A Data Mining Framework for the Analysis of Patient Arrivals into Healthcare Centers

Salam Abdallah, Mohsin Malik, Gurdal Ertek

A Data Mining Framework for the Analysis of Patient Arrivals into Healthcare Centers

Salam Abdallah, Mohsin Malik, Gurdal Ertek

A Data Mining Framework for the Analysis of Patient Arrivals into Healthcare Centers

Salam Abdallah
Abu Dhabi University
College of Business,
PO Box 1790, Abu Dhabi, U.A.E.
+971 2501 5710
salam.abdallah@adu.ac.ae

Mohsin Malik
The University of Melbourne
Melbourne, Victoria 3010
mohsin.malik@unimelb.edu.au

Gurdal Ertek
Abu Dhabi University
College of Business,
PO Box 1790, Abu Dhabi, U.A.E.
+971 3709 0714
gurdal.ertek@adu.ac.ae

ABSTRACT
We present a data mining framework that can be applied for analyzing patient arrivals into healthcare centers. The sequentially applied methods are association mining, text cloud analysis, Pareto analysis, cross-tabular analysis, and regression analysis. We applied our framework using real-world data from a one of the largest public hospitals in the U.A.E., demonstrating its applicability and possible benefits. The dataset used was eventually 110,608 rows in total for the regression models, covering the most utilized 14 hospital units. The dataset is at least 10-fold larger than datasets used in closely-related research. The developed data mining framework can provide the input for a subsequent optimization model, which can be used to optimally assign appointments for patients, based on their arrival patterns.

CCS Concepts
• Health informatics • Data mining • Information systems • Association rules • Healthcare information systems.

Keywords
Data Mining; Health Informatics; Healthcare information systems; Patient Arrival Patterns.

1. INTRODUCTION

1.1 Demand for Healthcare
The world's population is growing and aging [1]. The immediate consequence of this demographic trend is an increase in the demand for healthcare services throughout the world, accompanied with considerable increases in costs. The general trend of increased demand for healthcare services is also valid for almost every region and country, and especially pronounced for regions with higher population growth [1]. Specifically, in countries throughout the world, healthcare spending is expected to increase 2.4-7.5% per year until 2020 [2].

This paper presents a study motivated by a real-world project and conducted in the United Arab Emirates (U.A.E.), using data from one of the largest public hospitals in the country. The Middle East and North Africa (MENA) region is estimated to have a shortage of 150,000 physicians, 326,000 dentists, and 1.8 million nurses and midwifery personnel by 2020 [3]. The increasing demand for skilled healthcare professionals is resulting in higher staff costs for healthcare providers. The United Arab Emirates (U.A.E.) healthcare market is expected to show an annual average growth of 12.7% until 2020 [4]. The demand may grow even further due to medical tourism, which is an economic priority for U.A.E.'s tourism sector [3].

1.2 Healthcare Operations
Recent reports cite "improving operational efficiencies" as one of the measures that can be adopted by healthcare providers throughout the world [2] and in the U.A.E. [5], to cope with increased competition and rising staff costs. Various analytical tools, including data mining and optimization, and adoption of information technologies, can greatly contribute to achieving such operational efficiencies. This opportunity is readily recognized by the healthcare industry, resulting in an expected growth of 15.9% per year in the healthcare information technology (IT) market until 2021 [6].

Healthcare providers (hospitals, clinics) typically follow a service process where, as the first step, the patient contacts the provider to schedule an appointment. The alternative triggering of the service process is through the arrival of walk-in patients, who directly show up at the healthcare center without prior notice. Once the patient is admitted into the system, vitals are checked and the doctor appointment and the other steps of the process follow.

There are opportunities for improving operational efficiencies in every step of the service process. In this study, the aspect of healthcare operations we are interested in is the patient arrivals. Specifically, we are interested in quantifying and analyzing the timing of patient arrivals, with the ultimate goal of improving the patient admission process step, through elimination of wasted time.

1.3 Performance Measure: Lateness
The main performance measure in our study is "Lateness", calculated as the difference between the arrival time of the patient and the appointment time in the system. If a patient arrives at the exact minute as of the appointment, then the Lateness value is equal to 0. If the patient arrives later than the appointment time, then the Lateness value is positive. Conversely, if patient arrives earlier than the appointment time then the Lateness value is negative.

1.4 Contributions
In this study, we present a data mining framework, consisting of six steps that follow data cleaning, that can be adopted for analyzing patient arrival data. We applied our framework using real-world data from a large public hospital in the U.A.E., demonstrating its applicability and possible benefits. The dataset we used was eventually 110,608 rows in total for the regression models of the most active 14 hospital units. This is at least 10-fold larger than in comparison to datasets used in closely-related research.

While the data mining study presented in this paper is unique in the literature, it is valuable especially if it can be applied to improve patient appointment assignment and eliminate waste of time. To

this end, we also present in this paper, in Section 7, how the results of this study can be used in a subsequent optimization model for direct improvement in operational efficiency.

2. LITERATURE

There has been considerable amount of research in patient flow modeling, where the patient arrival distribution and service time distribution are used to compute waiting times in the various steps of the process steps [7]. In contrast to this developed line of research, our study focuses solely on Lateness and how Lateness is associated with various factors. In our study, analyzing waiting times was not possible because our dataset did not include any attributes pertaining to the resources (doctors, nurses, rooms, etc.).

Another common line of research related to our study is the analysis of the patient-related factors behind delays/no-shows in seeking care [8][9][10]. For example, in the mentioned line of research, survey data posterior to patient arrivals can be merged with patient data in the information systems, including time of initial contact and arrival, to discover why patients ended up seeking out for treatment much later in time than they should have.

Other relevant research can be summarized as follows:

- [11] applies lean principles, specifically root cause analysis, to identify sources of operational inefficiency. Other lean-focused studies show that operational inefficiencies can be significantly eliminated by very simple acts, such as making telephone reminders [12].
- [13] analyzes punctuality of arriving patients and derives statistical distributions that characterize lateness. [13] assumes that the population of patients is uniform. We, on the contrary, accept that Lateness is very much dependent on the attributes of each patient, and focus on coming up with a predictive regression model for characterizing this dependency relation.
- [14] predicts arrival time and no shows in outpatient clinics, as we do. However, [14] employs a very different set of independent variables than we do.
- [15] applies association mining to predict no-shows (patients not showing up at their appointments) and conduct set covering optimization to reduce the vast number of rules to a manageable size. While [15] applies data mining, specifically association mining, and optimization together, the objective is not to optimize appointment assignments.
- [16] develops an appointment scheduling algorithm. Our study and the optimization model we propose (as future work), on the other hand, assume that Lateness is dependent on the scheduled time, rather than being independent of it.
- [17] predicts the duration of an appointment, and identifies late arrival of the surgeon as the most important factor. We, on the other hand, do not consider any resource-related factors, and predict Lateness using other variables.

To summarize, in our study, we develop predictive regression models for characterizing Lateness of patients for each hospital unit. Our study focuses solely on Lateness and how Lateness is associated with various factors. In our study, we could not analyze waiting times because our dataset did not include any attributes pertaining to the resources (doctors, nurses, rooms, etc.). Furthermore, we do not consider any resource-related factors in regression modeling, due to same reason. Our study and the optimization model we propose (as future work) assume that Lateness is dependent on the scheduled day and time, rather than being independent of it. This last contribution, namely the proposition of an optimization model where assigned appointment day and time are decision variables, is unique in the literature.

3. METHODOLOGY

The methods applied in this paper all fall under the general field of "data mining". Data mining, increasingly being referred to as "data science" (with subtle differences in between), is the rapidly growing field of computer science and informatics that aims at discovering new and useful information and knowledge from data [18][19][20][21].

Data mining is akin to a toolbox: Just as a toolbox contains a multitude of tools suitable for various tasks, data mining consists of a multitude of analytical methods (and algorithms), where each method or combination of methods are most suitable for a given data with unique characteristics. The selection of these methods and the particular order in which these methods could be applied, are a result of experience in the field, as well as empirical experience with the specific data at hand (through both an exhaustive application of various techniques, as well as several trial-error cycles).

The data analyzed in this study was a structured tabular database (consisting of rows and columns) recorded the log of patients arriving to a major public hospital in the U.A.E. The applied methods were association mining, text cloud analysis, Pareto analysis, cross-tabular analysis, and regression analysis, in sequence. These methods and the tools we used in applying the methods are described in this section. The developed data mining framework can provide the input for an optimization model, which can be used to optimally assign appointments to patients, based on their attributes as explained in Section 7. Therefore, even though not applied in this study, optimization is also introduced and discussed in this section.

3.1 Association Mining

Association mining is a data mining method for identifying associations between elements (items) of a set (set of items), based on how these elements appear in multiple subsets (transactions) of the set [22][23][24]. Association mining takes as input a transaction data, where each transaction contains a subset of items, and all item subsets coming from the same superset.

Association mining gives as output the list of itemsets that appear together frequently in transactions (frequent itemsets), and the rules that describe how these associations affect each other (association rules). An association rule is a rule in the form "IF [Antecedent A] THEN [Consequent B]" (or simply as "A⇒B).

There are many metrics related with an association rule, and the most popular metrics are support and confidence. Support of an itemset (e.g.: {A,B}) or a rule (e.g. A⇒B) is the percentage of transactions that the items in the itemset or the rule appear in. Confidence is defined for association rules (and not for frequent itemsets). Confidence of a rule A⇒B is the conditional probability of item B appearing in a transaction, given that item A readily appears in that transaction.

The standard (even though not fastest) algorithm for association mining is the Apriori algorithm, which was first introduced by [25] and has been used extensively since. Association mining, conducted through apriori algorithm or another alternative algorithm, is a standard function in almost every data mining platform (SAS, RapidMiner, WEKA). Association mining can also be conducted through specialized software [22][26].

In our research, we used the ARuleGUI [1] , which conducts computations using the Apriori[2] software library [26][27][28][29], to carry out association mining.

3.2 Text Cloud Analysis

A text cloud is a visualization of frequent textual terms in a document, where the size of each term reflects its frequency of appearance in the document [30]. While text cloud is typically applied for text documents in natural language, such as news, online messages, emails, etc., the visualization is flexible enough to analyze any type of text. To this end, in our research, we used text cloud visualization (using the Wordle.net[3] online service) for analyzing the association rules obtained from association mining.

3.3 Pareto Analysis

Pareto principle is a basic principle in business management, which states that a majority of effects are due to a minority of factors. While the origins of the principle is economics (first suggested by Italian economist Vilfredo Pareto in 1896 [31]), the principle is almost globally applicable in any applied field of knowledge [32][33]. Given that only a small percentage of factors are responsible for a major percentage of effects, one can prioritize identifying these most influential factors and analyzing their effects, rather than analyzing the complete system.

In our study, Pareto analysis was essential to reduce the number of values (for categorical values) that we would include in our detailed analysis. We conducted Pareto analysis using MS Excel's Pivot Table functionality.

3.4 Cross-Tabular Analysis

Cross-tabulation (also referred to as contingency tables or cross tabs) is a quantitative method for analyzing the relations between multiple variables of interest [34]. While cross-tabular analysis is typically done when one or more of the variables is categorical, it can also be conducted by discretizing numerical variables and converting them into categorical variables. In our research, we conducted cross-tabular analysis using MS Excel's Pivot Table functionality.

3.5 Regression Analysis

Regression analysis is a fundamental technique for estimating the relation between one or more independent variables and a dependent variable, whose values are assumed to be determined by the independent variable(s) [34]. In our study, for each hospital unit, the best linear regression model and the regression function (including confounding effects) was obtained through extensive search using genetic algorithm (GA).

The R statistical language and system[4] and RStudio[5], an open-source integrated development environment (IDE) for R, were utilized for conducting regression analysis. The *glmulti* R package [35] was utilized for automatically conducting regression analysis and systematically trying out different models. Specifically, the genetic algorithm (GA) built into *glmulti* package[6] was used for extensively searching for the best model for each hospital unit.

3.6 Optimization

Optimization is the selection of best element from among a set of alternatives, based on one or more objectives to be optimized (maximized or minimized). Linear programming is an optimization method, where there is a single linear objective function to be optimized under a set of linear constraints [38]. Linear programming models with only integer variables are referred to as integer programming models, and those only with binary (0/1) variables are referred to as binary programming models. Models that encompasses different combinations of variable types (continuous, integer, binary) are referred to as mixed-integer programming models.

Until this point in our research, we did not use optimization, however, the results obtained here can be used within an optimization model in later stages of the research, as described in Section 7.

4. FRAMEWORK

The developed data mining framework consist of the following steps:

Step 0. Data Cleaning

Perform data cleaning

Step 1. Association Mining

1.1. Construct histograms to identify the values at which to discretize Lateness values.

1.2. Discretize hour of the day (Hour) variable and other variables if necessary.

Conduct association mining to obtain association rules.

Filter the rules with Lateness in the consequent.

Step 2. Text Cloud Analysis

2.1. Visualize the antecedents of the filtered association rules, for both very early and very late patients using text cloud visualization.

2.2. From the text cloud, identify which values of which variables are most frequently observed for very early and very late patients.

Step 3. Pareto Analysis

3.1. Perform Pareto analysis to identify the values for each attribute that account for the overwhelming majority of observations.

3.1. Filter out only the rows having those values, for subsequent analysis

Step 4. Cross-Tabular Analysis

4.1. Conduct cross-tabular analysis to observe how average and standard deviation of Lateness varies with respect to different values of each variable.

4.2. Perform statistical hypothesis (ex: t-test, Mann-Whitney test) testing if necessary.

4.3. Use these observations to identify which variables and which variable values to focus on.

[1] http://www.borgelt.net/argui.html

[2] http://www.borgelt.net/apriori.html

[3] http://www.wordle.net/

[4] https://www.r-project.org/

[5] https://www.rstudio.com/

[6] https://cran.r-project.org/web/packages/glmulti/index.html

Step 5. Regression Analysis

5.1. Construct a multi-linear regression model for each hospital unit.

5.2. Search for the best regression model and its parameter set through a search algorithm (ex: through a genetic algorithm, GA), with a limit on the number of iterations or computational time.

5.3. Use the regression results to understand which factors are most influential in Lateness

Step 6. Optimization

6.1. Construct an optimization model, where day of the week (DayOfWeek) and hour of the day (Hour) are decision variables, and the objective is to assign each patient a day and hour such that the average waiting time over all patients is minimized.

6.2. Schedule patients on a rolling-horizon of up to one week, by using the optimization model.

5. CASE STUDY

5.1 Data Collection

The data used in the study comes from a large public hospital, in the United Arab Emirates (U.A.E.). The data includes date and time for appointment check in, vitals, patient service start time, as well as information about the hospital unit, nationality and appointment and insurance information of the patients. The original data consisted of 168,360 rows and 23 columns. The data fully covers a time interval of approximately six months, which is a large and representative sample.

5.2 Ethical Considerations

The study was conducted with complete respect regarding the privacy of patients. The participating hospital provided the data without any identifiable information about the patients. The gender information was not provided by the hospital. Personal information regarding nationality, health plan, and insurance company were immediately anonymized by the authors at the beginning of the study. The data was shared with the authors by the hospital under a strict and legally-binding non-disclosure agreement, and resided only on the authors' computers at all time. The research consultant who worked on the project also signed a non-disclosure agreement and was not provided with the data. Instead, he assisted with the data cleaning and analysis through secure remote connection, and only under the presence of one of the authors.

5.3 Data Cleaning

As in every data mining study, data cleaning and anonymization were carried out before any analysis.

1. For each time and date column, where both date and time were given as a single value of string type, multiple columns were created as numeric variables, for year, month, day, hour, and minute.

2. The appointment week within the year and the appointment day within the week (where 1 refers to Sunday, 2 refers to Monday, etc., consistent with the week days and weekends in the U.A.E.) were derived using existing columns. This step required using MS Excel functions for manipulating text and converting data types, and due to the large number of rows, was conducted through creating a separate file for each such date and time column in the original data. Later in the study, following some trials, these columns were excluded from the analysis.

3. For hospital unit, nationality, the appointment being a new one or a follow up, the health plan category, health plan type, specific health plan code, and the name of the insurance company were all anonymized using MS Excel's lookup functions. The anonymization for each column was again conducted on a separate file due to the large data size for a laptop.

4. Lateness was calculated for each patient.

5. Rows where the lateness was more than 3 hours or less than -3 hours (at least 3 hours early) were eliminated.

6. All cells in the spreadsheet were reduced to "Values Only" for reducing the file size and speed in human computer interaction (HCI).

6. ANALYSIS

6.1 Association Mining

The first conducted analysis was association mining, where the objective was to identify the variables and variables' values which are most associated with extreme arrival behaviors, namely arriving too early or too late.

In order to come up with cut-off values and conduct such an analysis, it was necessary to first observe the statistical distribution of lateness. The histograms were constructed for some of the hospital units (Figures 1 and 2 for units U23 and U43) and were observed to highly resemble the normal distribution. Formal statistical hypothesis testing regarding the goodness of fit of the normal distribution was not conducted due to very large number of observations, which would certainly result in rejection of the fit. Eventually, the cut-off values were selected as -60 and +60 minutes and Lateness<-60 and Lateness>60 were identified as values of interest.

For conducting association mining analysis, the hour of the day (Hour) and Lateness were discretized.

Association mining analysis was conducted to obtain association rules in the form "IF Antecedent THEN Consequent", using minimum support of 0.1% and minimum confidence of 40%, and with at most 4 items in each rule. Once the association rules were generated, only the rules with a value of E_60_INF (early at least 1 hour) and L_60_INF (late at least 1 hour) for the hour of the day in the consequent were filtered out, so that we could identify which attribute values are associated with very late arrivals or very early arrivals.

6.2 Text Cloud Analysis

The words in the antecedents of the association rules were visualized through text visualization as word clouds, and the attribute values which are most associated with very late or early (more than 60 minutes) arrivals were visually observed. Figures 3 and 4 display the attribute values associated with E_60_INF (early at least 1 hour) and L_60_INF (late at least 1 hour), respectively.

Figure 3, which shows associations with very early arrivals, suggest that very early arrivals do not take place frequently from 12:00 until 15:00 but are observed at other times. Patients arriving to hospital units U13 are frequently observed to come very early. Units U43 and U05 come after U13 with respect to such behavior. With respect to healthcare plan, patients with healthcare plan category HPC94, healthcare plan type HPT9, healthcare plan HP165 are observed to come very early much more frequently compare to patients with other healthcare plan. With respect to insurance supplier, patients who has healthcare plan under insurance

company C7 come very early the most frequently. Patients with nationality N31 come very early, before other nationalities.

Figure 4, which shows associations with very late arrivals, suggest that very late arrivals take place frequently from 06:00 until 09:00, and especially during week days. With respect to healthcare plan, patients with healthcare plan category HPC66, healthcare plan type HPT2, healthcare plan HP133 are observed to come very late much more frequently compare to patients with other healthcare plan. With respect to insurance supplier, patients who has healthcare plan under insurance company C2 come very early the most frequently. In contrast to the result from Figure 3, patients with nationality N31 do not come very late frequently.

6.3 Pareto Analysis

The next descriptive analysis would be cross-tabular analysis (pivot table analysis in MS Excel). However, before conducting cross-tabular analysis, Pareto analysis was conducted to identify the significant few values for each attribute. This way, cross-tabular analysis would include the few most significant attribute values. For example, out of the 46 hospital units, 17 of them were observed to account for more than 90% of all the rows. Out of 110 different nationalities, 9 of them were observed to account for more than 92% of all the rows. In the Pareto analysis, threshold values of around 90% were determined for each variable. Row corresponding to week days account for more than 97% of all rows, thus only week days were considered in further analysis. Similarly, only the rows with attribute values that account for the majority were included and the rows that contain the less significant attribute values for any of the attributes were eliminated.

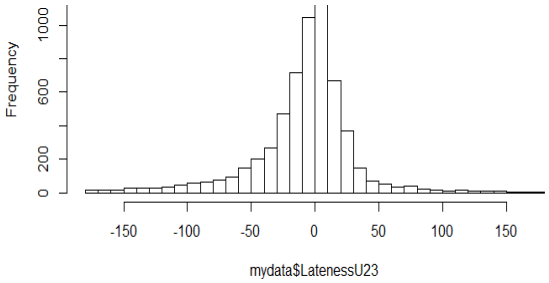

Figure 1. Histogram of **Lateness** for hospital unit **U23**.

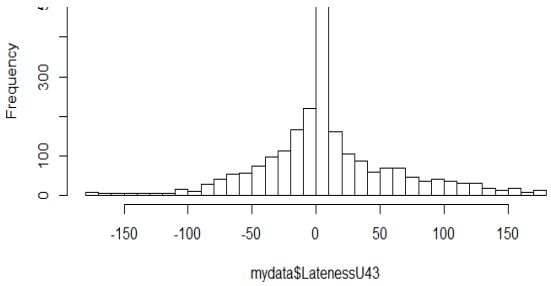

Figure 2. Histogram of **Lateness** for hospital unit **U43**.

6.4 Cross-Tabular Analysis

Cross-tabular analysis was conducted to analyze the effects of different values of each attribute on the selected performance measure namely lateness. This was conducted only for the significant values of the attributes coming from the earlier Pareto Analysis. After removing walk-in patients, pivot tables were constructed to calculate the average and the standard deviation of lateness for each value of each attribute. The results, shown in Figures 5 and 6, suggest that average lateness differs considerably among different hospital units and health plan categories. Lateness is color-coded such that darker green values correspond to earliest arrivals and darker red values correspond to latest arrivals.

From Figure 5, it is observed that average lateness for Unit U23 is -7.65 minutes (meaning that the patients arrive 7.65 minutes early on average, whereas average lateness for unit U43 is 9.53 minutes. Furthermore, there are considerable differences in average lateness values with respect to hours of the day.

From Figure 6, considerable differences are observed in average lateness values, especially when health plan categories HPC3 (earlier arrivals) and HPC8 (later arrivals) are compared. The hours of the day again seem to be influential: For HPC3, time interval 09:00-12:00 has average Lateness of 3.66 minutes, whereas time interval 12:00-15:00 has average Lateness of -3.60 (earliness of 3.60 minutes).

6.5 Regression Analysis

The predictive model applied in this study was regression analysis with categorical factors (independent variables) and the numerical response (dependent variable) of Lateness. Before conducting regression analysis, any row with any missing value for any of the attributes was eliminated. Also, only the rows with Pareto-significant values of the attributes were considered. Eventually only 110,608 rows corresponding to 14 hospital units were included in the regression analysis.

Regression analysis was conducted separately for each hospital unit, where the same set of factors and response were included. The included factors were Hour, DayOfWeek, Nationality, NewOrFollowUp, HealthPlanCategory, HealthPlan, HealthPlanType, InsuranceCompany. The response in each regression model was Lateness, which could take any value between -180 and 180 corresponding to a maximum of 3 hours earliness and 3 hours lateness, respectively. For each hospital unit, the best linear regression model (including confounding effects) was found through exhaustive search through a genetic algorithm (GA). The R command corresponding to the regression was as follows:

```
results <- glmulti(Lateness ~ Hour + DayOfWeek + Nationality + NewOrFollowUp + HealthPlanCategory + HealthPlan + HealthPlanType + InsuranceCompany, data=mydata, level=2, fitfunction=glm, confsetsize=100, method="g", conseq=5)
```

The best regression model with the optimal subsets of the factors and confounding effects were found to be considerably different for different hospital units. The regression analysis results are summarized in Table 1, illustrating which factors appear in which of the 14 models (corresponding to the 14 hospital units).

The number of models where each variable appears in the “best” regression equation is also given, in the last row of Table 1. From here, the factors (independent variables) which appear in most models are observed to be Hour, HealthPlanCategory, HealthPlan, and HealthPlanType, well before others. Other variables that

appear frequently include DayOfWeek, InsuranceCompany, and the confounded variable HealthPlanType:HealthPlanCategory. These results suggest that data regarding these attributes should be collected and included while predicting Lateness, regardless of which hospital unit is being analyzed.

As an example of regression results, Table 2 displays the best model identified, through genetic algorithm (GA) search, for predicting Lateness at hospital unit U23. It can be observed that both the variables and the confounded effects can play significant role in predicting Lateness.

7. CONCLUSIONS AND FUTURE WORK

Our case study has shown that our data mining framework (Section 4) can reveal various hidden patterns and can yield insights into patient arrival patterns. Our framework also can identify the significant factors (independent variables) that affect Lateness, as well as the form and the parameters of the "best" regression function (obtained after a given limited number of search iterations).

While these contributions are mostly novel in the literature (especially for the particular set of factors employed), the most significant benefit of our study is that it can serve as an engine for determining the parameters of a subsequent optimization model, which can optimize appointment day and time. In this section, we will outline such a model.

The regression model we discussed considers Lateness as a function of various factors, including hour of the day (Hour) and the day of the week (DayOfWeek) in which the patient arrived. None of the considered factors, except Hour and DayOfWeek, can be controlled by the healthcare provider, as these factors depend on the attributes or the medical condition of the patient. However, the healthcare provider *can* minimize the Lateness of the patient by selecting the "best" DayOfWeek-Hour combination. This is the main idea behind the proposed optimization model, which is described next:

Let

$\mathcal{P}$: set of patients who will visit the selected hospital unit, p= $1 \cdots P$

$\mathcal{D}$: set of days (DayOfWeek) $d = 1 \cdots 5$

$\mathcal{T}$: set of time periods (Hour) $t = 1 \cdots 5$

be the *sets* of the optimization model.

Next, let

$C_{d,t}$: capacity of the hospital unit for time period (Hour) t on day (DayOfWeek) d

be the *parameters* of the optimization model.

Let the *decision variables* of the optimization model be

$$Z_{p,d,t} = \begin{cases} 1 & \text{if patient } p \text{ is scheduled to time period } t \text{ on day } d \\ 0 & \text{o/w} \end{cases}$$

The *optimization model* designed to *minimize average Lateness*, is as follows:

$$\min_{p,d,t} \Lambda = \frac{1}{P} \sum_{p=1}^{P} \hat{L}\left(Z_{p,d,t}\right)$$

$$\sum_p Z_{p,d,t} = C_{d,t} \qquad \forall (d,t)$$

$$\sum_{d,t} Z_{p,d,t} = 1 \qquad \forall p$$

$$Z_{p,d,t} \quad \text{binary}$$

This model is to be constructed and solved independently for each hospital unit (it is assumed that no relation exists between independent hospital units, which may not be realistic in certain cases). The decision to be made is the assignment of a patient to a time interval for appointment. The appointment will be within a rolling horizon of a given number of days (the number of days is a modeling decision that needs to be done beforehand). The decision variable $Z_{p,d,t}$ takes the value of 1 if patient p is scheduled to time period t on day d, and 0 otherwise.

The objective is to minimize Λ, which is defined as the average expected lateness of the patients who will be scheduled in available set of time intervals in the upcoming days (up to seven days ahead). The expected lateness of a patient, given the time interval and day s/he is assigned to, is estimated using the fitted regression function $\hat{L}\left(Z_{p,d,t}\right)$ for that hospital unit.

The first constraint specifies that, for a given hospital unit, the number of scheduled patients for a specific time interval on a specific day can not exceed the capacity (patients that can be served) during that time interval. The second constraint specifies that each patient should be assigned to exactly one time interval within the given set of days and time intervals.

When this model is populated with the parameters coming from regression analysis, it can compute the optimal DayOfWeek-Hour combination that each patient should be assigned to. While we are referring to this decision as "optimal", it is only "near-optimal", because the parameters are only estimates under a stochastic setting. Yet still, this is a model that has not been proposed in the literature earlier and can serve greatly in improving operational efficiency in the patient arrival process.

Besides developing and testing the described optimization model, our current work can also be extended by considering additional factors in predictive modeling. These factors can include city of residence, presence of pain, marital status, living arrangement, and social support [10][39].

8. ACKNOWLEDGMENTS

This study was supported by Abu Dhabi Education Council (ADEC) under ADEC Award for Research Excellence (AARE 2015). The authors thank the public hospital in the United Arab Emirates (U.A.E.) for sharing their data for the research. The authors also thank Ayaz Salman for his help in the data cleaning and analysis process, as well as assisting in the writing of the paper.

9. REFERENCES

[1] World Health Organization, 2015. Ageing and health. Fact Sheet No 404. https://goo.gl/XD4UoT

[2] Deloitte, 2017. 2017 Global healthcare sector outlook. https://goo.gl/3KLJt6

[3] Al Masah Capital, 2014. Al Masah Capital: MENA Healthcare Sector. https://goo.gl/sfZMDF

[4] Gulf News, 2016. U.A.E. Health care Market to hit $19.5b by 2020. February 16, 2016. Babu Das Augustine. https://goo.gl/eXC1N7

[5] Deloitte, 2016, U.A.E. healthcare industry. https://goo.gl/4ynWKB

[6] MarketsAndMarkets, 2017. Healthcare IT Market by Product (EHR, RIS, PACS, VNA, CPOE, HIE, Telehealth, Healthcare Analytics, Population Health Management, Supply Chain Management, CRM, Fraud Management, Claims Management) End User (Provider, Payer) - Global Forecast to 2021. April 2017. https://goo.gl/2Z5rjV

[7] Bhattacharjee, P. and Ray, P.K., 2014. Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections. *Computers & Industrial Engineering,* 78, pp.299-312.

[8] Darawad, M.W., Alfasfos, N., Saleh, Z., Saleh, A.M. and Hamdan-Mansour, A., 2016. Predictors of delay in seeking treatment by Jordanian patients with acute coronary syndrome. *International Emergency Nursing*, 26, pp.20-25.

[9] Faiz, K.W., Sundseth, A., Thommessen, B. and Rønning, O.M., 2013. Prehospital delay in acute stroke and TIA. *Emerg Med J,* 30(8), pp.669-674.

[10] El-Din, M.M.N., Al-Shakhs, F.N. and Al-Oudah, S.S., 2008. Missed appointments at a university hospital in eastern Saudi Arabia: magnitude and association factors. *The Journal of the Egyptian Public Health Association,* 83(5-6), pp.415-433.

[11] Niveditha, M.S., 2015. Re-engineering the outpatient process flow of a multi-specialty hospital. https://goo.gl/cozXiW

[12] Reti, S., 2003. Improving outpatient department efficiency: a randomized controlled trial comparing hospital and general-practice telephone reminders. *The New Zealand Medical Journal (Online),* 116(1175).

[13] Tai, G. and Williams, P., 2009. A novel characterization of patient arrival process in healthcare facilities. https://goo.gl/uz4eEh

[14] Kawczynski, L. and Taisch, M., 2009, September. Health Care Provider Processes Analysis. In *IFIP International Conference on Advances in Production Management Systems* (pp. 595-602). Springer, Berlin, Heidelberg.

[15] Glowacka, K.J., Henry, R.M. and May, J.H., 2009. A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational Research Society,* 60(8), pp.1056-1068.

[16] Denton, B. and Gupta, D., 2003. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions,* 35(11), pp.1003-1016.

[17] Strahl, J., 2015. Patient appointment scheduling system: with supervised learning prediction. https://goo.gl/X6tXXy

[18] Han, J., Kamber, M. and Pei, J. 2011. *Data Mining: Concepts and Techniques.* Morgan Kaufmann; 3rd edition.

[19] Ertek, G., Chi, X. and Zhang, A.N., 2017. A framework for mining RFID data from schedule-based systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 47(11), pp. 2967-2984.

[20] Asian, S., Ertek, G., Haksoz, C., Pakter, S. and Ulun, S., 2017. Wind turbine accidents: A data mining study. *IEEE Systems Journal,* 11(3), pp.1567-1578.

[21] Ertek, G., Tokdemir, G., Sevinç, M. and Tunç, M.M., 2017. New knowledge in strategic management through visually mining semantic networks. *Information Systems Frontiers,* 19(1), pp.165-185.

[22] Çinicioğlu, E.N., Ertek, G., Demirer, D. and Yörük, H.E., 2011. A framework for automated association mining over multiple databases. In *2011 International Symposium on Innovations in Intelligent Systems and Applications (INISTA),* (pp. 79-85). IEEE.

[23] Ertek, G., Demiriz, A. and Cakmak, F., 2012. Linking behavioral patterns to personal attributes through data re-mining. In *Behavior Computing* (pp. 197-214). Springer, London.

[24] Ertek, G. and Tunc, M.M., 2012. Re-mining association mining results through visualization, data envelopment analysis, and decision trees. *Computational Intelligence Systems in Industrial Engineering,* pp.601-622.

[25] Agrawal, R., Imelinski, T. and Swami, A.N., 1993. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, (Eds), *Proceedings of the ACM SIGMOD International Conference on Management of Data.* (Washington, D.C., USA, May 25 - 28 1993). SIGMOD '93. ACM, New York, NY, 207-216.

[26] Borgelt, C. and Kruse, R., 2002. Induction of association rules: Apriori implementation. In *COMPSTAT* (pp. 395-400). Physica, Heidelberg.

[27] Borgelt, C., 2003. Efficient implementations of apriori and eclat. In *FIMI'03: Proceedings of the IEEE ICDM workshop on Frequent Itemset Mining Implementations*.

[28] Borgelt, C., 2004, November. Recursion Pruning for the Apriori Algorithm. In *FIMI* (Vol. 126).

[29] Borgelt, C., 2012. Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 2(6), pp.437-456.

[30] Hu, M., Wongsuphasawat, K. and Stasko, J., 2017. Visualizing social media content with sententree. *IEEE Transactions on Visualization and Computer Graphics,* 23(1), pp.621-630.

[31] Pareto, V., 1964. *Cours d'économie politique* (Vol. 1). Librairie Droz.

[32] Ertek, G., 2009. Visual Data Mining for Developing Competitive Strategies in Higher Education, in *Data Mining for Business Applications.* Editors: Longbing Cao, Philip S. Yu, Chengqi Zhang and Huaifeng Zhang. Springer.

[33] Bugalho, M.N., Dias, F.S., Briñas, B. and Cerdeira, J.O., 2016. Using the high conservation value forest concept and Pareto optimization to identify areas maximizing biodiversity and ecosystem services in cork oak landscapes. *Agroforestry systems*, 90(1), pp.35-44.

[34] Anderson, D.R., Sweeney, D.J. and Williams, T.A., 2011. *Essentials of Modern Business Statistics with Microsoft Excel*. Cengage Learning (page 67, pages 620-631).

[35] Calcagno, V., and de Mazancourt, C., 2010. glmulti: an R package for easy automated model selection with (generalized) linear models. *Journal of Statistical Software* 34.12 (2010): 1-29.

[36] Winston, W.L. 2004. *Operations Research: Applications and Algorithms* (4th edition). Cengage.

[37] DiMatteo, M.R., 2004. Social support and patient adherence to medical treatment: a meta-analysis. https://goo.gl/aGWk5A

Figure 3. Attribute values when the patients are **early** at least 60 minutes, with a confidence of at least 40%.

Figure 4. Attribute values when the patients are **late** at least 60 minutes, with a confidence of at least 40%.

Units	Hour_06_09	Hour_09_12	Hour_12_15	Hour_15_18	Hour_18_21	Hour_21_24	Average for Lateness
U8	8.66	-5.32	-5.23	-12.68		-2.71	**-3.94**
U9	6.68	-8.96	-6.63	-16.75		-0.53	**-5.99**
U14	15.11	-1.43	-1.90	-12.81	-19.62	-1.60	**-0.95**
U19	14.28	-5.42	-8.05	-17.70	-5.60	-2.57	**-3.95**
U20	6.78	-9.04	-6.54	-11.89		0.88	**-5.61**
U21	7.06	-5.90	-5.00	-13.32	-31.00	-1.55	**-4.83**
U23	8.18	-7.20	-7.40	-31.42		0.09	**-7.65**
U25	26.15	-5.20	-19.45	-35.91		-13.51	**-7.27**
U26	8.94	-2.76	-3.79	-17.66		-0.74	**-2.02**
U27	5.52	-6.16	-7.23	-13.10	-14.70	-2.14	**-6.03**
U29	10.94	-0.14	-1.80	-8.97	-4.33	1.34	**0.13**
U31	17.20	12.03	6.06	1.34		1.24	**8.26**
U41	23.98	0.48	-1.10	-12.29		6.50	**4.40**
U43	12.99	-5.93	-9.75	11.24	35.24	-10.71	**9.53**
Average for Lateness	**11.69**	**-2.85**	**-4.06**	**-11.06**	**16.50**	**-1.04**	**-1.67**

Figure 5. Cross-tabular analysis of **Lateness** with respect to **hospital units**.

HealthPlanCategory	Hour_06_09	Hour_09_12	Hour_12_15	Hour_15_18	Hour_18_21	Hour_21_24	Average for Lateness
HPC3	22.48	3.66	-3.60	-13.15		-14.90	**-2.18**
HPC35	12.03	-6.72	1.94	-8.33	7.44	3.31	**-0.56**
HPC8	10.44	3.73	5.80	-8.18	61.00	11.75	**1.71**
HPC94	11.62	-2.92	-5.13	-11.42	16.37	-1.05	**-1.80**
Average for Lateness	**11.69**	**-2.85**	**-4.06**	**-11.06**	**16.50**	**-1.04**	**-1.67**

Figure 6. Cross-tabular analysis of **Lateness** with respect to **health plan category**.

Table 1. The factors appearing in the best model for each hospital unit and the appearance frequency of each factor. 1 denotes that the factor is present in that model.

Unit	Intercept	Hour	DayOfWeek	NewOrFollowUp	HealthPlanCategory	HealthPlanType	InsuranceCompany	HealthPlan	Nationality	HealthPlanType:Hour	HealthPlanType:NewOrFollowUp	InsuranceCompany:HealthPlanCategory	HealthPlan:HealthPlanCategory	HealthPlanType:HealthPlanCategory	InsuranceCompany:HealthPlan	NewOrFollowUp:DayOfWeek	InsuranceCompany:HealthPlanType	DayOfWeek:Hour	HealthPlanType:DayOfWeek	HealthPlanType:Nationality	NewOrFollowUp:Hour	HealthPlanCategory:NewOrFollowUp
U8	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0
U9	1	1	0	1	1	0	1	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0
U14	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	0	0
U19	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	0	0	0	0	0
U20	1	1	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
U21	1	1	0	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0
U23	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	0	0	1	1	0	0	0
U25	1	1	1	1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0
U26	1	1	0	0	1	1	1	1	1	0	0	1	0	1	1	0	0	0	0	0	0	0
U27	1	1	1	1	1	1	1	1	0	1	0	1	0	0	1	0	1	0	0	0	1	1
U29	1	1	1	0	1	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0
U31	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	0	1	0
U41	1	1	1	0	1	1	1	1	0	0	0	0	0	1	1	0	1	0	0	0	0	0
U43	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	0	1	0	0	1	0
$\sum$	**14**	**14**	**11**	**9**	**14**	**13**	**10**	**14**	**3**	**7**	**3**	**8**	**1**	**10**	**9**	**3**	**3**	**2**	**1**	**1**	**3**	**1**

Table 2. The best regression model obtained for Hospital Unit U23.

Number of GA Generations: 830

Lateness ~ 1 + Hour + DayOfWeek + NewOrFollowUp + HealthPlanCategory + HealthPlanType + InsuranceCompany + DayOfWeek:Hour + HealthPlanType:DayOfWeek + HealthPlanType:NewOrFollowUp + HealthPlanType:HealthPlanCategory + InsuranceCompany:HealthPlanCategory

	Estimate	Std. Error	df
(Intercept)	3.3791639	5.117549	6119
HourHour_09_12	-15.9692562	3.504481	6119
HourHour_12_15	-14.430737	3.490735	6119
HourHour_15_18	-41.4217847	3.902486	6119
HourHour_21_24	-7.0373843	4.235951	6119
DayOfWeekSaturday	2.2369168	6.748343	6119
DayOfWeekSunday	-2.8651697	4.691285	6119
DayOfWeekThursday	-1.5603118	5.196154	6119
DayOfWeekTuesday	0.2219676	4.377265	6119
DayOfWeekWednesday	0.1872806	4.282621	6119
NewOrFollowUpN	4.3659576	1.066314	6119
HealthPlanCategoryHPC8	-11.948427	37.658373	6119
HealthPlanCategoryHPC94	4.0305095	4.101377	6119
HourHour_09_12:DayOfWeekSaturday	-2.2177596	7.680761	6119
HourHour_12_15:DayOfWeekSaturday	-24.023832	8.547821	6119
HourHour_15_18:DayOfWeekSaturday	-40.6290609	12.287082	6119
HourHour_21_24:DayOfWeekSaturday	-17.4552306	9.518936	6119
HourHour_09_12:DayOfWeekSunday	4.4908297	5.319504	6119
HourHour_12_15:DayOfWeekSunday	-3.5734008	5.36087	6119
HourHour_15_18:DayOfWeekSunday	2.1058863	6.210354	6119
HourHour_21_24:DayOfWeekSunday	-1.5541954	6.606346	6119
HourHour_09_12:DayOfWeekThursday	0.4066116	5.854857	6119
HourHour_12_15:DayOfWeekThursday	4.5757485	6.064936	6119
HourHour_15_18:DayOfWeekThursday	7.6470221	7.153453	6119
HourHour_21_24:DayOfWeekThursday	3.3355738	7.568839	6119
HourHour_09_12:DayOfWeekTuesday	-1.054235	4.965162	6119
HourHour_12_15:DayOfWeekTuesday	2.0431524	5.120432	6119
HourHour_15_18:DayOfWeekTuesday	5.7022637	6.220606	6119
HourHour_21_24:DayOfWeekTuesday	1.7907581	6.368967	6119
HourHour_09_12:DayOfWeekWednesday	1.1728756	4.847133	6119
HourHour_12_15:DayOfWeekWednesday	-3.2089929	5.065909	6119
HourHour_15_18:DayOfWeekWednesday	3.0012861	5.827958	6119
HourHour_21_24:DayOfWeekWednesday	-1.7185086	6.429327	6119