STUDYING EMPLOYEE ABSENTEEISM DUE TO HEALTH-RELATED FACTORS: A DATA-SCIENCE APPROACH

By

KENNETH GRIFNO

Bachelor of Science in Business Administration
The University of Texas at Dallas
Richardson, Texas
1999

Master of Science in Management and Administrative Sciences
The University of Texas at Dallas
Richardson, Texas
2001

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2022

STUDYING EMPLOYEE ABSENTEEISM DUE TO HEALTH-RELATED FACTORS: A DATA-SCIENCE APPROACH

Dissertation Approved:

Dursun Delen, Ph.D.

Dissertation Advisor

Bryan Edwards, Ph.D.

Stephanie Royce, Ph.D.

Chenzhang Bao, Ph.D.

Name: KENNETH GRIFNO

Date of Degree: JULY, 2022

Title of Study: STUDYING EMPLOYEE ABSENTEEISM DUE TO HEALTH-RELATED FACTORS: A DATA-SCIENCE APPROACH

Major Field: BUSINESS ADMINISTRATION

Abstract: United States employers are spending approximately $950 billion on healthcare benefits, and these costs are impeding their ability to compete in their respective markets. Furthermore, these costs do not include employee absenteeism—the cost of failing to show up for scheduled work. Research has shown that the primary reason for employee absenteeism is poor health. However, management research has primarily focused on controllable factors related to avoidable absences (e.g., job burnout, work attitudes, and personality characteristics). Therefore, the critical issue I address in this dissertation is: How can employers understand, predict and decrease the effect of absenteeism related to the health conditions of their workforces?

A data-science approach was used to explore this critical question, focusing on the leading cause of disability, musculoskeletal disorders (MSDs), and how they impact employee absenteeism. First, I created a well-formed combined dataset using advanced data preparation methods on the datasets of three self-insured employers, their medical claims, pharmacy claims, human resource records, and attendance data. Next, I ran machine learning algorithms to examine the prediction accuracy and the most probable risk factors influencing employee absenteeism related to the health condition. For example, factors influencing the risk of increased absence related to poor health include demographic features of the employees and their position (e.g., age, gender, salary, department, and workload), existing health conditions at the time of absence (e.g., diabetes, behavior health, arthritis, cardiac, and gastrointestinal), treatments for the health condition (e.g., drug, physical therapy, non-surgical procedures, and surgical procedures), and other medical-related variables (e.g., provider types, locations, imaging, labs, and tests). The impact of time was also investigated to obtain treatment information because research indicates that shorter wait times correlate with better outcomes for MSD treatments. A post hoc analysis was conducted to compare the essential variables that predict long-term employee absenteeism to the critical variables that predict high medical costs. It provides important insights into which sorts of healthcare services are connected with a quality outcome (e.g., lower employee absence).

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

.

CHAPTER I

INTRODUCTION

> *"If healthcare costs to corporations are imagined as an iceberg, the proportion representing medical care is only the tip of the iceberg; the major portion is out of sight…the impact of absenteeism and presenteeism on productivity is enormous."*
>
> - Richard Ilka, MD, MPH

Healthcare costs impede companies' ability to compete in their respective marketplace. Finding ways to minimize this burden is critical (Stempel, 2018). Last year, the United States (U.S.) spent $3.8 trillion on healthcare, with employers spending approximately $950 billion on healthcare benefits (Centers for Medicare and Medicaid Services, 2021a; Integrated Benefits Institute, 2020). Furthermore, these costs do not include employee absenteeism—the cost of failing to show up for scheduled work (Johns, 2008). Excluding these lost productivity costs significantly understates the true cost of healthcare because they cost U.S. organizations $575 billion annually (Integrated Benefits Institute, 2020). The critical issue addressed in this dissertation is: How can employers understand, predict and decrease the effect of absenteeism related to the health conditions of their workforces?

Decreasing employee absenteeism is critical because it is a widespread problem that costs employers financial and productivity losses (Martocchio & Harrison, 1993; OECD, 2021; Singer & Cohen, 2020). The financial and productivity losses from employee absenteeism include (1) the cost of workers to take their place, (2) idle equipment, (3) disrupted production schedules that inconvenience customers, (4) increased inventory levels due to process delays, and (5) waste caused by substitute workers doing jobs for which they were not trained (Felton & Cole, 1963). Aside from increased absenteeism, poor health can lead to employees making lower-quality decisions (Boyd, 1997) and decreasing their overall contributions to their employer (Price & Hooijberg, 1992).

Management research on employee absenteeism in organizational behavior (OB) has been focused on factors related to five broadly defined cohorts: personality, demographics, attitudes, social context, and decision-making (Harrison & Martocchio, 1998) because OB is the study of behaviors of individuals and groups within organizations (Heath & Sitkin, 2001). Although these studies revealed invaluable insights by developing and testing a wide range of theories and questions, they did not provide an instrument to accurately predict and improve employee absenteeism related to poor health, which researchers have found to be the most important cause of employee absenteeism (Chadwick-Jones, Nicholson, & Brown, 1982; Hackett, Bycio, & Guion, 1989; Hedges, 1973, 1975, 1977; Morgan & Herman, 1976; Nicholson & Payne, 1987; Paringer, 1983), accounting for one-half to two-thirds of all absences (Brooke, 1986; Hedges, 1977; Miner & Brewer, 1976).

I, therefore, explore a critical question, focusing on musculoskeletal disorders (MSDs) and how they impact employee absenteeism. MSDs are the leading cause of disability and affect more than 1.7 billion people worldwide (World Health Organization

[WHO], 2021). MSDs, such as lower back pain, are highly prevalent in the working-age population, costly and a leading cause of occupational risks and absences (Forouzanfar, Afshin, Alexander et al., 2016; Hartvigsen, Hancock, Kongsted et al., 2018). Furthermore, MSDs such as sprains or strains caused by over-exertion in lifting accounted for 31% of all workers' compensation cases and required a median of 12 days to recover (Bureau of Labor Statistics, 2016). Research suggests that improved management of these conditions may improve productivity, benefiting both employers and employees (McDonald, daCosta DiBonaventura, & Ullman, 2011). Thus, the considerable burden and prevalence of MSDs highlight the need for an improved understanding of the various factors in their management.

Algorithms predicting employee absenteeism due to poor health may provide numerous benefits to organizations, such as identifying at-risk individuals and improving employee absenteeism management. However, machine learning prediction models have yielded mixed outcomes (Burdorf, 2019; Montano, Marques, Alonso et al., 2020). Two gaps in the existing machine learning prediction research have led to mixed outcomes. The first gap is the data itself. Although the retrospective data used in the studies is convenient, it has limitations. For example, researchers have recommended that employee absenteeism be measured to assess the impact of the various treatments for health conditions (Johns, 1997; Johnston, Harvey, Glozier et al., 2019). However, the current data does not contain the information needed for such an investigation.

This leads us to the second gap—a lack of medical-related variables to predict employee absenteeism. What if an employer wants to decrease employee absences attributable to a specific health condition? A more thorough set of facts is required to address this issue, allowing for an assessment that extends beyond the health problem itself. For

example, should employees with lower back pain be expected to be absent at the same rate as others? This dissertation examined a more extensive dataset to analyze treatment pathways (e.g., drug, physical therapy, non-surgical procedures, and surgical procedures) and other medical-related variables (e.g., treating providers, provider specialties, imaging, labs, and tests).

Given the excessive financial and productivity burden of employees with poor health, why don't employers have a better handle on this problem? According to Pfeffer (2018), one of the main reasons is that "there is little to no systematic (or even non-systematic) attention to measuring employee health and well-being in companies" (p. 194). One solution to overcome this issue has been self-reported questionnaires, such as the World Health Organization's Health and Work Performance Questionnaire (HPQ). Questionnaires such as the HPQ are designed to help researchers and employers better understand employee outcomes and associated variables, such as employee absence due to health conditions (Johnston et al., 2019).

Although polling employees may assist employers in measuring employee health and well-being in companies initially, self-reported questionnaires are not without faults. For example, the HPQ employee questionnaire has 21 pages of questions (Kessler, Barber, Beck et al., 2003). Not only does it take an employee a lot of time to fill out a questionnaire, but the questionnaire also excludes essential factors that we theorize are crucial in predicting employee absenteeism. Three examples are that the questionnaire does not contain all MSD health conditions, excludes all information regarding the employee's treatments, and excludes who ordered those treatments. Another potential issue with using self-reported questionnaires is short recall periods, which require employees to be polled often (Mattke,

Balakrishnan, Bergamo et al., 2007). Employees also may regard the polling as invasive (Follmer & Jones, 2018).

To overcome the challenges of self-reported questionnaires, different retrospective datasets available to employers were used to study employee absenteeism due to poor health—data from their healthcare claims and human resources information systems (HRIS). Self-insured employers can get information on paid claims for these services from their health plan administrators. However, the claims have a major flaw: they provide no information about what happened to the subject employees (Pfeffer, 2018). One measure for employee health outcomes would be employee absences related to those healthcare conditions and the effect of various treatments. However, workforce productivity outcomes are rarely measured on a comprehensive scale by employers (Skrepnek, Nevins, & Sullivan, 2012). This is where the HRIS data comes into play. By pooling the employer's healthcare claims data with their HRIS data, employers can examine the most prevalent risk factors and their relative importance in predicting employees at risk for prolonged absences.

Why aren't employers already using these datasets? First, many employers do not possess complete and integrated data or the resources needed to analyze the data correctly. For example, advanced data methods may be required to combine medical and pharmacy claims data with human resources demographic and attendance data. Therefore, in addition to the technical know-how, the employer will have to possess in-depth industry knowledge on how to use the healthcare data properly. This healthcare knowledge gap is understandable as most employers are not in the business of collecting and analyzing healthcare data. However, some large employers are now recruiting health management positions such as Chief Health Officer to assist in bridging this knowledge gap (Ilka, 2016).

This dissertation followed the Cross Industry Standard Process for Data Mining (CRISP-DM) data process and methods and utilized machine learning to provide a quantitative method to overcome the previously addressed gaps (Shearer, 2000). First, I applied advanced data preparation methods to the datasets of three self-insured employers, their medical claims, pharmacy claims, human resource records, and attendance data. To construct a well-formed combined dataset, I first consolidated it, sanitized it, transformed it, and used data reduction methods. Next, I ran machine learning algorithms to examine the prediction accuracy and the most probable risk factors influencing employee absenteeism related to health conditions. For example, factors influencing the risk of increased absence related to poor health include demographic features of the employee and their position (e.g., age, gender, salary, position, and workload), existing health conditions at the time of absence (e.g., diabetes, behavior health, arthritis, cardiac, and gastrointestinal), treatments for the health condition (e.g., drug, physical therapy, non-surgical procedures, and surgical procedures), and other medical-related variables (e.g., treating providers, provider specialties, locations, procedure severity, imaging, labs, and tests). I also investigated the impact of time to obtain treatment information because research indicates that shorter wait times correlate with better outcomes for MSD treatments (Lewis, Harding, Snowdon et al., 2018). Employers must understand how treatments affect employee absenteeism and how quickly the employee obtains that treatment.

As a result, this research aimed to identify and analyze the various employee health-related factors and investigate how these individual factors influence employee absenteeism. This goal was pursued explicitly through the two research questions guiding this study: (1) How can employers understand, predict and decrease the effect of absenteeism related to the

health conditions of their workforces? (2) And post hoc, how can employers compare the critical variables that predict long-term employee absence to the critical variables that predict high medical costs? The second question may provide important insights into which types of healthcare services (e.g., treatments, tests, and labs) are associated with a quality outcome (e.g., a lower rate of employee absenteeism).

Chapter II presents a conceptual background and reviews the literature on employee absenteeism definitions, types, and reasons. I then examine the existing research on machine learning approaches to predict employee absenteeism. I next investigate how individual factors influence employee absenteeism, what variables were lacking in previous models, and how those missing variables could improve the prediction of employee absenteeism through the lens of the Conservation of Resources (COR) theory (Hobfoll, 1989, 2011). The COR theory posits that poor health negatively impacts organizations and employee outcomes by reducing employees' cognitive and emotional resources (Hobfoll, Vinokur, Pierce et al., 2012). Next, I use the COR framework to explain how individual-level factors can help obtain, retain, foster, and protect against further resource losses; and be resource-restoring by identifying functional relationships between health preservation and absenteeism. Finally, I employed a Model Theoretic approach in this dissertation, in which empirical research was not used to validate or contradict the COR theory but to inform the theory (Harris, Johnson, & Souder, 2013).

CHAPTER II

CONCEPTUAL BACKGROUND & REVIEW OF LITERATURE

The concept of absenteeism is actually quite simple to understand. When it is used in everyday conversations, there is no need to pause and inquire what it means because it is utilized contextually with the story being told. While this frequently occurs in academic literature, researchers will also pause to define absenteeism plus any additional labels connected to their definition. As a result, there is no universally accepted definition for absenteeism in academia (Durand, 1985; Kohler & Mathieu, 1993). Therefore, before diving into the machine learning predictive models of employee absenteeism, both the definition and dimensionality of employee absenteeism is discussed. Equally important is ascertaining what is already known about absenteeism as well as what remains unknown. Finally, it is vital to understand why this specific construct is worthy of further research.

Defining Absenteeism

Since there is no commonly agreed definition for employee absenteeism, I start with how it has been defined. When reviewing the absenteeism literature, various

definitions have been used (see Table 1). This study's focus, in particular, may alter or impose restrictions on the concept and how it was defined. For example, a broad definition of absenteeism would be a person's lack of physical presence in a particular area and at a specific time (Gibson, 1966). Martocchio and Harrison (1993) extend this broad definition by attaching a social norm constraint to the meaning. They define absenteeism as the absence of physical presence at a specific location and time when there is a social expectation that they are present. Johns (1994) put it both simple and pure; absenteeism is failing to show up for scheduled employment. In both examples, the researchers use constraints aligned to the research question—not being at work when the employee is scheduled and expected to be present.

Table 1

Definitions of Absenteeism

Definitions	**Source**
"Failure to be present at the appropriate time and in the appropriate place to meet the terms of the contract."	Gibson (1966)
"Absence can be defined as missing work for a single day."	Fichman (1984)
"Simply stated, absenteeism occurs when an employee does not report for work, when he or she was scheduled or expected to be present."	Brooke (1986)
"An absence is an individual's lack of physical presence at a given location and time when there is a social expectation for him or her to be there."	Martocchio & Harrison (1993)
"Absenteeism is the failure to report for work as scheduled."	Johns (1994)
"Habitual absence from work for one or more days, usually justified by medical certificate but, actually, due to personal interests and poor sense of duty."	Cucchiella, Gastaldi & Ranieri (2014)
"Any failure by an employee to report for work as scheduled or to stay at work when scheduled."	Mathis, Jackson & Valentine (2015)
"Broadly defined as failure to attend scheduled work as a result of ill health."	Lawrance, Petrides & Guerry (2021)

Furthermore, numerous descriptive labels have been used to further constrain the definition of employee absenteeism, which is discussed next.

Types of Absenteeism

Much of the subsequent literature has adapted and expanded on Johns's (1994) definition, taking into account the study's parameters, encompassing descriptive labels and types and even reasons for employee absence. Table 2 lists commonly used descriptive labels that have been attached to the construct, employee absenteeism. It is crucial to recognize these descriptors since they can alter the meaning of employee absenteeism, perhaps leading to inaccurate assumptions and inferences when comparing similar research. However, not all of the descriptors shown in Table 2 carry the same impact on the definition, as I will highlight next.

Table 2

Types of Employee Absence

Types	**Source**
authorized—unauthorized	Clegg (1983)
avoidable—unavoidable	Johns (1994)
excused—unexcused	Muchinsky (1977)
legitimate—illegitimate	Gibson (1966)
physical—mental	Sagie, Birati, & Tziner (2002)
short term—long term	Hedges (1973)
voluntary—involuntary	Steers & Rhodes (1978)

The designations "short term" and "long term" are sensible descriptors because they focus on the measurement length of the construct (Marocchio & Harrison, 1993). Whereas qualifiers like unauthorized, unexplained, and illegitimate go beyond

measurement and may change the construct's definition because they imply that the absence must be approved. In addition, the approval-based descriptors are commonly used in conjunction with avoidable absences. Management researchers, for example, have investigated how poor job satisfaction and organizational commitment affect unauthorized employee absences. However, they frequently include the additional boundary for avoidable absences to eliminate other explanations, such as absences from poor health. They include the classification because the most prevalent reasons for unavoidable absences include illness, injuries, and accidents, which is not the primary focus of this study's research question.

Aside from the types of absenteeism discussed above, two types may lend themselves to different constructs. According to Sagie, Birati, and Tziner (2002), employee absenteeism could be physical or mental. Thus, a person can be physically present but mentally absent, either totally or partially. The authors are correct, but this definition aligns better with two different constructs—employee absenteeism and employee presenteeism. Employee presenteeism can be defined as attending work despite illness, regarded as a counterpart of employee absenteeism (Johns, 2010; Lohaus & Habermann, 2019).

Reasons for Absenteeism

The hypothesized reasons for employee absences have also influenced how the construct has been defined. Lawrance, Petrides, and Guerry (2021) adopted a narrower definition by including poor health as part of their criteria in their study on employee absenteeism. I begin with this example because employee health is at the heart of this dissertation. Research has shown that poor health is the leading cause of employee

absenteeism (Chadwick-Jones et al., 1982; Hackett et al.,1989; Hedges, 1973, 1975, 1977; Morgan & Herman, 1976; Nicholson & Payne, 1987; Paringer, 1983), accounting for up to two-thirds of all absences (Brooke, 1986; Hedges 1977; Miner & Brewer, 1976). However, since poor health has not been the focus of many previous studies on employee absenteeism, other critical factors in the research domain were reviewed (see Table 3). Due to the number of studies on employee absenteeism, I attempted to be representative rather than exhaustive, and I recommend Martocchio and Harrison (1993), Harrison and Martocchio (1998), and Johns (1997) for narrative analyses.

Organizational behavior (OB) is the study of the behaviors of individuals and groups within organizations (Heath & Sitkin, 2001). Therefore, it makes sense that research on employee absenteeism has focused on controllable factors related to five broadly defined groups: personality, demographics, attitudes, social context, and decision-making (Harrison & Martocchio, 1998). This focused view of the causes and correlates has yielded essential insights, but it has not proven a tool for adequately predicting and improving employee absence. For example, an early process model by Steers and Rhodes (1978) contained 24 variables related to personal characteristics, values, job situation, job satisfaction, pressures to attend, attendance motivation, ability to attend, and attendance.

Table 3

Reasons for Absenteeism

Reasons	**Source**
Personal characteristics, employee values and job expectations, job situation, satisfaction with job situation, pressures to attend, attendance motivation, ability to attend, and employee attendance	Steers & Rhodes (1978)

Routinization, centralization, pay, distributive justice, role ambiguity, role conflict, role overload, work involvement, organizational permissiveness, kinship responsibility	Brooke (1986)
Industry, white-collar/blue-collar, category of job (sales, managerial, clerical), type of sick leave plan, age, gender	Hedges (1977)
Work strain, psychological and physical illness (anxiety, acute stress/illness, burnout, emotional exhaustion, depersonalization, lack of personal accomplishment, depression, fatigue, negative mood, physical composite, psychological composite, psychosomatic/ill health), individual (attribution, disposition, gender), social (occupational status, macro context)	Darr & Johns (2008)
Job scope + satisfaction, time-value, ideal workweek, weekend overtime, weekday overtime, work-nonwork interaction, education, income, tenure, children, sex, race	Youngblood (1984)
Minor illness (self), mental health day, illness in the family, family social function, work to do at home, emotional problems, school work to do, bereavement, physical fatigue, professional appointments, obnoxious patients, hangover-partying late, frustrated with work, snow storm-weather, misread time sheet, bought a house-moving, nice day, too little time off, hard to concentrate, no permanent ward, shift change-tired, worked overtime, missed bus-car problem, extend holiday, compassionate leave, ward is overstaffed, house was robbed, pregnancy, religious holiday, mad at the supervisor, and peace rally	Hackett, Bycio & Guion (1989)
Job-satisfaction (overall, promotion, co-workers, pay, work, and supervision)	Hackett & Guion (1985)

Although Steers and Rhodes's (1978) model contained many variables, Fichman (1984) found no empirical evidence for the model, and later, Rhodes and Steers (1990) themselves found limited support for the model. Brooke (1986) then improved on the Steers and Rhodes model. However, the subsequent study revealed a poor fit between data and model (Hendrix & Spencer, 1989), with only limited support for the model discovered (Brooke & Price, 1989). Furthermore, researchers who re-examined the association between job satisfaction and employee absenteeism discovered that job

satisfaction scores explained less than 4% of the variance in absence measurements (Hackett & Guion, 1985).

The one area in the literature that has consistently reported significant relationships and the most considerable portion on variance to employee absenteeism is poor health. Hedges (1977) included variables beyond the five categories mentioned above and found that illness and injury accounted for the bulk of all hours lost. Darr and Johns (2008) conducted a meta-analysis on work strain, health, and absenteeism and found a significant link between absenteeism, work stress, mental disease, and physical illness. Youngblood's (1984) research supports the notion that absenteeism results from motivation processes in both the work and non-work domains. Furthermore, illness absence and the importance a person placed on non-work time were significantly related. Hackett, Bycio, and Guion (1989) reported that poor health was the leading predictor of absenteeism. Poor health and tiredness were found to be associated with absenteeism. Employees stated that poor health was one of the most prevalent reasons for past absences.

This leads us to the question: Why has a majority of the management research ignored sources of variances related to poor health? One reason is that management researchers have focused on employee absenteeism deemed avoidable, and absences related to poor health have been viewed as unavoidable. However, are employee absences related to poor health unavoidable? For argument's sake, let us assume the reason for the absence is unavoidable, then my next question becomes: Are all the days in the absence episode unavoidable? I propose that this answer is "no" since the medical treatments and

when the employee receives the treatments for their health issue may affect the length of their absence from work.

Prediction Models for the Study of Absenteeism

As previously noted, the management literature on employee absenteeism has primarily focused on the explanation and relationship of numerous avoidable risk factors with employee absence rather than the accuracy of predictions (Lawrance et al., 2021). For example, commonly researched factors in the management literature have included job burnout, work attitudes, personality characteristics, mental health, time off entitlements expiration, environment, and even equipment (Brooke, 1986; Darr & Johns, 2008; Hackett et al., 1989; Hedges, 1977; Steers & Rhodes, 1978). Notably, employee absenteeism from job burnout and poor health do have some crossover; however, their focus is different. The most significant difference is that employee absenteeism due to job burnout is focused on the social environment at work. On the other hand, employee absenteeism due to poor health is broad and affects all aspects of our lives. As a result, in my literature review, I found that most contributions in predicting employee absenteeism were from fields other than management, such as occupational health and medicine (see Appendix 1).

The next section summarizes my findings from the literature review on applying various prediction modeling techniques to predict employee absenteeism. In reviewing the literature, I concentrated on four criteria: model performance, variables, data, and algorithms. However, my primary focus was on model performance and the factors needed to develop a useful and transparent tool for identifying employee absenteeism attributable to poor health.

Area Under the Curve (AUC)

I used the metric Area Under the Curve (AUC) to examine model performance and not the model's overall prediction accuracy. The AUC is generated from the Receiver Operating Characteristics (ROC) curve, which assesses classifier performance across a complete range of class distributions and error costs and has been empirically proven to be a better measure for machine learning applications (Bradley, 1997; Ling, Huang & Zhang, 2003). Additionally, measuring the AUC on a ROC chart is a more straightforward method for evaluating models and comparing different classifiers. The AUC allows values ranging from 0.0 to 1.0, with a baseline of 0.5, which means that classifiers with an AUC less than 0.5 perform worse than a random guess (e.g., coin toss). An AUC of 0.5 to 0.7 is regarded as poor, 0.7 to 0.8 good, 0.8 to 0.9 excellent, and greater than 0.9 remarkable. The higher a classifier's performance, the closer its AUC is to 1.0 (Hosmer, Lemeshow & Sturdivant, 2013).

I start with the AUCs that have been reported. Prediction models had mixed results in my examination of the literature, which is similar to the findings of other researchers (Burdorf, 2019; Montano et al., 2020). I reviewed eleven papers containing 61 individual models. A frequency histogram of reported AUC values from prior studies is depicted in Figure 1. Thirty-one models were classified as poor, twenty-two as good, three as excellent, and the remaining did not meet the baseline of 0.50 (see Appendix 1). It should be noted that the AUC for the three excellent models ranged from 0.80 to 0.81, barely crossing the threshold.

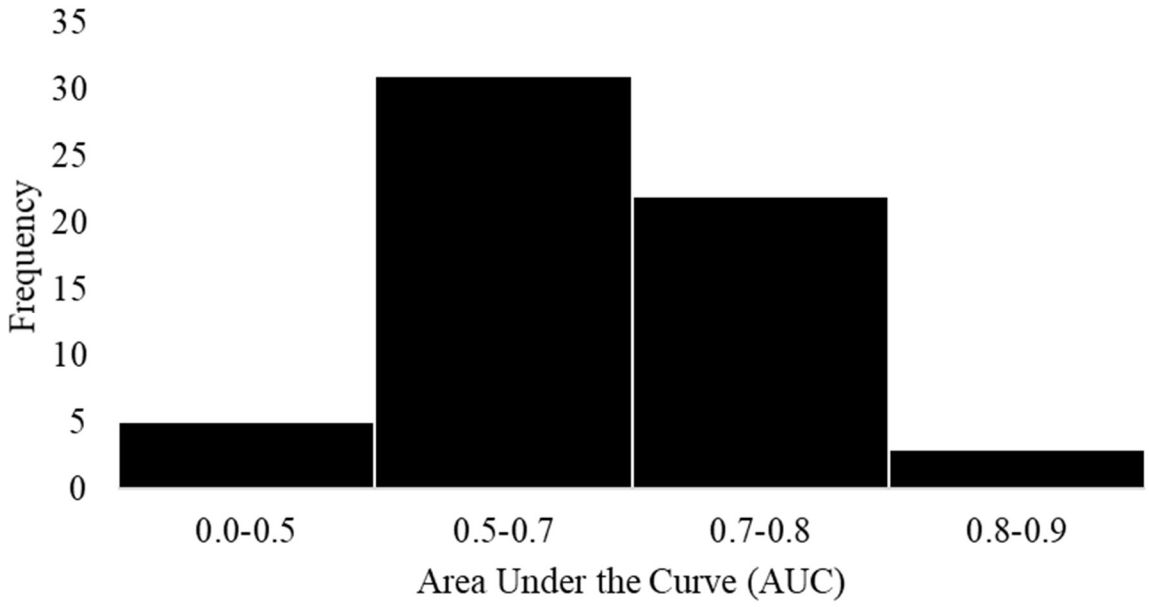


Figure 1. Frequency histogram of reported AUC values from referenced prior studies.

Prediction Models

The reviewed papers all had one thing in common: predicting employee absence. They each, however, took a different systematic approach to achieve their aim. For example, the prediction models deployed included statistical techniques such as logistic regression and many algorithmic models that included numeric, classification, and ensemble models. However, the critical distinction between the publications was the number of individual variables included in the models and how those variables were operationalized (see Appendix 1). These distinctions apply to both the independent and dependent variables. For example, employee absence was measured differently in eight of the eleven papers. In six papers, the dependent variable definition was long-term absence; however, the long-term absence was defined using different ranges, for example, $>$ 9 days, $\geq$ 30 days, $\geq$ 42 days, $\geq$ 90 days. The remaining models' focus was

not on long-term employee absence and used either hours or days to operationalize the dependent variable.

In addition, a wide range of predictor variables has been used in prior research. For example, one study used data from the Finnish Public Sector (FPS) and Health and Social Support (HeSSUp) to build and test a risk prediction model for long-term (> 9 days and ≥ 90 days) sickness absence (Airaksinen, Jokela, Virtanen et al., 2018). Their model for > 9 days was poor (AUC = 64.7%), but their model for ≥ 90 days was good (AUC = 73.5%). Predictor variables for sociodemographic, health status, lifestyle behaviors, and working conditions were included in the dataset and their models. Poor self-rated health, age, gender, smoking, depression, past illness absence, chronic diseases, and socioeconomic position were linked to sickness absence. Airaksinen et al.'s study found that none of the work-related factors improved prediction after demographics and lifestyle variables were included in the model.

Notenbomer, van Rhenen, Groothoff, and Roelen (2019) aimed to predict long-term sick absence (≥ 42 days) from illness or injury using a dataset from The Netherlands Occupational Health Register. They examined two prediction models; the first model focused on job demands and resources (AUC = 62.3%), and the second model focused on job burnout and work engagement (AUC = 62.4%). Both models included additional variables such as age, gender, education, prior long-term sickness absence, which were significant predictors, but the models reflected poor overall performance. Although they aimed to predict long-term sick absence from illness or injury, they did not include any healthcare-related variables in their models.

Many of the researchers did include health-related factors in their studies but did not focus on predicting employee absenteeism for a specific health-related condition. Could adding the boundary help fine-tune and improve the prediction results? In a recent study, researchers examined the Örebro Musculoskeletal Pain Screening Questionnaire (ÖMPSQ) concerning long-term (≥ 30 days) sick leave for 185 employees with back pain (Bergström, Hagberg, Busch et al., 2014). Bergström et al. reported excellent prediction results (AUC = 81%) for the first six months following the baseline, noting that it was the highest reported AUC in their review. The primary focus of the ÖMPSQ tool is on long-term pain and disability for those suffering from neck and back pain. It includes factors related to the individual's pain, functioning (both physical and psychological), and fear-avoidance beliefs. The researchers included additional factors such as those related to the employees' demographics, prior absence, job strain, co-workers support, manager support, and monotonous. The ÖMPSQ was significant, and when the total score was above 90, it indicated that the employee had a five times increase in risk of long-term absences. In addition, prior year long-term sick absence was significant along with two workplace factors: co-worker and manager support.

However, simply adding boundaries does not guarantee good prediction results. For example, a cohort study on Dutch employees aimed to predict long-term employee absenteeism due to lower back pain based on characteristics gathered in occupational health exams (Bosman, Dijkstra, Joling et al., 2018). Bosman et al.'s prediction model included age, pain and stiffness variables, number of musculoskeletal conditions, work health complaints, work stress, fatigue, and work factors such as satisfaction. They also

looked at the prediction results for employees with manual and non-manual jobs, but both models had poor discrimination (manual AUC = 65.9%, non-manual AUC = 69.2).

Van Hoffen, Roelen, van Rhenen et al. (2018) took a similar approach. However, they looked at predicting long-term employee absenteeism (≥ 42 days) due to mental disorders for Dutch employees, emphasizing psychosocial work variables. They examined employees' workload, work pace, changes in work, variety in work, autonomy in work, participation in work decisions, learning opportunities, feedback about one's performance, support from supervisor, and support from co-workers for psychosocial work factors. They also included employee demographics (such as age, gender, and education), job demographics (such as job type, tenure, and worked hours), and mental health. Along with predicting long-term employee absenteeism for mental disorders, they also looked at all-causes of long-term employee absenteeism. Both models showed poor model performance (e.g., mental health AUC = 65%, all-cause AUC = 59%), which shows that the added specificity did help in the outcome. However, although the emphasis was on psychosocial work characteristics, the models did not help predict mental health-related long-term absenteeism.

Van Hoffen, Norder, Twisk, and Roelen's (2020) research looked at predicting long-term employee absenteeism (≥ 42 days) due to mental disorders for Dutch employees but with an emphasis on sociodemographic and work-related variables. They ran two prediction models, logistic regression and decision tree models, using occupational health survey data, including 27 variables. The logistic regression model performed good (AUC = 71.3%) and used 11 variables (e.g., gender, marital status, economic sector, company tenure, role clarity, cognitive demands, learning opportunities,

co-worker support, social support from family/friends, work satisfaction, and distress). The decision tree model performed almost as well (AUC = 70.9%), which used 3-nodes (e.g., distress, gender, work satisfaction, and work pace). However, the decision tree model has the advantage of providing better transparency and use in practice.

Neisse, de Oliveira, F.L.P., de Oliveira, A.C.S., and Neto (2021) conducted a cross-sectional study on 621 mine workers in Brazil focused on assessing the risk of chronic fatigue syndrome (CFS) on employee absenteeism. The available variables in their dataset included employee demographics (e.g., age, height, weight, gender) and many medical markers (e.g., body fat percent, blood pressure, cholesterol, triglycerides, calcium, glucose). Neisse et al. ran three different prediction models which all had similar results that showed poor model performance (model 1 AUC = 58.43%, model 2 AUC = 56.97%, model 3 AUC = 57.46%). Although the models did not perform well, they were able to show that there are effects from both sodium and total cholesterol on employee absenteeism and partial agreement on LDL and Triglycerides.

Skorikov, Hussain, Khan et al. (2020) examined many data mining techniques (e.g., zeroR, tree-based J48, Naïve Bayes, and KNN) using a dataset on a Brazilian courier company that contained 20 variables to understand and predict employee absenteeism (see Appendix 1). Employee absenteeism was examined in hours using three classes (class A = 0, class B = 1-15, and class C = 16-120) and three different studies of the available predictor variables. Skorikov et al. applied the Correlation Feature Set for study one, which identified the month of the absence, employee age, disciplinary failure, and social drinker as the most influential variables to predict employee absence. For study two, they included all the available predictor variables. Study three only included

one disciplinary failure because it had the highest information gain and feature score. Study two included all 19 predictor variables using KNN-Euclidean, which performed the best (AUC = 81%), and Naïve Bayes (AUC = 80%) was right behind it. Study one overall did not perform as well as study two did, with its best model being Naïve Bayes (AUC = 77%). Study three performed the worst and did not have any model exceeding an AUC more significant than 69%.

Singer and Cohen (2020) used the Brazilian courier company dataset to demonstrate interpretable classification algorithms for understanding factors and predicting employee absenteeism. There were a few deviations from the Skorikov et al. (2020) study. First, Singer and Cohen examined a different set of algorithms because their goal was to compare interpretable ordinal and non-ordinal classifiers. Second, they excluded the variable, reason for the absence, from the model because it would not be known in advance of the absence. Third, employee absenteeism was grouped into four classes (not absent = 0 hours, hours = 1-8 hours, days = 8-39 hours, and weeks = ≥ 40 hours). Overall, they examined two ordinal (CART OBE(c^{mode}) and CART OBE(c^{max})) and six non-ordinal (XGBoost, Multilayer Perceptron, KNN, naïve Bayes, Random Forest, and CART) classifiers across the four classes. Singer and Cohen's results show that OBE(c^{max}) performed the best of the eight models run (hours AUC = 75%, days AUC = 72%, and weeks AUC = 65%,). Although the model performed well for the hours and days classes, their primary class of interest (weeks) performed poorly.

de Oliveira, Torres, Moreira, and de Lima (2019) also used the Brazilian courier company dataset and applied various existing machine learning methods to predict employee absenteeism. They applied seven machine learning models (Random Forest,

Multilayer Perceptron, Support Vector Machine, Naïve Bayes, XGBoost, and Long Short-Term Memory) and used 241 variables to compare employee absenteeism predictions. The available variables in the dataset included employee features, work activities, social and administrative platform features, absenteeism features, but it lacked any health-related features. Two of the seven models performed well (XGBoost AUC = 72.6%, Random Forest AUC = 71.0%), and the remaining performed poorly.

Lawrance, Petrides, and Guerry (2021) applied a machine learning approach to identify groups of employees who were at risk of being absent due to illness so that interventions could be tailored to decrease or avoid absences. The data used for their study was Human Resources and payroll records from 280 Belgian employers across many sectors. The predictors they used from the dataset included employee features (e.g., age, gender, education), work environment (e.g., wage, contract type, shift irregularities), and historic absence patterns. However, though they focused on employee absenteeism due to sickness, their study was void of any health-related predictors. Instead, employee absences by month were aggregated and classified as zero or one on whether they met their assigned threshold value. Lawrance et al. ran many combinations of decision tree ensembles and used a decision tree (CART) as the base algorithm for each prediction period (see Appendix 1). On average, their models performed poorly, but there were three prediction periods with a reported AUC between 70% and 71%.

Conservation of Resources Theory

The independent studies had the same goal but used slightly different approaches to predict employee absenteeism. However, the studies did have a few things in common. First, most of the studies only included a limited set of health conditions. This study

addressed this gap and included a richer set of health conditions and comorbidities-based variables in the models. Second, the reviewed studies lacked healthcare variables that could influence a person's resource gains and further resource depletion beyond previously examined healthcare factors (e.g., chronic diseases, comorbidities, mental health, and substance abuse). For example, none included variables that examined the medical treatments received, who administered the treatments, when the treatments first started, or how long. Each one of these variable types could very well impact the variability in a person's resource pools affecting the length of their absence from work.

Hobfoll's (1989) Conservation of Resources (COR) theoretical framework was used for this current study to explain how the factors may affect the length of an employee's absence. According to Hobfoll (2011), people "strive to obtain, retain, foster, and protect the things they centrally value," which are resources (p. 117). According to the COR theory, a significant cause of poor mental and physical health outcomes is the loss of resources, and the critical goal is to address these losses (Hobfoll et al., 2012). Resources have been generally characterized and grouped into four cohorts: items, conditions, personal, and energy, and recently have expanded to include health and coping. The basic argument is that the COR theory focuses on what people centrally value most and the incentive they utilize to attain, maintain, and guard those things. Thus, centrally valued is universal and includes health, well-being, peace, family, self-preservation, amongst other things. These factors are at the heart of this current research, examining factors that may help obtain, retain, foster, and protect against further resource loss and may even be resource-restoring by identifying functional relationships between health preservation and absenteeism.

The COR theory comprises two principles and four corollaries, but its basic foundation is resource losses (Hobfoll, 1989). The secondary principle of COR addresses resource investments, stating that "that people must invest resources in order to protect against resource loss, recover from losses, and gain resources" (Hobfoll, 2001). Poor health is a primary source of resource loss in that it can deplete an individual's mental, emotional, and even physical resources. According to COR theory's Corollary 3, if a person has resources, they are more likely to gain, and the initial gain simply fosters additional gains. It also claims that loss cycles have a more significant impact and occur faster than gain cycles since a loss of resources carries more weight. In this dissertation, I link the employee absence episodes to these cycles to acquire insights into the loss and gain of resources from individual health and treatment-related variables. Also, a model-theoretic perspective was utilized whereas empirical research was used to inform the COR theory rather than validate or refute it (Harris et al., 2013).

The Need for New Factors

Let us examine the need to address these gaps using three scenarios: (1) Diabetes, (2) Behavioral health, and (3) Musculoskeletal disorders. I chose these health conditions because all three are prevalent and costly to the working-age population. For example, diabetes estimated costs exceeded $325 billion in 2017, up by 26% from 2012. However, most of these costs were for medical care and not employee absenteeism (American Diabetes Association, 2018). For example, one vial of Humalog (insulin lispro) in 2019 cost $332, up by 1480% since 1999 (Rajkumar, 2020).

Furthermore, unlike diabetes, behavioral health and musculoskeletal disorders are two conditions that have influenced employees' ability to work and are the two leading

causes of disability in the United States (Forouzanfar et al., 2016). This relationship can be seen in our dataset. For example, Figure 2 below shows that the medical costs in the treatment of diabetes make up the majority of the total cost (medical cost + absence cost). This makes sense; just because an employee has a chronic condition such as diabetes does not mean it is the cause for employees to miss work, especially if they are managing the condition. However, when looking at musculoskeletal disorders (e.g., back, knee, and neck), we see a different relationship between employee medical and absence cost; a large proportion of the cost is for absenteeism.

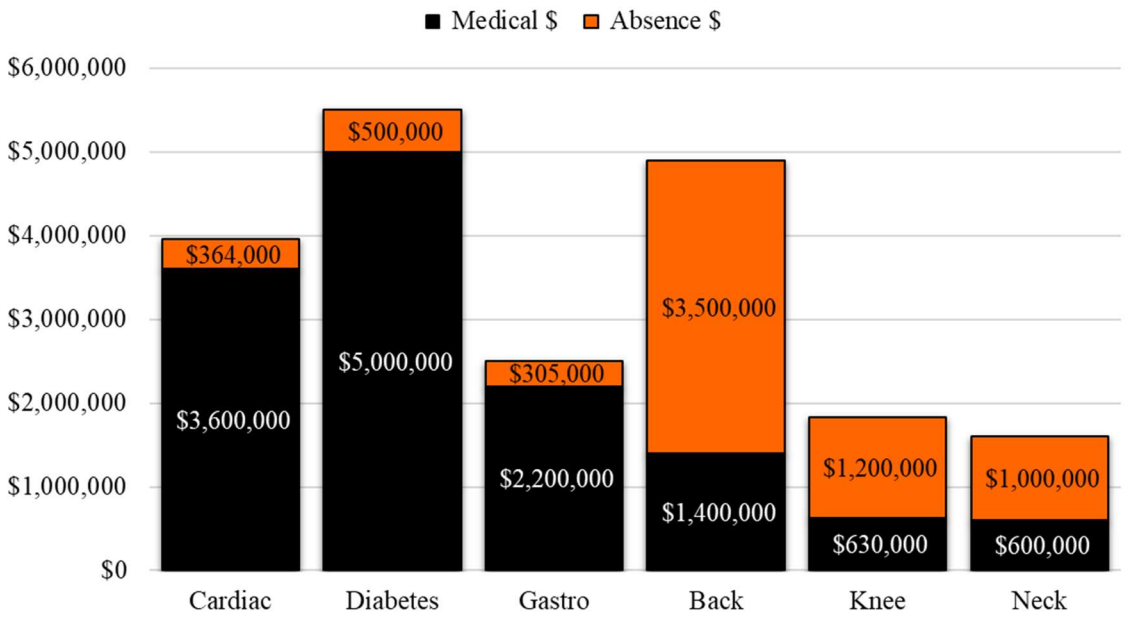


Figure 2. Sample of relationships between employee medical and absence costs for cardiac, diabetes, gastrointestinal, back, knee, and neck diseases and disorders.

Due to its prevalence and high medical costs, it is understandable why diabetes has garnered so much interest. Diabetes has also been found to have a significant association with employee absence (Moore & Buschbom, 1974; Pell & D'Alonzo, 1967). Having a significant relationship with absence makes sense because, according to the

American Diabetes Association (ADA), an employee effectively managing their diabetes will visit their primary care provider four times per year (ADA, 2020). Therefore, the patients in this population will miss more work time in aggregate when compared to employees without diabetes. However, was it diabetes that led to those employees with long-term absenteeism? It could be diabetes, but it is also plausible that the employee could be suffering other health conditions that contributed to their long-term absence. For example, could it be related to one of the two leading causes of disability, behavioral health or musculoskeletal conditions (World Health Organization, 2019, 2021)?

The point is that there are many other health conditions beyond diabetes that could lead to the cause of employees' absence and could be misconstrued by an employer that is not in the business of healthcare. Thus, it was crucial to determine what is causing their absence from poor health because knowing the cause could lead to more informed investments in employee health benefits and management programs. Therefore, expanding the scope of factors used to predict employee absenteeism can lead to decisions based on data-driven results instead of investing based on information that was not a possible cause of their absence.

According to Johns (1997), it is becoming evident that depth is required to comprehend the construct employee absence, and expecting a complicated set of linkages to hold up in any given sample may be asking too much. To address this notion, in this dissertation, I add depth to the dataset by including additional health conditions and examining the medical treatments received, who treated them, and the timing related to the person's care. What do I mean by adding depth with medical treatments of a health condition? To demonstrate, let us explore this query using a single condition of

behavioral health such as depression. Then, at a high level, three common therapy types are used in treating depression (e.g., drug therapy, talk therapy, and hybrid therapy), assuming the employee seeks treatment.

A recent study examined medical care expenditures and quality-adjusted life-years for talk therapy and antidepressants separately (Ross, Vijan, Miller et al., 2019). Ross et al. concluded that pharmaceutical therapies are more cost-effective than talk therapy treatments. Their study, however, did not evaluate combination medical therapies and did not investigate their effect on employee absenteeism. What if employees treated with drug therapy miss less work than those treated with talk therapy? What if employees treated with a hybrid approach miss less work than those getting drug or talk therapies? If the treatment does influence the length of the employee's absence, as it does medical costs, then the duration of employee absence from poor health is not unavoidable.

In addition, some treatments have a wide variety of eligible professionals that can prescribe and administer treatments. For example, those that are seeking talk therapy (e.g., CPT code 90834: Psychotherapy, 45 minutes with the patient) for a mental health issue may receive treatment from many different eligible healthcare providers, which include medical doctors, doctors of osteopathy, clinical psychologist, clinical social workers, and even clinical nurse specialists, nurse practitioners, and physician assistants, to name a few (Centers for Medicare and Medicaid Services [CMMS], 2021b). It may be essential to know the type of health professional who treats the employees as they have different training, licenses, and cost for their services. What if employees treated by psychologists miss less work than those treated by a social worker? An employer armed with this information could enhance their employee benefit offerings to improve this

outcome. In this scenario, I used the depression health condition to walk through how different treatments and those providing the treatment may influence the variability in the length of employee absence.

For the third scenario, musculoskeletal disorders (MSDs) was used because it is the leading cause of disability worldwide and an issue that many can relate to (WHO, 2021). MSDs are not made up of a single health problem but are comprised of many conditions (e.g., osteoarthritis, bones, and muscles) across multiple body areas (e.g., back, neck, and knees) that impact the musculoskeletal system of individuals (WHO, 2021). Therefore, I created individual variables that make up the MSDs group. Having individual variables is important because they inform us which MSDs the employee has and how many. For example, regarding employee absenteeism, will an employee miss more work if they have multiple MSDs such as neck and lower back pain versus only neck pain?

Much like depression, MSDs can involve a single type of treatment, multiple types of treatment, and those treatments can be from different types of licensed providers. For example, lower back problems are commonly treated by primary care providers, psychiatrists, chiropractors, physical therapists, and specialists, including pain management, orthopedic surgeons, and neurosurgeons (Shalen, 2000). For example, take an employee undergoing a treatment plan that includes physical therapy, typically provided by a physical or occupational therapist. Additionally, this employee may have their health condition treated with manipulative and adjustment therapy, commonly performed by chiropractors but can also be provided by a physical therapist. In another example, an employee could receive drug treatments to reduce inflammation or pain,

which different types of eligible physicians can prescribe. Also, knowing if an employee is being treated with a high-risk drug such as opioids and who is prescribing those high-risk treatments is important because their misuse has often led to adverse outcomes, including employee absence (Van Hasselt, Keyes, Bray, & Miller, 2015).

Surgical specialists (e.g., orthopedic surgeons, neurosurgeons) also provide procedural-based treatments such as injections, infusions, implants, and surgical repairs. Research has shown evidence that health outcomes for a person are influenced when they have utilized different treatments (Hurwitz, Morgenstern, Harber et al., 2002). Researchers have also found that referral patterns are different for health providers in treating musculoskeletal conditions, which is influenced by their specialty, which can directly impact the person's treatment regimen (Freburger, Holmes, & Carey, 2003). Therefore, treatment variables are needed in the model to help understand how the varying treatments received influenced the employee's resource pools.

How does the aspect of time influence employee absence? Other than time being used to define employee absence (e.g., short-term, long-term) or employee tenure, it has been void in the literature. Time may play a critical role in the length of an employee's absence from work. An employer can provide all the medical benefit products and programs needed for their employees' health and well-being. However, can they get the care they need when they need it? For example, if an employee has a health issue, how long does it take to see a healthcare provider? If the employee cannot get an appointment to see a healthcare provider for a week, will waiting one week impact the length of the employee's absence? The simple argument here is that if access to care is inadequate, the employee may miss additional time from work waiting to get care; their health may

deteriorate further as the employee waits because the untreated condition may deplete their resource pools. Prior research has shown evidence that how long a person has to wait for care is associated with health outcomes, specifically for the referral to an initial musculoskeletal visit, which influences employee participation (Lewis et al., 2018; Solomon, Bates, Punish et al., 1997).

Each of the above scenarios was designed to show the importance and need for this research. How can employers better understand and predict employee absenteeism concerning their workforce's health? Each scenario emphasizes the necessity for a richer dataset that makes available new healthcare-related factors that may better understand this complex construct. The addition of new variables will aid in effectively predicting and improving long-term employee absence from poor health. In addition to the primary predictor variables of interest, I included employee demographics (e.g., employee type, employee status, age, gender) and job attributes (e.g., department, position, tenure, compensation) found in the literature review.

Employers have employee populations that require different levels of physical effort to perform their job duties. The level of physical activity needed for their job may influence how much time an employee misses work. For example, an employer may see similar utilization of absence and the length of absence concerning MSDs from jobs with high physical effort (e.g., nurses, firemen, police officers, construction workers) compared to lower physical effort positions (e.g., computer programmers, accountants). Consider a nurse treating patients in a hospital with lower back pain; this MSD may significantly impact the nurse's ability to do their job than a computer programmer and may require them to miss more time from work. The two positions in this example have

different resource requirements, which may influence the overall impact of certain health conditions on employee absence. The above example is crucial because if the nurse cannot work due to lower back pain, the hospital may be required to cover their shift with a temporary nurse from a nurse pool, which adds cost and potentially lowers healthcare service levels.

In addition to the primary study on predicting and improving the understanding of long-term absence from poor health, I conducted a post hoc study. Two primary data sources (medical and pharmacy claims) are routinely used to study both medical costs and the quality of care with machine learning techniques for the health insurance and medical industries (Doupe, Faghmous, & Basu, 2019; Kshirsagar, Hsu, Chaturvedi et al., 2020; Maisog, Li, Xu et al., 2019; Obermeyer & Emanuel, 2016; Saxena, Das, Rubens et al., 2019). To expand upon the initial study, I ran the same machine learning prediction models using medical costs as the dependent variable instead of employee absence. Therefore, the goal of Study 2 was to examine if the critical variables that predict high medical costs also predict long-term employee absenteeism. In addition, Study 2 offers an employer's view on the adage that higher medical care costs do not always imply a better healthcare outcome (Anderson & Chalkidou, 2008). Using an employer's perspective was essential because it could provide insights into which types of healthcare costs (e.g., treatments) are associated with a quality outcome (e.g., lower employee absence). Chapter III presents the methodology for this study.

CHAPTER III

METHODOLOGY

This dissertation tests several prediction models for employees with musculoskeletal disorders (MSDs) using factors related to employees' demographics, job, and medical care received to examine the impact on employee absenteeism and medical costs. I analyzed a combined dataset using KNIME 4.5.2 Advanced Analytics platform to accomplish this (see Figure 3). In KNIME, I evaluated eight machine learning models (e.g., Artificial Neural Network, Decision Tree, Gradient Boosted Trees, K Nearest Neighbor, Logistic Regression, Naïve Bayes, Random Forest, and XGBoost) using *k*-Fold Cross-Validation estimation methodology. In addition, I conducted two studies using retrospective datasets from three employers in two sectors (healthcare and government). The primary study focused on the prediction accuracy for employees' long-term absenteeism and the relative importance of each study variable. The second study was post hoc, and I ran the same predictive models using the dependent variable, cost of medical care. The purpose of Study 2 was to see if the variables of importance that predicted high medical costs would also predict long-term absenteeism. The outcome was expected to either support or not support the adage that more medical care costs do not necessarily relate to a good healthcare outcome (Anderson & Chalkidou, 2008). This is

important from an employer's viewpoint because it may lend insights into what types of healthcare costs (e.g., treatments) are related to high quality and others with low quality.

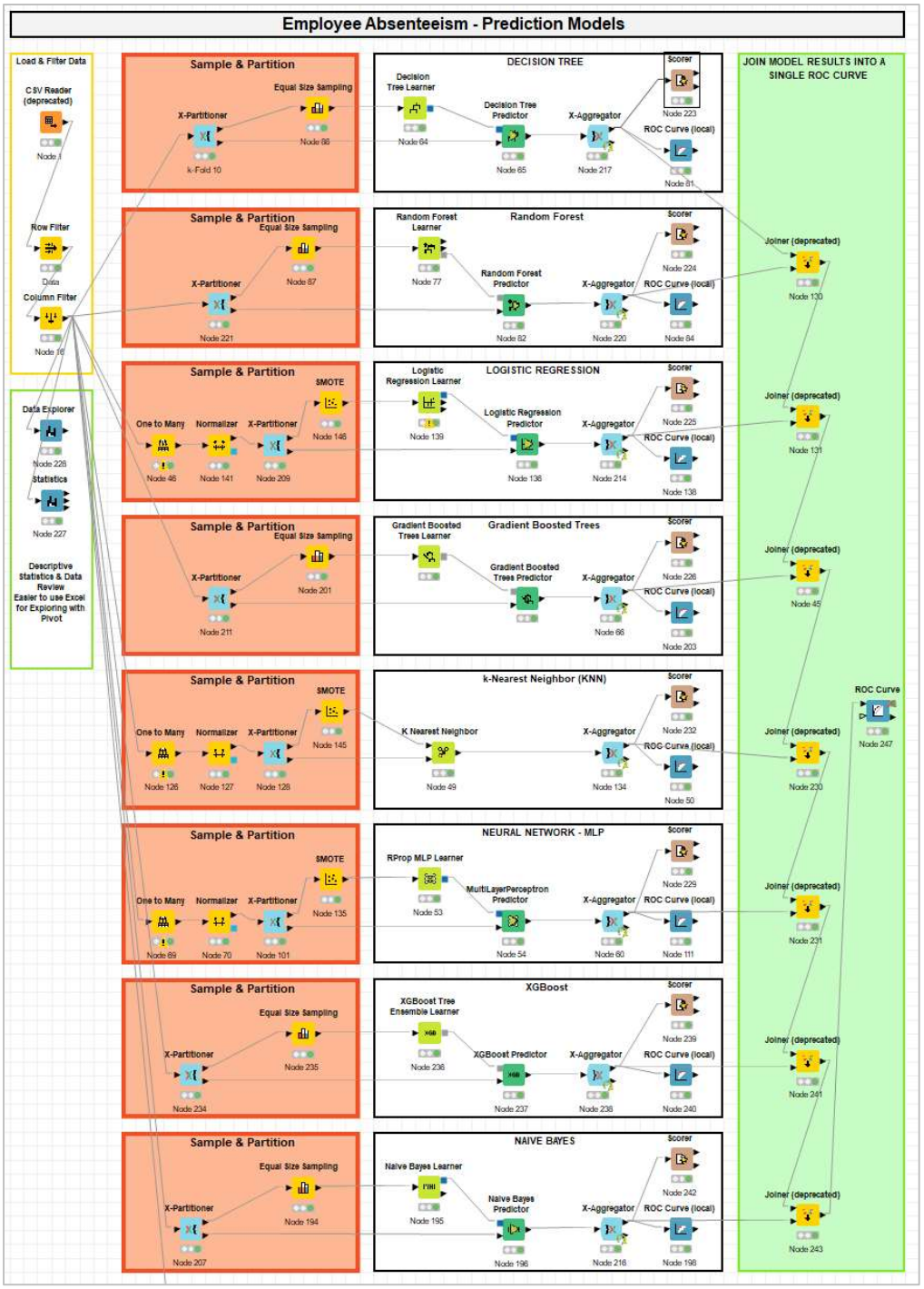


Figure 3. Graphical representation of prediction model workflows inside KNIME.

Each employer self-insures their health benefits (e.g., medical and pharmacy care), combined total 27,000 employees (employer 1 = 18,000, employer 2 = 6,000, and employer 3 = 3,000), and are located in the United States. Self-insuring their health benefits allows the employers access to their employees' medical and pharmacy administrative claims data representing the billing of rendered medical services and treatments. The employers have a combined annual medical and pharmacy burden of at least $215 million (employer 1 = $115 million, employer 2 = $70 million, and employer 3 = $30 million), which excludes the costs due to employee absenteeism related to health conditions (e.g., depression, back pain, diabetes). Each employer uses a human resources time and attendance system to track employees' time, capturing workforce schedules and their status (e.g., normal, sick, vacation, holiday, and bereavement). However, in this study, the primary focus is on employees' sick hours related to their musculoskeletal disorders and treatments.

Sample and Procedures

For this dissertation, I focused on a subset of the employee population that had been diagnosed with musculoskeletal disorders (N = 8,162). Among participants, 67.75% were women, and 32.25% were men with an average age of 47. I defined employees with musculoskeletal disorders using the medical claims data. I identified all employees with a primary ICD-10 (International Statistical Classification of Diseases and Related Health Problems, 10th revision) diagnosis code within the range of M00 through M99, which are codes used to diagnose diseases of the musculoskeletal system and connective tissue. The scope was limited to employees with a diagnosis in a single year because I wanted to make sure I had the longitudinal data to expand 18 months past the employee's initial

diagnosis (first visit by a treating provider for the primary diagnosis of a musculoskeletal disorder). Each of the datasets contained 3 to 4 years of data; no data past 2019 was used due to the possible influence of the COVID-19 outbreak. I examined the six months before the employee's initial diagnosis to validate that there were no prior claims for the documentation or treatments of the musculoskeletal disorder.

Next, I excluded employees on several job-related filters. First, the employees had to be full-time employees of the organization. Part-time employees were excluded because their health benefits and time off work benefits are different from full-time employees. Second, the employees identified had to be actively employed during the measurement period (12-months prior to and 18-months post the initial diagnosis). Lastly, I excluded employees who were not enrolled in the employer's medical and pharmacy health benefit programs for the study's duration.

Medical Claims Data

The administrative medical claims data represents the rendered medical services and treatments that were billed for each insured participant in the employer-sponsored health plan. The employer self-insures their medical benefits, but a third-party administrator (TPA) manages their medical claims processing, data, and healthcare provider network. Therefore, this dataset contains important attributes about each service (e.g., diagnosis codes, treatment codes, service date, location, who provided the treatment and paid amounts). This dataset was essential because it is used to identify MSDs and create many of the healthcare variables used for the studies. However, this dataset does

not contain drug treatment data, which contains essential information needed to create many of the study's predictor variables related to musculoskeletal disorders.

Pharmacy Claims Data

The administrative pharmacy claims data represents the filled prescriptions that were billed for each insured participant in the employer-sponsored pharmacy benefit plan. The employer self-insures their pharmacy benefits, but a TPA manages their pharmacy claims processing, data, and pharmacy network. This dataset contains important attributes about each filled prescription (e.g., drug names, national drug codes, date prescription filled, location prescription filled, who wrote the prescription, quantity, and paid amounts). It was used to create medical treatment variables related to drug therapy.

Human Resources Data

The human resources data contains employee demographics (e.g., employee type, employee status, age, gender), job attributes (e.g., department, position, tenure, compensation), and employee time (in hours). Each employer uses different vendors for their human resources systems (e.g., PeopleSoft and Work Day), but each collect the data types noted above. Each of the human resources information systems includes a module that manages and collects the time and attendance for each employee. This data includes the individual employee schedules, time, and status of entered time (e.g., normal, sick, vacation, holiday, and bereavement) used to calculate employee absence hours.

Secondary Data

In this study, I included publicly available datasets that I used to create many healthcare variables based on accepted standards in the healthcare industry. I used the International Classification of Diseases (ICD-10-CM) to classify medical diagnoses into health condition boundaries and individual variables (Centers for Disease Control and Prevention [CDC], 2021). I used the Medicare Severity Diagnosis Related Groups (MS-DRGs) to classify inpatient hospital stays into health conditions and identify inpatient stays (CMMS, 2021c). During an inpatient stay, there can be many ancillary medical services provided, which I used in variable creation and are documented by Revenue Codes (UB-Rev-Code) that are defined by the National Uniform Billing Committee (Research Data Assistance Center, 2021)

For the classification of healthcare providers, I used the National Plan and Provider Enumeration System (NEPPES) database (CMMS, 2021d). Each healthcare provider has a National Provider Identification Number (NPI) stored in NEPPES used for linkage between the claims (medical and pharmacy) and used to link to the specialty and taxonomy crosswalk (CMMS, 2021e) for their provider type classification. In addition, I used the National Drug Code Directory for pharmacy claims, which I linked to using the National Drug Code (NDC), which serves as the FDA's unique identifier for drugs (Center for Drug Evaluation and Research, 2020). This database contains additional attributes about the drugs that assisted in classifying drug treatment variables.

Combined Data

Next, I created a well-formed combined dataset following the Cross-Industry Standard Process for Data Mining (CRISP-DM) data process and methods (see Figure 4)

(Shearer, 2000). Data consolidation was the first step of the CRISP-DM data preparation process, which involved collecting, selecting, and integrating the data. The second step was the data cleaning process, where I imputed values, reduced noise, and eliminated duplicates. The third step was data transformation, where I normalized the data, discretized data, and created data. Finally, in the last step of the process, I focused on data reduction, including reducing data dimensionality, reducing the volume of data, and balancing the data. The critical linkage between the primary sources was the employee's identification number (EIN); each dataset contained the employee's identification number and dates to link the datasets together across time.

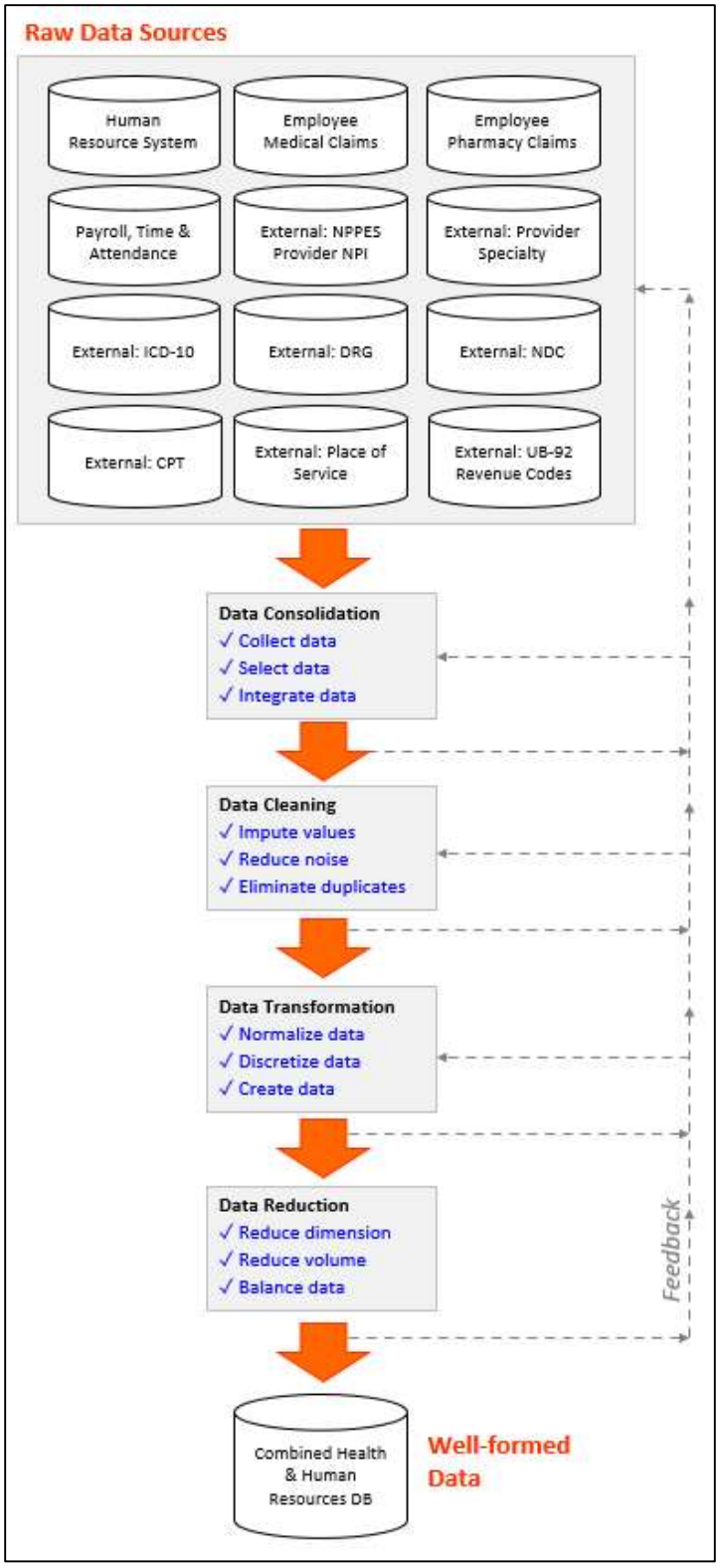


Figure 4. Individual datasets combined to create a well-formed dataset (Adapted from Delen, 2020).

Measures

Three primary variable groups were used to explore this study's research questions (employee demographics, job demographics, and healthcare variables). Variables in the employee and job demographics groups have been widely researched, and many were found to be significant when exploring employee absence. Although certain healthcare variables have been studied in employee absence, variables focused on how the health condition was treated, who prescribed or performed the treatment, and access (time) to treatment is lacking in the research of employee absenteeism. Therefore, many known and new variables were included to evaluate the prediction model's performance to improve model accuracy and better understand variable importance. See Appendix 2 for individual distribution charts for each measure used in the studies.

Predictor Variables

Health Conditions. Health conditions are a primary variable of interest for Studies 1 and 2. Although health conditions have been included in past studies on employee absence, they have been limited in scope. Therefore, the number of health conditions have been expanded. Each health condition is defined using the diagnostic Hierarchical Condition Categories (HCCs) except for individual musculoskeletal disorders (CMMS, 2021h). The Department of Health and Human Services created HCCs to group various diagnosis codes into indicators for different health conditions (Kautter, Pope, Ingber et al., 2014; Pope, Kautter, Ellis et al., 2004). Over 69,000 ICD-10 diagnosis codes and a subset of approximately 9,700 are focused on acute and chronic health conditions that map to 131 HCCs (based on the Final 2021 Benefit Year Risk

Adjustment Coefficients), each representing a single medical condition (see Appendix B for distributions). Out of the 131 HCC variables currently included in the HCC Risk Adjustment Model, 23 did not have any participants that met the diagnosis HCC criteria. In addition to the HHS-HCC risk model, I used the risk score established by the Chronic Illness and Disability Payment System (CDPS) as another indicator of the person's overall health. The CDPS system is also a diagnostic-based risk adjustment model commonly utilized to alter capitated payments for health plans that enroll Medicaid beneficiaries (CDPS, 2022).

However, the models mentioned above did not offer the specificity needed to define musculoskeletal disorders for this study. Many of the models focused on chronic health conditions and excluded non-diagnostic diagnoses (e.g., a diagnosis of abdominal pain), clinically insignificant diagnoses (e.g., a sprain), or diagnoses that are definitively treated (e.g., acute appendicitis) because they are not likely to impact a person's long-term health expenditures (Yeatts & Sangvai, 2016). Therefore, I used a two-step procedure to define the individual groups by body area (e.g., back, shoulder, knee, arm/elbow, hand/wrist, foot/ankle, and leg/hip). For the first step, I leveraged the Clinical Classifications Software Refined (CCSR), which aggregates each ICD-10 diagnosis code into clinically meaningful categories (Agency for Healthcare Research and Quality [AHRQ], 2021). Using the CCSR, I flagged all participants who had a CCSR equal to two different classifications (1. Diseases of the Musculoskeletal System and Connective Tissue, and 2. Injury, Poisoning, and Certain Other Consequences of External Causes). In the second step, I organized each of the individual codes into cohorts by body area.

Treatments. Treatment variables are one of the primary variables of interest in

this dissertation (see Table 4). The treatments that employees receive may directly influence their ability to recover from a health issue and reduce their time at work. However, the time they miss work may vary depending on how the condition is treated and how soon they received treatment. Therefore, knowing the importance and variability between the treatment variables is critical to employers, employees, and all involved in the healthcare industry. I define drug-based treatments with specific National Drug Codes based on their assigned drug classes reported by the FDA (Center for Drug Evaluation and Research, 2020). For treatments derived from medical claims, a combination of different industry-wide standard codes and available groupers were used. For example, many of the treatment codes are based on the procedure code listed on the medical claim, which is referred to as the Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS) (American Medical Association [AMA], 2021). There are approximately 13,500 procedure codes; therefore, I leveraged the Restructured BETOS Classification System (RBCS), which has grouped each procedure code into clinically meaningful categories, subcategories, and families (CMMS, 2020).

Table 4

Treatment Variables

Treatment Variables	Treatment Defined
Visit Types	RBCS Category = E&M
Outpatient	RBCS Subcategory = Office/outpatient services
Emergency	RBCS Subcategory = Emergency department services; also include any claim with Revenue Code Group = "Room & Board" or CMS Place of Service Code = 21, 31, or 61
Inpatient	RBCS Subcategory = Hospital inpatient services
Critical Care/ICU	RBCS Subcategory = Critical care services; also include any claim with Revenue Code Group = "Intensive Care"

Therapy Types	RBCS Category = Treatment
Physical Therapy	RBCS Subcategory = Physical, occupational, and speech therapy; RBCS Family = PT treatment; also include any claim with Revenue Code = 420, 421, 422, 423, 424, 429
Occupational Therapy	RBCS Subcategory = Physical, occupational, and speech therapy; RBCS Family = Occupational therapy; also include any claim with Revenue Code = 430, 431, 432, 433, 434, 439
Adjustments & Manipulations	RBCS Subcategory = Spinal manipulation
Surgical Procedure Types	RBCS Category = Procedure; RBCS Subcategory - Musculoskeletal
General	RBCS Family = No RBCS Family
Arthrodesis Spine	RBCS Family = Arthrodesis Spine
Arthroplasty	RBCS Family = Arthroplasty - hip, Arthroplasty - knee
Arthroscopy	RBCS Family = Arthroscopy - lower extremity, Arthroscopy - upper extremity
Destruction by Neurolytic Agent	RBCS Family = Destruction by neurolytic agent - back
Joint Injection	RBCS Family = Joint injection
Laminotomy/Laminectomy	RBCS Family = Laminotomy or Laminectomy - Lumbar
Nerve Block Injection	RBCS Family = Nerve block injection - back
Neurostimulator	RBCS Family = Neurostimulator - back
Percutaneous Vertebroplasty	RBCS Family = Percutaneous Vertebroplasty
Global Surgical Package	Centers for Medicare & Medicaid Services, 2021g
Minor Procedure (0 days)	CPT codes assigned a global period = 000
Minor procedure (10 days)	CPT codes assigned a global period = 010
Major procedure (90 days)	CPT codes assigned a global period = 090
Devices and Durable Medical Equipment	RBCS Category = DME
Prosthetic Orthotic	RBCS Subcategory = Orthotic devices
Medical Surgical Supplies	RBCS Subcategory = Medical/Surgical Supplies, Hospital Beds, Oxygen & Supplies, Other DME, Drugs Administered through DME
Wheelchair	RBCS Subcategory = Wheelchairs
Imaging	RBCS Category = Imaging
X-ray	RBCS Subcategory = Standard X-ray
CT	RBCS Subcategory = CT scan

MR	RBCS Subcategory = MR
Ultrasound	RBCS Subcategory = Ultrasound
Nuclear	RBCS Subcategory = Nuclear
Tests	RBCS Category = Test
General	RBCS Subcategory = General laboratory, Test - Miscellaneous
Anatomic	RBCS Subcategory = Anatomic pathology
Cardiography	RBCS Subcategory = Cardiography
Molecular	RBCS Subcategory = Molecular testing
Neurological	RBCS Subcategory = Neurologic
Pulmonary	RBCS Subcategory = Pulmonary function
Drugs	Center for Drug Evaluation and Research, 2020
NSAIDs	Drug Class = NSAID
Opioids	Drug Class = Opioid (codeine, fentanyl, hydrocodone-acetaminophen, morphine, oxycodone, and oxycodone-acetaminophen)
Transportation for Treatment	RBCS Category = Other
Ambulance (Air & Ground)	RBCS Subcategory = Ambulance; also include any claim with Revenue Code between 540 and 549
Time to Treatment	Indexing Service Date to First Treatment Service Date

Provider Types. There are ten provider-type variables. Each variable is defined based on the combination of the treating provider's CMS specialty and taxonomy corresponding to the provider's NPI stored in the NEPPES database. Once I had the treating provider's combined specialty and taxonomy, I bucketed them into ten logical groups (see Table 5 for examples) based on common provider types that treat musculoskeletal conditions. These individual buckets were derived from the lookup table provided by CMS, which crosswalks the provider's CMS specialty and CMS taxonomy codes (CMMS, 2021e).

Table 5

Provider Type Variables

Provider Type Variables	Provider Specialty & Taxonomy Names
Primary Care	e.g., Family Medicine, Internal Medicine
Physical & Occupational Therapy	e.g., Physical Therapist, Occupational Therapist
Neurological Surgery	e.g., Neurological Surgery
Orthopedic Surgery	e.g., Orthopedic Surgery
Pain Medicine	e.g., Pain Medicine
Physical Medicine & Rehabilitation	e.g., Physical Medicine & Rehabilitation
Podiatrist	e.g., Podiatrist
Chiropractor	e.g., Chiropractor
Non-surgical Specialist	All others binned (e.g., Sports Medicine, Neurology)
General Surgery	e.g., General Surgery

Employee and Job Demographics. Each employee and job demographic variable was defined from the human resources information system dataset. Employee gender for participants had two values (F, M) and was dummy coded (F=0, M=1). Employee age was calculated using their date of birth and the indexing date of service related to the evaluation year in the study. Then, I binned each age into five categories (0-29, 30-39, 40-49, 50-59, and 60+). Employee salary was calculated based on hourly compensation rate multiplied by annual work hours. Then, I binned the employee salary into four categories (0-50,000, 50,000-100,000, 100,000-150,000, and 150,000+).

Each job factor is a categorical variable that must be normalized and synced within each customer's data and between each dataset. For example, the Human Resources Department had many different spellings, abbreviations, and aliases in the

datasets. Therefore, for each functional area, I mapped each to discrete values. Job titles specific to an individual were binned into a broader category. For example, since many executive-level positions are a single unique role (e.g., CEO, CFO, COO, and CIO), I binned them into a single group called management. Job workload, which represents the level of physical activity, is based on the definition from the U.S. Department of Labor (2021), whereas a particular position contains five categories ranging from sedentary to very heavy.

Dependent Variables

Employee Absenteeism. Employee absenteeism is the dependent variable in Study 1. Employee absenteeism is a categorical variable derived from the number of absence hours and was calculated from the human resources and time management system. However, the aim was to include employee absence from work related to musculoskeletal disorders. Therefore, I queried the combined dataset to include time from work related to MSDs and treatments. I attributed the healthcare provided services and treatments for each employee in the sample for all possible health conditions. Mapping all services and treatments for each health condition was critical because I did not want to attribute absence hours to MSDs when they belonged to a different health condition. For example, if I attributed employee absence to depression or childbirth, it would lead to contaminated results. I examined the combined dataset using key attributes (e.g., diagnosis codes, procedures codes, drug codes, and MS-DRG codes). Each service and treatment was attributed to the correct health condition by date and employee.

Next, I mapped the employees' time off work that maps to their MSDs. Since I did not want to attribute employee absence to a health condition unrelated to MSDs, I

flagged employee absence that had been used for an unrelated reason. When there were medical or pharmacy claims for two or more conditions on the same day, I only attributed a fraction of the absence hours equal to the percent of the total paid in claims for that day. Having multiple health conditions on the same day was uncommon for fewer than 3% of the sample participants.

Long-term absence is the primary interest in this study. To better understand long-term absence, there was a need to examine variable importance in different ranges of absence utilization. However, there is no standard definition for the thresholds (Airaksinen et al., 2018; Skorikov et al., 2020; van Hoffen et al., 2020). Therefore, I created two classifications using data analysis (Low = < 120 hours, High = ≥ 120 hours). Musculoskeletal disorder-related absence time episodes for each participant may have a different start date because the indexing diagnosis sets it. The initial diagnosis date can be any start date in the treatment year.

Medical Cost. The dependent variable in Study 2 is medical cost. Medical cost is a continuous variable and is the total dollars paid by the employer for employees with MSDs from the medical and pharmacy claims. These are the transaction-level paid amounts aggregated to the episode of care by condition, including individual treatments across time. For example, suppose an employee is being treated for lower back pain. In that case, the episode for that health condition and employee may have medical costs from primary care, labs, tests, imaging, physical therapy, non-surgical procedures, surgical procedures, and hospitalizations. It does not include out-of-pocket costs that the employee was responsible for paying.

However, since the interest was in high-cost patients, I needed to examine variable importance in different ranges of medical costs. However, the best way to determine thresholds for high-cost patients is still up for debate. For the purpose of defining "high cost," various thresholds (top 5%, 10%, 20%, and 50%) have been utilized (Luo, Li, Lian et al., 2020). Because there is no agreement on defining high cost, I used data analysis and the Pareto principle (also known as the 80-20 rule), which indicates that about 80% of the effects result from 20% of the causes (Pareto, 1897). As a result, the high-cost threshold is defined as: Low = approximately < $3,000, High = approximately ≥ $3,000.

Data Analysis

Data Preparation

I filtered the overall combined dataset to focus the data on employees with MSDs because it contains all billed medical conditions. I generated a dataset appropriate to the study on MSDs. Only records from the combined dataset for full-time employees participating in the employer's medical plan with one or more MSDs were included, limiting the dataset to 10,042 records. The dataset was then filtered to 8,162 items for two key reasons. First, I was only interested in employees who had complete absence data during the study's observation period in the time and attendance system since the dependent variable in Study 1 is employee absence. Second, I filtered all participants who had only regular time (i.e., not sick time) reported on days when the person was in the hospital for inpatient care.

The impact of the various independent variables on employee absenteeism and medical expenditures is the focus of the research. The dependent variable in Study 1 is employee absence, divided into two categories (Low = < 120 hours, High = ≥ 120 hours), with a focus on the minority class (High). Employees who missed less than 120 hours of work attributable to MSDs fall into the first category, Low (N = 6,550). Employees who missed 120 hours or more of work attributed to MSDs fall into the second category, High (N = 1,612).

Employee medical costs is the dependent variable in Study 2, and there are two classes (Low = approximately < $3,000, High = approximately ≥ $3,000) with a focus on the minority class (High). The number of employees who experienced medical costs linked to MSDs falls into the first category, Low (N = 6,509). Additionally, the number of employees who had medical costs related to MSDs falls into the minority class, High (N = 1,653). There was some overlap between the two dependent variables. Still, they are different, crediting the argument that high medical costs do not imply high absenteeism. The number of employees in each group consisted of: high absence hours and high medical cost = 873, high absence hours and low medical cost = 739, low absence hours and high medical cost = 780, and low absence hours and low medical cost = 5,770.

Due to the uneven nature of the dependent variables (see Appendix 2), two balancing nodes were used depending on the prediction model chosen. Delen (2020) suggests balancing the data to avoid anticipated outcomes being skewed toward the most frequent events, implying that if the models are not adequately balanced, they may reward employees with short-term absence, dubbed “Fools Gold.” That is, you will frequently obtain a better overall accuracy for the majority class but a lower precision for

the minority class. As such, the data were balanced because this study was mainly concerned with the minority population (high medical cost and high absenteeism). Equal Size Sampling was used for models that required categorical data, such as Decision Trees, Random Forest, Naive Bayes, and Gradient Boosted Trees.

In contrast, the Synthetic Minority Oversampling Technique (SMOTE) was used for models that required numeric data, such as Logistic Regression and Neural Network–MLP. I configured the SMOTE so that it oversampled the minority classes. The objective was to avoid biased prediction models that focused exclusively on the majority class.

Study 1 and 2 used approximately 200 independent variables (see Appendix B). For each independent variable, I ran and evaluated each variable's distributions, averages, ranges, and medians within KNIME. To avoid redundant levels, I used a combination of frequencies and business logic (*combining similar levels into similar groups*) to group each independent variable into five or fewer categories. There were three exceptions, employees' age, job title, and department. When binned in five or fewer classes, I lost the specificity needed to understand the impact of each variable in the tested models.

I partitioned the final dataset for training, calibrating, and testing all prediction models. To begin, I utilized the *k*-Fold with ten validations (as illustrated in Figure 5). The purpose of running each model with the *k*-Fold was to reduce the bias introduced by random sampling of the training data. Bias is an error, while variance is an inconsistency. I wanted the models to be low in bias and variance, but they behave like a teeter-totter. As one side improves, the other side regresses. Therefore, the *k*-Fold cross-validation approach can be used to reduce sampling bias. This rotational method tests for and

eliminates bias through repetition and stratification. The variable (k) denotes the number of folds in this testable method. The k-Fold cross-validation procedure was chosen mainly to reduce any potential bias in the model.

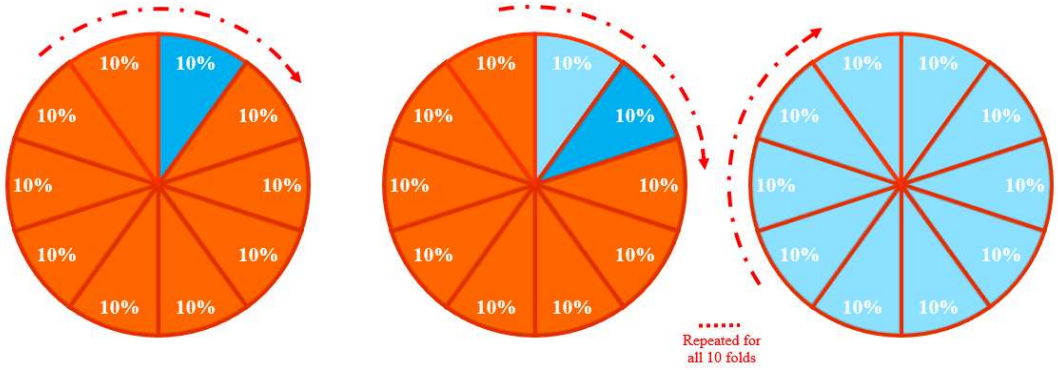


Figure 5. A graphical illustration of the 10-fold cross-validation methodology (Delen, Tomak, Topuz, & Eryarsoy, 2017).

Nominal data types were used for all variables in the model. Therefore, I changed all nominal variables (excluding the dependent variable) to numeric values for the numeric-based models (Logistic Regression and Artificial Neural Networks). I accomplished this operation by utilizing two unique KNIME nodes. First, I began using the "One to Many" node, which converts all potential nominal values to a single column. I then used the "Normalizer" node to normalize the new numeric columns using the min-max setup, transforming the numbers linearly.

Prediction Models

I employed several machine learning techniques—five categorical and three numeric. The categorical models used for my analysis included the Decision Tree, Gradient Boosted Trees, Naïve Bayes, Random Forest, and XGBoost (as illustrated in Figure 6). Three categorical models are ensemble models (Random Forest, Gradient

Boosted Trees, and XGBoost). An ensemble model is a collection of models that have been merged to yield a single prediction outcome. Ensemble models have also been shown to improve the models' resilience, stability, and reliability (Delen, 2020). Additionally, research conducted over the last two decades has demonstrated that ensembles almost always improve predictive accuracy for a given problem and rarely predict worse than a single model (Abbott, 2014).

Numerical models included in this research are K Nearest Neighbor, Logistic Regression, and Artificial Neural Networks. A partitioning node was used for the five categorical models, whereas "one to many" and normalizer nodes were required to convert the independent variables to numeric coefficients for the numerical models to perform correctly. Each model was evaluated using a stratified *k*-Fold with ten cross-validations methods. In addition, confusion matrixes and ROC curves were examined for each model. The following sections summarize various prediction (i.e., classification) methodologies and their precise specifications for this dissertation.

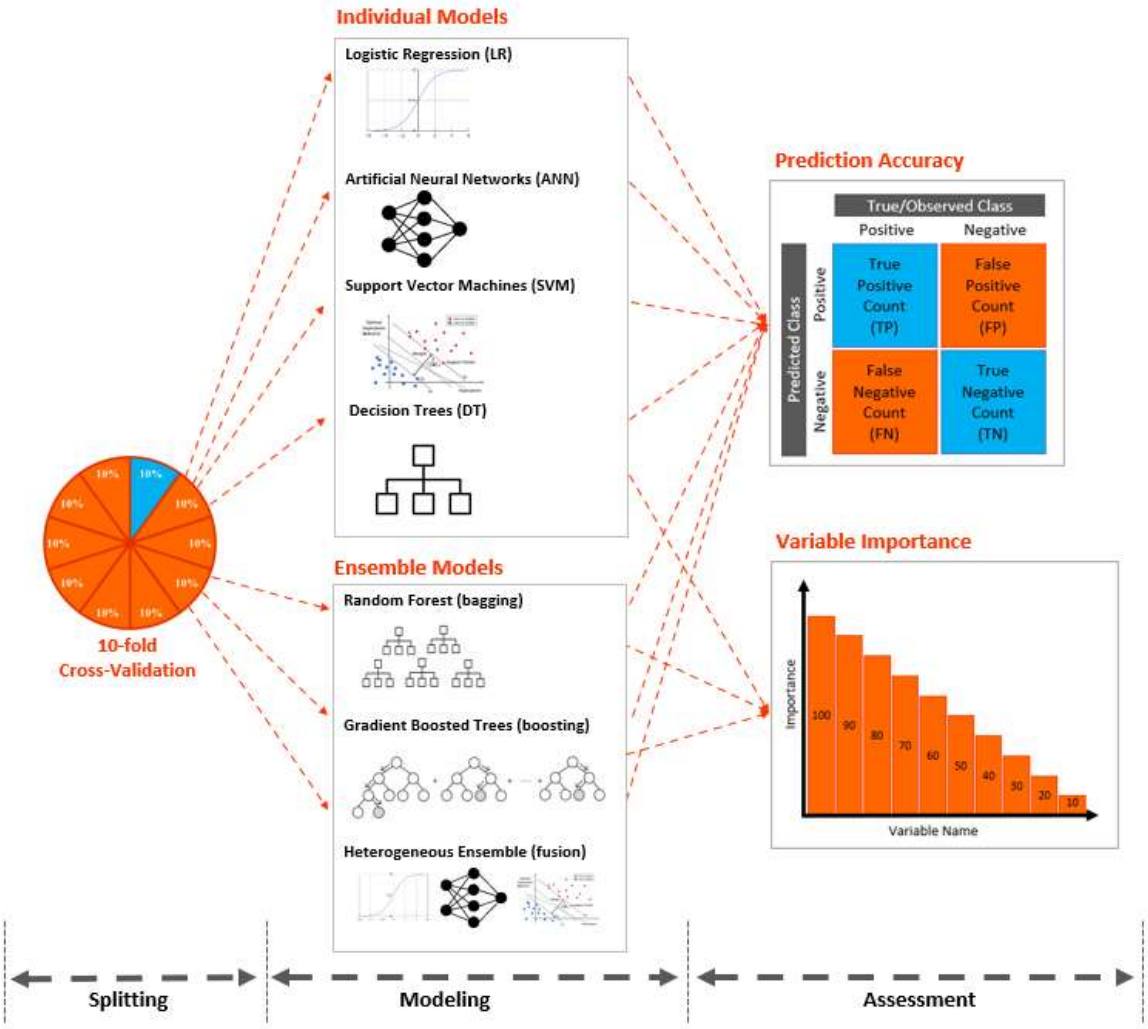


Figure 6. A graphical illustration of the three of four phases (splitting, modeling, and assessment) (Adapted from Delen, 2020).

Artificial Neural Network – Multi-Layer Perceptron (MLP). The first model tested was a machine learning classification technique called a Neural Network (ANN). Neural networks have been around since the early 1960s and were termed after resembling the biological human brain's neural network. Neural networks use artificial intelligence (AI) to create linear relationships and connections between the variables to predict outcomes better. After conducting a process termed "learning" from historical

data, predictive models constructed with neural networks can predict the result of new observations/cases based on the patterns recorded from previous observations/cases (Haykin, 2008). Neural networks have a reputation for being more accurate in prediction when a model contains many variables; however, they are less transparent, slower, and tend to be limited to smaller datasets.

This study used the multi-layer perceptron (MLP) with a back-propagation supervised learning algorithm, a feedforward ANN class. The MLP is a collection of processing elements organized in multiple layers. The input layer obtains the signal and passes it on to the next hidden layer, potentially to another hidden layer, and finally to the output layer. The output signal is then compared to the actual observation, and the error/difference is given back to the network as part of the learning process. Prior research demonstrated that an MLP model could learn highly complicated non-linear correlations to optimal accuracy levels if designed optimally with suitable model parameter values (Hornik, Stinchcombe, & White, 1990).

Decision Tree. Decision trees, the first classification model, in some form, have been used for various decisions since the 1930s. Because decision trees are relatively transparent, their fields may be easily retrieved and adapted for use in rule-based information systems. Decision trees are used to segment data in a determined series of stages. First, the decision tree selects the best independent variable and uses it to split the data. This procedure was done with additional independent variables until the data were split into numerous branches. KNIME software enabled easy visualization of decision trees in action. Then, based on the values of the independent variables, a decision tree model classified the data into an infinite number of groups. Finally, it separates them

using a hierarchical listing of if-then statements based on an index indicating the goodness of the split.

In decision trees, the Gini Index determines how well the model splits the data (Sharda, Delen, & Turban, 2018). The Gini Index was utilized to determine the goodness of the split in the proposed model since it works well with both category and numeric data. The splits created by the Gini Index are referred to as branches, while the final node is referred to as a leaf node. The decision tree's first two stages are depicted graphically in Appendix 3. The view depicted is only a sample view; the model's actual tree view descends numerous levels depending on its current pruning and node limit settings. Different decision tree algorithms are used in research and practice (e.g., ID3, C4.5, C5, CART) (Breiman, Friedman, Olshen et al., 1984; Quinlan, 1986). As a decision tree prediction approach, I employed the C5 algorithm (an upgraded variant of C4.5 and ID3).

Gradient Boosted Trees and XGBoost. The third and fourth models I evaluated were Gradient Boosted Trees, a boosting decision tree ensemble model, and XGBoost, an upgraded version. The boosting ensemble was first introduced by researchers Freund and Schapire (1996), who created the AdaBoost boosting prediction method, which utilizes a weighting procedure. They begin with a simple classification model that is just required to be slightly more than 50% accurate in its prediction. It then analyzes the findings, weights the correct predictions equally or less than the incorrect predictions, and weights the errors. This informs the model to evaluate the misclassified records (i.e., boosted) thoroughly. This method can be repeated hundreds of times with weighted averages used to make final forecasts.

Unlike bagging algorithms, which rely on bootstrapping, boosting algorithms use all available training data. Additionally, unlike bagging, which generates independent modes, boosting is a set of dependent models that learn from one another, explicitly weighing the errors to focus more strongly on them to enhance prediction. Because boosting is error-focused, it works best with inexperienced learners or simple models. Boosting, on average, yields more accurate models than single decision trees and bagging prediction models.

Jerry Friedman of Stanford University invented the stochastic gradient boosting technique, which has been demonstrated to perform well. He also upgraded the AdaBoost algorithm, which he originally called multiple additive regression trees (MART) but has since been renamed TreeNet (Friedman, 2001). These methods construct small trees (about six nodes) and then join them. After the initial tree is constructed, the residuals are computed, and subsequent trees use the residuals as a target variable. The trees will then look for patterns connecting the inputs to the small and large errors. The benefit of this is that each tree will examine the more significant errors created by poor prediction and utilize them to improve prediction in the subsequent tree, while the small errors operate similarly. This becomes an added benefit when working with incompletely cleansed data. The slight inaccuracies or accurate predictions are then used in the subsequent tree. Following the construction of all the trees (which often number in the hundreds), a final prediction is made using an approach that is an additive combination of the outcomes (Abbott, 2014).

In addition, I put XGBoost, a newer version of the algorithm, to the test. XGBoost has emerged as one of the most popular machine learning approaches, with significant

success in machine learning and data mining challenges. XGBoost has achieved success due to its scalability, which was achieved through the system and algorithmic optimizations. XGBoost, for example, addresses sparse data handling, provides a theoretically justified weighted quantile sketch for an efficient proposal calculation, a sparsity-aware algorithm for parallel tree learning, and a cache-aware block structure for out-of-core tree learning (Chen & Guestrin, 2016).

***K*-Nearest Neighbor.** The fifth model examined was the *k*-nearest neighbors (k-NN), a basic non-parametric supervised learning technique (Fix & Hodges, 1951). It was developed initially to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to calculate. Since its inception, many improvements to its features have been made, including rejection approaches, error rate, distance weight approaches, soft computing, and fuzzy methods (Peterson, 2009).

K-NN is a supervised machine learning technique capable of solving both classification and regression prediction problems. K-NN is a form of the lazy learner algorithm that does not instantly learn from the training set, instead deferring all calculations until the actual prediction. K-NN works by determining the similarity (closeness) between the new case and the available cases, picking the specified number of *k*-nearest neighbors, and voting and placing the new case in the category most similar to the available categories for classification-type prediction. The k-NN algorithm then stores all existing data and uses similarity to classify a new data point. When fresh cases are generated, they can be quickly classified into a well-suited category using the k-NN algorithm (Delen, 2020).

Logistic Regression. The sixth model tested was a statistical classification technique called Logistical Regression which was one of the most popular methods found in my literature review. The logistic regression model was created in the 1940s and makes predictions based on the mathematical relationship between the dependent and independent variables. A regression looks at the effect of an independent variable on a dependent variable by regressing the dependent variables on the independent variables. I chose logistical regression versus another regression type because the dependent variable is categorical. Logistic regression is designed to predict binary dependent variables, but its extended version can also handle multi-class classification problems. In addition, logistic regression does not use the least-squares method to model linear functions; instead, it uses a heuristic strategy to represent discrete outputs. Finally, logistical regression makes the following assumptions: the data is normally distributed, and the relationship between the independent and dependent variables is linear. Also relevant to this analysis is that a logistical regression model assumes that variables are not correlated. These few assumptions have helped push machine-learning techniques. However, it was less desirable as machine learning approaches for real-world prediction problems became more capable due to tight assumptions about independence, normalcy, and multicollinearity.

Naïve Bayes. The seventh model tested and one of the most well-known perdition algorithms was a Bayesian classifier called Naïve Bayes. The Bayes classifier uses past incidences to predict future outcomes by classifying predictions based on probable results (Chen, Webb, Liu, & Ma, 2020). For instance, when a new sample is provided for classification, the Bayes classifier will first look for all existing examples that are

identical to it. This is accomplished by identifying all predictor variables with the same values as the classification sample. Second, the Bayes classifier assigns them all to the same class label. Finally, the new sample is classified into the most representative class using the Bayes classifier. Suppose no sample includes an exact match for the new class. In that case, the classifier will reject assigning the sample to a class label due to a lack of solid evidence. It is considered one of the less accurate prediction models and is based on the conditional probability theory (Delen, 2020).

Random Forest. The last model tested I used was a bagging decision tree ensemble model. Bagging, short for bootstrap aggregation, is an approach that aggregates projected values from several decision trees using resampled data (Breiman, 1996). On average, bagging will not utilize 37% of the records in the training dataset (Abbott, 2014). Additionally, it is compatible with classification and regression estimation and prediction models. Typically, between 10 and 25 bootstraps are employed, and it is recommended that when the dependent variable is numerical, fewer bootstraps are required; however, when the dependent variable is classified, as the number of classes rises, the number of bootstraps should increase as well (Breiman, 1996). Finally, bagging is a variance reduction technique that smooths the predictions by averaging them, making the outcomes less variable based on the incoming data.

Random Forest was developed to replace the simple bagging algorithm to increase prediction accuracy (Breiman, 2000). The fundamental difference is in the manner in which the split happens. In Random Forest, each split considers a random subset of variables, unlike the initial bagging technique, which treated all variables as candidates. As a result, you end up with a bootstrapping strategy that incorporates

random case selection and variable selection (Delen, 2020). In contrast to the Decision Tree model, the Random Forest model generates a forest of small trees rather than a single bigger tree. For this study's Random Forest, I set the number of trees in the forest (models) to 1,000.

Like a decision tree, another feature of the Random Forest model, similar to the decision tree model, may be used to determine the significance of the predictor variables employed in the model. For example, the Random Forest finds that three variables (Variable1, Variable2, and Variable3) are the most important and superior at predicting long-term employee absenteeism. Also, Random Forest has consistently outperformed both simple bagging and simple boosting (AdaBoost) strategies in terms of prediction accuracy.

Testing and Evaluation

Each model was evaluated using a stratified *k*-Fold with ten cross-validation methods. Also, confusion matrixes (see Figure 7) and ROC curves were examined for each model.

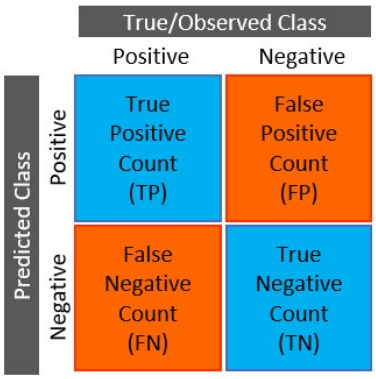


Figure 7. A graphical illustration of a Confusion Matrix (Adapted from Delen, 2020).

Model performance is commonly evaluated in balanced datasets using the accuracy metric (see Equation 1). However, Study 1 and 2 were focused on minority classes (long-term absence and high medical costs); thus, using the accuracy metric may have been misleading in evaluating model performance. For example, if the minority class represents 15% of the population, the model's accuracy may be as high as 85% without correctly predicting any observations in the minority class. Therefore, it is recommended that more appropriate evaluation metrics are used, which include sensitivity (see Equation 2), specificity (see Equation 3), and AUC (Chawla, 2009). Finally, I reviewed variable performance to understand which factors were most likely to influence the outcome variables.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+F} \quad (1)$$

$$\text{Sensitivity (True Positive Rate)} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity (True Negative Rate)} = \frac{TN}{TN+F} \quad (3)$$

The first step in evaluating model performance was examining the above three metrics: Eq. 1. Accuracy, Eq. 2. Sensitivity, and Eq. 3. Specificity. The variables True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) were used to calculate the performance metric results. Accuracy assesses how effectively the model predictions perform to determine the overall likelihood of accurate predicting performance. The True Positive Rate (TPR) and True Negative Rate (TNR) describe how accurately the model predicts minority and majority classes.

Next, I examined model performance using Area Under the Curve (AUC). The AUC is derived from the Receiver Operating Characteristics (ROC) curve, which evaluates classifier performance across a wide variety of class distributions and error costs and has been empirically demonstrated to be a more accurate measure for machine learning applications (Bradley, 1997; Ling et al., 2003). Furthermore, assessing the AUC on a ROC chart is a more straightforward way of evaluating models and comparing different classifiers. The AUC ranges from 0.0 to 1.0, with a baseline of 0.5, implying that classifiers with an AUC less than 0.5 outperform a random guess. An AUC of 0.5 to 0.7 is considered poor, 0.7 to 0.8 is considered good, 0.8 to 0.9 is considered excellent, and greater than 0.9 is considered noteworthy. The higher the classifier's performance, the closer its AUC is to 1.0 (Hosmer et al., 2013).

The ROC chart (see Figure 8) illustrates the trade-off between sensitivity (TPR and specificity by changing the decision cut-off. For example, the ROC curve A has a better AUC and thus better classification performance than B (whose AUC is depicted in blue). However, C is the baseline, showing the ROC curve performing no better than random chance. Therefore, the closer the charted line in the ROC chart is to the top-left point, the better the classifier's performance.

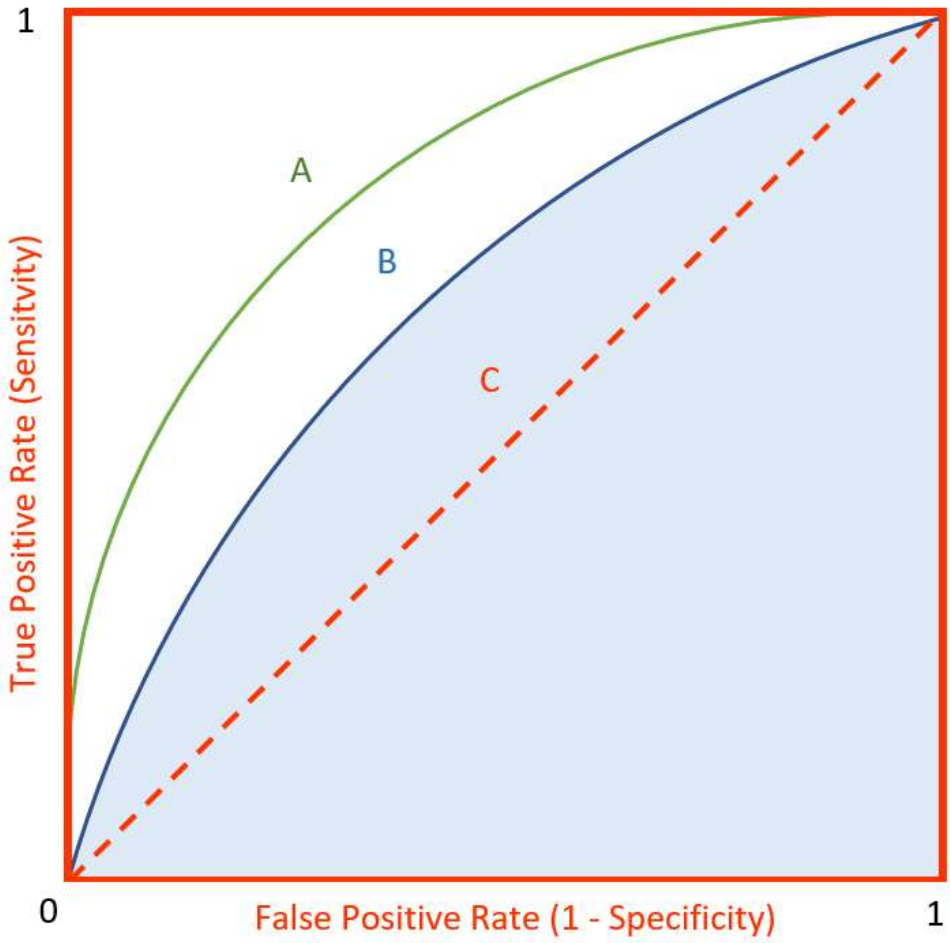


Figure 8. A graphical illustration of a ROC chart (Adapted from Delen, 2020).

A primary goal was to identify and prioritize the critical risk factors associated with long-term employee absenteeism and high medical costs due to poor health. Since the mechanisms for many machine learning algorithms are difficult for human brains to comprehend (black box), I opted to use the actual splitting rate (ASR) as a Random Forest-based heuristic method to determine the order of variables' importance. In essence, the ASR calculates a variable's importance by computing the ratio of actual splits on that variable to the number of times that variable is detected as a candidate for splitting inside the forest. In addition, my model computes 10,000 trees to eliminate the

Random Forest algorithm's potential for bias. This technique made sure that all variables suitable for splitting were appropriately split. Chapter IV presents the findings/results of this study.

CHAPTER IV

RESULTS

The fundamental goal of this dissertation was to identify and analyze numerous employee health-related factors and examine how the individual factors influence long-term employee absenteeism. I followed these objectives through two studies. Study 1 focused on predicting and understanding long-term employee absence to provide employers with insights on how to reduce the effect of long-term absenteeism related to the health conditions of their workforces. Study 2 compared the critical variables that predict long-term employee absence to those that predict high medical costs.

Study 1

The process of model training (i.e., constructing), testing (i.e., validating), and importance measurement (i.e., assessing variable importance from the Random Forest heuristic) is depicted below in Figure 9. First, the input data was pre-processed and transformed into a flat file in Microsoft Excel format, as seen on the left side of the figure. Then, as part of the cross-validation technique, the dataset was randomly partitioned into ten mutually exclusive partitions that would be utilized as training and testing sets for the eight prediction models. All models' prediction accuracy and

sensitivity analysis results were gathered and depicted using the aforementioned performance measures, as shown on the right side of the figure.

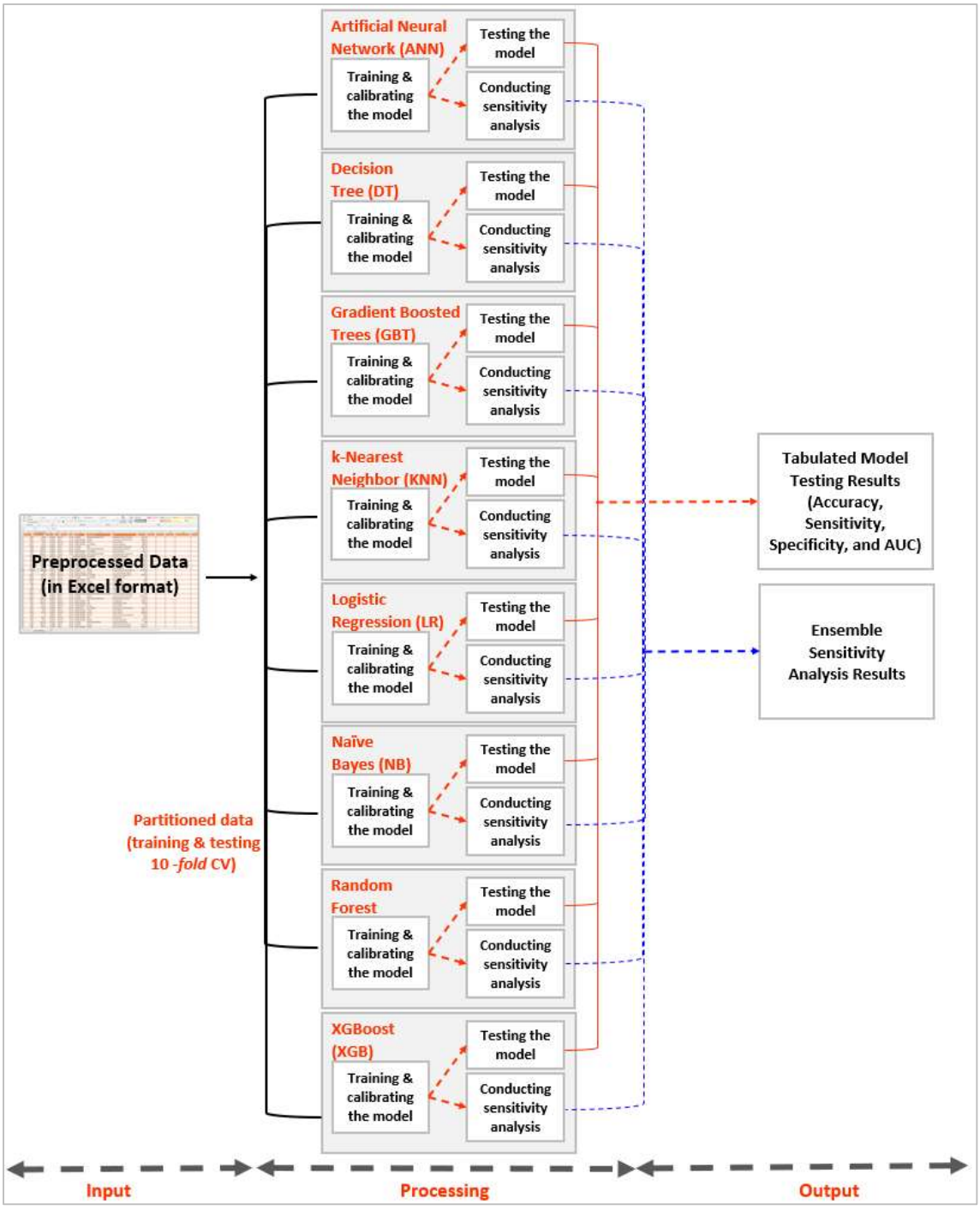


Figure 9. The process of model building, testing, and validation (Adapted from Delen, Tomak, Topuz, & Eryarsoy, 2017).

The prediction accuracies of all eight model types are shown in Table 6. The table explicitly displays the confusion matrices, overall accuracy, sensitivity, specificity, and area under the curve (AUC) of the receiver operating characteristic (ROC) metrics. Notably, the Naïve Bayes was the most accurate classification technique according to the overall accuracy measure (81.51%) and specificity (84.02%). However, the goal was to predict the minority class (high absence hours); thus, I concentrated on the sensitivity and AUC measures.

Table 6

Prediction Results Based on 10-fold Cross-Validation for High Absence Hours

Model Type		**Confusion Matrices**		**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
		Yes	**No**	**(%)**	**(%)**	**(%)**	**(%)**
Artificial Neural Networks (ANN)	Yes	961	651	79.04	59.62	83.82	78.43
	No	1,060	5,490				
Decision Tree	Yes	1,199	413	72.74	74.38	72.34	79.66
	No	1,812	4,738				
Gradient Boosted Trees	Yes	1,278	334	76.7	79.28	76.06	86.08
	No	1,568	4,982				
k-Nearest Neighbor (KNN)	Yes	1,196	416	70.69	74.19	69.83	78.27
	No	1,976	4,574				
Logistic Regression	Yes	1,081	531	72.49	67.06	73.83	76.15
	No	1,714	4,836				
Naïve Bayes	Yes	1,150	462	81.51	71.34	84.02	86.54
	No	1,047	5,503				
Random Forest	Yes	1,387	225	74.86	86.04	72.11	87.20
	No	1,827	4,723				
XGBoost	Yes	1,245	367	76.27	77.23	76.03	85.33
	No	1,570	4,980				

As seen in Table 6, the three homogeneous ensemble prediction models fared the best out of the eight techniques tested when the sensitivity and AUC measures were used. The Random Forest prediction model came out on top, with a sensitivity of more than 86.04% and an AUC of 87.20% (from a possible total of 1.00). Gradient Boosted Trees came in second, having slightly higher assessed sensitivity and AUC than XGBoost. Figure 10 shows the charted AUCs achieved by 10-*fold* cross-validation for each mode type.

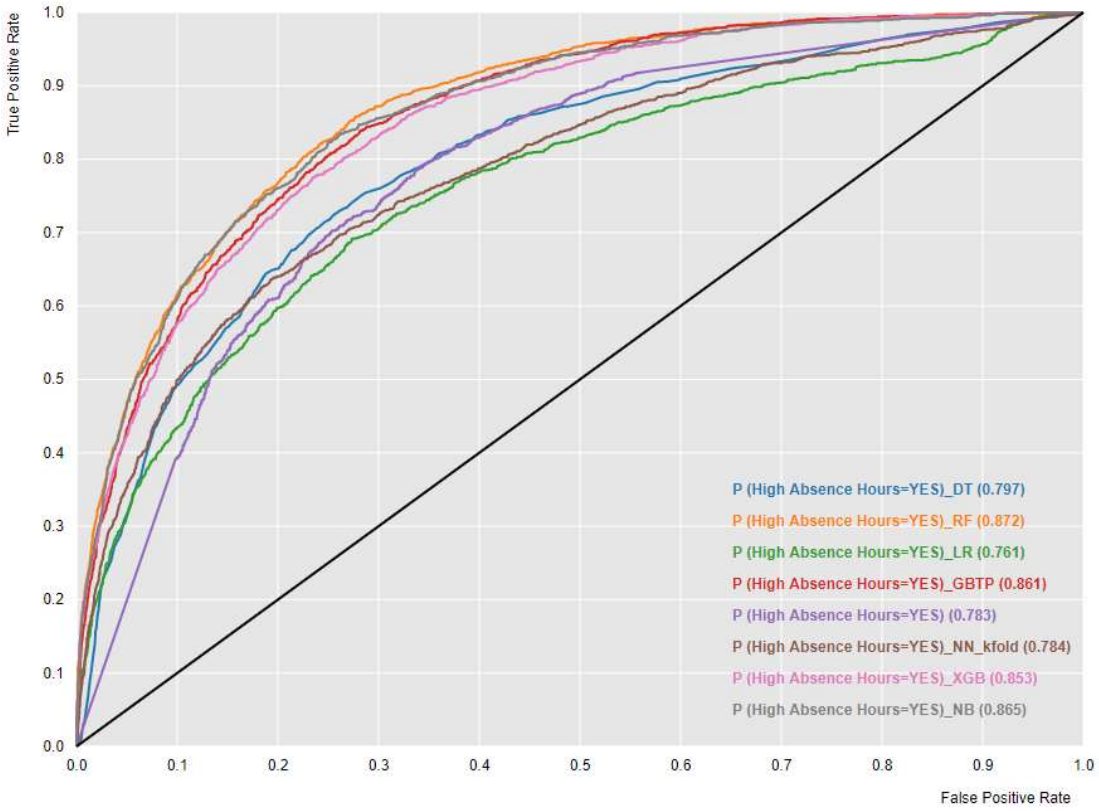


Figure 10. The area under the curve (AUC) of the Receiver Operating Characteristic (ROC) for the eight prediction models.

Although the accuracy measures obtained from all eight model types are sufficient to validate our methodology, the primary goal of this study was to identify and prioritize the significant health-related risk factors influencing employees' long-term absence from musculoskeletal conditions. Therefore, I performed sensitivity analysis on all of the created prediction models to accomplish this. Then, I executed the procedure described in the previous chapter, which focused on the Random Forest variable importance heuristic since it yielded the best sensitivity results.

Due to many variables in the analysis, I reduced the results to the top 30 most critical risk factors (see Figure 11). The top 30 factors reported in the variable importance results suggest four fairly distinct risk groupings, each with four to twelve variables. *Imaging*: X-Ray (how many x-rays did a person receive to manage their condition), *Visits*: Outpatient (how many office visits in an outpatient setting did the person have), *Global Surgical Package*: 90 Days (did the person have a surgical procedure with a 90 day bundled services), and *Number of Musculoskeletal Conditions* were the top groups, in order of importance. According to the Random Forest model's variable importance analysis results, these four risk factors appear to be much more relevant than the rest for predicting long-term employee absenteeism.

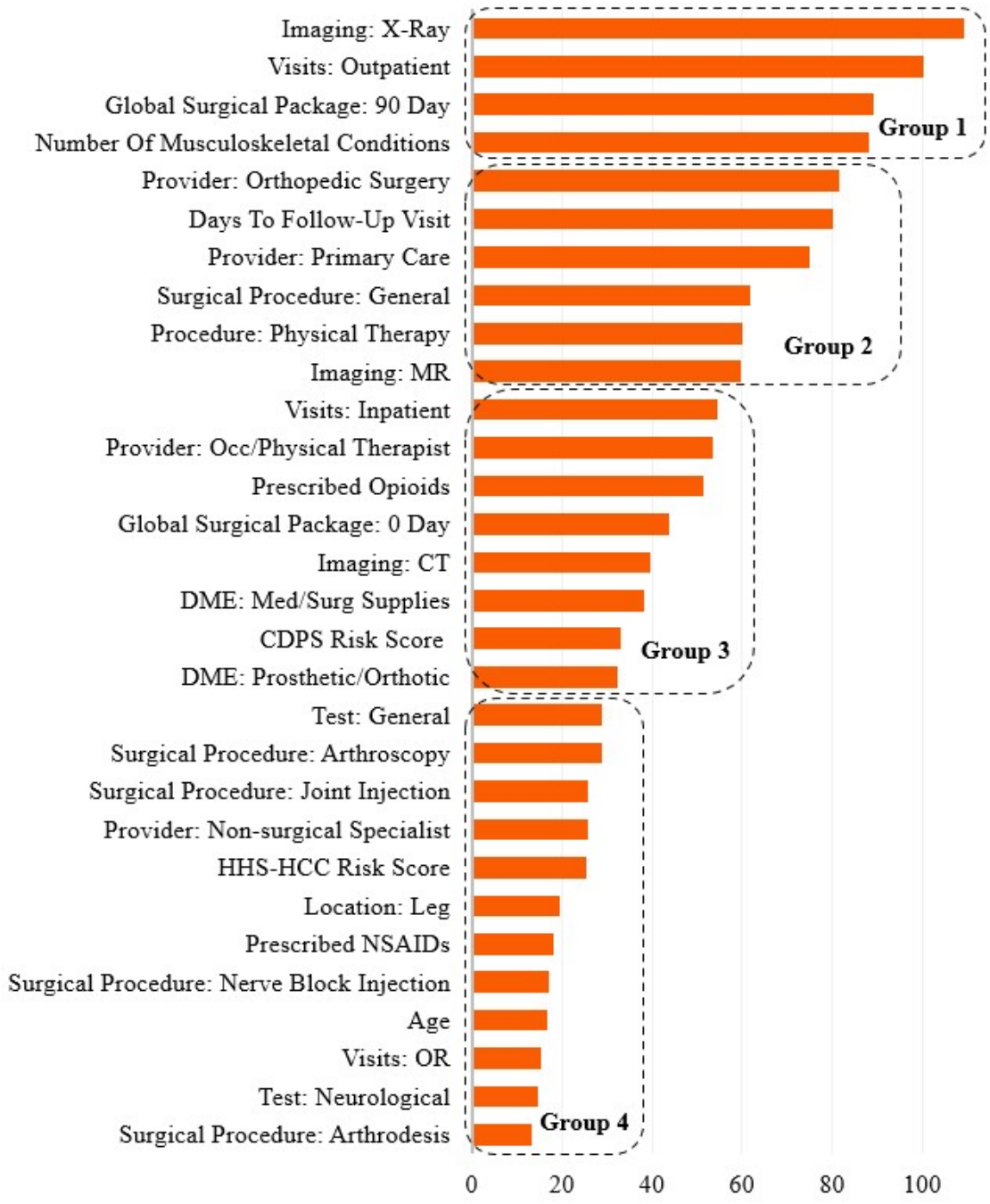


Figure 11. Top 30 normalized variable importance measures for high absence hours.

Next, I further explored the individual factors comprising the first three groupings. Group 1 consists of four variables. The first two (the number of x-ray images and outpatient office visits) are clear leaders in the Random Forest variable importance

heuristic. For example, the number of x-ray images variable has a value of two or more images 67.37% of the time when absence hours are high, but only 26.76% when absence hours are low (see Figure 12).

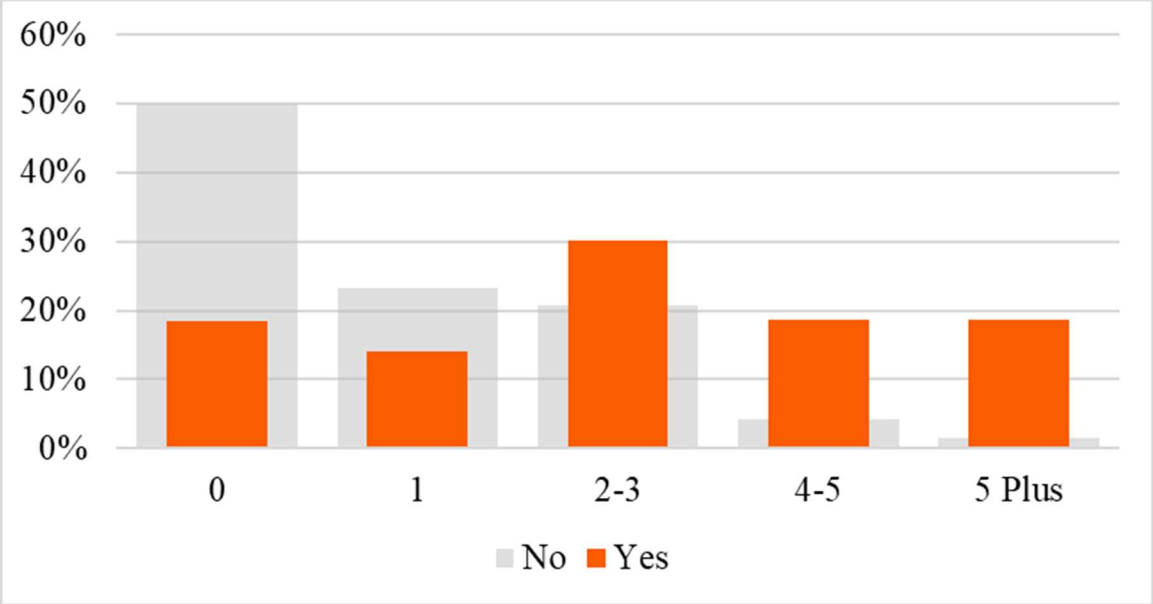


Figure 12. Distribution of absence hours by the number of x-ray images.

In addition, the number of outpatient office visits variable has a value of two or more visits 67.69% of the time when absence hours are high, but when absence hours are low then, only 29.25% had two or more outpatient office visits (see Figure 13).

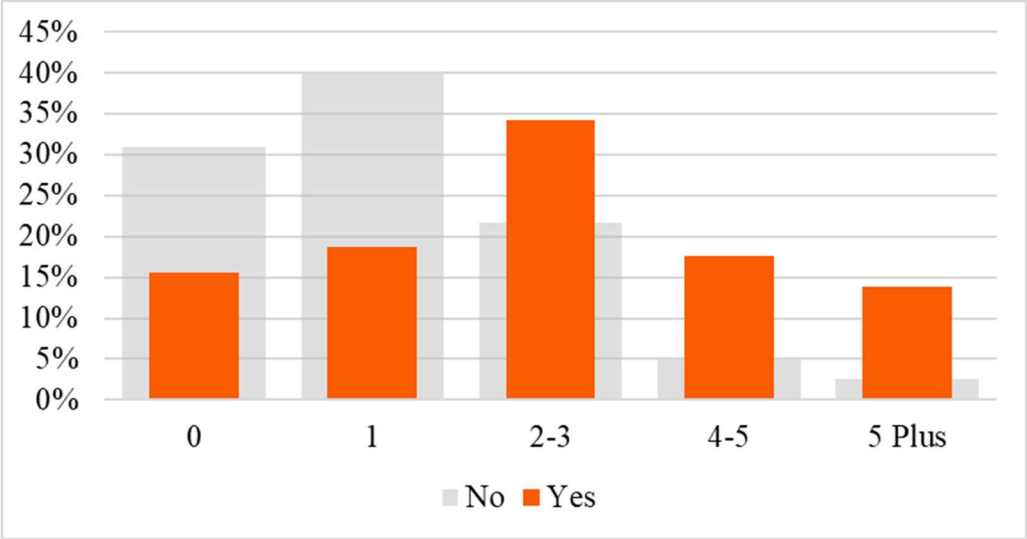


Figure 13. Distribution of absence hours by the number of outpatient visits.

The global 90-day variable is an indicator of surgical treatment utilization. Additionally, it is a proxy indicator of the severity of the treatment received. The global 90-day variable was designed primarily for the CMS National Physician Fee Schedule to cover normal follow-up post-operative care bundled into a global fee (CMMS, 2021g). Therefore, when absence hours are high, the global 90-day variable has a value greater than one 28.78% of the time (see Figure 14). However, when the number of absence hours is low, the variable is only 3.30% when the value is greater than one.

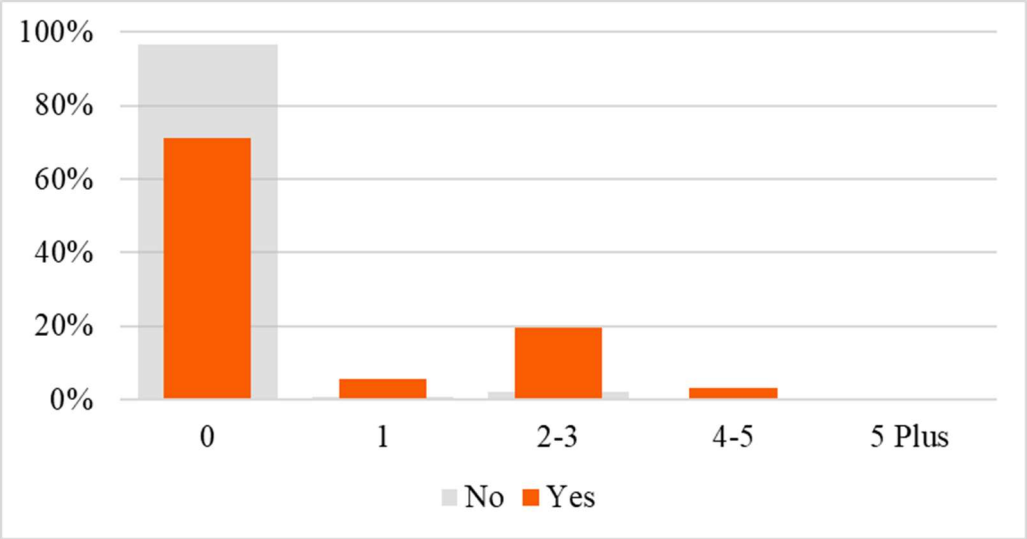


Figure 14. Distribution of absence hours by the global 90-day variable.

The number of musculoskeletal conditions variable has a value greater than one 61.10% of the time when absence hours are high (see Figure 15). However, when the number of absence hours is low, the variable is 24.35% when the value is more than one. In addition, when absence hours are high, 22.70% have three or more musculoskeletal conditions compared to just 4.85% when low. The number of musculoskeletal conditions variable being a top variable in predicting long-term employee absence is important since it can be utilized as an indicator of complexity for the person's health condition and is not accounted for in the HHS-HCC or CDPS risk scoring systems.

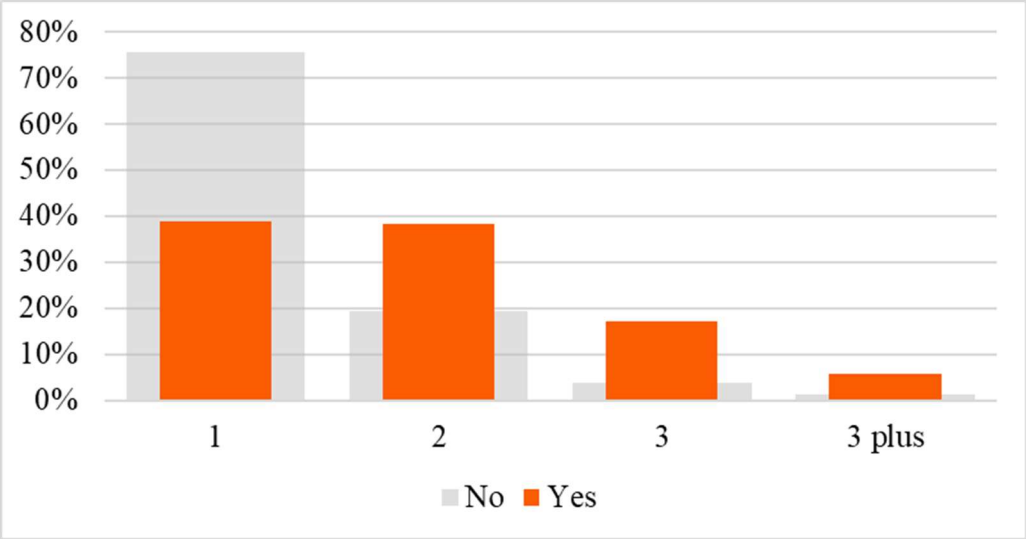


Figure 15. Distribution of absence hours by the number of musculoskeletal conditions.

According to researchers, the length of time a person must wait for care is related to health outcomes, specifically a referral to the initial MSD appointment, which influences employee participation (Lewis et al., 2018; Solomon et al., 1997). However, it has received little attention in the employer health benefit area, despite findings indicating that it is a significant determinant in predicting employee long-term absence. For example, the days to follow-up visit variable has a value greater than 14 days 52.61% of the time when absence hours are high, compared to 35.37% of the time when absence hours are low (see Figure 16). In addition, when absence hours are high, the days to follow-up visit variable has a value greater than 28 plus days 34.68% of the time compared to 20.37 when absence hours are low.

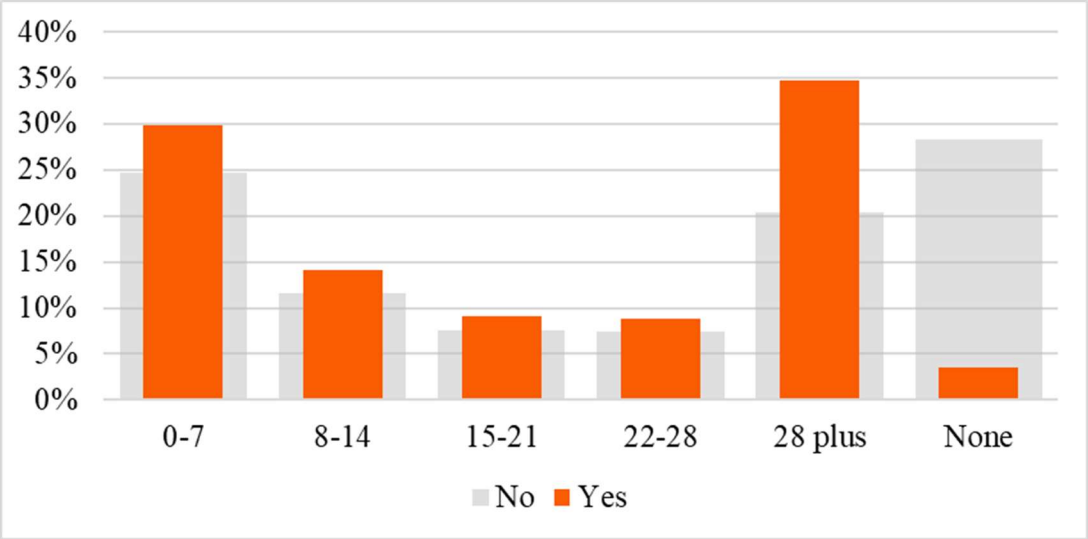


Figure 16. Distribution of absence hours by the days to follow-up visit variable.

The variable importance results show that the type of healthcare provider is essential. For example, when absence hours are high, the orthopedic provider type variable has at least one visit; 51.86% compared to only 23.47% when absence hours are low (see Figure 17). In addition, the number of visits for the orthopedic provider type variable has a value of two or more visits 40.01% of the time when absence hours are high, but only 11.28% when absence hours are low.

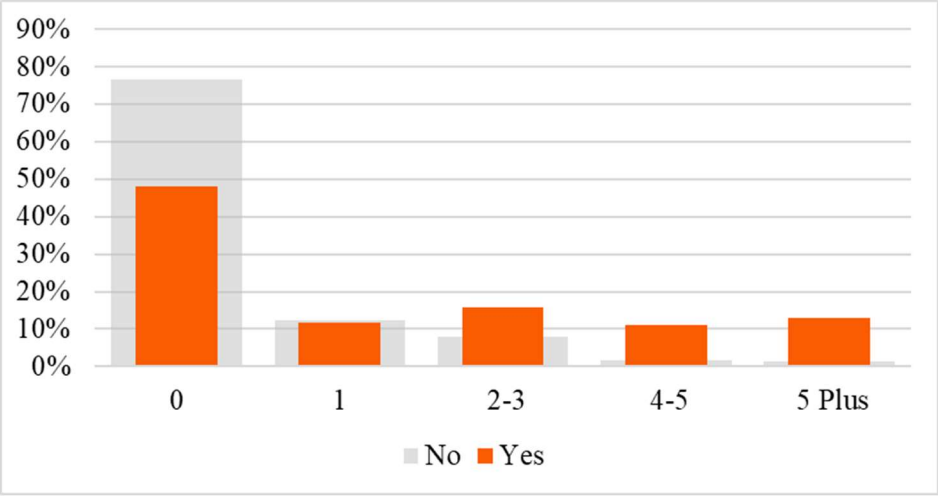


Figure 17. Distribution of absence hours by the orthopedic provider type variable.

The primary care provider type variable has similar findings but at a higher visit rate (see Figure 18). For example, the number of visits for the primary care provider variable has a value of four or more visits 46.28% of the time when absence hours are high. Still, only 17.77% had four or more visits when absence hours were low. These two variables highlight that knowing more about the types of providers is important because they can have different referral and treatment practices.

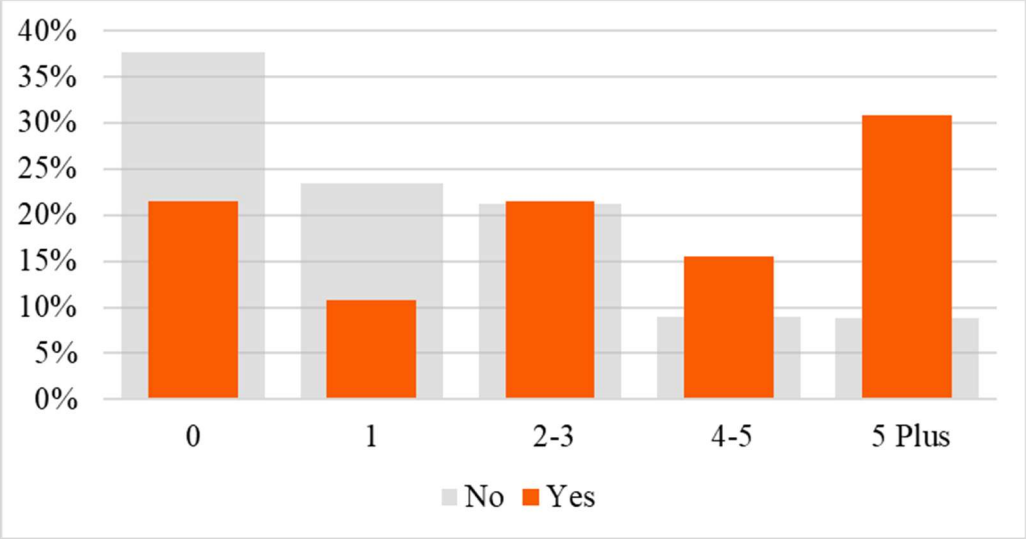


Figure 18. Distribution of absence hours by the primary care provider type variable.

The last three variables of Group 2 are general surgical procedures, physical therapy, and MR imaging. When absence hours are high, the general surgical procedures variable has a value of one or more 30.58% of the time compared to only 8.12% when absence hours are low (see Figure 19). The physical therapy variable has a value greater than zero 51.61% of the time when absence hours are high (see Figure 20). However, when the number of absence hours is low, the variable is 30.38% when the value is more than zero. In addition, when absence hours are high, 35.30% have five or more physical therapy visits compared to just 11.57% when low. The last variable in Group 2 is the MR

imaging variable, and when absence hours are high, it has a value of at least one 36.79% of the time compared to only 11.42% when absence hours are low (see Figure 21).

Each logically makes sense since having a general surgical procedure does not guarantee that an individual will be in the high absence hours cohort compared to someone who does not, but a person usually will need rest and therapy, which greatly improves their chance. In addition, a person who needs physical therapy and how many therapy sessions are both an indicator that there is an increased chance of missing more time from work than someone that does not require it. Also, it is not surprising to see MR imaging to show up at a lower level than x-ray imaging since it is not ordinarily ordered as an initial scan due to many things such as cost to the patient, access, and other imaging modalities such as x-rays and ultrasounds (Dean-Deyle, 2011; Jacobson, 2009).

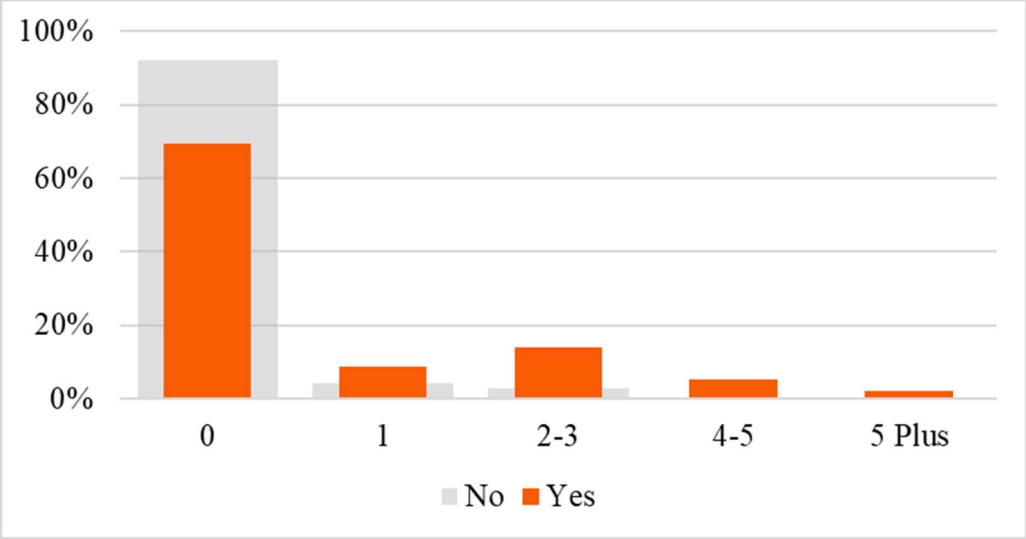


Figure 19. Distribution of absence hours by the general surgical procedure variable.

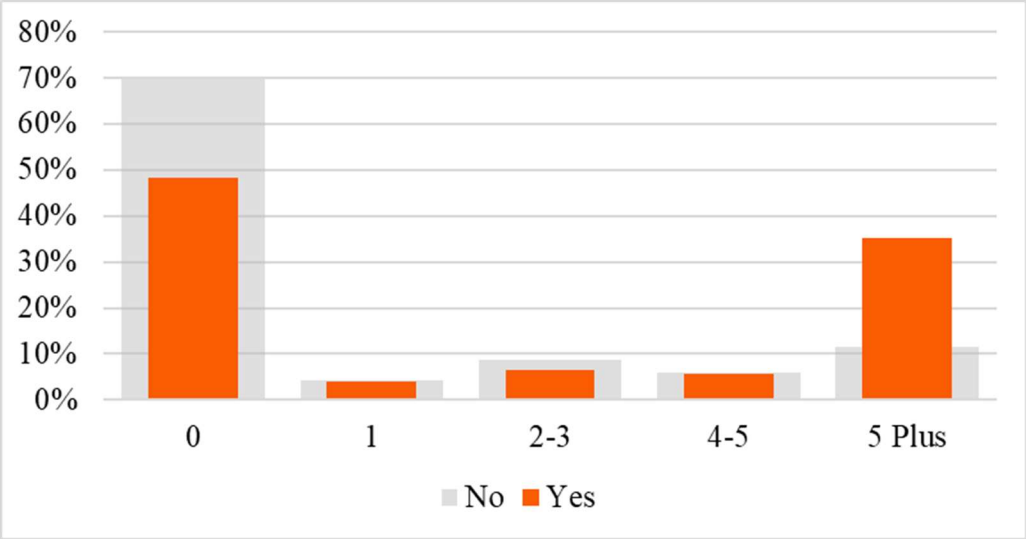


Figure 20. Distribution of absence hours by the physical therapy variable.

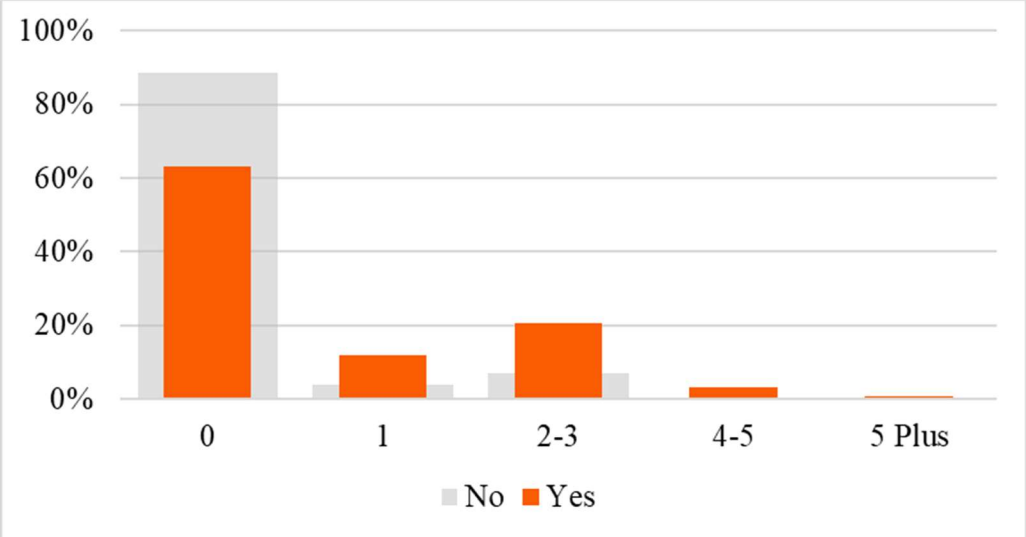


Figure 21. Distribution of absence hours by the MR imaging variable.

Inpatient visits lead off the third group of variables. When absence hours are high, the inpatient visits variable has a value of at least one 17.25% of the time compared to only 2.02% when absence hours are low (see Figure 22). Therefore, I suspected that this variable would be a top predictor. However, I am a bit surprised that it is in Group 3

instead of Group 1 or 2 because a person who gets hospitalized for their health condition would likely miss more time than those that do not require hospitalization.

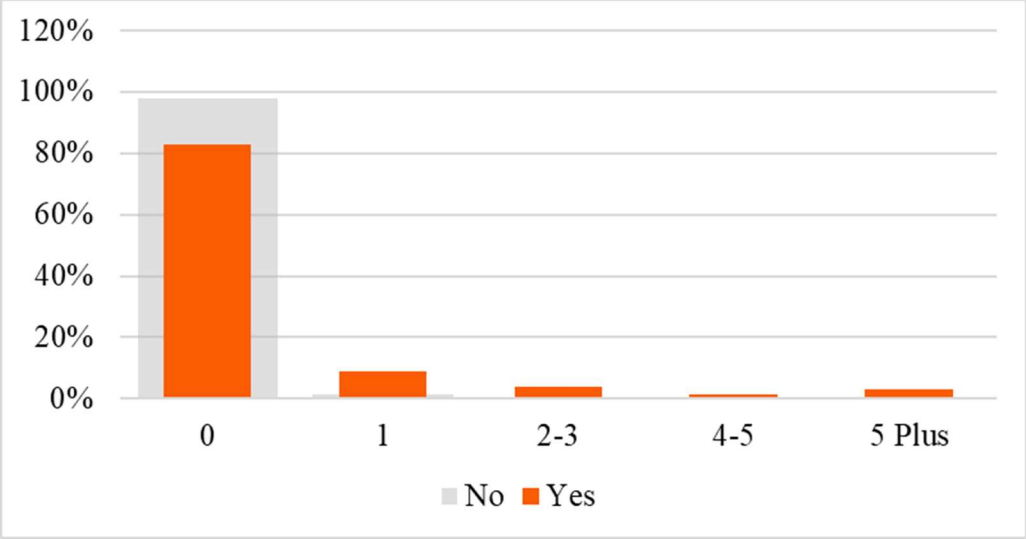


Figure 22. Distribution of absence hours by the inpatient visit variable.

The second variable in Group 3 is the physical and occupational therapy provider type variable. When absence hours are high, the physical and occupational therapy provider type variable has a value of one or more 26.86% of the time compared to only 7.71% when absence hours are low (see Figure 23). In addition, the number of visits for the physical and occupational therapy provider type variable has a value of five or more visits 19.17% of the time when absence hours are high, but only 3.83% when absence hours are low.

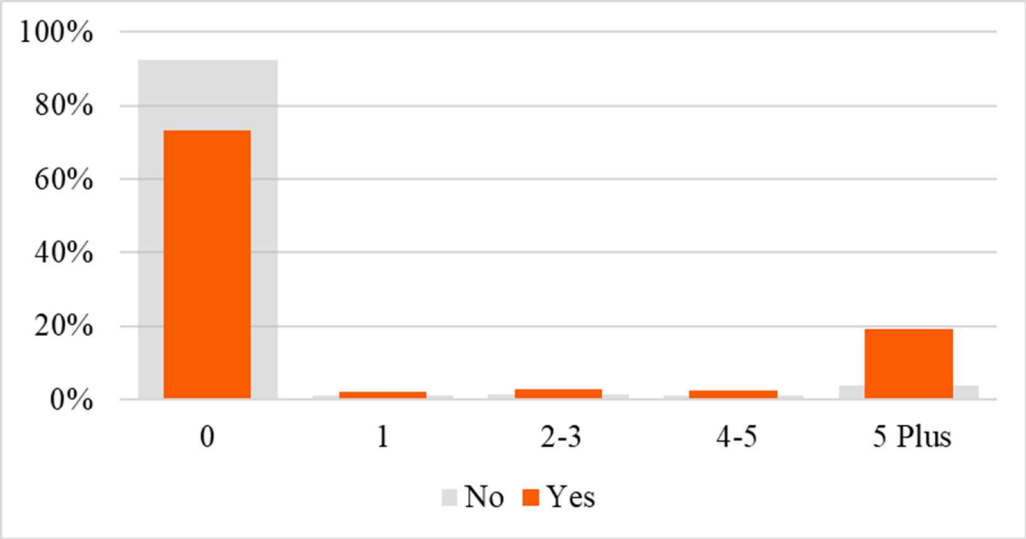


Figure 23. Distribution of absence hours by the physical and occupational therapy provider type variable.

The third variable in Group 3 is the opioid variable which is based on the number of prescriptions for the patient during their episode. When absence hours are high, the opioid variable has a value of one or more 24.75% of the time compared to only 5.04% when absence hours are low (see Figure 24). In addition, the number of prescriptions for the opioid variable has a value of two or more prescriptions 12.16% of the time when absence hours are high, but only 1.82% when absence hours are low. Furthermore, 112 people who were prescribed opioids had high absence hours and no surgical procedure.

The global 0-day variable is the fourth in Group 3 and the fourteenth most important variable in the list. It is related to the third most important overall variable, the global 90-day variable, but it includes treatments that are less severe and have a shorter recovery period on average. When absence hours are high, the global 0-day variable has a value of one or more 46.40% of the time (see Figure 25). However, when the number of absence hours is low, the variable is only 24.17% when the value is one or more.

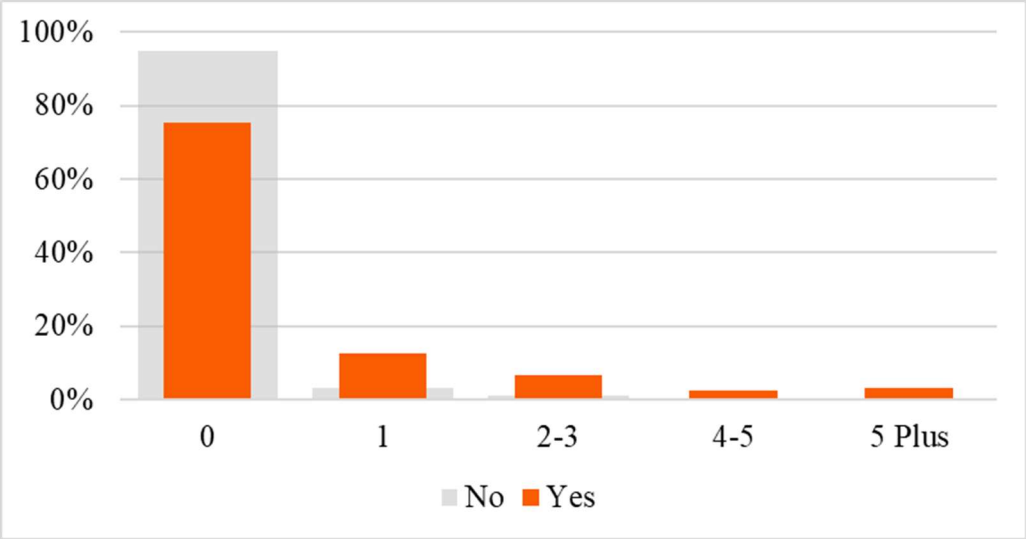


Figure 24. Distribution of absence hours by the opioid variable.

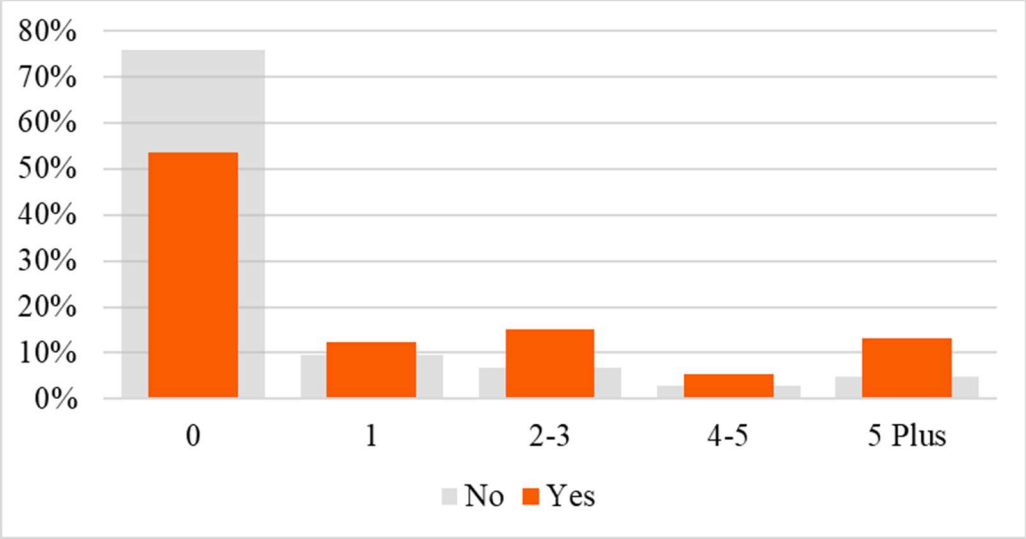


Figure 25. Distribution of absence hours by the global 0-day variable.

Group 3, like Groups 1 and 2, has an imaging variable in the cohort. The fifth variable in Group 3 is the CT imaging variable, and when absence hours are high, it has a value of at least one 17.80% of the time compared to only 4.35% when absence hours are low (see Figure 26). Furthermore, when absence hours were high, the CT variable (N =

287) with one or more people was 21.86% of those who had x-rays (N= 1,313) and 48.40% of those who had MRs (N = 593).

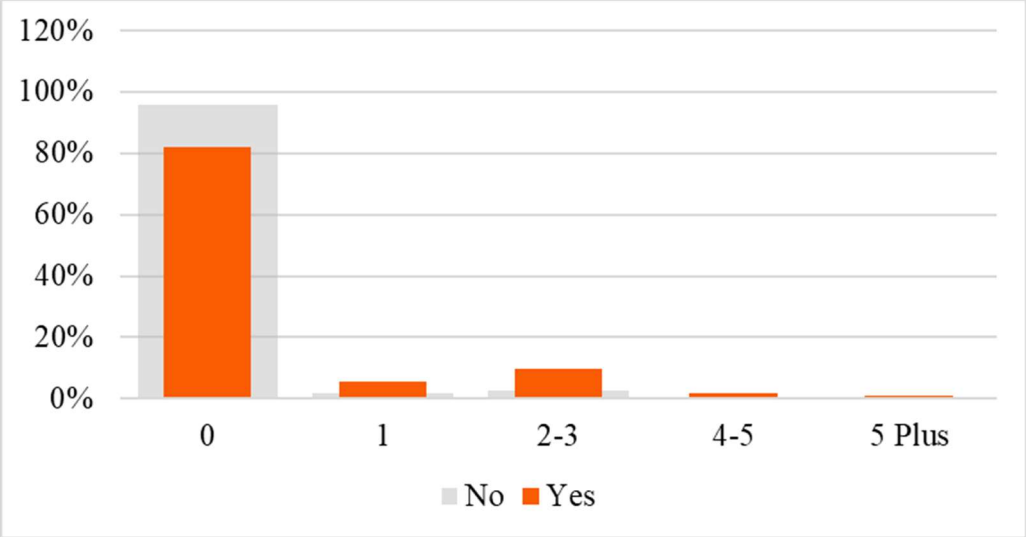


Figure 26. Distribution of absence hours by the CT imaging variable.

Two of the last three variables that make up Group 3 are durable medical equipment (DME) variables—medical and surgical supplies; prosthetic and orthotics. The medical and surgical supplies variable is based on the utilization of supplies for the patient during their episode. When absence hours are high, the medical and surgical supplies variable has a value of one or more 13.15% of the time compared to only 1.66% when absence hours are low (see Figure 27). Furthermore, when absence hours are high, the medical and surgical supplies variable has a value of one or more 20.78% of the time compared to only 7.44% when absence hours are low (see Figure 28).

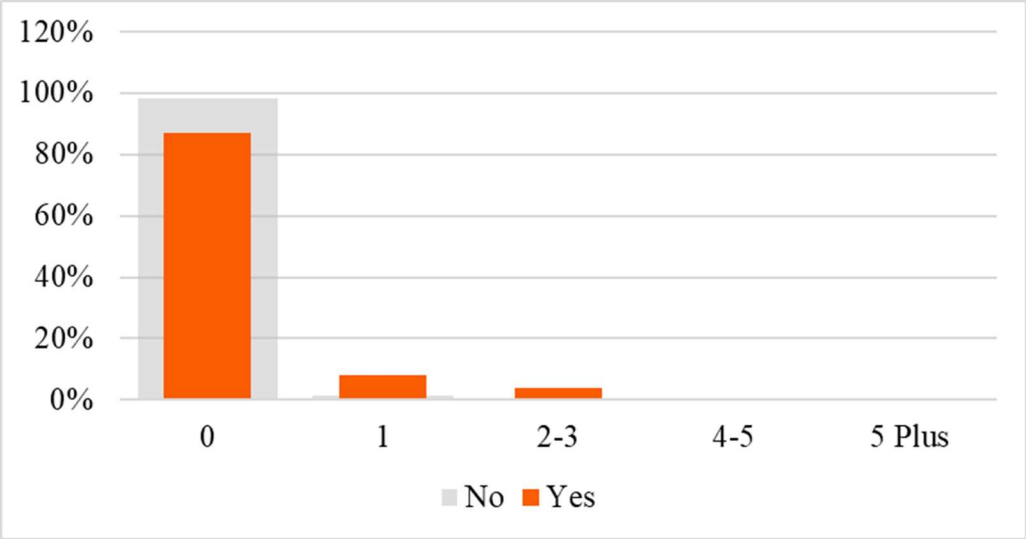


Figure 27. Distribution of absence hours by the durable medical equipment for medical and surgical supplies variable.

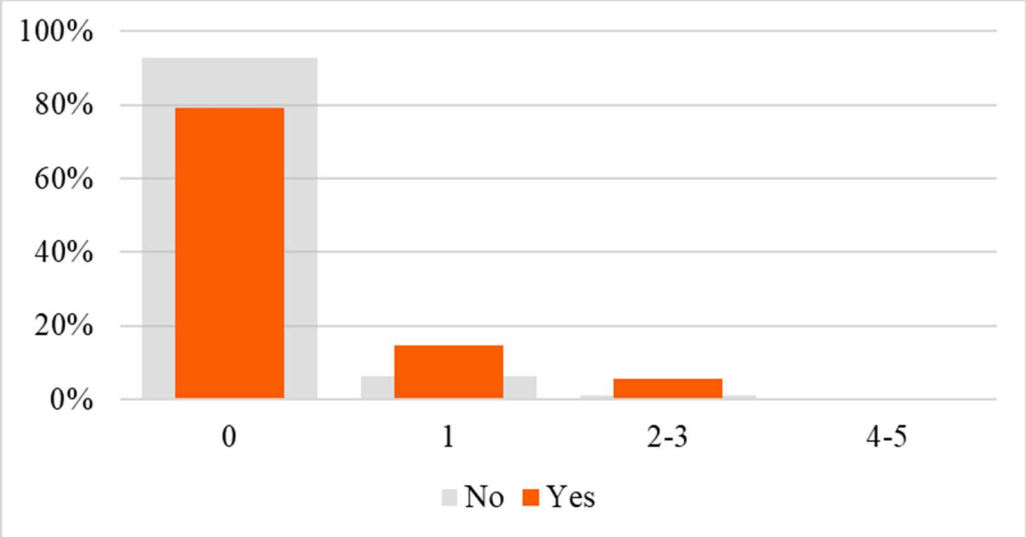


Figure 28. Distribution of absence hours by the durable medical equipment for prosthetic and orthotics variable.

The CDPS risk score variable is the third and last variable in Group 3. The CDPS risk score is an accumulation of coefficients assigned to each diagnostic condition that the individual has (e.g., type 2 diabetes, ischemic heart disease, cystic fibrosis). The

greater the aggregated risk score, the more health issues reported in claims that can be utilized as an indicator of health. As a result, the higher the risk score, the worse their health. When absence hours are high, the CDPS risk score variable has a value of one or more 28.29% of the time compared to only 13.69% when absence hours are low (see Figure 29).

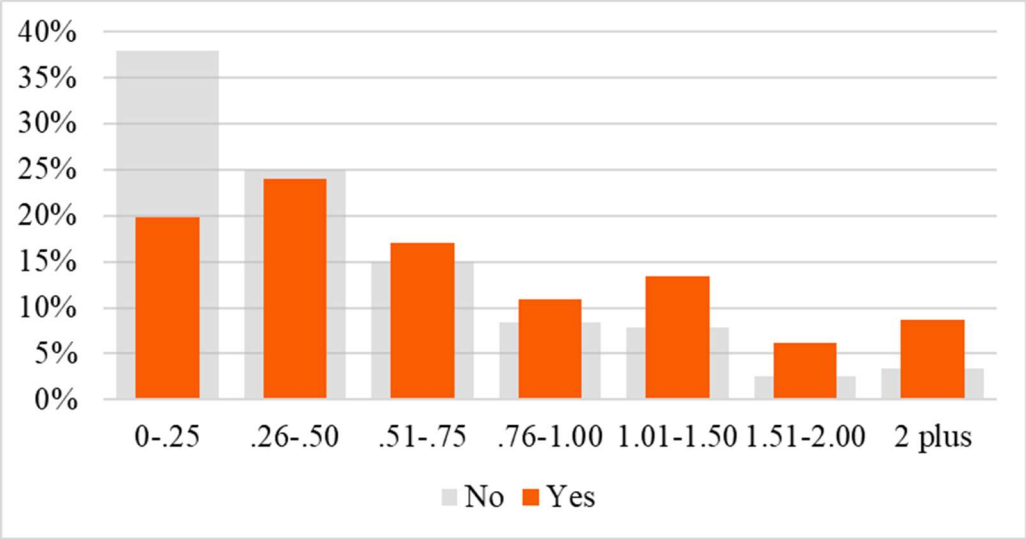


Figure 29. Distribution of absence hours by the CDPS risk score variable.

Study 2

The purpose of Study 2 was to better understand the variables that predict long-term employee absenteeism compared to those that predict high medical costs. I utilized the same method as in Study 1, but this time I solely concentrated on the Random Forest model. First, the Random Forest prediction model outcomes were collected and compared to Study 1. Second, the variable importance results were generated and compared to determine whether the same variables that indicated long-term employee absence were also those that predicted high medical costs. The prediction results for high

medical costs were better when compared to the high absence hours for all measures except sensitivity (see Table 7).

Table 7

Prediction Results Based on 10-fold Cross-Validation for High Medical Costs

Model Type		Confusion Matrices		Accuracy	Sensitivity	Specificity	AUC
		Yes	No	(%)	(%)	(%)	(%)
Random Forest	Yes	1,307	346	88.99	79.07	91.52	93.83
	No	552	5,957				

As can be seen in Table 7, the overall prediction accuracy was 88.99% for high medical costs and 74.86% for high absence hours. The sensitivity analysis was 79.07% for high medical costs and 86.04 for high absence hours. As a result, the model better predicted the minority class for high absence hours. The specificity was 91.52% for high medical costs and 72.11% for high absence hours, which resulted in an AUC of 93.83% for high medical costs and 87.20% for high absence hours.

The variable importance for high medical costs was then computed because the primary purpose of this study was to identify and prioritize the substantial health-related risk variables affecting high medical costs and compare them to the significant risk factors of employees' long-term absence. Therefore, I performed the same variable importance methodology as in Study 1, except I used high medical cost as the dependent variable. Furthermore, I limited the results to the top 30 most critical risk factors.

The results suggest four fairly distinct risk groupings, each with four to twelve variables (see Figure 30). *Imaging*: MR (how many MRIs did a person receive to manage

their condition), *Global Surgical Package*: 90 Days (did the person have a surgical procedure with a 90-day bundled service), *Imaging*: X-Ray (how many x-rays did a person receive to manage their condition), and *Visits*: Outpatient (how many office visits in an outpatient setting did the person have) were the top groups, in order of importance. According to the Random Forest model's variable importance analysis results, these four risk factors appear to be much more relevant than the rest for predicting high medical costs.

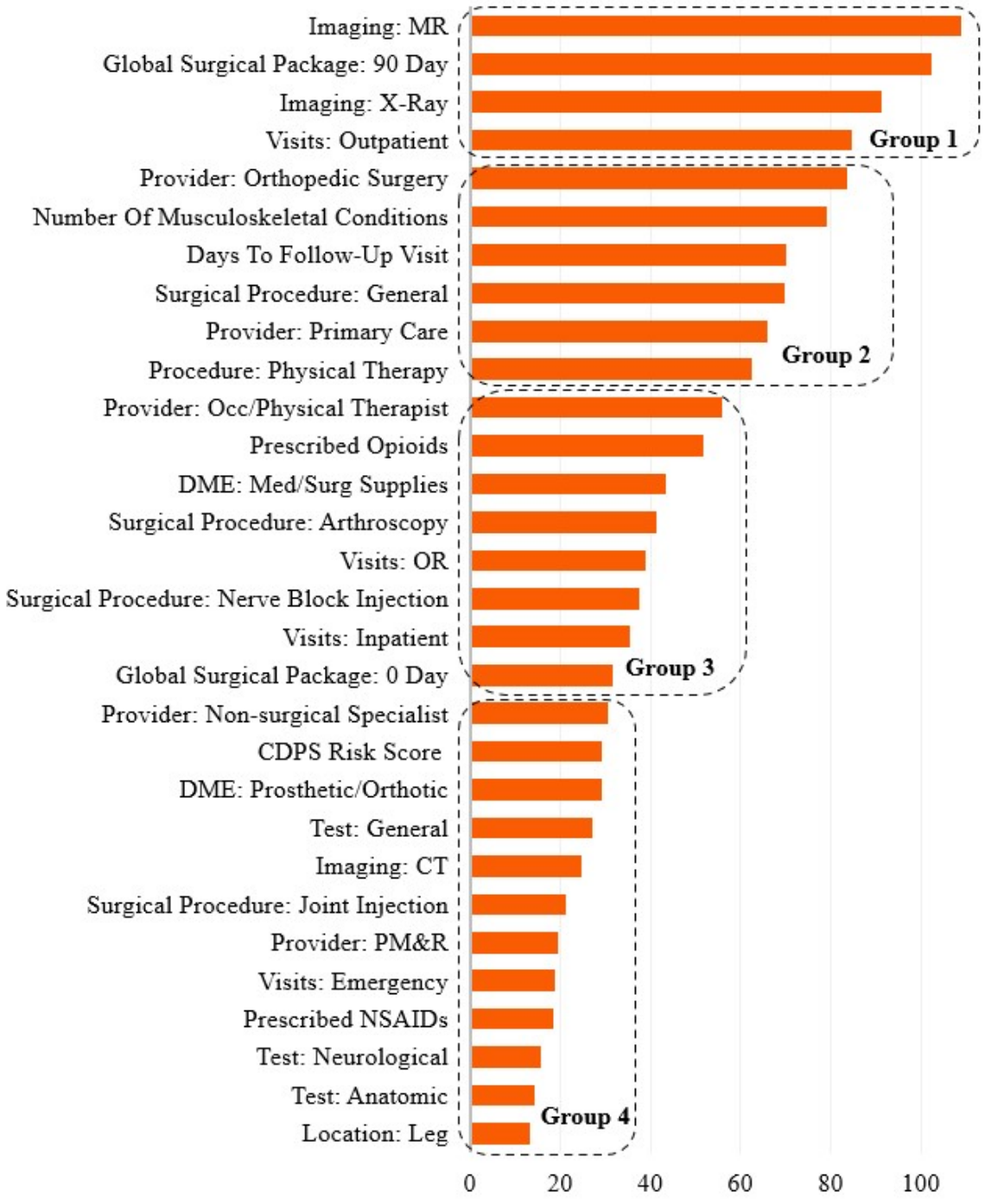


Figure 30. Top 30 normalized variable importance measures for high medical costs.

Last, the variable importance results were compared between Study 1 and Study 2, highlighting that the two dependent variables shared many characteristics. For example, looking at the top 30 factors between the two studies, I observed that high absence hours had three variables (e.g., HHS-HCC Risk Score, Surgical Procedure: Arthrodesis, and Age) that made the top 30 for high absence hours but did not make the top 30 for high medical costs. Furthermore, three variables (e.g., Provider: PM&R, Visits: Emergency, and Test: Anatomic) made the top 30 for high medical costs but did not make the top 30 for high absence hours.

Aside from those six factors, the rest of the top 30 variables overlapped between the two studies, although their importance was ranked differently. Figure 31 shows the variables that account for the top 30 for the two dependent variables, sorted by high medical costs. The difference in variable importance for the predictor variables regarding the dependent variable is indicated on the right side of the chart. If the difference bar is negative, the variable is more important for high absence hours and vice versa.

The first two ranked variables make sense in their variable rank position. For example, the MR imaging variable (ranks 49 points lower and ninth for high absence hours) is commonly more expensive and not ordinarily ordered as an initial scan (Berger & Czypionka, 2021). The second-ranked variable, Global Surgical Package: 90 Days (ranked third in Study 1), is comprised of more complex surgical procedures that are typically more expensive. However, it was surprising not to see inpatient visits in the top group instead of near the bottom of Group 3. Hospitalizations usually carry higher costs (Hessel, 2021; Stull, Bhat, Kane, & Raiken, 2017) for many health issues; however, we

are looking at episodes of care, not specific hospitalizations, contributing to the difference here.

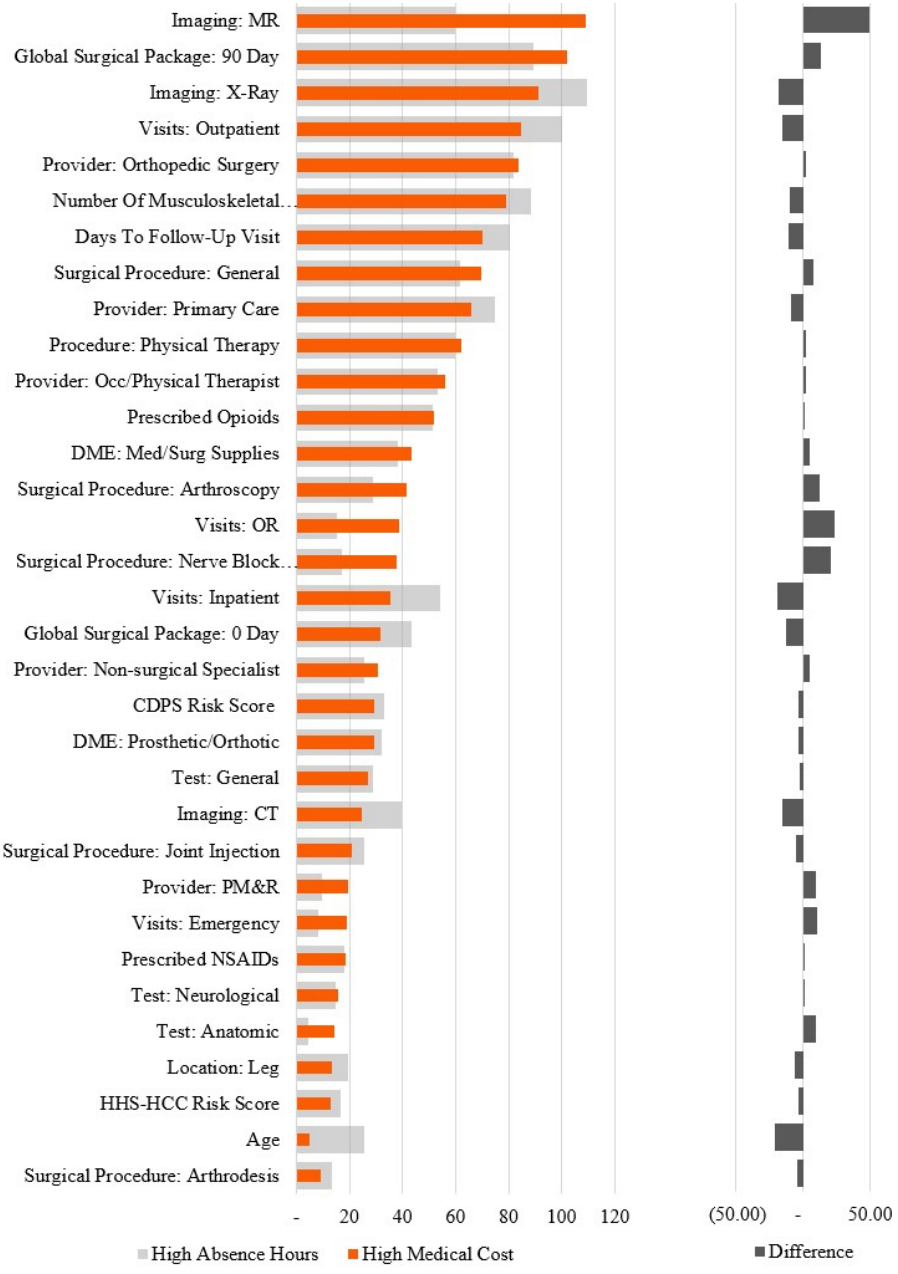


Figure 31. Variable importance differences between high absence hours and high medical costs.

Chapter V concludes with a discussion of this study's purpose and summary of findings.

CHAPTER V

DISCUSSION AND CONCLUSION

Research Summary

Poor health is the primary cause of employee absenteeism (Chadwick-Jones et al., 1982; Hackett et al., 1989; Hedges, 1973, 1975, 1977; Morgan & Herman, 1976; Nicholson & Payne, 1987; Paringer, 1983), accounting for up to two-thirds of all absences (Brooke, 1986; Hedges 1977; Miner & Brewer, 1976). However, most management research has centered on controllable factors associated with five roughly defined groups: personality, demographics, attitudes, social context, and decision-making (Harrison & Martocchio, 1998). This narrow perspective on the causes and correlations has generated important insights, but it has not shown to be a reliable method for predicting employee absence. As a result, I expanded the list of variables in this dissertation to include health factors related to employees' health conditions. For example, what are the individual's pre-existing conditions, what medical or pharmaceutical treatments did they receive for the issue, by who has it been treated, and how long has it taken to receive the treatment they were referred to?

The fundamental goal of this dissertation was to improve management's understanding of how businesses might better understand, predict, and reduce the influence of employee absenteeism due to poor health. Two research questions were used to achieve this goal: (1) How can businesses understand, predict, and mitigate the impact of employee absence due to health issues? (2) What differences exist between the risk factors that predict long-term employee absence and those that predict high medical costs? I studied these questions using data-science approaches, focusing on the leading cause of employee disability, musculoskeletal disorders (MSDs), and how they influence long-term absence and high medical costs. Furthermore, I used *k*-fold cross-validation estimates to explore the influence of eight machine learning models (e.g., Artificial Neural Network, Decision Tree, Gradient Boosted Trees, K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Random Forest, and XGBoost). But first, I created a combined dataset from three key data sources—medical claims, pharmacy claims, and human resources—that many self-funded companies have access to.

To ensure data reliability and compliance with the Cross Industry Standard Process for Data Mining (CRISP-DM) data process and methods, a thorough assessment of the combined dataset utilized critical risk factors related to employees' demographics, job, health, and medical care received from 8,162 full-time employees. Each employee was from one of three large organizations, was a member of the employer's medical and pharmacy benefit programs, and had a diagnosed MSD. I used the combined dataset that was meticulously pre-processed so that all eight prediction models could be evaluated and compared correctly.

Research Contributions

Management research has mainly concentrated on the explanation and relationship of risk factors for employee absenteeism rather than the accuracy of predictions. However, there have been numerous prediction studies outside the management domain (e.g., occupational health and medicine) on long-term employee absence (see Appendix 1), but the highest level of accuracy attained had an AUC of 81%. As a result, the first research contribution is the development of a prediction methodology that improves prediction accuracy, not just for any employee absence but for employee long-term absences related to poor health.

To begin, I used a combined dataset that contained medical and pharmacy claims data. The depth of the included claims data allowed me to build on the limited collection of health indicators used in previous investigations. First, I increased the number of health conditions and comorbidity variables in the models. Next, I included variables for the medical treatments received, how many they received, who administered the treatments, and how long it took to begin treatment. I included these additional risk factors because the medical care received should have an essential role in an individual's ability to recover and how long it takes to recover, influencing the individual's resource pool and length of absence from work.

Using *k*-fold cross-validation estimates, I developed and trained eight machine learning models (e.g., Artificial Neural Network, Decision Tree, Gradient Boosted Trees, K-Nearest Neighbor, Logistic Regression, Nave Bayes, Random Forest, and XGBoost) using KNIME 4.5.2 Advanced Analytics against the combined dataset. The Random

Forest approach performed the best and was the most accurate (AUC = 87.20%; Sensitivity = 86.04%), followed by Gradient Boosted Trees (AUC = 86.08%; Sensitivity = 79.28%) and XGBoost (AUC = 85.33%; Sensitivity = 77.28%). According to the data and results, the homogeneous ensemble machine learning algorithms beat individual models in predicting long-term employee absence. Furthermore, the findings show that incorporating new healthcare-related factors effectively predicted long-term employee absenteeism and outperformed previous results (see Appendix 1).

While long-term employee absenteeism due to poor health is interesting and challenging to predict, it may not have much practical benefit on its own. More valuable is the proper identification and understanding of employee health-related factors that can increase (or decrease) the risk of absence length. As a result, data-driven insights might point to the need for new health programs, service-level agreements, and technological advancements to address the broader issue of long-term employee absences. Therefore, the primary goal of the next research contribution was to determine the significant health-related risk variables impacting long-term employee absence and discover their patterns.

Using the Random Forest variable importance heuristic, I narrowed the numerous variables to the top 30 most important. The top 30 variables fall into four distinct risk groups, each with four to twelve variables (see Figure 11). No single factor by itself appeared to be a key determinant of long-term employee absence; however, it can act as a catalyst or hindrance in combination with other factors in affecting the level of employee absences. For example, in order of importance, the top variables in Group 1 were Imaging: X-Ray, Visits: Outpatient, Global Surgical Package: 90 Days, and Number of Musculoskeletal Conditions. According to the Random Forest variable importance

results, these four risk factors appear far more critical than the rest in predicting long-term employee absences. Therefore, I was not surprised to learn that the Global Surgical Package: 90 Days variable and the Number of Musculoskeletal Conditions variable made Group 1. The 90-Day Global Surgical Package indicates that an individual is undergoing a more complicated treatment with a longer recovery time, and the Number of Musculoskeletal Conditions variable suggests that the individual has more health issues that may require more care.

One variable in Group 2 stands out above the rest: Days to Follow-Up Visit. Conventional logic would suggest that the longer an individual waits for additional care, the more likely the individual's recovery may be slowed or worsen. However, the remaining five variables in Group 2 complement the variables in Group 1. For example, two of the five remaining variables are provider type variables (e.g., Orthopedic Surgery and Primary Care), representing how many times an individual visited that type of healthcare provider. The outpatient visits variable being in Group 1 describes whom the individuals have seen for their outpatient care. Furthermore, the general surgical procedure, orthopedic surgery, and MR imaging variables in Group 2 combined to complement the 90-day Global Surgical Package variable in Group 1. Finally, physical therapy completes Group 2 and complements the other variables in that individuals with MSDs will most likely require therapy to aid in their recovery process.

The other two groups mostly include additional treatment variables centered on specific surgical procedures (e.g., Arthroscopy, Joint Injections, Nerve Block Injections, and Arthrodesis), drugs (e.g., Opioids, NSAIDs), and visits (e.g., Inpatient, OR). In addition, the CDPS risk score is in Group 3, and the HHS-HCC risk score is in Group 4,

which was used as an indicator of the individual's health. I was surprised that neither was ranked higher, but the focus of this study was on MSDs, and the risk score variables were proved important. Furthermore, the number of Musculoskeletal Conditions variable appeared in Group 1. It might complement an individual's risk score because both risk models focus mainly on chronic conditions that are likely to impair an individual's long-term health and health-related expenditures (Yeatts & Sangvai, 2016).

There were a couple of variables that I expected to make the Top 30 that did not. For example, while the risk score factors demonstrated the importance of individuals' overall health, no single health condition indicator did. Furthermore, I was surprised that neither the department nor job-related factors made the Top 30. Their rank was not bad, but I expected departments such as Fire, particularly active jobs such as nurses, to play an essential role in the variable importance of employee absenteeism. The department variable, for example, ranked 34, whereas the job workload variable ranked 57.

While the variable importance analysis performed can provide invaluable insight into the ranked importance of the study's independent variables, it does not capture and/or explain the variables' directional contribution to the dependent variable (e.g., long-term employee absence, high medical costs). I used data mining techniques to explore the top-ranked variables for new useful information based on their patterns to overcome this limitation. As discussed in Chapter IV, when I examined the patterns for x-ray imaging and outpatient visits, each variable had a value of two or more approximately 68% of the time when absence hours were high. Conversely, when the absence hours were low, approximately 27% of patients had two or more x-rays, and 29% had outpatient visits. Furthermore, when absence hours are high, an individual is more likely

to have more than one MSD (61% vs. 24%). Additionally, when absence hours are high, around 53% of people have days to follow-up visit of more than 14 days, compared to 35% when they are low. These are just a few examples; more can be found in Chapter IV.

In Study 2, I focused on this study's final contribution. I examined the distinctions between the health risk factors that predict long-term employee absenteeism and high medical costs. Both outcome variables are important for an employer because they can impact the organization's financial well-being. However, employee absenteeism is not a typical target variable in healthcare-related research. Primary for two reasons. First, most health-related studies utilize publicly-available CMS datasets primarily focused on retirees and not the population of working-age individuals. Second, healthcare databases in the U.S. mostly do not contain absence information. Absence data is not commonly made publicly available at the employer and employee levels. Thus, a company's human resource data must be combined with their healthcare data using advanced data processing techniques. More commonly used measures as dependent variables in healthcare-related research include medical costs and healthcare quality of care (e.g., infection rates, readmission rates, mortality rates).

Before comparing the healthcare-related variables, I utilized the same Random Forest prediction model to predict high absence hours because it produced the best results. Then I compared the two Random Forest prediction findings based on AUC (high absence hours = 87.20%; high medical costs = 93.83%), and the high medical costs result was superior. I did not anticipate, however, that the sensitivity analysis (high absence hours = 86.04%; high medical costs = 79.07%) would show that high absence hours performed better. This indicates that the model was better at accurately predicting the

minority class for employee absences. Ordinary reasoning may indicate that since this was a time-based study (episodes of care) and absence hours are a time-based variable, there is a stronger relationship between the two when compared to costs. For example, if two individuals receive the same treatment, but one receives it in three days and the other in three weeks, the medical costs may be the same, but the absence hours from work could be significantly different. This is assuming the individual did not get worse while waiting for the referred care.

Comparing the risk factors that predict long-term employee absenteeism and high medical costs, I found they are comparable but not identical. As noted in Chapter IV, the results of both investigations were influenced by characteristics representing greater utilization of healthcare services (e.g., imaging, office visits), more complex treatments (e.g., 90-day global surgical package), and more complicated conditions (e.g., number of musculoskeletal conditions). This result seemed surprising, not because of the variables, which may indicate more complex treatments or conditions, but the utilization measures themselves.

Take Group 1, for example; for high absence hours, the first two most important variables are x-rays and outpatient visits. Three of the four most important variables for high medical costs are MRIs, x-rays, and outpatient visits. For high medical costs, ordinary reasoning would be if an individual gets more MRIs or x-rays, they will have higher costs. However, can we say the same about high absence hours? For example, ordinary reasoning may suggest that more complex treatments and conditions would have higher absence hours, not just because they had more x-rays or office visits. However, the utilization measure may be acting as a proxy measure. More utilization may indicate that

the individual has a harder-to-diagnose problem, the problem is being monitored over time, or receiving subpar care. These are speculative, and further research is needed to understand better the relationships for both outcome and utilization measures.

Assumptions and Limitations

Because the sample only included three organizations that self-insure their healthcare products, the findings may not apply to all employers and employees. Administrative claims data lacks essential clinical information required to assess service quality (Pincus, Scholle, Spaeth-Rublee et al., 2016). This clinical data would allow for additional risk factors, such as lab, imaging, procedure results, and the individual's physical characteristics, such as height and weight. In addition, we may be missing important historical patient information, which could put an individual at greater risk of complications, such as prior injuries or surgical treatments of the same area, which happened before becoming an employee or prior to them being part of the company's benefits program. Furthermore, in the studies, I assume that if the medical or pharmaceutical treatment was billed through claims, it was followed through as prescribed.

In addition, the employers used a variety of vendors that sourced their medical claims, pharmaceutical claims, and human resource demographics and attendance data. Because the companies used different suppliers, there were inconsistencies in file formats, attribute names, and attribute values, which elevated the processing complexity and standardizing the data across employers and their vendors. Where ever possible, I

overcame vendor and employer-specific categories and cohorts utilizing publicly available industry-standard secondary data files.

Human resources information systems data has additional limitations. For example, employers may have unique policies or business logic that may not align with another employer on how they capture employees' time. For example, organizations may have different time off work policies for salaried versus hourly employees or civil versus non-civil service. Therefore, before combining, detailed attention must be given to the data between employers and within an organization's data. One method to overcome these obstacles is to have conversations with leadership or their subject matter experts in the human resources department.

Future Directions

Future research should focus on other prevalent health conditions for the employer population, such as diabetes, cardiovascular disease, obesity, and various behavioral health issues (e.g., stress, depression, anxiety, and substance abuse). Furthermore, as new health conditions are investigated, new treatment factors common to the health conditions should be developed. For example, suppose the focus is on depression. You may remove existing variables, create new treatment variables related to pharmacological or talk therapies, and adapt the provider-type variables to those who treat depression (e.g., family medicine, psychologist, counselor, psychiatrist). Since the dependent variables in this study were designed as binary classifications, future studies could examine employee absence at several levels rather than the binary classification.

This could help determine the critical explanatory variables at each level of differentiation between employee absenteeism.

Many of the top health-related factors in this study showed how higher utilization of specific services might indicate that the individual is not as healthy. However, determining the severity of an individual's health compared to another was not the goal of this study, but to show how these prediction results, variable importance, and the patterns of the variables may be used for new insights and even as proxy measures for future research. For example, the Centers for Medicare and Medicaid Services (2021i) has a fraud and abuse initiative that focuses on numerous types of abuse and fraud, ranging from submitting false claims to ordering unnecessary items or services for patients. When individual absence hours are low, but the individual has high utilization for variables that predict high absence hours, can absence hours be utilized to indicate possible unneeded treatments? This is only an example, but it would be intriguing to see how employee absences due to poor health could be utilized to identify unnecessary care and reduce the cost of patient care.

Potential Implications and Conclusion

Employee health is complex for an employer to evaluate beyond aggregate medical and productivity costs. Creating a system that can be specific to their employee's health gives employers an evidence-based tool that can then be used to improve employee well-being and cost-savings strategies that lead to competitive advantages (Miller, 1995; Nielsen, Nielsen, Ogbonnaya et al., 2017). Employers and their vendors, in particular, can gain a better understanding of their employees' prevalent health issues and

how health-related aspects affect costs and productivity. Furthermore, this type of system could be used to gain a better understanding of health-related factors such as medical treatments on employee health using occupational outcomes (e.g., absenteeism, presenteeism, turnover), and these new insights could pave the way for future research in management, occupational health, medicine, and psychology.

According to research, traditional absence management systems may be less effective than more focused corporate initiatives (Kohler & Mathieu, 1993). A system like the one described in this paper would enable more targeted absence management initiatives. For example, let us assume that a large proportion of employee absences are caused by musculoskeletal conditions. Further analysis indicates that it is primarily related to back pain for police officers and knee pain for nurses. In that instance, companies could use these data insights for targeted interventions for their employees based on those unique conditions. One potential organizational intervention is for a company to contract with a case management vendor who could use the insights to educate and work with employees on available medical care to improve their current health.

A second potential organizational intervention would be collaborating with their vendors to improve access to timely healthcare. According to the findings of this study, those who have to wait longer for care are more likely to have excessive absence hours and medical costs. As a result, organizations may enter into service-level agreements with its medical benefits vendors to ensure that all of its employees receive medical care as soon as possible. A third practical use for such a system might be to assess the effectiveness of employee health programs that employers invest in to reduce overall

healthcare costs and improve employees' health. Knowing which programs and health conditions employees are subjected to would allow employers to assess the health program's impact by comparing pre- and post-program medical costs and occupational outcomes. Employers could also utilize the system to identify employees who would benefit from a health program but are not participating. Such a data-rich system has numerous practical applications for employers. However, organizations would be unable to target such interventions if they did not grasp the dimensionality of absenteeism related to their employee's health.

REFERENCES

Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.

Agency for Healthcare Research and Quality. (2021). *Clinical classifications software refined (CCSR)*. AHRQ. Https://Www.Hcup-Us.Ahrq.Gov/. Retrieved December 13, 2021, from https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp

Airaksinen, J., Jokela, M., Virtanen, M., Oksanen, T., Koskenvuo, M., Pentti, J., Vahtera, J., & Kivimäki, M. (2018). Prediction of long-term absence due to sickness in employees: development and validation of a multifactorial risk score in two cohort studies. *Scandinavian Journal of Work, Environment & Health*, *44*(3), 274-282.

American Diabetes Association. (2018). Economic costs of diabetes in the U.S. in 2017. *Diabetes Care*, *41*(5), 917-928.

American Diabetes Association. (2020). Standards of medical care in diabetes—2020. *Diabetes Care, 43*(Supplement 1), S1-S212.

American Medical Association. (2021). *CPT®*. Retrieved December 16, 2021, from https://www.ama-assn.org/practice-management/cpt

Anderson, G. F., & Chalkidou, K. (2008). Spending on medical care: More is better?. *Jama*, *299*(20), 2444-2445.

Berger, M., & Czypionka, T. (2021). Regional medical practice variation in high-cost healthcare services. *The European Journal of Health Economics, 22*(6), 917-929.

Bergström, G., Hagberg, J., Busch, H., Jensen, I., & Björklund, C. (2014). Prediction of sickness absenteeism, disability pension and sickness presenteeism among employees with back pain. *Journal of Occupational Rehabilitation*, *24*(2), 278-286.

Bosman, L. C., Dijkstra, L., Joling, C. I., Heymans, M. W., Twisk, J. W., & Roelen, C. A. (2018). Prediction models to identify workers at risk of sick leave due to low-back pain in the Dutch construction industry. *Scandinavian Journal of Work, Environment & Health*, *44*(2), 156-162.

Boyd, A. (1997). Employee traps—Corruption in the workplace. *Management Review*, *86*(8), 9-10.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145-1159.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140.

Breiman, L. (2000). Randomizing outputs to increase prediction accuracy. *Machine Learning*, *40*(3), 229-242.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Brooks/Cole Publishing.

Brooke, P. P. (1986). Beyond the Steers and Rhodes model of employee attendance. *Academy of Management Review*, *11*(2), 345-361.

Brooke Jr, P. P., & Price, J. L. (1989). The determinants of employee absenteeism: An empirical test of a causal model. *Journal of Occupational Psychology*, *62*(1), 1-19.

Burdorf, A. (2019). Prevention strategies for sickness absence: Sick individuals or sick populations?. *Scandinavian Journal of Work, Environment & Health*, *45*(2), 101-102.

Bureau of Labor Statistics. (2016, November 10). *Nonfatal occupational injuries and illnesses requiring days away from work*, *2015*. https://www.bls.gov/news.release/pdf/osh2.pdf

Center for Drug Evaluation and Research. (2020). *Information by drug class*. U.S. Food and Drug Administration. Retrieved December 8, 2021, from https://www.fda.gov/drugs/drug-safety-and-availability/information-drug-class

Centers for Disease Control and Prevention. (2021). *ICD - ICD-10-CM - International classification of diseases, 10th revision, clinical modification*. National Center for Health Statistics. Retrieved December 13, 2021, from https://www.cdc.gov/nchs/icd/icd10cm.htm

Centers for Medicare & Medicaid Services. (2020). *Restructured BETOS classification system (RBCS) final report*. Centers for Medicare & Medicaid Services. Retrieved December 8, 2021, from https://data.cms.gov/sites/default/files/2021-08/rbcs_base_year_final_report_0.pdf

Centers for Medicare and Medicaid Services. (2021). *NHE fact sheet*. Retrieved from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet

Centers for Medicare and Medicaid Services. (2021b). *Medicare mental health*. Retrieved from https://www.cms.gov/files/document/medicare-mental-health.pdf

Centers for Medicare and Medicaid Services. (2021c). *MS-DRG classifications and software | CMS*. Retrieved December 13, 2021, from https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/MS-DRG-Classifications-and-Software

Centers for Medicare and Medicaid Services. (2021d). *NPI files*. Retrieved December 13, 2021, from https://download.cms.gov/nppes/NPI_Files.html

Centers for Medicare & Medicaid Services. (2021e). *Taxonomy crosswalk*. CMS. Retrieved December 13, 2021, from https://www.cms.gov/medicare/provider-enrollment-and-certification/medicareprovidersupenroll/downloads/taxonomycrosswalk.pdf

Centers for Medicare & Medicaid Services. (2021f). *Place of service codes | CMS*. Retrieved December 13, 2021, from https://www.cms.gov/Medicare/Coding/place-of-service-codes

Centers for Medicare & Medicaid Services. (2021g). *Chapter IV Surgery: Musculoskeletal system CPT codes (20000–29999) for national correct coding initiative policy manual for medical association*. Retrieved December 8, 2021, from https://www.cms.gov/files/document/chapter4cptcodes20000-29999final112021.pdf

Centers for Medicare & Medicaid Services. (2021h). *Final 2021 benefit year final HHS risk adjustment model coefficients*. Retrieved December 13, 2021, from https://www.cms.gov/CCIIO/Resources/Regulations-and-Guidance/Downloads/Final-2021-Benefit-Year-Final-HHS-Risk-Adjustment-Model-Coefficients.pdf

Centers for Medicare & Medicaid Services. (2021i). *Medicare fraud & abuse*. Retrieved May 22, 2022, from https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/Fraud-Abuse-MLN4649244.pdf

Chadwick-Jones, J. K., Nicholson, N., & Brown, C. (1982). *Social psychology of absenteeism*. New York: Praeger.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In Maimon, O., & Rokach, L. (eds.), *Data mining and knowledge discovery handbook* (pp. 875-886). Boston, MA: Springer.

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems, 192*, 105361.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Chronic illness and disability payment system. (2022). (n.d.). UCSanDiego. Retrieved January 6, 2022, from https://hwsph.ucsd.edu/research/programs-groups/cdps.html

Clegg, C. W. (1983). Psychology of employee lateness, absence, and turnover: A methodological critique and an empirical study. *Journal of Applied Psychology*, *68*(1), 88.

Cucchiella, F., Gastaldi, M., & Ranieri, L. (2014). Managing absenteeism in the workplace: The case of an Italian multiutility company. *Procedia-Social and Behavioral Sciences*, *150*, 1157.

Darr, W., & Johns, G. (2008). Work strain, health, and absenteeism: A meta-analysis. *Journal of Occupational Health Psychology*, *13*(4), 293.

de Oliveira, E. L., Torres, J. M., Moreira, R. S., & de Lima, R. A. F. (2019, April). Absenteeism prediction in call center using machine learning algorithms. In *World Conference on Information Systems and Technologies* (pp. 958-968). Springer, Cham.

Dean-Deyle, G. (2011). The role of MRI in musculoskeletal practice: A clinical perspective. *Journal of Manual & Manipulative Therapy, 19*(3), 152-161.

Delen, D. (2020). *Predictive analytics: Data mining, machine learning and data science for practitioners* (2nd ed.), Pearson Business Analytics Series. Pearson FT Press.

Delen, D., Tomak, L., Topuz, K., & Eryarsoy, E. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health*, *4*, 118-131.

Doupe, P., Faghmous, J., & Basu, S. (2019). Machine learning for health services researchers. *Value in Health*, *22*(7), 808-815.

Durand, V. M. (1985). Employee absenteeism: A selective review of antecedents and consequences. *Journal of Organizational Behavior Management*, *7*(1-2), 135-168.

Felton, J. S., & Cole, R. (1963). The high cost of heart disease. *Circulation*, *27*(5), 957-962.

Fichman, M. (1984). A theoretical approach to understanding employee absence. In Goodman, P. S., & Atkin, R. S., *Absenteeism: New approaches to understanding, measuring, and managing employee absence* (pp. 1-46). Jossey-Bass Management Series.

Fix, E., & Hodges, J. L. (1951). *Nonparametric discrimination: Consistency properties*. Randolph Field, Texas, Project, 21-49.

Follmer, K. B., & Jones, K. S. (2018). Mental illness in the workplace: An interdisciplinary review and organizational research agenda. *Journal of Management*, *44*(1), 325-351.

Forouzanfar, M. H., Afshin, A., Alexander, L. T., Anderson, H. R., Bhutta, Z. A., Biryukov, S., ... & Carrero, J. J. (2016). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, *388*(10053), 1659-1724.

Freburger, J. K., Holmes, G. M., & Carey, T. S. (2003). Physician referrals to physical therapy for the treatment of musculoskeletal conditions. *Archives of Physical Medicine and Rehabilitation*, *84*(12), 1839-1849.

Freund, Y., & Schapire, R. E. (1996). *Experiments with a new boosting algorithm*. International Conference on Machine Learning, Bari, 3-6 July (pp. 148-156).

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232.

Gibson, R. O. (1966). Toward a conceptualization of absence behavior of personnel in organization. *Administrative Science Quarterly*, *11*(1), 107-133.

Hackett, R. D., Bycio, P., & Guion, R. M. (1989). Absenteeism among hospital nurses: An idiographic-longitudinal analysis. *Academy of Management Journal*, *32*(2), 424-453.

Hackett, R. D., & Guion, R. M. (1985). A reevaluation of the absenteeism-job satisfaction relationship. *Organizational Behavior and Human Decision Processes*, *35*(3), 340-381.

Harris, J. D., Johnson, S. G., & Souder, D. (2013). Model-theoretic knowledge accumulation: The case of agency theory and incentive alignment. *Academy of Management Review*, *38*(3), 442-454.

Harrison, D. A., & Martocchio, J. J. (1998). Time for absenteeism: A 20-year review of origins, offshoots, and outcomes. *Journal of Management*, *24*(3), 305-350.

Hartvigsen, J., Hancock, M. J., Kongsted, A., Louw, Q., Ferreira, M. L., Genevay, S., ... & Woolf, A. (2018). What low back pain is and why we need to pay attention. *The Lancet*, *391*(10137), 2356-2367.

Haykin, S. (2008). *Neural networks and learning machines,* (3rd ed.). Prentice Hall Publishing.

Heath, C., & Sitkin, S. B. (2001). Big-B versus Big-O: What is organizational about organizational behavior?. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *22*(1), 43-58.

Hedges, J. N. (1973). Absence from work—A look at some national data. *Monthly Lab. Rev.*, *96*, 24.

Hedges, J. N. (1975). Unscheduled absence from work—An update. *Monthly Labor Review*, 36-39.

Hedges, J. N. (1977). Absence from work-measuring the hours lost. *Monthly Lab. Rev.*, *100*, 16.

Hendrix, W. H., & Spencer, B. A. (1989). Development and test of a multivariate model of absenteeism. *Psychological Reports*, *64*(3), 923-938.

Hessel, F. P. (2021). Overview of the socio-economic consequences of heart failure. *Cardiovascular Diagnosis and Therapy, 11*(1), 254.

Hobfoll, S. E. (1989). Conservation of resources: A new attempt at conceptualizing stress. *American Psychologist*, *44*(3), 513.

Hobfoll, S. E. (2001). The influence of culture, community, and the nested-self in the stress process: Advancing conservation of resources theory. *Applied Psychology, 50*(3), 337-421.

Hobfoll, S. E. (2011). Conservation of resources theory: Its implication for stress, health, and resilience. In S. Folkman (ed.), *The Oxford handbook of stress, health, and coping* (pp. 127-147). Oxford University Press.

Hobfoll, S. E., Vinokur, A. D., Pierce, P. F., & Lewandowski-Romps, L. (2012). The combined stress of family life, work, and war in Air Force men and women: A test of conservation of resources theory. *International Journal of Stress Management*, *19*(3), 217.

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multi-layer feedforward networks. *Neural Networks*, *3*(5), 551-560.

Hosmer Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hurwitz, E. L., Morgenstern, H., Harber, P., Kominski, G. F., Belin, T. R., Yu, F., & Adams, A. H. (2002). A randomized trial of medical care with and without physical therapy and chiropractic care with and without physical modalities for patients with low back pain: 6-month follow-up outcomes from the UCLA low back pain study. *Spine*, *27*(20), 2193-2204.

Ilka, R. (2016). Preventive care and the chief health officer. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/preventive-care-and-the-chief-health-officer-1453243425

Integrated Benefits Institute. (2020, December 8). *Poor health costs U.S. employers $575 billion and 1.5 billion days of lost productivity per Integrated Benefits Institute*. from https://www.ibiweb.org/poor-health-costs-us-employers-575-billion/

Jacobson, J. A. (2009). Musculoskeletal ultrasound: Focused impact on MRI. *American Journal of Roentgenology, 193*(3), 619-627.

Johns, G. (1994). Absenteeism estimates by employees and managers: Divergent perspectives and self-serving perceptions. *Journal of Applied Psychology*, *79*(2), 229.

Johns, G. (1997). Contemporary research on absence from work: Correlates, causes and consequences. *International Review of Industrial and Organizational Psychology*, *12*, 115-174.

Johns, G. (2008). Absenteeism and presenteeism: Not at work or not working well. In J. Barling & C. L. Cooper, *The Sage handbook of organizational behavior* (vol. 1, pp. 160-177). SAGE.

Johns, G. (2010). Presenteeism in the workplace: A review and research agenda. *Journal of Organizational Behavior*, *31*(4), 519-542.

Johnston, D. A., Harvey, S. B., Glozier, N., Calvo, R. A., Christensen, H., & Deady, M. (2019). The relationship between depression symptoms, absenteeism and presenteeism. *Journal of Affective Disorders*, *256*, 536-540.

Kautter, J., Pope, G. C., Ingber, M., Freeman, S., Patterson, L., Cohen, M., & Keenan, P. (2014). The HHS-HCC risk adjustment model for individual and small group markets under the Affordable Care Act. *Medicare & Medicaid Research Review*, *4*(3). https://doi.org/10.5600/mmrr2014-004-03-a03

Kessler, R. C., Barber, C., Beck, A., Berglund, P., Cleary, P. D., McKenas, D., ... & Wang, P. (2003). The World Health Organization health and work performance questionnaire (HPQ). *Journal of Occupational and Environmental Medicine*, *45*(2), 156-174.

Kohler, S. S., & Mathieu, J. E. (1993). Individual characteristics, work perceptions, and affective reactions influences on differentiated absence criteria. *Journal of Organizational Behavior*, *14*(6), 515-530.

Kshirsagar, R., Hsu, L. Y., Chaturvedi, V., Greenberg, C. H., McClelland, M., Mohan, A., ... & Alvarado, M. (2020). Accurate and interpretable machine learning for transparent pricing of health insurance plans. *arXiv preprint arXiv:2009.10990*. https://doi.org/10.48550/arXiv.2009.10990

Lawrance, N., Petrides, G., & Guerry, M. A. (2021). Predicting employee absenteeism for cost effective interventions. *Decision Support Systems*, *147*(1), 113539.

Lewis, A. K., Harding, K. E., Snowdon, D. A., & Taylor, N. F. (2018). Reducing wait time from referral to first visit for community outpatient services may contribute to better health outcomes: A systematic review. *BMC Health Services Research*, *18*(1), 1-14.

Ling, C. X., Huang, J., & Zhang, H. (2003, June). AUC: A better measure than accuracy in comparing learning algorithms. In *Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 329-341). Springer Berlin Heidelberg.

Lohaus, D., & Habermann, W. (2019). Presenteeism: A review and research directions. *Human Resource Management Review*, *29*(1), 43-58.

Luo, L., Li, J., Lian, S., Zeng, X., Sun, L., Li, C., ... & Zhang, W. (2020). Using machine learning approaches to predict high-cost chronic obstructive pulmonary disease patients in China. *Health Informatics Journal*, *26*(3), 1577-1598.

Maisog, J. M., Li, W., Xu, Y., Hurley, B., Shah, H., Lemberg, R., ... & Gutfraind, A. (2019). Using massive health insurance claims data to predict very high-cost claimants: A machine learning approach. *arXiv preprint arXiv:1912.13032*.

Martocchio, J. J., & Harrison, D. A. (1993). To be there or not to be there? Questions, theories and methods in absenteeism research. *Research in Personnel and Human Resources Management*, *11*(1), 259-328.

Mathis, R. L., Jackson, J. H., & Valentine, S. R. (2015). *Human resource management: Essential perspectives*. Cengage Learning.

Mattke, S., Balakrishnan, A., Bergamo, G., & Newberry, S. J. (2007). A review of methods to measure health-related productivity loss. *American Journal of Managed Care*, *13*(4), 211.

McDonald, M., daCosta DiBonaventura, M., & Ullman, S. (2011). Musculoskeletal pain in the workforce: The effects of back, arthritis, and fibromyalgia pain on quality of life and work productivity. *Journal of Occupational and Environmental Medicine*, *53*(7), 765-770.

Miller, R. J. (1995). Restructuring wages and benefits to gain a competitive edge. *Journal of the Healthcare Financial Management Association, 49*(2), 58-60.

Miner, J. B., & Brewer, J. F. (1976) The management of ineffective performance. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 995-1029). Chicago: Rand McNally.

Montano, I. H., Marques, G., Alonso, S. G., López-Coronado, M., & de la Torre Díez, I. (2020). Predicting absenteeism and temporary disability using machine learning: A systematic review and analysis. *Journal of Medical Systems*, *44*(9), 1-11.

Moore, R. H., & Buschbom, R. L. (1974). Work absenteeism in diabetics. *Diabetes*, *23*(12), 957-961.

Morgan, L. G., & Herman, J. B. (1976). Perceived consequences of absenteeism. *Journal of Applied Psychology*, *61*(6), 738.

Muchinsky, P. M. (1977). Employee absenteeism: A review of the literature. *Journal of Vocational Behavior*, *10*(3), 316-340.

Nielsen, K., Nielsen, M. B., Ogbonnaya, C., Känsälä, M., Saari, E., & Isaksson, K. (2017). Workplace resources to improve both employee well-being and performance: A systematic review and meta-analysis. *Work & Stress, 31*(2), 101-120.

Neisse, A. C., de Oliveira, F. L. P., de Oliveira, A. C. S., & Neto, R. M. N. (2021). Chronic fatigue syndrome and its relation with absenteeism: Elastic-net and stepwise applied to biochemical and anthropometric clinical measurements. *Revista Brasileira de Biometria*, *39*(1), 221-239.

Nicholson, N., & Payne, R. (1987). Absence from work: Explanations and attributions. *Applied Psychology, 36(2),*121-132.

Notenbomer, A., van Rhenen, W., Groothoff, J. W., & Roelen, C. A. (2019). Predicting long-term sickness absence among employees with frequent sickness absence. *International Archives of Occupational and Environmental Health*, *92*(4), 501-511.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, *375*(13), 1216.

Organisation for Economic Cooperation and Development. (2021, April 23). *OECD productivity statistics*. https://www.oecd-ilibrary.org/employment/data/oecd-productivity-statistics_pdtvy-data-en

Pareto, V. (1897). *Cours d'économie politique professeur á l'Université de Lausanne*. Paris: F. Rouge Lausanne.

Paringer, L. (1983). Women and absenteeism: Health or economics?. *The American Economic Review*, *73*(2), 123-127.

Pell, S., & D'Alonzo, C. A. (1967). Sickness absenteeism in employed diabetics. *American Journal of Public Health and the Nations Health, 57*(2), 253-260.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia, 4*(2), 1883.

Pfeffer, J. (2018). *Dying for a paycheck: How modern management harms employee health and company performance—and what we can do about it*. Harper Business.

Pincus, H. A., Scholle, S. H., Spaeth-Rublee, B., Hepner, K. A., & Brown, J. (2016). Quality measures for mental health and substance use: Gaps, opportunities, and challenges. *Health Affairs (Project Hope), 35*(6), 1000-1008. https://doi.org/10.1377/hlthaff.2016.0027

Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Iezzoni, L. I., ... & Robst, J. (2004). Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financing Review*, *25*(4), 119.

Price, R. H., & Hooijberg, R. (1992). Organizational exit pressures and role stress: Impact on mental health. *Journal of Organizational Behavior*, *13*(7), 641-651.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106.

Rajkumar, S. V. (2020). The high cost of insulin in the United States: An urgent call to action. In *Mayo Clinic Proceedings* (Vol. 95, No. 1, pp. 22-28). Elsevier.

Research Data Assistance Center. (2021). *UB-92 revenue code (20th) | ResDAC*. ResDAC. Retrieved December 13, 2021, from https://resdac.org/cms-data/variables/ub-92-revenue-code-20th#:%7E:text=CODE%20WHICH%20IDENTIFIES%20A%20SPECIFIC,AND%20BOARD%20(OR%20ACCOMMODATIONS).

Rhodes, S. R., & Steers, R. M. (1990). *Managing employee absenteeism*. Addison Wesley Publishing Company.

Ross, E. L., Vijan, S., Miller, E. M., Valenstein, M., & Zivin, K. (2019). The cost-effectiveness of cognitive behavioral therapy versus second-generation antidepressants for initial treatment of major depressive disorder in the United States: A decision analytic model. *Annals of Internal Medicine, 171*(11), 785-795.

Sagie, A., Birati, A., & Tziner, A. (2002). Assessing the costs of behavioral and psychological withdrawal: A new model and an empirical illustration. *Applied Psychology, 51*(1), 67-89.

Saxena, A., Das, S., Rubens, M., Salami, J., Bhatt, C., Tian, T., ... & Veledar, E. (2019). Predicting employee health and cost: Application of machine learning on employee health claims data, insights, and possibilities. *Circulation: Cardiovascular Quality and Outcomes*, *12*(Suppl_1), A178-A178.

Shalen, P. (2000). *Specialists who treat back pain*. Spine-Health. https://www.spine-health.com/treatment/spine-specialists/specialists-who-treat-back-pain

Sharda, R. Delen, D., & Turban, E. (2018). *Business intelligence, analytics, and data science: A managerial perspective* (4th ed.). New York, NY: Pearson.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*(4), 13-22.

Singer, G., & Cohen, I. (2020). An objective-based entropy approach for interpretable decision tree models in support of human resource management: The case of absenteeism at work. *Entropy*, *22*(8), 821.

Skorikov, M., Hussain, M. A., Khan, M. R., Akbar, M. K., Momen, S., Mohammed, N., & Nashin, T. (2020, December). Prediction of absenteeism at work using data mining techniques. In *2020 5th International Conference on Information Technology Research (ICITR)* (pp. 1-6). IEEE.

Skrepnek, G. H., Nevins, R., & Sullivan, S. (2012). An assessment of health and work productivity measurement in employer settings. *Pharmaceuticals Policy and Law*, *14*(1), 37-49.

Solomon, D. H., Bates, D. W., Panush, R. S., & Katz, J. N. (1997). Costs, outcomes, and patient satisfaction by provider type for patients with rheumatic and musculoskeletal conditions: A critical review of the literature and proposed methodologic standards. *Annals of Internal Medicine*, *127*(1), 52-60.

Steers, R. M., & Rhodes, S. R. (1978). Major influences on employee attendance: A process model. *Journal of Applied Psychology*, *63*(4), 391.

Stempel, J. (2018, May 6). *Buffett targets CEO for Berkshire-Amazon-JPMorgan healthcare venture soon*. Reuters. Retrieved from https://www.reuters.com/article/us-berkshire-buffett-healthcare/buffett-targets-ceo-for-berkshire-amazon-jpmorgan-healthcare-venture-soon-idUSKBN1I60RG

Stull, J. D., Bhat, S. B., Kane, J. M., & Raikin, S. M. (2017). Economic burden of inpatient admission of ankle fractures. *Foot & Ankle International*, *38*(9), 997-1004.

U.S. Department of Labor. (2021). *Dictionary of occupational titles.* DOT Dictionary of Occupational Titles. Retrieved December 13, 2021, from https://occupationalinfo.org/

Van Hasselt, M., Keyes, V., Bray, J., & Miller, T. (2015). Prescription drug abuse and workplace absenteeism: Evidence from the 2008–2012 national survey on drug use and health. *Journal of Workplace Behavioral Health*, *30*(4), 379-392.

van Hoffen, M. F., Norder, G., Twisk, J. W., & Roelen, C. A. (2020). Development of prediction models for sickness absence due to mental disorders in the general working population. *Journal of Occupational Rehabilitation*, *30*(3), 308-317.

van Hoffen, M. F. A., Roelen, C. A. M., van Rhenen, W., Schaufeli, W. B., Heymans, M. W., & Twisk, J. W. R. (2018). Psychosocial work characteristics and long-term sickness absence due to mental disorders. *Journal of Mental Health, 29*(6), 649-656. https://doi.org/10.1080/09638237.2018.1437603

World Health Organization. (2019). *Mental health: A state of well-being*. WHO. Retrieved from https://www.who.int/features/factfiles/mental_health/en/

World Health Organization. (2021). *Musculoskeletal conditions*. WHO. https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions

Yeatts, J. P., & Sangvai, D. (2016). HCC coding, risk adjustment, and physician income: What you need to know. *Family Practice Management*, *23*(5), 24-27.

Youngblood, S. A. (1984). Work, nonwork, and withdrawal. *Journal of Applied Psychology*, *69*(1), 106-117.

APPENDICES

APPENDIX A: Employee Absenteeism Prediction Models

Source	Variables	Results
Airaksinen et al. (2018)	Dependent Variables: Long-term sick leave (> 9 days), Long-term sick leave (≥ 90 days) Independent Variables: Self-rated health, Depression, Sex, Age, Socioeconomic status, Previous sickness absences, Chronic diseases, Smoking, Shift work, Night shift, Self-rated health (squared), Body mass index (squared), age (squared), Jenkins sleep scale (squared)	Long-term sick leave (> 9 days): Logistic Regression - AUC < 65%; Long-term sick leave (≥ 90 days): Logistic Regression - AUC = 73%
Bergström, Hagberg, Busch, Jensen, & Björklund (2014)	Dependent Variables: Long-term sick leave (>= 30 days) Independent Variables: Age, Gender, Education, Employment (Blue or White collar), Pain localization, Pain Duration, Pain Intensity during last week, Sickness absence due to neck-/back pain during the previous year, General health, Job strain, Perceived physical exertion at work, and Heavy lifting	Long-term sickness absence (>= 30 days): 0-6 months follow-up - AUC = 81%; 0-12 months - AUC = 75%; 13-24 months - AUC = 69%; Self-reported sickness absenteeism due to neck-/back pain: AUC = 77%

Notenbomer, van Rhenen, Groothoff, & Roelen (2019)	Dependent Variable: Long-term sick leave (6 weeks or longer) Model 1 Independent Variables: Age, Gender, Education, Marital status, Job demands (work pace, cognitive demands, emotional demands, work-home-interference), Job resources (role clarity, task variety, learning opportunities, supervisor support, co-worker support), Sickness absence spells in the year prior to the survey, Long-term sickness absence in the year prior to the survey, Long-term sickness absence in the year following the survey Model 2 Independent Variables: Age, Gender, Education, Marital status, Burnout, Work engagement, Sickness absence spells in the year prior to the survey, Long-term sickness absence in the year prior to the survey, Long-term sickness absence in the year following the survey	Model 1: Logistic Regression - AUC = 62.3%; Model 2: Logistic Regression - AUC = 62.4%
Neisse, de Oliveira, de Oliveira, & Neto (2021)	Dependent Variable: Absent in any giving day in the year related to the condition Independent Variables:Sex, Age, Height, Weight, BMI, Waiste to Hip Ratio, Total Body Fat,Visceral Fat, Blood Pressure (Diag., Sist.), Cholesterol (HDL, LDL), Triglycerides, Total Cholesterol, Calcium Ion, Phosporus Kinetic UV, Vitamin D, PTH Hormone, Fasting Glucose, Sodium, Potassium	Model 1: Stepwise Regression - AUC = 58.4%; Model 2: Lasso Regression - AUC = 56.9% Model 2: Elastic-Net Regression - AUC = 57.4%

Bosman et al. (2018)	Dependent Variable: Long-term sick leave; LBP sick leave had a median duration of 52 days (range of 3-730 days) Independent Variables: Gender, Age, Employed Years (construction industry, current company, current position), Work hours, Back Pain/Stiffness, Pain/Stiffness (other), Physician diagnosed musculoskeletal disorders/injuries, Health complaints caused/worsened by work, Feeling healthly, Sport activities, Stress, Fatigue, Vitality, Work satisfactions, Work organization, Psychological work demands, Autonomy, Workability, Postural physical work demands, Dynamic physical work demands, and Sick leave due to low-back pain	Logistic Regression - AUC = 69.2%;
van Hoffen, Norder, Twisk, & Roelen (2020)	Dependent Variable: Long-term sick leave (6 weeks or longer) Independent Variables: Sociodemographic (age, gender, marital status, care for children at home, education, years employed at company, years in present job, work hours per week, prior mental LTSA) Psychosocial work factors (work pace, cognitive demands, emotional demands, variety in work, role clarity, learning opportunities, support supervisor, support co-workers, organizational commitment) Social support family/friends, Work-family interference, Intrinsic work motivation, Work satisfaction, Work ability, Work engagement, Burnout, and Distress	Logistic Regression - AUC = 71.3%; 3-node Decision Tree - AUC = 70.9%

van Hoffen et al. (2018)	Dependent Variable: Long-term sick leave (42 days or longer) Independent Variables: Age, Gender, Education Level, Job type, Job Tenure, Work hours per week, Mental Health, Psychosocial work characteristics (workload, work pace, changes in work, variety in work, autonomy in work, participation in decisions about work, learning opportunities, feedback about one's performance, support from supervisor, and support from co-workers)	Logistic Regression - AUC = 65%
Skorikov et al. (2020)	Dependent Variable: Absenteeism hours Class A: Absence Hours (0 hours): Class B: Absence Hours (1-15 hours): Class C: Absence Hours (16-120 hours): Independent Variables: Month of absence, Day of absence, Seasons, Travel expense, Distance, Service time, Age, Workload per day, Hit target, Disciplinary failure, Education, Children, Social Drinker, Social Smoker, Pet, Weight, Height, BMI, and Reasons for absence (Certain infectious and parasitic diseases, Neoplasms, Blood-forming organ & immune mechanism, Endocrine, Nutritional and metabolic diseases, Mental and behavioral disorders, Diseases of the nervous system, Diseases of the eye and adnexa, Diseases of the ear and mastoid process, Diseases of the circulatory system, Diseases of the respiratory system, Diseases of the digestive system, Diseases of the skin and subcutaneous tissue, Diseases of musculoskeletal system & tissue, Diseases of the	Study A: 4 variables from CFS method zeroR - AUC = 50% naïve Bayes - AUC = 77% J48 - AUC = 72% KNN-Euclidean - AUC = 73% KNN-Manhattan - AUC = 70% KNN-Chebyshev - AUC = 64% Study 2: All variables zeroR - AUC = 50% naïve Bayes - AUC = 80% J48 - AUC = 76% KNN-Euclidean - AUC = 81% KNN-Manhattan - AUC = 76% KNN-Chebyshev - AUC = 70% Study 3: Most influential variable zeroR - AUC = 50% naïve Bayes - AUC = 69% J48 - AUC = 69% KNN-Euclidean - AUC = 69% KNN-Manhattan - AUC =

	genitourinary system, Pregnancy, childbirth and the puerperium, Conditions originating in the perinatal period, Congenital malformations and chromosomal abnormalities, Abnormal clinical and laboratory findings, injury, poisoning and consequences of external causes, External causes of morbidity and mortality, Factors to health status and health services, Patient follow-up, Medical consultation, Blood Donation, Laboratory examination, Unjustified absence, Physiotherapy, and Dental Consultation)	69% KNN-Chebyshev - AUC = 69%
Singer & Cohen (2020)	Dependent Variable: Categorical (not absent, hours, days, weeks) Not Absent: Absence Hours = 0 Hours: Absence Hours = between 1 and 7 Days: Absence Hour = between 8 and 39 Week: Absence Hours >= 40 Independent Variables: Month of absence, day of the week, Season, Transportation expense, Distance from residence to work (km), Service time, Age, Workload (average daily), Hit target, Disciplinary failure, Education, Number of Children, Social Drinker, Social Smoker, Number of Pets, Weight, Height, BMI, and Reason for Absence (21 categories according to the International Code of Diseases (ICD))	Extreme Gradient Boosting (XGBoost) - AUC = 73% Multi-Layer Perceptron - AUC = 50% KNN - AUC = 60% naïve Bayes - AUC = 56% Random Forest - AUC = 70% CART - AUC = 69% Ordinal CART OBE (c^{mode}) - AUC = 72% Ordinal CART OBE (c^{max}) - AUC = 76%

Lawrance, Petrides, & Guerry (2021)	Dependent Variable: Absenteeism hours (binary target based on threshold) Independent Variables: Month, Employee Demographics (age, gender, marital status, education), Work Environment (wage, contract type, fatigue, shift irregularities (weekend work, overtime, night shift), holiday applications (frequency, duration, number of rejected holiday applications, time since last holiday of certain duration)), Historic Absence Patterns (time since last absence, frequency of illnesses, average hours of sickness absences in the 12 month period prior to the prediction period)	Top model selection by target period: 2018/03 AdaBoost - AUC = 67% 2018/04 Easy Ensemble - AUC = 67% 2018/05 Easy Ensemble - AUC = 68% 2018/06 Easy Ensemble - AUC = 70% 2018/07 AdaBoost - AUC = 69% 2018/08 AdaBoost - AUC = 71% 2018/09 Easy Ensemble - AUC = 68% 2018/10 AdaBoost - AUC = 68% 2018/11 AdaBoost - AUC = 68% 2018/12 AdaBoost - AUC = 69% 2019/01 Easy Ensemble - AUC = 66% 2019/02 Easy Ensemble - AUC = 67% 2019/03 Easy Ensemble - AUC = 70%
de Oliveira, Torres, Moreira, & de Lima (2019)	Dependent Variable: Employee Absence (missing more than 50% of daily working hours) Independent Variables: Personal Features (individual registration, has landline, dependents, internal questionnaire, individual and management assessment, education, marital status, origin of person, age, sex), Work Activities Features (city of work, distance from work, work sector, work shift, instant manager, productivity level, worked hours, business time, hiring disclosure), Social and Admin Platform Features	Models: Random Forest - AUC = 71% Multilayer Perception - AUC = 64% Support Vector Machine - AUC = 56% naïve Bayes - AUC = 63% XGBoost - AUC = 73% Long Short-Term Memory - AUC = 53% Logistic Regression - AUC = 60%

	and Administrative Platform (productivity, mood, system access, virtual store and character interaction, interaction messages, friends on the social network, login feature), Absenteeism Related Features (holidays, absences at work, weekly absences at work, absence of friends at work, days to last rest, last day worked)	

APPENDIX B: Frequency Histograms

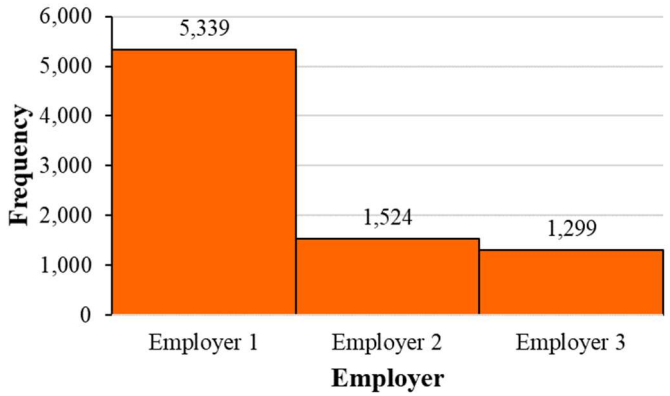


Figure 32. Frequency histogram by Employer.

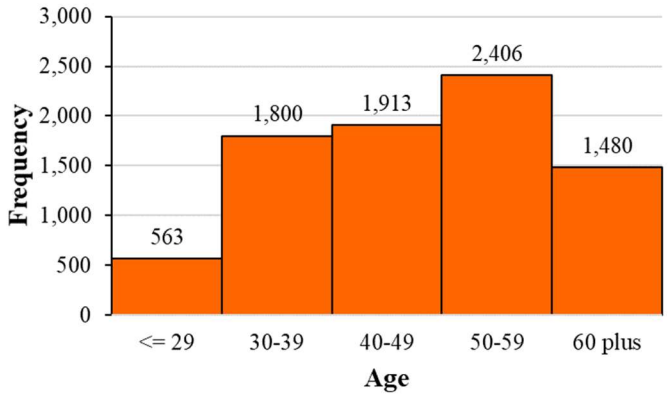


Figure 33. Frequency histogram by Employee Age.

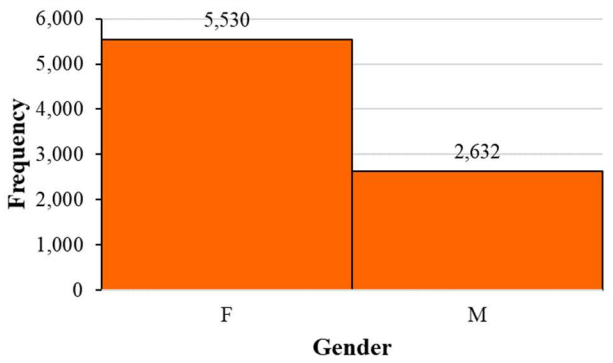


Figure 34. Frequency histogram by Employee Gender.

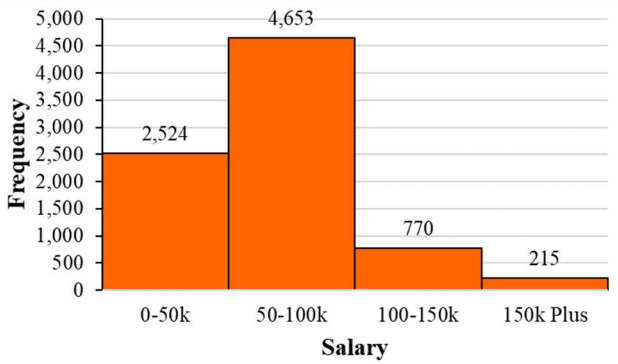


Figure 35. Frequency histogram by Employee Salary.

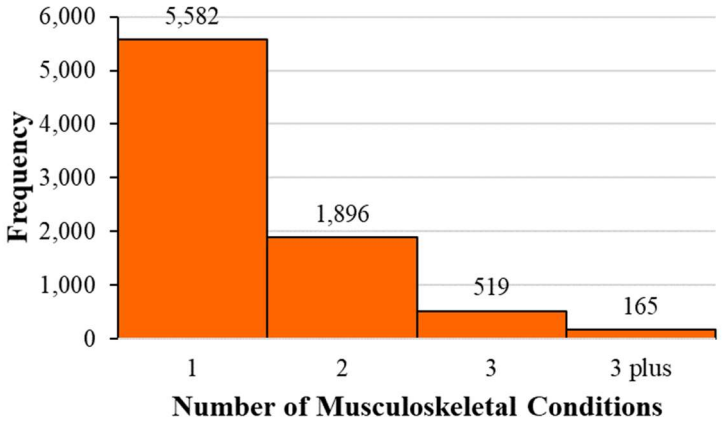


Figure 36. Frequency histogram by the Number of Musculoskeletal Conditions.

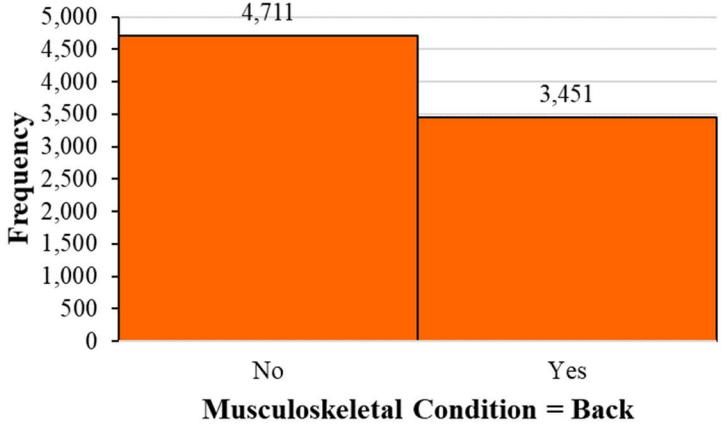


Figure 37. Frequency histogram by Musculoskeletal Condition for the Back.

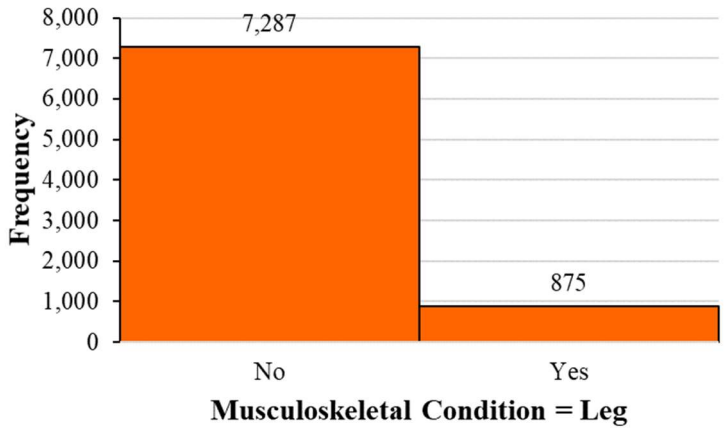


Figure 38. Frequency histogram by Musculoskeletal Condition for the Leg.

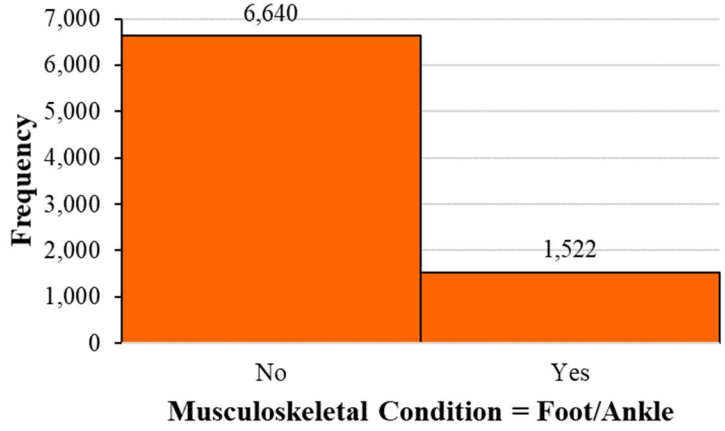


Figure 39. Frequency histogram by Musculoskeletal Condition for the Foot and Ankle.

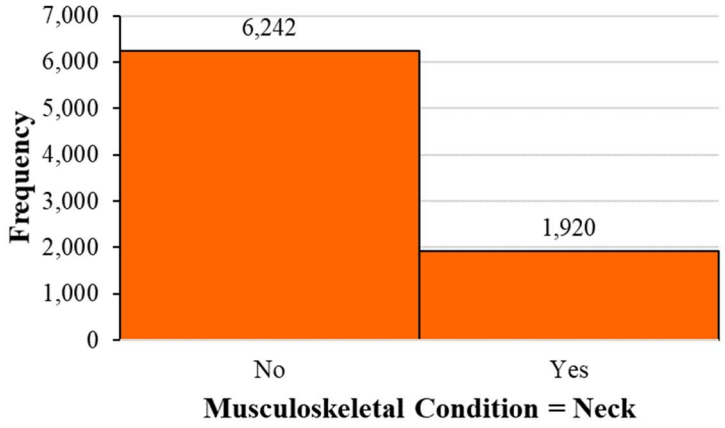


Figure 40. Frequency histogram by Musculoskeletal Condition for the Neck.

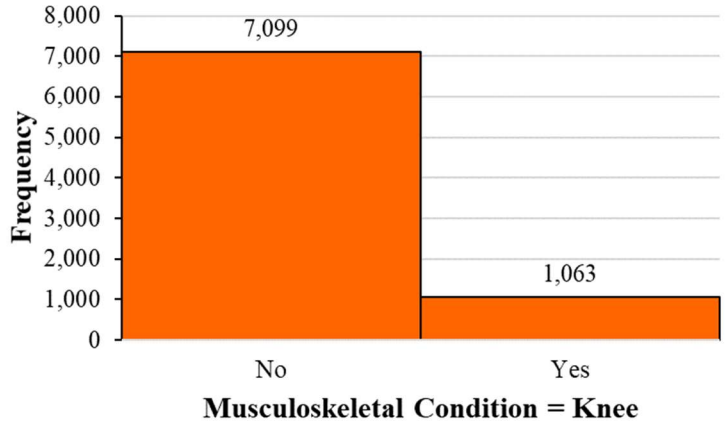


Figure 41. Frequency histogram by Musculoskeletal Condition for the Knee.

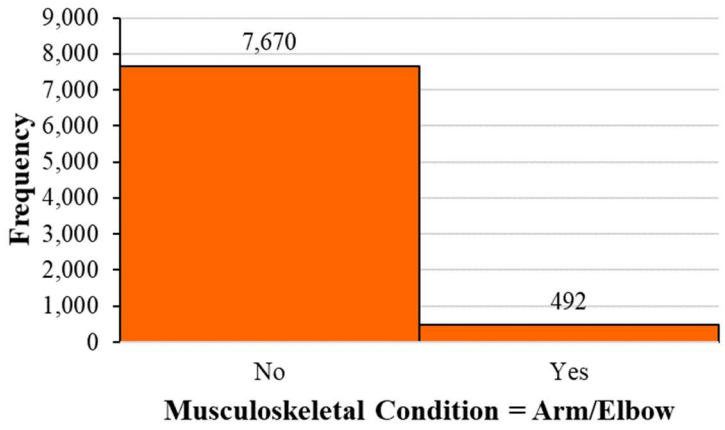


Figure 42. Frequency histogram by Musculoskeletal Condition for the Arm and Elbow.

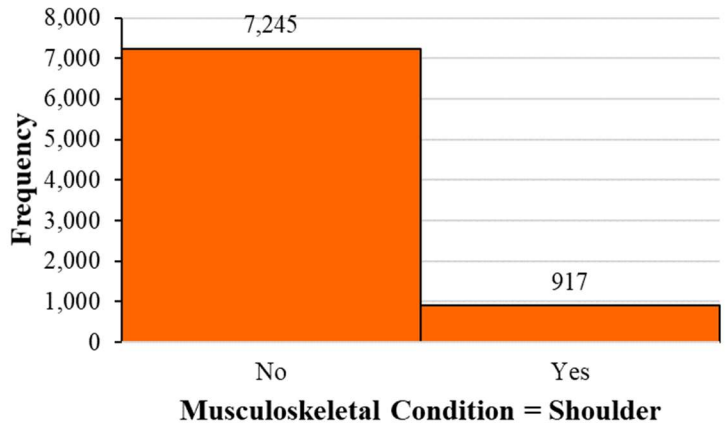


Figure 43. Frequency histogram by Musculoskeletal Condition for the Shoulder.

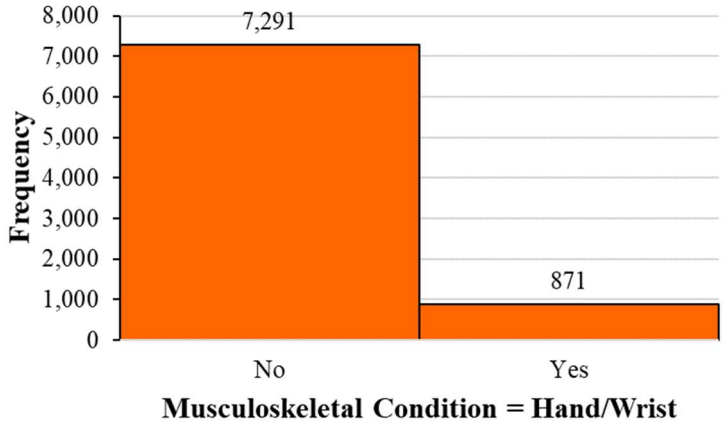


Figure 44. Frequency histogram by Musculoskeletal Condition for the Hand and Wrist.

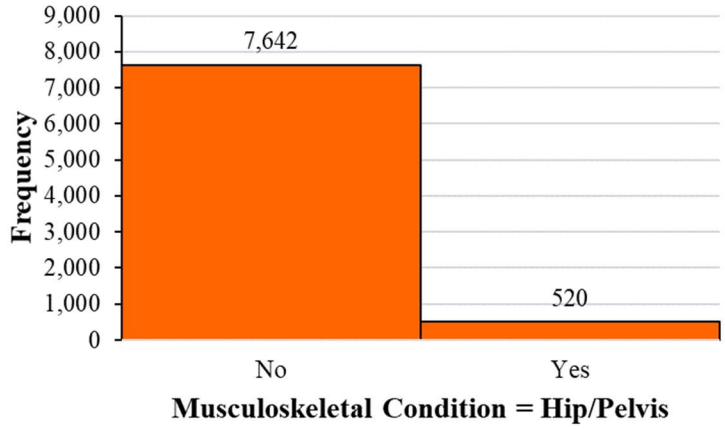


Figure 45. Frequency histogram by Musculoskeletal Condition for the Hip and Pelvis.

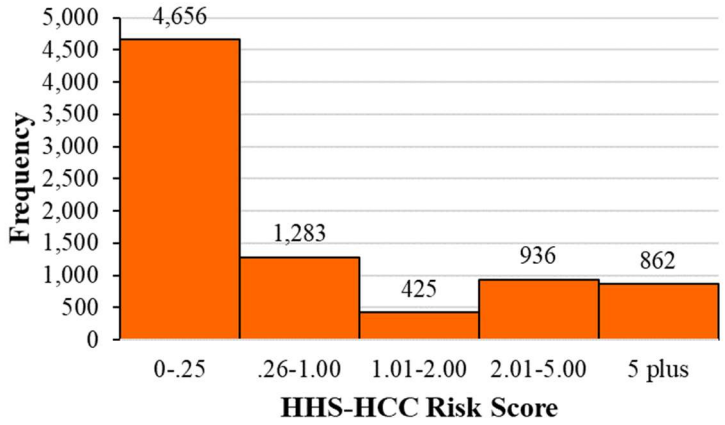


Figure 46. Frequency histogram by the HHS-HCC Risk Score.

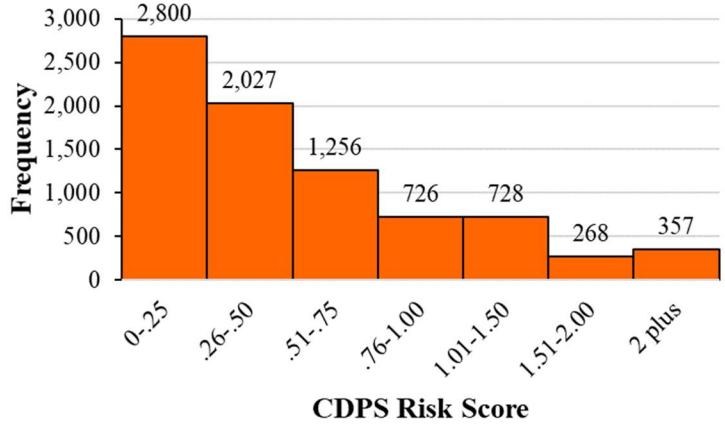


Figure 47. Frequency histogram by the CDPS Risk Score.

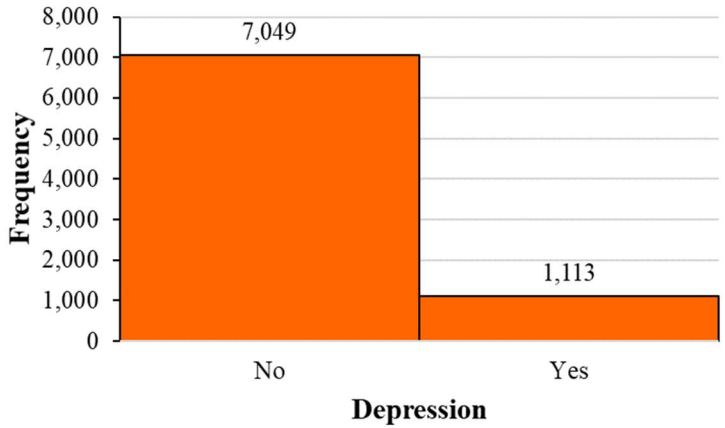


Figure 48. Frequency histogram by Depression.

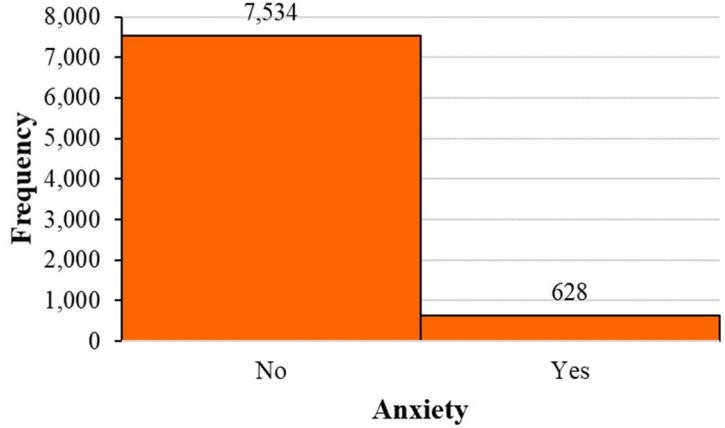


Figure 49. Frequency histogram by Anxiety.

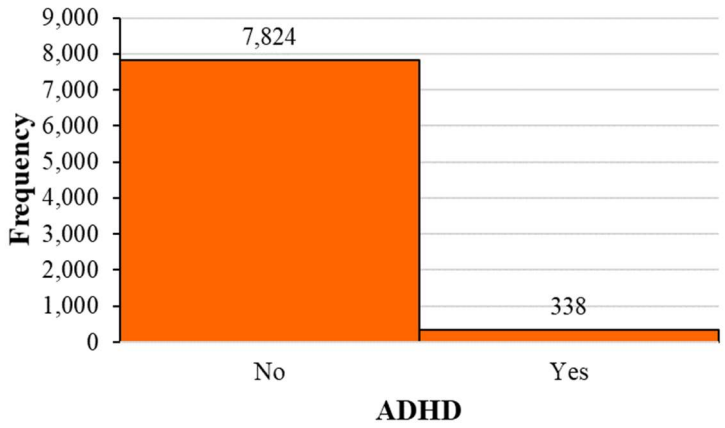


Figure 50. Frequency histogram by ADHD.

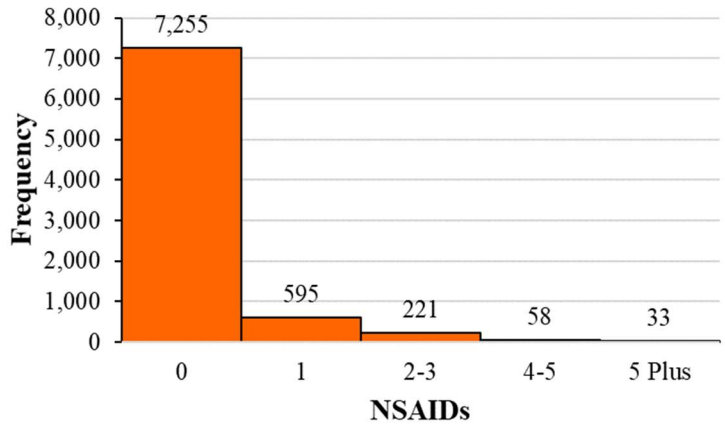


Figure 51. Frequency histogram by NSAIDs.

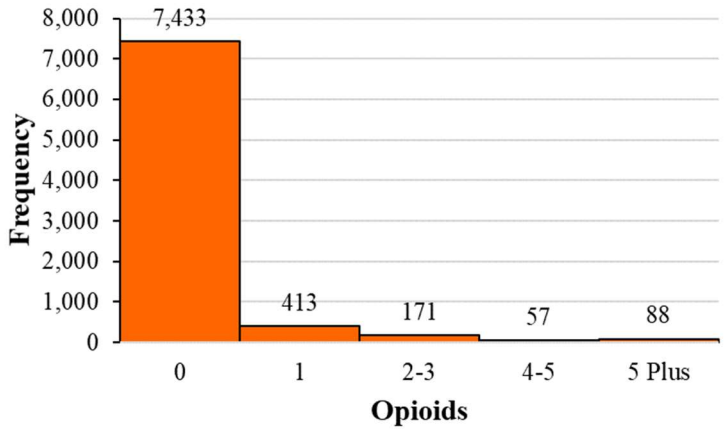


Figure 52. Frequency histogram by Opioids.

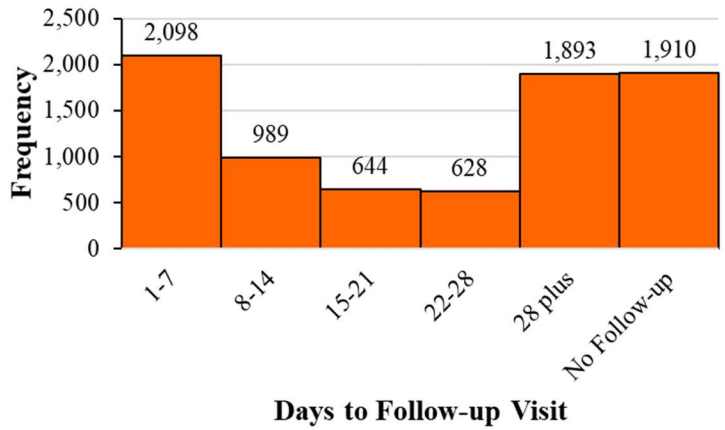


Figure 53. Frequency histogram by Days to Follow-up Visit.

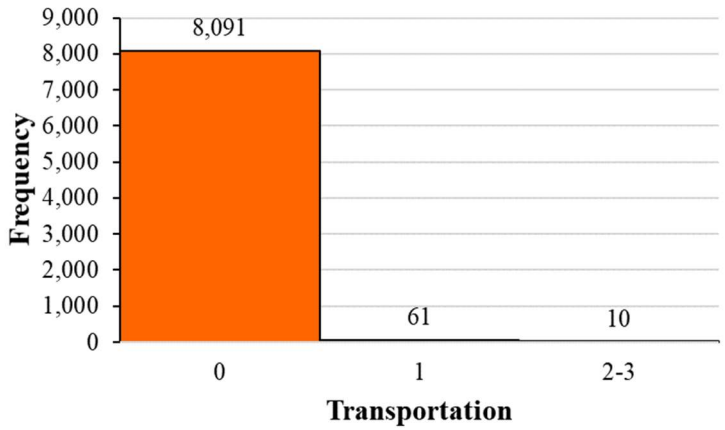


Figure 54. Frequency histogram by Transportation.

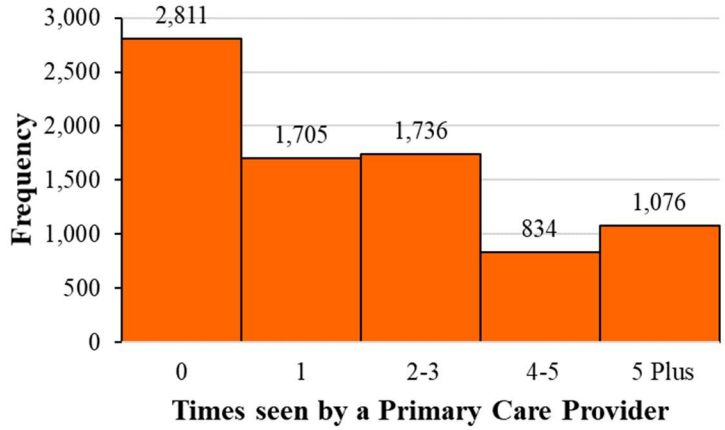


Figure 55. Frequency histogram by Times seen by a Primary Care Provider.

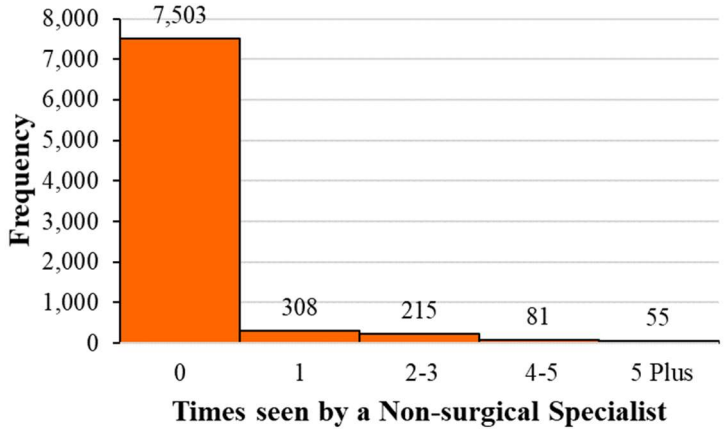


Figure 56. Frequency histogram by Times seen by a Non-surgical Specialist.

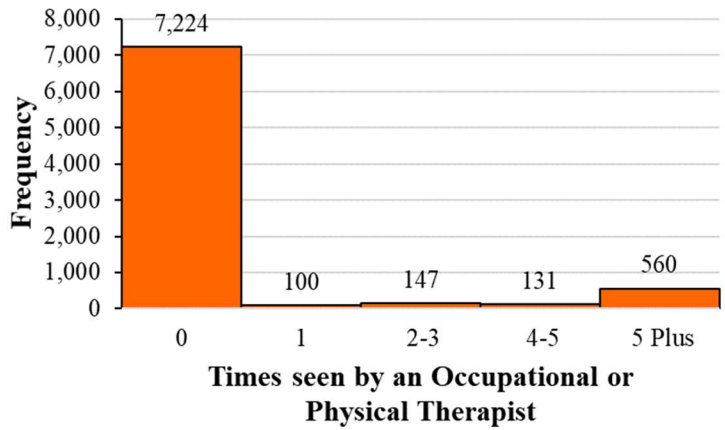


Figure 57. Frequency histogram by Times seen by an Occupational or Physical Therapist.

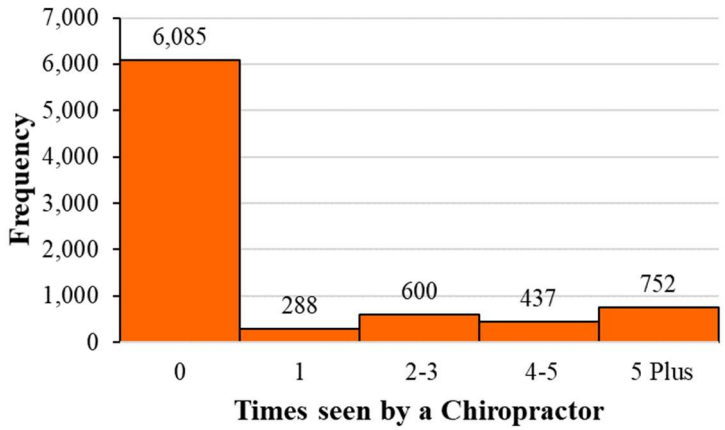


Figure 58. Frequency histogram by Times seen by a Chiropractor.

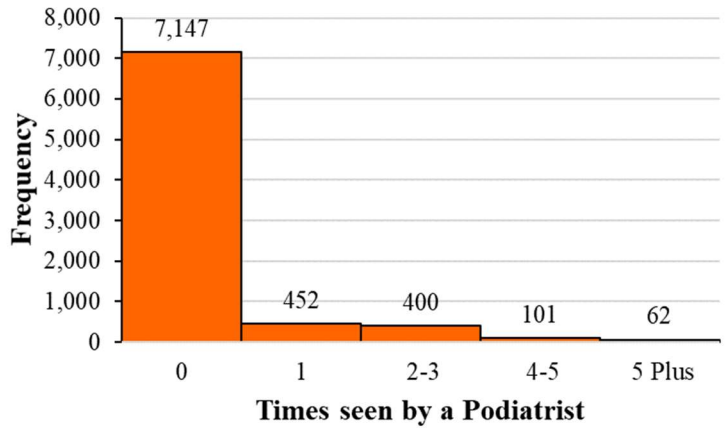


Figure 59. Frequency histogram by Times seen by a Podiatrist.

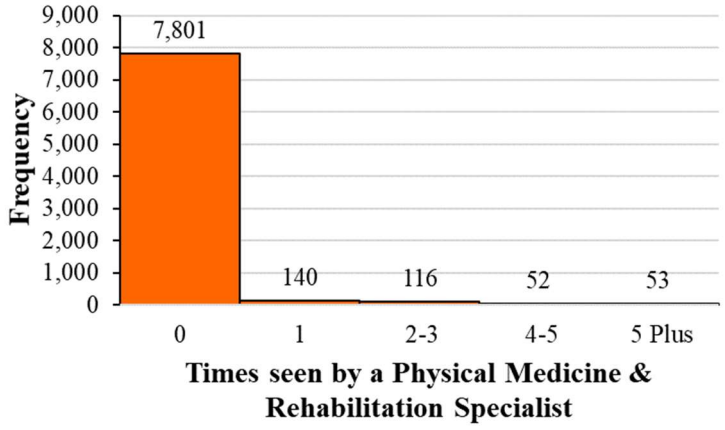


Figure 60. Frequency histogram by Times seen by a Physical Medicine and Rehabilitation Specialist.

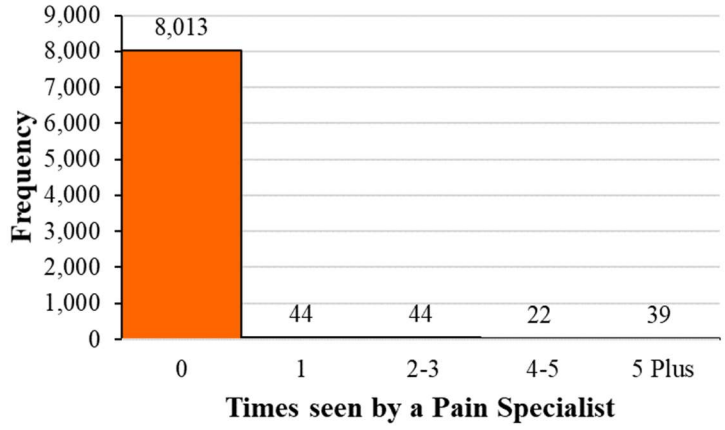


Figure 61. Frequency histogram by Times seen by a Pain Specialist.

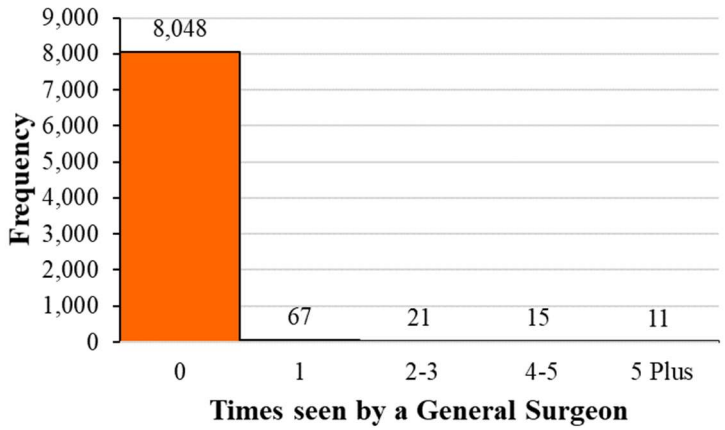


Figure 62. Frequency histogram by Times seen by a General Surgeon.

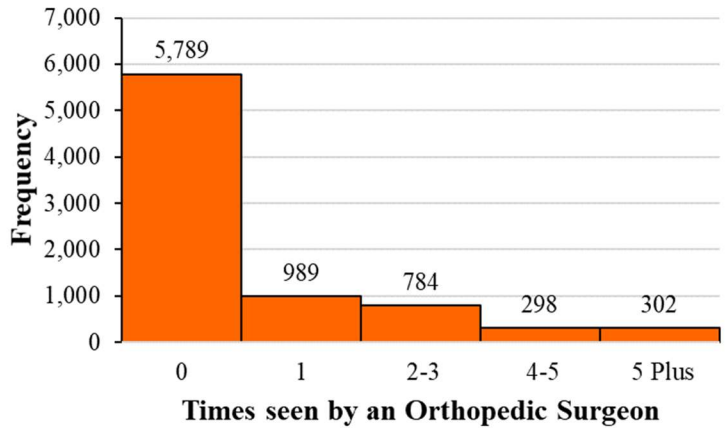


Figure 63. Frequency histogram by Times seen by an Orthopedic Surgeon.

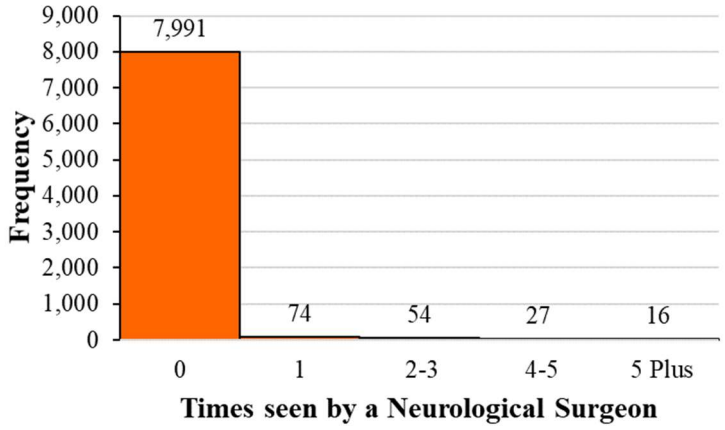


Figure 64. Frequency histogram by Times seen by a Neurological Surgeon.

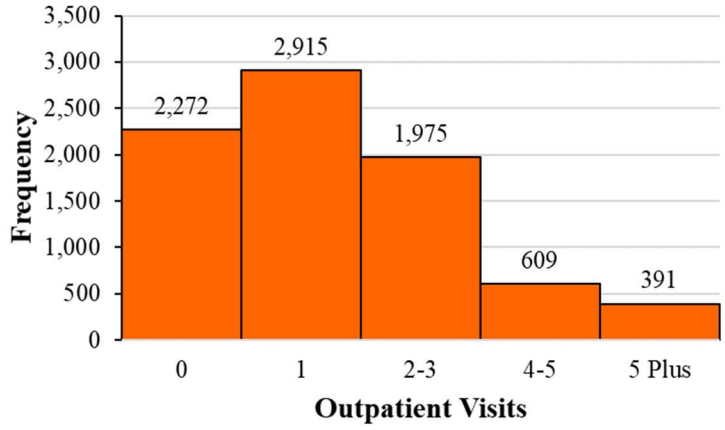


Figure 65. Frequency histogram by Outpatient Visits.

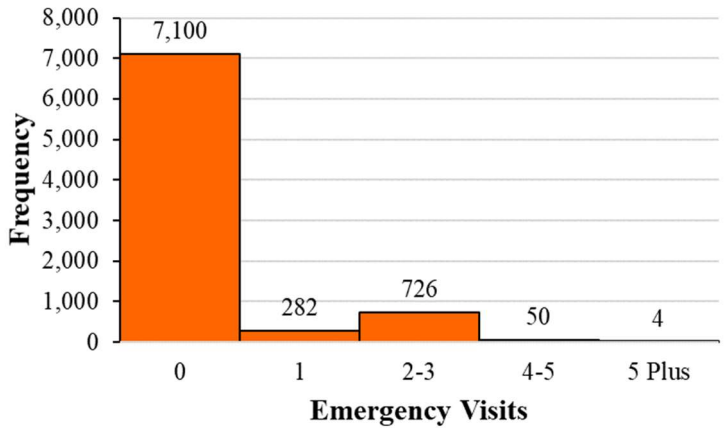


Figure 66. Frequency histogram by Emergency Visits.

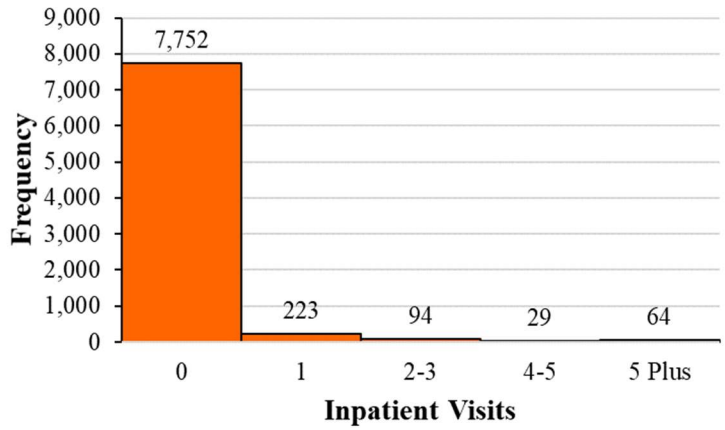


Figure 67. Frequency histogram by Inpatient Visits.

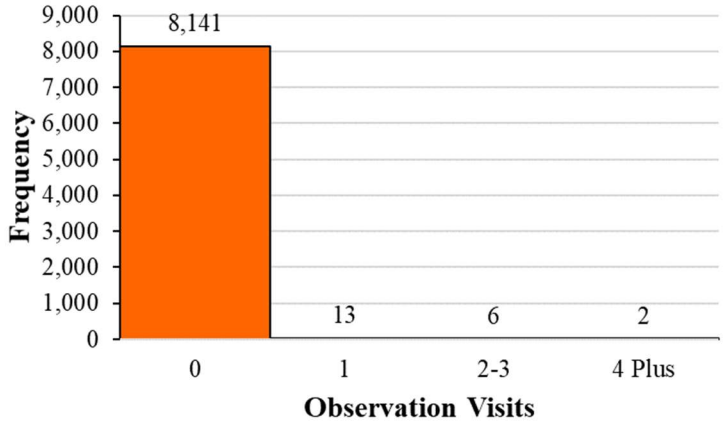


Figure 68. Frequency histogram by Observation Visits.

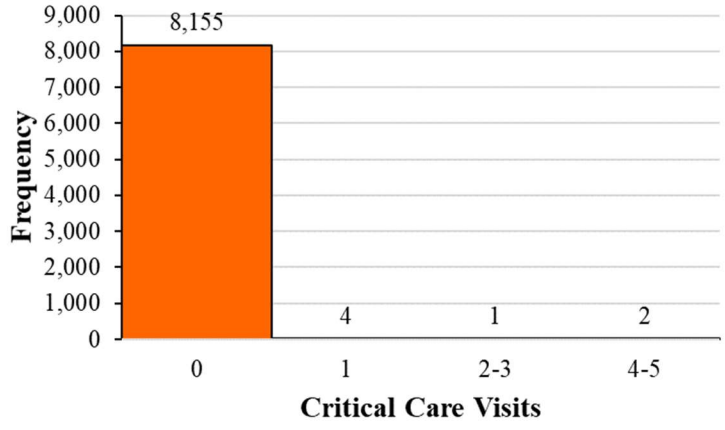


Figure 69. Frequency histogram by Critical Care Visits.

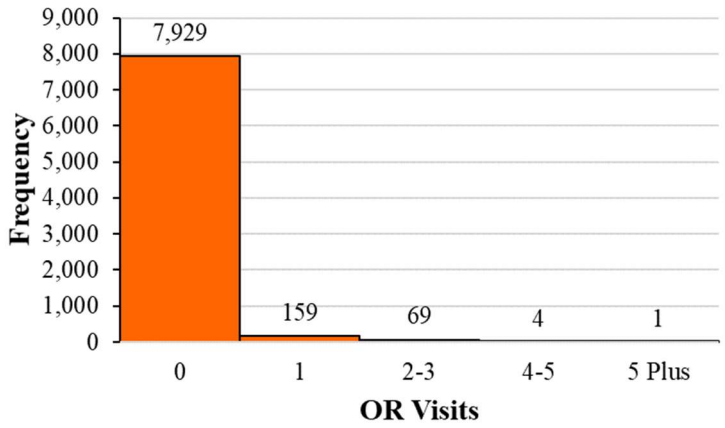


Figure 70. Frequency histogram by OR Visits.

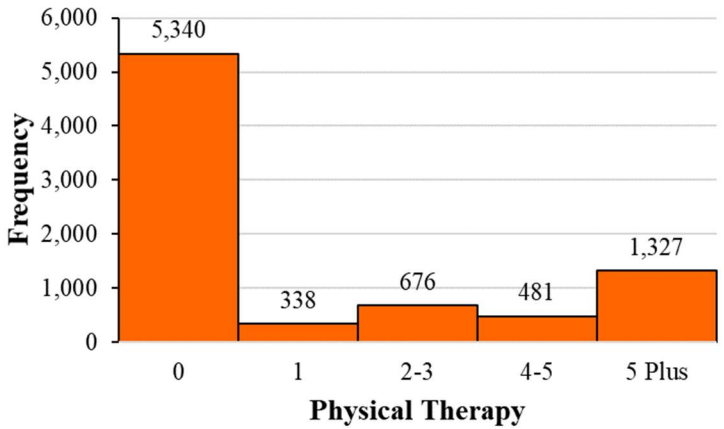


Figure 71. Frequency histogram by Physical Therapy.

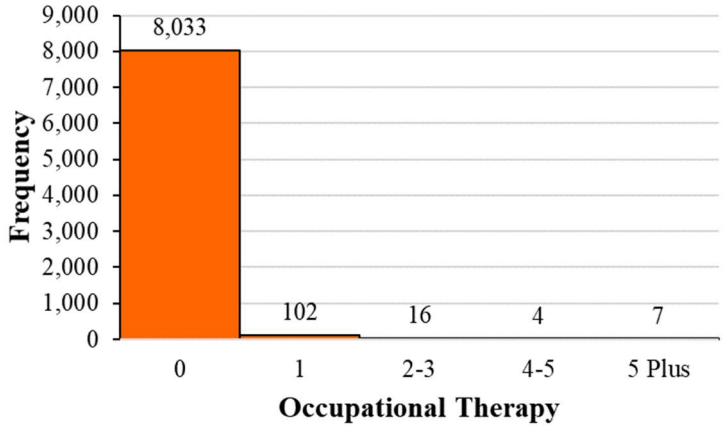


Figure 72. Frequency histogram by Occupational Therapy.

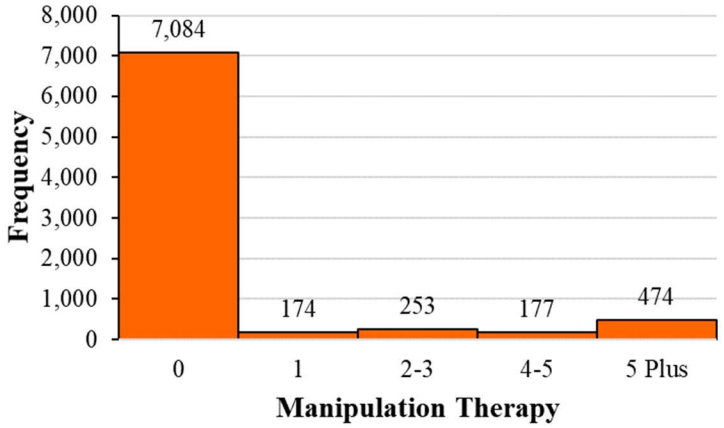


Figure 73. Frequency histogram by Manipulation Therapy.

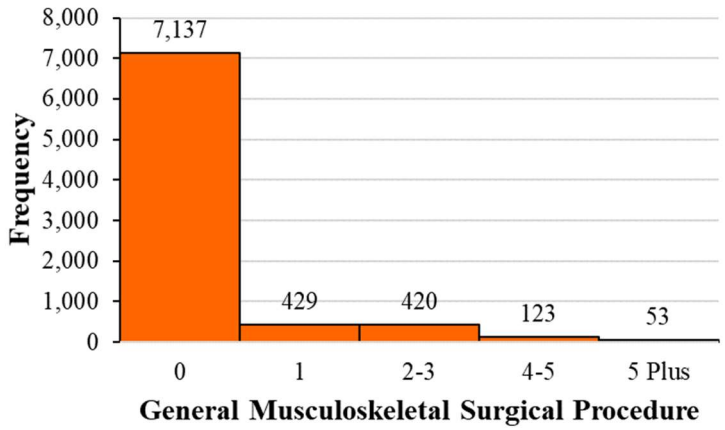


Figure 74. Frequency histogram by General Musculoskeletal Surgical Procedures.

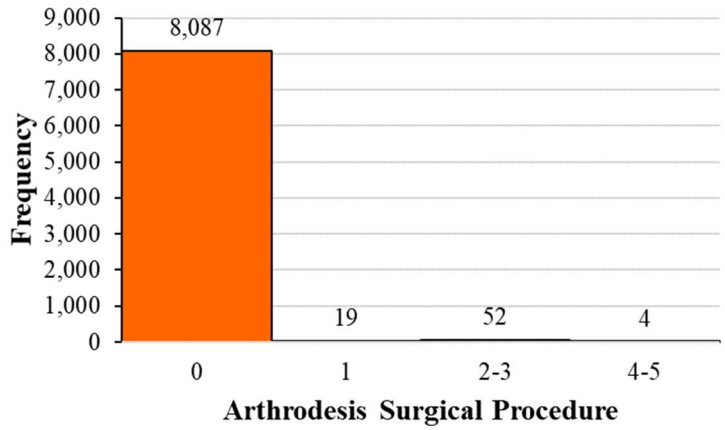


Figure 75. Frequency histogram by Arthrodesis Surgical Procedures.

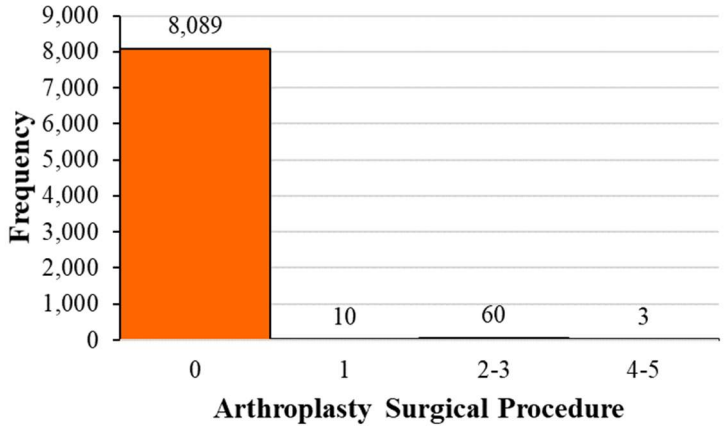


Figure 76. Frequency histogram by Arthroplasty Surgical Procedures.

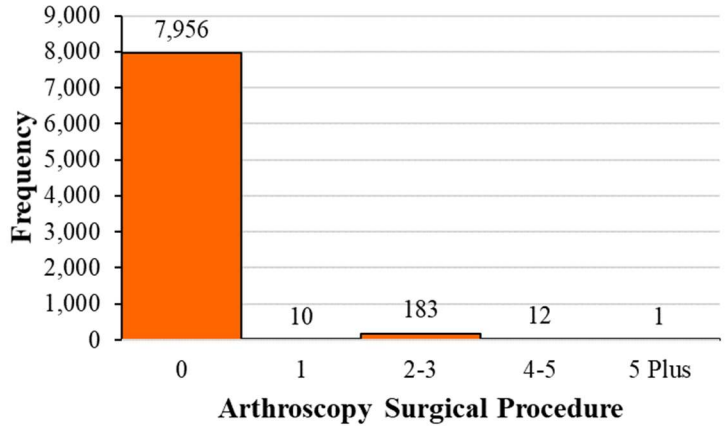


Figure 77. Frequency histogram by Arthroscopy Surgical Procedures.

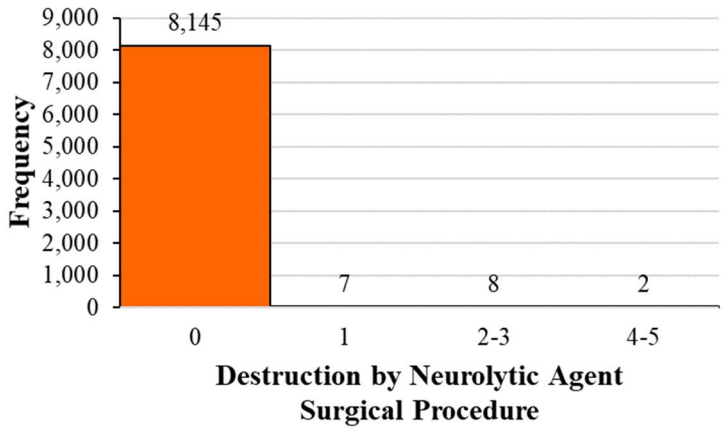


Figure 78. Frequency histogram by Destruction by Neurolytic Agent Surgical Procedures.

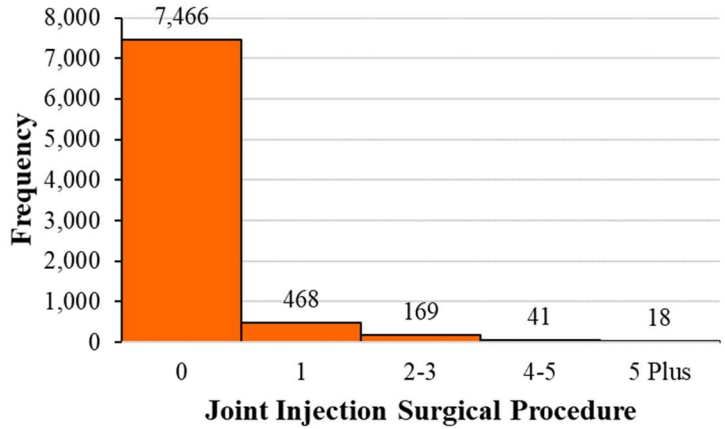


Figure 79. Frequency histogram by Joint Injection Surgical Procedures.

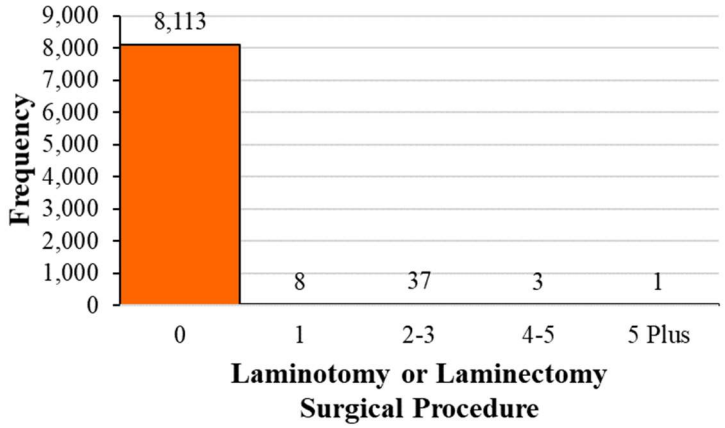


Figure 80. Frequency histogram by Laminotomy or Laminectomy Surgical Procedures.

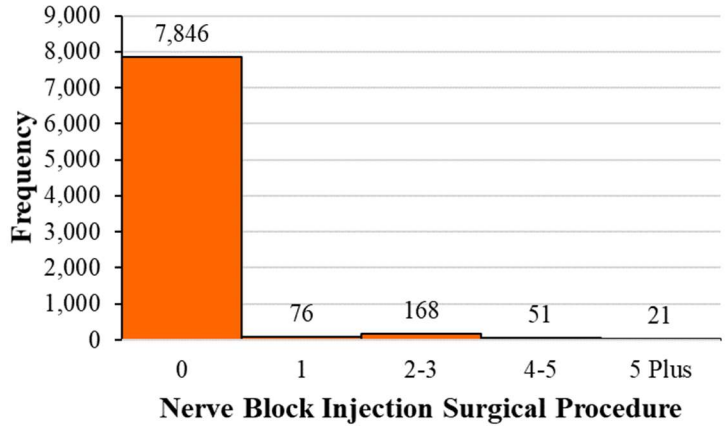


Figure 81. Frequency histogram by Nerve Block Injection Surgical Procedures.

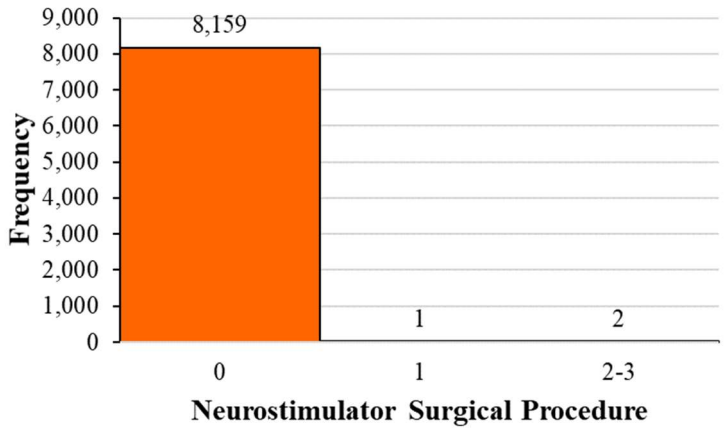


Figure 82. Frequency histogram by Neurostimulator Surgical Procedures.

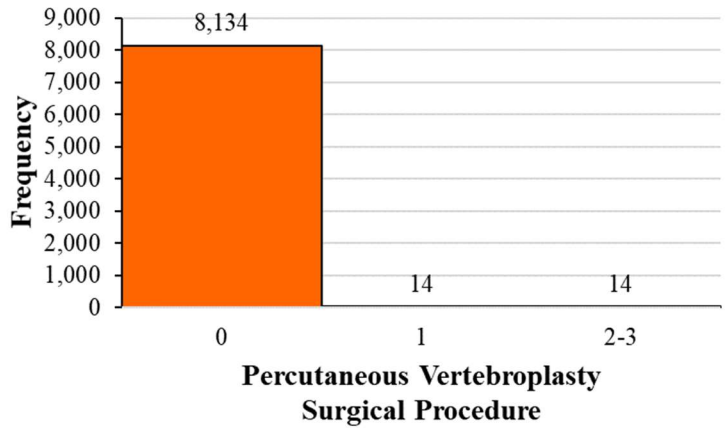


Figure 83. Frequency histogram by Percutaneous Vertebroplasty Surgical Procedures.

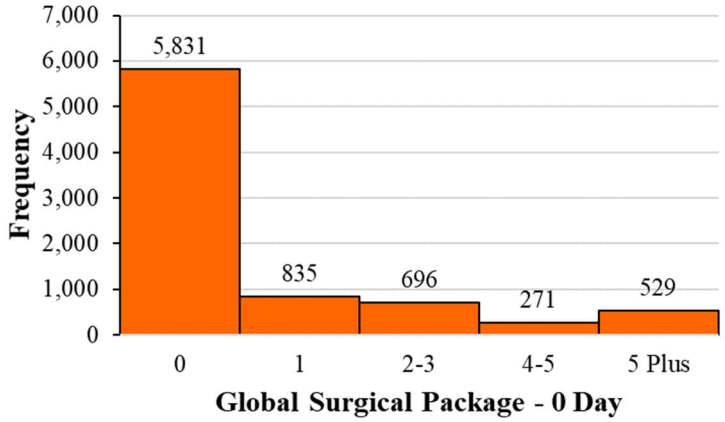


Figure 84. Frequency histogram by 0 Day Global Surgical Package.

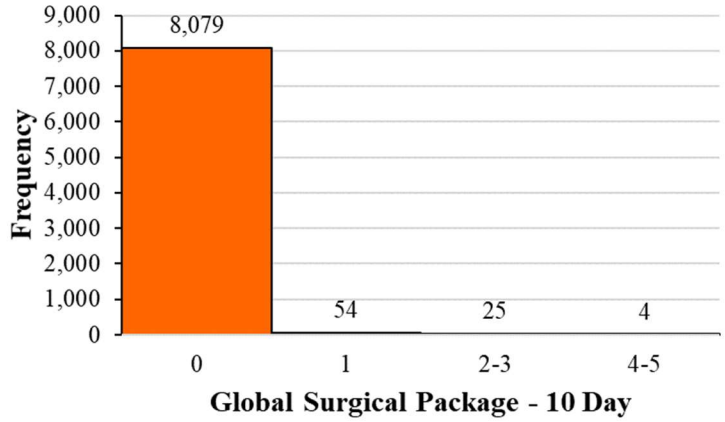


Figure 85. Frequency histogram by 10 Day Global Surgical Package.

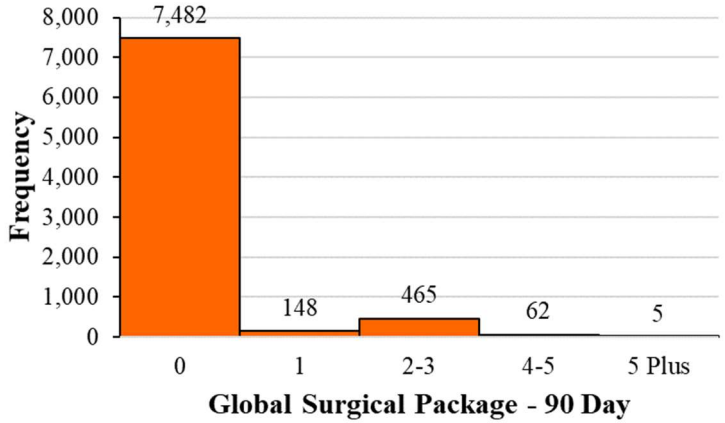


Figure 86. Frequency histogram by 90 Day Global Surgical Package.

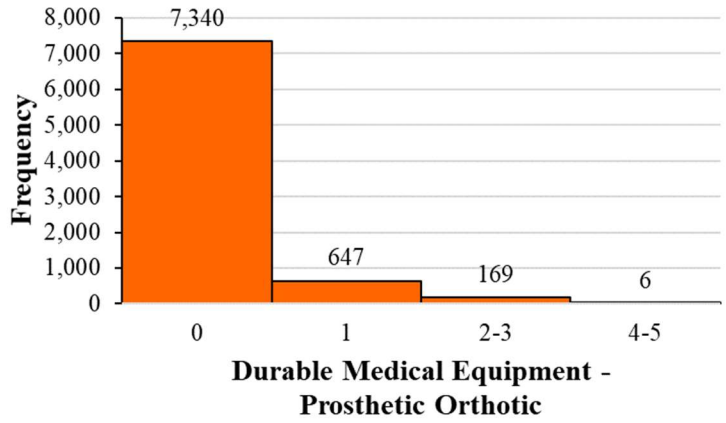


Figure 87. Frequency histogram by Durable Medical Equipment (DME) – Prosthetic and Orthotics.

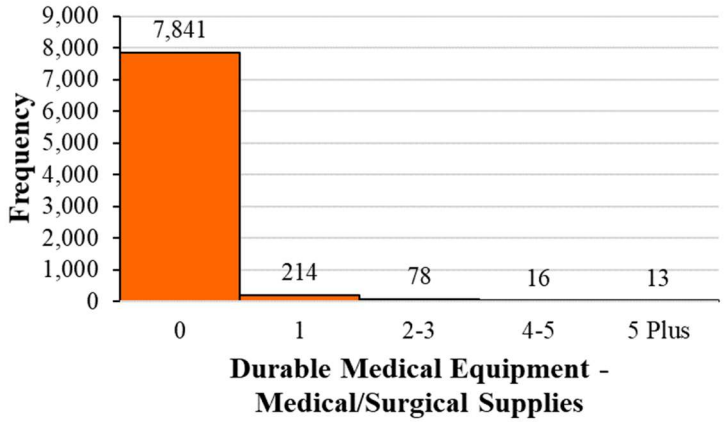


Figure 88. Frequency histogram by Durable Medical Equipment (DME) – Medical and Surgical Supplies.

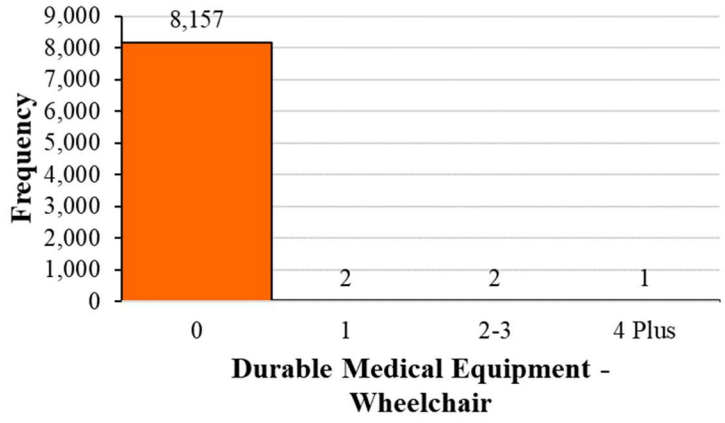


Figure 89. Frequency histogram by Durable Medical Equipment (DME) - Wheelchair.

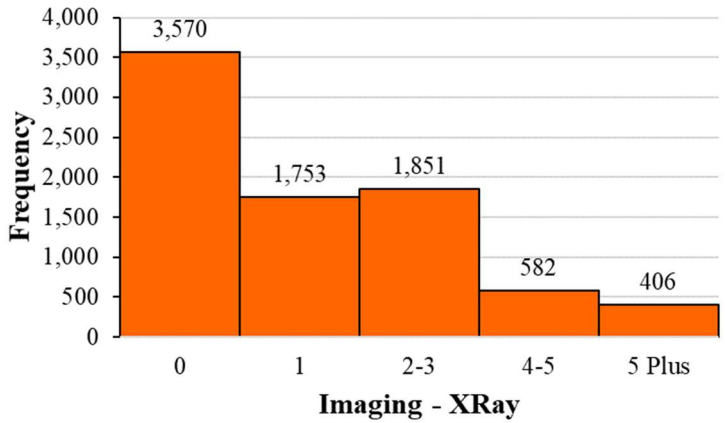


Figure 90. Frequency histogram by X-Ray Imaging.

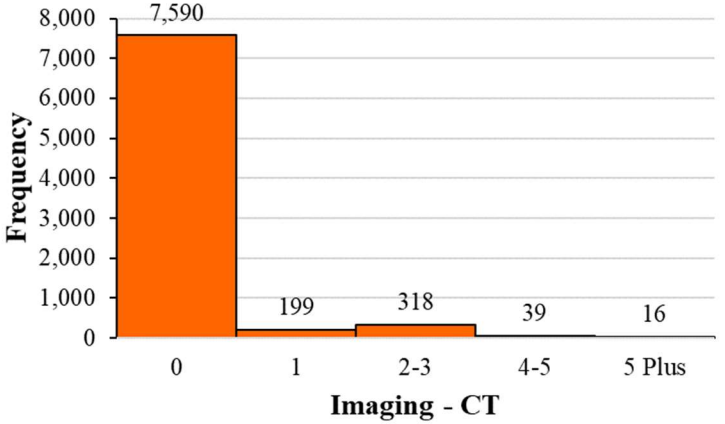


Figure 91. Frequency histogram by Computed Tomography (CT) Imaging.

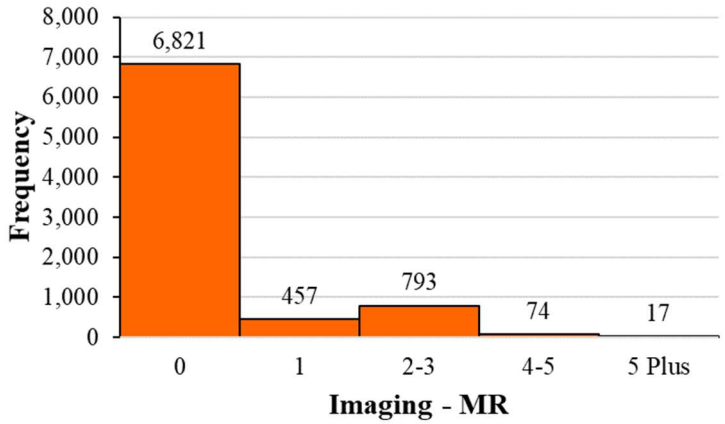


Figure 92. Frequency histogram by Magnetic Resonance (MR) Imaging.

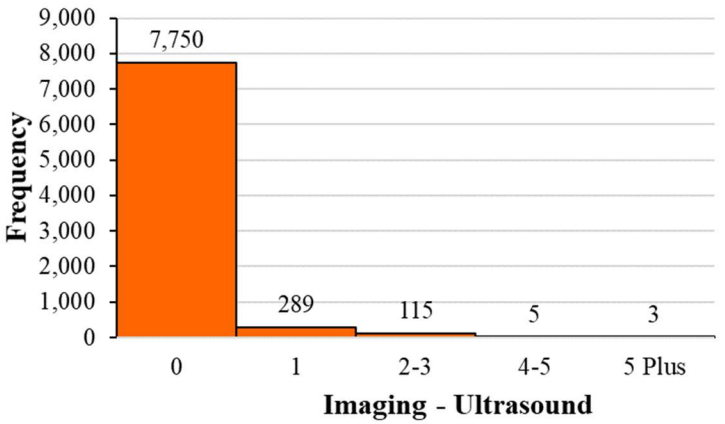


Figure 93. Frequency histogram by Ultrasound Imaging.

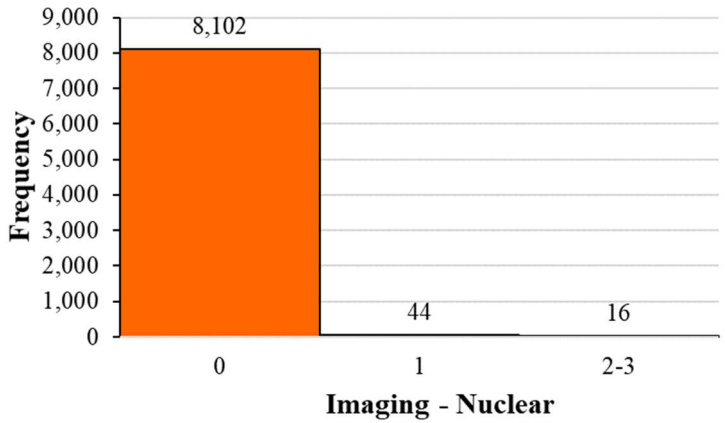


Figure 94. Frequency histogram by Nuclear Imaging.

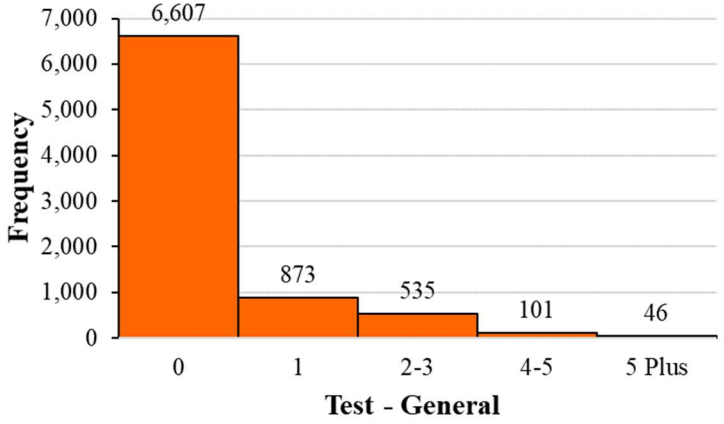


Figure 95. Frequency histogram by General Test.

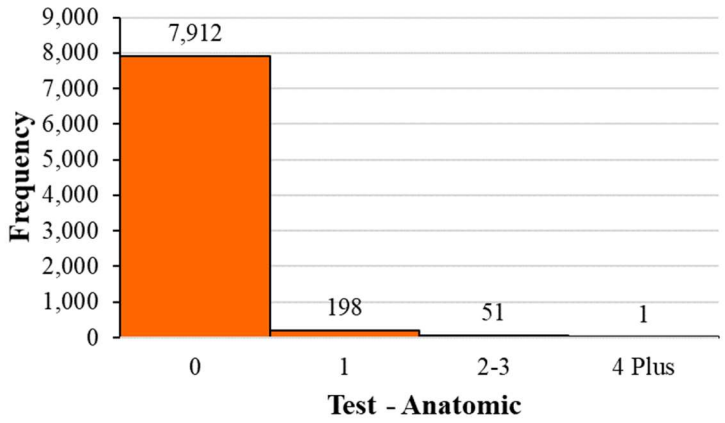


Figure 96. Frequency histogram by Anatomic Test.

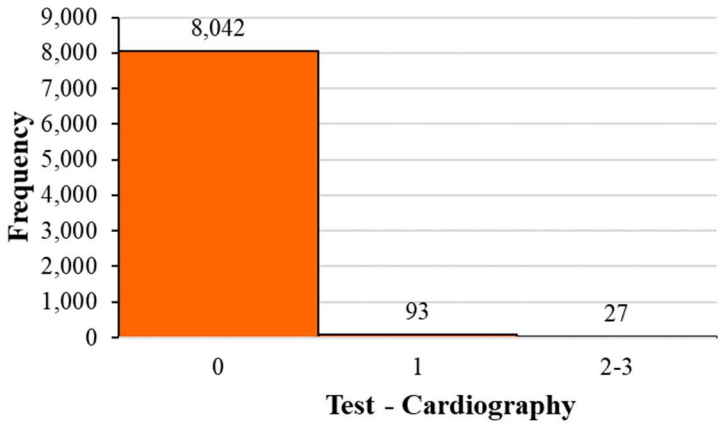


Figure 97. Frequency histogram by Cardiography Test.

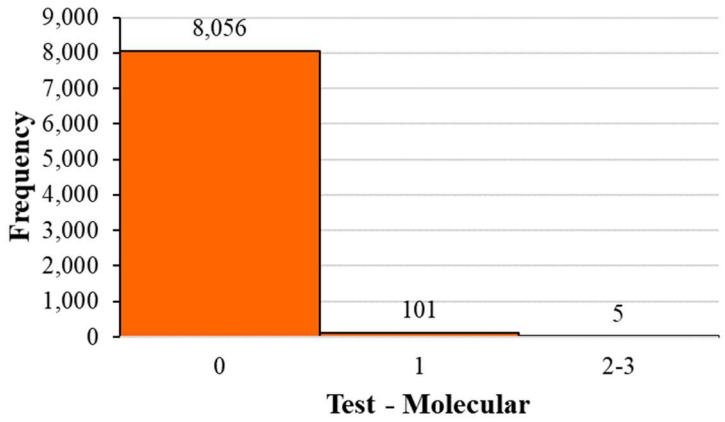


Figure 98. Frequency histogram by Molecular Test.

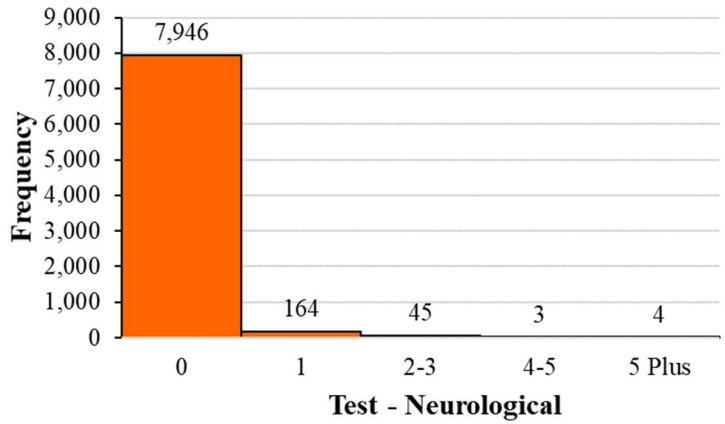


Figure 99. Frequency histogram by Neurological Test.

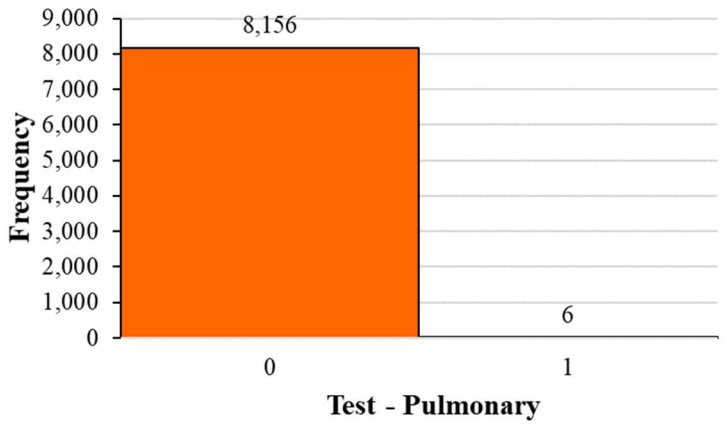


Figure 100. Frequency histogram by Pulmonary Test.

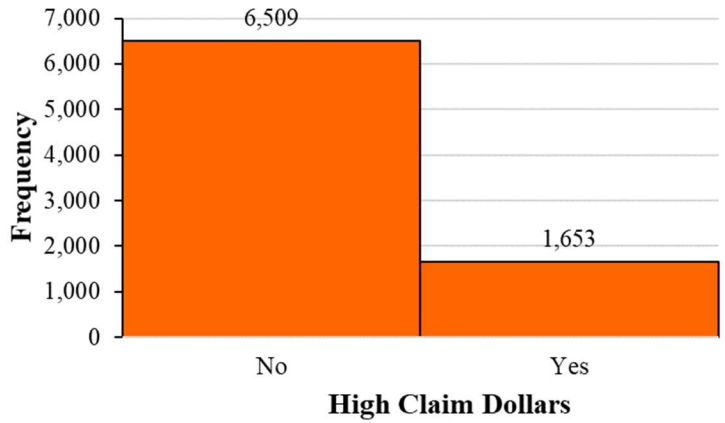


Figure 101. Frequency histogram by High Claim Dollars.

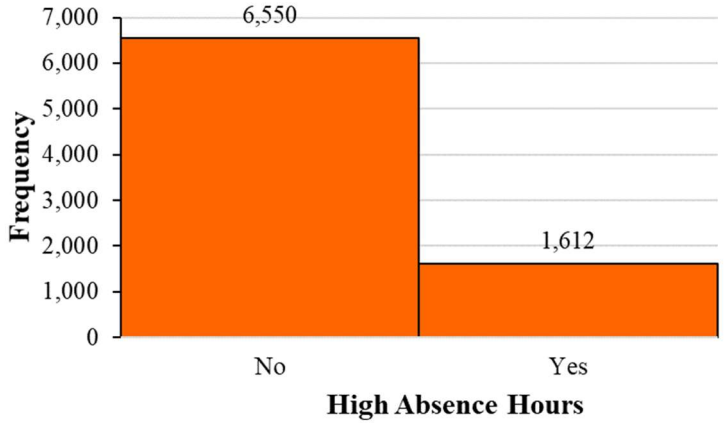


Figure 102. Frequency histogram by High Absence Hours.

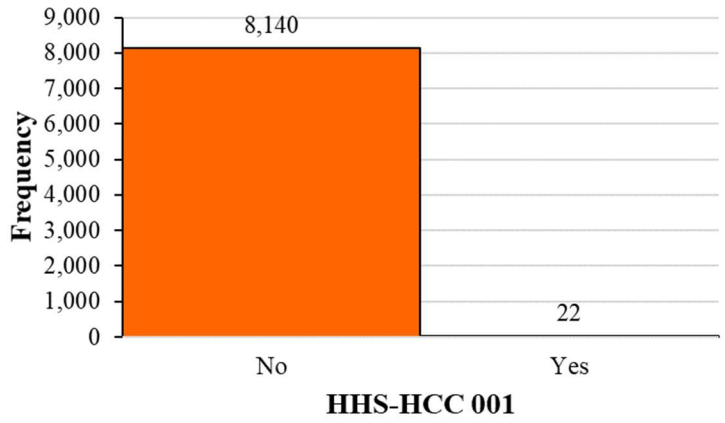


Figure 103. Frequency histogram by HIV/AIDS.

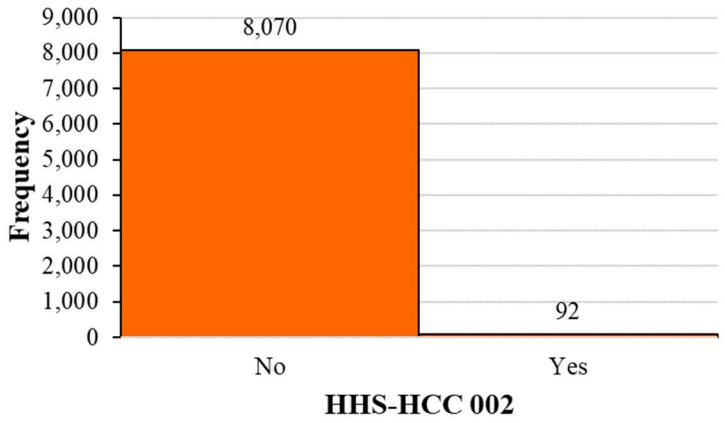


Figure 104. Frequency histogram by Septicemia, Sepsis, Systemic Inflammatory Response Syndrome/Shock.

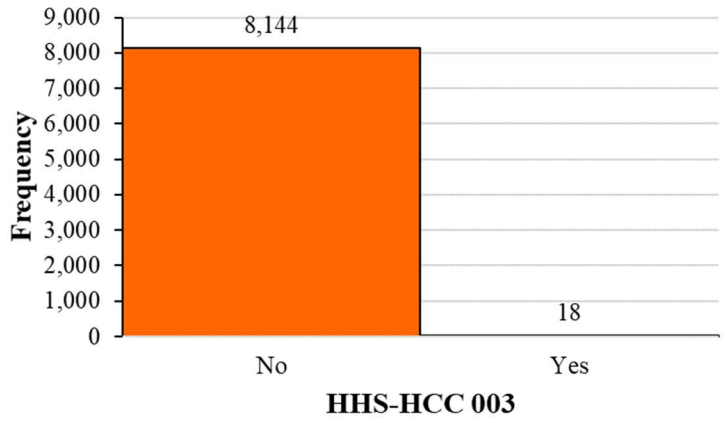


Figure 105. Frequency histogram by Central Nervous System Infections, Except Viral Meningitis.

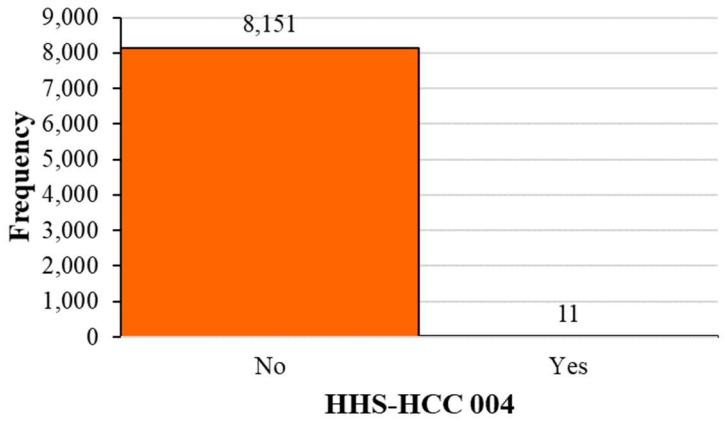


Figure 106. Frequency histogram by Viral or Unspecified Meningitis.

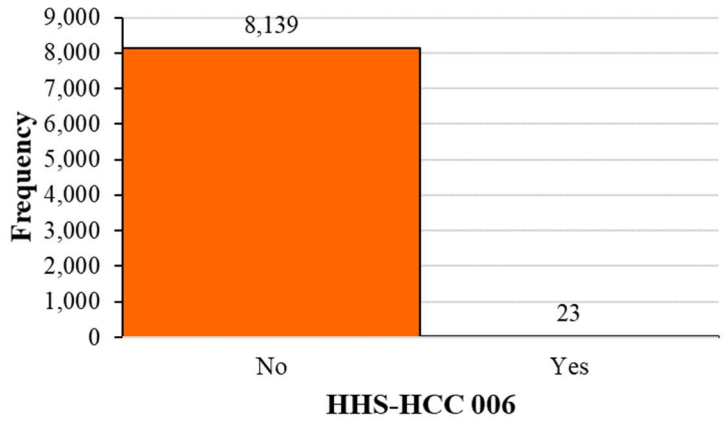


Figure 107. Frequency histogram by Opportunistic Infections.

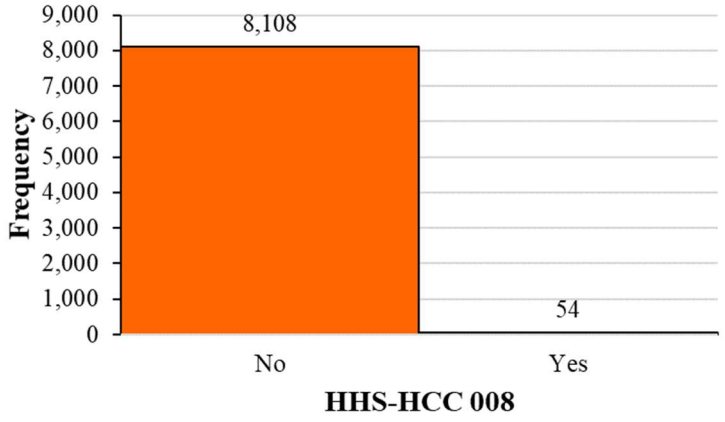


Figure 108. Frequency histogram by Metastatic Cancer.

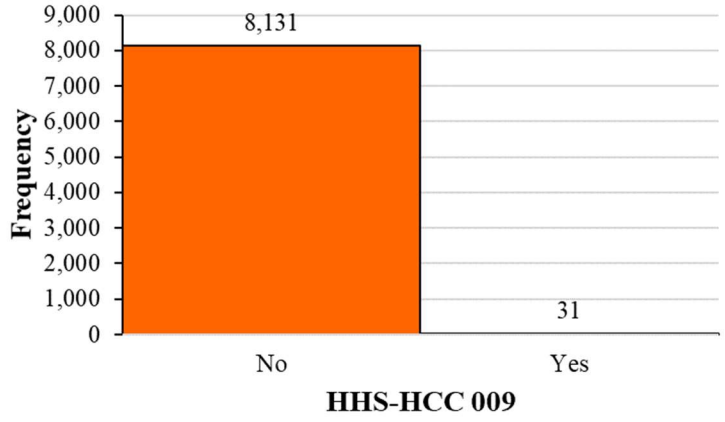


Figure 109. Frequency histogram by Lung, Brain, and Other Severe Cancers, Including Pediatric Acute Lymphoid Leukemia.

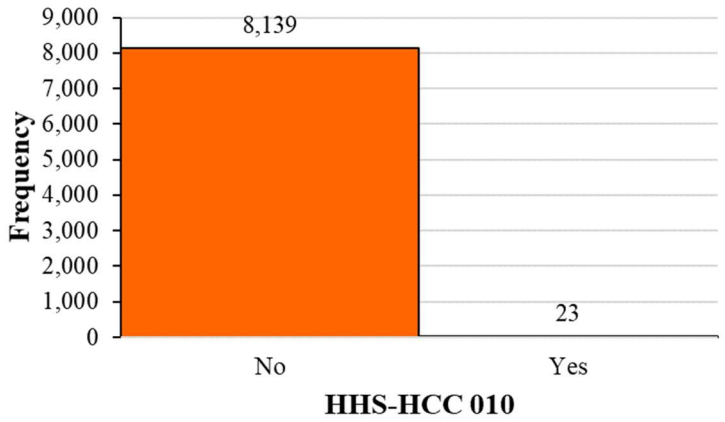


Figure 110. Frequency histogram by Non-Hodgkin Lymphomas and Other Cancers and Tumors.

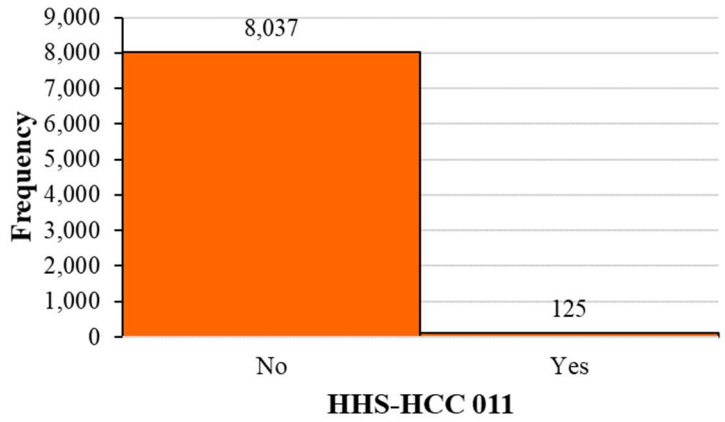


Figure 111. Frequency histogram by Colorectal, Breast (Age < 50), Kidney, and Other Cancers.

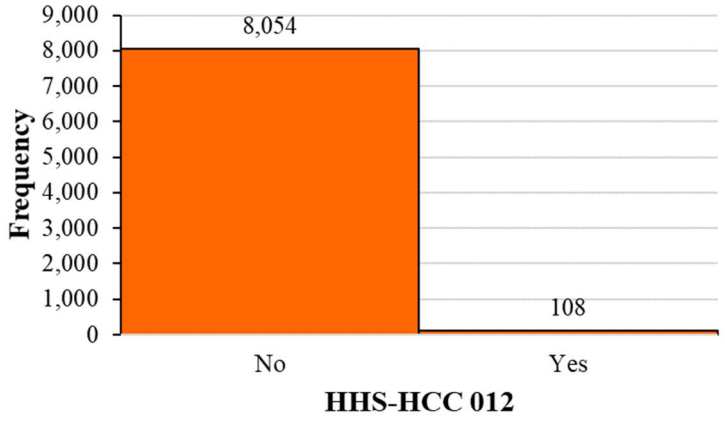


Figure 112. Frequency histogram by Breast (Age 50+) and Prostate Cancer, Benign/Uncertain Brain Tumors, and Other Cancers and Tumors.

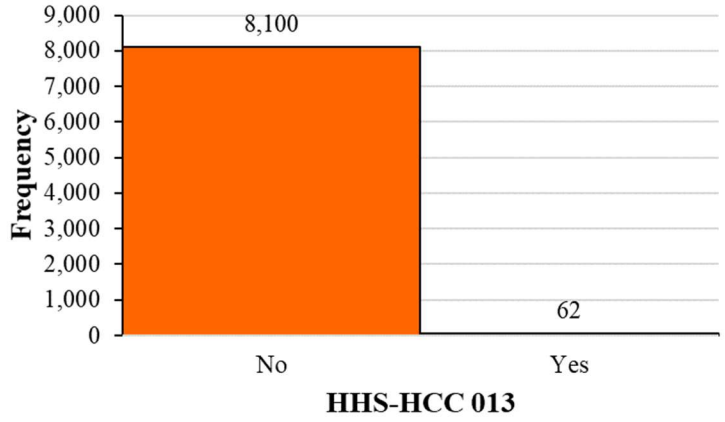


Figure 113. Frequency histogram by Thyroid Cancer, Melanoma, Neurofibromatosis, and Other Cancers and Tumors.

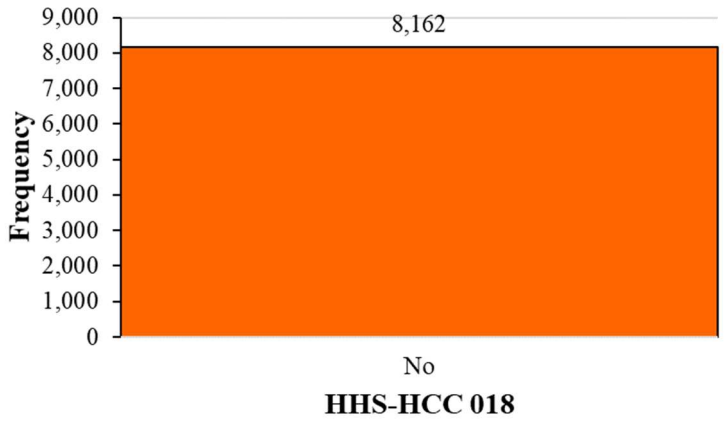


Figure 114. Frequency histogram by Pancreas Transplant Status.

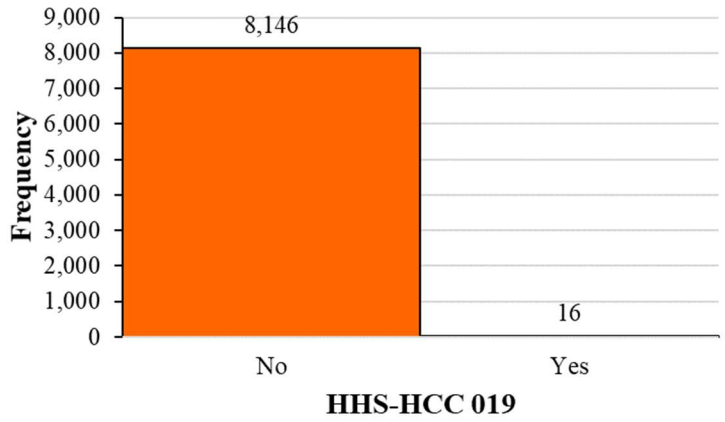


Figure 115. Frequency histogram by Diabetes with Acute Complications.

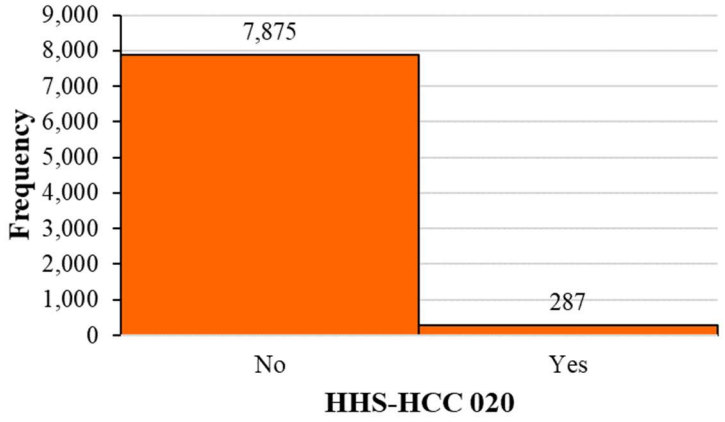


Figure 116. Frequency histogram by Diabetes with Chronic Complications.

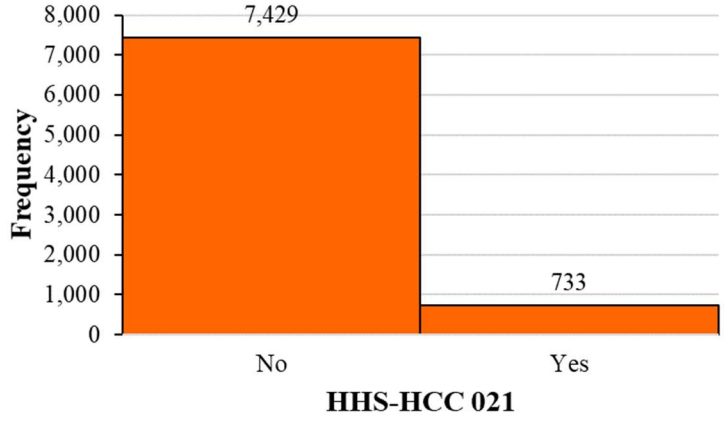


Figure 117. Frequency histogram by Diabetes without Complication.

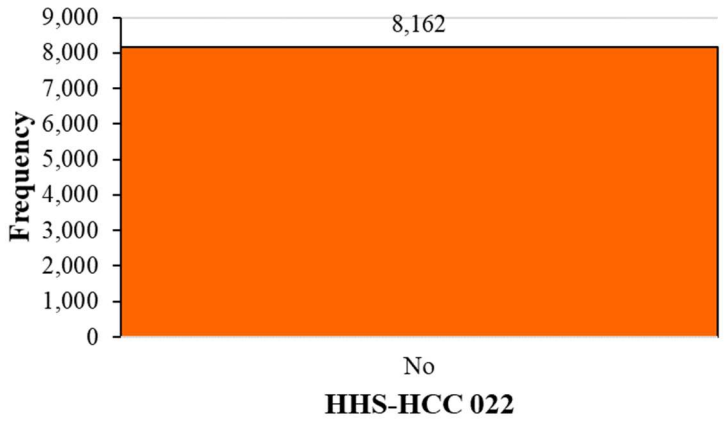


Figure 118. Frequency histogram by Type 1 Diabetes Mellitus, add-on to Diabetes HCCs 19-21.

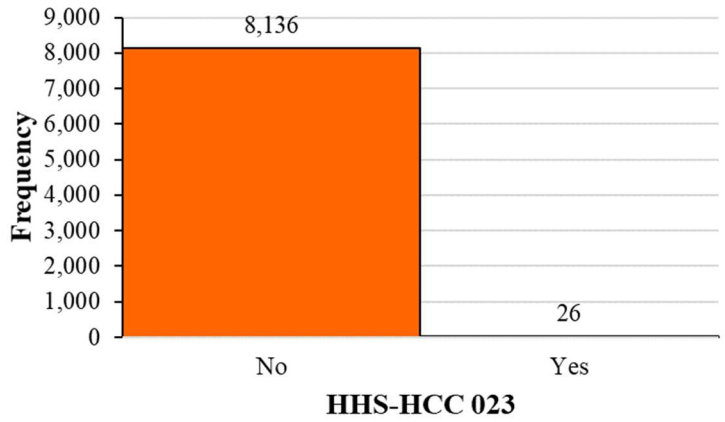


Figure 119. Frequency histogram by Protein-Calorie Malnutrition.

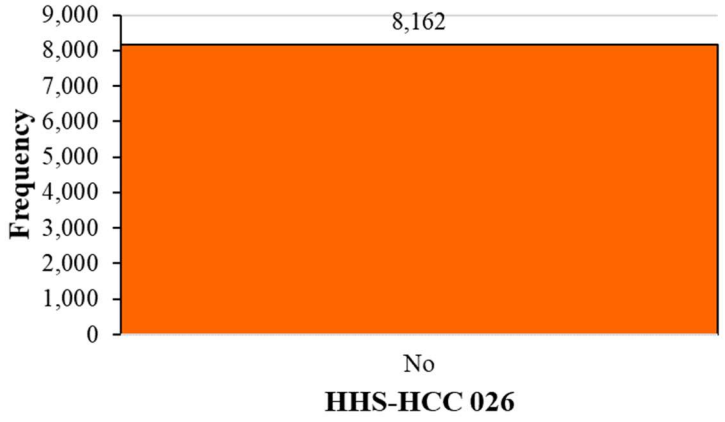


Figure 120. Frequency histogram by Mucopolysaccharidosis.

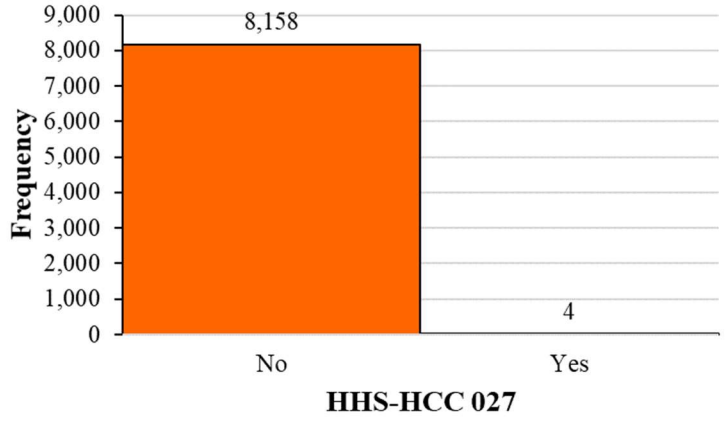


Figure 121. Frequency histogram by Lipidoses and Glycogenosis.

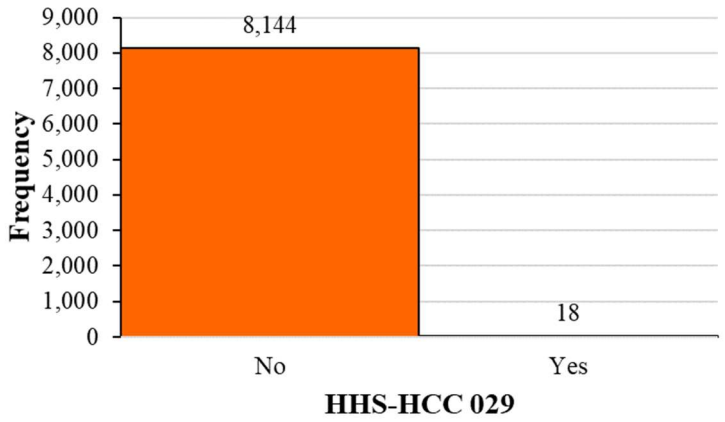


Figure 122. Frequency histogram by Amyloidosis, Porphyria, and Other Metabolic Disorders.

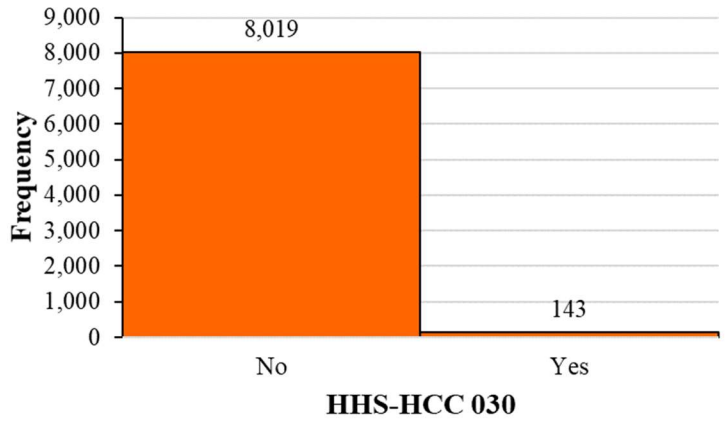


Figure 123. Frequency histogram by Adrenal, Pituitary, and Other Significant Endocrine Disorders.

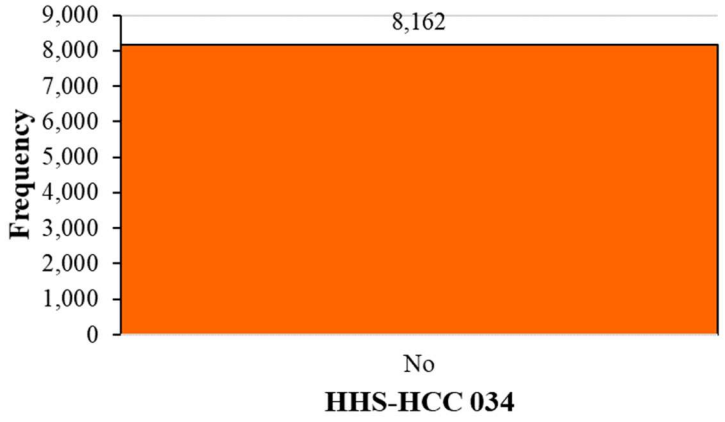


Figure 124. Frequency histogram by Liver Transplant Status/Complications.

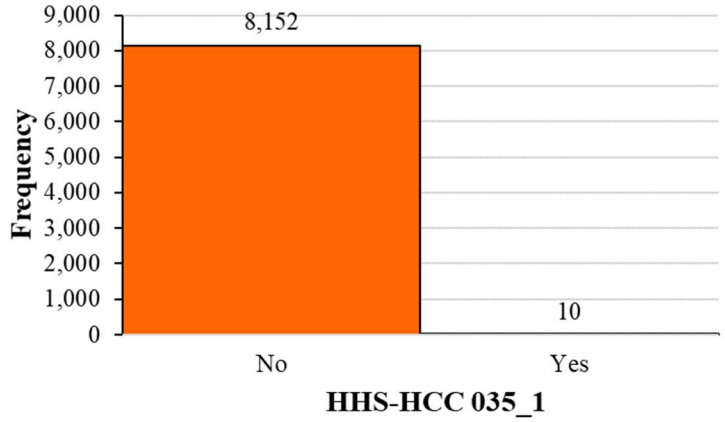


Figure 125. Frequency histogram by Acute Liver Failure/Disease, Including Neonatal Hepatitis.

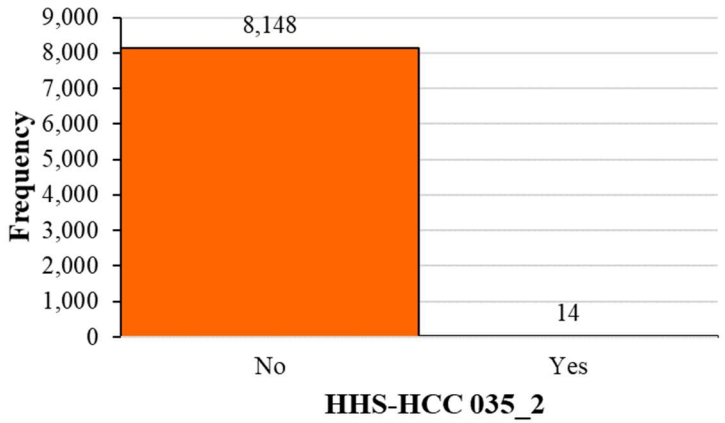


Figure 126. Frequency histogram by Chronic Liver Failure/End-Stage Liver Disorders.

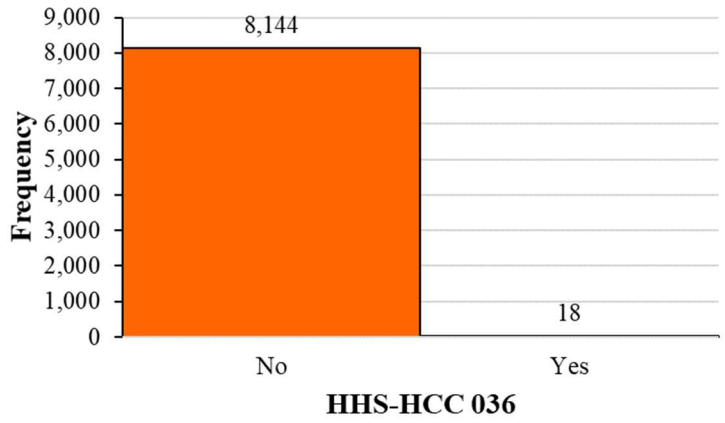


Figure 127. Frequency histogram by Cirrhosis of Liver.

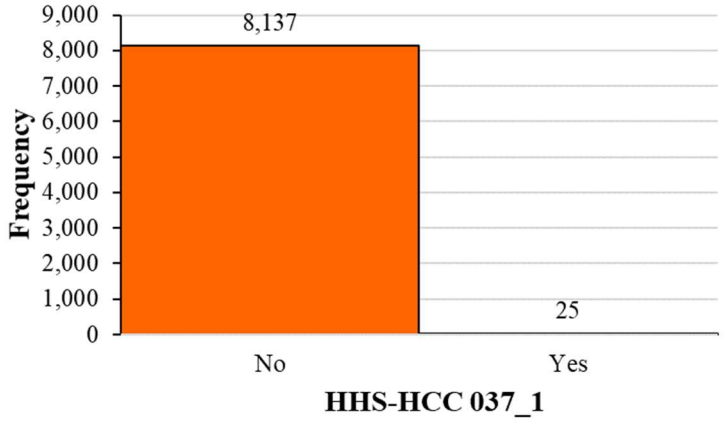


Figure 128. Frequency histogram by Chronic Viral Hepatitis C.

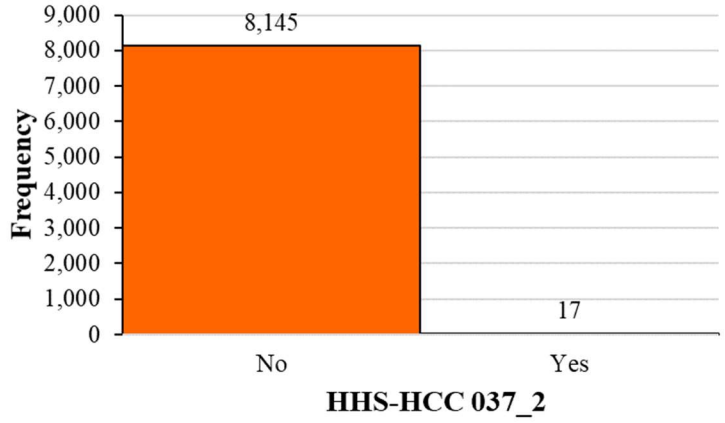


Figure 129. Frequency histogram by Chronic Hepatitis, Except Chronic Viral Hepatitis C.

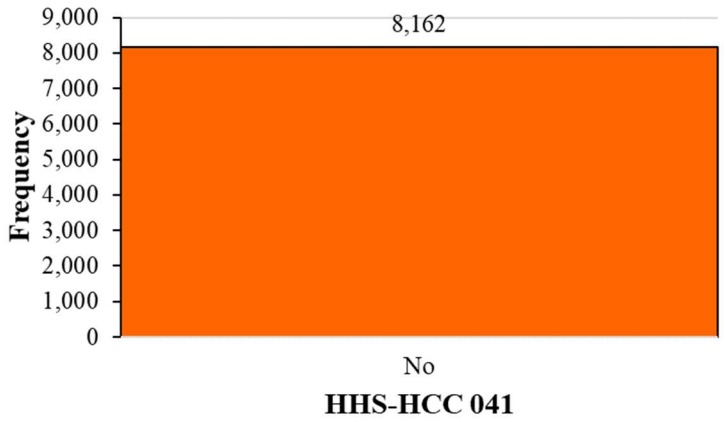


Figure 130. Frequency histogram by Intestine Transplant Status/Complications.

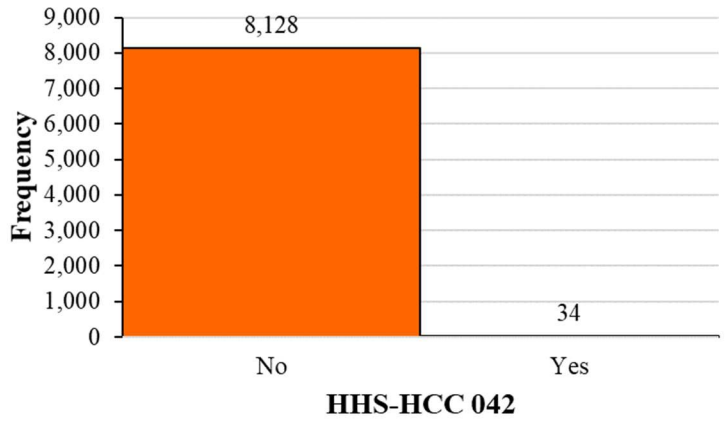


Figure 131. Frequency histogram by Peritonitis/Gastrointestinal Perforation/Necrotizing Enterocolitis.

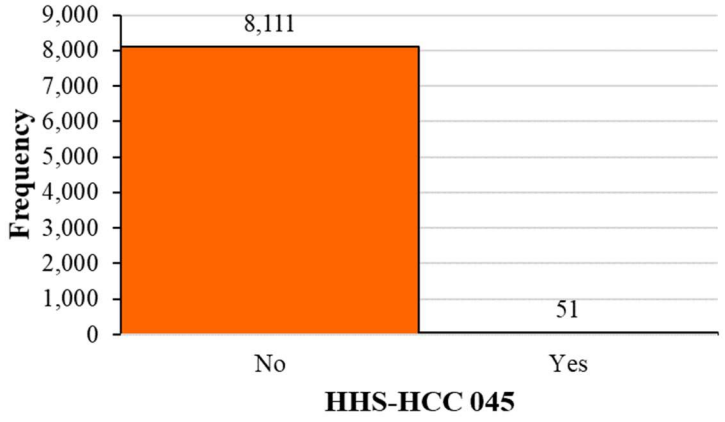


Figure 132. Frequency histogram by Intestinal Obstruction.

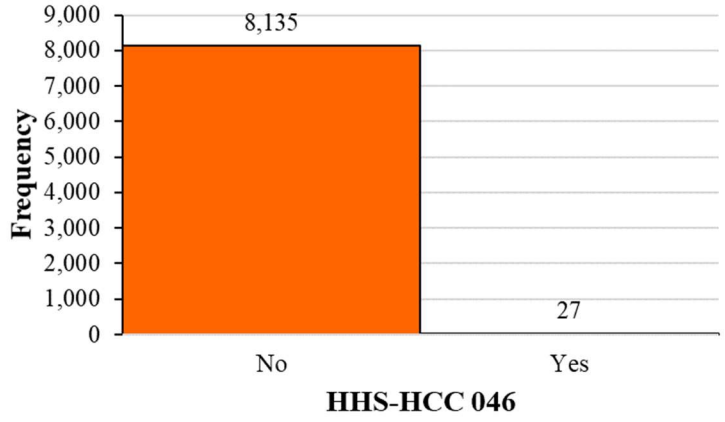


Figure 133. Frequency histogram by Chronic Pancreatitis.

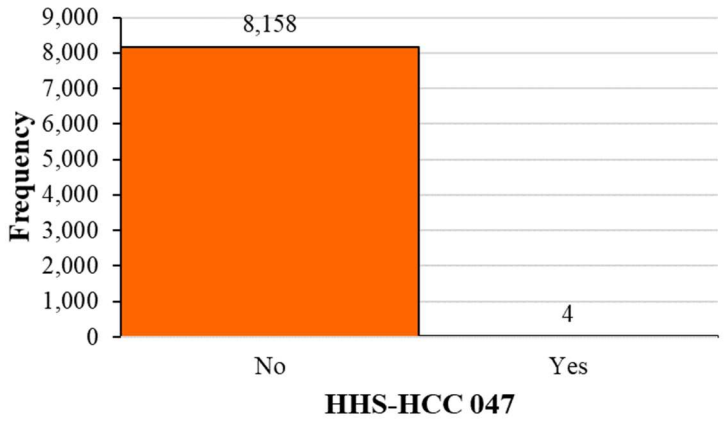


Figure 134. Frequency histogram by Acute Pancreatitis.

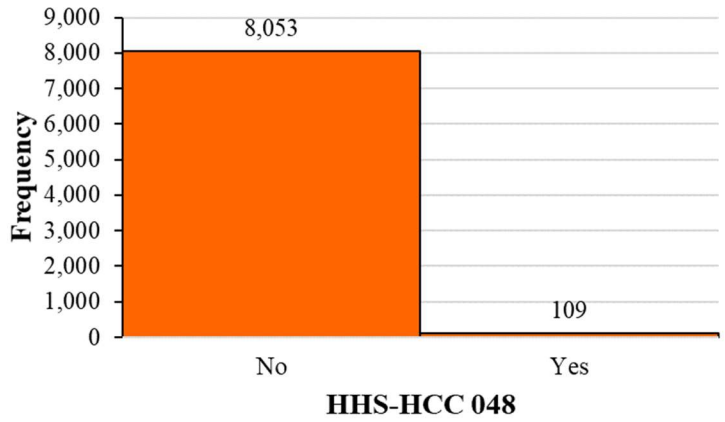


Figure 135. Frequency histogram by Inflammatory Bowel Disease.

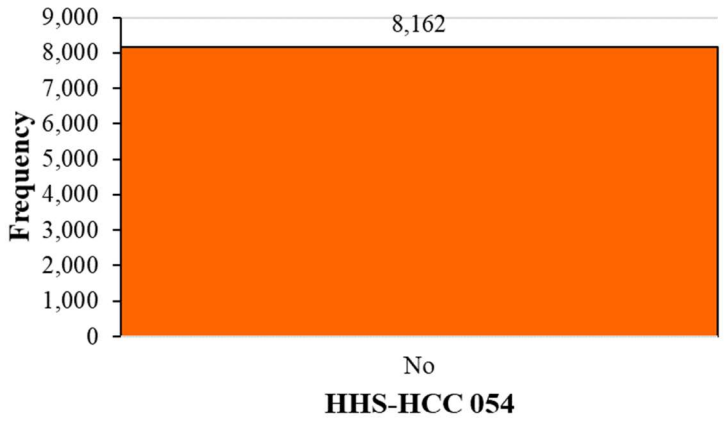


Figure 136. Frequency histogram by Necrotizing Fasciitis.

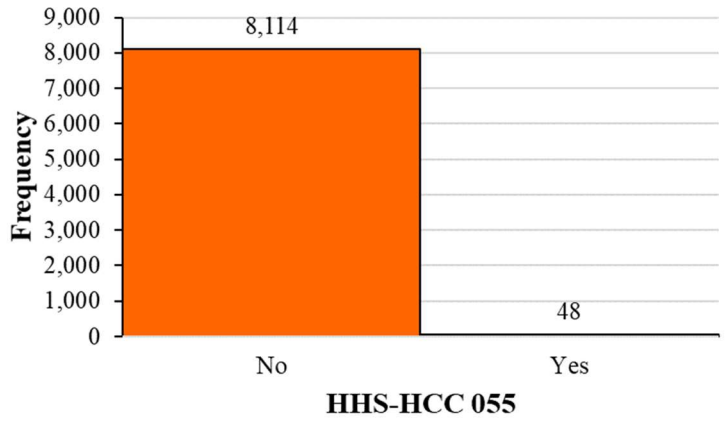


Figure 137. Frequency histogram by Bone/Joint/Muscle Infections/Necrosis.

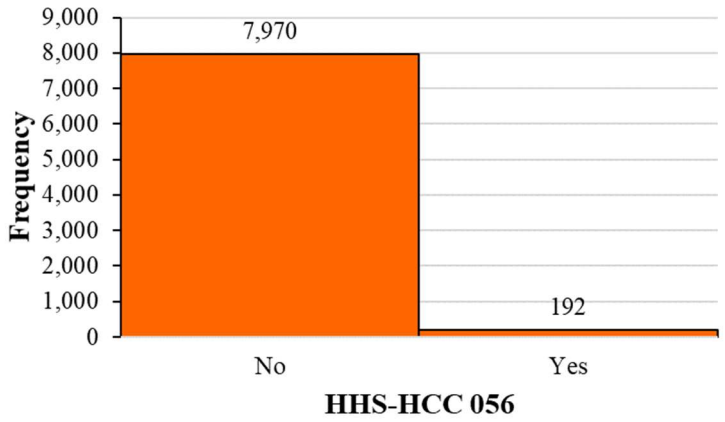


Figure 138. Frequency histogram by Rheumatoid Arthritis and Specified Autoimmune Disorders.

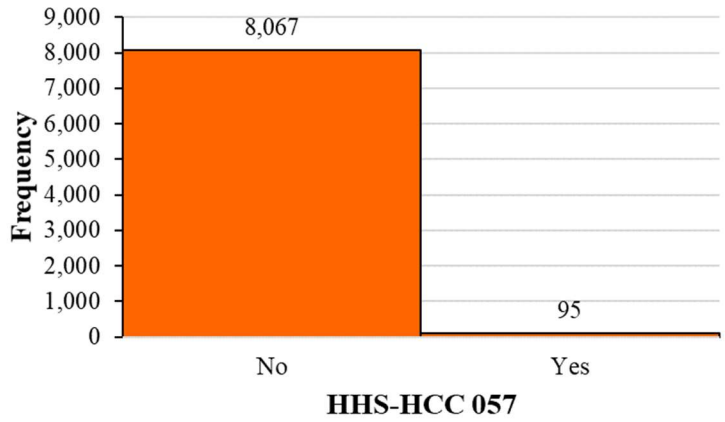


Figure 139. Frequency histogram by Systemic Lupus Erythematosus and Other Autoimmune Disorders.

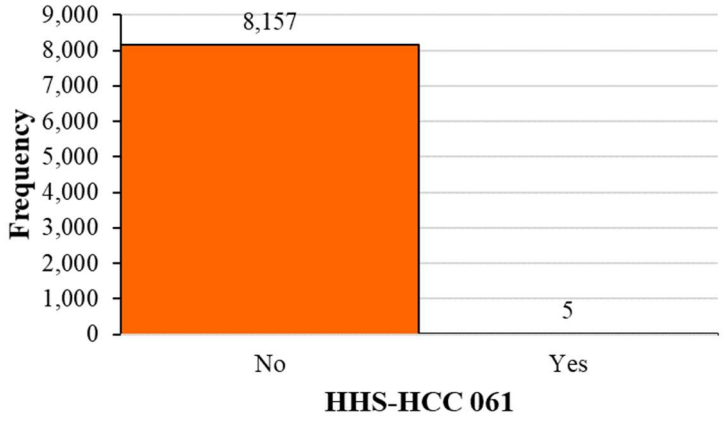


Figure 140. Frequency histogram by Osteogenesis Imperfecta and Other Osteodystrophies.

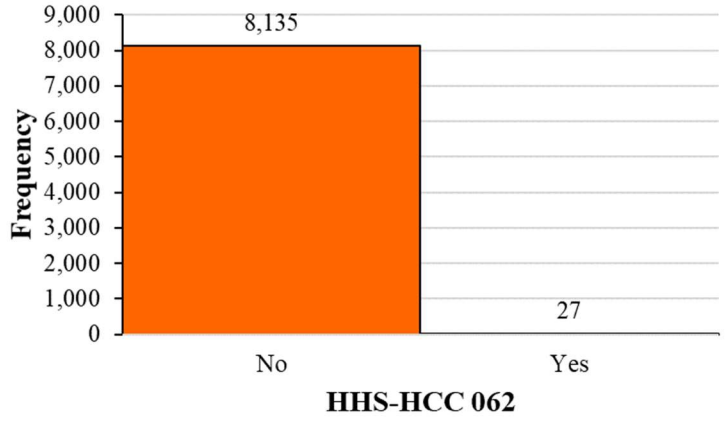


Figure 141. Frequency histogram by Congenital/Developmental Skeletal and Connective Tissue Disorders.

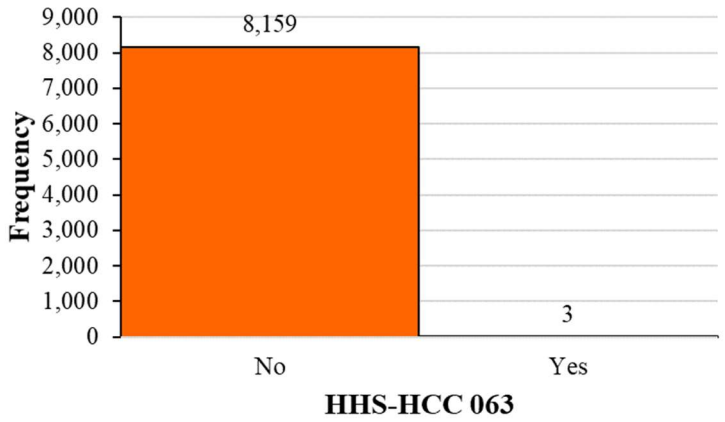


Figure 142. Frequency histogram by Cleft Lip/Cleft Palate.

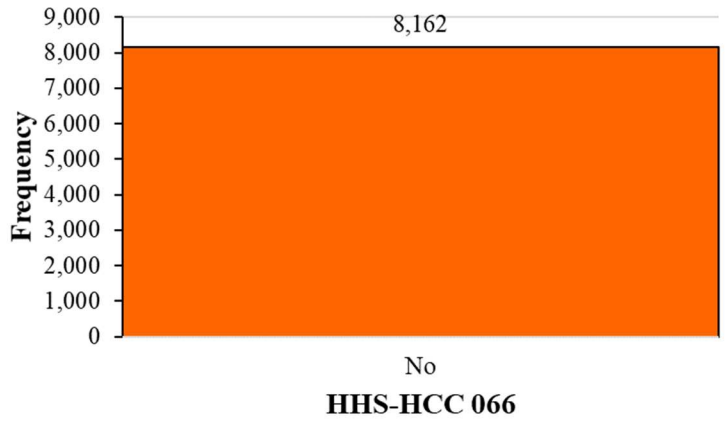


Figure 143. Frequency histogram by Hemophilia.

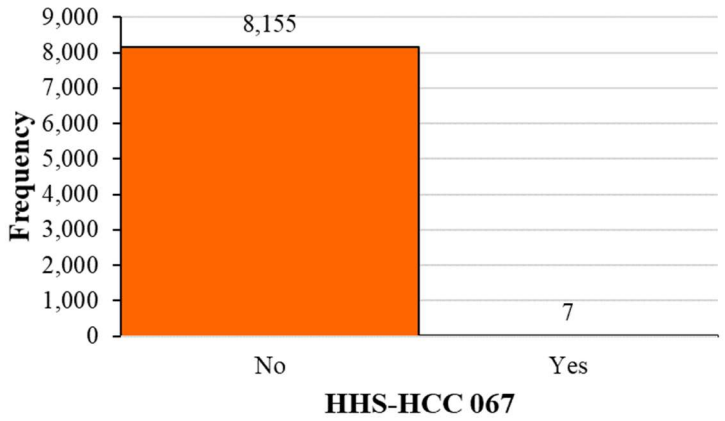


Figure 144. Frequency histogram by Myelodysplastic Syndromes and Myelofibrosis.

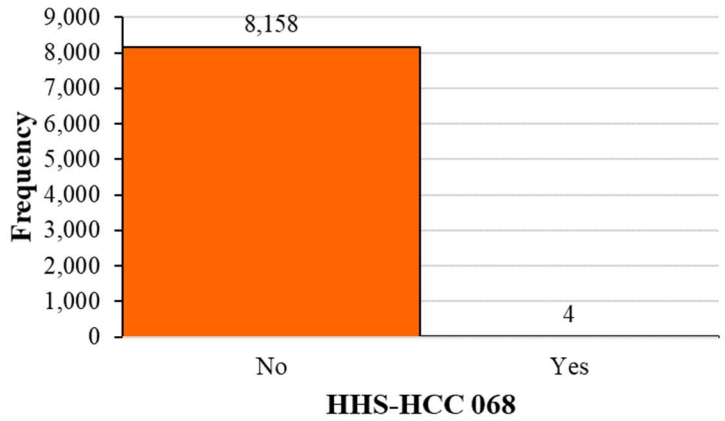


Figure 145. Frequency histogram by Aplastic Anemia.

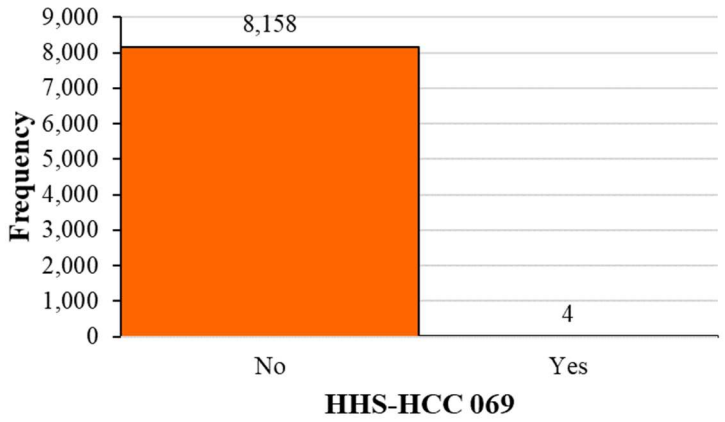


Figure 146. Frequency histogram by Acquired Hemolytic Anemia, Including Hemolytic Disease of Newborn.

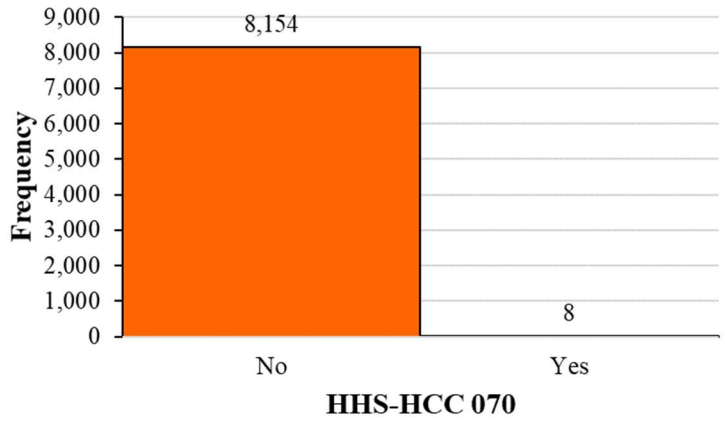


Figure 147. Frequency histogram by Sickle Cell Anemia (HbSS).

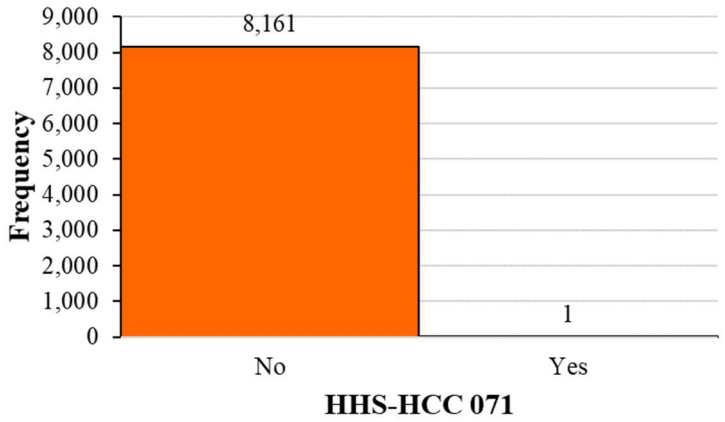


Figure 148. Frequency histogram by Beta Thalassemia Major.

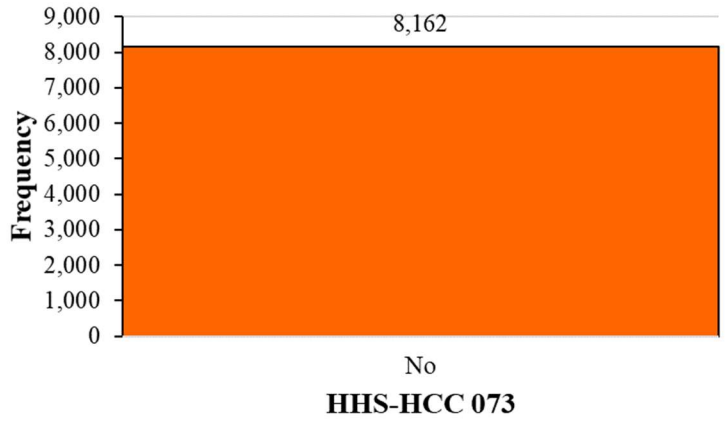


Figure 149. Frequency histogram by Combined and Other Severe Immunodeficiencies.

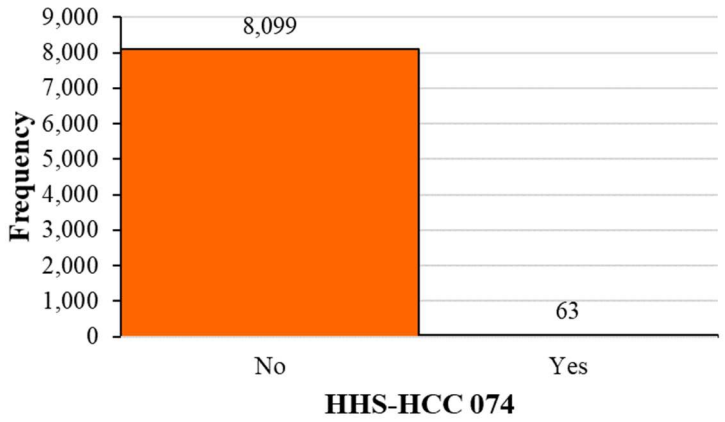


Figure 150. Frequency histogram by Disorders of the Immune Mechanism.

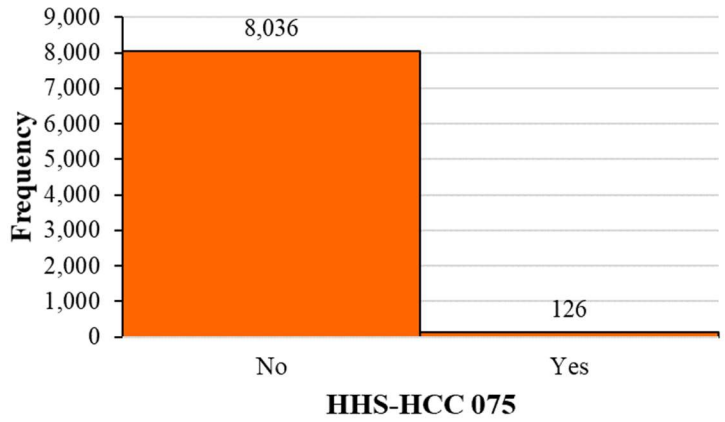


Figure 151. Frequency histogram by Coagulation Defects and Other Specified Hematological Disorders.

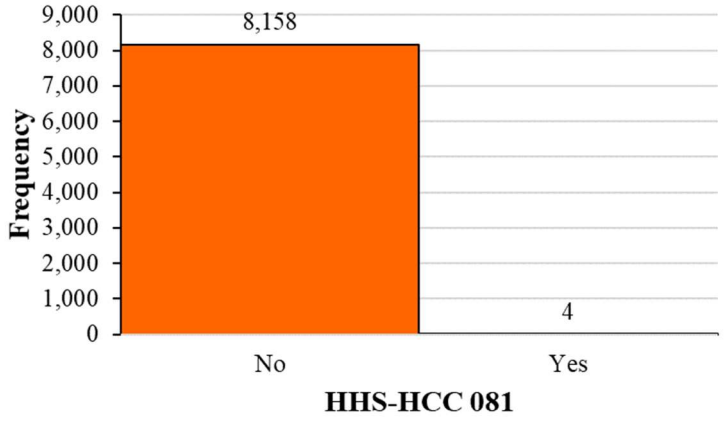


Figure 152. Frequency histogram by Drug Use with Psychotic Complications.

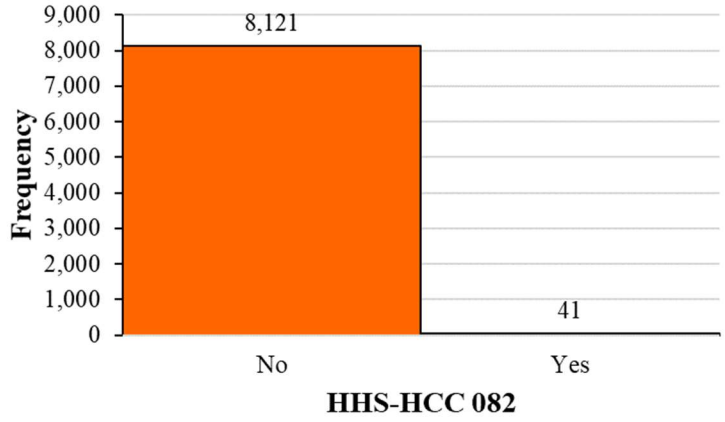


Figure 153. Frequency histogram by Drug Use Disorder, Moderate/Severe, or Drug Use with Non-Psychotic Complications.

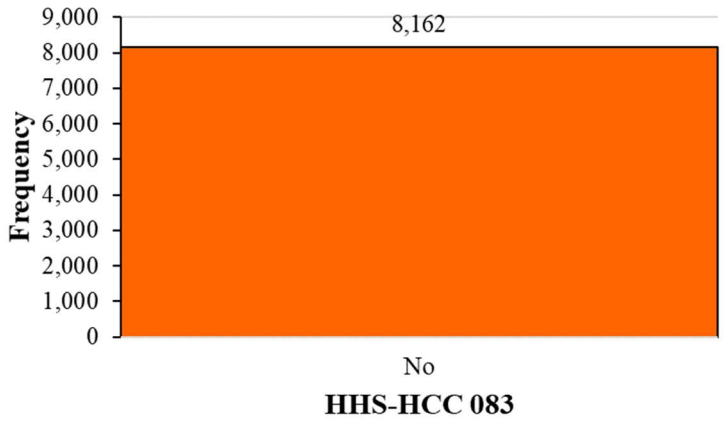


Figure 154. Frequency histogram by Alcohol Use with Psychotic Complications.

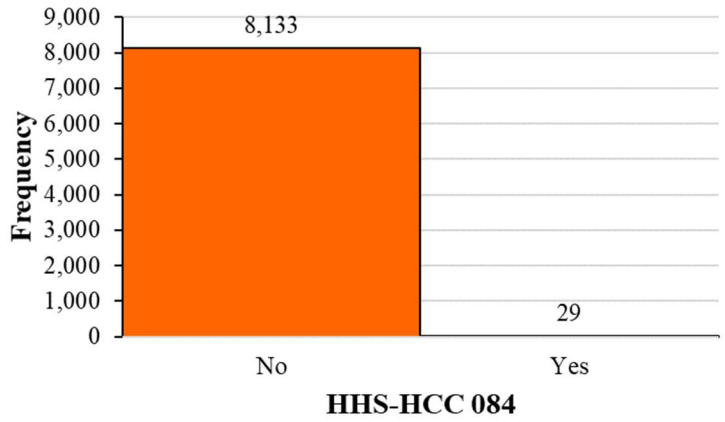


Figure 155. Frequency histogram by Alcohol Use Disorder, Moderate/Severe, or Alcohol Use with Specified Non-Psychotic Complications.

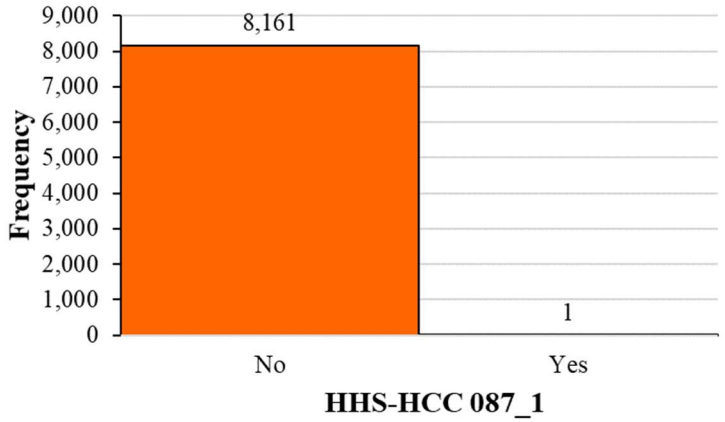


Figure 156. Frequency histogram by Schizophrenia.

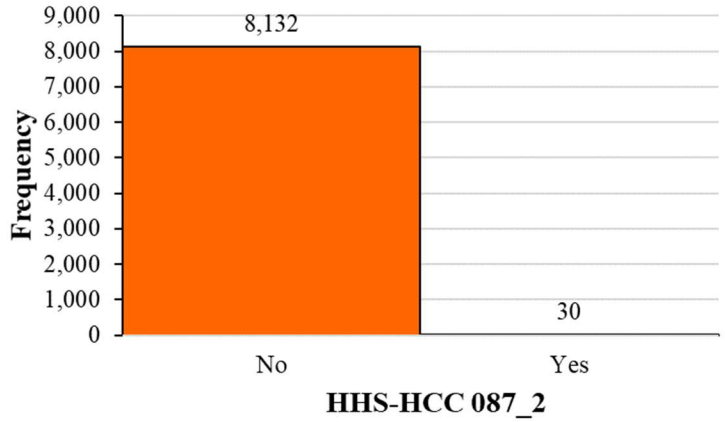


Figure 157. Frequency histogram by Delusional and Other Specified Psychotic Disorders, Unspecified Psychosis.

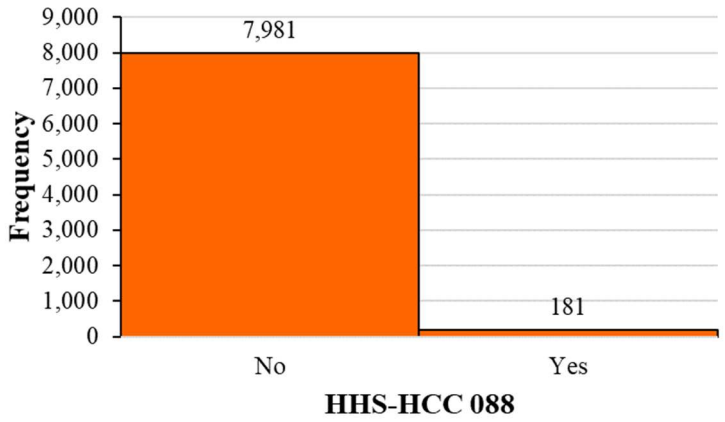


Figure 158. Frequency histogram by Major Depressive Disorder, Severe, and Bipolar Disorders.

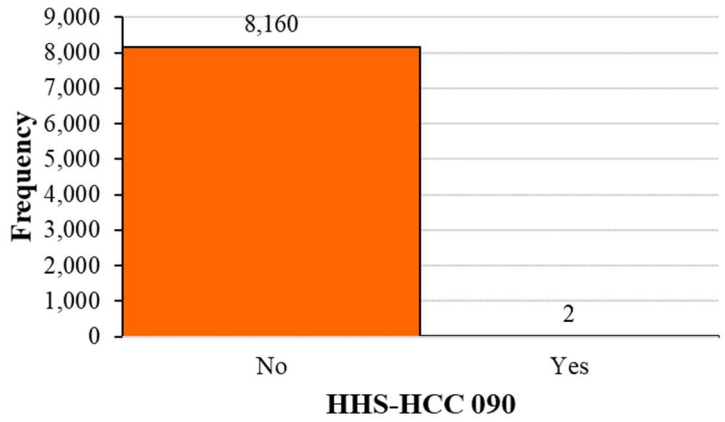


Figure 159. Frequency histogram by Personality Disorders.

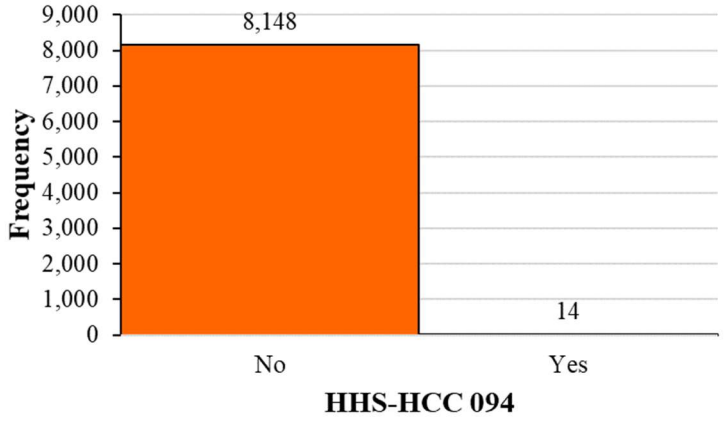


Figure 160. Frequency histogram by Anorexia/Bulimia Nervosa.

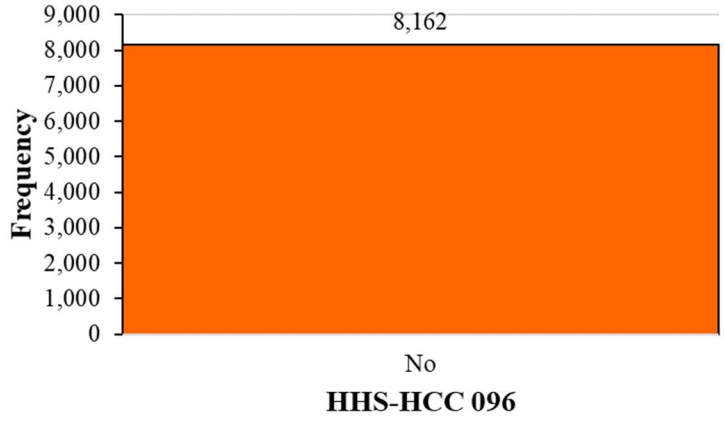


Figure 161. Frequency histogram by Prader-Willi, Patau, Edwards, and Autosomal Deletion Syndromes.

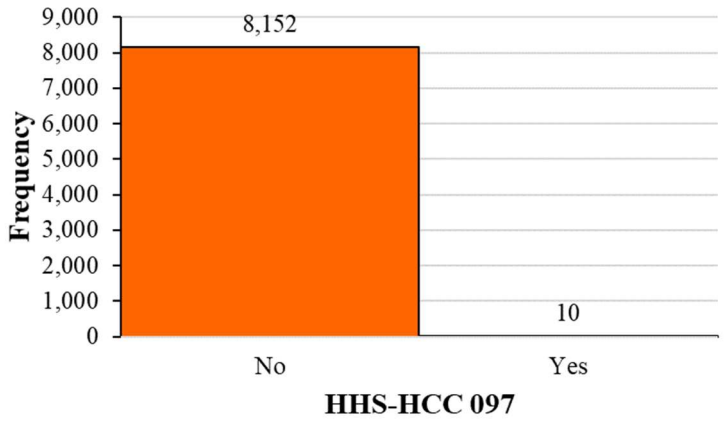


Figure 162. Frequency histogram by Down Syndrome, Fragile X, Other Chromosomal Anomalies, and Congenital Malformation Syndromes.

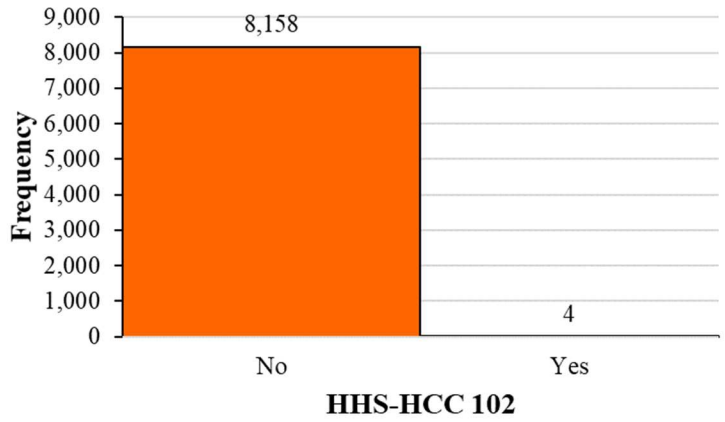


Figure 163. Frequency histogram by Autistic Disorder.

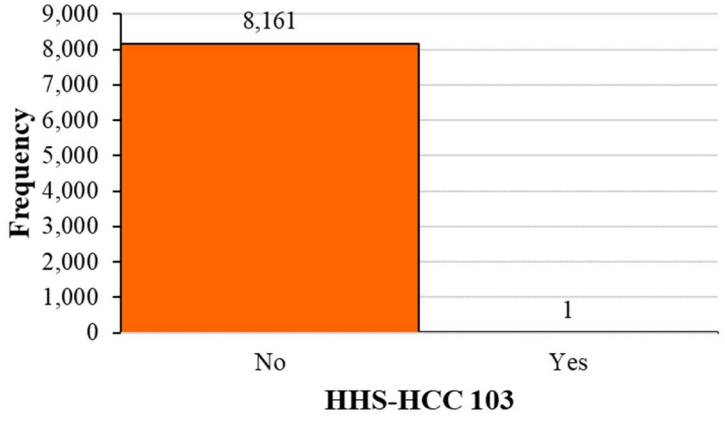


Figure 164. Frequency histogram by Pervasive Developmental Disorders, Except Autistic Disorder.

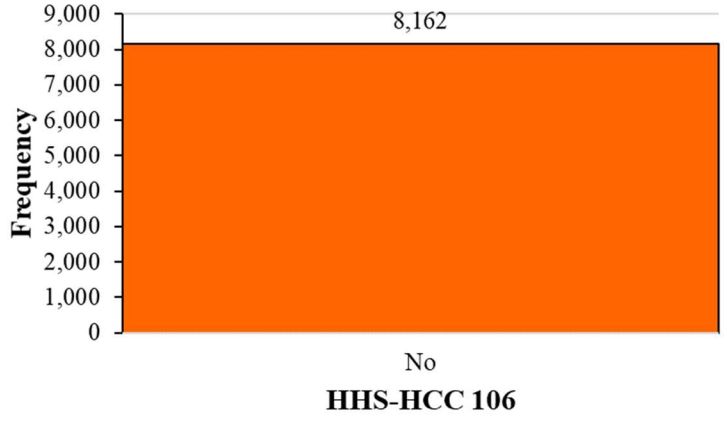


Figure 165. Frequency histogram by Traumatic Complete Lesion Cervical Spinal Cord.

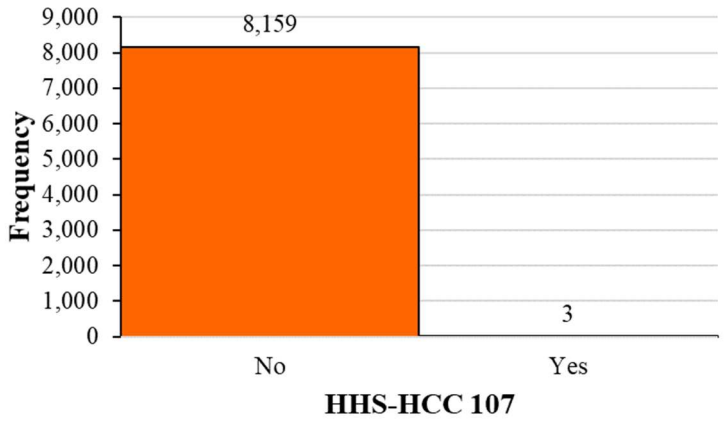


Figure 166. Frequency histogram by Quadriplegia.

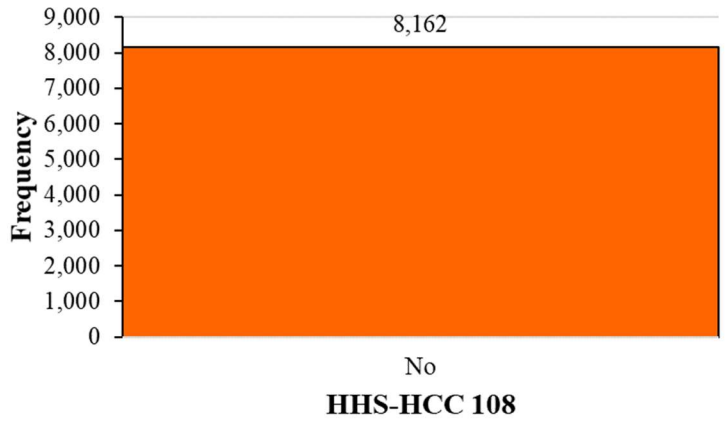


Figure 167. Frequency histogram by Traumatic Complete Lesion Dorsal Spinal Cord.

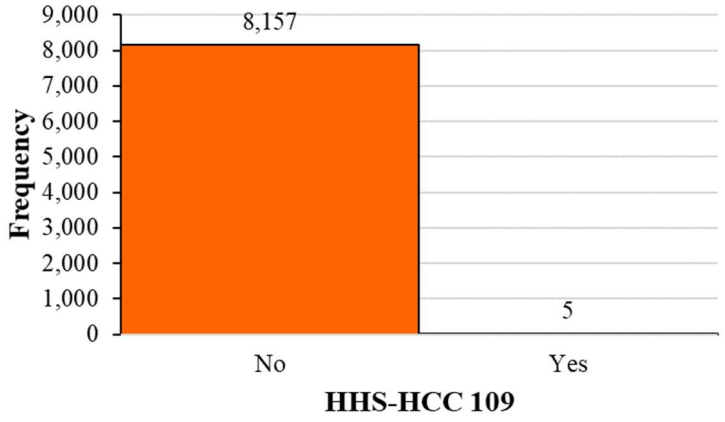


Figure 168. Frequency histogram by Paraplegia.

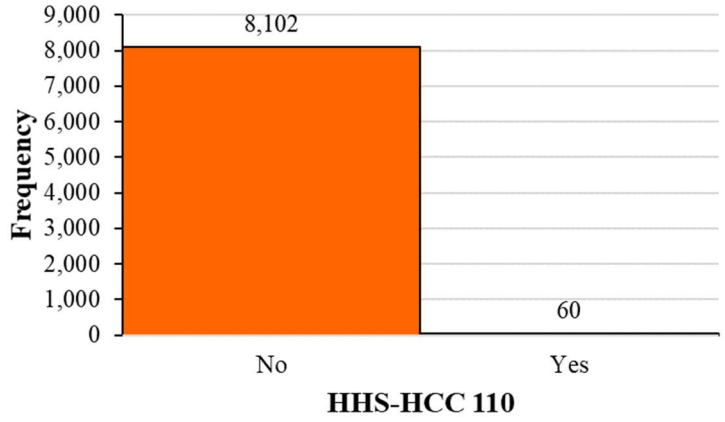


Figure 169. Frequency histogram by Spinal Cord Disorders/Injuries.

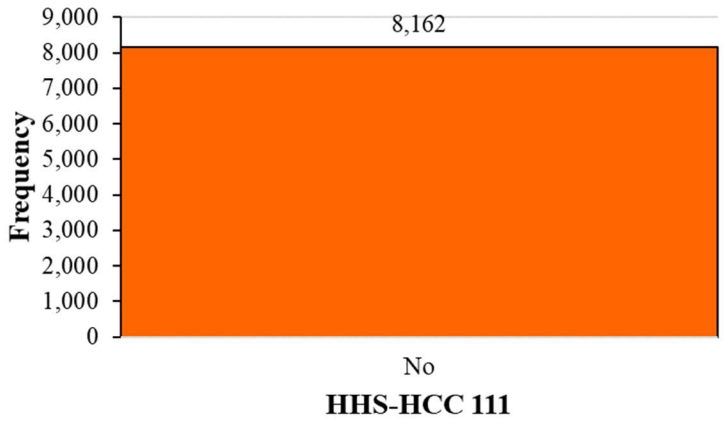


Figure 170. Frequency histogram by Amyotrophic Lateral Sclerosis and Other Anterior Horn Cell Disease.

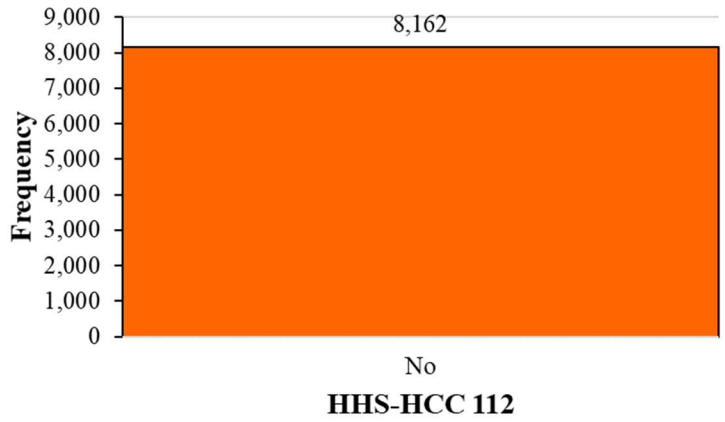


Figure 171. Frequency histogram by Quadriplegic Cerebral Palsy.

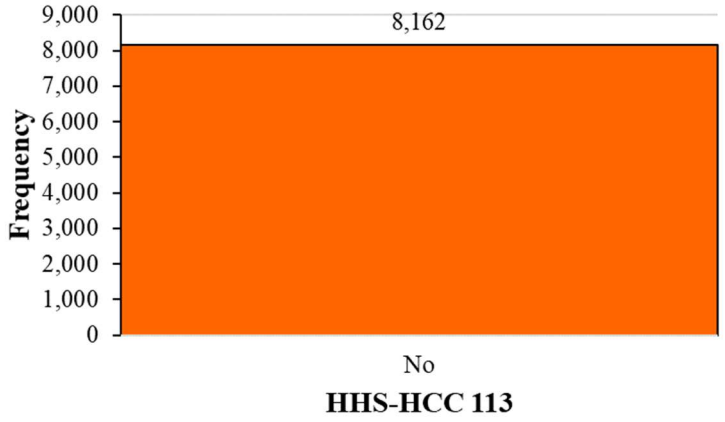


Figure 172. Frequency histogram by Cerebral Palsy, Except Quadriplegic.

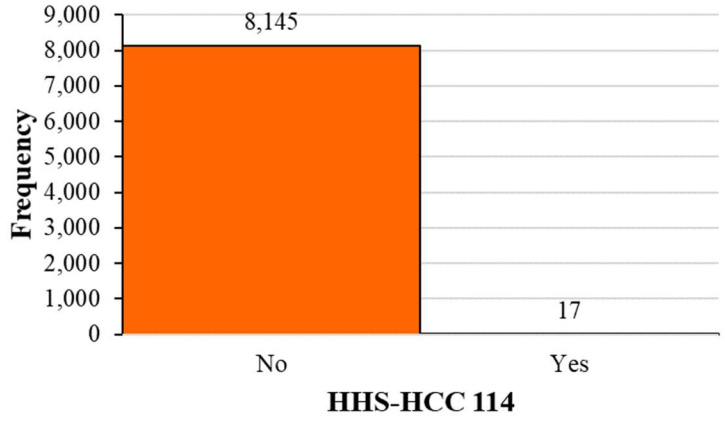


Figure 173. Frequency histogram by Spina Bifida and Other Brain/Spinal/Nervous System Congenital Anomalies.

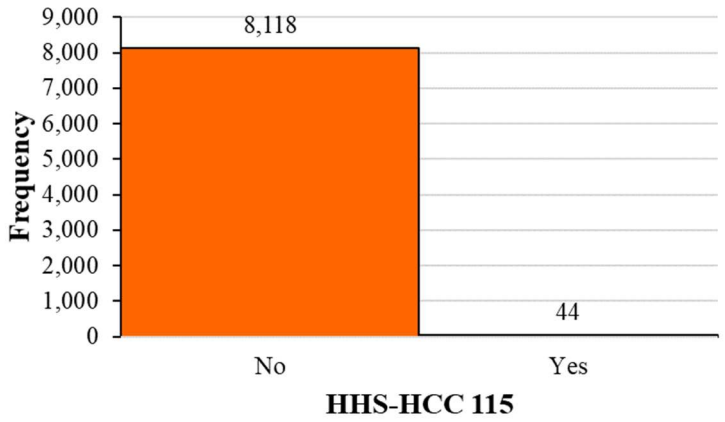


Figure 174. Frequency histogram by Myasthenia Gravis/Myoneural Disorders and Guillain-Barre Syndrome/Inflammatory and Toxic Neuropathy.

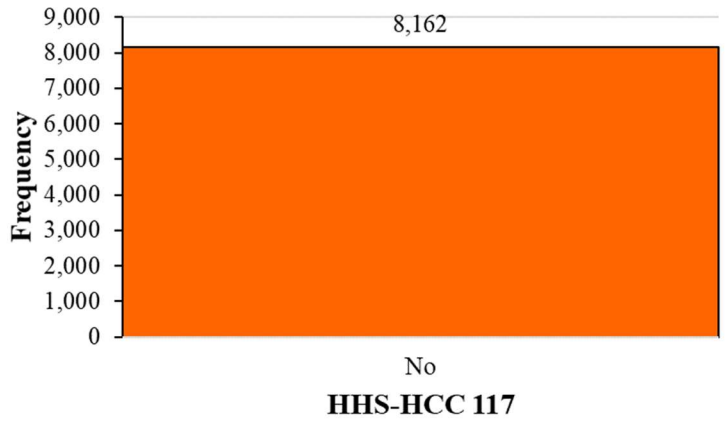


Figure 175. Frequency histogram by Muscular Dystrophy.

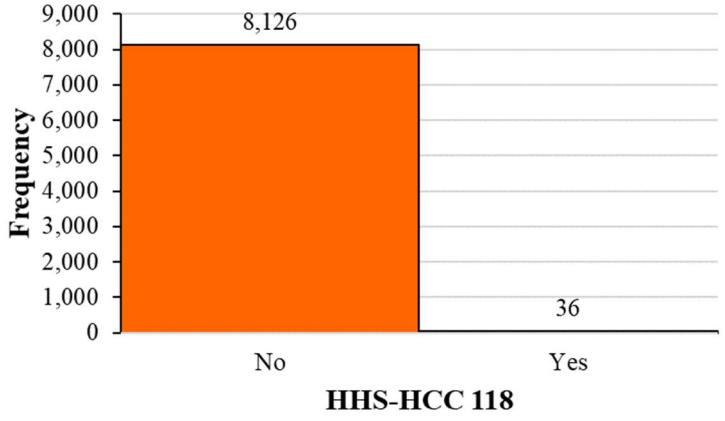


Figure 176. Frequency histogram by Multiple Sclerosis.

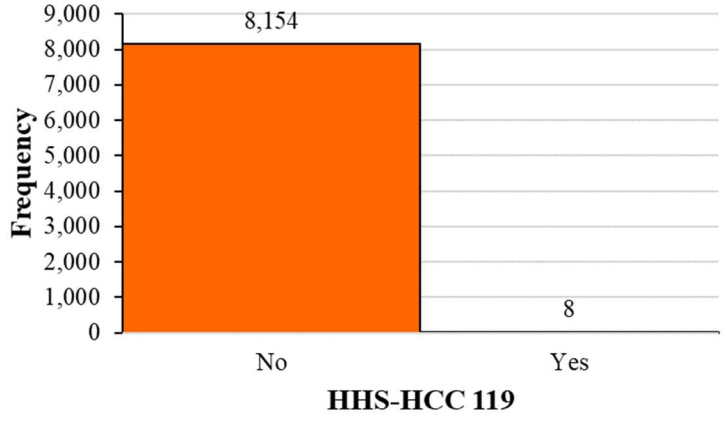


Figure 177. Frequency histogram by Parkinson's, Huntington's, and Spinocerebellar Disease, and Other Neurodegenerative Disorders.

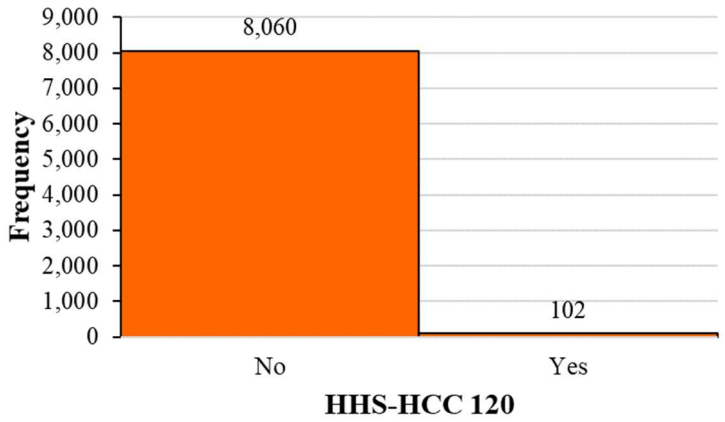


Figure 178. Frequency histogram by Seizure Disorders and Convulsions.

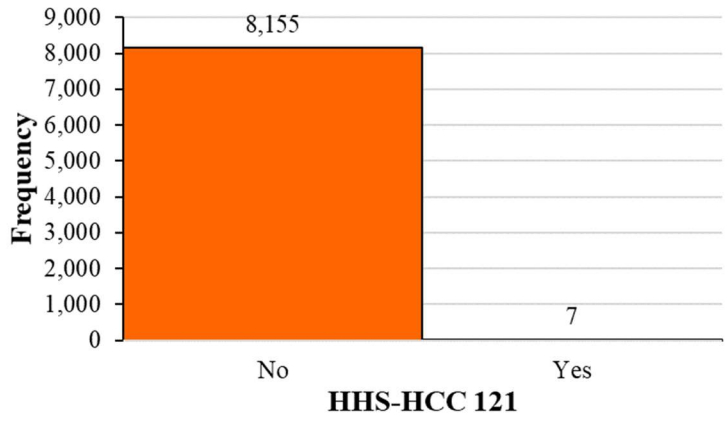


Figure 179. Frequency histogram by Hydrocephalus.

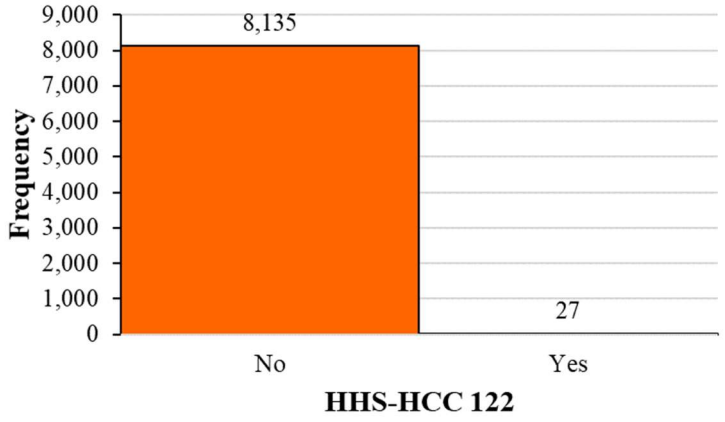


Figure 180. Frequency histogram by Coma, Brain Compression/Anoxic Damage.

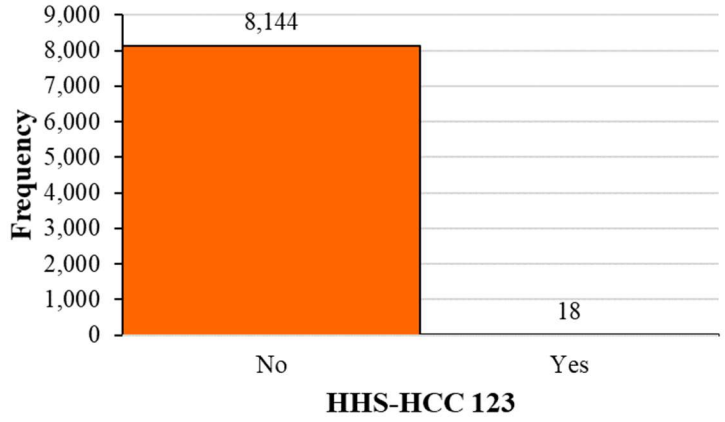


Figure 181. Frequency histogram by Narcolepsy and Cataplexy.

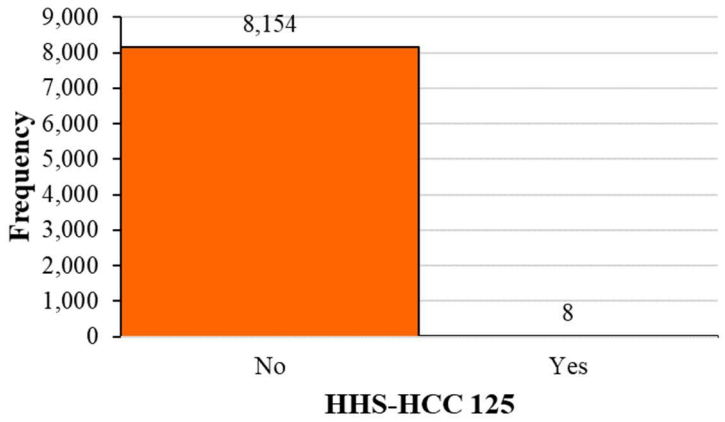


Figure 182. Frequency histogram by Respirator Dependence/Tracheostomy Status.

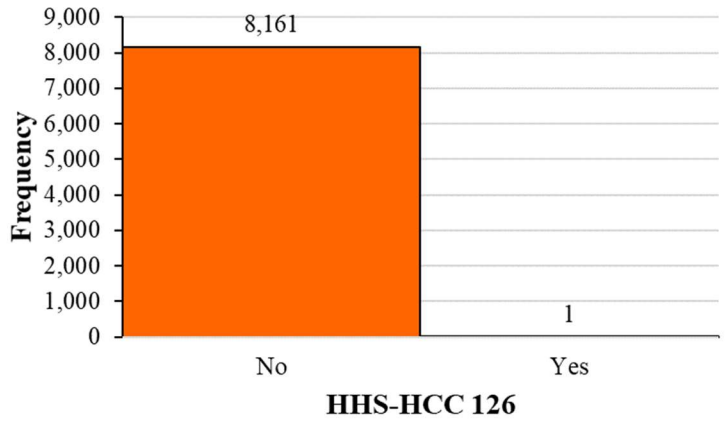


Figure 183. Frequency histogram by Respiratory Arrest.

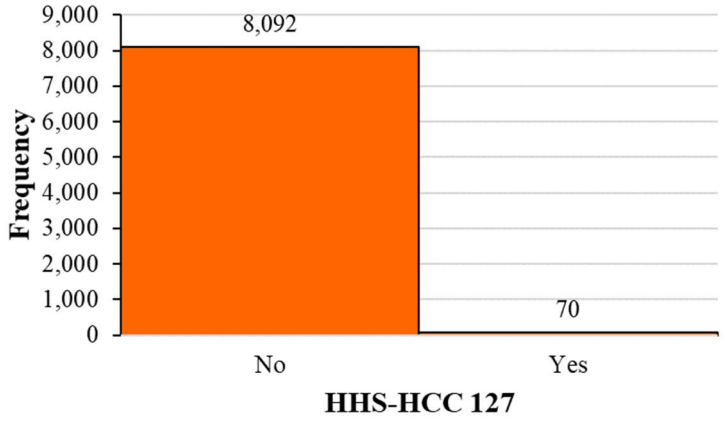


Figure 184. Frequency histogram by Cardio-Respiratory Failure and Shock, Including Respiratory Distress Syndromes.

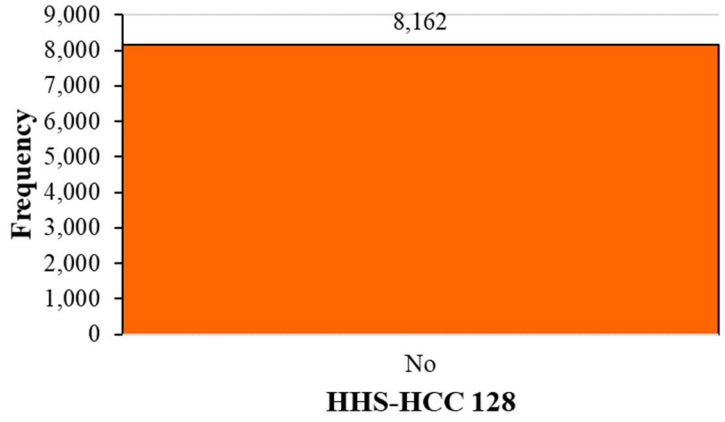


Figure 185. Frequency histogram by Heart Assistive Device/Artificial Heart.

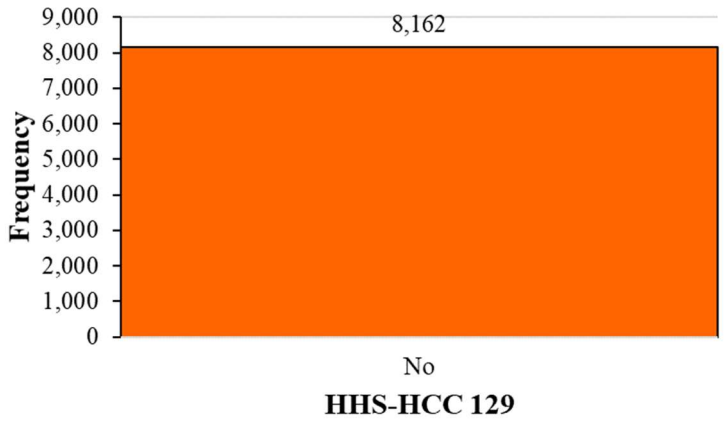


Figure 186. Frequency histogram by Heart Transplant Status/Complications.

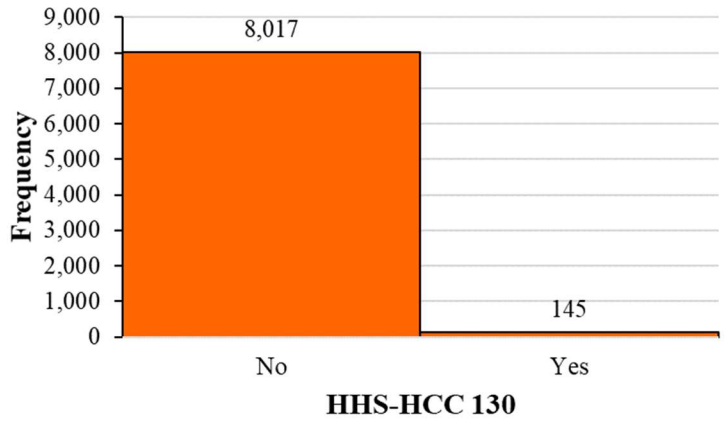


Figure 187. Frequency histogram by Heart Failure.

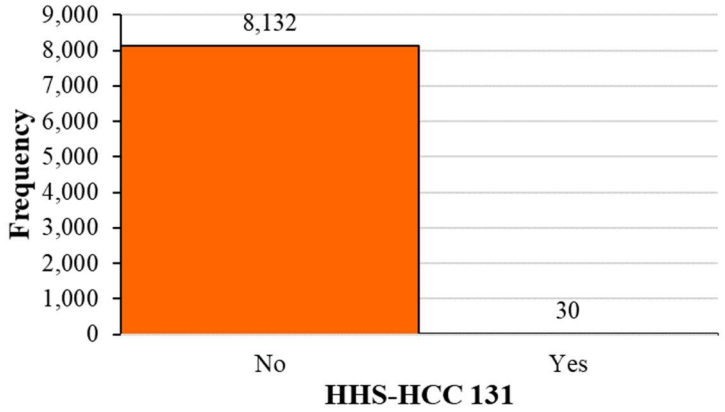


Figure 188. Frequency histogram by Acute Myocardial Infarction.

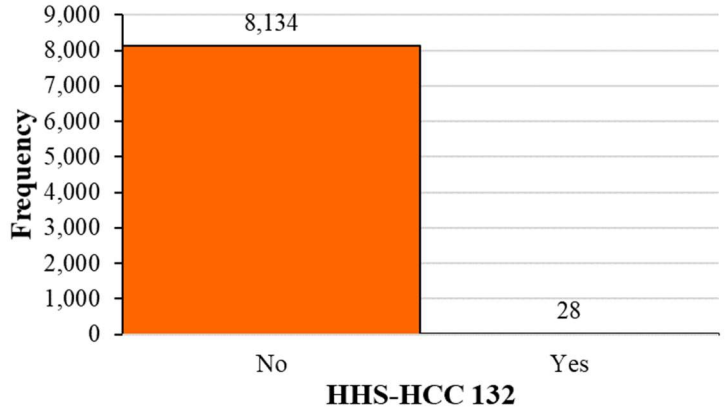


Figure 189. Frequency histogram by Unstable Angina and Other Acute Ischemic Heart Disease.

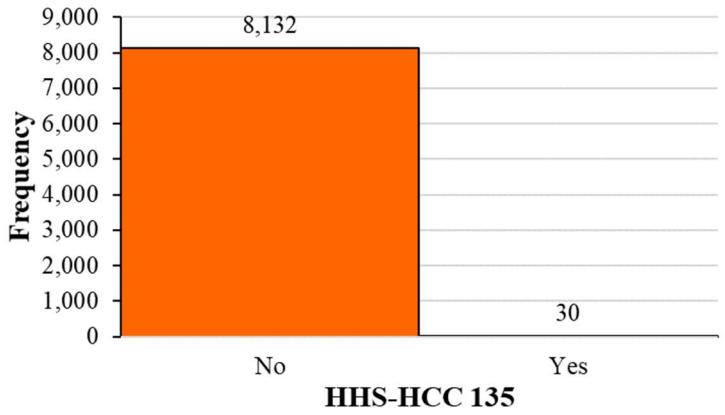


Figure 190. Frequency histogram by Heart Infection/Inflammation, Except Rheumatic.

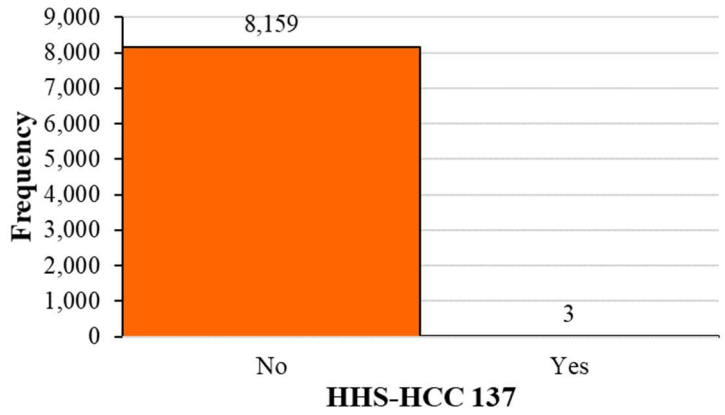


Figure 191. Frequency histogram by Hypoplastic Left Heart Syndrome and Other Severe Congenital Heart Disorders.

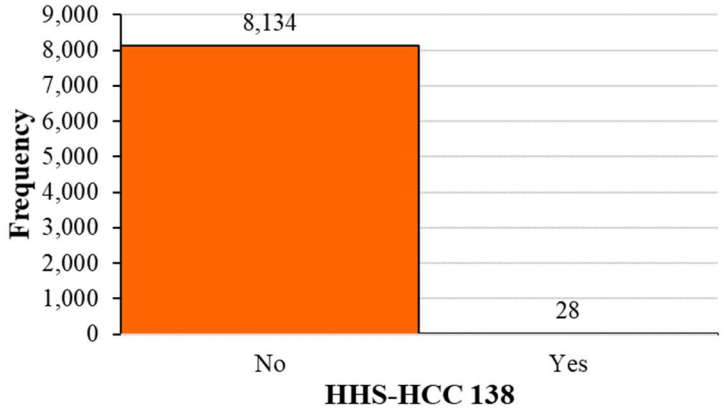


Figure 192. Frequency histogram by Major Congenital Heart/Circulatory Disorders.

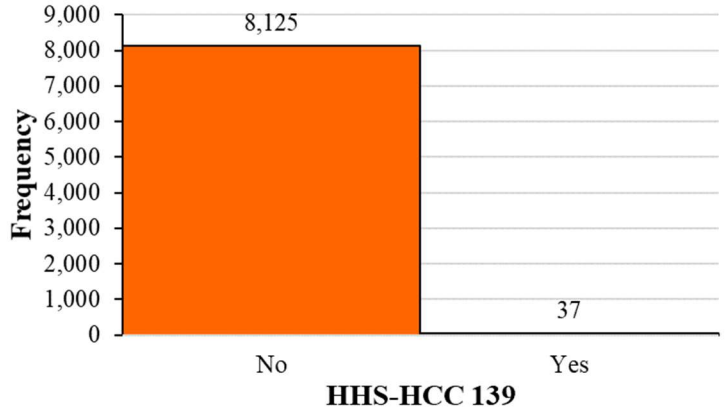


Figure 193. Frequency histogram by Atrial and Ventricular Septal Defects, Patent Ductus Arteriosus, and Other Congenital Heart/Circulatory Disorders.

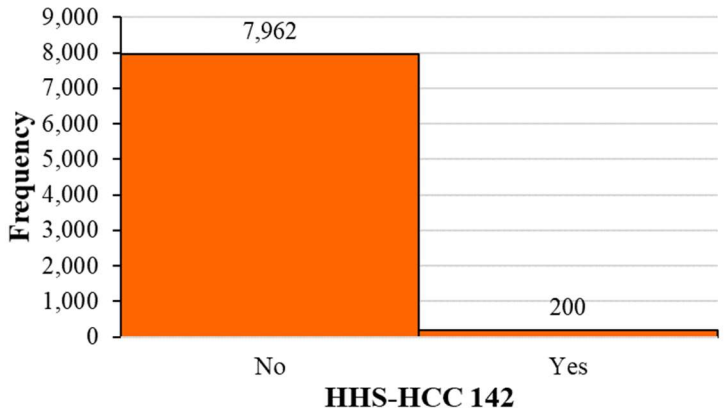


Figure 194. Frequency histogram by Specified Heart Arrhythmias.

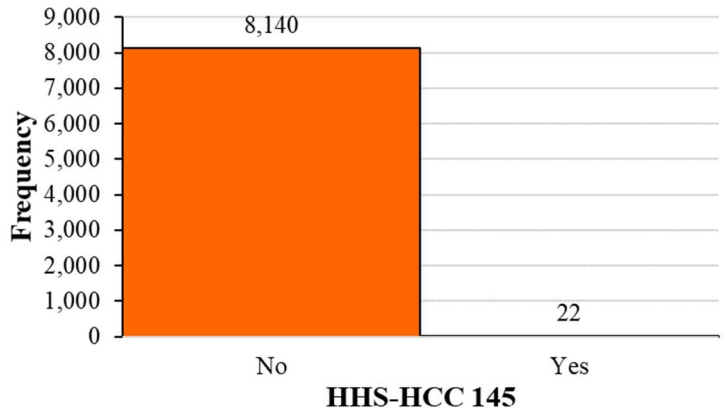


Figure 195. Frequency histogram by Intracranial Hemorrhage.

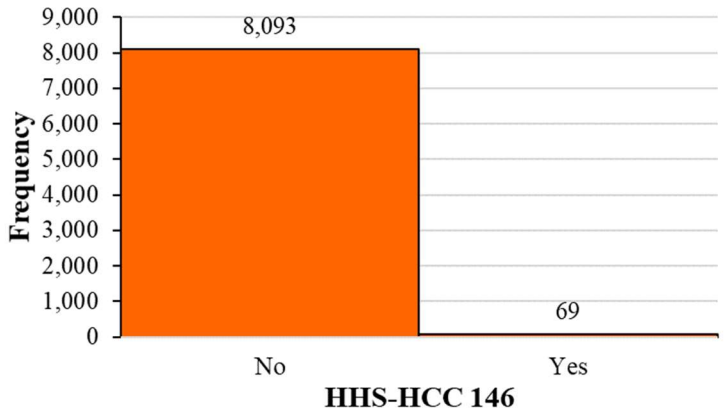


Figure 196. Frequency histogram by Ischemic or Unspecified Stroke.

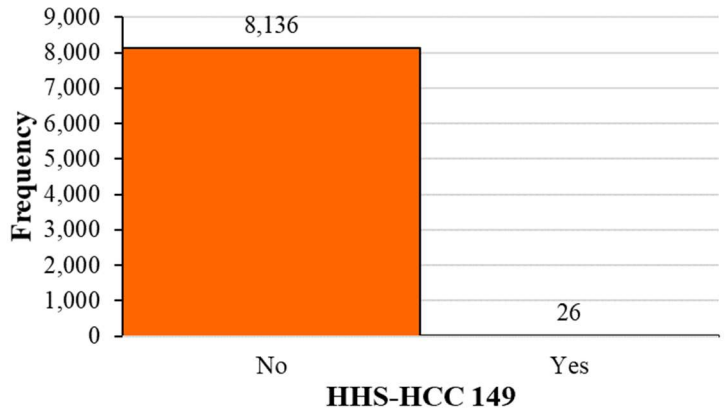


Figure 197. Frequency histogram by Cerebral Aneurysm and Arteriovenous Malformation.

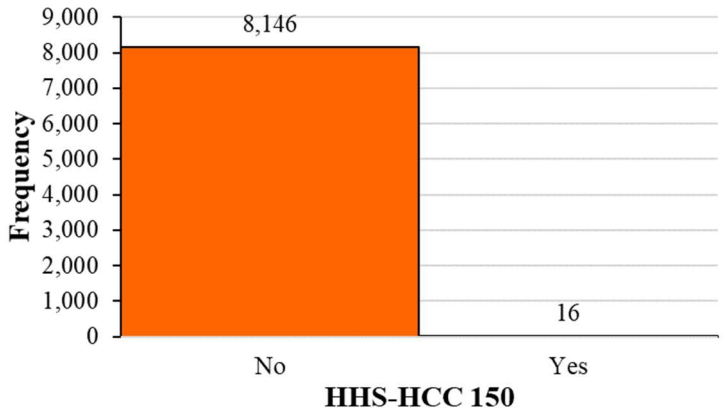


Figure 198. Frequency histogram by Hemiplegia/Hemiparesis.

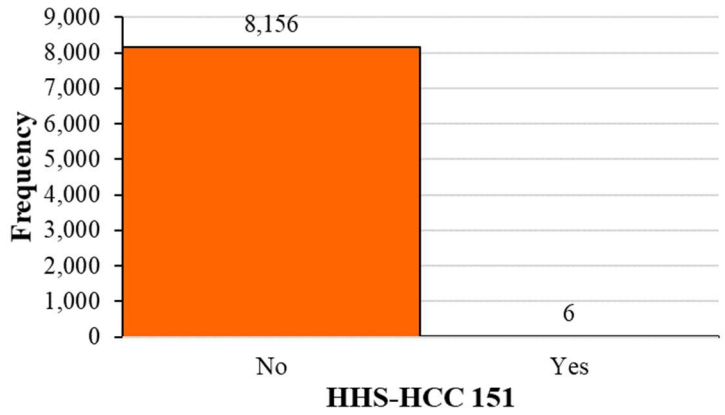


Figure 199. Frequency histogram by Monoplegia, Other Paralytic Syndromes.

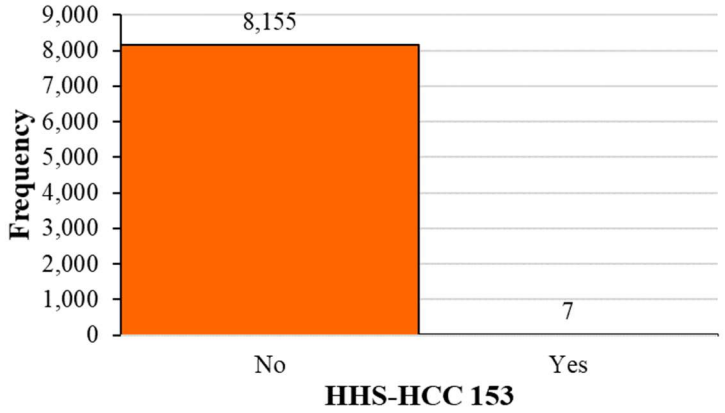


Figure 200. Frequency histogram by Atherosclerosis of the Extremities with Ulceration or Gangrene.

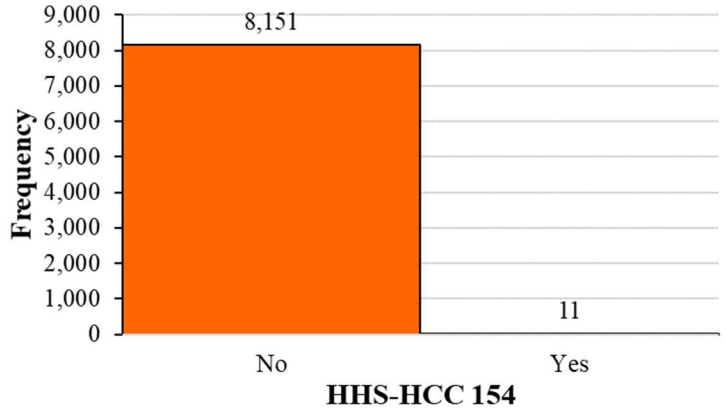


Figure 201. Frequency histogram by Vascular Disease with Complications.

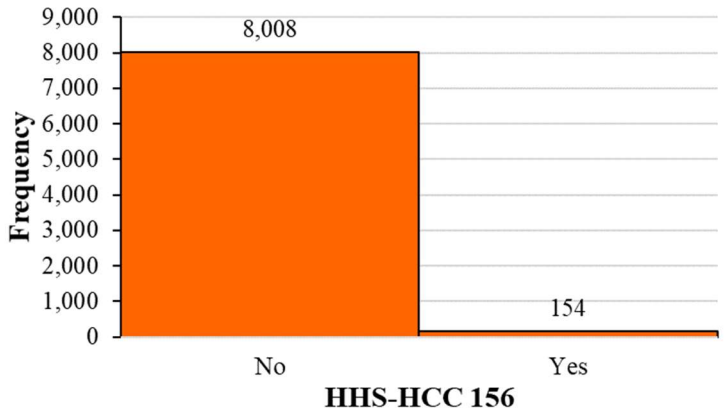


Figure 202. Frequency histogram by Pulmonary Embolism and Deep Vein Thrombosis.

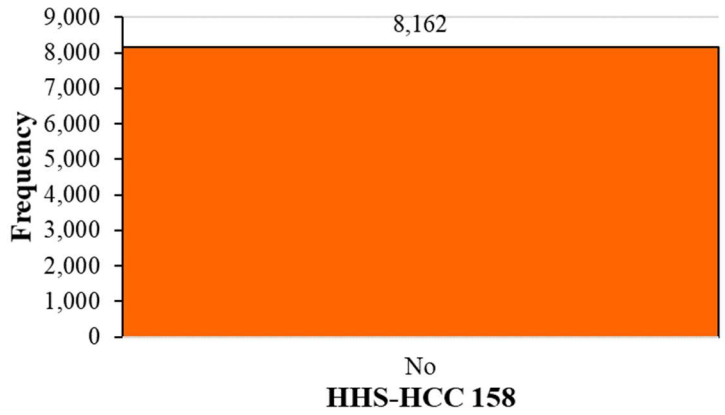


Figure 203. Frequency histogram by Lung Transplant Status/Complications.

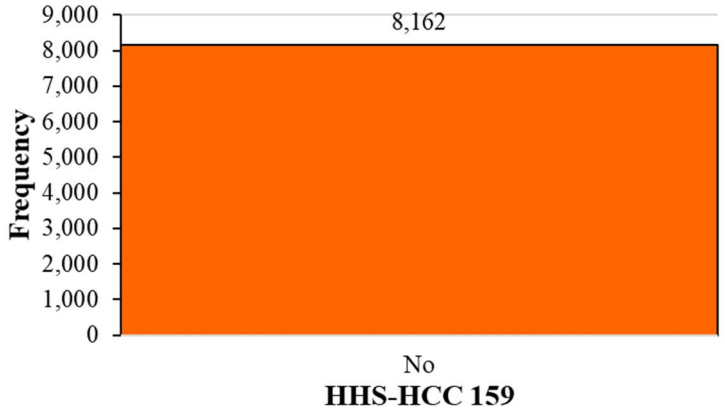


Figure 204. Frequency histogram by Cystic Fibrosis.

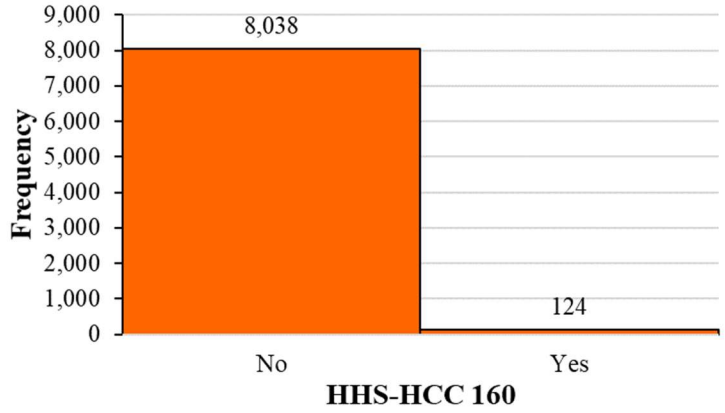


Figure 205. Frequency histogram by Chronic Obstructive Pulmonary Disease, Including Bronchiectasis.

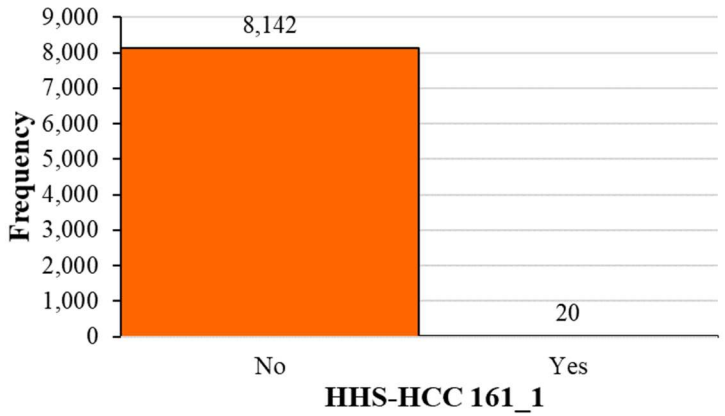


Figure 206. Frequency histogram by Severe Asthma.

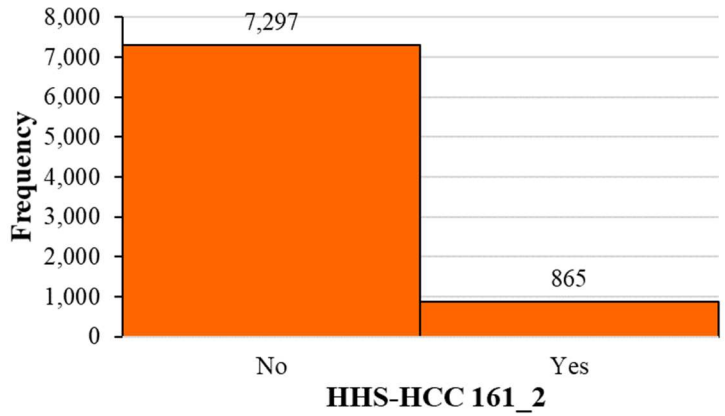


Figure 207. Frequency histogram by Asthma, Except Severe.

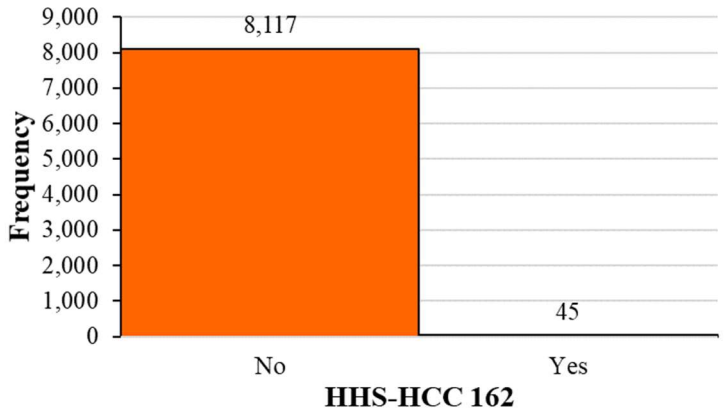


Figure 208. Frequency histogram by Fibrosis of Lung and Other Lung Disorders.

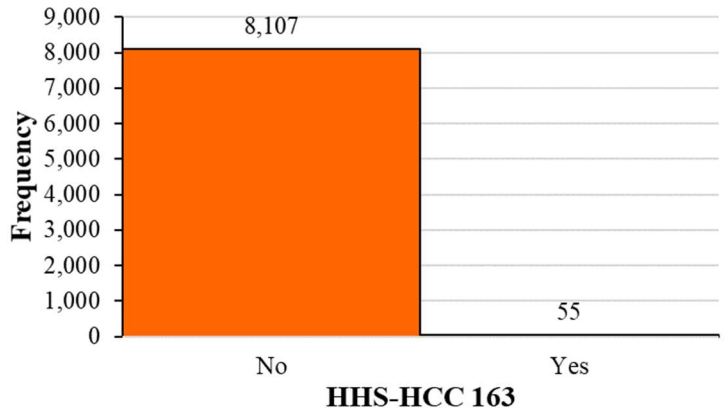


Figure 209. Frequency histogram by Aspiration and Specified Bacterial Pneumonias and Other Severe Lung Infections.

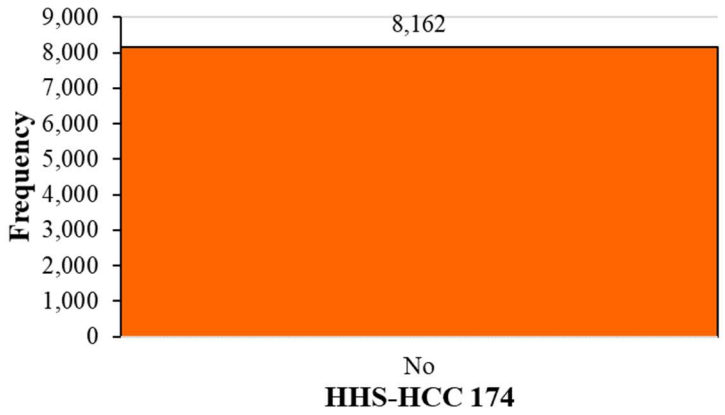


Figure 210. Frequency histogram by Exudative Macular Degeneration.

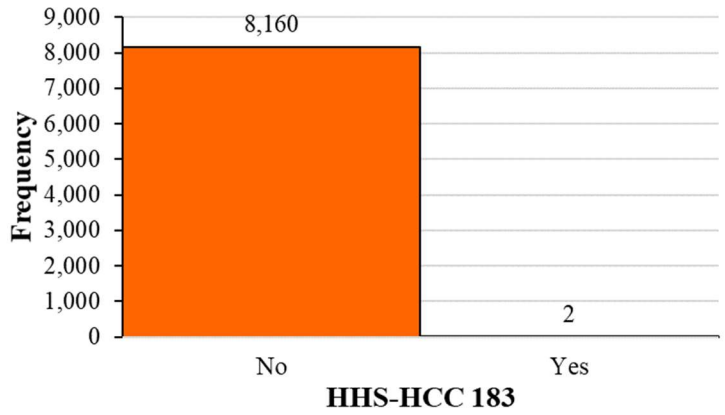


Figure 211. Frequency histogram by Kidney Transplant Status/Complications.

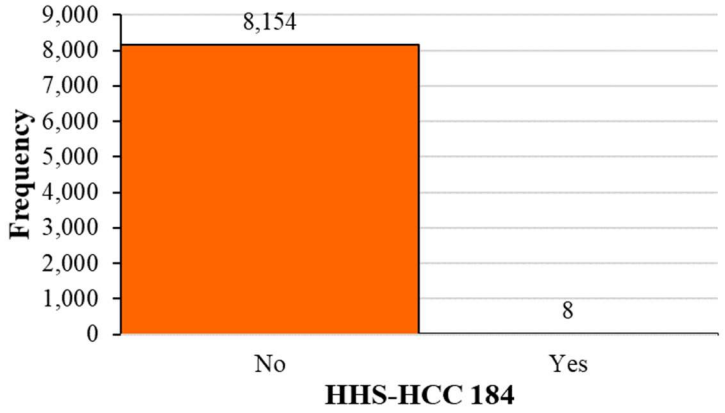


Figure 212. Frequency histogram by End Stage Renal Disease.

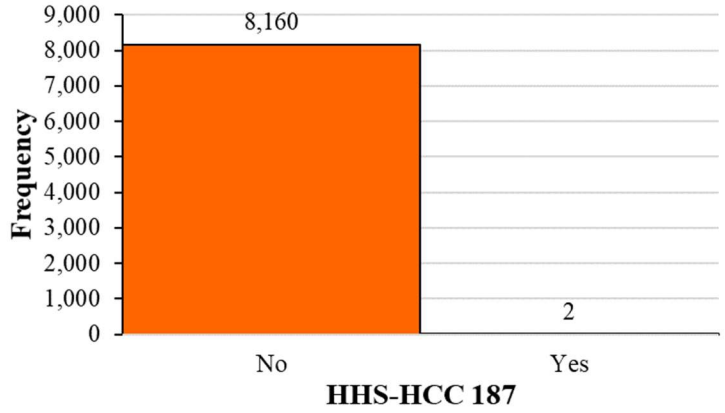


Figure 213. Frequency histogram by Chronic Kidney Disease, Stage 5.

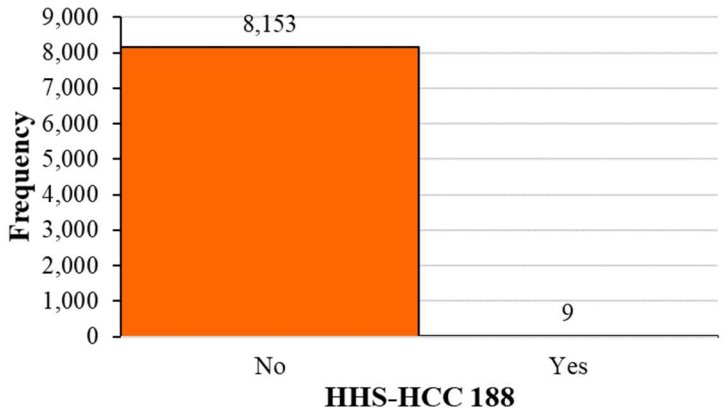


Figure 214. Frequency histogram by Chronic Kidney Disease, Severe (Stage 4).

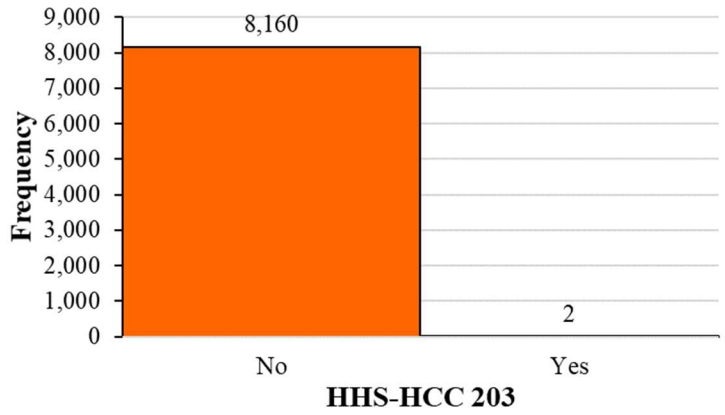


Figure 215. Frequency histogram by Ectopic and Molar Pregnancy.

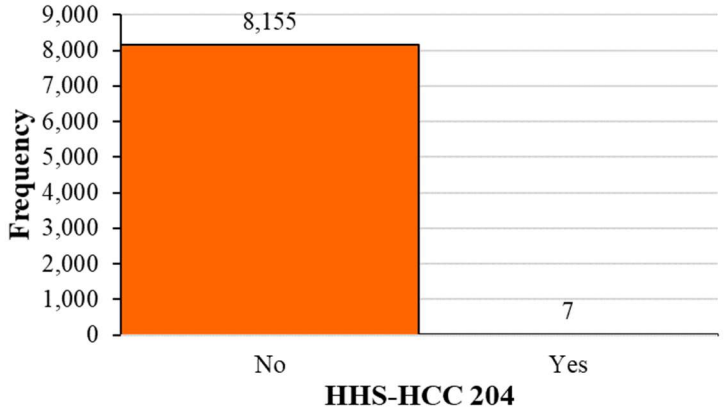


Figure 216. Frequency histogram by Miscarriage with Complications.

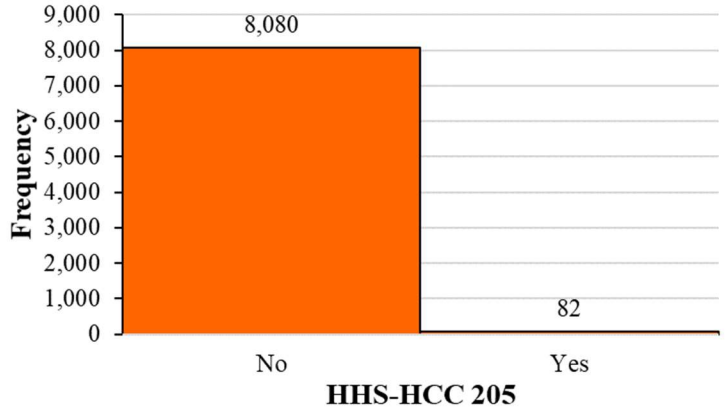


Figure 217. Frequency histogram by Miscarriage with No or Minor Complications.

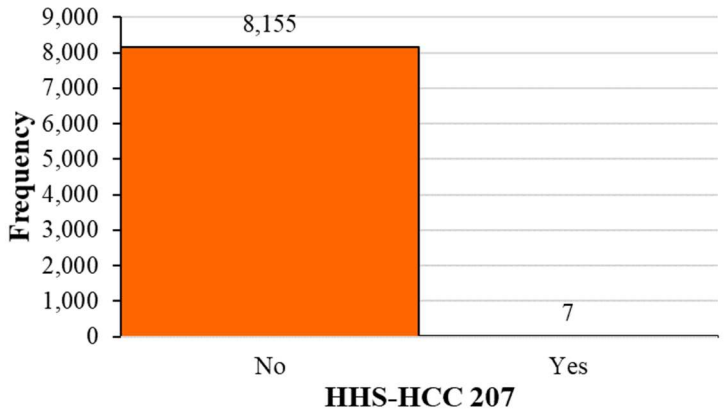


Figure 218. Frequency histogram by Pregnancy with Delivery with Major Complications.

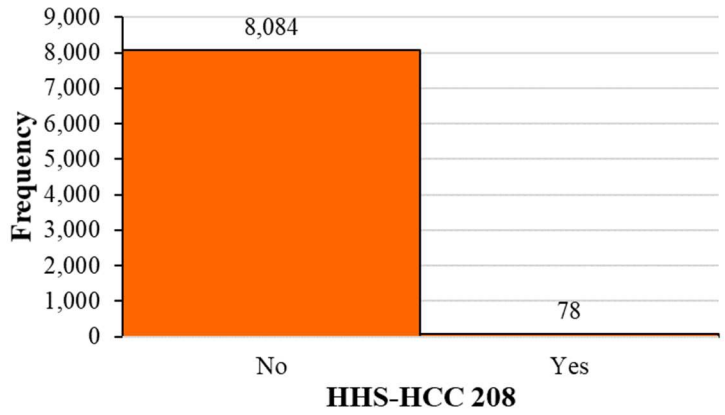


Figure 219. Frequency histogram by Pregnancy with Delivery with Complications.

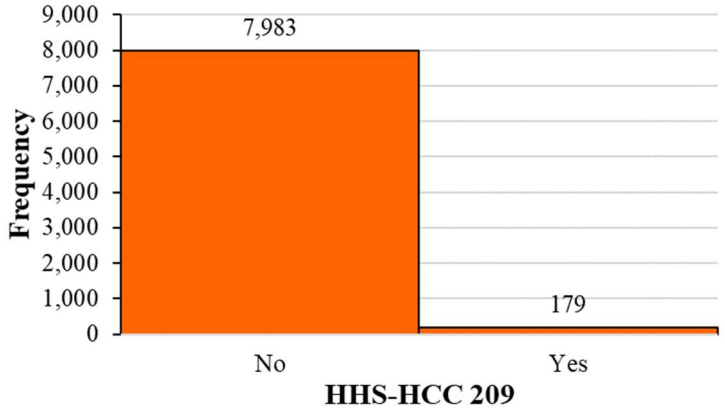


Figure 220. Frequency histogram by Pregnancy with Delivery with No or Minor Complications.

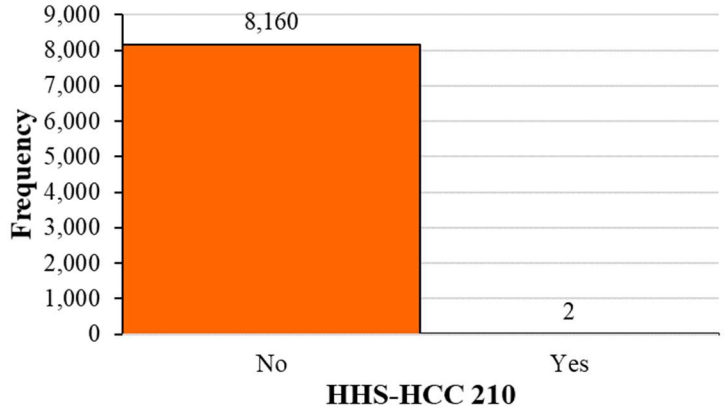


Figure 221. Frequency histogram by (Ongoing) Pregnancy without Delivery with Major Complications.

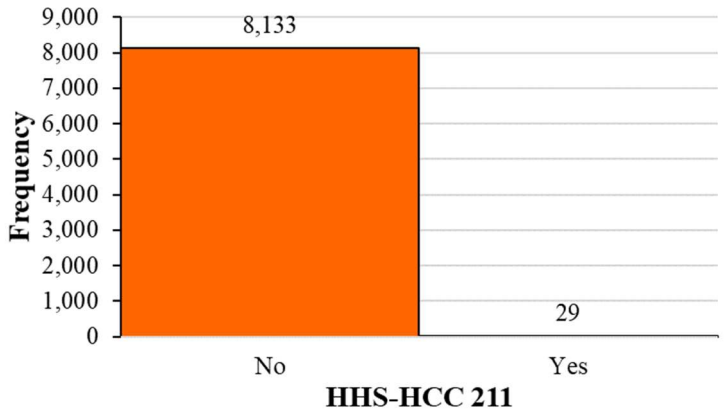


Figure 222. Frequency histogram by (Ongoing) Pregnancy without Delivery with Complications.

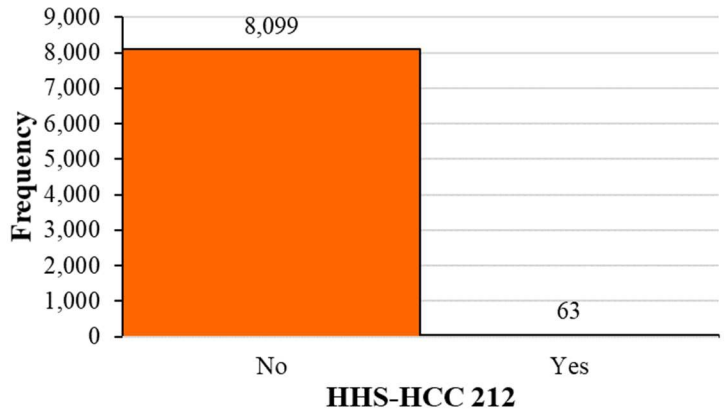


Figure 223. Frequency histogram by (Ongoing) Pregnancy without Delivery with No or Minor Complications.

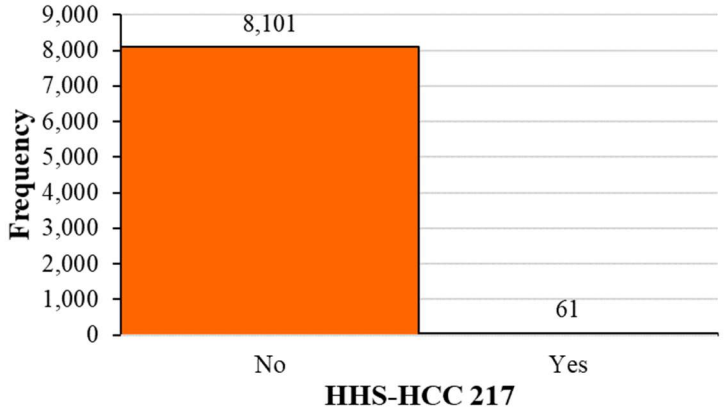


Figure 224. Frequency histogram by Chronic Ulcer of Skin, Except Pressure.

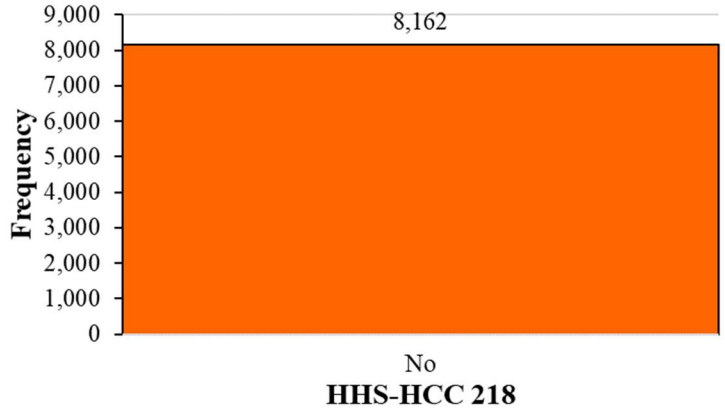


Figure 225. Frequency histogram by Extensive Third Degree Burns.

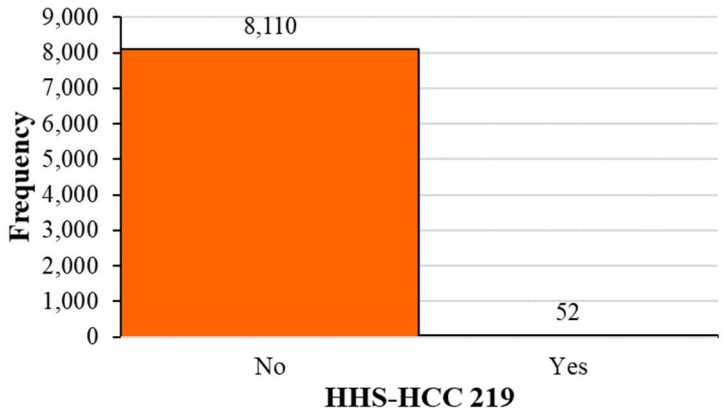


Figure 226. Frequency histogram by Major Skin Burn or Condition.

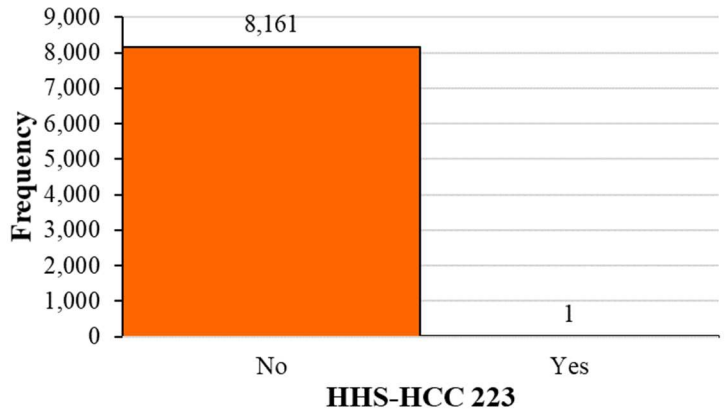


Figure 226. Frequency histogram by Severe Head Injury.

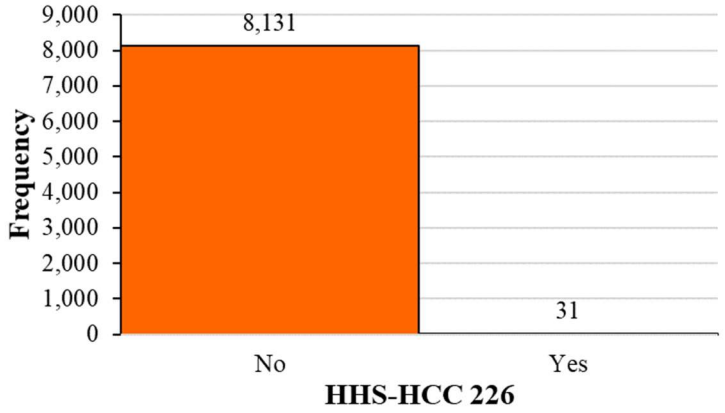


Figure 228. Frequency histogram by Hip and Pelvic Fractures.

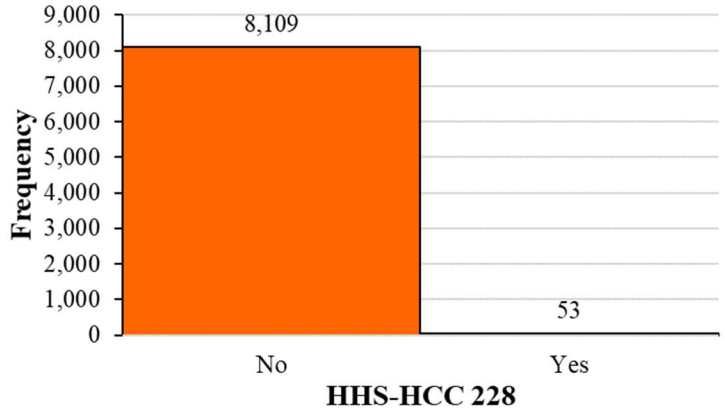


Figure 229. Frequency histogram by Vertebral Fractures without Spinal Cord Injury.

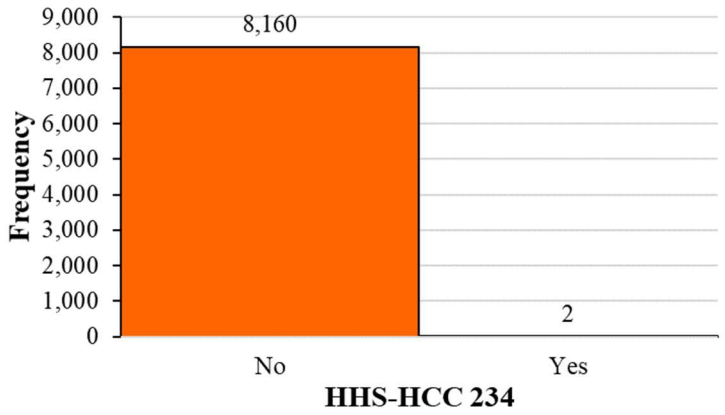


Figure 230. Frequency histogram by Traumatic Amputations and Amputation Complications.

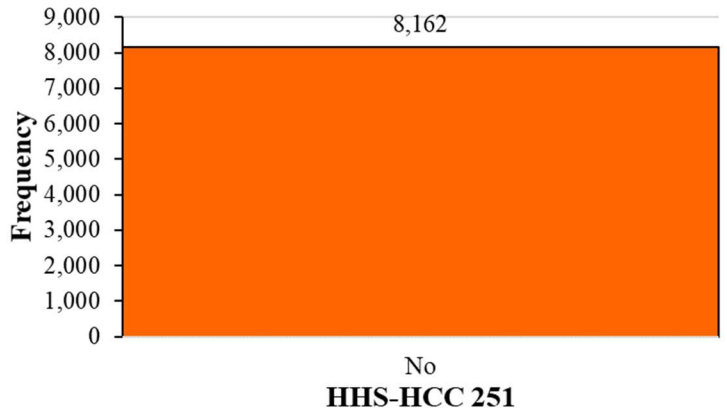


Figure 231. Frequency histogram by Stem Cell, Including Bone Marrow, Transplant Status/Complications.

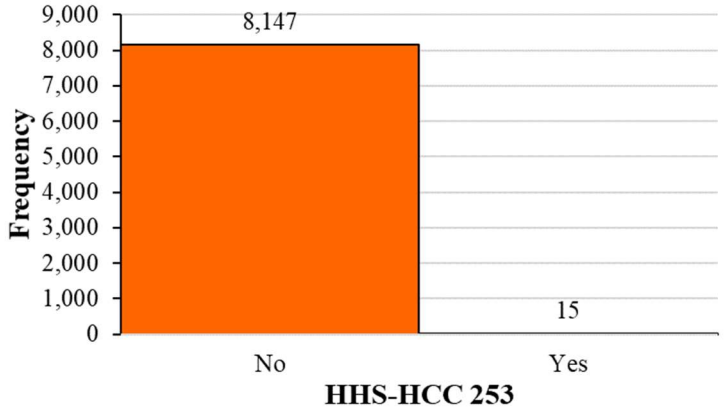


Figure 232. Frequency histogram by Artificial Openings for Feeding or Elimination.

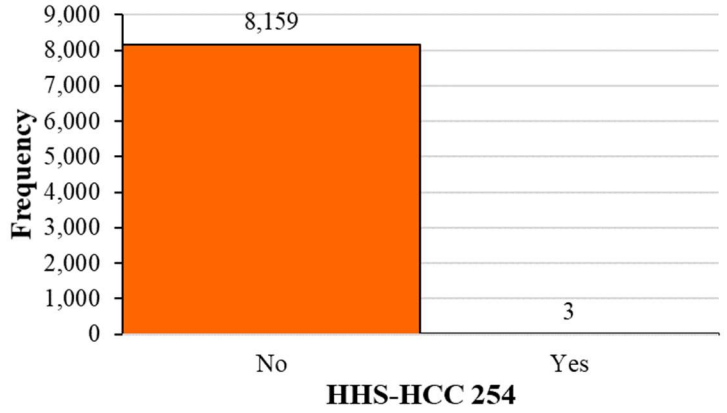


Figure 233. Frequency histogram by Amputation Status, Upper Limb or Lower Limb.

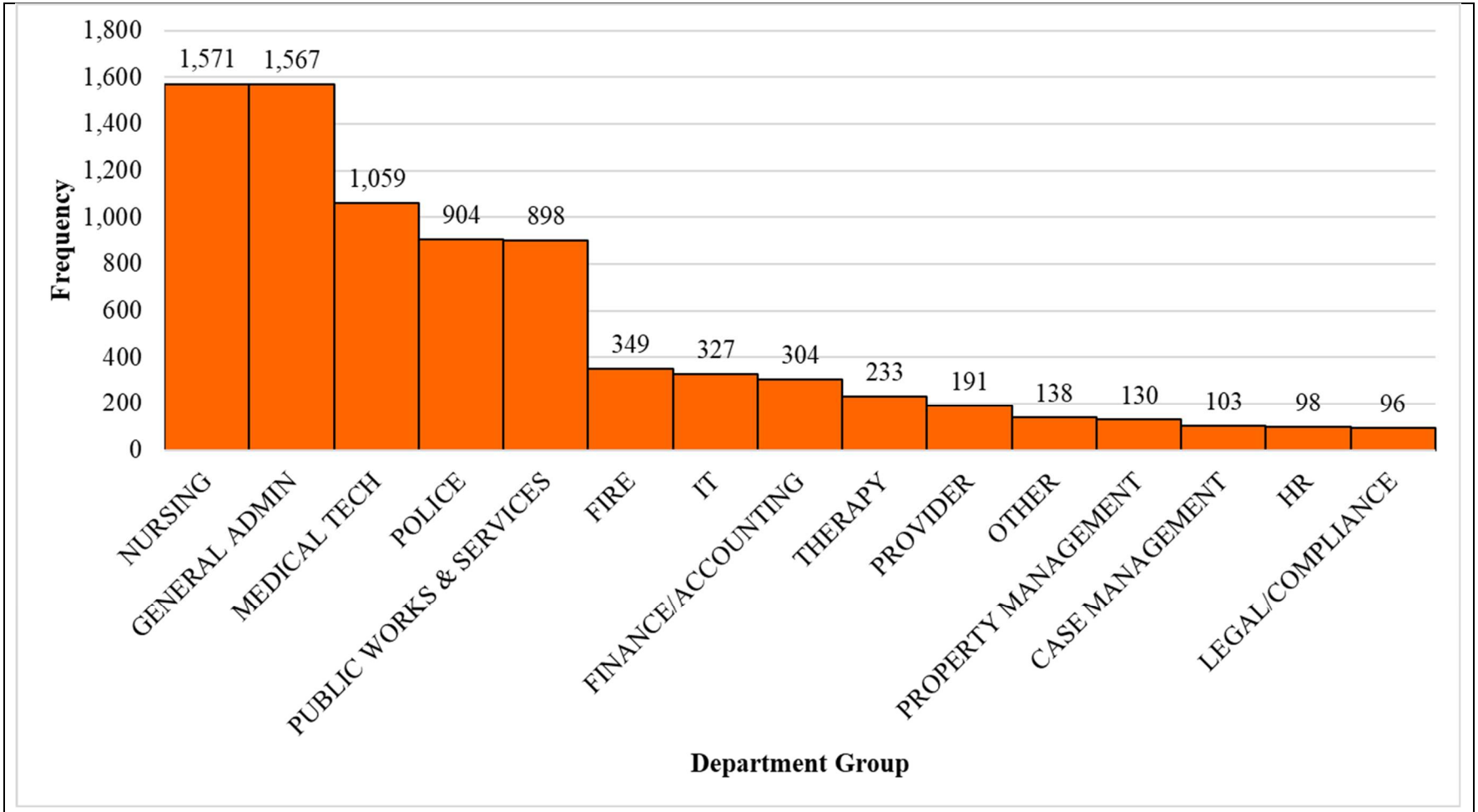


Figure 234. Frequency histogram by the Top 15 Department Groups.

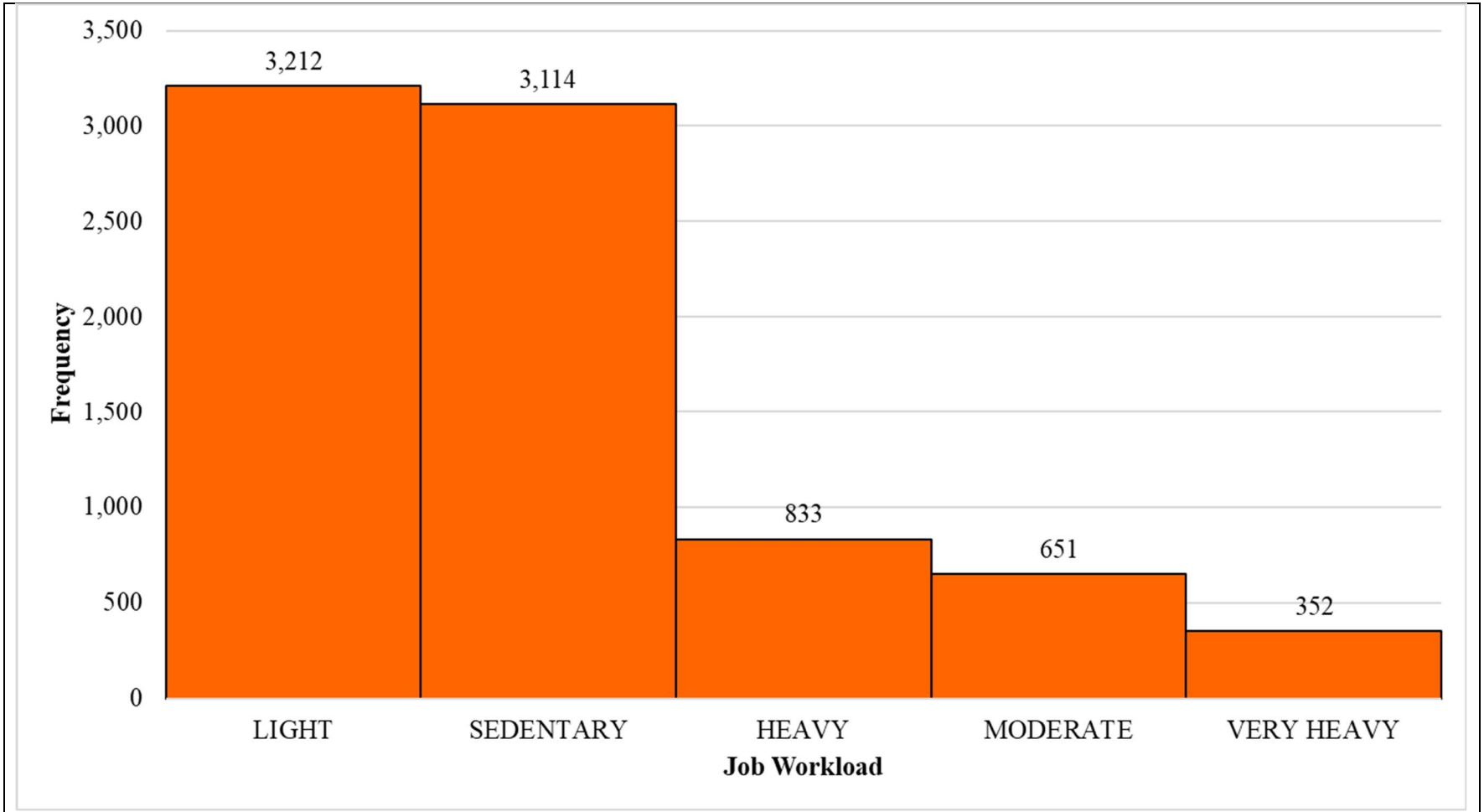
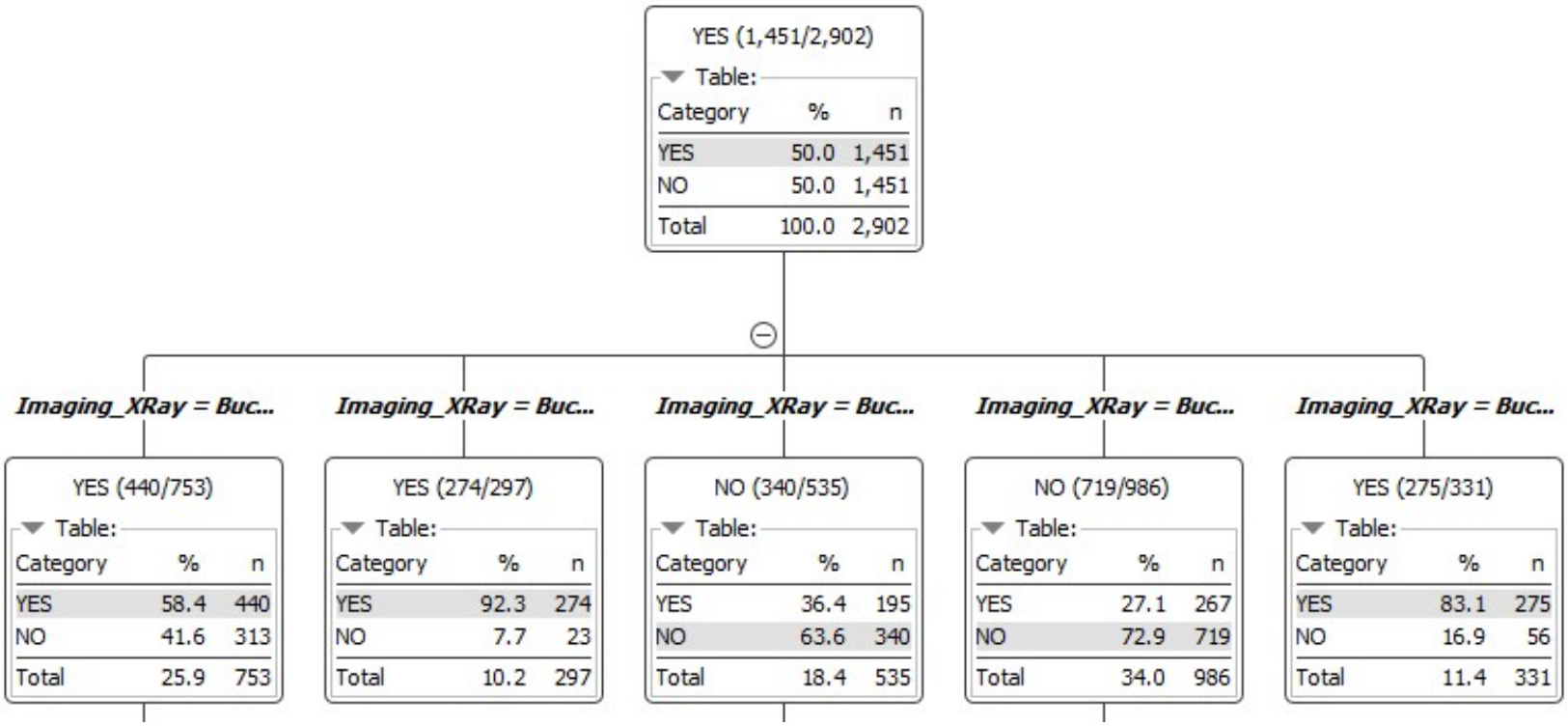


Figure 235. Frequency histogram by the Job Workload.

APPENDIX C: Decision Tree Diagram



VITA

Kenneth Grifno

Candidate for the Degree of

Doctor of Philosophy

Dissertation: STUDYING EMPLOYEE ABSENTEEISM DUE TO HEALTH-RELATED FACTORS: A DATA-SCIENCE APPROACH

Major Field: Business Administration

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Business Administration at Oklahoma State University, Stillwater, Oklahoma in July, 2022.

Completed the requirements for the Master of Science in Management and Administrative Sciences at The University of Texas at Dallas, Richardson, Texas in 2001.

Completed the requirements for the Bachelor of Science in Business Administration at The University of Texas at Dallas, Richardson, Texas in 1999.

Experience:

Chief Analytics Officer, IntegerHealth Technologies, Fort Worth, Texas, 2016 – Current.