative diagnostic subcategories within *Category 18*, an in-depth exploration was conducted using the XGBoost model, known for its robustness and stable results (Hastie et al., 2009).

Table 3-9
Summary of Variable Informativeness in Three Machine Learning Models in NACRS Dataset

	Decision tree	Random forest	XGBoost tree
1st category	*Category 18*: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	*Category 18*: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	*Category18*: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
2nd category	*Category 9*: diseases of the circulatory system	*Category 9*: diseases of the circulatory system	*Category13*: diseases of the musculoskeletal system and connective tissue
3rd category	*Category 11*: diseases of the digestive system	*Category 11*: diseases of the digestive system	*Category 9*: diseases of the circulatory system

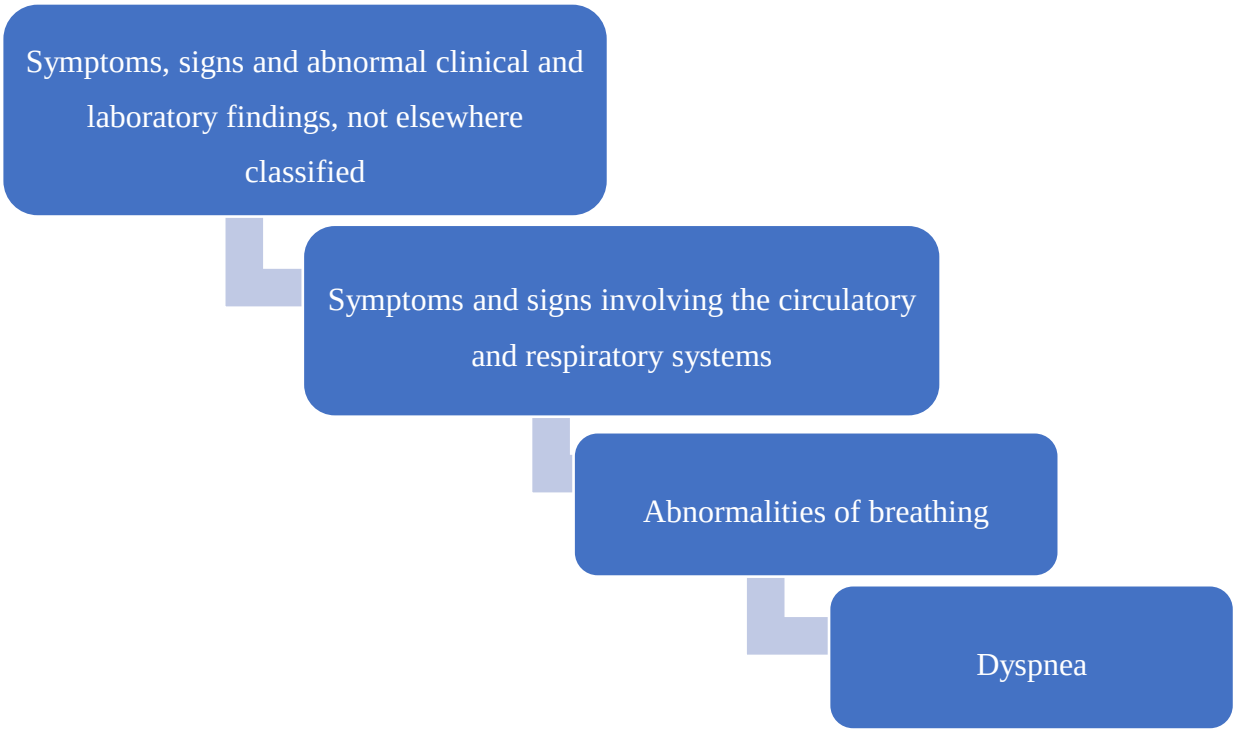
Interestingly, all three models consistently ranked *Category 18*: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified as the most informative category. This suggests that this category plays a crucial role in predicting the outcome variable across all three models. As the second most informative category, the decision tree and random forest models identified *Category 9*: diseases of the circulatory system, while the XGBoost tree model identified *Category 13*: diseases of the musculoskeletal system and connective tissue. Similarly, in the third most informative category, the decision tree and random forest models once again agreed on *Category 11*: diseases of the digestive system as the most informative category, while the XGBoost tree model highlighted *Category 9:* diseases of the circulatory

system instead.

3.1.6 Diagnostic Subcategories

To identify the most informative diagnostic subcategories within *Category 18*, an in-depth exploration was conducted. The XGBoost model was utilized for this purpose, as it is known for its robustness and stable results (Hastie et al., 2009). *Category 18* was comprised of 13 subcategories of codes from *R00* to *R99*. The results show that dyspnea or shortness of breath was the most informative ICD-10-CA code (Figure 3-6). The complete list of subcategories of *Category 18* can be found in Appendix G and the graph of informativeness of each subcategory can be found in Appendix H.

Figure 3-6
The Most Informative Subcategories in Category 18: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified



3.2 Hospital Level of Care

3.2.1 Demographics

The current study analyzed a total of 688,868 observations in the DAD dataset (hospitals) out of which 304,495 observations were identified as FRIs. Table 3-10 provides age and Table 3-11 provides sex information for the identified FRI observations in the DAD dataset. The age group of 75 to 79 years had the highest incidence of FRIs, while the age group over 90 experienced the lowest number of FRIs. The overall mean age of patients was 79.8 years. Moreover, females experienced more FRIs compared to males.

Table 3-10
Age Distribution of FRI Observations in Hospitals of Ontario (DAD Dataset)

Age group	Number of observations	Percentage of total frequency (%)
65-69	37,819	12.4
70-74	37,948	12.5
75-79	78,688	25.9
80-84	59,629	19.6
85-89	53,426	17.5
Over 90	36,985	12.1
Total	304,495	100.0

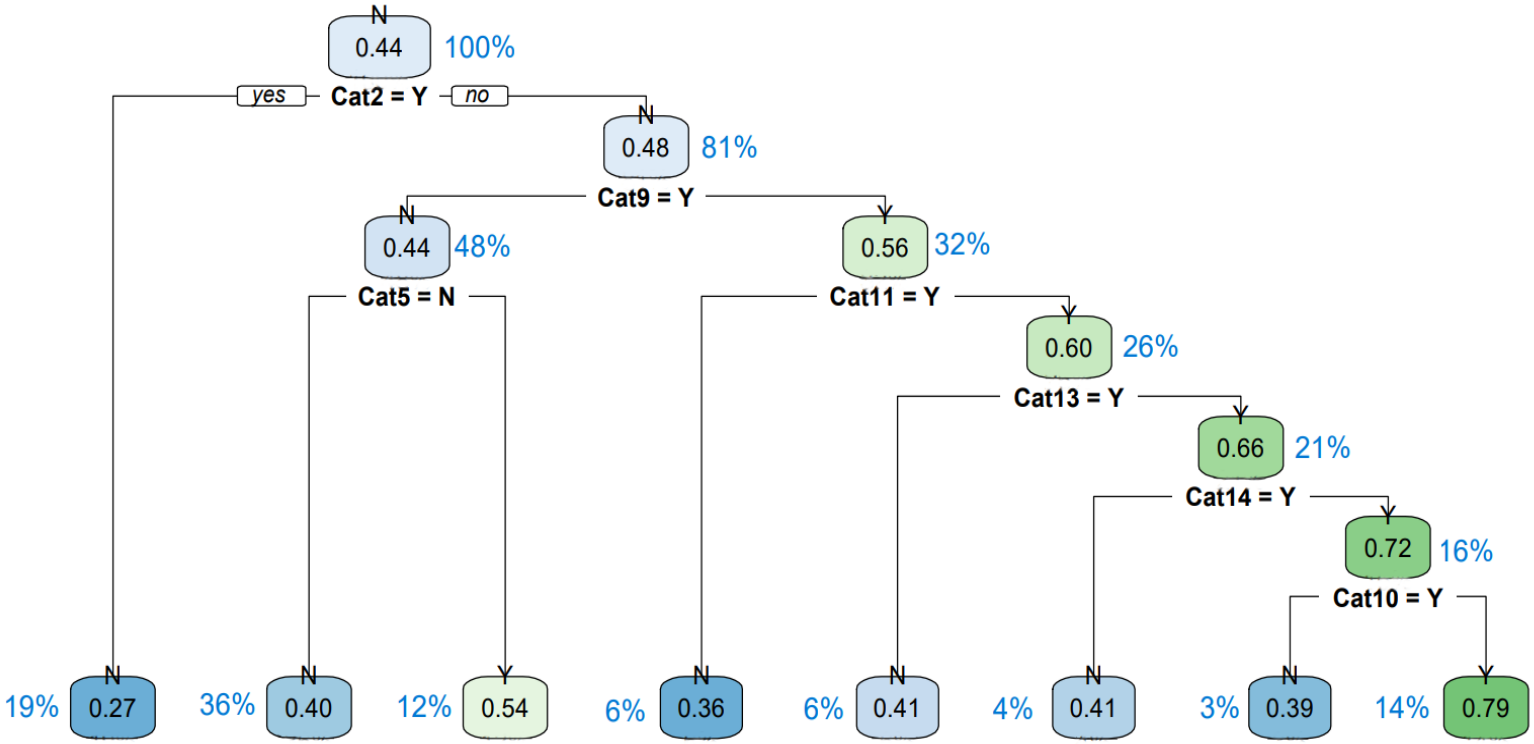
Table 3-11
Sex Distribution of FRI Observations in Hospitals of Ontario (DAD Dataset)

Sex	Number of observations	Percentage of total frequency (%)
Females	187,387	61.5
Males	117,108	38.5
Total	304,495	100.0

3.2.2 Decision Tree

The target variable of the tree was the occurrence or non-occurrence of FRIs over a 9-year period. Input variables were composed of 21 categories of diagnostic codes listed in Appendix F. The final decision tree model is presented in Figure 3-7.

Figure 3-7
Visualization of the Decision Tree to Predict FRIs Using Diagnostic Categories of ICD-10-CA Codes in DAD Dataset



Note. The tree consists of various components: starting from the root node, each internal node represents a decision based on specific input variables and their conditions (Y=Yes or N=No). The branches emanating from each internal node depict the possible outcomes based on the input variables' conditions. The leaf nodes, or terminal nodes, signify the final predictions, representing specific combinations of input variables. Within each leaf node, the percentages indicate the proportion of observations falling into that category (Y=Yes or N=No). By examining the tree's structure, splitting criteria, branches, leaf nodes, and the percentages in each leaf, we can gain a comprehensive understanding of the relationship between the input variables and the predicted outcomes of FRIs based on the ICD-10-CA diagnostic categories. The use of lighter and darker shades of blue and green conveys information about the certainty or probability associated with the predicted outcomes. Darker shades of blue or green indicate a higher probability or a stronger association with a particular outcome, representing a higher confidence in the prediction being made. On the other hand, lighter shades of blue or green suggest lower probabilities or weaker associations with the predicted outcomes, indicating a lower confidence in the corresponding predictions. This color scheme helps to distinguish between more reliable or significant associations depicted by darker colors and less robust relationships represented by lighter colors, allowing for a visual assessment of the model's confidence in specific conditions or categories within the decision tree; Cat2= neoplasms; Cat9= diseases of the circulatory system; Cat5= mental and behavioural disorders; Cat11=; diseases of the digestive system Cat13= diseases of the musculoskeletal system and connective tissue; Cat14= diseases of the genitourinary system; Cat10= diseases of the respiratory system.

The performance of the model was evaluated by generating a confusion matrix. Table 3-12 shows a confusion matrix that resulted from testing the results of the decision tree, and Table 3-13 also shows the performance metric of the model.

Table 3-12
Confusion Matrix of the Decision Tree Model in DAD Dataset

N=206,817	Actual yes	Actual no	Totals
Predicted yes	57,292	34,179	91,471
Predicted no	21,711	93,635	115,346
Totals	79,003	127,814	206,817

Table 3-13
Performance Metric of the Decision Tree Model in DAD Dataset

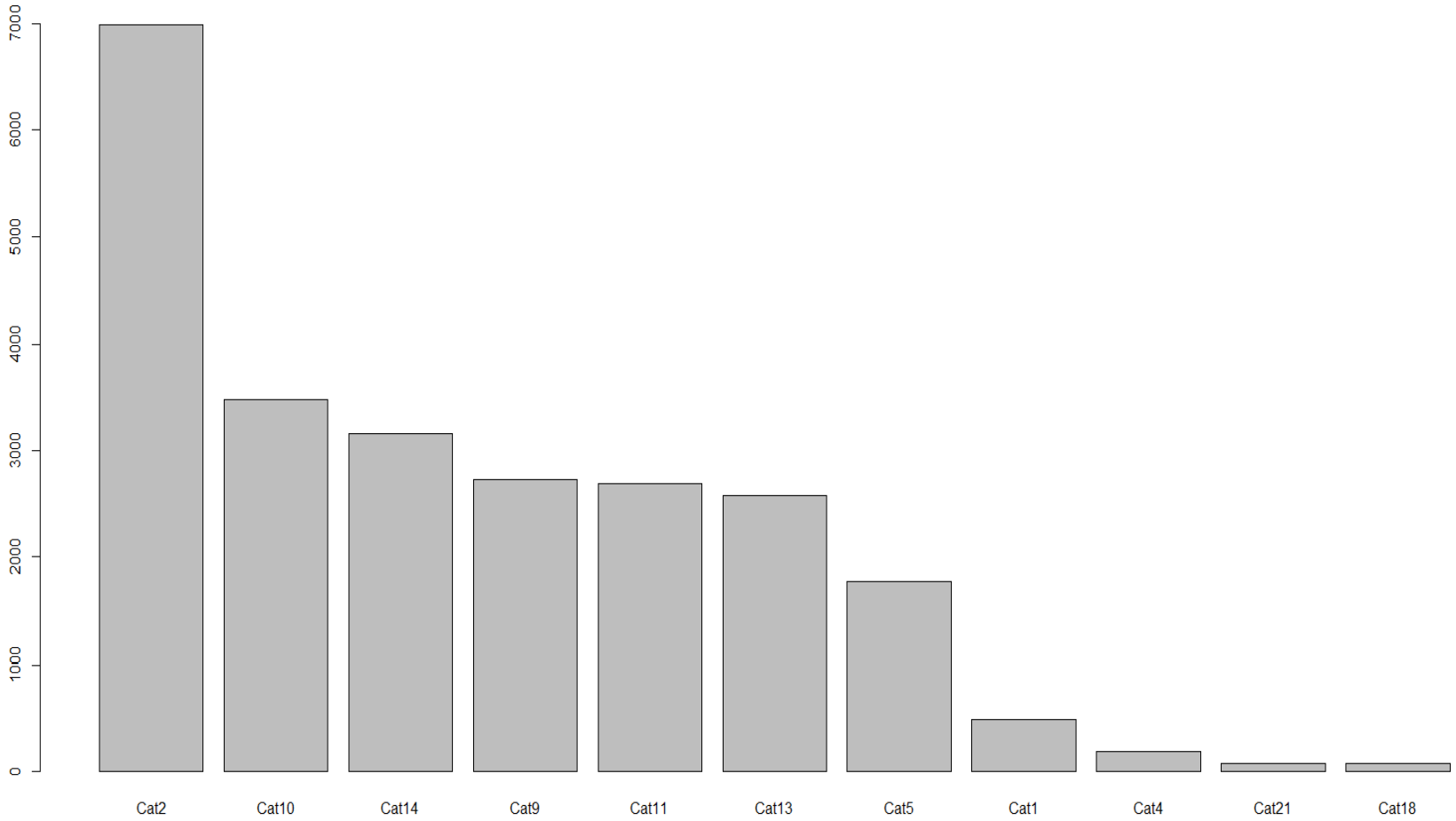
Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
73.0	72.5	73.2	62.6	67.2

The confusion matrix (Table 3-12) reveals that out of the total 206,817 instances, the model correctly predicted 57,292 true positive cases and 93,635 true negative cases. However, it also made 21,711 false positive predictions and 34,179 false negative predictions. The overall accuracy of the model is reported at 73.0%, indicating that approximately three-fourths of the predictions were correct. Moving on to the performance metrics (Table 3-13), the model exhibited a sensitivity of 72.5%, indicating its performance in correctly identifying positive cases. The specificity stands at 73.2%, suggesting the model's ability to accurately identify negative cases. The precision (positive predictive value) achieved by the model was 62.6%, demonstrating the accuracy of positive predictions among all instances labeled as positive. Moreover, the F1-Score, which considers both precision and sensitivity, stood at 67.2%, providing a balanced evaluation of the model's overall performance.

The three most informative variables were identified in the following order: *Category 2*, which represents neoplasms; *Category 10*, denoting diseases of the respiratory system; and

Category 14, corresponding to diseases of the genitourinary system. Figure 3-8 displays the most prominent diagnostic categories, ranked according to their degree of informativeness.

Figure 3-8
The Importance of Diagnostic Categories of the Decision Tree Model in DAD Dataset



Note. The x-axis represents the name of each category and the y-axis represents the entropy values associated with each variable. Cat2=neoplasms; Cat10=diseases of the respiratory system; Cat14= diseases of the genitourinary system; Cat9= diseases of the circulatory system; Cat11=diseases of the digestive system; Cat13=diseases of the musculoskeletal system and connective tissue; Cat5= mental and behavioural disorders; Cat1= certain infectious and parasitic diseases; Cat4= endocrine, nutritional and metabolic diseases; Cat21= factors influencing health status and contact with health services; Cat18= symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified.

3.2.3 Random Forest

The random forest is an ensemble of multiple trees, and visualizing it is not an option. However, the variable importance can provide some information about the nature of the model and its relation to the input data. To evaluate the performance of the model, accuracy, sensitivity, specificity, precision, and F1 score were chosen among the results of the confusion matrix (Table 3-14). Table 3-15 shows the performance metric of the random forest model.

Table 3-14
Confusion Matrix of the Random Forest Model in DAD Dataset

N=206,676	Actual yes	Actual no	Totals
Predicted yes	52,659	38,956	91,615
Predicted no	13,228	101,833	115,061
Totals	65,887	140,789	206,676

Table 3-15
Performance Metrics of the Random Forest Model in DAD Dataset

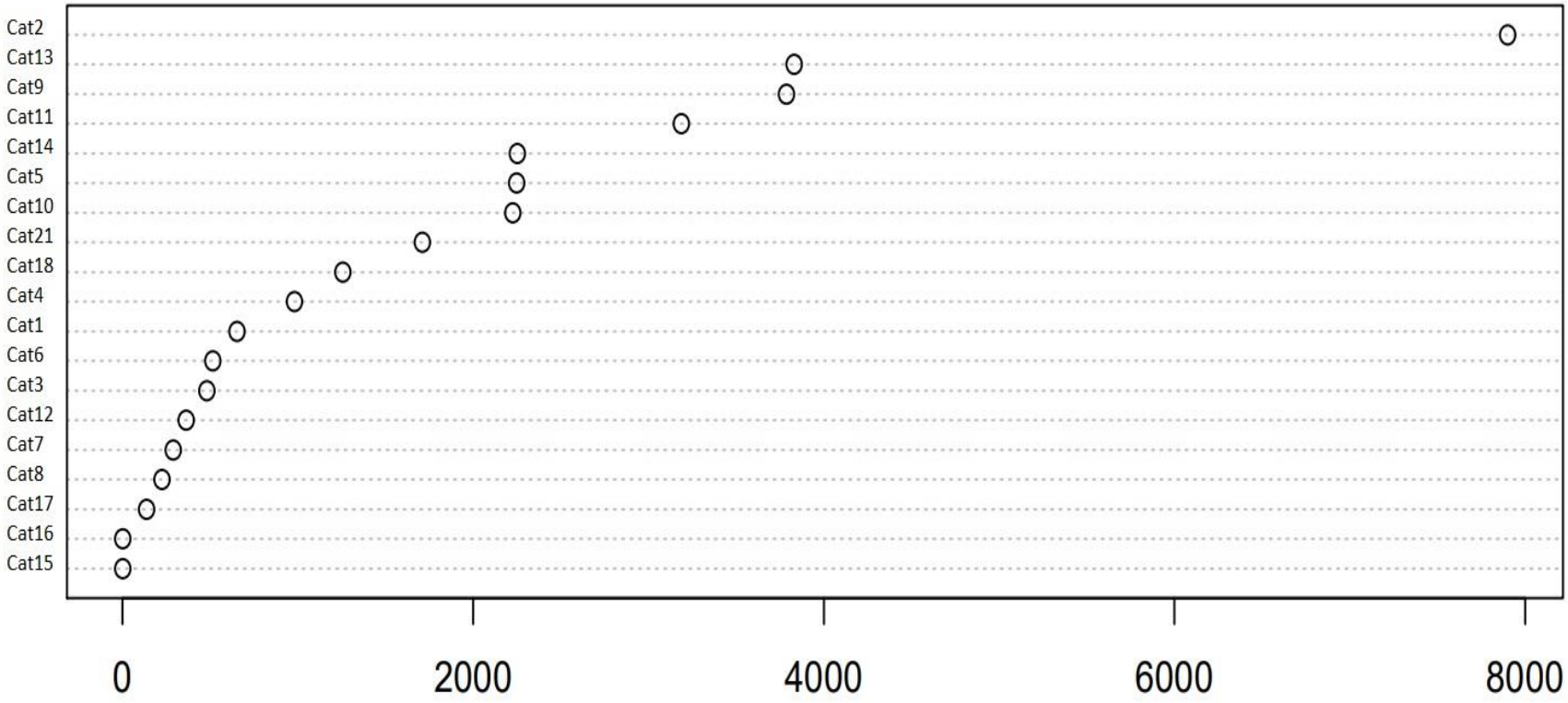
Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
74.8	79.9	72.3	57.5	66.9

With a total of 206,676 instances, the model correctly predicted 52,659 cases as yes and 101,833 cases as no, while misclassifying 13,228 and 38,956 instances, respectively (Table 3-14). This demonstrates the model's ability to capture a substantial number of true positives and true negatives. However, there were instances where the model made false positive and false negative predictions. The total accuracy of the model stood at 74.8%, indicating that it classified around 75% of the cases correctly. According to Table 3-15 the sensitivity was 79.9%, indicating the model's effectiveness in correctly identifying positive cases. The specificity was 72.3%, representing the model's proficiency in correctly identifying negative cases. The precision of the model was 57.5%, reflecting the proportion of true positive predictions among all positive

predictions. The F1-score, which takes both precision and sensitivity into account, stood at 66.9%, providing an overall evaluation of the model's predictive ability.

Next, the three most informative variables were identified in the following order: *Category 2*, which represents neoplasms; *Category 13*, denoting diseases of the musculoskeletal system and connective tissue; and *Category 9*, corresponding to diseases of the circulatory system. The list of diagnostic categories in terms of their informativeness is presented in Figure 3-9.

Figure 3-9
The Informativeness of Diagnostic Categories of the Random Forest Model in DAD Dataset

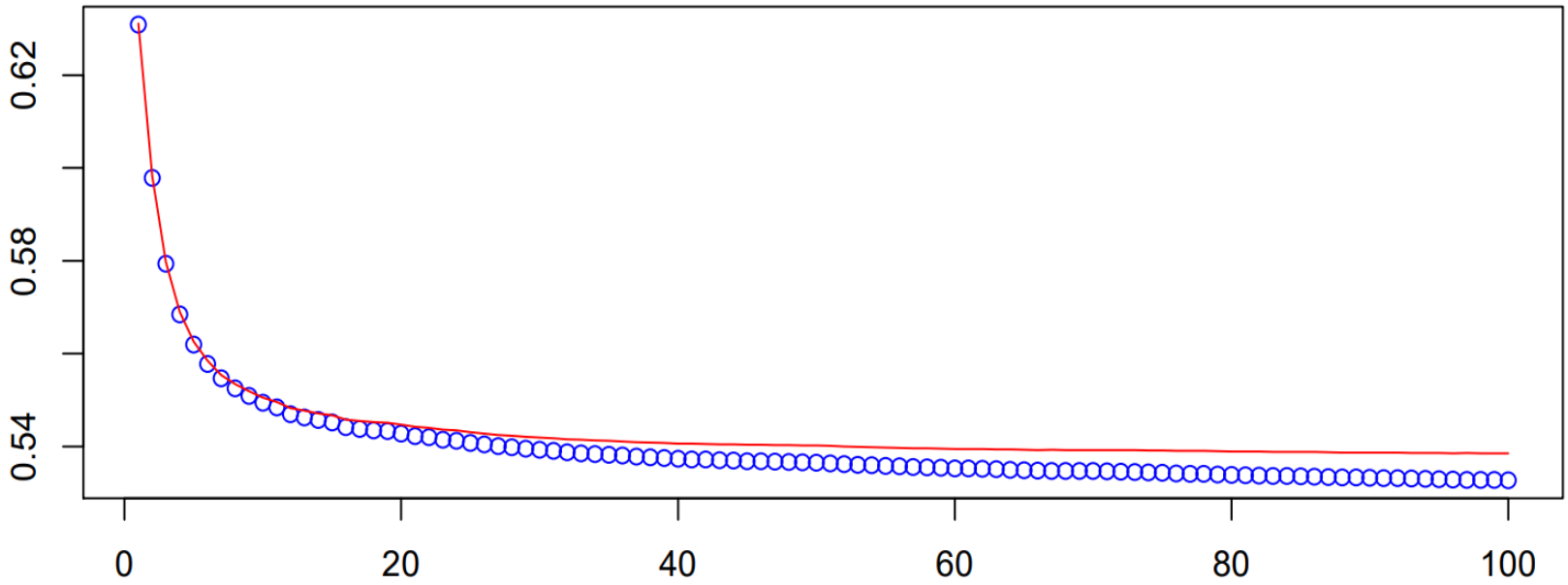


Note. The y-axis represents the name of each category and the x-axis represents the Gini index values associated with each variable. Cat2= neoplasms; Cat13=diseases of the musculoskeletal system and connective tissue; Cat9=diseases of the circulatory system; Cat11=diseases of the digestive system; Cat14=diseases of the genitourinary system; Cat5= mental and behavioural disorders; Cat10= diseases of the respiratory system; Cat21=factors influencing health status and contact with health services; Cat18=symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; Cat1=certain infectious and parasitic diseases; Cat6=diseases of the nervous system; Cat3=diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism; Cat12=diseases of the skin and subcutaneous tissue; Cat7=diseases of the skin and subcutaneous tissue; Cat8=diseases of the ear and mastoid process; Cat17= congenital malformations, deformations, and chromosomal abnormalities; Cat16= certain conditions originating in the perinatal period; Cat15= pregnancy, childbirth and the puerperium.

3.2.4 Extreme Gradient Boosting Tree (XGBoost Tree)

A parameter search was conducted to determine the optimal number of iterations, and it was determined that 100 iterations yielded the highest level of model performance, as shown in Figure 3-10. To evaluate the performance of the model, a confusion matrix was generated. The results of the confusion matrix are presented in Table 3-16 and the performance metrics is also provided in Table 3-17.

Figure 3-10
Parameter Search for XGBoost Model in DAD: Number of Trees



Note. The graph displays the performance of a trained XGBoost tree model during the training process. The y-axis represents the train-mlogloss, which is a measure of the model's multiclass log loss on the training data. The x-axis represents the number of iterations (in this case number of trees) or rounds during the training process.

Table 3-16
Confusion Matrix of the XGBoost Model in DAD Dataset

N=206,817	Actual yes	Actual no	Totals
Predicted yes	52,837	14,174	67,011
Predicted no	38,634	101,172	139,806
Totals	91,471	115,346	206,817

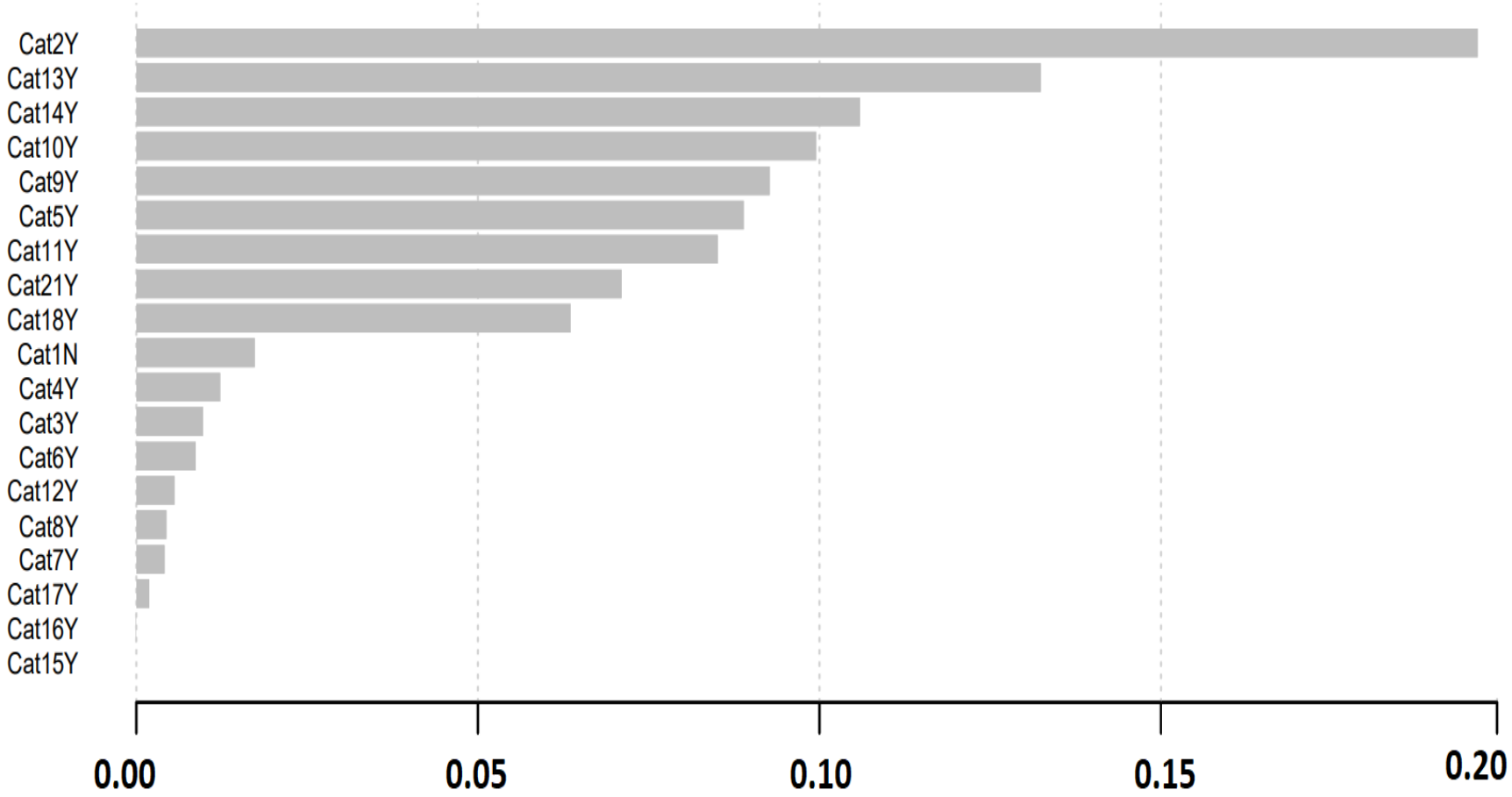
Table 3-17
Performance Metrics of the XGBoost Model in DAD Dataset

Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
74.5	57.8	87.7	78.9	66.7

The model achieved 74.5% accuracy, correctly predicting 52,837 instances as positive and 101,172 instances as negative. However, it misclassified 14,174 instances as false positives and 38,634 instances as false negatives. The sensitivity of 57.8% indicates the model's ability to accurately identify true positives in relation to the actual positive cases, while the high specificity of 87.7% signifies its proficiency in correctly identifying true negatives among the actual negative cases. Moreover, the precision of 78.9% highlights the proportion of true positive predictions relative to all positive predictions, indicating the model's ability to minimize false positives. The F1-Score of 66.7% reflected a balanced trade-off between precision and sensitivity.

Further analysis was conducted to determine the informativeness of the variables. The first three most informative variables were identified, in order, as *Category 2*, which pertains to neoplasms, *Category 13*, which relates to diseases of the musculoskeletal system and connective tissue, and *Category 14* which is for diseases of the genitourinary system. A visual representation of the full list of diagnostic categories in terms of informativeness is shown in Figure 3-11.

Figure 3-11
The Informativeness of Diagnostic Categories of the XGBoost Model in DAD Dataset



Note. The numbers on the x-axis correspond to the variable importance scores. These scores quantify the relative importance of each variable in influencing the model's predictions. Higher values indicate greater importance, while lower values indicate lesser importance. Cat2= neoplasms; Cat13= diseases of the musculoskeletal system and connective tissue; Cat14= diseases of the genitourinary system; Cat10= diseases of the respiratory system; Cat9= diseases of the circulatory system; Cat5= mental and behavioral disorders; C11= diseases of the digestive system; C21= factors influencing health status and contact with health services; C18= symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; C1=certain infectious and parasitic diseases; C4= endocrine, nutritional and metabolic diseases; Cat6= diseases of the nervous system; C12= diseases of the skin and subcutaneous tissue; C8= diseases of the ear and mastoid process; C7=diseases of the eye and adnexa; ; Cat17= congenital malformations, deformations, and chromosomal abnormalities; Cat16= certain conditions originating in the perinatal period; Cat15= pregnancy, childbirth and the puerperium.

3.2.5 Summary of Outcomes at the Hospital Level of Care

Table 3-18 shows the results of the first three most informative variables in different models to determine which categories of ICD-10-CA diagnostic codes were deemed more informative in FRIs. The variable identified as the most informative among all three models was *Category 2*. To identify the most informative diagnostic subcategories within *Category 2*, an in-depth exploration was conducted. The XGBoost model was utilized for this purpose, as it is known for its robustness (Hastie et al., 2009).

Table 3-18
Summary of Variable Informativeness in Three Machine Learning Models in DAD Dataset

	Decision tree	Random forest	XGBoost tree
1st category	*Category 2*: neoplasms	*Category 2*: neoplasms	Category 2: neoplasms
2nd category	*Category 10*: diseases of the respiratory system	*Category 13*: diseases of the musculoskeletal system and connective tissue	*Category 13*: diseases of the musculoskeletal system and connective tissue
3rd category	*Category 14*: diseases of the genitourinary system	*Category 9*: diseases of the circulatory system	*Category 14*: diseases of the genitourinary system

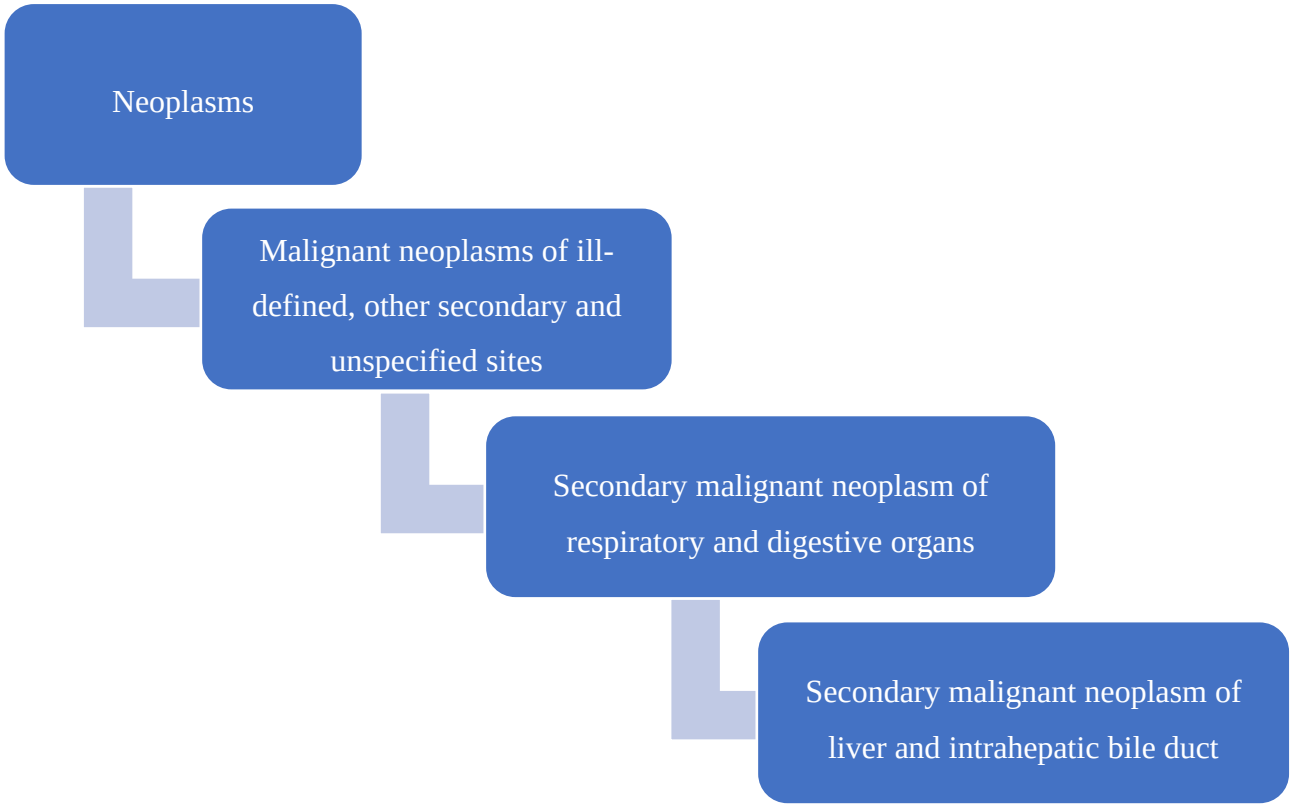
Across the three models, *Category 2*: neoplasms consistently emerged as the most informative variable, indicating its significant impact on the model's predictions in all three cases. For the decision tree model, the second most informative category was *Category 10*: diseases of the respiratory system. In contrast, for the random forest model, the second most informative category was *Category 13*: diseases of the musculoskeletal system and connective tissue. The XGBoost tree model shared this second spot with *Category 13*, showcasing its importance in both random forest and XGBoost models. The third most informative category varied among the models. For the decision tree model, it was *Category 14*: diseases of the genitourinary system, and for the random forest model it was *Category 9*: diseases of the circulatory system. Finally, the XGBoost tree model, much like the decision tree model, identified *Category 14*: diseases of the genitourinary system as the third most informative

category. This summary provides valuable insights into the variable importance across different machine learning models applied to the DAD dataset. The consistent presence of *Category 2*: neoplasms as the most informative variable underlined its significance in predicting the target variable.

3.2.6 Diagnostic Subcategories

Category 2 was comprised of 20 subcategories of codes from C00 to D48. The results show that secondary malignant neoplasm of liver and intrahepatic bile duct was the most informative disease (Figure 3-12). The complete list of subcategories of *Category 2* can be found in Appendix G and the graph of informativeness of each subcategory can be found in Appendix H.

Figure 3-12
The Most Informative Subcategories in Category 2 (neoplasms)



3.3 Summary of Model Performances

To answer the third research question, a comparison was carried out to identify similarities and differences in performance of the three machine learning models (Table 3-19). First informative variable In NACRS was *Category 18* while in DAD it was *Category 2*. DAD has a more comprehensive data collection for diagnostic codes (25) and thus a larger suite of options. NACRS data was collected in the EDs and therefore has fewer diagnostic codes (10). Table 3-19 also shows the performance metrics of machine learning models in NACRS and DAD datasets. The first informative variable used in all the models is *Category 18* for the NACRS dataset and *Category 2* for the DAD dataset.

Table 3-19
Comparing the Performance Metrics of Machine Learning Models in NACRS and DAD Datasets

	Accuracy	Sensitivity	Specificity	Precision	F1 Score	First informative variable
			NACRS			
Decision tree	78.8	86.0	73.9	69.5	76.8	Cat18
Random forest	78.8	86.3	73.8	69.3	76.9	Cat18
XGBoost	78.6	70.6	86.8	84.6	77.0	Cat18
			DAD			
Decision tree	73.0	72.5	73.2	62.6	67.2	Cat2
Random forest	74.8	79.9	72.3	57.5	66.9	Cat2
XGBoost	74.5	57.8	87.7	78.9	66.7	Cat2

Looking at the results for the NACRS dataset, both decision tree and random forest achieved the same accuracy of 78.8%. The improvement in accuracy was not substantial after deploying ensemble learning algorithms. This is because of the categorical nature of predictive variables, as discussed further in the discussion chapter. They exhibited similar sensitivity and specificity values, with XGBoost showing slightly lower sensitivity but higher specificity. Precision and F1 score, which measure the model's ability to correctly classify positive instances, are quite close for all three models, with XGBoost slightly outperforming the other two.

In DAD dataset, random forest achieved the highest accuracy at 74.8%, followed closely

by XGBoost at 74.5%, and decision tree at 73.0%. Sensitivity for random forest was the highest at 79.9%, indicating its ability to detect true positive instances. However, XGBoost achieved the highest specificity at 87.7%, suggesting its capability to identify true negative instances effectively. Precision and F1 Score for random forest and XGBoost were comparable, with decision tree lagging in these metrics. In summary, random forest was the most accurate and sensitive model in both NACRS and DAD datasets.

Chapter 4

4. Discussion

This study utilized healthcare administrative data and ICD-10-CA diagnostic codes to explore the power of three machine learning algorithms to determine the informative variables associated with FRIs in older adults. The findings revealed that one diagnostic category emerged as the most informative when associated with FRIs in ED level of care which was *Category 18* or symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified with it's the most informative ICD-10-CA code dyspnea. At hospital level of care, *Category 2* or neoplasms was the most informative category with secondary malignant neoplasm of liver and intrahepatic bile duct being its most informative disease. There is a notable difference in ICD-10-CA diagnostic categories associated with FRIs reported in EDs and hospitals with minimal overlap in informativeness of the top three code categories. The most accurate and sensitive machine learning model for determining associations between ICD-10-CA diagnostic codes and FRIs in both ED level of care and hospitals was the random forest model.

At the ED level of care, dyspnea, which refers to shortness of breath and is a key symptom of chronic obstructive pulmonary disease (Marciniuk et al., 2011), was the most informative ICD-10-CA code associated with FRIs. Dyspnea has been considered a risk factor for falls and FRIs in previous studies. The association between dyspnea and chronic obstructive pulmonary disease with falls and injuries due to falls has been explored previously. For example, a study by Ozalevli and colleagues examined the connection between disease-related factors and balance, as well as a history of falls in patients with chronic obstructive pulmonary disease. The findings revealed that hypoxemia, dyspnea, and fatigue were correlated with balance impairment and falls. Authors recommended that assessing and enhancing balance should be incorporated into pulmonary rehabilitation programs for older adult patients with chronic obstructive pulmonary disease (Ozalevli et al., 2011). Roig et al. (2011) also found that people with chronic obstructive pulmonary disease have a high susceptibility to falls related to worsening of dyspnea. Lawlor et al. (2003) discovered that older adult women aged 60 to 79 living with circulatory diseases (6.2%), chronic obstructive pulmonary disease (8%), arthritis (17.4%), and depression (9.4%) have an elevated susceptibility to falling, with chronic illnesses potentially contributing to as much as 30% of falls within this demographic. Furthermore, earlier cross-sectional studies

conducted in individuals with chronic obstructive pulmonary disease have shown a heightened prevalence of falls in this group, ranging from 44% to 51% (Beauchamp et al., 2009, 2012; Roig et al., 2011), when contrasted with community-dwelling older adults, where the prevalence of falls typically falls within the range of 29% to 33% (Bongue et al., 2011; O'loughlin et al., 1993; Tromp et al., 2001). Prabhakaran et al. (2020) indicated that chronic obstructive pulmonary disease increased the risk of fall-related readmissions in older adults. Roo et al.,(2015) indicated that pulmonary disease was a leading cause of readmission for people aged 50 years or older who had been hospitalized for a fall-related fracture. Additionally, Choi et al. (2019) revealed that among older adults who received medical attention for fall injuries, lung disease was associated with a higher risk of ED visits and hospital stay.The connection between falls and injurious falls can also warrant investigation in the context of medications commonly used by individuals with chronic obstructive pulmonary disease. Notable examples of such medications include albuterol (Bone et al., 1994), which has been associated with tremors (Vogelmeier, 2014), ipratropium (Bone et al., 1994) , known for its potential to cause blurred vision (Sharafkhaneh et al., 2013) , and fluticasone (Wedzicha et al., 2016), which may lead to long-term effects on bone density (Caramori et al., 2019), potentially increasing the risk of falling among affected individuals. It is worth noting that further research is essential to delve deeper into this relationship and better understand its nuances. All this evidence confirms that the results of this study are in agreement with the existing literature. However, more research is needed to investigate the relationship between pulmonary diseases and their related symptom, dyspnea with falls and FRIs.

At the hospital level, the most informative disease associated with FRIs was secondary malignant neoplasm of the liver and intrahepatic bile duct. Intrahepatic bile duct cancer specifically refers to cancer that develops within the bile ducts located inside the liver (National Cancer Institute, 2022). The literature provides some evidence on the association between cancer and falls where individuals diagnosed with advanced cancer face a higher risk of falling and a potential for sustaining injuries due to falls (Goodridge & Marr, 2002; Healy & Scobie, 2007; Pautex et al., 2008; Pearse et al., 2004). Frith et al. (2012) revealed a high prevalence of falls (47%) among older individuals diagnosed with chronic liver disease, and Frith et al. (2010) discovered that falls and associated injuries were widespread in individuals with primary biliary cirrhosis, an autoimmune liver disease. Perhaps, it is worthwhile to investigate the relationship

between falls and injurious falls in individuals with cirrhosis as well. Cirrhosis is a condition characterized by liver scarring and dysfunction as the disease progresses (Slivinski, 2022). Murphy et al. (2019) delved into the underlying factors contributing to falls within this specific population. Their study emphasized the necessity of comprehensively understanding why individuals with cirrhosis face an elevated risk of falling and introduces a mechanistic model for assessing, treating, and researching falls in this context. The authors' research unveiled several critical insights. They underscored the multifaceted nature of falls in cirrhotic patients, attributing these incidents to a range of factors. IMPORTANT These include muscle weakness, which can stem from complications related to liver disease and a lack of physical conditioning; impaired balance, which may be influenced by changes in proprioception and neuromuscular function; and altered mental status, potentially arising from conditions like hepatic encephalopathy or side effects of medications. IMPORTANT The article also highlighted the prevalence of orthostatic hypotension among cirrhotic patients, which further contributes to fall risk during position changes, such as moving from a seated or lying position to standing. This article underscores the critical importance of promptly identifying and evaluating fall risk in individuals with cirrhosis and offers potential strategies for mitigating the occurrence of falls in this vulnerable population. Further research is required to improve our understanding of the link between FRIs and cancers of the digestive system, liver, and bile duct.

Investigating the symptoms of liver and bile duct diseases can provide insights into the potential reasons for their association with FRIs. The frequent symptoms of this diseases are peripheral neuropathy (Knill-Jones et al., 1972), jaundice (Wittig et al., 1978), and malnutrition (Purnak & Yilmaz, 2013). Certain liver diseases can cause peripheral neuropathy, resulting in numbness and decreased sensation in the feet, which might increase the risk of tripping. Jaundice can affect vision, leading to decreased visual acuity and depth perception, which also might contribute to falls. Malnutrition and vitamin deficiencies associated with liver diseases can weaken bones and muscles, making individuals more susceptible to falls and injuries. What is more, medications, such as interferon, used to manage liver diseases can have side effects such as dizziness and impaired coordination (Dusheiko, 1997). Acknowledging the current state of knowledge about the association between advanced stages of liver disease and FRIs, additional research is required.

The four diagnostic categories that were mutual among the findings of EDs and hospitals were *Category 9,* diseases of the circulatory system, *Category 13*, diseases of the musculoskeletal system and connective tissue, *Category 14,* diseases of the genitourinary system, and *Category10,* diseases of the respiratory system. The following is an overview of what is known about this association from previous research. Hypertension, low bone density due to osteoporosis, and urinary incontinence which fall into these disease categories are prevalent conditions in older adults (Buford, 2016; Jackson et al., 2004; Shaw & Wagg, 2017; Silver & Einhorn, 1995). Impaired physical and mental function in older adults with hypertension is linked to a higher risk of falls (Chu et al., 2015). Mitchell et al. (2013) found that 25% of the participants who reported falling within the last year, had several common factors that included being 85 years or older, having cataracts, having musculoskeletal and connective tissue disorders, having major circulatory, respiratory, and nervous system diseases, using four or more medications, relying on mobility aids, and being overweight. Furthermore, the study revealed that individuals aged 85 years or older, those with circulatory diseases, individuals using four or more medications, and those dependent on mobility aids were particularly prone to experiencing multiple falls.

The association between diseases of musculoskeletal systems, such as osteoporosis and fall, is well documented in the literature. Quigley et al. (2007) found that one of the most effective interventions for reducing FRIs is the reduction of fracture risks. This risk reduction can be achieved through primary prevention of osteoporosis. Berk et al. (2019) found that as individuals age, osteoporosis can increase the likelihood of balance problems and falls, which significantly diminishes their quality of life. They also discovered a correlation between the reduction in bone mineral density and balance issues. Osteoporosis not only decreases bone mineral density but also heightens the risk of fractures by increasing the likelihood of falling. Women aged 60 or older with postmenopausal osteoporosis have a higher likelihood of experiencing one or more falls within the year compared to women without osteoporosis, and they face an increased risk of recurrent falls (Beserra Da Silva et al., 2010).

Urinary incontinence emerged as one of the most significant predictors for fall-related risk factors in a study aimed at developing a fall-risk model for older adults (Tromp et al., 2001). Another study confirmed that urge urinary incontinence was linked to a rise in falls in older

adults, prompting researchers to recommend that this association needs to be considered when developing fall prevention programs (Chiarelli et al., 2009).

Chronic obstructive pulmonary disease and asthma were identified as important risk factors for falls and FRIs among middle-aged adults with kidney disease (Kistler et al., 2019). A study in India found that a majority of older adult patients who had experienced falls also had asthma (Mane et al., 2014), while a study on the frequency of dizziness in general population reported that individuals with chronic bronchitis were considerably more likely to experience feelings of dizziness (Tamber & Bruusgaard, 2009). Another explanation for why people with respiratory diseases are more prone to falling and injurious falls could be reduced lung function (Martinez-Pitre et al., 2022), shortness of breath (Mayo Clinic, 2020), medication side effects such as antihistamines (Harvard health publishing, 2021; Meltzer et al., 2010), muscle weakness (Wüst & Degens, 2007), and increased vulnerability to infections (American lung association, 2022). These conditions might lead to fatigue, weakness, impaired balance, and reduced physical stamina, all of which might be risks for falling.

The overlaps in diagnostic categories associated with FRIs in ED and hospital level of care, reported in this study, could potentially be attributed to common mechanisms of falls, the higher prevalence of certain health conditions in older adults, and the organizational structure of the ICD-10-CA coding system. Falls can lead to injuries affecting the circulatory system, musculoskeletal system, genitourinary system, and respiratory system, which are relevant across both levels of care. Older adults' increased susceptibility to certain health conditions and comorbidities further contributes to these similarities. The broad categorization of diagnoses in the coding system allows for overlap in diagnostic categories based on clinical presentation or anatomical location. This emphasizes the importance of considering specific diagnostic categories when managing FRIs in older adults across different healthcare settings.

The differences in findings between EDs and hospitals could be attributed to several factors. One possible explanation is that the severity and complexity of FRIs treated in hospitals may be higher than those in EDs. Hospitalized patients are likely to have more advanced or pre-existing health conditions, such as cancer, which can contribute to FRIs in different ways. Another factor could be the differences in the patient population served by EDs and hospitals.

EDs often see a broader range of patients, including those who may not require hospitalization, while hospitals treat patients with more severe conditions requiring specialized care. These different patient populations could lead to variations in the types of FRIs and their underlying causes. It is important to highlight the distinct patterns observed in the two settings and emphasize the significance of understanding these differences. This difference in findings suggests that interventions and preventive strategies targeting FRIs should be tailored to the specific care setting. Additionally, the identification of different informative diagnostic code categories can help healthcare providers prioritize interventions and allocate resources more effectively. For example, in the ED, focusing on symptoms and abnormal clinical findings not classified elsewhere might be crucial for early detection and prompt management of FRIs. In hospitals, considering the association with neoplasms, especially secondary malignant neoplasms of the liver and intrahepatic bile duct, can aid in identifying high-risk patients and implementing appropriate preventive measures.

The findings of this study can significantly contribute to researchers' understanding of future falls and FRIs and improve preventive guidelines for falls. For instance, the latest "World Guidelines for Falls Prevention and Management for Older Adults: A Global Initiative" by Montero-Odasso and colleagues (2022) recommends important steps for clinicians and physicians when dealing with older adults who experience falls or related injuries. Firstly, the guidelines suggest that clinicians inquire about the details of the event and its consequences, previous falls, any episodes of transient loss of consciousness or dizziness, and any existing impairments of mobility or concerns about falling that might limit their usual activities. Furthermore, the fall-prevention model proposed in the guidelines emphasizes the importance of opportunistic case findings. It advises clinicians and physicians to assess the individual's gait and balance. If gait and balance impairments are identified, they should recommend exercises aimed at improving balance and gait to prevent future falls and related injuries. Additionally, for older adults who have sustained a hip fracture, the guidelines recommend offering an individualized and progressive exercise program focused on enhancing mobility (such as standing up, balance, walking, and climbing stairs) as a fall prevention strategy. While these methods and exercise interventions for falls and related injuries are crucial, they may not address all aspects of the issue. Therefore, combining these interventions with extending the knowledge on the association

between different diagnoses, disease control, and symptom management through collaborative efforts among healthcare professionals could yield more comprehensive and effective results. To create a higher-quality and holistic guideline for fall and injury prevention, researchers should focus more on injury investigation and explore the importance and relationships of ICD-10 diagnostic categories. This approach will lead to a better understanding of the mechanisms behind falls and related injuries, ultimately facilitating the development of stronger and more efficient guidelines.

This study provided new insights and contributed to the body of knowledge on the use of machine learning in gerontology, reviewed in chapter 1, specifically by deploying ICD-10-CA codes. Previous research mainly focused on demographic features, such as age and sex (Ateeq, 2018) or socio-economic variables, such as income and marital status (Speiser et al., 2021). This study explored ICD-10-CA codes, making the results more reproducible. The use of different factors or predictive variables by different studies makes it difficult to compare the most important variables across different studies. For example, variables such as race and income level could have different meanings in different contexts while ICD-10-CA diagnostic codes are standardized variables that have an objective meaning regardless of the context of the study.

Regarding algorithms, while Ateeq (2018) decided to use and compare logistic regression and random forest, the current study relied on decision tree and its ensemble counterparts (i.e., random forest and XGBoost) to make the comparison more objective. This is because the base learner for XGBoost and random forest is the decision tree. Speiser et al.(2021) utilized decision tree and random forest without offering an explanation for the reason behind superior performance of decision tree over random forest. Furthermore, their study lacked sufficient details on data cleaning and data characteristics, both of which have been comprehensively addressed in this research.

When comparing the performance of different algorithms in Tables 3-19, it is observed that the XGBoost, which is known as one of the ensemble learning models with the best performance (Hastie et al., 2009), did not show highest accuracy and sensitivity in this study. This has been observed in previous literature, where in some cases even the base learner outperformed the ensemble algorithm (Smith et al., 2013; Speiser et al., 2021).This is usually

observed when predictive variables are categorical rather than numeric, which results in developing trees that are skewed (e.g., not symmetric) and thus less accurate (Kumar, 2020). This is because discrete variables provide fewer number of options for splitting (two in the case of a binary variable) which leads to having skewed and sparse decision trees (Ravi, 2019). More specifically, when comparing the sensitivity of XGBoost and random forest, it is observed that the XGBoost has a considerably lower sensitivity in DAD dataset compared to its other counterparts. This can be attributed to the characteristics of the data, with DAD having a larger ratio of variable to observation when compared to NACRS (with DAD having 25 diagnostic variables and 688,868 observations while NACRS having only 10 diagnostic variables against 1,248,029 observations). It is reported that XGBoost can have difficulty finding patterns in high-dimensional datasets (Hastie et al., 2009), a phenomenon known as curse of dimensionality (Das et al., 2023), as it may struggle to generalize effectively, leading to lower sensitivity. On the other hand, its simpler counterparts (i.e., decision trees and random forest) can behave more robustly when mining high-dimensional data since they do not rely heavily on variable interactions (e.g., decision tree simply splits data based on single variables (Provost & Fawcett, 2013)). Furthermore, the relatively higher precision of the XGBoost can be associated to its more nonlinear nature and capability in effectively identifying important features and assigning higher weights to them during the boosting process. This ability can lead to a more precise model, as it concentrates on the most discriminative variables for positive label predictions (Hastie et al., 2009).

This study contributes new knowledge to the application of machine learning to exploration of FRIs in older adults. This is important because machine learning is becoming more popular in all disciplines, and application of machine learning in rehabilitation sciences field is expected to increase (Frontiers, 2023). The current study explained the data structure of the training and test set, data preparation procedure and the model training steps. The framework used in this study could be applied to comparable research questions or datasets in the field of gerontology.

4.1 Strengths and Limitations

This study has several strengths worth highlighting. First, the study utilized a large

population-based dataset with 1,248,029 observations in NACRS and 688,868 observations in DAD datasets, representative of the entire population of older adults who experienced FRIs and were admitted to EDs and hospitals and their controls over a period of nine years. The use of administrative databases eliminated recall bias and self-reporting bias frequent in prospective studies. Moreover, the use of both ED and hospital-level data provided a more comprehensive estimate over the two levels of care from the entry point into the healthcare system at ED to hospital discharge. Previous studies that have only examined FRIs hospitalizations have contributed limited evidence about the impact of FRIs on the healthcare systems (Alexander et al., 1992; Hartholt et al., 2011; Rau et al., 2014). Therefore, the use of ED data in conjunction with hospital-level data was needed to obtain a more accurate estimate of the total number of FRIs and their impact on a larger setting of healthcare systems.

Another strength of this study is inclusion of only a single observation per person to eliminate the possibility of double registration. Because including multiple observations from the same person can lead to biased estimates of model performance. For example, if one individual has experienced multiple FRIs, their repeated observations would disproportionately influence the model's learning and evaluation process, potentially leading to an overestimation of the model's performance. Furthermore, each observation in the dataset could have up to 10 (ED level) and up to 25 (hospital level) ICD-10-CA diagnosis codes. This enabled an evaluation of all recorded diagnoses at the same time, instead of focusing on a singular, initial, primary, or most responsible diagnosis. The method of utilizing multiple diagnosis codes as an approach to study of FRIs caries a potential to identify more effective treatments and injury prevention methods (Nilson et al., 2016; Watson & Mitchell, 2011)

Several study limitations also require acknowledgement. First, the analysis was conducted on secondary data obtained from administrative healthcare databases for the population of Canadian province of Ontario. The ICD-10-CA coding is routinely carried out by trained coders based on a manual that accompanies ICD-10-CA (CIHI, 2022). Hence, data collection was out of the author's control, precluding the exclusion of errors in the data collection and coding processes. This is a limitation acknowledged in many studies using machine learning (Intellspot, 2022). Second, the dataset was derived from two extensive healthcare databases, NACRS and DAD, by a trained IC/ES analyst following the specific dataset creation plan which

allowed for the data analyst's discretion in interpreting the request. Third, IC/ES implemented an age limit of 91 years, categorizing all individuals above this threshold as 91 years old. Fourth, included data does not fully reflect the extent of FRIs' impact on Ontario's healthcare system since the dataset did not encompass observations from individuals who: (1) received treatment for an FRI from their primary care physician or at a walk-in clinic, (2) received treatment outside the province, (3) failed to identify a fall as the cause of injury during ED admission, (4) passed away before admission to the ED, or (5) were directly admitted to a hospital without undergoing registration at the ED. It is recommended that future FRIs research incorporate these medical services. The dataset used in this study consisted only of population records for older adults from one Canadian province. This limits generalizability of the findings due to potential differences in demographic characteristics and healthcare protocols. Five, the nine-year observation window (2006-2015) is somewhat outdated, but it offers an advantageous alignment with the retirement age entry of the baby boomer generation, in North America, providing a baseline for future comparisons and trend analysis.

A brief reflection on the limitations of machine learning is also warranted. Machine learning is an effective tool in prediction, but its results should be interpreted with care as they do not always imply causality. Machine learning algorithms require a lot of data for training. While the culture of open-source software[3] and code has eliminated the issue of software requirement, in health research protection of personal health information (even when de-identified) and data security restrictions have created a less open-data culture, which remains a barrier to the implementation of machine learning.

This study offered limited socio-demographic information regarding FRI observations, as deeper analysis was beyond the study's scope. Future research could delve deeper into this aspect to gain a more comprehensive understanding of older adults experiencing FRIs. Exploring the socio-demographic characteristics in context of different diagnoses could provide a broader perspective on risk factors for injury.

Finally, the IC/ES platform has some limitations, especially for researchers working with

[3] SAS is not open-source or free. It is a proprietary software.

machine learning powered solutions. The secure environment did not allow the use of more sophisticated and newer machine learning algorithms and programing languages. Open-source and more versatile software options, such as Python, were not available, which limited inclusion of some algorithms and visualization of results and graphs. Software use was restricted to a basic version of R.

4.2 Implications for Future Research

It is recommended that future studies incorporate machine learning algorithms to effectively identify the most informative risk factors for FRIs in older adults. Machine learning and artificial intelligence are new and developing fields. During completion of this study many breakthroughs were made in the field such as text to video artificial intelligence platform or artificial intelligence based automated employment decision tool (Press, 2022). Perhaps the most famous example is the introduction of more powerful language models, such as ChatGPT based on GPT (Biswas, 2023; Hill-Yardin et al., 2023). This study was limited to SQL binary data and several algorithms. Emerging advanced algorithms and data structures should be investigated in future FRIs research. Given that the investigation of ICD-10-CA codes in association with FRIs is a relatively new area of inquiry, there is a need for continuing exploration. For example, additional research is necessary to examine the subcategories of all 21 categories, to obtain a more comprehensive understanding of the role of different diagnoses might play in occurrences of FRIs.

Chapter 5

5. Conclusion

FRIs represent a significant challenge for older adults and impose a substantial burden on the healthcare systems. To address this issue, a population-based retrospective study was conducted, analyzing medical data from 631,339 older adults who were admitted to the EDs and 304,495 older adults who were hospitalized after FRIs. Although a considerable body of research has focused on the prediction of falls and FRIs, no peer-reviewed literature could be found on the use of machine learning algorithms for associating FRIs with ICD-10-CA diagnostic codes. The objective of this thesis was to investigate the most informative categories of ICD-10-CA diagnostic codes associated with FRIs, compare the differences in the most informative categories between EDs and hospitals, and determine the most accurate and sensitive machine learning model for establishing associations between ICD-10-CA codes and FRIs.

The present study demonstrated that dyspnea and secondary malignant neoplasm of liver and intrahepatic bile duct are the most informative ICD-10-CA code and disease associated with FRIs on ED and hospital level of care. There was a notable difference between the primary informative diagnostic categories for FRIs in these two settings. Methodologically, the random forest models exhibited the highest accuracy and sensitivity, outperforming the decision trees and the XGBoost models in both datasets . Further research is necessary to examine the subcategories of all 21 diagnostic categories and to include numerical and other informative variables, such as the prescribed medications or different age groups, for a more comprehensive understanding of the topic. In conclusion, this study provided evidence that machine learning models are capable of handling larger datasets, generating visualizations, and delivering a convincing performance with accuracy and sensitivity higher than 60%.

References

Alexander, B. H., Rivara, F. P., & Wolf, M. E. (1992). The cost and frequency of hospitalization for fall-related injuries in older adults. *American Journal of Public Health*, *82*(7), 1020–1023.

Ali, M. R., Myers, T., Wagner, E., Ratnu, H., Dorsey, E. R., & Hoque, E. (2021). Facial expressions can detect Parkinson's disease: preliminary evidence from videos collected online. *Npj Digital Medicine*, *4*(1), 1–4. https://doi.org/10.1038/s41746-021-00502-8

Ambrose, A. F., Paul, G., & Hausdorff, J. M. (2013). Risk factors for falls among older adults: A review of the literature. *Maturitas*, *75*, 51–61. https://doi.org/10.1016/j.maturitas.2013.02.009

American lung association. (2022). *Who Is at Risk? | American Lung Association*. https://www.lung.org/clean-air/outdoors/who-is-at-risk

Aoyagi, K., Ross, P. D., Davis, J. W., Wasnich, R. D., Hayashi, T., & Takemoto, T. (1998). Falls among community-dwelling elderly in Japan. *Journal of Bone and Mineral Research*, *13*(9), 1468–1474.

Ateeq, S. (2018). *Machine Learning Approach on Evaluating Predictive Factors of Fall-Related Injuries*.

Awais, M., Chiari, L., Ihlen, E. A. F., Helbostad, J. L., & Palmerini, L. (2021). Classical machine learning versus deep learning for the older adults free-living activity classification. *Sensors*, *21*(14), 1–13. https://doi.org/10.3390/s21144669

Badgujar, S., & Pillai, A. S. (2020). Fall Detection for Elderly People using Machine Learning. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–4. https://doi.org/10.1109/ICCCNT49239.2020.9225494

Balash, Y., Peretz, C. H., Leibovich, G., Herman, T., Hausdorff, J. M., & Giladi, N. (2005). Falls in outpatients with Parkinson's disease: frequency, impact and identifying factors. *Journal of Neurology, 252*, 1310–1315.

Barbour, K. E., Sagawa, N., Boudreau, R. M., Winger, M. E., Cauley, J. A., Nevitt, M. C., Fujii, T., Patel, K. V, & Strotmeyer, E. S. (2019). Knee osteoarthritis and the risk of medically treated injurious falls among older adults: a community-based US cohort study. *Arthritis Care & Research, 71*(7), 865–874.

Beauchamp, M. K., Hill, K., Goldstein, R. S., Janaudis-Ferreira, T., & Brooks, D. (2009). Impairments in balance discriminate fallers from non-fallers in COPD. *Respiratory Medicine, 103*(12), 1885–1891. https://doi.org/10.1016/J.RMED.2009.06.008

Beauchamp, M. K., Sibley, K. M., Lakhani, B., Romano, J., Mathur, S., Goldstein, R. S., & Brooks, D. (2012). Impairments in Systems Underlying Control of Balance in COPD. *Chest, 141*(6), 1496–1503. https://doi.org/10.1378/CHEST.11-1708

Bergland, A., & Wyller, T. B. (2004). Risk factors for serious fall related injury in elderly women living at home. *Injury Prevention, 10*(5), 308–313.

Berk, E., Koca, T. T., Güzelsoy, S. S., Nacitarhan, V., & Demirel, A. (2019). Evaluation of the relationship between osteoporosis, balance, fall risk, and audiological parameters. *Clinical Rheumatology, 38*, 3261–3268.

Beserra Da Silva, R., Costa-Paiva, L., Siani Morais, S., Mezzalira, R., Oliveira Ferreira, N. De, & Mendes Pinto-Neto, A. (2010). Predictors of falls in women with and without osteoporosis. *Journal of Orthopaedic & Sports Physical Therapy*, *40*(9), 582–588.

Betts, K. S., Kisely, S., & Alati, R. (2019). Predicting common maternal postpartum complications: leveraging health administrative data and machine learning. *BJOG: An International Journal of Obstetrics & Gynaecology*, *126*(6), 702–709.

Biswas, S. S. (2023). Role of chat gpt in public health. *Annals of Biomedical Engineering, 51*(5), 868–869.

Bloem, B. R., Grimbergen, Y. A. M., Cramer, M., Willemsen, M., & Zwinderman, A. H. (2001). Prospective assessment of falls in Parkinson's disease. *Journal of Neurology, 248*, 950–958.

Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. CRC press.

Bolourani, S., Thompson, D., Siskind, S., Kalyon, B. D., Patel, V. M., & Mussa, F. F. (2021). Cleaning up the MESS: can machine learning be used to predict lower extremity amputation after trauma-associated arterial injury? *Journal of the American College of Surgeons, 232*(1), 102–113.

Bone, R., Boyars, M., Braun, S. R., Buist, A. S., Campbell, S., Chick, T., Cohen, B. M., Conway, W., Cugell, D. W., DeGraff, A., Friedman, M., George, R. B., Gershwin, E., Kram, J. A., Levin, D. C., Levine, B., Petty, T. L., Rennard, S., & Repsher, L. (1994). In Chronic Obstructive Pulmonary Disease, a Combination of Ipratropium and Albuterol Is More Effective Than Either Agent Alone: An 85-Day Multicenter Trial. *Chest*, *105*(5), 1411–1419. https://doi.org/10.1378/CHEST.105.5.1411

Bongue, B., Dupré, C., Beauchet, O., Rossat, A., Fantino, B., & Colvez, A. (2011). A screening tool with five risk factors was developed for fall-risk prediction in community-dwelling elderly. *Journal of Clinical Epidemiology*, *64*(10), 1152–1160. https://doi.org/10.1016/J.JCLINEPI.2010.12.014

Borges, S. de M., Radanovic, M., & Forlenza, O. V. (2015). Fear of falling and falls in older adults with mild cognitive impairment and Alzheimer's disease. *Aging, Neuropsychology, and Cognition*, *22*(3), 312–321.

Buford, T. W. (2016). Hypertension and aging. *Ageing Research Reviews*, *26*, 96–111.

Caramori, G., Ruggeri, P., Arpinelli, F., Salvi, L., & Girbino, G. (2019). Long-term use of inhaled glucocorticoids in patients with stable chronic obstructive pulmonary disease and risk of bone fractures: a narrative review of the literature. *International Journal of Chronic Obstructive Pulmonary Disease, 14*, 1085. https://doi.org/10.2147/COPD.S190215

CDC. (2022). *Keep on Your Feet—Preventing Older Adult Falls | CDC*. Centers for Disease Control and Prevention. https://www.cdc.gov/injury/features/older-adult-falls/index.html

Centers for Medicare & Medicaid Services. (2023). *Centers for Medicare & Medicaid Services*. https://www.cms.gov/Medicare/Coding/ICD10

Chan, V., Zagorski, B., Parsons, D., & Colantonio, A. (2013). Older adults with acquired brain injury: A population based study. *BMC Geriatrics*, *13*(1). https://doi.org/10.1186/1471-2318-13-97

Chiarelli, P. E., Mackenzie, L. A., & Osmotherly, P. G. (2009). Urinary incontinence is associated with an increase in falls: a systematic review. *Australian Journal of Physiotherapy*, *55*(2), 89–95.

Choi, Lee, W., Yoon, J.-H., Won, J.-U., & Kim, D. W. (2018). Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. *Journal of Affective Disorders*, *231*, 8–14.

Choi, N. G., Choi, B. Y., DiNitto, D. M., Marti, C. N., & Kunik, M. E. (2019). Fall-related emergency department visits and hospitalizations among community-dwelling older adults: examination of health problems and injury characteristics. *BMC Geriatrics*, *19*(1), 1–10.

Chu, J.-J., Chen, X.-J., Shen, S.-S., Zhang, X.-F., Chen, L.-Y., Zhang, J.-M., He, J., & Zhao, J.-F. (2015). A poor performance in comprehensive geriatric assessment is associated with increased fall risk in elders with hypertension: a cross-sectional study. *Journal of Geriatric Cardiology: JGC*, *12*(2), 113.

CIHI. (2011). *Data Quality Documentation, Discharge Abstract Database*.

CIHI. (2022). International statistical classification of diseases and related health problems ; 2022. *Geneva: World Health Organization*, *10*.

Colombo, P. J., Crawley, M. E., East, B. S., & Hill, A. R. (2012). Aging and the Brain. In *Encyclopedia of Human Behavior: Second Edition* (pp. 53–59). https://doi.org/10.1016/B978-0-12-375000-6.00006-9

Cowling, T. E., Cromwell, D. A., Bellot, A., Sharples, L. D., & van der Meulen, J. (2021). Logistic regression and machine learning predicted patient mortality from large sets of diagnosis codes comparably. *Journal of Clinical Epidemiology*, *133*, 43–52.

Cuaya-Simbro, G., Perez-Sanpablo, A. I., Morales, E. F., Quiñones Uriostegui, I., & Nuñez-Carrera, L. (2021). Comparing Machine Learning Methods to Improve Fall Risk Detection in Elderly with Osteoporosis from Balance Data. *Journal of Healthcare Engineering*, *2021*. https://doi.org/10.1155/2021/8697805

Das, S., Sultana, M., Bhattacharya, S., Sengupta, D., & De, D. (2023). XAI–reduct: accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI. *Journal of Supercomputing*. https://doi.org/10.1007/s11227-023-05356-3

Deandrea, S., Lucenteforte, E., Bravi, F., Foschi, R., La Vecchia, C., & Negri, E. (2010). Risk factors for falls in community-dwelling older people:" a systematic review and meta-analysis". *Epidemiology*, 658–668.

Deschepper, M., Eeckloo, K., Vogelaers, D., & Waegeman, W. (2019). A hospital wide predictive model for unplanned readmission using hierarchical ICD data. *Computer Methods and Programs in Biomedicine*, *173*, 177–183.

Di Martino, F., Delmastro, F., & Dolciotti, C. (2021). Malnutrition Risk Assessment in Frail Older Adults using m-Health and Machine Learning. *IEEE International Conference on Communications*, 1–6. https://doi.org/10.1109/ICC42927.2021.9500471

Dusheiko, G. (1997). Side effects of α interferon in chronic hepatitis C. *Hepatology*, *26*(S3), 112S-121S.

Edgcomb, J. B., Shaddox, T., Hellemann, G., & Brooks III, J. O. (2021). Predicting suicidal behavior and self-harm after general hospitalization of adults with serious mental illness. *Journal of Psychiatric Research*, *136*, 515–521.

Edgcomb, J. B., Thiruvalluru, R., Pathak, J., & Brooks III, J. O. (2021). Using machine learning to differentiate risk of suicide attempt and self-harm after general medical hospitalization of women with mental illness. *Medical Care, 59*, S58.

Ergen, E., Guven, G., Kurc, O., Erberik, M. A., Ergin, T., Birgonul, M. T., & Akinci, B. (2015). Blockage assessment of buildings during emergency using multiple types of sensors. *Automation in Construction*, *49*, 71–82.

Erkal, S. (2010). Home Safety, Safe Behaviors of Elderly People, and Fall Accidents at Home. *Educational Gerontology*, *36*(12), 1051–1064. https://doi.org/10.1080/03601277.2010.482482

Florence, C. S., Bergen, G., Atherly, A., Burns, E., Stevens, J., & Drake, C. (2018). Medical Costs of Fatal and Nonfatal Falls in Older Adults. *Journal of the American Geriatrics Society*, *66*(4), 693–698. https://doi.org/10.1111/JGS.15304

Fricke, C., Alizadeh, J., Zakhary, N., Woost, T. B., Bogdan, M., & Classen, J. (2021). Evaluation of Three Machine Learning Algorithms for the Automatic Classification of EMG Patterns in Gait Disorders. *Frontiers in Neurology*, *12*(May), 1–11. https://doi.org/10.3389/fneur.2021.666458

Frith, J., Kerr, S., Robinson, L., Elliott, C., Ghazala, C., Wilton, K., Pairman, J., Jones, D. E. J., & Newton, J. L. (2010). Primary biliary cirrhosis is associated with falls and significant fall related injury. *QJM: An International Journal of Medicine*, *103*(3), 153–161.

Frith, J., Kerr, S., Robinson, L., Elliott, C. S., Wilton, K., Jones, D. E. J., Day, C. P., & Newton, J. L. (2012). Falls and fall-related injury are common in older people with chronic liver disease. *Digestive Diseases and Sciences*, *57*, 2697–2702.

Frontiers. (2023). *Machine Learning in Rehabilitation Sciences | Frontiers Research Topic*. https://www.frontiersin.org/research-topics/51675/machine-learning-in-rehabilitation-sciences

Galet, C., Zhou, Y., Eyck, P. Ten, & Romanowski, K. S. (2018). Fall injuries, associated deaths, and 30-day readmission for subsequent falls are increasing in the elderly US population: a query of the WHO mortality database and National Readmission Database from 2010 to 2014. *Clinical Epidemiology*, 1627–1637.

Geithner, C. A., & McKenney, D. R. (2010). Strategies for Aging Well. *Strength & Conditioning Journal*, *32*(5). https://journals.lww.com/nsca-scj/Fulltext/2010/10000/Strategies_for_Aging_Well.4.aspx

Gillespie, L. D., Robertson, M. C., Gillespie, W. J., Sherrington, C., Gates, S., Clemson, L. M., & Lamb, S. E. (2012). Interventions for preventing falls in older people living in the community. *Cochrane Database of Systematic Reviews*, *2012*(9). https://doi.org/10.1002/14651858.CD007146.PUB3/INFORMATION/EN

Gimigliano, F. (2020). Are interventions for preventing falls in older people in care facilities and hospitals effective? A Cochrane Review summary with commentary. *International Journal of Rheumatic Diseases*, *23*(6), 833–836.

Goodridge, D., & Marr, H. (2002). Factors associated with falls in an inpatient palliative care unit: an exploratory study. *International Journal of Palliative Nursing*, *8*(11), 548–556.

Government of Ontario Ministry for Seniors and Accessibility. (2017). *AGING WITH CONFIDENCE: Ontario's Action Plan for Seniors*.

Greene, B. R., McManus, K., Ader, L. G. M., & Caulfield, B. (2021). Article unsupervised assessment of balance and falls risk using a smartphone and machine learning. *Sensors*, *21*(14). https://doi.org/10.3390/s21144770

Harrison, J. E., Weber, S., Jakob, R., & Chute, C. G. (2021). ICD-11: an international classification of diseases for the twenty-first century. In *BMC Medical Informatics and Decision Making* (Vol. 21). BioMed Central Ltd. https://doi.org/10.1186/s12911-021-01534-6

Hartholt, K. A., Stevens, J. A., Polinder, S., van der Cammen, T. J. M., & Patka, P. (2011). Increase in fall-related hospitalizations in the United States, 2001–2008. *Journal of Trauma and Acute Care Surgery, 71*(1), 255–258.

Harvard health publishing. (2021). *Medications that increase your risk of falling - Harvard Health*. https://www.health.harvard.edu/staying-healthy/medications-that-increase-your-risk-of-falling

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. https://doi.org/10.1007/b94608

Healy, F., & Scobie, S. (2007). *Slips, trips and falls in hospital. Third report from the Patient Safety Observatory*. London: National Patient Safety Agency.

Heo, J. N., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. *Stroke*, *50*(5), 1263–1265. https://doi.org/10.1161/STROKEAHA.118.024293

Hill-Yardin, E. L., Hutchinson, M. R., Laycock, R., & Spencer, S. J. (2023). A Chat (GPT) about the future of scientific publishing. *Brain Behav Immun, 110*, 152–154.

Hogan, H., Perez, D., & Bell, W. (2008). *Who (Really) are the first baby boomers*.

Huda, A., Castaño, A., Niyogi, A., Schumacher, J., Stewart, M., Bruno, M., Hu, M., Ahmad, F. S., Deo, R. C., & Shah, S. J. (2021). A machine learning model for identifying patients at risk for wild-type transthyretin amyloid cardiomyopathy. *Nature Communications, 12*(1), 2725.

Intellspot. (2022). *Secondary Data: Advantages, Disadvantages, Sources, Types*. https://www.intellspot.com/secondary-data/

Iron, K., & Sykora, K. (2015). Health services data, sources and examples: the Institute for clinical evaluative sciences data repository. *Data and Measures in Health Services Research. Boston, MA: Springer US*, 1–13.

Jackson, R. A., Vittinghoff, E., Kanaya, A. M., Miles, T. P., Resnick, H. E., Kritchevsky, S. B., Simonsick, E. M., Brown, J. S., & Health and Body Composition Study, A. (2004). Urinary incontinence in elderly women: findings from the Health, Aging, and Body Composition Study. *Obstetrics & Gynecology*, *104*(2), 301–307.

Jang, H., Soroski, T., Rizzo, M., Barral, O., Harisinghani, A., Newton-Mason, S., Granby, S., Stutz da Cunha Vasco, T. M., Lewis, C., Tutt, P., Carenini, G., Conati, C., & Field, T. S. (2021). Classification of Alzheimer's Disease Leveraging Multi-task Machine Learning Analysis of Speech and Eye-Movement Data. *Frontiers in Human Neuroscience*, *15*(September), 1–15. https://doi.org/10.3389/fnhum.2021.716670

Jehu, D. A., Davis, J. C., Falck, R. S., Bennett, K. J., Tai, D., Souza, M. F., Cavalcante, B. R., Zhao, M., & Liu-Ambrose, T. (2021). Risk factors for recurrent falls in older adults: A systematic review with meta-analysis. *Maturitas*, *144*, 23–28.

Kalatzis, A., Stanley, L., Karthikeyan, R., & Mehta, R. K. (2020). Mental stress classification during a motor task in older adults using an artificial neural network. *UbiComp/ISWC 2020 Adjunct - Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, 244–248. https://doi.org/10.1145/3410530.3414360

Källstrand-Ericson, J., & Hildingh, C. (2009). Visual impairment and falls: a register study. *Journal of Clinical Nursing*, *18*(3), 366–372. https://doi.org/10.1111/j.1365-2702.2008.02516.x

Kim, I., Won, S., Lee, M., & Lee, W. (2018). A risk-factor analysis of medical litigation judgments related to fall injuries in Korea. *Medicine, Science and the Law*, *58*(1), 16–24.

Kistler, B. M., Khubchandani, J., Jakubowicz, G., Wilund, K., & Sosnoff, J. (2018). Peer Reviewed: Falls and Fall-Related Injuries Among US Adults Aged 65 or Older With Chronic Kidney Disease. *Preventing Chronic Disease*, *15*(6). https://doi.org/10.5888/PCD15.170518

Kistler, B. M., Khubchandani, J., Wiblishauser, M., Wilund, K. R., & Sosnoff, J. J. (2019). Epidemiology of falls and fall-related injuries among middle-aged adults with kidney disease. *International Urology and Nephrology*, *51*, 1613–1621.

Knill-Jones, R. P., Goodwill, C. J., Dayan, A. D., & Williams, R. (1972). Peripheral neuropathy in chronic liver disease: clinical, electrodiagnostic, and nerve biopsy findings. *Journal of Neurology, Neurosurgery & Psychiatry*, *35*(1), 22–30.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Kumar, S. (2020). *Does Ensemble Models Always Improve Accuracy?* Medium. https://medium.com/analytics-vidhya/does-ensemble-models-always-improve-accuracy-c114cdbdae77

Lagiewka, K. (2012). European innovation partnership on active and healthy ageing: triggers of setting the headline target of 2 additional healthy life years at birth at EU average by 2020. *Archives of Public Health*, *70*(1), 1–8.

Lamb, S. E., Jrstad-Stein, E. C., Hauer, K., & Becker, C. (2005). *Development of a Common Outcome Data Set for Fall InjuryPrevention Trials: The Prevention of Falls Network EuropeConsensus*. https://doi.org/10.1111/j.1532-5415.2005.53455.x

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

Lappan, N. (2021). *Descriptive Analysis of Fall-Related Injuries Among Older Adults in Descriptive Analysis of Fall-Related Injuries Among Older Adults in Ontario Ontario* [Western University]. https://ir.lib.uwo.ca/etdhttps://ir.lib.uwo.ca/etd/7910

Lawlor, D. A., Patel, R., & Ebrahim, S. (2003). Association between falls in elderly women and chronic diseases and drug use: cross sectional study. *Bmj*, *327*(7417), 712–717.

Lee, S., Kang, W. S., Seo, S., Kim, D. W., Ko, H., Kim, J., Lee, S., & Lee, J. (2022). Model for Predicting In-Hospital Mortality of Physical Trauma Patients Using Artificial Intelligence

Techniques: Nationwide Population-Based Study in Korea. *Journal of Medical Internet Research*, *24*(12), e43757.

Li. (2016). Alzheimer's disease increases the incidence of hospitalization due to fall-related bone fracture in elderly Chinese. *International Journal of Gerontology*, *10*(4), 227–231.

Li, I.-F., Hsiung, Y., Hsing, H.-F., Lee, M.-Y., Chang, T.-H., & Huang, M.-Y. (2016). Elderly Taiwanese's intrinsic risk factors for fall-related injuries. *International Journal of Gerontology*, *10*(3), 137–141.

Lord, S. R., Menz, H. B., & Sherrington, C. (2006). Home environment risk factors for falls in older people and the efficacy of home modifications. *Age and Ageing*, *35 Suppl 2*, ii55–ii59. https://doi.org/10.1093/ageing/afl088

Makino, K., Lee, S., Bae, S., Chiba, I., Harada, K., Katayama, O., Tomida, K., Morikawa, M., & Shimada, H. (2021). Simplified decision-tree algorithm to predict falls for community-dwelling older adults. *Journal of Clinical Medicine*, *10*(21). https://doi.org/10.3390/jcm10215184

Mane, A. B., Sanjana, T., Patil, P. R., & Sriniwas, T. (2014). Prevalence and correlates of fear of falling among elderly population in urban area of Karnataka, India. *Journal of Mid-Life Health*, *5*(3), 150.

Marchetti, G. F. (1994). *Risk factors for falls and fall-related injuries in older adults with signs and symptoms of vestibular dysfunction*.

Marciniuk, D. D., Goodridge, D., Hernandez, P., Rocker, G., Balter, M., Bailey, P., Ford, G., Bourbeau, J., O'Donnell, D. E., & Maltais, F. (2011). Managing dyspnea in patients with advanced chronic obstructive pulmonary disease: a Canadian Thoracic Society clinical practice guideline. *Canadian Respiratory Journal*, *18*, 69–78.

Martinez-Pitre, P. J., Sabbula, B. R., & Cascella, M. (2022). Restrictive Lung Disease. *Preoperative Assessment: A Case-Based Approach*, 101–106. https://doi.org/10.1007/978-3-030-58842-7_16

Mayo Clinic. (2020). *Shortness of breath Causes - Mayo Clinic*. https://www.mayoclinic.org/symptoms/shortness-of-breath/basics/causes/sym-20050890

McCann-Pineo, M., Ruskin, J., Rasul, R., Vortsman, E., Bevilacqua, K., Corley, S. S., & Schwartz, R. M. (2021). Predictors of emergency department opioid administration and prescribing: a machine learning approach. *The American Journal of Emergency Medicine*, *46*, 217–224.

McClure, R. J., Turner, C., Peel, N., Spinks, A., Eakin, E., & Hughes, K. (2005). Population-based interventions for the prevention of fall-related injuries in older people. *Cochrane Database of Systematic Reviews*, *1*.

McMaster, C., Liew, D., Keith, C., Aminian, P., & Frauman, A. (2019). A machine-learning algorithm to optimise automated adverse drug reaction detection from clinical coding. *Drug Safety*, *42*, 721–725.

Meltzer, E. O., Caballero, F., Fromer, L. M., Krouse, J. H., & Scadding, G. (2010). Treatment of congestion in upper respiratory diseases. *International Journal of General Medicine*, 69–91.

Milat, A. J., Watson, W. L., Monger, C., Barr, M., Giffin, M., & Reid, M. (2011). Prevalence, circumstances and consequences of falls among community-dwelling older people: results of the 2009 NSW Falls Prevention Baseline Survey. *New South Wales Public Health Bulletin*, *22*(4), 43–48.

Ming, Y. (2020). *Population-based Studies on Medications and Fall-related Injury in Population-based Studies on Medications and Fall-related Injury in Older Adults Older Adults*. https://ir.lib.uwo.ca/etdhttps://ir.lib.uwo.ca/etd/7559

Mitchell, R. J., Watson, W. L., Milat, A., Chung, A. Z. Q., & Lord, S. (2013). Health and lifestyle risk factors for falls in a large population-based sample of older people in Australia. *Journal of Safety Research*, *45*, 7–13.

Mohile, S. G., Fan, L., Reeve, E., Jean-Pierre, P., Mustian, K., Peppone, L., Janelsins, M., Morrow, G., Hall, W., & Dale, W. (2011). Association of cancer with geriatric syndromes in older Medicare beneficiaries. *Journal of Clinical Oncology*, *29*(11), 1458.

Montero-Odasso, M., van der Velde, N., Martin, F. C., Petrovic, M., Tan, M. P., Ryg, J., Aguilar-Navarro, S., Alexander, N. B., Becker, C., & Blain, H. (2022). World guidelines for falls prevention and management for older adults: a global initiative. *Age and Ageing*, *51*(9), afac205.

Moreland, B., Kakara, R., & Henry, A. (2020). Trends in Nonfatal Falls and Fall-Related Injuries Among Adults Aged ≥65 Years — United States, 2012–2018. *MMWR. Morbidity and Mortality Weekly Report*, *69*(27), 875–881. https://doi.org/10.15585/MMWR.MM6927A5

Murphy, S. L., Tapper, E. B., Blackwood, J., & Richardson, J. K. (2019). Why Do Individuals with Cirrhosis Fall? A Mechanistic Model for Fall Assessment, Treatment, and Research. *Digestive Diseases and Sciences*, *64*(2), 316–323. https://doi.org/10.1007/S10620-018-5333-8/TABLES/1

Nachreiner, N. M., Findorff, M. J., Wyman, J. F., & McCarthy, T. C. (2007). Circumstances and consequences of falls in community-dwelling older women. *Journal of Women's Health*, *16*(10), 1437–1446.

National Cancer Institute. (2022). *What Is Bile Duct Cancer (Cholangiocarcinoma)? - NCI*. https://www.cancer.gov/types/liver/bile-duct-cancer

Nilson, F., Moniruzzaman, S., & Andersson, R. (2016). Hospitalized fall-related injury trends in Sweden between 2001 and 2010. *International Journal of Injury Control and Safety Promotion*, *23*(3), 277–283.

Nolan, D., & Lang, D. T. (2015). *Data science in R: A case studies approach to computational reasoning and problem solving*. CRC Press.

O'loughlin, J. L., Robitaille, Y., Boivin, J. F., & Suissa, S. (1993). Incidence of and Risk Factors for Falls and Injurious Falls among the Community-dwelling Elderly. *American Journal of*

Epidemiology, *137*(3), 342–354. https://doi.org/10.1093/OXFORDJOURNALS.AJE.A116681

Olsavszky, V., Dosius, M., Vladescu, C., & Benecke, J. (2020). Time series analysis and forecasting with automated machine learning on a national ICD-10 database. *International Journal of Environmental Research and Public Health*, *17*(14), 4979.

Ontario Ministry of Health and Long-Term Care. (2012). *Health Analyst's Toolkit*.

Opher, T., & Ostfeld, A. (2011). A coupled model tree (MT) genetic algorithm (GA) scheme for biofouling assessment in pipelines. *Water Research*, *45*(18), 6277–6288.

Osoba, M. Y., Rao, A. K., Agrawal, S. K., & Lalwani, A. K. (2019). Balance and gait in the elderly: A contemporary review. *Laryngoscope Investigative Otolaryngology*, *4*(1), 143–153.

Overcash, J. (2007). Prediction of falls in older adults with cancer: a preliminary study. *Number 2/March 2007*, *34*(2), 341–346.

Overcash, J. A., & Beckstead, J. (2008). Predicting falls in older patients using components of a comprehensive geriatric assessment. *Clinical Journal of Oncology Nursing*, *12*(6).

Ozalevli, S., Ilgin, D., Narin, S., & Akkoclu, A. (2011). Association between disease-related factors and balance and falls among the elderly with COPD: A cross-sectional study. *Aging Clinical and Experimental Research*, *23*(5–6), 372–377. https://doi.org/10.1007/BF03325235

Parachute. (2021). *The Cost of Injury in Canada*.

Park, J. H. (2020). Machine-Learning Algorithms Based on Screening Tests for Mild Cognitive Impairment. *American Journal of Alzheimer's Disease and Other Dementias*, *35*, 153331752092716. https://doi.org/10.1177/1533317520927163

Pautex, S., Herrmann, F. R., & Zulian, G. B. (2008). Factors associated with falls in patients with cancer hospitalized for palliative care. *Journal of Palliative Medicine*, *11*(6), 878–884.

Pearse, H., Nicholson, L., & Bennett, M. (2004). Falls in hospices: a cancer network observational study of fall rates and risk factors. *Palliative Medicine*, *18*(5), 478–481.

Pedroso, R. V., de Melo Coelho, F. G., Santos-Galduróz, R. F., Costa, J. L. R., Gobbi, S., & Stella, F. (2012). Balance, executive functions and falls in elderly with Alzheimer's disease (AD): a longitudinal study. *Archives of Gerontology and Geriatrics*, *54*(2), 348–351.

Peel, N. M. (2011). Epidemiology of falls in older age. *Canadian Journal on Aging/La Revue Canadienne Du Vieillissement*, *30*(1), 7–19.

Pengpid, S., & Peltzer, K. (2018). Prevalence and risk factors associated with injurious falls among community-dwelling older adults in Indonesia. *Current Gerontology and Geriatrics Research*, *2018*.

Peterson, E. W., Cho, C. C., von Koch, L., & Finlayson, M. L. (2008). Injurious falls among middle aged and older adults with multiple sclerosis. *Archives of Physical Medicine and Rehabilitation*, *89*(6), 1031–1037.

Pi, H.-Y., Gao, Y., Wang, J., Hu, M.-M., Nie, D., & Peng, P.-P. (2016). Risk factors for in-hospital complications of fall-related fractures among older Chinese: a retrospective study. *BioMed Research International*, *2016*.

Piryonesi, S. M., & El-Diraby, T. E. (2020a). Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index. *Journal of Infrastructure Systems*, *26*(1), 04019036. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000512

Piryonesi, S. M., & El-Diraby, T. E. (2020b). Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *Journal of Transportation Engineering, Part B: Pavements*, *146*(2), 04020022. https://doi.org/10.1061/JPEODX.0000175

Piryonesi, S. M., Rostampour, S., & Piryonesi, S. A. (2021). Predicting falls and injuries in people with multiple sclerosis using machine learning algorithms. *Multiple Sclerosis and Related Disorders*, *49*(4), 102740. https://doi.org/10.1016/J.MSARD.2021.102740

Prabhakaran, K., Gogna, S., Pee, S., Samson, D. J., Con, J., & Latifi, R. (2020). Falling again? Falls in geriatric adults—risk factors and outcomes associated with recidivism. *Journal of Surgical Research*, *247*, 66–76.

Press. (2022). *What Happened To AI In 2022?* Forbes. https://www.forbes.com/sites/gilpress/2022/12/30/what-happened-to-ai-in-2022/?sh=7e63310c64d1

Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media, Inc.

Purnak, T., & Yilmaz, Y. (2013). Liver disease and malnutrition. *Best Practice & Research Clinical Gastroenterology*, *27*(4), 619–629.

Quigley, P. A., Bulat, T., & Hart-Hughes, S. (2007). Strategies to reduce risk of fall-related injuries in rehabilitation nursing. *Rehabilitation Nursing*, *32*(3), 120–125.

Rau, C.-S., Lin, T.-S., Wu, S.-C., Yang, J. C.-S., Hsu, S.-Y., Cho, T.-Y., & Hsieh, C.-H. (2014). Geriatric hospitalizations in fall-related injuries. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *22*(1), 1–8.

Ravi, R. (2019). *One-Hot Encoding is making your Tree-Based Ensembles worse, here's why*. Towards Data Science. https://towardsdatascience. com/one-hot-encoding ….

Rodríguez, M. D., Beltrán, J., Valenzuela-Beltrán, M., Cruz-Sandoval, D., & Favela, J. (2021). Assisting older adults with medication reminders through an audio-based activity recognition system. *Personal and Ubiquitous Computing*, *25*(2), 337–351. https://doi.org/10.1007/s00779-020-01420-4

Roig, M., Eng, J. J., MacIntyre, D. L., Road, J. D., FitzGerald, J. M., Burns, J., & Reid, W. D. (2011). Falls in people with chronic obstructive pulmonary disease: An observational cohort study. *Respiratory Medicine*, *105*(3), 461–469. https://doi.org/10.1016/j.rmed.2010.08.015

Roo, S. Van, Johnston, T., Petersen, L., & Branch, H. S. (2015). *Readmission rates for fall-related injuries*. www.health.qld.gov.au

Rubenstein, L. Z. (2006). Falls in older people: epidemiology, risk factors and strategies for prevention. *Age and Ageing*, *35*(suppl_2), ii37–ii41.

Saeed, M. A., Hashem Almourish, M., Alqady, Y. A., Alsharabi, H., Alkhorasani, H., Alsorori, S., & Saeed, A. Y. A. (2021, July). Predicting fall in elderly people using machine learning. *2021 International Congress of Advanced Technology and Engineering, ICOTEN 2021*. https://doi.org/10.1109/ICOTEN52080.2021.9493442

Sander, M., Oxlund, B., Jespersen, A., Krasnik, A., Mortensen, E. L., Westendorp, R. G. J., & Rasmussen, L. J. (2015). The challenges of human population ageing. *Age and Ageing*, *44*(2), 185–187.

Scott, V., Peck, S., & Kendall, P. (2004). Prevention of falls and injuries among the elderly: A special report from the Office of the Provincial Health Officer. *Victoria: BC Ministry of Health Planning*.

Shah, A. A., Devana, S. K., Lee, C., Bugarin, A., Lord, E. L., Shamie, A. N., Park, D. Y., van der Schaar, M., & SooHoo, N. F. (2021). Prediction of major complications and readmission after lumbar spinal fusion: a machine learning–driven approach. *World Neurosurgery*, *152*, e227–e234.

Sharafkhaneh, A., Majid, H., & Gross, N. J. (2013). Safety and tolerability of inhalational anticholinergics in COPD. *Drug, Healthcare and Patient Safety*, *5*(1), 49. https://doi.org/10.2147/DHPS.S7771

Shaw, C., & Wagg, A. (2017). Urinary incontinence in older adults. *Medicine*, *45*(1), 23–27.

Shuto, H., Imakyure, O., Matsumoto, J., Egawa, T., Jiang, Y., Hirakawa, M., Kataoka, Y., & Yanagawa, T. (2010). Medication use as a risk factor for inpatient falls in an acute care hospital: a case-crossover study. *British Journal of Clinical Pharmacology*, *69*(5), 535–542.

Silver, J. J., & Einhorn, T. A. (1995). Osteoporosis and Aging: Current Update. *Clinical Orthopaedics and Related Research (1976-2007)*, *316*, 10–20.

Slivinski, N. (2022). *Cirrhosis: Symptoms, Causes, Stages, Diagnosis, and Treatment*. https://www.webmd.com/digestive-disorders/understanding-cirrhosis-basic-information

SMARTRISK. (2009). *The Economic Burden of Injury in Canada*.

Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods*, *220*(1), 85–91. https://doi.org/10.1016/J.JNEUMETH.2013.08.024

Speiser, J. L., Callahan, K. E., Houston, D. K., Fanning, J., Gill, T. M., Guralnik, J. M., Newman, A. B., Pahor, M., Rejeski, W. J., & Miller, M. E. (2021). Machine Learning in Aging: An Example of Developing Prediction Models for Serious Fall Injury in Older Adults. *The Journals of Gerontology: Series A*, *76*(4), 647–654. https://doi.org/10.1093/gerona/glaa138

Spoelstra, S. L. (2013). Do older adults with cancer fall more often? A comparative analysis of falls in those with and without cancer. *Number 2/March 2013*, *40*(2), E69–E78.

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, *10*(1), 1–10. https://doi.org/10.1038/s41598-020-77220-w

Statistics Canada. (2017). *Census profile, 2016 census*. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/search-recherche/lst/results-resultats.cfm?Lang=E&TABID=1&G=1&Geo1=&Code1=&Geo2=&Code2=&GEOCODE=35&type=0

Stevens, J. A., Mack, K. A., Paulozzi, L. J., & Ballesteros, M. F. (2008). Self-reported falls and fall-related injuries among persons aged≥ 65 years–United States, 2006. *Journal of Safety Research*, *39*(3), 345–349. https://doi.org/https://doi.org/10.1016/j.jsr.2008.05.002

Stone, C., Lawlor, P. G., Nolan, B., & Kenny, R. A. (2011). A prospective study of the incidence of falls in patients with advanced cancer. *Journal of Pain and Symptom Management*, *42*(4), 535–540.

Su, C., Aseltine, R., Doshi, R., Chen, K., Rogers, S. C., & Wang, F. (2020). Machine learning for suicide risk prediction in children and adolescents with electronic health records. *Translational Psychiatry*, *10*(1), 413.

Tamber, A.-L., & Bruusgaard, D. (2009). Self-reported faintness or dizziness—comorbidity and use of medicines. An epidemiological study. *Scandinavian Journal of Public Health*, *37*(6), 613–620.

Tarekegn, A., Ricceri, F., Costa, G., Ferracin, E., & Giacobini, M. (2020). Predictive modeling for frailty conditions in Elderly People: Machine learning approaches. *JMIR Medical Informatics*, *8*(6). https://doi.org/10.2196/16678

Tran, Z., Verma, A., Wurdeman, T., Burruss, S., Mukherjee, K., & Benharash, P. (2022). ICD-10 based machine learning models outperform the Trauma and Injury Severity Score (TRISS) in survival prediction. *Plos One*, *17*(10), e0276624.

Tran, Z., Zhang, W., Verma, A., Cook, A., Kim, D., Burruss, S., Ramezani, R., & Benharash, P. (2021). The Derivation of an ICD-10-based Trauma-related Mortality Model Utilizing Machine Learning. *The Journal of Trauma and Acute Care Surgery*.

Tromp, A. M., Pluijm, S. M. F., Smit, J. H., Deeg, D. J. H., Bouter, L. M., & Lips, P. (2001). Fall-risk screening test: a prospective study on predictors for falls in community-dwelling elderly. *Journal of Clinical Epidemiology*, *54*(8), 837–844.

Van Doorn, C., Gruber-Baldini, A. L., Zimmerman, S., Richard Hebel, J., Port, C. L., Baumgarten, M., Quinn, C. C., Taler, G., May, C., & Magaziner, J. (2003). Dementia as a risk factor for falls and fall injuries among nursing home residents. *Journal of the American Geriatrics Society*, *51*(9), 1213–1218.

Vogelmeier, C. F. (2014). Systemic steroids in COPD-the beauty and the beast. *Respiratory Research*, *15*(1). https://doi.org/10.1186/1465-9921-15-38

Watson, W. L., & Mitchell, R. (2011). Conflicting trends in fall-related injury hospitalisations among older people: variations by injury type. *Osteoporosis International*, *22*, 2623–2631.

Wedzicha, J. A., Banerji, D., Chapman, K. R., Vestbo, J., Roche, N., Ayers, R. T., Thach, C., Fogel, R., Patalano, F., & Vogelmeier, C. F. (2016). Indacaterol–Glycopyrronium versus Salmeterol–Fluticasone for COPD. *New England Journal of Medicine*, *374*(23), 2222–2234. https://doi.org/10.1056/NEJMOA1516385/SUPPL_FILE/NEJMOA1516385_DISCLOSURES.PDF

Weegar, R., & Sundström, K. (2020). Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations. *PloS One*, *15*(8), e0237911.

WHO. (2008). *WHO global report on falls prevention in older age*. World Health Organization; World Health Organization. https://extranet.who.int/agefriendlyworld/wp-content/uploads/2014/06/WHo-Global-report-on-falls-prevention-in-older-age.pdf

Wittig, J. H., Burns, R., & Longmire Jr, W. P. (1978). Jaundice associated with polycystic liver disease. *The American Journal of Surgery*, *136*(3), 383–386.

Wong, C. K., Chihuri, S. T., & Li, G. (2016). Risk of fall-related injury in people with lower limb amputations: A prospective cohort study. *Journal of Rehabilitation Medicine*, *48*(1), 80–85. https://doi.org/10.2340/16501977-2042

Wu, Q., Nasoz, F., Jung, J., Bhattarai, B., & Han, M. V. (2020). Machine Learning Approaches for Fracture Risk Assessment: A Comparative Analysis of Genomic and Phenotypic Data in 5130 Older Men. *Calcified Tissue International*, *107*(4), 353–361. https://doi.org/10.1007/s00223-020-00734-y

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10

algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

Wu, Y., Xiao, Y., & Ge, H. (2020). Fall detection monitoring system based on MEMS sensor. *Journal of Physics: Conference Series*, *1650*(2). https://doi.org/10.1088/1742-6596/1650/2/022037

Wüst, R. C. I., & Degens, H. (2007). Factors contributing to muscle wasting and dysfunction in COPD patients. *International Journal of Chronic Obstructive Pulmonary Disease*, *2*(3), 289–300.

Zaki, M. J., & Meira Jr, W. (2020). *Data mining and machine learning: Fundamental concepts and algorithms*. Cambridge University Press.

Zaninotto, P., Huang, Y.-T., Di Gessa, G., Abell, J., Lassale, C., & Steptoe, A. (2020). Polypharmacy is a risk factor for hospital admission due to a fall: evidence from the English Longitudinal Study of Ageing. *BMC Public Health*, *20*, 1–7.

Zecevic, A. A., Chesworth, B. M., Zaric, G. S., Huang, Q., Salmon, A., McAuslan, D., Welch, R., & Brunton, D. (2012). Estimating the cost of serious injurious falls in a Canadian acute care hospital. *Canadian Journal on Aging/La Revue Canadienne Du Vieillissement*, *31*(2), 139–147.

Appendices

Appendix A. Dataset Creation Plan for IC/ES

Project Initiation **This Section must be Completed Prior to Project Dataset(s) Creation**		
Project Title:	The association between selected health indicators, prescription medications and serious fall-related injuries in older adults (REVISED July 2019)	
Project TRIM number:		
Research Program:	DAS	
Site:	ICES Western	
Project Objectives:	*Insert Project Objectives as listed in the approved ICES Project PIA*	
	The objective of this study is to investigate the association between selected health indicators, prescription medications and serious fall-related injuries in older adults	
ICES Project PIA Initial Approval Date:	*The ICES Employee or agent who is responsible for creating the Project Dataset(s) is responsible for ensuring there is an approved ICES Project PIA and verifying the date of approval prior to creating the Project Dataset(s)*	
	2017-Jul-12	
Principal Investigator (PI):	Aleksandra Zecevic	
Check the applicable box if the PI is an ICES Student/Trainee	☐ ICES Student ☐ ICES Fellow ☐ ICES Post-Doctoral Trainee ☐ Visiting Scholar	
Responsible ICES Scientist:	*Name the Responsible ICES Scientist if the PI is not a Full Status ICES Scientist*	
Project Team Member(s) Responsible for Project Dataset Creation and/or Statistical Analysis and date joined (list all):	*All person(s) (ICES Analyst, Appointed Analyst, Analytic Epidemiologist, PI, and/or Student) responsible for creating the Project Dataset(s) and/or statistical analysis on the Research Analytics Environment (RAE) and the date they joined the project must be recorded*	
	Aleksandra Zecevic, Yu Ming	2017-Jul-12
	~~Tyson Schierholtz~~	~~2017-Dec-01~~
	Nicolette Lappan	2018-Nov-6
Other ICES Project Team Members and date joined (list all):	*All other Research Project Team Members (e.g., Research Administrative Assistants, Research Assistants, Project Managers, Epidemiologists) and the date they joined the project must be recorded*	
		yyyy-mon-dd
Confirmation that DCP is consistent with Project Objectives:	*The following individuals must confirm that the ICES Data provided for in this DCP is relevant (e.g., with respect to cohort, timeframe, and variables) and required to achieve the Project Objectives stated in the ICES Project PIA prior to initial Project Dataset creation: 1) PI; 2) Responsible ICES Scientist if the PI is not a Full Status ICES Scientist, or a second ICES Scientist or the Scientific Program Lead if the PI is creating both the DCP and the Project Dataset[s]; 3) ICES Research and Analysis Staff creating the DCP; and 4) ICES Analytic Staff (ICES Employee or agent responsible for creating the Project Dataset[s]). This may be delegated either verbally or via e-mail.*	
	Principal Investigator Aleksandra Zecevic ☐	2017-Aug-21
	Responsible ICES Scientist or Second ICES Scientist/Lead ☐	yyyy-mon-dd

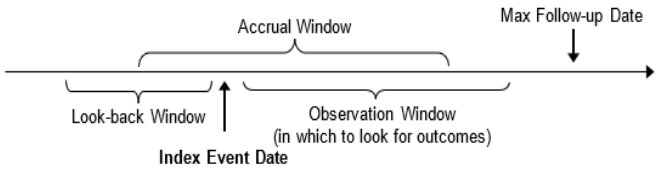
Project Initiation This Section must be Completed Prior to Project Dataset(s) Creation			
	ICES Research and Analysis Staff Creating the DCP	☐	yyyy-mon-dd
	ICES Analytic Staff	☐	yyyy-mon-dd
Designated ICES Research and Analysis Staff accountable for Project Documentation:	*The person named (ICES staff) is accountable for ensuring that the approved ICES Project PIA, ICES Project PIA Amendments, and DCP are saved on the T Drive, ensuring ICES Project PIA Amendments are submitted as required, ensuring DCP Amendments are documented, and sharing the final DCP with the PI/Responsible ICES Scientist at project completion*		
DCP Creation Date and Author:	*Date DCP was finalized prior to Project Dataset(s) creation*	*Name of person who created the DCP*	
	Date	***Name***	
	2017-Aug-20	***Yu Ming, Tyson Schierholtz, Aleksandra Zecevic***	
	2019-Jul-8	***Nicolette Lappan***	

ICES Data This Section must be Completed Prior to Project Dataset(s) Creation	
The ICES Employee or agent who is responsible for creating the Project Dataset(s) must ensure that this list includes only data listed in the ICES Project PIA *Changes to this list after initial ICES Project PIA approval require an ICES Project PIA Amendment*	*Mandatory for all datasets that are available by individual year*
General Use Datasets – Health Services	***Years (where applicable)***
See list	
See list	
General Use Datasets – Care Providers	
See list	
See list	
General Use Datasets – Population	
See list	
See list	
General Use Datasets – Coding/Geography	
See list	
See list	
General Use Datasets - Facilities	
See list	
General Use Datasets - Other	
See list	
See list	
Controlled Use Datasets	
See list	
See list	
Other Datasets	

Project Amendments and Reconciliation			
ICES Project PIA Amendment History (add additional rows as needed):	*Privacy approval date*	*Person who submitted amendment*	*Note that any changes to the list of ICES Data or Project Objectives require an ICES Project PIA Amendment*
	Date	***Name***	***Amendment***
	yyyy-mon-dd		
DCP Amendment History (add additional rows as needed):	*Date DCP amended*	*Person who made the DCP amendment*	*Note that any DCP amendments involving changes to the list of ICES Data or Project Objectives require an ICES Project PIA Amendment*
	Date	***Name***	***Amendment***
	yyyy-mon-dd		
Date Programs/DCP reconciled	*The person(s) creating the dataset and/or analyzing the data are responsible for ensuring that the final DCP reflects the final program(s) when the project is completed*		
	yyyy-mon-dd		

Project Cohort	
Study Design	☐ Cohort study ☐ Matched cohort study ☒ Case-control study ☐ Cross-sectional study ☐ Other (specify):
Index Event / Inclusion Criteria	• Serious Fall-Related Injuries and Death (Appendix A – ICD-10 W, S and T codes). When cutting a cohort of cases, please keep information on all 10 diagnostic codes from NACRS (e.g., dx10code1 – dx10code10), so we can confirm W code is combined with S or T code (e.g., S and W, T and W, S&T and W) to make it "fall-related". **Case group** inclusion criteria: a) Older adults 65 years and older b) Residents of Ontario c) Presented to Emergency Department d) Diagnosed with a serious fall-related injury or death due to fall-related injury. Fall-related injury is defined by combining ICD-10 codes for falls W00-W19 with ICD-10 codes for injuries S00-S99 or T00-T14. The time between codes W00-W19 and codes S00-S99 or T00-T14 should be the same day. Descriptive information (e.g., brakedown of S00-S99 or T00-T14 and W00-W19 codes) should be included into the dataset to allow analysis of injury types.) e) Diagnosed between Jan 1 2006 and Dec 31 2015. **Control group**

Project Cohort	
	a) Matched to the case group by sex, age Charlson Comorbidity Index score, and LHIN, with a ratio of 1.5 to 1. b) Exclude the patients having serious fall-related injuries (codes W00-W19, codes S00-S99 or T00-14) between Jan 1, 2006 and Dec 31, 2015. • The date of visiting Emergency Department due to serious fall-related injury will be defined as the index event date. • For individuals with repetitive serious fall-related injuries during the observation period, the first time and consecutive visits to ED due to serious fall-related injury will be taken into account.
Estimated Size of Cohort (if known)	
Exclusions (in order)	*Step* Description
	1 Invalid IKN
	2 Incomplete information (e.g, missing age or missing sex information, non-Ontario residents, died before ED visits)
	3 <65 years old at the index event date
	4 Patients who have experienced in-hospital serious fall-related injuries
	5 Excluding diagnoses that are suspected, questionable, rule out.
	6 Excluding transferred ED visits.
	7 Excluding ED visits from which a patient left without being seen.
	8 Excluding scheduled ED visits.

Project Time Frame Definitions	
	
Accrual Start/End Dates	2006 Jan 1st to 2015 December 31st
Max Follow-up Date	2017 Jan 1st

Project Time Frame Definitions	
When does observation window terminate?	12 months after the index date
Lookback Window(s)	12 months (till 2005 Jan 1st)

Variable Definitions (add additional rows as needed)	
Main Exposure or Risk Factor	For the full list of variables, please refer to Appendix B – Variables from NACRS, CIHI-DAD, RAI-HC and ODB.
Primary Outcome Definition	Serious fall-related injuries (ICD-10 codes W00-W19 combined with S00-S99 or T00-T14 ICD-10 codes)
Secondary Outcome Definition(s)	Death
Baseline Characteristics	Case and Control: age, sex, Charlson Comorbidity Index score, LHIN
Other Variables	Variables needed from NACRS, CIHI-DAD, RAI-HC and ODB databases are provided in attached file (Appendix B). IMPORTANT: to reduce number of observations we decided NOT to proceed with analysis of OHIP data, hence we excluded it in this revision.

Analysis Plan and Dummy Tables (expand/modify as needed)	
Descriptive Tables (insert or append dummy tables), e.g.: See Appendix C and Appendix D	
Table 1. Baseline characteristics according to primary/secondary exposure Descriptive statistics, measures of central tendency and dispersion for continuous variables, and frequency tables for categorical variables. Trends in the data across time will also be analyzed.	
Table 2. Outcomes according to primary/secondary exposure: See dummy tables in Appendix D	
Table 3. Covariates (baseline characteristics) according to outcomes : See dummy tables in Appendix D	
Statistical Model(s)	
Type of model	Cox Proportional Hazard Regression
Primary independent variable	Risk factors mentioned above
Dependent variable	Serious fall-related injuries and death

Analysis Plan and Dummy Tables (expand/modify as needed)	
Covariates	
Sensitivity Analyses	
Type of model	
Primary independent variable	
Dependent variable	
Covariates	

Quality Assurance Activities		
RAE Directory of SAS Programs		
RAE Directory of Final Dataset(s)	*The final analytic dataset for each cohort includes all the data required to create the baseline tables and run all the models. It should include all covariates for all models such as patient risk factors, hospital characteristics, physician characteristics, exposure measures (continuous, categorical) and outcomes. It should include covariates that were considered but didn't make the final cut. This would permit an analyst to easily re-run the models in the future.*	
RAE README file available: ☐Yes ☐No		
Date results of quality assurance tools for final dataset shared with project team (where applicable):		
	%assign	yyyy-mon-dd
	%evolution	yyyy-mon-dd
	%dinexplore	yyyy-mon-dd
	%track / %exclude	yyyy-mon-dd
	%codebook	yyyy-mon-dd
Additional comments:		

Appendix B. Ethics Approval

Western Research

Date: 4 August 2022

To: Dr. Aleksandra Zecevic

Project ID: 121336

Study Title: Machine Learning in Prediction of Fall-Related Injuries in Older Adults

Short Title: Machine Learning in Prediction of FRIs

Application Type: HSREB Initial Application

Review Type: Delegated

Meeting Date / Full Board Reporting Date: 23/Aug/2022

Date Approval Issued: 04/Aug/2022 16:26

REB Approval Expiry Date: 04/Aug/2023

Dear Dr. Aleksandra Zecevic

The Western University Health Sciences Research Ethics Board (HSREB) has reviewed and approved the WREM application form for the above mentioned study, as of the date noted above. HSREB approval for this study remains valid until the expiry date noted above, conditional to timely submission and acceptance of HSREB Continuing Ethics Review.

This research study is to be conducted by the investigator noted above. **All other required institutional approvals and mandated training must also be obtained prior to the conduct of the study.**

Documents Approved:

Document Name	Document Type	Document Date	Document Version
Study Procedures	Supplementary Tables/Figures		
REB 121336_03082022_CLEAN	Protocol	03/Aug/2022	121336

Documents Acknowledged:

Document Name	Document Type	Document Date
ICES-Privacy-Report	Technology Review document	27/Jun/2022
REB References	References	

No deviations from, or changes to the protocol should be initiated without prior written approval from the NMREB, except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario. Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB. The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Patricia Sargeant, Research Ethics Officer (psargean@uwo.ca) on behalf of Dr. Emma Duerden, HSREB Vice Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

Appendix C. Codes Used in SAS

Table C1

Finding Diagnostic Categories (Category 2 is Provided as an Example)

Step	SAS code
Create table Cat2_C with observations starting with C00 to C99	`CREATE TABLE Cat2_C AS SELECT Study_id, 'Y' AS Cat2 FROM nacrs_recs WHERE (dx10code1 LIKE 'C00%' OR dx10code1 LIKE 'C01%' OR dx10code1 BETWEEN 'C02' AND 'C99') OR (dx10code2 LIKE 'C00%' OR dx10code2 LIKE 'C01%' OR dx10code2 BETWEEN 'C02' AND 'C99')` `#Add similar conditions for dx10code3 to dx10code10 if necessary#`
Create table Cat2_D with observations starting with D00 to D48	`CREATE TABLE Cat2_D AS SELECT Study_id, 'Y' AS Cat2 FROM nacrs_recs WHERE (dx10code1 LIKE 'D00%' OR dx10code1 LIKE 'D01%' OR dx10code1 BETWEEN 'D02' AND 'D48') OR (dx10code2 LIKE 'D00%' OR dx10code2 LIKE 'D01%' OR dx10code2 BETWEEN 'D02' AND 'D48')` `# Add similar conditions for dx10code3 to dx10code10 if necessary#`

Merge Cat2_C and Cat2_D with nacrs_recs using Study_id	`proc sort data=Cat2_C;`
	`  by Study_id;`
	`run;`
	
	`proc sort data=Cat2_D;`
	`  by Study_id;`
	`run;`
	
	`proc sort data=nacrs_recs;`
	`  by Study_id;`
	`run;`
	`proc sql;`
	`  data cat2_merged;`
	`  merge Cat2_C Cat2_D nacrs_recs ;`
	`  by key_variable;`
	`run;`

Remove duplicate rows by Study_id	`CREATE TABLE cat2_final AS SELECT DISTINCT Study_id, <other columns> FROM cat2_merged;`

Table C2

Codes Used for Descriptive Analysis of FRI Observations

Step	SAS code

Divide the variable "age" into 6 groups	`proc sql;`
	`  create table nacrs_fri_with_groups as`
	`  select *,`
	`  case`
	`  when age between 65 and 69 then '65-69'`
	`  when age between 70 and 74 then '70-74'`
	`  when age between 75 and 79 then '75-79'`
	`  when age between 80 and 84 then '80-84'`
	`  when age between 85 and 89 then '85-89'`
	`  when age >= 90 then 'Over 90'`
	`  else 'Unknown'`
	`  end as age_group  from nacrs_fri;`
	`quit;`

Calculate the frequency of each age group	`proc sql;`
	`  create table age_group_frequency as`
	`  select age_group, count(*) as frequency`
	`  from nacrs_fri_with_groups`
	`  group by age_group;`
	`quit;`

Calculate the percentage of total frequency for each age group	`proc sql;`
	`  create table age_group_percentage as`
	`  select age_group, frequency, frequency*100/sum(frequency) as percentage`
	`  from age_group_frequency`
	`  group by age_group;`
	`quit;`

Find the number of values in the "sex" column where the value is 'm' (males) and 'f' (females)	`proc sql;`
	`  select sex, count(*) as count`
	`  from nacrs_fri`
	`  where sex in ('m', 'f')`
	`  group by sex;`
	`quit;`

Calculate the percentage of total frequency for each female and male	`proc sql;`
	`  create table sex_percentage as`
	`  select sex, count(*) as frequency, count(*)*100/sum(count(*)) as percentage`
	`  from nacrs_fri`
	`  where sex in ('m', 'f')`
	`  group by sex;`
	`quit;`

Appendix D. Codes Used in R

Table D1
R Codes Used for Decision Tree Model

Step	R code
Install necessary packages	`install.packages(c("rpart", "rpart.plot", "ggplot2", "lattice", "caret", "e1071"))`

Load the required libraries	`library(rpart)`
	`library(rpart.plot)`
	`library(ggplot2)`
	`library(lattice)`
	`library(caret)`
	`library(e1071)`

Load the dataset	`data(nacrs_final)`

Split the dataset into training and testing sets	`set.seed(123)`
	`trainIndex <- createDataPartition(nacrs_final$FRI, p = 0.7, list = FALSE)`
	`trainData <- iris[trainIndex, ]`
	`testData <- iris[-trainIndex, ]`

Train the decision tree model	`dt_model <- rpart(FRI ~ ., data = trainData, method = "class")`
Plot the decision tree	`rpart.plot(dt_model)`
Predict using the trained model	`predictions <- predict(dt_model, testData, type = "class")`
Create a confusion matrix	`confusionMatrix(predictions, testData$FRI)`
Calculate variable importance	`var_importance <- varImp(dt_model)`
Plot variable importance graph	`print(var_importance)`

Table D2
R Codes Used for Random Forest Model

Step	R code
Install necessary packages	`install.packages(c("stats", "dplyr", "randomForest"))`

Load the required libraries	`library(stats)`
	`library(dplyr)`
	`library(randomForest)`

Load the dataset	`data(nacrs_final)`

Split the dataset into training and testing sets	`set.seed(123)`
	`trainIndex <- sample(1:nrow(nacrs_final), nrow(nacrs_final) * 0.7)`
	`trainData <- nacrs_final[trainIndex, ]`
	`testData <- nacrs_final[-trainIndex, ]`

Train the random forest model	`rf_model <- randomForest(FRI ~ ., data = trainData, ntree = 100)`

Table D3
R Codes Used for XGBoost Model

Step	R code
Install necessary packages	`install.packages(c("xgboost", "magrittr", "dplyr", "Matrix"))`

Load the required libraries	`library(xgboost)`
	`library(magrittr)`
	`library(dplyr)`

Load the dataset	`data(nacrs_final)`
Convert the dataset to a DMatrix object	`dtrain <- xgb.DMatrix(as.matrix(nacrs_final[, -5]), label = as.numeric(nacrs_final$FRI))`

Set XGBoost parameters	`params <- list(`
	`  objective = "multi:softprob",`
	`  eval_metric = "mlogloss",`
	`  num_class = 3`
	`)`

Train the XGBoost model	`xgb_model <- xgboost(data = dtrain, params = params, nrounds = 10)`
Predict using the trained model	`predictions <- predict(xgb_model, as.matrix(nacrs_final[, -5]))`
Convert predicted probabilities to class labels	`predicted_labels <- max.col(predictions) - 1`
Create a confusion matrix	`confusionMatrix(predicted_labels, as.numeric(nacrs_final$FRI))`
Calculate variable importance	`var_importance <- xgb.importance(model = xgb_model)`
Plot variable importance graph	`xgb.plot.importance(importance_matrix = var_importance)`

Table D4
R Codes Used for Finding FRI Observations

Step	R code
Merging Full and Master datasets by Study_ID	`merged_dataset <- merge(Full, Master, by = "Study_ID")`
Filtering observations with case groups	filtered_dataset <- filter(Groups == 'case')
Filtering observations with "Days from index date to registration date" = 0	filtered_dataset <- filter(filtered_dataset, days_to_regdate == 0)
Merging NACRS and DAD datasets separately with the Master-Full dataset	NACRS_merged_dataset <- merge(filtered_dataset, NACRS, by = "Study_ID") DAD_merged_dataset <- merge(filtered_dataset, DAD, by = "Study_ID")
Removing multiple observations for the same Study_ID	`NACRS_final_dataset <- distinct(NACRS_merged_dataset, Study_ID, .keep_all = TRUE)` `DAD_final_dataset <- distinct(DAD_merged_dataset, Study_ID, .keep_all = TRUE)`

Appendix E. ICD-10-CA Codes Used for FRIs

Table E1

Codes for Fall Types based on the International Classification of Diseases Tenth Edition with Canadian Enhancements

Fall type	ICD-10-CA code
Fall on same level involving ice and snow	W00
Fall on same level from slipping, tripping and stumbling	W01
Fall involving skates, skis, sport boards and in-line skates	W02
Other fall on same level due to collision with, or pushing by, another person	W03
Fall while being carried or supported by other persons	W04
Fall involving wheelchair and other types of walking devices	W05
Fall involving wheelchair	W0500
Fall involving walker	W0501
Fall involving bed	W06
Fall involving chair	W07
Fall involving other furniture	W08
Fall involving playground equipment	W09
Fall on and from stairs and steps	W10
Fall on and from ladder	W11
Fall on and from scaffolding	W12
Fall from, out of or through building or structure	W13
Fall from tree	W14
Fall from cliff	W15
Diving or jumping into water causing injury other than drowning or submersion	W16
Other fall from one level to another	W17
Other fall on same level	W18
Unspecified fall	W19

Note. CIHI, 2015. ICD 10-CA = International Classification of Diseases Tenth Edition with Canadian Enhancements.

Table E2

Codes for Body Locations of Injuries based on the International Classification of Diseases Tenth Edition with Canadian Enhancements

Body location	ICD 10-CA codes
Head	S00-S09
Neck	S10-S19
Trunk	S20-S29 (thorax) S30-S39 (abdominal, lower back, lumbar spine and pelvis) T08, T09 (spine, trunk)
Upper limb	S40-S49 (shoulder and upper arm) S50-S59 (elbow and forearm) S60-S69 (wrist and hand) T10, T11 (level unspecified)
Lower limb	S70-S79 (hip and thigh) S80-S89 (knee and lower leg) S90-S99 (ankle and foot) T12, T13 (level unspecified)
Multiple regions	T00-T07
Unspecified level	T14 (unspecified body region)

Note. CIHI, 2015. ICD 10-CA = International Classification of Diseases Tenth Edition with Canadian Enhancements.

Appendix F. ICD-10-CA Codes Used for Diagnostic Categories

Table F1

Codes for Sprains, Strains, or Tears by Body Location based on the International Classification of Diseases Tenth Edition with Canadian Enhancements

Chapter	Title	ICD-10-CA Code
Chapter I	Certain infectious and parasitic diseases (A00-B99)	A00-B99
Chapter II	Neoplasms (C00-D48)	C00-D48
Chapter III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (D50-D89)	D50-D89
Chapter IV	Endocrine, nutritional, and metabolic diseases (E00-E90)	E00-E90
Chapter V	Mental and behavioural disorders (F00-F99)	F00-F99
Chapter VI	Diseases of the nervous system (G00-G99)	G00-G99
Chapter VII	Diseases of the eye and adnexa (H00-H59)	H00-H59
Chapter VIII	Diseases of the ear and mastoid process (H60-H95)	H60-H95
Chapter IX	Diseases of the circulatory system (I00-I99)	I00-I99
Chapter X	Diseases of the respiratory system (J00-J99)	J00-J99
Chapter XI	Diseases of the digestive system (K00-K95)	K00-K95
Chapter XII	Diseases of the skin and subcutaneous tissue (L00-L99)	L00-L99
Chapter XIII	Diseases of the musculoskeletal system and connective tissue (M00-M99)	M00-M99
Chapter XIV	Diseases of the genitourinary system (N00-N99)	N00-N99
Chapter XV	Pregnancy, childbirth, and the puerperium (O00-O99)	O00-O99
Chapter XVI	Certain conditions originating in the perinatal period (P00-P99)	P00-P99
Chapter XVII	Congenital malformations, deformations, and chromosomal abnormalities (Q00-Q99)	Q00-Q99
Chapter XVIII	Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99)	R00-R99
Chapter XIX	Injury, poisoning, and certain other consequences of external causes (S00-T88)	S00-T88
Chapter XX	External causes of morbidity and mortality (V01-Y99)	V01-Y99
Chapter XXI	Factors influencing health status and contact with health services (Z00-Z99)	Z00-Z99

Appendix G. ICD-10-CA Codes Used for Diagnostic Subcategories

Table G1
Codes for Subcategories of Category 2: Neoplasms (C00-D48)

Title	ICD-10-CA Code
Malignant neoplasms of lip, oral cavity, and pharynx	C00-C14
Malignant neoplasms of digestive organs	C15-C26
Malignant neoplasms of respiratory and intrathoracic organs	C30-C39
Malignant neoplasms of bone and articular cartilage	C40-C41
Melanoma and other malignant neoplasms of skin	C43-C44
Malignant neoplasms of mesothelial and soft tissue	C45-C49
Malignant neoplasm of breast	C50
Malignant neoplasms of female genital organs	C51-C58
Malignant neoplasms of male genital organs	C60-C63
Malignant neoplasms of urinary tract	C64-C68
Malignant neoplasms of eye, brain, and other parts of CNS	C69-C72
Malignant neoplasms of thyroid and other endocrine glands	C73-C75
Malignant neoplasms of ill-defined, secondary, and unspec.	C76-C80
Malignant neoplasms of lymphoid, hematopoietic, and others	C81-C96
In situ neoplasms	D00-D09
Benign neoplasms	D10-D36
Neoplasms of uncertain behavior, polycythemia vera, etc.	D37-D48
Malignant neoplasm of other and ill-defined sites	C76
Secondary and unspecified malignant neoplasms of lymph nodes	C77
Secondary malignant neoplasm of respiratory and digestive organs	C78
Secondary malignant neoplasm of other and unspecified sites	C79
Malignant neoplasm without specification of site	C80

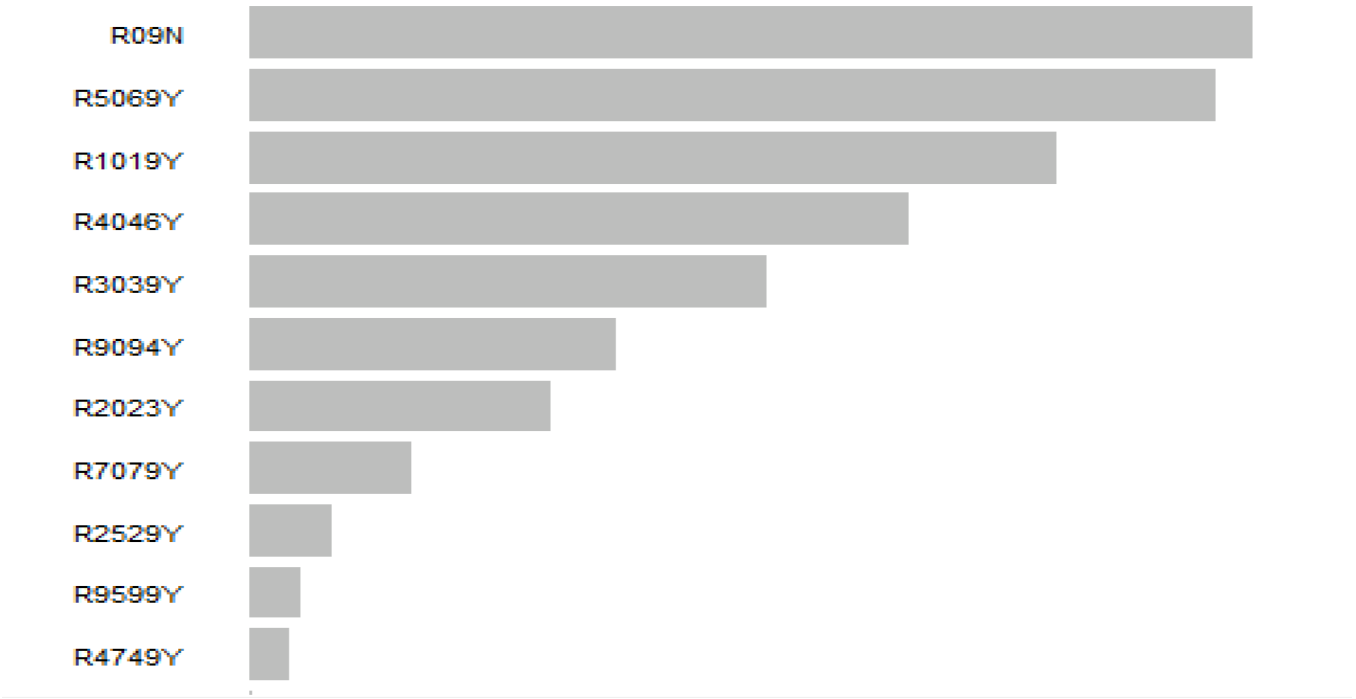
Table G2
Codes for Subcategories of Category 18: Symptoms, Signs, and Abnormal Clinical and Laboratory Findings, not Elsewhere Classified (R00-R99)

Title	ICD-10-CA Code
R00-R09	Symptoms and signs involving the circulatory and respiratory systems
R00	Abnormalities of heart rate and rhythm
R01	Cardiac murmurs and other cardiac sounds
R02	Gangrene, not elsewhere classified
R03	Abnormal blood-pressure reading, without diagnosis
R04	Hemorrhage from respiratory passages
R05	Cough
R06	Abnormalities of breathing
R06.0	Dyspnea
R06.1	Stridor
R06.2	Wheezing
R06.3	Acute respiratory distress
R06.4	Hyperventilation
R06.5	Mouth breathing
R06.6	Hiccough
R06.7	Sneezing
R06.8	Other abnormalities of breathing
R06.81	Apnea, not elsewhere classified
R06.82	Tachypnea, not elsewhere classified
R06.83	Snoring
R06.89	Other abnormalities of breathing
R09	Other symptoms and signs involving the circulatory and respiratory systems
R10-R19	Symptoms and signs involving the digestive system and abdomen
R20-R23	Symptoms and signs involving the skin and subcutaneous tissue
R25-R29	Symptoms and signs involving the nervous and musculoskeletal systems
R30-R39	Symptoms and signs involving the urinary system
R40-R46	Symptoms and signs involving cognition, perception, emotional state, and behavior
R47-R49	Symptoms and signs involving speech and voice
R50-R69	General symptoms and signs
R70-R79	Abnormal findings on examination of blood, without diagnosis
R80-R82	Abnormal findings on examination of urine, without diagnosis
R83-R89	Abnormal findings on examination of other body fluids, substances, and tissues, without diagnosis
R90-R94	Abnormal findings on diagnostic imaging and in function studies, without diagnosis
R95-R99	Ill-defined and unknown cause of mortality

Appendix H. Informativeness of Subcategories

Figure H1

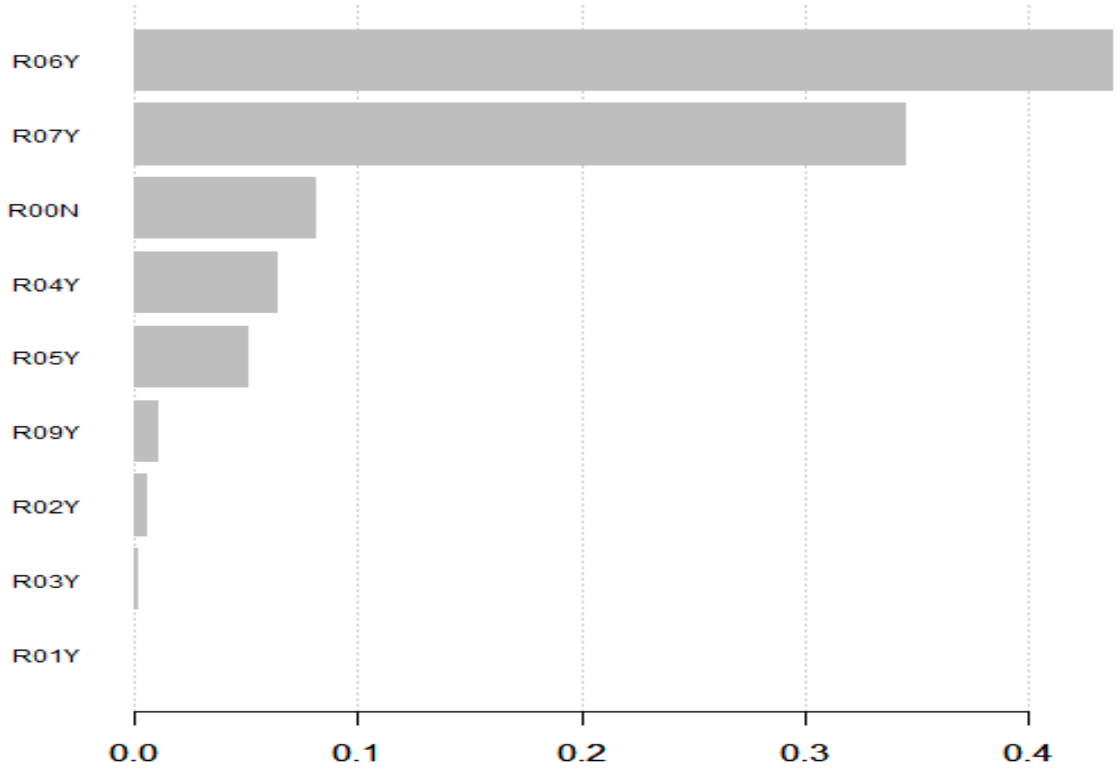
Informativeness of Subcategories of Category 18: Symptoms, Signs, and Abnormal Clinical and Laboratory Findings, not Elsewhere Classified (R00-R99)



Note. R09= Symptoms and signs involving the circulatory and respiratory systems; R5069= General symptoms and signs; R1019= Symptoms and signs involving the digestive system and abdomen; R4046= Symptoms and signs involving cognition, perception, emotional state and behaviour; R3039= Symptoms and signs involving the urinary system; R9094= Abnormal findings on diagnostic imaging and in function studies, without diagnosis; R2023= Symptoms and signs involving the skin and subcutaneous tissue; R7079= Abnormal findings on examination of blood, without diagnosis; R2529= Symptoms and signs involving the nervous and musculoskeletal systems; R9599= Ill-defined and unknown causes of mortality; R4749= Symptoms and signs involving speech and voice.

Figure H2

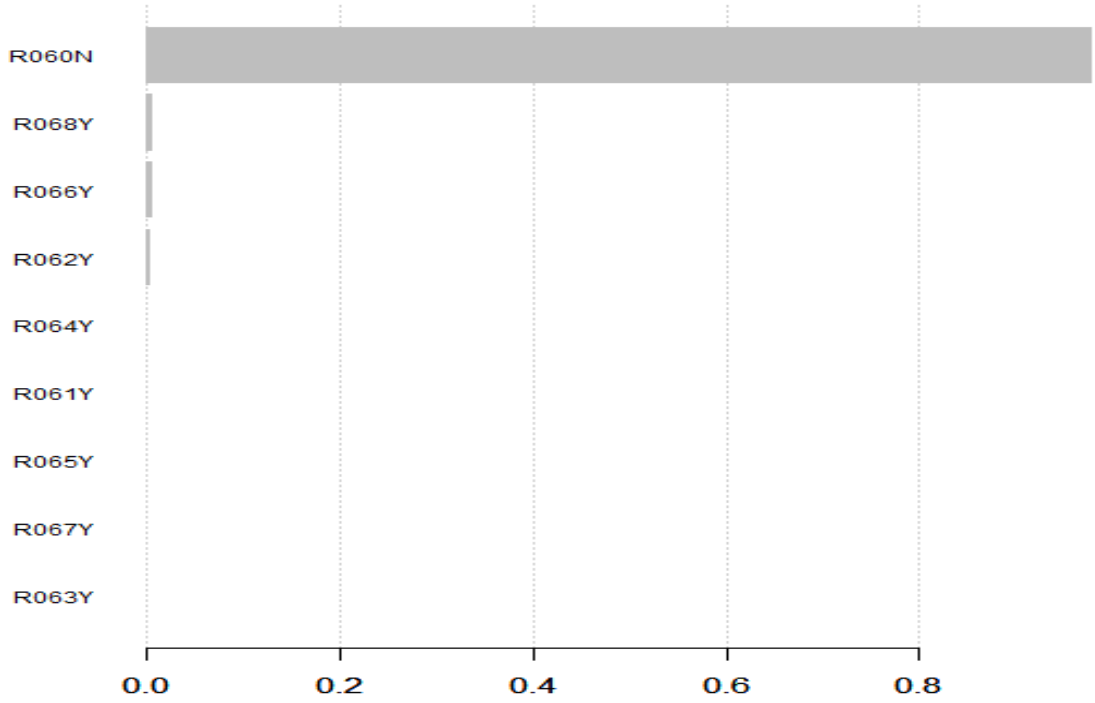
Informativeness of Subcategories of R00-R09: Symptoms and Signs Involving the Circulatory and Respiratory Systems



Note. R06= Abnormalities of breathing**;** R07= Pain in throat and chest; R00= Abnormalities of heartbeat; R04= Hemorrhage from respiratory passages; R05=Cough; R09= Other symptoms and signs involving the circulatory and respiratory systems; R02= Gangrene, not elsewhere classified; R03= abnormal blood-pressure reading, without diagnosis; R01= Cardiac murmurs and other cardiac sounds.

Figure H3

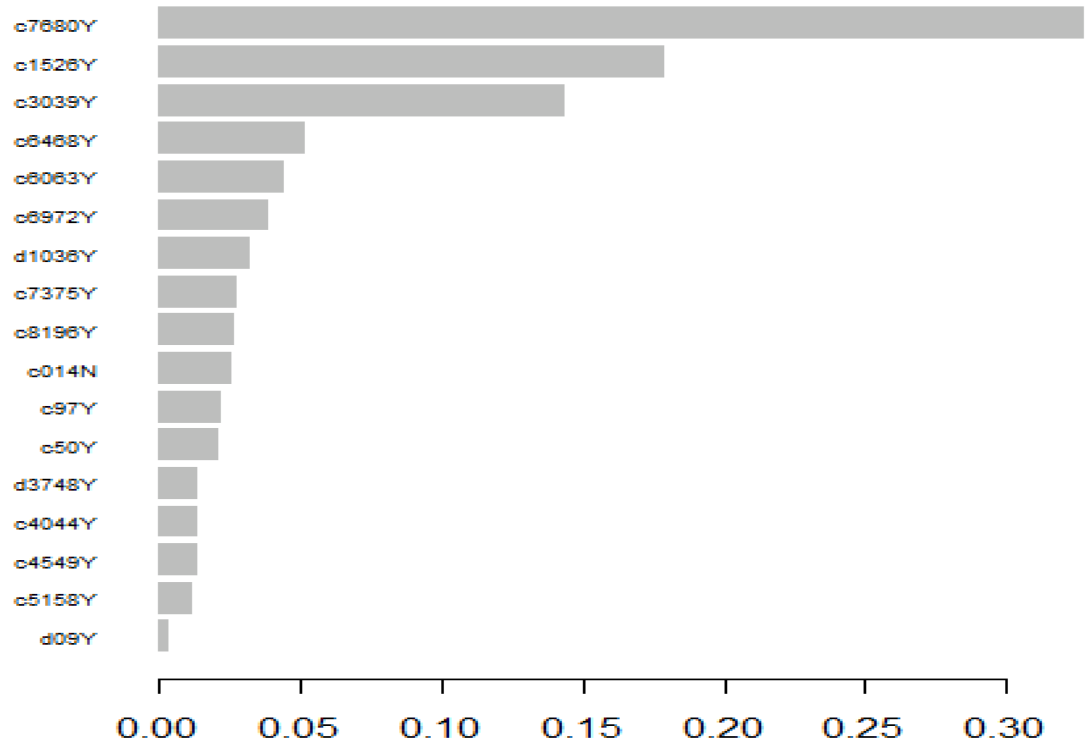
Informativeness of Subcategories of R06: Abnormalities of Breathing



Note. R060=Dyspnea; R068=Other and unspecified abnormalities of breathing; R066=Hiccough; R062=Wheezing; R064=Hyperventilation; R061=Stridor; R065=Mouth breathing; R067=Sneezing; R063=Periodic breathing.

Figure H4

Informativeness of Category 2: Neoplasms



Note. C7680= Malignant neoplasms of ill-defined, secondary and unspecified sites; C1526= Malignant neoplasms of digestive organs; C3039= Malignant neoplasms of respiratory and intrathoracic organs; C6468= Malignant neoplasms of urinary tract; C6063= Malignant neoplasms of male genital organs; C6972= Malignant neoplasms of eye, brain and other parts of central nervous system; C1036= Benign neoplasms; C7375= Malignant neoplasms of thyroid and other endocrine glands; C8196= Malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic and related tissue; C014= Malignant neoplasms of lip, oral cavity and pharynx; C97= Malignant neoplasms of independent (primary) multiple sites; C50= Malignant neoplasm of breast; D3748= Neoplasms of uncertain or unknown behaviour; C4044= Malignant neoplasms of bone and articular cartilage; C4549= Malignant neoplasms of mesothelial and soft tissue; C5158= Malignant neoplasms of female genital organs; D09= In situ neoplasms.

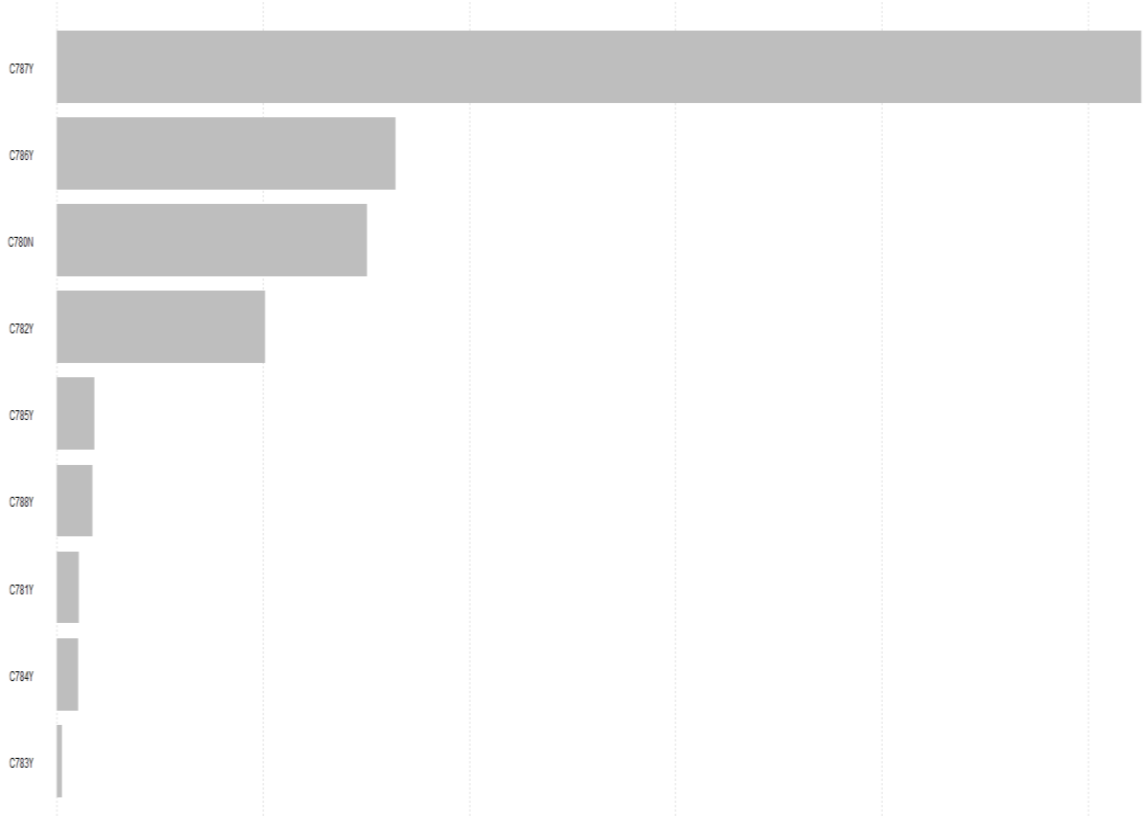
Figure H5

Informativeness of Subcategories of C76-80: Malignant Neoplasms of Ill-defined, Other Secondary and Unspecified Sites

*Note***.** C78= Secondary malignant neoplasm of respiratory and digestive organs; C79= Secondary malignant neoplasm of other and unspecified sites; C80= Malignant neoplasm, without specification of site; C77= Secondary and unspecified malignant neoplasm of lymph nodes.

Figure H6

Informativeness of Subcategories of C78: Secondary Malignant Neoplasm of Respiratory and Digestive Organs



*Note***.** C787= Secondary malignant neoplasm of liver and intrahepatic bile duct; C786= Secondary malignant neoplasm of retroperitoneum and peritoneum; C780= Secondary malignant neoplasm of lung; C782= Secondary malignant neoplasm of pleura; C785= Secondary malignant neoplasm of large intestine and rectum; C788= Secondary malignant neoplasm of other and unspecified digestive organs; C781= Secondary malignant neoplasm of mediastinum; C784= Secondary malignant neoplasm of small intestine; C783= Secondary malignant neoplasm of other and unspecified respiratory organs.

Curriculum Vitae

Name: Sorour Rostamour

Post-secondary Education and Degrees:
Shahid Beheshti University of Medical Sciences
Tehran, Tehran, Iran
2013-2017 BSc.

The University of Western Ontario
London, Ontario, Canada
2021-2023 MSc.

Honours and Awards:
Tehran, Iran, Student Bursary, Shahid Beheshti University of Medical Sciences
2013-2017

Western University Financial Support Package for Graduate Assistants
2021-2023

Related Work Experience
Teaching Assistant
The University of Western Ontario
2021-2023

Publications:

Piryonesi, S. M., Rostampour, S., & Piryonesi, S. A. (2021). Predicting Falls and Injuries in People with Multiple Sclerosis Using Machine Learning Algorithms. Multiple Sclerosis and Related Disorders, 49 (4), 102740. https://doi.org/10.1016/j.msard.2021.102740

Rostampour, S. (2020). The Alphabet of Stretching Exercises, Dastan Publication, ISBN: 978-600-481-214-60, Tehran, [in Persian].