Prediction Model for Hospital-acquired Influenza Using Electronic Medical Records

Young Hee Cho

The Graduate School
Yonsei University
Department of Nursing

Prediction Model for Hospital-acquired Influenza Using Electronic Medical Records

A Dissertation
Submitted to the Department of Nursing
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Nursing

Young Hee Cho

June 2022

This certifies that the dissertation
of Young Hee Cho is approved.

_____

Thesis Supervisor: Mona Choi

_____

Hyang Kyu Lee: Thesis Committee Member

_____

Joungyoun Kim: Thesis Committee Member

_____

Ki-Bong Yoo: Thesis Committee Member

_____

Jongrim Choi: Thesis Committee

The Graduate School
Yonsei University
June 2022

감사의 글

한 분야에서 오랜 직장생활로 축적한 지식을 체계화하고 싶다는 생각으로 시작한 도전이었습니다. 오랜만에 듣는 수업은 너무 즐거웠고 공부가 재미있었습니다. 하지만 곧 그저 저의 부족함을 한없이 느끼는 초라한 시간들의 연속이었습니다. 학위논문을 마치면서 많이 아쉽지만, 많이 배우는 시간이었고 그 크기를 떠나 어제보다 한 발자국 발전이 있었기에 만족합니다.

공부하는 내내 어느 한 분 고맙지 않은 분들이 없고 그 도움으로 끝까지 완주할 수 있었습니다. 부족한 학생을 입학부터 졸업까지 이끌어 주신 최모나 교수님께 진심으로 감사드립니다. 부족한 논문이지만 따듯한 마음으로 지도해주신 이향규 교수님, 김정연 교수님, 유기봉 교수님, 최종림 교수님께도 너무 감사드립니다. 어려울 때마다 독려해주고 모르는 거 많은 저를 이끌어준 동기들 덕에 끝까지 올 수 있었고, 일에 소홀하다고 질타하기보다 응원해 준 삼성 SDS 동료 여러분께도 진심으로 감사드립니다. 친구들의 응원은 넘어지려 할 때마다 저를 일으켜 주었고, 먼저 경험했기에 친한 친구의 고생길이 보여 더 걱정이 많았던 친구들의 조언은 정말 큰 도움이었습니다. 그리고, 가족들의 지지로 직장생활과 함께하는 학업을 무사히 마칠 수 있었습니다.

힘든 시간이었지만, 분에 넘치는 사랑과 따듯한 마음들을 느낄 수 있어 행복한 시간이었습니다.

그리고, 어머니, 아버지 사랑하고 존경합니다.

꾸준히 새로운 것을 공부하시는 모습을 본받아 저도 늦은 박사공부를 시작할 수 있었고, 언제나 성실하신 모습을 본받아 박사과정을 마칠 수 있었습니다.

철없던 대학시절 저에게 간호학은 참 재미없는 학문이었습니다. 조금 철들어서 다시 만난 간호학은 아주 훌륭한 학문이었습니다. 제가 간호학을 선택한 것이 기쁘고, 간호학이 더 많이 발전하기를 바랍니다.

그리고, 이제 저도 간호학에 조금이나마 기여할 수 사람이 되겠습니다.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Prediction Model for Hospital-acquired Influenza Using Electronic Medical Records

Young Hee Cho

Department of Nursing

The Graduate School

Yonsei University

Background: Hospital-acquired influenza (HAI) is under-recognized in spite of its high morbidity and poor health outcomes. It is important that nurses detect influenza infections early to prevent its spread in hospitals. This study was conducted to identify characteristics and factors associated with HAI and develop HAI prediction models based on electronic medical records (EMR) using machine learning.

Methods: This study was a retrospective observational study that included 111 HAI patients and 73,748 non-HAI patients of a tertiary hospital in South Korea. General characteristics, comorbidities, vital signs, laboratory results, chest X-ray results, and room information in their EMR were analyzed. Chi-square and t-test univariate analyses were performed to identify HAI infection characteristics and logistic regression (LR), random forest (RF), extreme gradient boosting (XGB) and artificial neural network (ANN) were used to develop the prediction model.

Result: HAI patients had significantly differences in general characteristics, comorbidities, vital signs, laboratory results, chest X-ray results and room status from non-HAI patients. All prediction models had AUC over 70% (LR: 84.9%, RF: 83.4%, XGB:71.1%, ANN: 76.5%). Staying in a double room contributed most to prediction power followed by vital signs, laboratory results.

Conclusion: All of the prediction models developed in this study exceeded acceptable performance criteria. They would help nurses detect HAI patients earlier and take better infection prevention strategies.

Keywords: Influenza, Hospital-acquired influenza, Prediction model, Machine learning, Logistic regression, Random Forest, Extreme gradient boosting, Artificial neural network, Double room, Vital sign

I. Introduction

1. Background

Hospital-acquired influenza (HAI) has high morbidity and mortality and causes high medical costs due to longer hospital stays (Maltezou, 2008; Enstone et al., 2011; Taylor et al., 2014). Prior studies reported nearly a quarter of all inpatients diagnosed with influenza in hospitals had HAI (Mitchell et al., 2013; Huzly et al., 2015). Mortality rates were reported from 9% (Huzly et al., 2015) to 18.8% (Godoy et al., 2020), furthermore it goes up to 39.2% in critical illness patients (Alvarez-Lerma et al., 2017).

Nevertheless, most healthcare providers think of influenza as a community-acquired infection (Choi et al., 2017), and HAI is under-recognized because they are discharged before being diagnosed with influenza due to the incubation period (Macesic et al., 2013). However, HAI patients have longer LoS, stay longer in intensive care units (ICUs), and have higher mortality rates than community-acquired influenza (CAI) patients (Alvarez-Lerma et al., 2017; Godoy et al., 2020). In addition, these poor outcomes of HAI requires medical resources that could be used to treat other patients.

Inpatients can be transmitted by influenza from infected family members, visitors, healthcare workers, and other inpatients through direct or indirect contact (Chow &

Mermel, 2017; Parkash et al., 2019). In South Korea, 77% of rooms in tertiary hospitals and 79% of rooms in general hospitals are multi-occupancy rooms with an average of 4.2 beds per room (Korea Healthcare Bigdata Hub, n.d). In addition, it is common for family members or professional caregivers to stay with patients in the hospital room to care for them and many others come to visit. Thus, patients are more vulnerable to influenza infection in this environment.

Furthermore, influenza has an incubation period and is the most contagious in the first 3 to 4 days after symptoms begin. However, some individuals can spread the virus even when they have no or weak symptoms, which leads to outbreak of influenza in hospital settings (Keilman, 2019). Therefore, it is important that nurses detect influenza infections early regardless of whether patients show symptoms or not and provide the preventive care to infected patients.

There are few nursing studies about hospital-acquired respiratory virus infections and many fewer about HAI infection (Choi et al., 2017). Nursing studies about respiratory infection are limited to caring for patients undergoing ventilation therapy in the ICU. There are dearth studies about influenza infection prevention. In aspect of nursing, it is important that nurses provide the preventive nursing care as well as respiratory care after infection

occurs. Therefore, nurses should understand the risk factors for influenza protection to provide sufficient preventative care.

Traditionally, a hypothesis is formed, and data is collected to determine if it is supported when conducting academic research. As data mining has become more popular, hypotheses could be developed based on patterns observed in data (Hey et al., 2009). Data science has been used in nursing both research and practice, but it is still relatively rare (Westra et al., 2017; Linnen et al., 2019). Data analysis studies have become increasingly common around the world since 2014, but they are still relatively rare in the context of nursing studies in South Korea (Jeong, 2020).

The knowledge discovered by data mining would help nurses to make better decisions improving the quality of nursing care (Courtney et al., 2005; Linnen et al., 2019). This study was conducted to develop an HAI prediction model using data mining that helps nurses make better decisions to prevent the spread of influenza in hospitals.

2. Purpose and Specific Aims of Study

This study was conducted to develop an HAI infection prediction model using Electronic Medical Records (EMR) to help nurses detect HAI patients early to ultimately reduce the spread of HAI.

The specific aims of this study are:

1) To identify characteristics and factors associated with HAI based on EMR data.

2) To develop and evaluate the prediction models to identify best model for HAI.

3. Definitions of Terms

Conceptual definition: HAI infection refers to the case of patients who do not have any influenza-like symptoms, such as high fever, when they are admitted to the hospital but later exhibit such symptoms a certain amount of time after admission. However, definitions of high fever are heterogenous with a common threshold of 37.5–38 ℃ in the literature. The definitions for the minimum amount time between admission and symptom onset required for HAI diagnosis are heterogenous as well, from 24–96 hours in various studies and countries (Munier-Marion, Benet, & Vanhems, 2017).

Operational definition: In this study, HAI infections refer to cases of patients who are admitted to the hospital with a diagnosis for a condition other than influenza and confirmed to have influenza by laboratory testing performed more than four days after admission, considering the incubation period of influenza (Kimberlin et al., 2015). CAI infections refer to cases of patients who are admitted with only a diagnosis of influenza or confirmed to have influenza by laboratory testing performed within four days of admission. Non-hospital-acquired influenza (non-HAI) infection refers to cases of both CAI infections and non influenza infection among inpatients who stay more than four days in a hospital.

II. Literature review

A systematic search for studies published between 2011 and 2021 was undertaken in the following biomedical databases: PubMed, cumulative index for nursing and allied health literature (CINAHL), Medline Complete, Research Information Sharing Service (RISS), and Korean Studies Information Service System (KISS). Medical Subject Heading search terms were “influenza, human,” “cross infection,” “influenza,” “hospital acquired infection,” and “nosocomial”, a general search was made using the search term “flu”, and the term “avian” was excluded (Fig. 1).

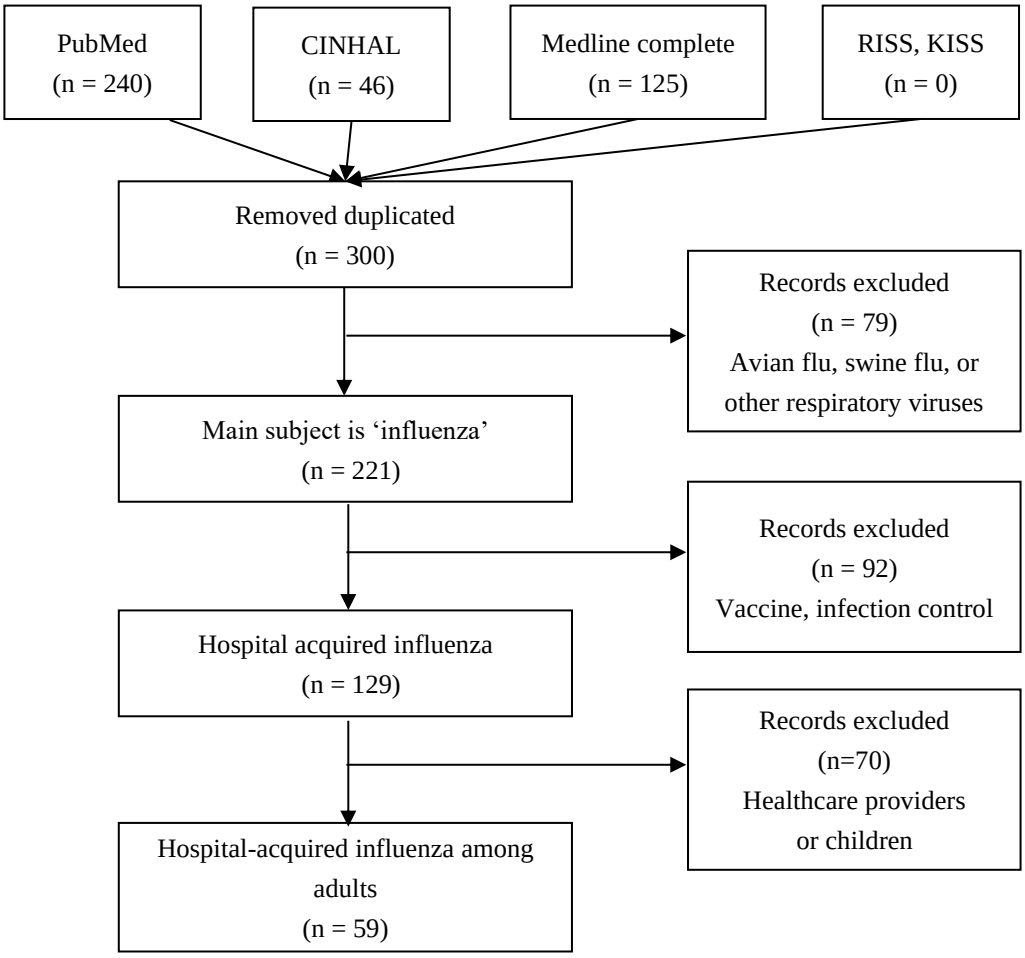


Figure 1. Literature review flow.

1. Hospital-acquired Influenza

There are four types of influenza viruses: A, B, C, and D. Influenza A, B, and C infect humans. Influenza A and B are the cause of seasonal influenza that are epidemic from late

fall to early spring in every year (Paules & Subbarao, 2017). Influenza A infections may develop into a global outbreak when new or major changed influenza A virus are spread among and infect humans quickly (CDC, n.d.). Influenza A can infect animals as well, but influenza B can only affect humans. Influenza C infection usually results in mild respiratory symptoms and occurs in sporadic cases or localized outbreaks, so it does not cause epidemics (Keilman, 2019). Influenza D infects cows, but not humans (CDC, n.d.). Influenza A and B are significantly more common cause of both CAI and HAI than influenza C (Macesic et al., 2013; Parkash et al., 2019; Godoy et al., 2020).

Influenza symptoms can range from mild to severe. Age and comorbidities of chronic disease are associated with symptom severity (Keilman, 2019). Signs and symptoms of influenza include chills, dry cough, persistent cough, diaphoresis, discomfort, fever or feeling feverish, headache, myalgia, sneezing, joint pain, sore throat, nasal congestion, and rhinorrhea (Killingley & Nguyen-Van-Tam, 2013). These signs and symptoms are similar to those of the common cold, but influenza symptoms arise suddenly and include pain and high fever that can last for 3 to 4 days. Young children, weak people, and the elderly may experience nausea or vomiting that can lead to viral or bacterial secondary pneumonia or diarrhea (CDC, n.d.). However, HAI patients may not present symptoms because of other treatments they are receiving which can delay the detection of HAI (Maltezou, 2008).

HAI infections are those in which the patient does not have any influenza symptoms when they are admitted to the hospital, but they do eventually show symptoms after an incubation period and test positive for influenza infection. However, the definition of HAI is not standardized. Various incubation periods are used in studies (Munier-Marion, Benet, Dananche, et al., 2017). Studies have found HAI infection rates in a wide range of 3–24% of influenza patients, partially because they used different incubation periods (Parkash et al., 2019). Countries that do have national HAI surveillance systems use different delay time from admission to symptom onset. Australia's delay time is 48 hours (Macesic et al., 2013) while Canada's is 96 hours (Taylor et al., 2014). Munier-Marion, Benet, Dananche, et al. (2017) found that the delay time used by studies had a range of 48–196 hours and that 75% of studies used the median 72 hours.

HAI patients have worse outcomes after treatment than CAI patients. HAI patients have longer LoS than CAI patients (Salgado et al., 2002; Maltezou, 2008; Macesic et al., 2013; Alvarez-Lerma et al., 2017; Godoy et al., 2020), stayed longer in ICUs (Maltezou, 2008; Alvarez-Lerma et al., 2017; Godoy et al., 2020; Li et al., 2021) and have higher mortality rate (Maltezou, 2008; Enstone et al., 2011; Taylor et al., 2014; Alvarez-Lerma et al., 2017; Godoy et al., 2020).

2. Risk Factors for Hospital-acquired Influenza

The risk factors for HAI include all risk factors for influenza infection, HAI's characteristics, and hospital environment characteristics. Young children (Hall, 2001; Paes et al., 2011; Kondrich & Rosenthal, 2017) and the elderly over 65 years old (Falsey et al., 2005; Murata & Falsey, 2007; Walsh, 2011); the immunosuppressed (Alvarez-Lerma et al., 2017); those receiving treatment for cancer, HIV or AIDS; those taking corticosteroid medication; and those taking medications for long periods of time (Agarwal et al., 2018) are generally greater risk of influenza infection than the general population. The elderly are particularly vulnerable because the immune system weakness with age and the elderly generally take medication over long periods for time (Agarwal et al., 2018). Those who are pregnant; obese; or have a chronic disease, such as asthma; a hematologic disorder; chronic obstructive pulmonary disease (COPD); heart disease; renal disease; or liver disease are more likely to be infected by influenza than others (Keilman, 2019). Smoking is also with influenza (Han et al., 2019). However, the risk factors associated with infection are not clearly distinguished from the factors associated with severity of symptoms or severe complication after infection in many studies of influenza.

Table 1 summarizes 16 studies about characteristics and risk factors for HAI. Eleven studies compared HAI patients with CAI patients and one compared inpatients who

got HAI with those who did not despite the fact that both groups were exposed to influenza in a hospital. Although the other studies did not compare infected and non-infected groups, they all identified common characteristics among of HAI patients. The minimum average age of infection was 52 years old except for one study though most reported an average of over 70 years old. Eleven studies of the 13 reporting sex ratios reported that more men were infected than women. Although pregnancy and smoking are risk factors for influenza infection generally, however they were not generally included HAI studies. Immunosuppression was found to be a risk factor in 10 studies. Comorbidities, such as diabetes, obesity, cardiovascular disease, and malignancy, were reported as risk factors for HAI although studies reported different orders or rate of them.

Yang et al. (2020) collected the data about HAI patients (n = 93) and selected a control group (n = 93) who stayed in the same units with HAI patients for at least seven days from the date on which HAI patients were diagnosed with an HAI infection. They compared age, sex, smoking status, pregnancy, comorbidities, laboratory findings, radiology findings, corticosteroids consumption, influenza vaccine status, length of stay, and mortality of the HAI and control groups. They found that HAI group had higher rate of lymphocytopenia, hypoalbuminemia, and pleural effusion than the control group. Bischoff et al. (2020) examined vital signs, namely temperature, systolic blood pressure (SBP), and diastolic

blood pressure (DBP); laboratory test results, namely white blood cell (WBC) count and creatinine blood levels; age; sex; and comorbidities. They found that the HAI group had a higher WBC count than the control group.

The environmental risk factors associated with influenza infection are the group activities with high possibilities of exposing to viruses or bacteria like going to school, work or daycare, living in a group home, nursing home, or dormitory, and army. Hospitalization is also an environmental risk factor for HAI. Being an inpatient in a multi-occupancy room or room share with influenza patients are also environmental risk factors for HAI. Those in double room are more likely to get an HAI infection than those in single rooms (Munier-Marion et al., 2016; Luque-Paz et al., 2020).

Parkash et al. (2019) mapped HAI patients' rooms with rooms where influenza patients stayed for the incubation period preceding their diagnosis date, and from their diagnosis date to their discharge date. They found that 22 of the 28 HAI patients stayed in the same room as an influenza patients and 17 stayed in the same unit with an influenza patient for the incubation period. Sansone et al. (2019) found that HAI infection occurred in the same units as well as in the same rooms where influenza patients were staying.

In conclusion, inpatients are a high-risk group for influenza infection. They are generally susceptible to influenza infection because they are more likely to have low

immunity and related comorbidities and often share rooms or units with other patients who may be infected with influenza. However, there are still insufficient number of studies about HAI, so it is necessary to study its characteristics and risk factors and develop a prediction model based on them.

Table 1. Characteristics of HAI patients.

Comparison group	Authors	Subject	HAI patients' rate†	HAI definition‡	Characteristics of HAI patients	Comparison results
Matched-case control study	Yang et al. (2020)	2018–2019 season Tertiary hospital (n = 186)	23%	Seven days or more after admission	- Age: 58 years, Male: 53.8% - Immunosuppression: 16.1% - Comorbidities: Hypertension (41.9%) Coronary heart disease (21.5%) Cerebrovascular disease (20.4%) Diabetes (17.2%) - Laboratory results: Lymphocytopenia (51.6%) Anemia (55.9%) Hypoalbuminemia (78.5%) - Radiology results: Pleural effusion (26.9%)	Age, Male Immunosuppression Hypertension Coronary heart Cerebrovascular Lymphocytopenia Hypoalbuminemia Pleural effusion
CAI	Taylor et al. (2014)	2006–2012 season Canadian Nosocomial Infection Surveillance Program (n = 3,299)	17.3%	96 hours or more after admission or readmission within 96 hours after discharge with symptoms	- Age: 81 years, Male: 48.8% - Immunosuppression (14.6%) - Comorbidities: Chronic heart disease (31.7%) Chronic lung disease (23.1%) Chronic kidney disease (12.2%)	Age Immunosuppression Chronic heart disease Chronic lung disease Diabetes Chronic kidney disease

Table 1. Characteristics of HAI patients (continued).

Comparison group	Authors	Subject	HAI patients' rate†	HAI definition‡	Characteristics of HAI patients	Comparison results
CAI	Alvarez-Lerma et al. (2017)	Jan.2009-Dec. 2015 Registry of patient with influenza A (n = 1,327)	9.3%	Seven days or more after admission	- Age: 53 years, Male: 63.3% - Immunosuppression: 20.5% - Comorbidities: Obesity (37.9%) COPD (21%)	Age Immunosuppression Influenza vaccine Hematologic disease Pregnancy APACHE II SOFA
CAI	Jhung et al. (2014)	2010–2011 US Influenza Hospitalization Surveillance Network (n = 6,171)	2.8%	Four days or more after admission	- Age: 54years, Male: 50% - Comorbidities: Cardiovascular disease (40%) Metabolic disease (39%) Asthma or chronic lung disease (39%)	Chronic lung disease Cardiovascular disease Metabolic disease Renal disease Immunosuppression
CAI	Macesic et al. (2013)	2010–2011 Australian Sentinel Surveillance System (n = 598)	4.3%	Symptom onset two or more days after admission or tested positive seven days	- Age: 52years, Male: 53.8% - Smoking: 3.8% - Immunosuppression: 50% - Comorbidities: Diabetes (23.1%) Malignancy (15.4%)	Immunosuppression Malignancy

Table 1. Characteristics of HAI patients (continued).

Comparison group	Authors	Subject	HAI patients' rate†	HAI definition‡	Characteristics of HAI patients	Comparison results
				or more after admission	Renal disease (15.4%)	
CAI	Godoy et al. (2020)	2010–2015 season 12 hospitals in Spain (n = 1,722)	5.6%	Two days or more after admission	- Age (≥75 years): 40.6%, Male: 55.2% - Immunodeficiency: 71.9% - Comorbidities: Obesity (85.4%) COPD (72.9%) Diabetes (64.6%) Chronic renal disease (70.8%) Heart disease (60.4%) Liver disease (91.7%)	Immunodeficiency Diabetes Chronic renal disease Heart disease
CAI	Huzly et al. (2015)	Jan. 2013–Apr. 2014 Academic hospital (n = 218)	23.8%	Three days or more after admission	- Age: 55.2 years, Male: 52% - Immunosuppression: 78% - Comorbidities: Blood malignancy (44%) Organ transplantation (40%) Cardiovascular disease (25%) Chronic lung disease (25%)	Age Immunosuppression Blood malignancy Organ transplantation

Table 1. Characteristics of HAI patients (continued).

Comparison group	Authors	Subject	HAI patients' rate†	HAI definition‡	Characteristics of HAI patients	Comparison results
					Renal impairment (28%)	
CAI	Hagel et al. (2016)	2014–2015 season 1400-bed tertiary hospital (n = 197)	35.5%	Four days or more after admission or readmission within 48 hours after discharge with influenza symptom	(All influenza patients) - Age: 72 years, Male: 54.8% - Immunosuppression: 22.8% - Comorbidities: Diabetes (36%) Heart disease (43.1%) Chronic renal disease (28.4%) COPD (21.8%)	-
CAI	Sansone et al. (2020)	2016 season 1900-bed academic hospital (n = 435)	26%	Symptom onset two days or more after admission or within two days after discharge	- Age: 80 years - Charlson Score: 2	-
CAI	Naudion, Lepiller,	2016 season Tertiary	23.6%	Two days or more	- Age: 79 years, Male: 34.7% - Immunosuppression: 18.4%	Age

Table 1. Characteristics of HAI patients (continued).

Comparison group	Authors	Subject	HAI patients' rate†	HAI definition‡	Characteristics of HAI patients	Comparison results
	and Bouiller (2020)	hospital (n = 208)		after admission	- Comorbidities: Chronic heart disease (7.2%) Malignancy (5.8.4%) Diabetes (5.3%)	
CAI	Bischoff et al. (2020)	2017–2018 season 885-bed tertiary hospital (n = 111)	9.7%	Four days or more after admission	- Age: 62 years, Male: 62.1% - Immunosuppression: 24.3% - Comorbidities: Heart disease (44.4%) Diabetes (41.7%)	High WBC count
CAI	Parkash et al. (2019)	2017 Surveillance system in Canberra hospitals (n = 292)	9.6%	Symptom onset two days or more or tested positive seven days or more after admission	- Age: 79 years, Male: 64.3% - Pregnant: 10% - Ex-smoker: 43.8% - Current smoker: 18.8% - Immunosuppression (14.8%) - Comorbidities: Heart disease (50%) Diabetes (42.9%) Neurological disease (38.5%) Malignancy (30.8%)	Chronic respiratory disease Diabetes Malignancy Chronic liver disease

Table 1. Characteristics of HAI patients (continued).

Comparison group	Authors	Subject	HAI patients' rate†	HAI definition‡	Characteristics of HAI patients	Comparison results
					Obesity (18.2%) - Sharing a room or same unit with an influenza patient	
CAI	Luque-Paz et al. (2020)	2017–2018 season 1500-bed tertiary hospital (n = 860)	6.6%	Two days or more after admission	- Age: 82 years, Male: 54.4% - Comorbidities: Diabetes (21.1%) Heart disease (19.3%) - Double room: 68.4%	Age Influenza type (A or B)
Others	Munier-Marion et al. (2016)	2004–2011 season Academic hospital (n = 93)	-	Two days or more after admission	- Double room	-
Others	Veenith et al. (2012)	2010–2011 season Tertiary hospital (n = 83)	12%	Four days or more after admission	- Age: 44years, Male: 70% - Influenza vaccination 20% - Immunosuppression: 50%	-

Table 1. Characteristics of HAI patients (continued).

Comparison group	Authors	Subject	HAI patients' rate†	HAI definition‡	Characteristics of HAI patients	Comparison results
Others	Sansone et al. (2019)	2015–2016 season Sweden hospital (n = 20)	-	2 days or more after admission	- Age: 77years - Charlson score 4 - Same room with influenza patients (45%) - Same unit with influenza patients (30%)	-

CAI Community-acquired influenza, *COPD* Chronic obstructive pulmonary disease, *SOFA* Sequential organ failure assessment, *APACHE* Acute physiological assessment and chronic health evaluation

† Percentage of influenza patients with HAI

‡ All definition are no influenza-like-symptoms at admission and laboratory confirmed in common.

III. Conceptual Framework

The epidemiologic triangle model is frequently used to explain how communicable diseases spread. The model illustrates the interaction between the essential components causing infectious disease, namely agents, hosts, and environment, and a host becomes ill when one or more factors among three factors are changed (McEwen & Wills, 2017). In other words, a patient gets influenza when the patient with weak immunity against influenza virus (host) exposes to the influenza virus (agent) in a hospital room (environment). According to the epidemiologic triangle model, disease could be prevented by inhibiting exposure to agents or improving a host's physical condition to resist disease, or minimizing environmental factors developing a disease (McEwen & Wills, 2017).

Relevant host factors include age, sex, race, ethnicity, marital status, economic status, immunity status, and lifestyle factors like diet, exercise, hygiene, occupation, and sexual health. There are three types of agents: biologic organisms like bacteria, fungi, and viruses; physical agents like radiation, extremes of temperatures, and noise; and chemical agents like poisons, allergens, and gases. Environment factors include physical elements like climate, season, and geology; biological entities like animals, insects, food, and drugs; and

socio-economic elements like family status, public policy, and culture (McEwen & Pullis, 2009; McEwen & Wills, 2017).

Figure 2 illustrates this study's conceptual framework based on the epidemiologic triangle model. Agents in this study were influenza A and B viruses, which cause seasonal influenza outbreaks and can infect humans.

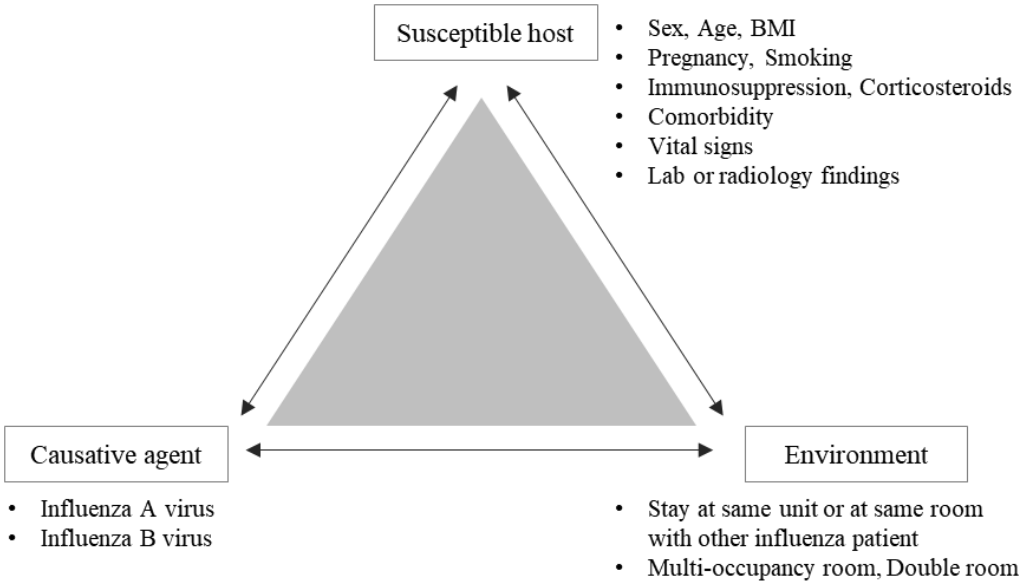


Figure 2. Conceptual framework.

Host factors were sex, age, body mass index (BMI), pregnancy status, smoking status, immunosuppression status, whether the person is taking corticosteroids, whether the person has comorbidities, vital signs, laboratory results, and radiology results. Vital signs reflect the patient's condition and can be used to identify clinical deterioration because they are collected and recorded regularly as a basic nursing activity. Churpek et al. (2014) state vital signs are the most accurate data to identify patient deterioration.

Environmental factors were whether a patient stayed in the same room or unit with an influenza patients. And whether the hospital room was multi-occupancy or a double room. When a susceptible patient is exposed to patients infected with influenza A or B virus by sharing with a room, they are likely to get influenza.

IV. Methods and Materials

This study used the knowledge discovery and data mining (KDDM) approach to develop the prediction model. It is often used to predict unknown value of other variables of interest from some variables in the database. The six steps are interactive and iterative, not separate from each other (Fig. 3) (Delen, 2014; Park et al., 2020). In this study, five steps were performed to develop the prediction model.

Figure 3. Knowledge discovery and data mining process.

1. Research Design

This study was designed as a retrospective observational study to identify characteristics of HAI and develop an HAI prediction model using EMR data.

2. Study Setting

A. Study Population

The population of this study was patients over 19 years old admitted in hospitals in South Korea. The accessible population was patients over 19 years old admitted in tertiary hospitals in South Korea. The sample of this study was drawn from patients over 19 years old admitted in Severance hospital in the Yonsei University Health System, a tertiary teaching hospital located in Seoul, South Korea.

Inclusion criteria were:

a) Patients were over 19 years old.

b) Patients stayed in general adult wards, because patients in other wards, such as the ICU, might have unstable vital signs and laboratory findings due to other conditions.

c) Patients stayed in the hospital for more than four days, because a shorter stay might mean that a patient could have gotten HAI but would not have tested due to incubation period of influenza (Kimberlin et al., 2015).

Exclusion criteria were:

a) Patients only had a diagnosis of J09 (Influenza due to identified zoonotic or pandemic influenza virus), J10 (Influenza due to seasonal influenza virus), or J11

(Influenza, virus not identified) because such patients would be considered to have CAI infections.

b) Patients had a positive polymerase chain reaction (PCR) test within four days of admission because they would be considered to have CAI infections.

c) Patients underwent surgery during this admission.

B. Study Period

The period examined in this study was from the 2011-2012 influenza season to the 2019-2020 season (Table 2), and influenza season lasts from October to April of the following year. Interest in influenza grew since the influenza A H1N1 outbreak in 2009, however the hospital of this study did not have any HAI patients in 2009 and 2010. The COVID-19 outbreak had a significant effect on influenza prevalence because people engaged in more preventative activities, such as washing their hands and wearing masks (Wong et al., 2021). Therefore, the 2019-2020 season excluded March and April 2020 because the COVID-19 pandemic outbreak began in March 2020. A total of 189,321 patients were included in this study, 117 of whom were HAI patients and 182,204 were non-HAI patients.

Table 2. Study period.

Season	From	To
2011 - 2012	October 2011	April 2012
2012 - 2013	October 2012	April 2013
2013 - 2014	October 2013	April 2014
2014 - 2015	October 2014	April 2015
2015 - 2016	October 2015	April 2016
2016 - 2017	October 2016	April 2017
2017 - 2018	October 2017	April 2018
2018 - 2019	October 2018	April 2019
2019 - 2020	October 2019	February 2020

This study was approved by Yonsei University Health System Institutional Review Board (IRB No. 4-2021-1252) and Data Review Board (DRB No. 2021300331). After approval, data was extracted and anonymized by authorized personnel of the hospital's records management department before being sent to the researcher.

3. Study Variables

This study was conducted to develop prediction models for HAI infections, so the variables examined were those agent, host, and environmental factors associated with influenza infection (Table 3).

A. Observation Period

The observation period for each patient was defined as the four days before symptoms presented given the incubation period of influenza (Kimberlin et al., 2015). To define the index date of observation period, it was necessary to know when symptom was presenting. However, it was difficult to get the data whether or when the patient had presented influenza symptoms because this study was retrospective and based on EMR. Thus, it was assumed that patients would take PCR tests after nurses had checked if they had presented symptoms. So PCR test date was designated as the index date for determining the observation period. Bischoff et al. (2020) and Jhung et al. (2014) also used the test date as the index date. In the case of non-HAI patients who did not take a PCR test, the observation period was defined the first four days after admission, and the index date was defined the fifth date after admission (Fig. 4).

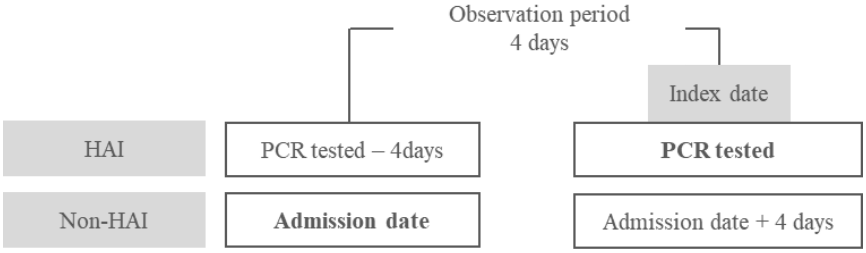


Figure 4. Observation period of vital signs and laboratory and radiology results.

B. Agent Factors

Agent factor was the seasonal influenza infection, i.e., influenza A or influenza B virus. The outcome variable was defined as the result of influenza A or influenza B virus PCR test which was taken more than four days after admission. The positive result PCR testing was categorized into the HAI group. If a patient did not take a PCR test, the patient was categorized into non-HAI. Patients whose PCR tests were negative would have been categorized as non-HAI. However, given that they took the test because they had symptoms and the test is not 100% accurate, they may actually have been HAI patients. Thus, these patients were excluded because they could have improperly influenced the training of the prediction model (Fig. 5).

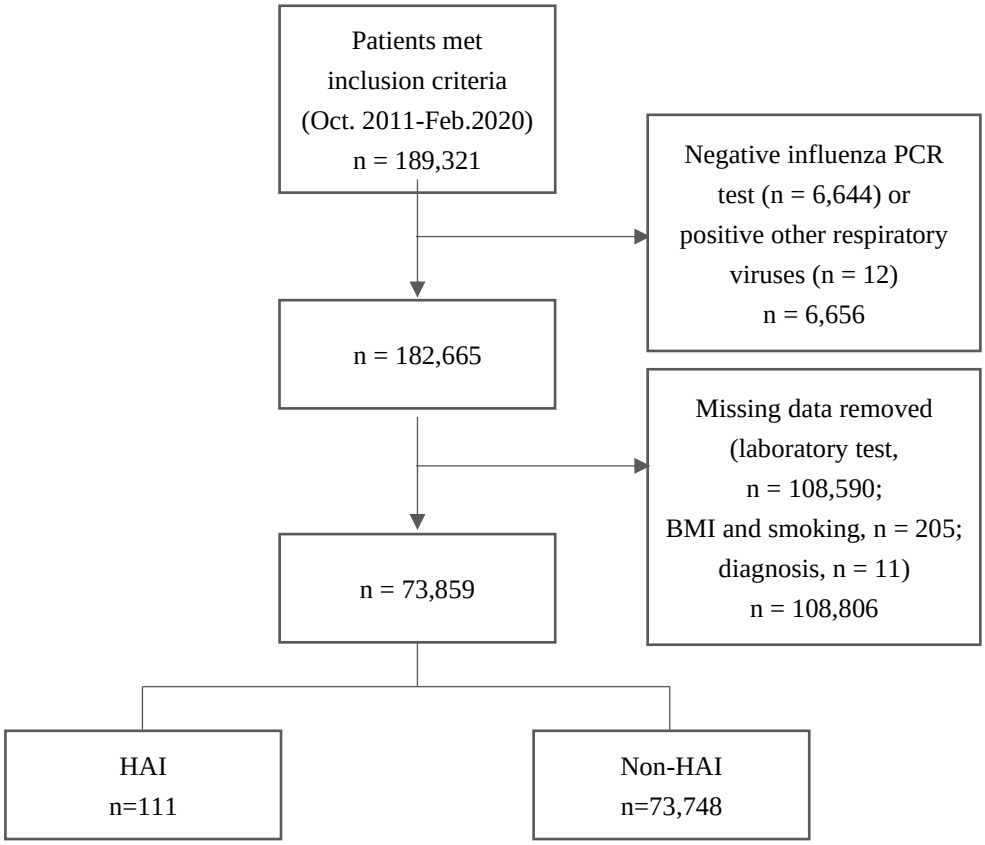


Figure 5. Study population selection.

C. Host Factors

The host factors were the patient's condition during the observational period for this study. Patient conditions were defined in terms of general characteristics, comorbidities, vital signs, laboratory test and radiology test results based on not only a literature review but also the availability of data.

General Characteristics

General characteristics were sex, age, BMI, pregnancy status, and a previous or current smoker. These data for the inpatient episode were used because it would not likely have changed significantly during their admission period. Immunosuppression status and the use of corticosteroids were also categorized as general characteristics. Naudion et al. (2020) defined patients as immunosuppressed if they took 10 mg or more of prednisolone-equivalent steroids, monoclonal antibodies, antimetabolite drugs, or T-cell inhibitors within 30 days preceding the index date. However, in this study, patients were classified as immunosuppressed if they took these medications during the observation period because data before admission were not consistently available in the EMR. Patients were classified as having taken corticosteroids in the same manner.

Patients' medications in their EMR only included brand names, so patient medications were mapped to immunosuppression-related drugs and corticosteroids (Fig. 6). First, generic names of these drugs were found in a list provided by the Korea Pharmaceutical Information Center (KPIC) and then brand names for them were found in a drug list provided by Health Insurance Review Assessment Service (HIRA). Finally, those brand names were mapped to the list of medications taken by patients in this study.

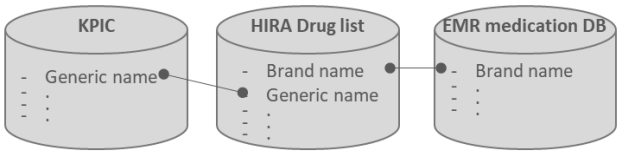


Figure 6. Medication mapping.

Comorbidities

Patients were defined as having comorbidities if they had diagnoses of diabetes, obesity, heart disease, liver disease, renal disease, hematologic disease, malignancy, organ transplantation, asthma, or COPD before the index date. Diagnoses were classified based on International Classification of Diseases 10 (ICD 10) codes (World Health Organization, n.d.).

Vital Signs

Bischoff et al. (2020) used vital signs to discover factors associated with HAI, and it did not show any significant difference in HAI patients. However, additional research using vital signs is required because there are insufficient HAI literature using vital signs and only one-time vital sign on the date testing PCR was used in Bischoff et al. (2020). Thus, in this study, temperature, heart rate, respiration rate, SBP, and DBP over the four days before the index date were analyzed.

Laboratory and Radiology Results

Laboratory and radiology results were included following Yang et al. (2020) who found that they were correlated with influenza infection. In addition, the hematological inflammatory parameters neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PNR), and platelet-to-lymphocyte ratio (PLR) were selected following Han et al. (2020), which reported the sensitivity of NLR to diagnose influenza was higher than the common systemic inflammatory makers (i.e., neutrophil and lymphocyte count). Hematological inflammatory indexes also had a significant predictive value for the diagnosis and prognosis of infections (Han et al., 2020). Chest X-ray result were selected for radiology results. The results closest to the index date were used.

D. Environment Factors

Hospital room type is a risk factor for HAI infection (Munier-Marion et al., 2016; Luque-Paz et al., 2020). Staying in the same room or unit with an influenza patient is also a risk factor (Parkash et al., 2019; Sansone et al., 2019). Therefore, patients' rooms and units during the observation period were included as environment factors.

Rooms were classified as double rooms or multiple-occupancy rooms which were rooms capable of housing more than two patients. Patients who stayed in the same room or

unit as an influenza patient were defined as having stayed with an influenza patient. Influenza patients were deemed to be such from four days before their index date to their discharge date, regardless of whether they had HAI or CAI.

4. Data Preparation

The data obtained from the hospital came in 12 datasets: a main patient list, nursing records, medication records, vital sign records, laboratory test results, chest X-ray results, BMI records, transfer records, diagnoses, a list of those tested for influenza, and a list of influenza patients. These records were deidentified and all patients were given research episode numbers instead. The 12 datasets were linked with the research episode number as a unique key.

The outcome variable was whether the patient had an HAI infection or not. The host variables of age and BMI were numerical variables while the other general characteristics were binary. Comorbidity variables were binary based on ICD 10 codes as well.

Using the simple value like the most recent vital signs or the difference from the previous ones can conceal clinical deterioration (Churpek et al., 2016). This study transformed vital signs using the method of mean and changes from the previous

values (Churpek et al., 2016). Vital signs were typically measured three times per day at the general wards in this study hospital. Patients' vital signs were calculated as the difference between the current and the average of the three preceding values (variation), δ_j. The largest variation, $max\ (\delta_j)$, was then selected from all of the variation values during the observation period and was defined as follows:

$$max(\delta_j) = max\left(\frac{\sum_{i=j-3}^{j-1} v_i}{3} - v_j\right)$$

The laboratory results were numerical variables. NLR was defined as the neutrophil count divided by the lymphocyte count, PNR was defined as the platelet count divided by the lymphocyte count, and PLR was defined as the platelet count divided by the lymphocyte count. Chest X-ray results were coded with normal (0), abnormal (1), and no result (9). All variables related to hospital rooms were binary.

Handling Missing Data

In this study, there were no laboratory results for 108,590 patients, smoking or BMI information for 205 patients, and diagnosis information for 11 patients. All 108,806 of

these patients were removed (Fig. 5). A total of 73,859 patients remained of which 111 had HAI infections.

The direct bilirubin variable was removed due to missing for 80.8% of patients. The variables with higher missing rate were calcium (4.6%) and total bilirubin (3.7%). Univariate analysis showed that these variables between HAI and non-HAI patients were significantly different. The missing rates of the laboratory results for alanine transaminase (ALT, 2%), albumin (1.1%), aspartate transaminase (AST, 0.9%), blood urea nitrogen (BUN, 0.4%), creatinine (0.3%), and tCo2 (0.02%) were less than 2%. Therefore, data for missing laboratory test variables other than direct bilirubin count were imputed. The fact that a laboratory test result is missing means that the doctor did not think that the patient required that test, so missing laboratory test results were not considered to be abnormal (Hu et al., 2017). The median value of the normal ranges of continuous laboratory variables was imputed when they were missing.

Table 3. Study variables.

Factor category	Category	Variables	Type /Level	Description
Agent	Seasonal influenza	HAI	Y/N	A positive PCR test for influenza A or B more than four days after admission
Host	General characteristics			
		Age	Number	Age on admission
		Sex	M/F	
		BMI	Number	
		Pregnancy status	Y/N	Removed after univariate analysis
		Ex-smoker status	Y/N	
		Current smoker status	Y/N	
		Immuno-suppressed	Y/N	Immunosuppressant administered during this admission
		Corticosteroid use	Y/N	Administered during this admission
	Comorbidities			ICD 10 codes
		Diabetes	Y/N	E10, E11, E13
		Obesity	Y/N	E66.9 Removed after univariate analysis
		Heart disease	Y/N	I05–I09, I20–I25, I27, I30–I52
		Liver disease	Y/N	K70–K77
		Renal disease	Y/N	N00–N08, N10–N16, N17–N19, N25–N29

Table 3. Study variables (continued).

Factor category	Category	Variables	Type /Level	Description
		Hematologic disease	Y/N	D50–D53, D55–D59, D60 –D69, D70–D77
		Malignancy	Y/N	C00–C97
		Organ transplantation	Y/N	Z94
		Asthma	Y/N	J45, J46
		COPD	Y/N	J44
	Vital signs			Largest differences between the current and the average of the three preceding values
		Temperature	Number	
		Heart rate	Number	
		Respiration rate	Number	
		Systolic blood pressure	Number	
		Diastolic blood pressure	Number	
	Laboratory test results			Latest value during the observation period
		Red blood cell (RBC) count	Number	
		Hemoglobin	Number	
		WBC count	Number	

Table 3. Study variables (continued).

Factor category	Category	Variables	Type /Level	Description
		Platelet count	Number	
		Hematocrits	Number	
		Red blood cell distribution width (RDW)	Number	
		Delta neutrophil index (DNI)	Number	
		Neutrophil count	Number	
		Lymphocyte count	Number	
		Neutrophil-to-lymphocyte ratio	Number	Neutrophil count / lymphocyte count
		Platelet-to-neutrophil ratio	Number	Platelet count / neutrophil count
		Platelet-to-lymphocyte ratio	Number	Platelet count / lymphocyte count
		Na	Number	
		K	Number	
		Cl	Number	
		tCO2	Number	
		Calcium	Number	
		Albumin	Number	

Table 3. Study variables (continued).

Factor category	Category	Variables	Type /Level	Description
		Total bilirubin	Number	
		Direct bilirubin	Number	Removed, missing rate: 80.8%
		BUN	Number	
		Creatinine	Number	
		ALT	Number	
		AST	Number	
	Radiology test results			
		Chest X-ray	(0/1/9)	Normal/Abnormal/None
Environ-mental factor	Room information			
		Same room	Y/N	Share a room with an influenza patient
		Same unit	Y/N	Stay a same unit with an influenza patient
		Multi-occupancy room	Y/N	
		Double room	Y/N	

HAI Hospital-acquired influenza, *PCR* Polymerase chain reaction,

ICD International Classification of Diseases 10, *COPD* Chronic obstructive pulmonary disease

Note: No HAI patients were pregnant or obese, so these were removed for prediction model development.

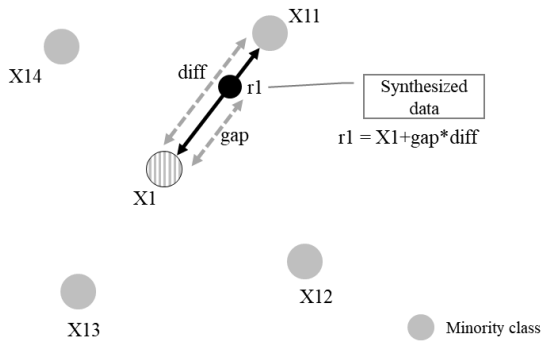
5. Data Analysis

To identify characteristics and factors associated with HAI, descriptive and univariate analyses were performed. After the raw data were processed, there were 53 variables in seven categories (Table 3). The prevalence of HAI was calculated. Chi-square tests and t-tests were used for the differences in HAI and non-HAI patients for categorical and continuous variables, respectively.

To develop and identify best prediction model for hospital-acquired influenza, classification modeling was used and is further explained in the following section. Of the 73,859 patients in this study, only 111 (0.15%) had HAI, making the data unbalanced. Unbalanced classes are often seen in real-world healthcare data and can lower the predictive power of prediction models (Park et al., 2020; Turlapati & Prusty, 2020).

The two main methods for dealing with imbalanced data are undersampling and oversampling. In undersampling, all of the minority cases are used but only a sample of the majority cases. In oversampling, all of the majority cases are used and more minority cases than in the original sample are used. Oversampling is preferred over undersampling because undersampling may leave out important data (Turlapati & Prusty, 2020). Sometimes, both methods are used together. The synthetic minority oversampling technique (SMOTE) is an oversampling method that creates new and relatively accurate

data based on existing minority cases (Chawla et al., 2002; Turlapati & Prusty, 2020). SMOTE creates data by calculating the Euclidean distance between any two randomly chosen k-nearest neighbors (KNN) from two minority samples and creating new data along the line between them (Fig. 7) (Turlapati & Prusty, 2020). For example, first, the X1 minority class is randomly selected. Then its k=4 KNNs are identified as X11, X12, X13, and X14. One of these k instances is chosen to interpolate new synthetic instance by calculating the distance between the feature vector, X1, and its neighbor, X11 using any distance metric. This difference is multiplied by any random value (gap) between zero and one and is then added to the previous feature vector, generating synthetic data, r1. SMOTE is only used in the training dataset, not in the test dataset.



Example case with k = 4 (Inoue, n.d.)

Figure 7. Example of how SMOTE generates data when k = 4.

Machine Learning Methods

In this study, random forest (RF), extreme gradient boosting (XGB), and artificial neural network (ANN) machine learning classification methods and the logistic regression (LR) method were tested. LR is used in the studies with binary outcome variables, like being infected with a disease or not (Delen, 2014). In addition, LR is widely used to predict patient outcomes, like death or disease onset, and compared with machine learning methods in healthcare data analysis studies (Dai et al., 2015; Bloch et al., 2019). LR is vulnerable to overfitting, however it showed robustness in many domains and effectiveness at estimating probabilities of binary variables (Long et al., 1993).

RF is an ensemble model of decision tree, which was introduced by Breiman (2001). Ensembling combines several weak classifier models into one strong classifier model that performs better than one of its component models (Lee, 2020; Adnan et al., 2022). Decision tree algorithms are sometimes sensitive to minor cases in datasets, but RF is not by aggregating the results of many different decision trees. It handles non-linear data well and is not at high risk of overfitting (Sahni et al., 2018). RF is also a bagging ensemble method. Bagging aggregates the results of several models built with several datasets generated by bootstrapping (Lee, 2020). Ensemble models generally take longer to train, but even simple ensemble models perform well (Lee, 2020; Adnan et al., 2022).

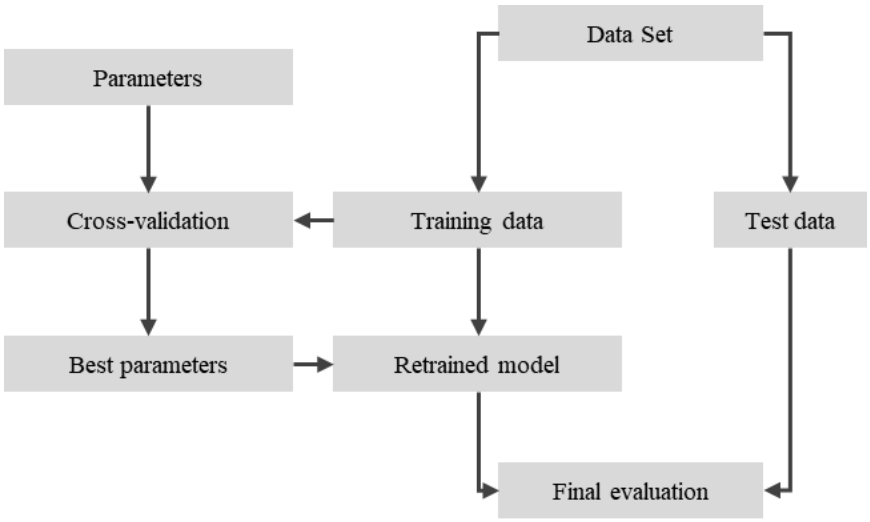
XGB is based on the gradient boosting model (GBM), which is introduced by Chen & Guestrin (2016). GBM is a tree-based ensemble method (Desai et al., 2020) that uses boosting which is another typical ensembling technique. GBM's performance is reliable but it takes a long time to train. XGB significantly reduces this training time. It is one of the most advanced supervised machine learning algorithms and is faster than other ensemble classifiers (Adnan et al., 2022).

ANN is widely used and have high predictive power among classification algorithms (Delen, 2014). It was inspired by the human nervous systems such that procession elements (PE) correspond to neurons (Lee, 2020). PEs accept input values and weight them. Groups of PEs form layers. Each ANN consists of an input layer, hidden layers, and an output layer. The input layer receives the values of predictor variables. The number of PEs in the input layer is usually equal to the number of predictors. Hidden layers connect the input and output layers and process the data. The number of hidden layers is determined when constructing the model. Output layers receives the weighted values from the hidden layers and return them to the user. The number of PEs in an output layer is equal to the number of outcome variables. Classification ANNs only have one output PE. The transparency and interpretability of models are important in healthcare (Bloch et al., 2019)

to show why the outcome makes. ANNs have the limitation of interpretability, however they have strong predictive power.

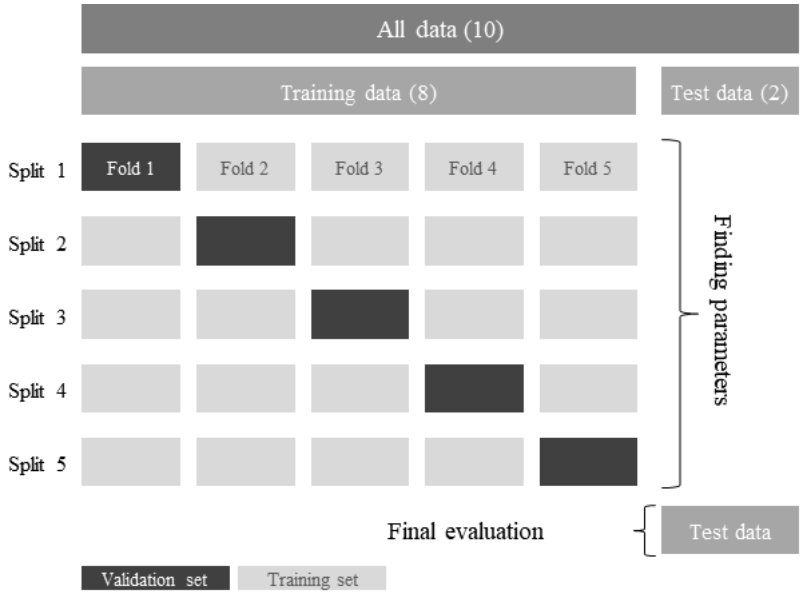
Models should not be trained and evaluated using the same dataset to determine their accuracy (Dreiseitl & Ohno-Machado, 2002). Datasets are generally split into two groups, one of which is used for training and the other is for testing (Penny & Chesney, 2006). In this study, 80% of the dataset was randomly allocated to the training set and the remaining 20% was allocated to the test set.

Five-fold grid search cross-validation (GSCV) was performed on the training set (Fig. 8). GSCV finds the best combination of hyper-parameters that optimizes model performance while avoiding overfitting (Barton et al., 2019; Adnan et al., 2022). Each machine learning technique has tuning parameters known as hyper-parameters (Golas et al., 2018). For example, the number of hidden layers in an ANN is a hyper-parameter. GSCV creates multiple models with unique combinations of hyper-parameters during model training process and evaluates their performance using cross-validation. In five-fold cross-validation, the training dataset is randomly split into five folds, then four of which are used to train the model and the fifth is used to validate it. This process is repeated five times with a different fold used as the validation dataset each time (Fig. 9). Finally, GSCV selects the model that shows the best performance (Adnan et al., 2022).



Müller (n.d.)

Figure 8. Data analysis process using grid search cross-validation.



Müller (n.d.)

Figure 9. How datasets are split and used during five-fold cross-validation.

The optimized hyper-parameters of each machine learning model tested in this study are as follows. The RF model had a maximum depth of 20, 2 as the minimum number of sample splits, and 100 n estimators. The XGB model had a maximum depth of 5, a learning rate of 0.2, a subsample of 0.75, and 10 n estimators. The ANN model had 50 and 100 activation-rectified linear units, a hidden layer size of 50, a learning rate of 0.005, and an Adam solver.

6. Model Evaluation

Discrimination ability is the main criterion by which a classification model is evaluated (Park, 2016). It is commonly measured in terms of sensitivity, specificity, accuracy, and the area under the receiver operating characteristics curve (AUC) (Beck & Shultz, 1986). Sensitivity (equal to recall), specificity, positive predictive value (PPV, equal to precision), negative predictive value (NPV), and accuracy are measured in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Positive or negative in binary classification is determined by a threshold probability which the prediction is classified as true (Freeman & Moisen, 2008). This study used the optimal threshold which gave the highest both sensitivity and specificity found by optimal threshold

option of GSCV. A model's F_1 score also reflects its sensitivity and PPV. F_1 scores are less sensitive to data imbalances while accuracy is sensitive to unbalanced classes (Raschka & Mirjalili, 2017). Sensitivity, specificity, PPV, NPV, accuracy, and the F_1 score are calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F_1\ score = 2 \times \frac{Sensitivity \times PPV}{Sensitivity + PPV}$$

AUC is the area under the receiver operating characteristics curve (ROC) which reflects a model's diagnostic ability in classification problem. ROC is the plot of true positive rate and false positive rate by changing threshold (Bloch et al., 2019). The value of AUC is between zero and one, and higher scores reflect greater discrimination. AUC is more than 0.7 are generally considered to be acceptable (Redon et al., 2010).

AUC and the number of FN were of primary interest in this study. AUC is the most widely used metric for evaluating prediction models (Demler et al., 2012). The number of FP is important in a hospital setting because it represents patients who are not being properly treated and so could be spreading the virus. Therefore, the number of FP has significant meaning in this study because the purpose of the developed models was to detect HAI patients early. Thus, the models in this study were evaluated in terms of their AUC, and the number of FN. Sensitivity, specificity, PPV, NPV, accuracy, and F_1 score were also presented in the evaluation result. In addition, feature importance was calculated according to permutation-based method (Altmann et al., 2010). Lastly, a Delong test was performed to compare the models' AUCs (DeLong et al., 1988).

Analysis Software

Data analysis was performed using SQL Server Management Studio v18.10 (Microsoft, Seattle, US) and Python 3.5. SQL was used to integrate, preprocess, and transform data. Python was used for univariate analyses and machine learning.

V. Results

1. HAI Characteristics

A total of 6,554 patients received PCR tests for influenza A or B virus while admitted to the hospital and 1,054 patients (16.1%) of them got positive results (Table 4). The 2011-2012 season had the highest (27.5%) while having the second lowest number of patients (n = 204) who took the test. The 2016-2017 season had the lowest rate of positive results (8.8%) with the highest number of patients (n = 1,173) who took the test.

Table 4. The Patient Number of Influenza Tested, Influenza Confirmed And HAI.

Season	Took PCR test (n)	Positive PCR result (n)	Positive PCR rate† (%)	HAI infections (n)	HAI rate‡ (%)
2011-2012	204	56	27.5	2	3.6
2012-2013	138	31	22.5	2	6.5
2013-2014	390	87	22.3	8	9.2
2014-2015	763	109	14.3	13	11.9
2015-2016	867	134	15.5	11	8.2
2016-2017	1,173	103	8.8	12	11.7
2017-2018	973	239	24.6	33	13.8
2018-2019	1,053	131	12.4	16	12.2
2019-2020	993	163	16.4	20	12.3
Total	6,554	1,053	16.1	117	11.1

† Percentage of test results that were positive ‡ Percentage of influenza infections that were HAI

Of the 1,053 patients with influenza infections, 117 (11.1%) had HAI infections. The lowest number of HAI infections was two in the 2011-2012 and 2012-2013 seasons, which were also the lowest rate of HAI at 3.6% and 6.5%, respectively. The highest number of HAI cases was 33 in the 2017-2018 season, which was also the highest rate of HAI at 13.8%. The greatest number of influenza patients was in the 2017-2018 season at 239 (24.6% of all influenza infections).

Of the HAI cases, 89 (76.1%) were due to influenza A and 28 (23.9%) were due to influenza B. Most HAI cases occurred in January (56 cases, 48%), followed by December (21 cases, 18%) and February (20 cases, 17.1%) (Fig. 10).

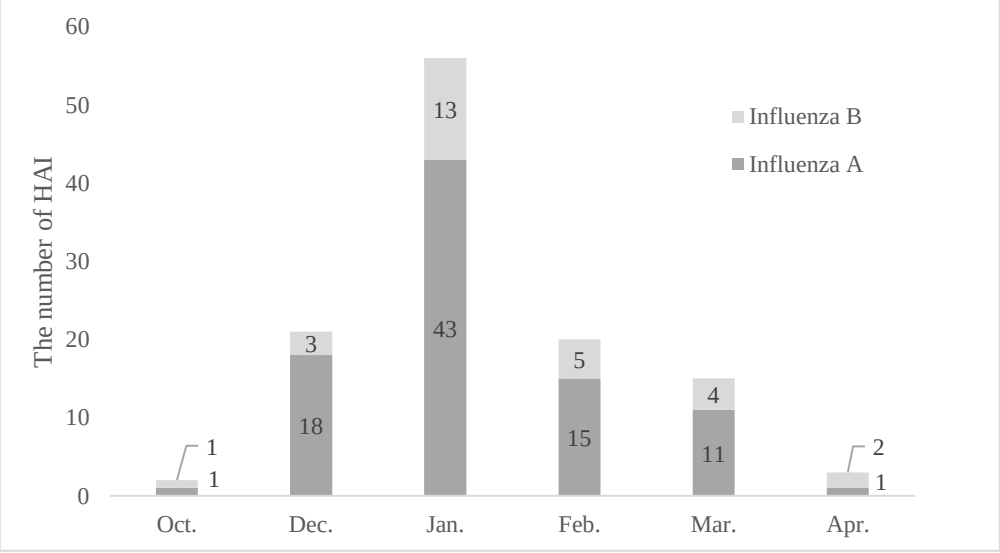


Figure 10. HAI prevalence by month.

2. Characteristics of HAI patients

Table5 shows the HAI patients' characteristics. The average LoS of HAI patients was 12.5 days (SD = 10.9 days) when they received a PCR test. The total LoS of HAI patients was statistically significantly longer than that of non-HAI patients ($p < 0.001$). HAI patients were significantly older than non-HAI patients ($p < 0.001$). The HAI group had a statistically significantly higher rate of immunosuppression and corticosteroid use than the non-HAI group (both $p < 0.001$).

Malignancy was the most common comorbidity in both HAI patients (43.2%) and non-HAI patients (50.4%) followed by heart disease (27.9%, 11.1% respectively), renal disease (27%, 12.3% respectively). There were statistically significant differences between two groups in the incidence of diabetes ($p < 0.001$), heart disease ($p < 0.001$), renal disease ($p < 0.001$), hematologic disease ($p = 0.037$), asthma ($p < 0.001$), and COPD ($p < 0.001$). HAI patients had significantly higher variation of temperature, heart rate, SBP, and DBP than non-HAI patients ($p < 0.001$, $p = 0.002$, $p < 0.001$, and $p < 0.001$ respectively).

With regard to laboratory results, RBC counts, hemoglobin levels, platelet counts, hematocrit levels, and lymphocyte counts were all significantly lower in HAI patients than in non-HAI patients (all: $p < 0.001$). RDW, DNI, and PLR were significantly higher in HAI patients than in non-HAI patients ($p = 0.007$, $p = 0.02$, and $p = 0.04$ respectively). Na, K,

and Cl levels were significantly lower in HAI patients than non-HAI patients ($p = 0.009$, $p < 0.001$, and $p < 0.001$ respectively). Calcium, albumin, and total bilirubin levels were significantly lower in HAI patients than non-HAI patients as well ($p < 0.001$, $p < 0.001$, and $p = 0.028$ respectively). BUN levels were statistically significantly higher in HAI patients than in non-HAI patients ($p = 0.024$). All HAI patients had chest X-ray results while 90.9% of non-HAI patients did. Of patients with X-ray results, statistically significantly more HAI patients had abnormal chest X-ray findings than non-HAI patients ($p < 0.001$). With regard to room status, higher rate of HAI patients stayed in the same room and the same unit with an influenza patient, and double room than that of non-HAI patients (all $p < 0.001$). In summary, HAI patients had significant differences from non-HAI patients in general characteristics, comorbidities, the variation of vital signs, laboratory results, radiology results, and room status.

Table 5. Characteristics of HAI and non-HAI patients.

Variable	Total (n = 73,859)	Non-HAI (n = 73,748)	HAI (n = 111)	t or χ^2	p-value
General characteristics					
LoS at PCR testing, days, mean (SD)	-	-	12.5 (10.9)	-	-
Total LoS, days, mean (SD)	12.5 (15.3)	12.5 (15.3)	27.0 (23.1)	-6.641	< 0.001***
Age, years, mean (SD)	58.9 (16.1)	58.9 (16.1)	68.8 (13.0)	-8.220	< 0.001***
Sex, male, n (%)	40,588 (55.0)	40,529 (55.0)	59 (53.2)	0.082	0.775
BMI, mean (SD)	23.0 (3.6)	23.0 (3.6)	22.9 (4.5)	0.332	0.740
Pregnant, n (%)	716 (1.0)	716 (1.0)	0 (0.0)	0.312	0.577
Ex-smoker, n (%)	15,161 (20.5)	15,135 (20.5)	26 (23.4)	0.408	0.523
Current smoker, n (%)	9,928 (13.4)	9919 (13.4)	9 (8.1)	2.278	0.131
Immunosuppressed, n (%)	19,543 (26.5)	19,495 (26.4)	48 (43.2)	15.240	< 0.001***
Corticosteroid use, n (%)	25,035 (33.9)	24,972 (33.9)	63 (56.8)	24.918	< 0.001***
Comorbidities, n (%)					
Diabetes	4,653 (6.3)	4,635 (6.3)	18 (16.2)	16.875	< 0.001***
Obesity	35 (0.0)	35 (0.0)	0 (0.0)	0.000	1
Heart disease	8,207 (11.1)	8,176 (11.1)	31 (27.9)	30.146	< 0.001***
Liver disease	4,825 (6.5)	4816 (6.5)	9 (8.1)	0.230	0.631

Table 5. Characteristics of HAI and non-HAI patients (continued).

Variable	Total (n = 73,859)	Non-HAI (n = 73,748)	HAI (n = 111)	t or $\chi 2$	p-value
Renal disease	9,104 (12.3)	9,074 (12.3)	30 (27.0)	20.890	< 0.001***
Hematologic disease	4,733 (6.4)	4,720 (6.4)	13 (11.7)	4.366	0.037*
Malignancy	37,214 (50.4)	37,166 (50.4)	48 (43.2)	1.991	0.158
Organ transplantation	3,600 (4.9)	3,596 (4.9)	4 (3.6)	0.161	0.688
Asthma	926 (1.3)	915 (1.2)	11 (9.9)	60.462	< 0.001***
COPD	1,095 (1.5)	1,081 (1.5)	14 (12.6)	86.809	< 0.001***
Vital signs, mean (SD)					
Largest variation for temperature, ºC	0.8 (0.4)	0.8 (0.4)	1.0 (0.4)	-6.117	< 0.001***
Largest variation for heart rate, beats/m	16.4 (10.1)	16.4 (10.1)	19.3 (10.5)	-3.082	0.002**
Largest variation for respiration rate, beats/m	2.4 (4.3)	2.4 (4.3)	2.4 (2.8)	-0.033	0.741
Largest variation for SBP (mmHg)	23.1 (12.6)	23.1 (12.6)	27.4 (12.5)	-3.610	< 0.001***
Largest variation for DBP (mmHg)	16.5 (8.4)	16.5 (8.4)	19.1 (8.5)	-3.340	< 0.001***
Laboratory test results, mean (SD)					
RBC count ($10^3/\mu L$)	3.7 (0.7)	3.7 (0.7)	3.3 (0.7)	5.431	< 0.001***

Table 5. Characteristics of HAI and non-HAI patients (continued).

Variable	Total (n = 73,859)	Non-HAI (n = 73,748)	HAI (n = 111)	t or $\chi 2$	p-value
Hemoglobin (g/dL)	11.3 (2.0)	11.3 (2.0)	10.2 (1.9)	5.723	< 0.001***
WBC count (10^3/μL)	7.6 (4.5)	7.6 (4.5)	6.9 (4.6)	1.571	0.116
Platelet count (10^3/μL)	221.0 (108.6)	221.0 (108.6)	184.8 (102.5)	3.509	< 0.001***
Hematocrits (%)	33.7 (5.8)	33.7 (5.8)	30.5 (5.8)	5.864	< 0.001***
RDW (%)	14.6 (2.2)	14.6 (2.2)	15.2 (2.3)	-2.703	0.007**
DNI	1.4 (3.2)	1.4 (3.2)	1.9 (2.0)	-2.366	0.020*
Neutrophil count (10^3/μL)	5.5 (4.1)	5.5 (4.1)	5.0 (3.4)	1.624	0.107
Lymphocyte count (10^3/μL)	1.3 (0.7)	1.3 (0.7)	0.9 (0.6)	6.711	< 0.001***
Neutrophil-to-lymphocyte ratio	7.1 (48.2)	7.1 (48.2)	8.7 (9.3)	-1.797	0.075
Platelet-to-neutrophil ratio	58.8 (145.5)	58.8 (144.7)	102.6 (427.0)	-1.080	0.283
Platelet-to-lymphocyte ratio	237.8 (356.5)	237.6 (356.2)	329.8 (468.0)	-2.073	0.040*
Na (mmol/L)	138.9 (3.9)	138.9 (3.9)	137.8 (4.5)	2.679	0.009***
K (mmol/L)	4.0 (0.5)	4.0 (0.5)	3.8 (0.5)	4.370	< 0.001***
Cl (mmol/L)	102.7 (4.4)	102.8 (4.4)	101.1 (5.0)	3.420	< 0.001***
tCo2 (mmol/L)	23.8 (3.3)	23.8 (3.3)	23.6 (3.6)	0.729	0.466
Calcium (mmol/L)	8.5 (0.7)	8.5 (0.7)	8.2 (0.8)	4.236	< 0.001***

Table 5. Characteristics of HAI and non-HAI patients (continued).

Variable	Total (n = 73,859)	Non-HAI (n = 73,748)	HAI (n = 111)	t or χ2	p-value
Albumin (mmol/L)	3.4 (0.6)	3.4 (0.6)	3.0 (0.6)	5.912	< 0.001***
Total bilirubin (mmol/L)	1.0 (1.9)	1.0 (1.9)	0.8 (1.0)	2.227	0.028*
BUN (mg/dL)	17.3 (13.5)	17.3 (13.5)	21.0 (16.8)	-2.296	0.024*
Creatinine (mg/dL)	1.1 (1.3)	1.1 (1.3)	1.3 (1.3)	-1.540	0.124
ALT (mmol/L)	36 (98.9)	36 (99)	33.8 (55.6)	0.398	0.692
AST (mmol/L)	41.4 (163.4)	41.4 (163.4)	45.6 (101.8)	-0.423	0.673
Radiology test result, n (%)					
Chest X-ray, Normal	25,121 (34)	25,111 (34)	10 (9.0)		
Abnormal	42,027 (56.9)	41,926 (56.9)	101 (91.0)		
None	6,711 (9.1)	6,711 (9.1)	0.0 (0.0)	53.237	< 0.001***
Room status, n (%)					
Same room	1,542 (2.1)	1,526 (2.1)	16 (14.4)	76.703	< 0.001***
Same unit	9,146 (12.4)	9,095 (12.3)	51 (45.9)	112.340	< 0.001***
Multi-occupancy room	64,858 (87.8)	64,763 (87.8)	95 (85.6)	0.328	0.567
Double room	38,325 (51.9)	38,240 (51.9)	85 (76.6)	26.158	< 0.001***

* p value ≤ 0.05, ** p value ≤ 0.01, *** p value ≤ 0.001

Note: No HAI patients were pregnant or obese, so these were removed for prediction model development.

3. Prediction Model Developments

Prediction models were developed using LR, RF, XGB, and ANN machine learning methods. Models were evaluated using the test dataset in terms of their sensitivity, specificity, PPV, NPV, accuracy, F_1 score, and the numbers of TP, TN, FP, and FN. The LR model had the highest AUC (84.9%) followed by RF (83.4%), ANN (76.5%), and XGB (71.1%) (Table 6). All models' AUCs were over 70%, which means that they all produced acceptable results (Redon et al., 2010), and were not significantly different according to the Delong test results (Table 7).

Table 6. Model evaluation results.

Model	AUC (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)	F_1 Score	Threshold
LR	84.9	72.7	75.0	0.4	99.9	75.1	0.9	0.28
RF	83.4	77.3	77.8	0.5	100.0	77.8	1.0	0.02
XGB	71.1	63.6	72.4	0.3	99.9	72.4	0.7	0.07
ANN	76.5	68.2	73.2	0.4	99.9	73.1	0.8	4.8.E-17

Table 7. Delong test results.

	LR	RF	XGB	ANN
LR	-	0.643	0.052	0.135
RF		-	0.068	0.189
XGB			-	0.535
ANN				-

Figure 11 shows the ROC curves and AUCs of all models. In addition, the RF model had the lowest number of FNs (5) followed by LR (6), ANN (7), and XGB (8) (Table 8).

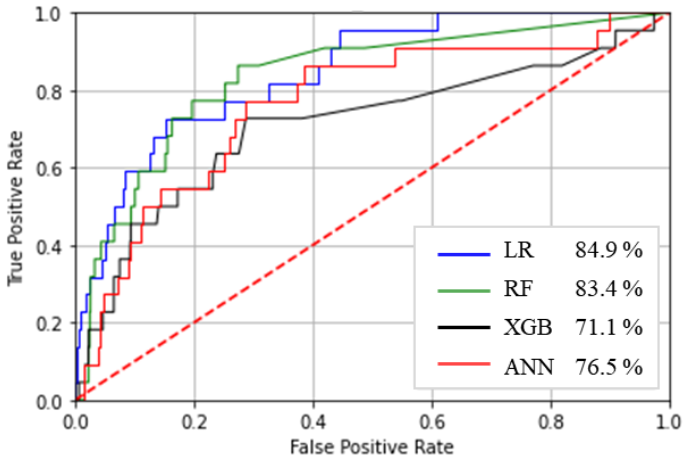


Figure 11. ROC curves and AUCs.

Table 8. Model evaluation results for TP, TN, FP, and FN.

Model	TP (n)	TN (n)	FP (n)	FN (n)
LR	16	11,074	3,676	6
RF	17	11,480	3,270	5
XGB	14	10,684	4,066	8
ANN	15	10,798	3,952	7

The feature importance analysis results are presented in Figure 12. Staying in a double room ranked first followed by DNI, Normal chest X-ray, temperature, and lymphocyte count. Three vital sign features, namely temperature, DBP, and SBP, and three laboratory results, namely DNI, lymphocyte count, and albumin levels, in the top 10 most important features as well.

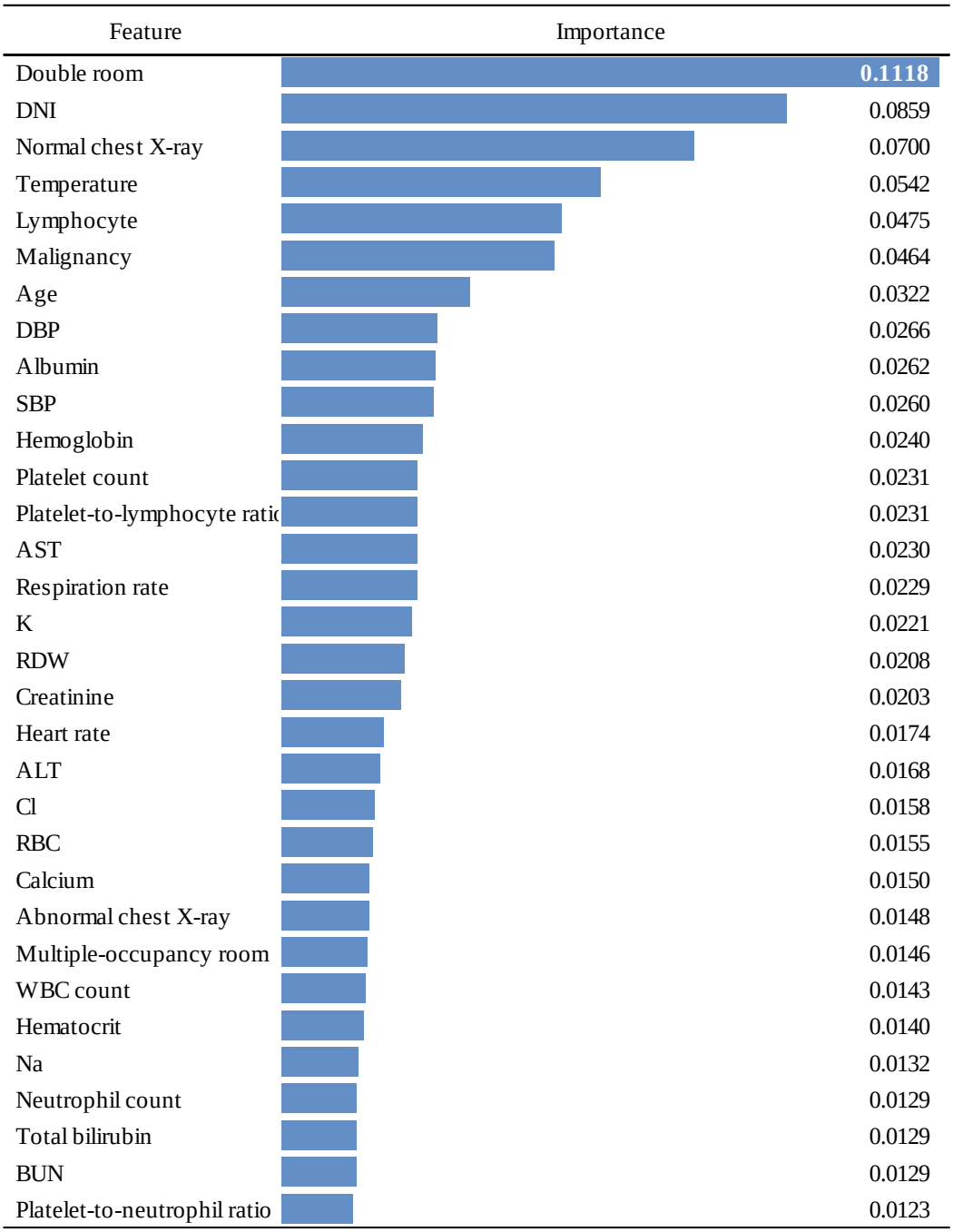


Figure 12. Feature importance analysis results.

Feature	Importance
Neutrophil-to-lymphocyte r	0.0116
Sex	0.0107
Current smoker	0.0103
BMI	0.0097
tCo2	0.0093
Liver disease	0.0081
Organ transplantation	0.0071
Ex-smoker	0.0065
Diabetes	0.0040
Hematologic disease	0.0039
Immunosuppression	0.0026
Corticosteroid	0.0019
Heart disease	0.0019
Renal disease	0.0017
Same unit	0.0016
Same room	0.0004
COPD	0.0002
Asthma	0.0000

Figure 12. Feature importance analysis results (continued).

VI. Discussion

Detecting HAI allows for infection prevention measures to be implemented in a timely fashion, which reduces the likelihood of an influenza outbreak in a hospital. Understanding HAI can help to identify HAI patients among hospitalized patients. This study was conducted to identify HAI characteristics and develop an HAI prediction model that uses EMR.

1. HAI Characteristics

Incidence of influenza of this study showed a similar trend of national influenza (Korea Healthcare Big data Hub, n.d.). Influenza infections in South Korea have increased consistently from 116,409 in the 2011-2012 season to 1,359,095 in the 2018-2019 season, except when it dropped in the 2012-2013 season (49,048). And it increased dramatically from the 2016-2017 season (490,201) to the 2017-2018 season (1,175,966). This study's data showed a similar trend from the 2011-2012 season to the 2017-2018 season but decreased in 2018-2019 season. This study showed that 11.1% of all influenza cases were HAI. Prior studies found that HAI accounted for a broad range of influenza cases, from 2.8% (Jhung et al., 2014) to 35.5% (Hagel et al., 2016). In this study, the highest number

of HAI cases occurred in January, which matches the findings of prior studies (Bischoff et al., 2020; Naudion et al., 2020).

2. Characteristics of HAI patients

In this study, HAI patients took a PCR test on average 12.5 days after admission, which was similar to the average of 12.4 days found by Bischoff et al. (2020). This result indicates that patients are more susceptible to HAI infection when they stay longer in a hospital. In addition, the total LoS of HAI patients in this study was 14.5 days longer on average than non-HAI patients. HAI patients stay longer in hospitals than both non-HAI patients (Yang et al., 2020) and CAI patients (Salgado et al., 2002; Maltezou, 2008; Macesic et al., 2013; Alvarez-Lerma et al., 2017; Bischoff et al., 2020; Godoy et al., 2020).

Most prior studies compared HAI patients with CAI patients, not non-HAI patients. The findings of this study showed similar to the findings of those studies. In this study, HAI patients were older on average than non-HAI patients, which was similar to prior studies (Taylor et al., 2014; Huzly et al., 2015; Alvarez-Lerma et al., 2017; Bischoff et al., 2020). HAI patients were more likely to be immunosuppressed (Macesic et al., 2013; Jhung et al., 2014; Taylor et al., 2014; Huzly et al., 2015; Alvarez-Lerma et al., 2017; Godoy et

al., 2020; Naudion et al., 2020; Yang et al., 2020), and have diabetes (Taylor et al., 2014; Parkash et al., 2019), heart disease (Jhung et al., 2014; Taylor et al., 2014; Godoy et al., 2020; Yang et al., 2020), renal disease (Jhung et al., 2014; Taylor et al., 2014; Godoy et al., 2020), hematologic disease (Alvarez-Lerma et al., 2017), and COPD (Jhung et al., 2014), all of which were similar to other studies.

This study showed the largest variation from the average value of the preceding 24 hours for temperature, heart rate, SBP, and DBP were higher in HAI patients than non-HAI patients. Although Bischoff et al. (2020) did not find a difference in the groups' vital signs, they used a raw value and compared HAI patients to CAI patients. Churpek et al. (2016) found that variations in vital signs were more relevant than their raw values. Churpek et al. (2016) used temperature, heart rate, respiration rate, SBP, DBP, and oxygen saturation over 24 hours to predict cardiac arrest, transfer to the ICU, and death. They transformed these to seven formats, i.e., the difference between consecutive measurements, means, standard deviations, slopes, maximums, minimums, and smoothed values for each vital sign. They found that slope, variation, and maximum were the most accurate predictors, and the difference from the previous value was the most inaccurate. This study found significance of vital signs using transformed formats to reflect variation. However, the relationship

between vital signs and HAI has not been sufficiently studied, so further research in this area should be conducted.

With regard to the hematological parameters, RBC, hemoglobin, platelet, hematocrit, and lymphocyte count were significantly lower and RDW, DNI, and PLR were significantly higher in HAI patients than non-HAI patients. The lymphocyte count result was consistent with Yang et al. (2020)'s while the hemoglobin and platelet results were not. The results of RBC, hemoglobin, platelet, lymphocyte, RDW, PLR in this study were similar to the findings of influenza patients compared with healthy people in Han et al. (2020)'s study. They did not examine the other two parameters, i.e. hematocrit and DNI. Han et al. (2020) conducted the study that influenza infection patients compared with three groups, namely healthy people, patients with bacterial infections, and patients who presented respiratory symptoms but did not have a positive result for either influenza or bacterial infection as a negative control. They found that platelet count of the influenza infection group was lower than those of the healthy and the negative control group, and that the influenza group's platelet count returned to normal when they were cured. Besides being involved in blood coagulation, platelets are also involved in inflammation (Hottz et al., 2018). Influenza virus infection increases platelets activation (Hottz et al., 2018) and excessive activation of platelets could lead to decrease platelet counts (Assinger, 2014; Han

et al., 2020). Thus, a low platelet count can be used to differentiate influenza infection from other types of infections (Han et al., 2020).

The other hematological inflammatory parameters, namely neutrophil and WBC count, were higher in influenza patients than in healthy people but lower than in bacteria-infected patients (Han et al., 2020). These results of this study were not significantly different, which were similar to those found by Yang et al. (2020) who compared HAI and non-HAI patients. These indicate that neutrophil count and WBC count would be more heterogeneous between those with and without influenza infections than platelet (Han et al., 2020). In addition, of the blood cell indexes, PLR was significant while NPR and NLR were not in this study. Both were calculated with neutrophil count, which was also found not to be significant. Nonetheless, hematological parameters may be associated with patients' conditions in other ways, so further research in this area should be conducted.

In this study, all HAI patients received chest X-rays while 90.9% of non-HAI patients did. Of the former, 91% had abnormal findings while only 56.9% of the latter did. Yang et al. (2020) also found that more HAI patients' chest X-ray results showed pleural effusion than non-HAI patients'. This result indicates that patients with abnormal result of chest X-rays are vulnerable to HAI infection.

Higher proportions of HAI patients stayed in the same rooms and units as influenza patients and stayed in double rooms than non-HAI patients. However, there was no difference in the proportions of patients who stayed in multi-occupancy rooms between the groups. Multi-occupancy rooms are more crowded than double rooms and the number of patients, caregivers, and visitors in a room would be a risk factor for influenza infection. However, patients in double rooms stay right next to the possible infection patients, while patients in multi-occupancy rooms could or could not stay right next to them. Although people should remain at least 1.8 meters from influenza patients to reduce the risk of infection (Keilman, 2019), hospital beds are less than 1.8 meters apart. Thus, patients in double room could be more vulnerable to influenza infection by droplet.

In addition, more people come and go from multi-occupancy rooms than double rooms, thus the room is more often ventilated. Influenza infections peak in December and January (Naudion et al., 2020) and people are not likely to open windows during these months, so opening doors is one of the only ways of ventilating the room. Wong et al. (2010) investigated influenza patients' location with the airflow and found the significant role of aerosol transmission of influenza in a hospital. In addition, Xiao et al. (2018) reported that more influenza was transmitted by airborne routes than fomite routes in

hospitals. The role of aerosols in influenza transmission would also explain this study's results.

Identifying how the characteristics of HAI and non-HAI patients differ is difficult because both have medical conditions that are severe enough to require hospital admission. However, this study identified characteristics that differ between them. Nurses can use them to develop infection prevention strategies to reduce the spread of influenza in hospitals.

3. HAI Prediction Model

To our knowledge, this is the first study to have developed an HAI prediction model using machine learning. The predictors were general characteristics, comorbidities, vital signs, laboratory results, radiology results, and room information in EMR. The performance of LR, RF, XGB, and ANN machine learning models was compared. All four models had AUCs of over 71%, but the LR model had the best AUC of 84.9% followed by the RF (83.5%), ANN (76.5%), and XGB (71.1%) models. However, these differences of AUCs were not statistically significant and the RF model generated the least number of FNs (5) followed by the LR (6), ANN (7), and XGB models (8). Therefore, the RF model

would be the best suited for clinical use. The threshold value used in this study was optimized value which both sensitivity and specificity showed best. Sensitivity is more important than specificity in preventing the spread of influenza in hospitals. Therefore, a lower threshold should be used to increase sensitivity in clinical practice.

Staying in double rooms was the most important feature in predicting HAI. This result makes sense because patients staying in double rooms are more vulnerable to influenza infection because of short distance between beds and less ventilation. DNI was the second-most important feature. During the early stages of infection, neutrophils are blocked from migrating to the infection site by the overproduction of cytokines and chemokines, causing immature neutrophils to enter the blood in a process known as left-shifting (Alves-Filho et al., 2010). DNI is the proportion of neutrophils accounted for by immature granulocytes in peripheral circulation (Kratz et al., 2006; Singer et al., 2016). Left-shifting is defined as an increase in DNI (Singer et al., 2016). DNI has more predictive power of infection and prognosis than WBC, C-reactive protein, or neutrophil counts (Bermejo-Martín et al., 2014; Kim et al., 2015; Lee et al., 2016; Kim et al., 2017). DNI is also useful for differentiating between low-grade community-acquired pneumonia from other upper respiratory infections, namely the common cold (Kim et al., 2015). Similarly, DNI was shown to be significant to predict HAI in this study.

HAI patients had greater variation in temperature, heart rate, SBP, and DBP than non-HAI patients. Temperature, SBP, and DBP were also among the 10 most important features while respiration rate and heart rate were ranked 15th and 19th, respectively. Thus, variations in vital signs can be useful to predict HAI infection. Vital signs are used to predict clinical deterioration (Churpek et al., 2014) and have been studied for use in predicting disease and prognoses, including for acute graft-versus-host disease (Tang et al., 2020) and sepsis (Mao et al., 2011; Escobar et al., 2012; Barton et al., 2019; Bloch et al., 2019). Although prior studies were not about influenza, this study found vital signs were also significant to predict HAI infection similarly.

Many vital signs, laboratory results and chest X-ray result were different between HAI and non-HAI patients and were found to be important features to predict HAI infection. Sex, smoking status, immunosuppression status, room status, and comorbidities have been shown to have less predictive ability for HAI infection than vital signs, laboratory results and chest X-ray result in terms of feature importance. This result may indicate that these reflect patients' immediate conditions while demographic and medical history variables do not. In addition, these variables were observed during the incubation period. This would indicate that influenza patients would have changes in vital signs, laboratory test, and chest X-ray before presenting influenza-like symptoms.

In summary, this study developed LR, RF, XGB, and ANN HAI infection prediction models using EMR data. All four models had acceptably good prediction performance. The RF model should be used in clinical practice because it had high AUC and the lowest number of FNs. It should be used with a low threshold to increase its sensitivity.

4. Implications

This study used data mining methods to help nurses detect HAI patients. Hospitalized patients have many influenza infection risk factors, such as relatively old age, low immunity, and comorbidities, all of which can cause difficulties in detecting possible HAI patients. Furthermore, hospitalized patients could present influenza-like symptoms, e.g., high fever, less than non-hospitalized influenza patients because the treatments they are receiving may suppress them. Therefore, nurses need to pay attention to subtle changes in patients' conditions. It was hypothesized that using big data analysis in research would be effective at achieving this goal. This study showed that the developed prediction models had sufficiently good prediction performance.

Vital signs are routinely checked as a fundamental tenet of nursing care (Considine et al., 2016) and are usually recorded in EMR for all inpatients. Vital signs are important for tracking patients' conditions, but they are difficult to use in research (Churpek et al., 2014)

because the most useful trends are not found in simple values and the amount of vital sign data can be relatively large to use in research. This study transformed vital signs to embrace variation and found the significance of vital signs in HAI research.

In practice, nurses can use this study's results to determine which influenza prevention measures to take and detect HAI patients earlier. First, the developed models in this study could be utilized as clinical decision support system in clinical practice. This study used easy-to-collect variables for inpatients as predictors so that the prediction models can be easily applied in practice. General characteristics, comorbidities, vital signs, and room information are standard parts of inpatient EMR. In addition, hematology test results used in this study are generally performed for inpatients and the chest X-ray results includes patients who do not undergo as well. Therefore the developed models in this study could be embedded in EMR systems after adjustment to that EMR data and provide the HAI prediction information to nurses in practice.

Second, this study showed that the distance of beds and ventilation play critical roles in influenza transmission in hospitals. A better understanding of influenza transmission can lead to better infection prevention measures even though the World Health Organization and the Centers for Disease Control and Prevention guidelines emphasize transmission prevention along all possible routes (Xiao et al., 2018). Hospitals should place beds more

than 1.8 m apart to reduce the spread of influenza. However, Article 34 of the Enforcement Rules of the Korea Medical Service Act only requires hospitals built before 2017 to put at least 1 meter between beds and those built after 2017 to put at least 1.5 meters between beds. These requirements need to be changed to require more space between beds to reduce influenza transmission. Furthermore, nurses should keep bedside curtains closed to reduce influenza spreading by droplet routes. Ultimately, rooms should be ventilated more regularly with a special focus on ventilating double rooms and all rooms during the winter.

5. Limitations

This study had four major limitations. The first limitation was that it was a single-center study so its results cannot be generalized. The hospital where it was conducted was a tertiary teaching hospital, so further research should be conducted in other hospital settings. Furthermore, the models developed in this study should be validated. The second limitation was that the dataset was imbalanced with HAI patients accounting for 0.15% of the sample. However, the SMOTE method was applied to compensate for this imbalance. The third limitation was that the EMR used in this study were from a single medical center. However, patients may have visited other medical centers, so the EMR used in this study

may not represent their complete medical data. Thus, in this study, data from selected inpatient visits were analyzed and so did not include influenza vaccine and home medication history. The fourth limitation was that this study did not examine data for healthcare providers, caregivers, or visitors because it was retrospective, although these people can be sources of influenza infection.

VII. Conclusion

This study was conducted to identify the characteristics and risk factors associated with HAI infection and develop HAI infection prediction machine learning models based on EMR. In this study, the differences in HAI and non-HAI patients' general characteristics, comorbidities, vital signs, laboratory findings, radiology findings, and room status were identified. Prediction models were developed using the LR, RF, XGB, and ANN machine learning algorithm, all of which exceeded acceptable performance criteria. Staying in double room contributed the most, and vital signs and laboratory result contributed considerably to prediction model performance.

This study's data mining approach is suitable for analyzing inpatients about whom significant amounts of data are generated. The models developed in this study could be used to support nurses in detecting influenza infection based on patient information that can include subtle changes in their condition. They can serve as the foundation for further research to produce models that can be used in clinical practice. Finally, they can be used to help nurses take better influenza infection prevention measures.

Reference

Adnan, M., Alarood, A. A. S., Uddin, M. I., & ur Rehman, I. (2022). Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. *PeerJ Computer Science, 8*, e803.

Agarwal, D., Schmader, K. E., Kossenkov, A. V., Doyle, S., Kurupati, R., & Ertl, H. C. (2018). Immune response to influenza vaccination in the elderly is altered by chronic medication use. *Immunity & Ageing, 15*(1), 19.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics, 26*(10), 1340-1347.

Alvarez-Lerma, F., Marin-Corral, J., Vila, C., Masclans, J. R., Loeches, I. M., Barbadillo, S., Gonzalez de Molina, F. J., Rodriguez, A., & Group, H. N. G. S. S. (2017). Characteristics of patients with hospital-acquired influenza A (H1N1)pdm09 virus admitted to the intensive care unit. *Journal of Hospital Infection, 95*(2), 200-206. https://doi.org/10.1016/j.jhin.2016.12.017

Alves-Filho, J. C., Spiller, F., & Cunha, F. Q. (2010). Neutrophil paralysis in sepsis. *Shock (Augusta, Ga.), 34*(7), 15-21.

Assinger, A. (2014). Platelets and infection – an emerging role of platelets in viral infection. *Frontiers in Immunology*, *5*, 649. https://doi.org/10.3389/fimmu.2014.00649

Barton, C., Chettipally, U., Zhou, Y., Jiang, Z., Lynn-Palevsky, A., Le, S., Calvert, J., & Das, R. (2019). Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Computers in Biology and Medicine*, *109*, 79-84.

Beck, J. R., & Shultz, E. K. (1986). The use of relative operating characteristic (ROC) curves in test performance evaluation. *Archives of Pathology & Laboratory Medicine*, *110*(1), 13-20.

Bermejo-Martín, J. F., Tamayo, E., Ruiz, G., Andaluz-Ojeda, D., Herrán-Monge, R., Muriel-Bombín, A., Fe Muñoz, M., Heredia-Rodríguez, M., Citores, R., & Gómez-Herreras, J. I. (2014). Circulating neutrophil counts and mortality in septic shock. *Critical Care*, *18*(1), 1-4.

Bischoff, W., Petraglia, M., McLouth, C., Viviano, J., Bischoff, T., & Palavecino, E. (2020). Intermittent occurrence of health care-onset influenza cases in a tertiary care facility during the 2017-2018 flu season. *American Journal of Infection Control*, *48*(1), 112-115. https://doi.org/10.1016/j.ajic.2019.06.020

Bloch, E., Rotem, T., Cohen, J., Singer, P., & Aperstein, Y. (2019). Machine Learning Models for Analysis of Vital Signs Dynamics: A Case for Sepsis Onset Prediction. *Journal of Healthcare Engineering, 2019*, 5930379. https://doi.org/10.1155/2019/5930379

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

CDC. (n.d.). *Influenza (Flu)*. https://www.cdc.gov/flu/symptoms/symptoms.htm

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

Choi, H. S., Kim, M. N., Sung, H., Lee, J. Y., Park, H. Y., Kwak, S. H., Lim, Y. J., Hong, M. J., Kim, S. K., Park, S. Y., Kim, H. J., Kim, K. R., Choi, H. R., Jeong, J. S., & Choi, S. H. (2017). Laboratory-based surveillance of hospital-acquired respiratory virus infection in a tertiary care hospital. *American Journal of Infection Control, 45*(5), e45-e47. https://doi.org/10.1016/j.ajic.2017.01.009

Chow, E. J., & Mermel, L. A. (2017). Hospital-acquired respiratory viral infections: incidence, morbidity, and mortality in pediatric and adult patients. Open Forum Infectious Diseases,

Churpek, M. M., Adhikari, R., & Edelson, D. P. (2016). The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation, 102*, 1-5.

Churpek, M. M., Yuen, T. C., Winslow, C., Robicsek, A. A., Meltzer, D. O., Gibbons, R. D., & Edelson, D. P. (2014). Multicenter development and validation of a risk stratification tool for ward patients. *American Journal of Respiratory and Critical Care Medicine, 190*(6), 649-655. https://doi.org/10.1164/rccm.201406-1022OC

Considine, J., Trotter, C., & Currey, J. (2016). Nurses' documentation of physiological observations in three acute care settings. *Journal of Clinical Nursing, 25*(1-2), 134-143.

Courtney, K. L., Demiris, G., & Alexander, G. L. (2005). Information technology: changing nursing processes at the point-of-care. *Nursing Administration Quarterly, 29*(4), 315-322.

Dai, W., Brisimi, T. S., Adams, W. G., Mela, T., Saligrama, V., & Paschalidis, I. C. (2015). Prediction of hospitalization due to heart diseases by supervised learning methods. *International Journal of Medical Informatics, 84*(3), 189-197.

Delen, D. (2014). Real-world data mining: applied business analytics and decision making. FT Press.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics, 44*(3), 837-845.

Demler, O. V., Pencina, M. J., & D'Agostino Sr, R. B. (2012). Misuse of DeLong test to compare AUCs for nested models. *Statistics in Medicine, 31*(23), 2577-2587.

Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T., & Schneeweiss, S. (2020). Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Network Open, 3*(1), e1918962. https://doi.org/10.1001/jamanetworkopen.2019.18962

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics, 35*(5-6), 352-359.

Enstone, J. E., Myles, P. R., Openshaw, P. J., Gadd, E. M., Lim, W. S., Semple, M. G., Read, R. C., Taylor, B. L., McMenamin, J., Armstrong, C., Bannister, B., Nicholson, K. G., & Nguyen-Van-Tam, J. S. (2011). Nosocomial pandemic (H1N1) 2009, United Kingdom, 2009-2010. *Emerging Infectious Diseases, 17*(4), 592-598. https://doi.org/10.3201/eid1704.101679

Escobar, G. J., LaGuardia, J. C., Turk, B. J., Ragins, A., Kipnis, P., & Draper, D. (2012). Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. *Journal of Hospital Medicine, 7*(5), 388-395.

Falsey, A. R., Hennessey, P. A., Formica, M. A., Cox, C., & Walsh, E. E. (2005). Respiratory syncytial virus infection in elderly and high-risk adults. *New England Journal of Medicine, 352*(17), 1749-1759.

Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling, 217*(1), 48-58. https://doi.org/https://doi.org/10.1016/j.ecolmodel.2008.05.015

Godoy, P., Torner, N., Soldevila, N., Rius, C., Jane, M., Martinez, A., Cayla, J. A., Dominguez, A., & Working Group on the Surveillance of Severe Influenza Hospitalized Cases in, C. (2020). Hospital-acquired influenza infections detected by a surveillance system over six seasons, from 2010/2011 to 2015/2016. *BMC Infectious Diseases, 20*(1), 80. https://doi.org/10.1186/s12879-020-4792-7

Golas, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., Hisamitsu, T., Kojima, G., Felsted, J., Kakarmath, S., Kvedar, J., & Jethwani, K. (2018). A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Medical Informatics and Decision Making, 18*(1), 44. https://doi.org/10.1186/s12911-018-0620-z

Hagel, S., Ludewig, K., Moeser, A., Baier, M., Loffler, B., Schleenvoigt, B., Forstner, C., & Pletz, M. W. (2016). Characteristics and management of patients with influenza in a German hospital during the 2014/2015 influenza season. *Infection, 44*(5), 667-672. https://doi.org/10.1007/s15010-016-0920-0

Hall, C. B. (2001). Respiratory syncytial virus and parainfluenza virus. *New England Journal of Medicine, 344*(25), 1917-1928.

Han, L., Ran, J., Mak, Y. W., Suen, L. K., Lee, P. H., Peiris, J. S. M., & Yang, L. (2019). Smoking and influenza-associated morbidity and mortality: a systematic review and meta-analysis. *Epidemiology, 30*(3), 405-417. https://doi.org/10.1097/EDE.0000000000000984

Han, Q., Wen, X., Wang, L., Han, X., Shen, Y., Cao, J., Peng, Q., Xu, J., Zhao, L., He, J., & Yuan, H. (2020). Role of hematological parameters in the diagnosis of influenza virus infection in patients with respiratory tract infection symptoms. *Journal of Clinical Laboratory Analysis, 34*(5), e23191. https://doi.org/10.1002/jcla.23191

Hey, A. J., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA.

Hottz, E. D., Bozza, F. A., & Bozza, P. T. (2018). Platelets in immune response to virus and immunopathology of viral infections. *Frontiers in Medicine, 5*, 121-121. https://doi.org/10.3389/fmed.2018.00121

Hu, Z., Melton, G. B., Arsoniadis, E. G., Wang, Y., Kwaan, M. R., & Simon, G. J. (2017). Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of Biomedical Informatics, 68*, 112-120. https://doi.org/10.1016/j.jbi.2017.03.009

Huzly, D., Kurz, S., Ebner, W., Dettenkofer, M., & Panning, M. (2015). Characterisation of nosocomial and community-acquired influenza in a large university hospital during two consecutive influenza seasons. *Journal of Clinical Virology, 73*, 47-51. https://doi.org/10.1016/j.jcv.2015.10.016

Inoue, M. (n.d.). *Oversampling with SMOTE with its relative algorithms*. https://github.com/minoue-xx/Oversampling-Imbalanced-Data

Jeong, G. H. (2020). Artificial intelligence, machine learning, and deep learning in women's health nursing. *Korean Journal of Women Health Nursing, 26*(1), 5-9.

Jhung, M. A., D'Mello, T., Perez, A., Aragon, D., Bennett, N. M., Cooper, T., Farley, M. M., Fowler, B., Grube, S. M., Hancock, E. B., Lynfield, R., Morin, C., Reingold, A., Ryan, P., Schaffner, W., Sharangpani, R., Tengelsen, L., Thomas, A., Thurston, D., . . . Chaves, S. S. (2014). Hospital-onset influenza hospitalizations--United States, 2010-2011. *American Journal of Infection Control, 42*(1), 7-11. https://doi.org/10.1016/j.ajic.2013.06.018

Keilman, L. J. (2019). Seasonal Influenza (Flu). *Nursing Clinics of North America, 54*(2), 227-243. https://doi.org/10.1016/j.cnur.2019.02.009

Killingley, B., & Nguyen-Van-Tam, J. (2013). Routes of influenza transmission. *Influenza and Other Respiratory Viruses, 7*, 42-51.

Kim, H., Kim, Y., Kim, K. H., Yeo, C. D., Kim, J. W., & Lee, H. K. (2015). Use of delta neutrophil index for differentiating low-grade community-acquired pneumonia from upper respiratory infection. *Annals of Laboratory Medicine, 35*(6), 647-650.

Kim, J. W., Park, J. H., Kim, D. J., Choi, W. H., Cheong, J. C., & Kim, J. Y. (2017). The delta neutrophil index is a prognostic factor for postoperative mortality in patients with sepsis caused by peritonitis. *PLoS One, 12*(8), e0182325.

Kimberlin, D. W., Brady, M. T., Jackson, M. A., & Long, S. S. (2015). *Red Book (2015): 2015 Report of the Committee on Infectious Diseases* (30th ed.). American Academy of Pediatrics. https://doi.org/10.1542/9781581109276

Kondrich, J., & Rosenthal, M. (2017). Influenza in children. *Current Opinion in Pediatrics, 29*(3), 297-302. https://doi.org/10.1097/MOP.0000000000000495

Korea Healthcare Bigdata Hub. (n.d.-a). *Disease Statstatics*. Health Insurance Review & Assessment Service. http://opendata.hira.or.kr/op/opc/olap3thDsInfo.do

Korea Healthcare Bigdata Hub. (n.d.-b). *Facility and Equipment*. http://opendata.hira.or.kr/op/opc/olapInfraEquipmentStatInfo.do

Kratz, A., Maloum, K., O'Malley, C., Zini, G., Rocco, V., Zelmanovic, D., & Kling, G. (2006). Enumeration of nucleated red blood cells with the ADVIA 2120 Hematology System: an International Multicenter Clinical Trial. *Laboratory Hematology, 12*(2), 63-70.

Lee, J.-H., Song, S., Yoon, S.-Y., Lim, C. S., Song, J.-W., & Kim, H.-S. (2016). Neutrophil to lymphocyte ratio and platelet to lymphocyte ratio as diagnostic markers for pneumonia severity. *British Journal of Biomedical Science, 73*(3), 140-142.

Lee, J. S. (2020). Data Analytics: Modeling Techniques, Data Analysis and Model Building Process by Examples. WIKIBOOKS.

Li, Y., Wang, L.-L., Xie, L.-L., Hou, W.-L., Liu, X.-Y., & Yin, S. (2021). The epidemiological and clinical characteristics of the hospital-acquired influenza infections: A systematic review and meta-analysis. *Medicine, 100*(11), e25142.

Linnen, D. T., Javed, P. S., & D'Alfonso, J. N. (2019). Ripe for disruption? Adopting nurse-led data science and artificial intelligence to predict and reduce hospital-acquired outcomes in the learning health system. *Nursing Administration Quarterly, 43*(3), 246-255. https://doi.org/10.1097/NAQ.0000000000000356

Long, W. J., Griffith, J. L., Selker, H. P., & D'agostino, R. B. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research, 26*(1), 74-97.

Luque-Paz, D., Pronier, C., Bayeh, B., Jouneau, S., Grolhier, C., Le Bot, A., Benezit, F., Thibault, V., & Tattevin, P. (2020). Incidence and characteristics of nosocomial influenza in a country with low vaccine coverage. *Journal of Hospital Infection, 105*(4), 619-624. https://doi.org/10.1016/j.jhin.2020.06.005

Macesic, N., Kotsimbos, T. C., Kelly, P., & Cheng, A. C. (2013). Hospital-acquired influenza in an Australian sentinel surveillance system. *Medical Journal of Australia, 198*(7), 370-372. https://doi.org/10.5694/mja12.11687

Maltezou, H. C. (2008). Nosocomial influenza: new concepts and practice. *Current Opinion in Infectious Diseases, 21*(4), 337-343. https://doi.org/10.1097/QCO.0b013e3283013945

Mao, Y., Chen, Y., Hackmann, G., Chen, M., Lu, C., Kollef, M., & Bailey, T. C. (2011). Medical data mining for early deterioration warning in general hospital wards. 2011 IEEE 11th International Conference on Data Mining Workshops,

McEwen, M., & Pullis, B. C. (2009). *Community-based nursing*. Saunders/Elsevier.

McEwen, M., & Wills, E. M. (2017). *Theoretical basis for nursing*. Lippincott Williams & Wilkins.

Mitchell, R., Taylor, G., McGeer, A., Frenette, C., Suh, K. N., Wong, A., Katz, K., Wilkinson, K., Amihod, B., Gravel, D., & Canadian Nosocomial Infection Surveillance, P. (2013). Understanding the burden of influenza infection among adults in Canadian hospitals: a comparison of the 2009-2010 pandemic season with the prepandemic and postpandemic seasons. *American Journal of Infection Control, 41*(11), 1032-1037. https://doi.org/10.1016/j.ajic.2013.06.008

Müller, A. C. (n.d.). *Cross Validation and Grid Search*. https://amueller.github.io/ml-training-intro/slides/03-cross-validation-grid-search.html#1

Munier-Marion, E., Benet, T., Dananche, C., Soing-Altach, S., Maugat, S., Vaux, S., & Vanhems, P. (2017). Outbreaks of health care-associated influenza-like illness in France: Impact of electronic notification. *American Journal of Infection Control, 45*(11), 1249-1253. https://doi.org/10.1016/j.ajic.2017.05.012

Munier-Marion, E., Benet, T., Regis, C., Lina, B., Morfin, F., & Vanhems, P. (2016). Hospitalization in double-occupancy rooms and the risk of hospital-acquired influenza: a prospective cohort study. *Clinical Microbiology and Infection, 22*(5), 461 e467-469. https://doi.org/10.1016/j.cmi.2016.01.010

Munier-Marion, E., Benet, T., & Vanhems, P. (2017). Definition of healthcare-associated influenza: A review and results from an international survey. *Influenza and Other Respiratory Viruses, 11*(5), 367-371. https://doi.org/10.1111/irv.12460

Murata, Y., & Falsey, A. R. (2007). Respiratory syncytial virus infection in adults. *Antiviral Therapy, 12*(4_part_2), 659-670.

Naudion, P., Lepiller, Q., & Bouiller, K. (2020). Risk factors and clinical characteristics of patients with nosocomial influenza A infection. *Journal of Medical Virology, 92*(8), 1047-1052. https://doi.org/10.1002/jmv.25652

Paes, B. A., Mitchell, I., Banerji, A., Lanctôt, K. L., & Langley, J. M. (2011). A decade of respiratory syncytial virus epidemiology and prophylaxis: translating evidence into everyday clinical practice. *Canadian Respiratory Journal, 18*(2), e10-e19.

Park, J. I. (2016). Developing a Predictive Model for Hospital-Acquired Catheter-Associated Urinary Tract Infections Using Electronic Health Records and Nurse Staffing Data [Thesis (Ph.D.), University of Minnesota].

Park, J. I., Bliss, D. Z., Chi, C. L., Delaney, C. W., & Westra, B. L. (2020). Knowledge discovery with machine learning for hospital-acquired catheter-associated urinary tract infections. *Computers, Informatics, Nursing : CIN, 38*(1), 28-35. https://doi.org/10.1097/CIN.0000000000000562

Parkash, N., Beckingham, W., Andersson, P., Kelly, P., Senanayake, S., & Coatsworth, N. (2019). Hospital-acquired influenza in an Australian tertiary Centre 2017: a surveillance based study. *BMC Pulmonary Medicine, 19*(1), 79. https://doi.org/10.1186/s12890-019-0842-6

Paules, C., & Subbarao, K. (2017). Influenza. *Lancet, 390*(10095), 697-708. https://doi.org/10.1016/s0140-6736(17)30129-0

Penny, K. I., & Chesney, T. (2006). Imputation methods to deal with missing values when data mining trauma injury data. 28th International Conference on Information Technology Interfaces, 2006.,

Raschka, S., & Mirjalili, V. (2017). Python machine learning : machine learning and deep learning with Python, scikit-learn, and TensorFlow (2nd ed.). Packt Publishing.

Redon, J., Coca, A., Lazaro, P., Aguilar, M. D., Cabanas, M., Gil, N., Sanchez-Zamorano, M. A., & Aranda, P. (2010). Factors associated with therapeutic inertia in hypertension: validation of a predictive model. *Journal of Hypertension, 28*(8), 1770-1777. https://doi.org/10.1097/HJH.0b013e32833b4953

Sahni, N., Simon, G., & Arora, R. (2018). Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: a proof-of-concept study. *Journal of General Internal Medicine, 33*(6), 921-928. https://doi.org/10.1007/s11606-018-4316-y

Salgado, C. D., Farr, B. M., Hall, K. K., & Hayden, F. G. (2002). Influenza in the acute hospital setting. *Lancet. Infectious Diseases, 2*(3), 145-155.

Sansone, M., Andersson, M., Gustavsson, L., Andersson, L. M., Norden, R., & Westin, J. (2020). Extensive hospital in-ward clustering revealed by molecular characterization of influenza a virus infection. *Clinical Infectious Diseases, 71*(9), e377-e383. https://doi.org/10.1093/cid/ciaa108

Sansone, M., Wiman, A., Karlberg, M. L., Brytting, M., Bohlin, L., Andersson, L. M., Westin, J., & Norden, R. (2019). Molecular characterization of a nosocomial outbreak of influenza B virus in an acute care hospital setting. *Journal of Hospital Infection, 101*(1), 30-37. https://doi.org/10.1016/j.jhin.2018.06.004

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., & Coopersmith, C. M. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA, 315*(8), 801-810.

Tang, S., Chappell, G. T., Mazzoli, A., Tewari, M., Choi, S. W., & Wiens, J. (2020). Predicting acute graft-versus-host disease using machine learning and longitudinal vital sign data from electronic health records. *JCO Clinical Cancer Informatics, 4*, 128-135.

Taylor, G., Mitchell, R., McGeer, A., Frenette, C., Suh, K. N., Wong, A., Katz, K., Wilkinson, K., Amihod, B., Gravel, D., & Canadian Nosocomial Infection Surveillance, P. (2014). Healthcare-associated influenza in Canadian hospitals from 2006 to 2012. *Infection Control and Hospital Epidemiology, 35*(2), 169-175. https://doi.org/10.1086/674858

Turlapati, V. P. K., & Prusty, M. R. (2020). Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19. *Intelligence-based Medicine, 3*, 100023. https://doi.org/10.1016/j.ibmed.2020.100023

Veenith, T., Sanfilippo, F., Ercole, A., Carter, E., Goldman, N., Bradley, P. G., Gunning, K., & Burnstein, R. M. (2012). Nosocomial H1N1 infection during 2010-2011 pandemic: a retrospective cohort study from a tertiary referral hospital. *Journal of Hospital Infection, 81*(3), 202-205. https://doi.org/10.1016/j.jhin.2012.04.010

Walsh, E. E. (2011). Respiratory syncytial virus infection in adults. *Seminars in Respiratory and Critical Care Medicine, 32*(4), 423-432. https://doi.org/10.1055/s-0031-1283282

Westra, B. L., Sylvia, M., Weinfurter, E. F., Pruinelli, L., Park, J. I., Dodd, D., Keenan, G. M., Senk, P., Richesson, R. L., Baukner, V., Cruz, C., Gao, G., Whittenburg, L., & Delaney, C. W. (2017). Big data science: A literature review of nursing research exemplars. *Nursing Outlook, 65*(5), 549-561.

Wong, B. C., Lee, N., Li, Y., Chan, P. K., Qiu, H., Luo, Z., Lai, R. W., Ngai, K. L., Hui, D. S., & Choi, K. (2010). Possible role of aerosol transmission in a hospital outbreak of influenza. *Clinical Infectious Diseases, 51*(10), 1176-1183.

Wong, S.-C., Lam, G. K.-M., AuYeung, C. H.-Y., Chan, V. W.-M., Wong, N. L.-D., So, S. Y.-C., Chen, J. H.-K., Hung, I. F.-N., Chan, J. F.-W., & Yuen, K.-Y. (2021). Absence of nosocomial influenza and respiratory syncytial virus infection in the coronavirus disease 2019 (COVID-19) era: implication of universal masking in hospitals. *Infection Control & Hospital Epidemiology, 42*(2), 218-221.

World Health Organization. (n.d.). *ICD-10 Version:2019*. https://icd.who.int/browse10/2019/en#/

Xiao, S., Tang, J. W., Hui, D. S., Lei, H., Yu, H., & Li, Y. (2018). Probable transmission routes of the influenza virus in a nosocomial outbreak. *Epidemiology and Infection, 146*(9), 1114-1122. https://doi.org/10.1017/S0950268818001012

Yang, K., Zhang, N., Gao, C., Qin, H., Wang, A., & Song, L. (2020). Risk factors for hospital-acquired influenza A and patient characteristics: a matched case-control study. *BMC Infectious Diseases, 20*(1), 863. https://doi.org/10.1186/s12879-020-05580-9

Abstract in Korean

병원획득 인플루엔자 감염 예측 모델 개발
: EMR 데이터 활용

조 영 희

연세대학교 대학원 간호학과

병원획득 인플루엔자(HAI)는 높은 유병률과 낮은 치료 결과에도 불구하고 그에 대한 인식이 부족하다. 간호사들이 인플루엔자 감염을 조기에 발견하여 병원내 확산을 예방하는 것이 중요하다. 본 연구의 목적은 HAI 와 관련된 특성과 위험 요인을 규명하고 머신 러닝을 이용하여 HAI 감염 예측 모델을 개발하는 것이다.

본 연구는 EMR 데이터를 이용한 후향적 관찰연구로 한국 상급병원에 입원한 111 명의 HAI 환자와 73,748 명의 non-HAI 환자를 대상으로

수행되었다. 일반적 특성, 동반질환, 활력징후, 진단검사 결과, 방사선검사 결과와 병실 정보를 활용하여 t-test 와 카이 제곱 검정을 수행하였으며, 예측 모델 개발을 위해 로지스틱 회귀분석(LR), 랜덤 포레스트(RF), extreme gradient boosting (XGB)와 인공신경망(ANN)을 이용하였다.

HAI 환자는 non-HAI 환자와 다음의 분석변수에서 유의한 차이를 보였다. 일반적 특성으로는 높은 연령, 높은 비율의 면역저하 상태와 Corticosteroid 사용이 있고, 동반 질환으로는 당뇨, 심장 질환, 신장 질환, 혈액 질환, 천식과 만성 패색성 폐질환에서 HAI 환자군이 더 높은 비율을 보였다. HAI 환자들이 체온, 심박수, 수축기 혈압과 이완기 혈압에서 더 큰 변화(variation)를 보였다. 진단검사로는 HAI 환자군에서 적혈구, 헤모글로빈, 혈소판, Hematocrit, Lymphocyte, Na, K, Cl, Calcium, Albumin, 총 빌리루빈은 낮은 결과를, BUN, RDW, Delta neutrophil index, Platelet-to-lymphocyte ratio 에서 높은 결과를 보였다. Chest X-ray 결과에서도 HAI 환자에서 더 높은 비율의 비정상 결과를 나타내었다. 병실 정보로는 더 높은 비율의 HAI 환자가 인플루엔자와 같은 병실 사용, 같은 병동 사용, 그리고 이인실을 사용하였다.

개발한 예측 모델의 성능 평가 결과 모든 모델에서 70% 이상의 AUC (LR: 84.9%, RF: 83.4%, XGB:71.1%, ANN: 76.5%)를 보였고, 위음성 환자수는 RF(5), LR(6), XGB(7), ANN(8) 순으로 적은 결과를 보였다. 이인실의 사용이

예측 모델에 가장 큰 영향을 미치는 요인이었으며 활력 징후와 진단검사 결과가 그 다음으로 영향을 미치는 요인이었다.

본 연구에서 개발한 모든 예측 모델이 수용 가능한 수준 이상의 성능을 보였다. 예측 모델은 간호사들의 HAI 환자를 조기에 발견하고 감염 예방 활동을 하도록 지원하는데 활용 가능할 것으로 기대된다.

Keywords: Influenza, Hospital-acquired influenza, Prediction model, Machine learning, Logistic regression, Random Forest, Extreme gradient boosting, Artificial neural network Double room, Vital sign