Analysis of Pediatric Respiratory Disease Trends Using the 2016 KIDs’ Inpatient Database

Shilpa Balan, California State University, Los Angeles, USA*
https://orcid.org/0000-0002-3582-2560

Shailja Pandit, California State University, Los Angeles, USA
https://orcid.org/0009-0004-4596-4521

ABSTRACT

Many infants and young children worldwide have been affected by chronic respiratory conditions. In this paper, the authors performed an exploratory and predictive analysis of the 2016 KID data set to examine respiratory disease trends among children. They applied the multiple linear regression and random forest regression methods to build a predictive model for the length of stay (LOS) for children with respiratory problems. The tree approach implemented using random forest is found to be a better approach for predicting the length of stay (LOS). In addition, they performed an exploratory analysis of significant fields from the data set. From the analysis, it is found that the winter season has the highest number of inpatient admissions of children having chronic respiratory illnesses. Further, it is found that newborns and infants are more prone to respiratory diseases, with bronchitis being the leading cause of respiratory diseases among children.

KEYWORDS
Bronchitis, KID Database, Length of Stay (LOS), Multiple Linear Regression, Random Forest Regression, Respiratory Disease

INTRODUCTION

Respiratory diseases have adversely had a worldwide impact. According to the World Health Organization (WHO, 2022), it is estimated that 235 million people suffer from asthma. More than one billion people suffer from chronic respiratory conditions (WHO, 2023). Each year, four million people die prematurely from chronic respiratory disease (WHO, 2023).

Infants and young children are particularly susceptible to respiratory diseases. Nine million children under five years of age die annually, and lung diseases are the most common causes of these deaths. Respiratory diseases such as asthma, bronchitis, pneumonia, allergic rhinitis, and sinusitis can impair a child’s ability to perform routine activities, causing missed school days (EPA, 2022).
Pneumonia is a major infectious cause of death in children worldwide (Walker et al., 2013). Asthma is the most common chronic disease, affecting about 14% of children globally with this number rising (Pearce et al., 2007). Factors such as air pollution increase a child’s risk of developing respiratory infections. Moreover, these lead to a large cost burden on patients and their families (Dieleman et al., 2020).

Asthma is a condition characterized by airway inflammation. It occurs in approximately 10% of children (Benton et al., 2022). Bronchiolitis, a lower respiratory tract infection, commonly affects infants and children younger than two years. Respiratory syncytial virus (RSV) is the most common cause. Each year in the United States, an estimated 58,000-80,000 children younger than five years old are hospitalized due to RSV infection (CDC, 2022).

Delayed hospital discharge can lead to increased costs and decreased patient satisfaction (Khechen et al., 2019). Particularly in the pediatric population, prolonged hospitalization is costly and poses a significant risk (Sasaki et al., 2016). In this study, we analyzed the KID data set for respiratory diseases and built a predictive model for the inpatient length of stay. Length of stay (LOS) means the number of days a patient stays in the hospital for treatment from the date of admission to the date of discharge (Kampstra et al., 2018).

Digital Health can have a positive impact on the management of asthmatic patients. According to the World Health Organization (WHO), digital health or eHealth is defined as the cost-effective and secure use of information and communication technologies for health (Tripodi et al., 2020). Digital solutions include utilizing smart devices like wearables and smartphones that will alert the patients to take action whenever attacks are sensed. Some of the medical devices for management of asthma include wearables, mobile applications, and smart inhalers. In addition, artificial intelligence and machine learning are rapidly growing fields with significance in airway management (Matava et al., 2019). The ability of artificial intelligence and machine learning algorithms to recognize patterns from large volumes of complex data makes them valuable for pediatric airway management. In this paper, we reviewed some machine learning algorithms for pediatric respiratory disease management.

In this paper, we attempted to identify key factors contributing to the length of stay (LOS) using machine learning methods; this is also a step toward effective management of medical resources. Prolonged hospital length of stay (LOS) remains one of the most important challenges facing hospitals and has important implications for health outcomes (Panis et al., 2003). Further, we performed an exploratory analysis on the KIDs’ inpatient database in this paper.

BACKGROUND

Diseases of the respiratory system account for approximately 10.6% of emergency department (ED) visits (Ashman et al., 2021). Improving the health of children is a primary goal of healthcare systems (Szilagyi & Schor, 1998). Previous studies provide evidence that exposure to allergens is associated with asthma and other respiratory diseases (Institute of Medicine, 2000).

Other studies suggest that children exposed to smoking have greater sensitivity to respiratory diseases (Ehrlich et al., 1996; Chilmonczyk et al., 1993; Strachan & Cook, 1998). Adult respiratory disease can have a direct negative impact on a child’s health. Children living in households that have adults with COPD are at increased risk of lower respiratory infection such as pneumonia (Zar & Ferkol, 2014). Passive exposure to tobacco smoke increases the risk of pneumonia and asthma and of having more severe respiratory illnesses (CDC, 2022).

Environmental tobacco smoke exposure is a high-risk factor for chronic respiratory illness (Cheraghi & Salvi, 2009). Tobacco use is the leading global cause of preventable death (WHO, 2013). Despite initiatives to reduce tobacco smoking, it is estimated that up to 40% of children are still exposed to tobacco smoke (Öberg et al., 2011) and approximately six million deaths are tobacco related (WHO, 2013).
Other factors also contribute to respiratory diseases. Air pollution exacerbates respiratory conditions such as asthma (Delfino, 2002; Eder et al. 2006). Previous studies show associations between traffic-related exposure and asthma or related symptoms (Gauderman et al. 2005; McConnell et al., 2006; Migliaretti & Cavallo, 2004).

A study by Duan et al. (2022), provides an analysis of medical expenditures for respiratory diseases. Their study analyzed trends in spending changes from 1996 to 2016, associations between these changes, and factors such as population growth, population aging, disease prevalence, healthcare utilization, and service price and intensity. Their findings show that the respiratory conditions with the highest spending in 2016 were asthma and chronic obstructive pulmonary disease.

Machine learning methods have been used frequently for the prediction of patients’ LOS (Barnes et al., 2016; Nouaouri et al., 2015; Taleb et al., 2017). Among the general machine learning based methods, support vector machine, random forest, and particularly decision tree-based methods were demonstrated to be effective in LOS prediction, along with other medical classification applications (Barnes et al., 2016; Nouaouri et al., 2015; Taleb et al., 2017). AI applications can process massive amounts of data and discover hidden patterns in the enormous quantity of medical data (Secinaro et al., 2021). Machine learning can play a major role because it supports evidence-based decision making (Neves et al., 2018).

To the best of our knowledge, there is only one study (Balan et al., 2019) that examined the LOS prediction using the 2016 KID data set. In their work, the authors applied the Random Forest method. For the LOS prediction, the authors applied the fields such as primary and secondary diagnoses and procedures, discharge status, patient demographics, hospital characteristics (e.g., ownership, size, census division), expected payment source, total charges, length of stay, and comorbidity measures.

The aim of this study is to investigate factors that can impact LOS for pediatric patients. In this paper, we limit the scope of our study to only respiratory diseases and examine the Multiple Linear Regression (MLR) and Random Forest (RF) approaches in the methodology section for the LOS prediction for pediatric patients. The methodology section describes the steps we applied for the exploratory and predictive analysis of the KID data set.

RESEARCH METHODOLOGY

In this paper, we utilized the KID's inpatient database (KID) developed for the healthcare cost and utilization project (HCUP, 2019). Figure 1 shows the methodology we used in this study. We analyzed the KID dataset using Python packages and Scikit Library. We applied the Multiple Linear Regression (MLR) and the Random Forest (RF) Regression algorithms to determine the length of stay (LOS) for children with respiratory problems. The Major Diagnostic Category appropriate for the date of discharge (MDC) is assigned by the Medicare DRG grouper during HCUP processing (HCUP, 2008). We filtered the data set on the basis of the ‘MDC in effect on discharge date’ field for the respiratory disease code.

![Figure 1. Research methodology](image-url)
The KID is a pediatric inpatient care database in the United States that includes approximately three million pediatric discharges each year (HCUP, 2019). The KID database was created periodically from 1997 to 2016, usually every three years. The data set we analyzed includes the inpatient pediatric information for the 2016 year. The KID database is based on hospital discharge data, including diagnoses and procedures, and focuses on national estimates of hospital admissions for patients under 21 years of age (HCUP, 2019). The KID database can be used to identify and analyze national trends in health care utilization, access, charges, quality, and health outcomes. The data set had 99 different variables from which to select. The principal variables we selected for this study are age in years at admission, admission month, admission day, number of diagnoses and procedures, patient location, total charges, and primary expected payer or insurance type.

Table 1 describes the categorical fields from the KID data set. The frequency of infants aged 0-1 is the highest among the various categories of kids suffering from respiratory diseases. Although several children were admitted to the hospital, there was only one inpatient death.
Data Pre-Processing

One field we excluded from the study is the indicator of sex. The data set when filtered for respiratory diseases included more male children. The data set had about 60% records containing male children with respiratory diseases, and 40% were those of female children. Thus, we excluded this field from the analysis since trends of admissions would illustrate greater number of male children with the inclusion of this field, which would not provide accurate analysis.

Feature Importance

A list of features having higher impact on the LOS prediction are shown in Figure 2. The most important features influencing the LOS prediction were patient location, number of diagnoses, admission month, age of the children, kid discharge weight, primary expected payer, and total charges. Since total charges would have obvious impacts on the LOS, we did not examine this further.

Heatmap Chart

We created a heatmap chart shown in Figure 3 to check for correlation of the selected data fields of the study with length of stay (LOS). We found that none of the data fields included in this study were correlated with LOS. Total charges indicated some correlation with length of stay, with a correlation value of 0.64. So, we did not further examine total charges.

Algorithm

For the study, we built a prediction model using Random Forest Regression and Multiple Linear Regression as described below. We set the size of the training data to 80% and the test data to 20%. The accuracy of a regression model can be determined by its $R^2$ value. The results from the algorithms are described in the analysis section.

Multiple Linear Regression (MLR)

Multiple linear regression is used to model the relationship between a continuous response variable and continuous or categorical explanatory variables (JMP, 2022). The pseudocode of MLR applied to the KID data set is illustrated in Figure 4. Multiple Linear Regression is represented as follows:
Figure 3. Heatmap chart of the KID data set (for the selected fields)

Figure 4. Pseudocode for MLR applied to the KID data set

Import Python Libraries-pandas, numpy, sklearn, matplotlib
from sklearn.linear_model import LinearRegression
from sklearn import metrics
Read KIDs’ dataset
from sklearn.model_selection import train_test_split
set x_train, x_test
x_train, x_test, y_train, y_test = train_test_split(x, y)
regressor.fit (x_train, y_train)
from sklearn.metrics import mean_absolute_error, mean_squared_error
Calculate R^2
Calculate RMSE
yi=β0+β1xi1+β2xi2+...+βpxip+ϵ

where:

yi=dependent variable
Xi=explanatory variables
β0=y-intercept (constant term)
βp=the model’s error term (also known as the residuals)

Random Forest

A Random Forest (RF) fits several classifying decision trees on various sub-samples of the dataset (Scikit, 2022). The Random Forest model makes predictions by combining decisions from a sequence of base models. Random Forest uses averaging to improve the predictive accuracy and control over-fitting. The pseudocode of RF applied to the KID data set is illustrated in Figure 5.

RESULTS AND ANALYSIS

This section describes the results of the analysis we conducted for the 2016 KID data set. A higher value of R² score indicates good performance with the model fitting the data better. For the study, we found that the model accuracy using Multiple Linear Regression is 66.21% whereas that of the Random Forest is 79.49%. The model accuracy is higher for Random Forest Regression, which is close to 80%. Random Forest Regression methodology is thus a better approach for predicting the length of stay (LOS) for pediatric inpatients suffering from respiratory illness. The value of 66.21% from the Multiple Linear Regression also supports its use for this task. In fact, linear models have the advantage of being easy to understand. The result of the impacts is in line with the literature. The result of previous studies have shown how age is a factor affecting procedures related to different DRGs (Liu et al., 2001). A significantly longer hospital stay was seen to be associated with pediatric surgery (Cheong & Emil, 2014).

From our study, we can infer that LOS is impacted by children’s age, admission month or season, and types of respiratory diseases (bronchitis, asthma, pneumonia, etc.) to name a few top indicators. LOS prediction is important to pediatric facilities because they are usually faced with daily or seasonal overcrowding, which may result in delayed medical treatment and higher risks of medical errors (Ma et al., 2020).

Further, we performed an exploratory analysis of the KID data set. We created the visuals for the analysis using the Tableau software. We examined in detail the variable indicators in the study...
such as the age group of children, the type of respiratory diseases, the different seasons, and metro vs. non-metro location in detail.

Figure 6 depicts the count of hospital admissions of kids in different seasons. From the figure, it can be seen that the number of children admitted in the winter season is more than those admitted in any of the other three seasons, which are fall, spring, and summer, with the summer season having the fewest risks for respiratory diseases. Chilly or cold air causes bronchitis issues.

Bronchiolitis is most often seen in children in the fall and winter (Nationwide Children’s, 2022). The inhalation of cold air has negative effects on the lungs of people with respiratory diseases and specifically on asthma patients, according to D’Amato, Molino, Calabrese, et al. (2018). Bradley (2022) explains that it is less common to get the flu in the summer season. In the fall and winter seasons, people spend more time indoors, enabling respiratory viruses to spread.

Figure 7 illustrates the leading cause of respiratory diseases in children. Bronchiolitis is the leading cause of infant hospitalization in the United States, with nearly 150,000 admissions annually (Pelletier et al., 2006). From Figure 7, it is seen that among various respiratory diseases, bronchitis is the dominant factor. In children, the most common cause of acute bronchitis is a virus (Stanford Medicine, 2021). The illness may develop after a cold or other viral infection in the nose, mouth, or throat. Bronchitis may also be caused by bacteria or dust, allergens, or tobacco smoke.

Digital technologies can provide an opportunity to develop elements to meet the individual patient’s needs, responding to both patient preferences and the clinician’s requirements (Farmer et al., 2017). Digital technologies could improve the precision of respiratory care. While digital interventions are generally acceptable to a wider population (Jeminiwa et al., 2019), specific attention is needed when evaluating them for children. Adolescents are a challenging group to treat, as they are at risk of not attending appointments (Lenney et al., 2018).

Figure 6. Kid admissions for different seasons
According to an article on Healthline (Watson, 2018), asthma and bronchitis have similar symptoms but different causes. In both asthma and bronchitis, the airways become inflamed. Viruses or environmental factors like tobacco smoke and pollution cause bronchitis. Gene changes and environmental triggers like pollen and dust in the air cause asthma. Asthma affects six million children in the United States, according to the Centers for Disease Control and Prevention (CDC, 2018). The proportion of children aged 0 to 17 years reported to currently have asthma increased from 8.7% in 2001 to 9.4% in 2010 which then decreased to 7.0% in 2019 (EPA, 2022). According to Stewart et al. (2018), among those who viewed asthma as a serious threat, electronic monitoring devices were found to be encouraging. These devices helped ensure that patients took their medication. However, many adolescents reported concerns that their healthcare providers did not trust them to take the timely medication prescribed. This exhibits the need to examine digital interventions tailored particularly at children and young people, as their needs and responses to the interventions may be different from the general population.

Figure 7 also shows that pneumonia is one of the top causes for hospital admission of children. As per Stanford Medicine (2021), pneumonia is most often caused by bacteria or viruses and is most common in children younger than five years old.

Figure 8 shows that newborns and infants are more prone to respiratory diseases. This is because infants are prone to respiratory infections (Peltola, 1993). Their inherent immunity diminishes within months after birth. Infants are grouped as less than two years old. Toddlers are grouped as less than four years old, and preschoolers are grouped from four to six years old. Respiratory conditions are the most common reason for admission to a neonatal unit for infants (Pramanik et al., 2015).
Figure 9 demonstrates that the number of children with a higher number of respiratory diseases are in the metropolitan area. Micropolitan communities are defined as mid-sized rural communities. They are in non-metro places with populations from 2,500-49,999 people (Iowa PRC, 2022).

In the KID data set, a small metropolitan area is defined as having fewer than one million residents, whereas a large metropolitan area has at least one million residents. In Figure 9, we included both small and large metropolitan areas together under the category of ‘Metropolitan’. Rates of childhood
asthma in the United States have been increasing (Maziak, 2005). Urbanization may increase asthma rates and symptoms due to higher levels of air pollutants (Hendryx et al., 2012). Previous studies have found lower asthma rates among farm children compared to urban children. As stated in a paper by Elliott et al. (2017), allergic and asthmatic patients benefit from telemedicine. For example, patient-doctor collaboration, easy access to doctor consultation, as well as simplified prescription procedures can be possible through telemedicine. This can positively impact patients, especially those living in rural or remote areas (Elliott et al., 2017).

Figure 10 indicates that most of the kid hospital admissions use Medicaid insurance. A pattern of children admissions was found dependent on different private insurances. Other insurances include workers’ compensation and other miscellaneous insurance types. Previous studies have shown that publicly insured children had a significantly longer hospital stay than privately insured children (Pati et al., 2012).

CONCLUSION

This study uses the pediatric inpatient KID database to analyze and create a predictive model for the patients’ LOS. The Multiple Linear Regression and Random Forest methods were applied in this study. The tree approach used by Random Forest gave a higher R² value of about 80%, indicating that the decision tree method of Random Forest is a good approach for LOS prediction. The proposed LOS prediction method in this study can be implemented to improve the decision making of bed management for pediatric patients.

The most important fields for the LOS prediction of pediatric patients are found to be age, season, location (metro vs non metro), and respiratory disease type (with the highest being bronchitis). From our analysis, they found that a greater number of inpatient hospitalizations of children suffering from respiratory diseases occurred in the winter season. Bronchitis, asthma, and pneumonia are the leading causes of respiratory diseases in children. Further, we found that newborns and infants are
more prone to respiratory diseases. Metropolitan areas have a high percentage of children struggling with respiratory diseases.

Fields such as external causes and kid discharge weight from the KID data set can be examined further. In the future, other data fields can be included in the analysis, such as parameters associated with specific hospitals (e.g., hospital services, patient satisfaction, and quality of patient care). To develop a prediction model for a specific hospital, these fields could be added into the model. Machine learning is the keystone in healthcare delivery.

The field of pediatric health has been pushed to embrace the use of digital technology and artificial intelligence. This looks optimistic due to the advantages of digital technology that enhance the standards of healthcare. With the expansions in digital health, many benefits and improvements have been achieved. However, there are still many socio-economic disparities including the digital divide (Evans & Eisenstein, 2021). This divide includes access to high-speed broadband internet and the equipment needed to access the services. The issue of digital divide can be examined further in future research.

**FUNDING STATEMENT**

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**STATEMENT OF INTEREST**

The authors of this publication declare that they have no competitive interests.
REFERENCES


**Shilpa Balan** is an Associate Professor at the College of Business and Economics at California State University, Los Angeles. Dr. Balan’s broad research interests are in Business Intelligence, Data Mining, and Health Analytics.

**Shailja Pandit** obtained her Master’s degree in Information Systems from the College of Business and Economics at California State University, Los Angeles. Her research interests include Business Intelligence, Information Systems, and Data Analytics.