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

Salam Abdallah

جامعة أبوظبي
ABU DHABI UNIVERSITY

Mohsin Malik

THE UNIVERSITY OF MELBOURNE

Gürdal Ertek

جامعة أبوظبي
ABU DHABI UNIVERSITY

Outline

- **Demand for Healthcare**
- **Healthcare Operations**
- **Performance Measure: Lateness**
- **Contributions**
- **Literature**
- **Research Gap ｜ Topic | Question | Methods**
- **Developed Framework**
- **Results**
- **Conclusions**
- **Optimization Model**
- **Acknowledgement**

Demand for Healthcare

- **Growing world population**
- **Growing demand for healthcare** services
- Healthcare **spending to increase**
  - 2.4-7.5% per year until 2020 in various countries
- Increase in costs

- **Real-world project**
- Conducted in the **United Arab Emirates (U.A.E.)**
- Data from **large public hospital**

Demand for Healthcare

- **Middle East and North Africa (MENA)** region
- **Shortage** in MENA by 2020:
  - 150,000 physicians
  - 326,000 dentists
  - 1.8 million nurses and midwifery personnel
- **U.A.E. healthcare** market to **grow 12.7% per year** until 2020.
- May grow even further due to **medical tourism**
  - an economic priority for U.A.E.'s tourism sector.
- Increasing demand resulting in **higher staff costs.**

http://www.un.org/sustainabledevelopment/sustainable-development-goals/
http://bit.ly/1Kjkn0B

Healthcare Operations

- **“Improving operational efficiencies”**
  - to cope with increased competition and rising staff costs.

- Adoption of **information technologies (IT)**
- **Analytical tools**
  - data mining
  - optimization
- Can greatly contribute to achieving such operational efficiencies.

- **Growth of 15.9% per year in healthcare IT** market until 2021.

Healthcare Operations

- **First step** in the service process
  - Patient contacting the provider to **schedule an appointment** and **arriving** around the scheduled time.
    OR
  - **Arrival of walk-in patients**, who directly show up.
- Patient admitted into the system
- Vitals checked
- Doctor appointment
- Other steps

- **This study:**
  - **Patient arrivals.**
  - Quantifying and analyzing the timing of patient arrivals.
- Ultimate **goal:**
  - **Improving the patient admission process step.**
  - Through elimination of wasted time.

Performance Measure: Lateness

- **Lateness = (Arrival Time) – (Appointment Time)**
- Main performance measure in our study.

- Arrival at exact appointment time:
  - **Lateness = 0**
- Arrival later than appointment time:
  - **Lateness > 0**
- Arrival earlier than appointment time:
  - **Lateness < 0**

Contributions

- A **data mining framework** for **analyzing patient arrival data.**

- Unique data mining study
- **Real world application:**
  - applicability
  - possible benefits.

- **Data size:**
  - 110,608 rows for 14 hospital units.
  - **At least 10-fold larger** compared to similar research.

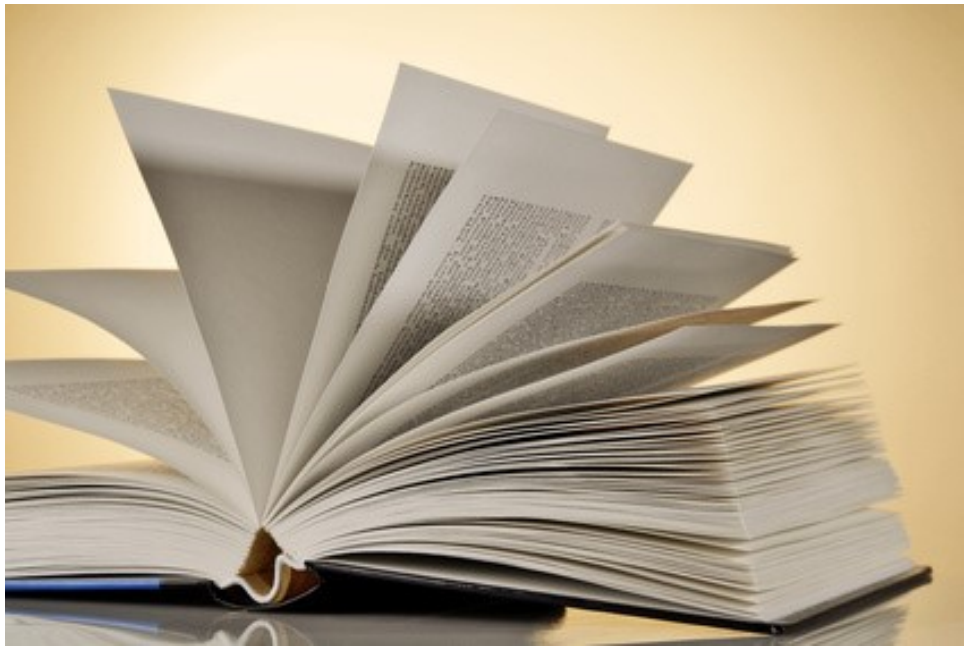
- We also present in this paper, **a subsequent optimization model for**
  - **scheduling appointments** for
  - direct improvement in operational efficiency.

Literature (1 of 3)

- **Patient flow modeling [7]**
  - Patient arrival and service time distributions are used **to compute waiting times.**

- **Our focus**
  - **Lateness**
  - **Association of Lateness with various factors.**

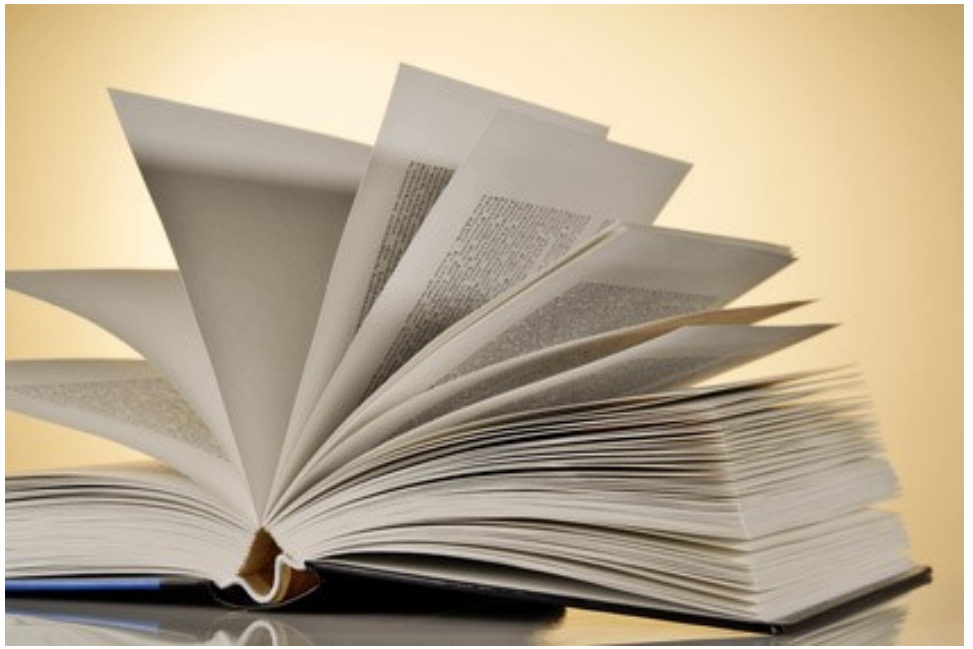
- **Our dataset**
  - No attributes pertaining to the resources (number of doctors, nurses, rooms, etc. over time).



Literature (2 of 3)

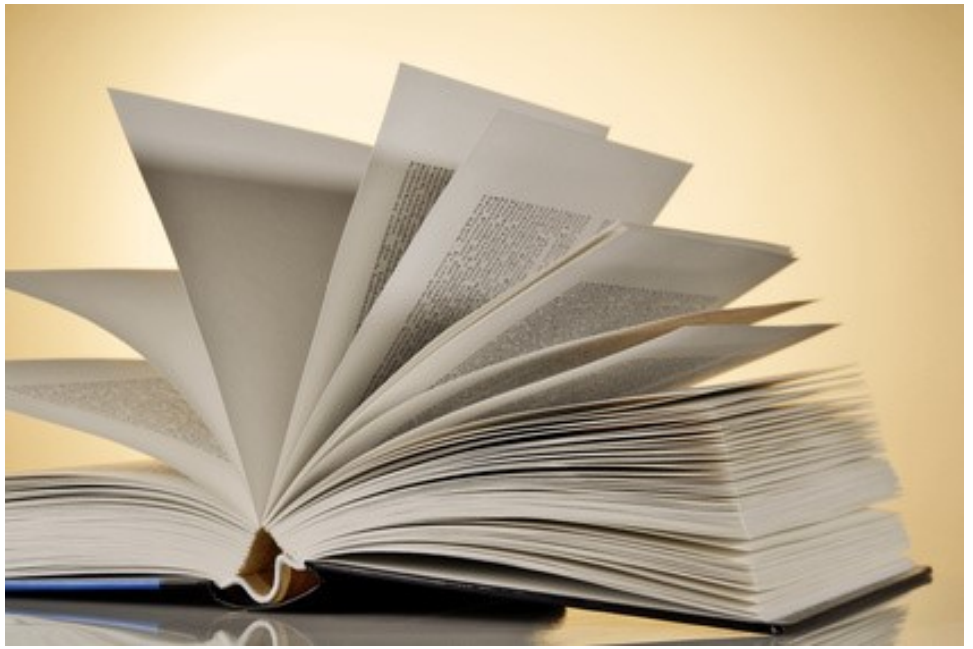
- Analysis of **patient-related factors behind delays/no-shows** in seeking care **[8][9][10].**

- **[11] applies lean principles**,
  - specifically root cause analysis,
  - to identify sources of operational inefficiency.
- **[13] 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.



Literature (3 of 3)

- **[15] applies association mining to predict no-shows**
  - [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.



Research Gap

- **No academic studies** on **predicting Lateness as a function of schedule day and time.**

- **Significant gap of knowledge and insights** with regards to the **understanding of Lateness and the factors underlying Lateness** in healthcare centres.

Research Topic

- **<u>Predicting Lateness</u> as a function of schedule day and time** (as well as other factors)**.**

- **Developing an optimization model**
  - for **scheduling appointments,**
  - to **minimize average lateness.**

Research Questions

1. **How can** the **patient arrivals** into healtcare centres **be analyzed** to come up with insights into Lateness?

2. **Which factors** are **associated with Lateness**?

3. **How can patient appoinments be scheduled** to minimize Lateness?

Research Methods (1 of 6)

- **Data Mining**
  - growing field of computer science and informatics
  - aims at **discovering new and useful information and knowledge** from data.
  - multitude of analytical methods (and algorithms)
  - each method or combination of methods are most suitable for a given data with unique characteristics.

Research Methods (2 of 6)

- **Association Mining**
  - 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.
  - gives as output
    - list of **itemsets appearing together frequently** in transactions (frequent itemsets), and
    - **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"**).

Research Methods (3 of 6)

- **Text Cloud Analysis**
  - **visualization of frequent textual terms** in a document, where
  - the **size of each term reflects its frequency** of appearance in the document

- **Our Study**
  - we used text cloud visualization
  - **for analyzing the association rules** obtained from association mining.
  - using **Wordle.net online service**.

Research Methods (4 of 6)

- **Pareto Analysis**
  - "Majority of effects are due to a minority of factors."
  - Suggested by Italian economist Vilfredo Pareto in 1896
  - Globally applicable in any applied field of knowledge.
  - One can prioritize identifying these most influential factors, rather than all factors.

- **Cross-Tabular Analysis**
  - Contingency tables or cross tabs
  - Quantitative method for analyzing the relations between multiple variables of interest
  - Using MS Excel's Pivot Table functionality.

Research Methods (5 of 6)

- **Regression Analysis**
  - Technique for estimating **the relation between one or more independent variables and a dependent variable**.
  - Our study, **for each hospital unit,**
    - **best linear regression model** and the regression function (including confounding effects)
    - obtained through extensive **search using genetic algorithm (GA).**

Research Methods (6 of 6)

- **Optimization**
  - Selection of best element from among a set of alternatives, based on one or more objectives to be optimized (maximized or minimized).
- **Linear programming**
  - an optimization method
  - single linear objective function to be optimized under a set of linear constraints.

Developed Framework

- **Step 0.** Data Cleaning
- **Step 1.** Association Mining
- **Step 2.** Text Cloud Analysis
- **Step 3.** Pareto Analysis
- **Step 4.** Cross-Tabular Analysis
- **Step 5.** Regression Analysis
- **Step 6.** Optimization

Results: **Association Mining**

- Coming up with cut-off values
- Cut-off values were selected as -60 and +60 minutes
- **Lateness < -60** and **Lateness > 60** were identified as values of interest.

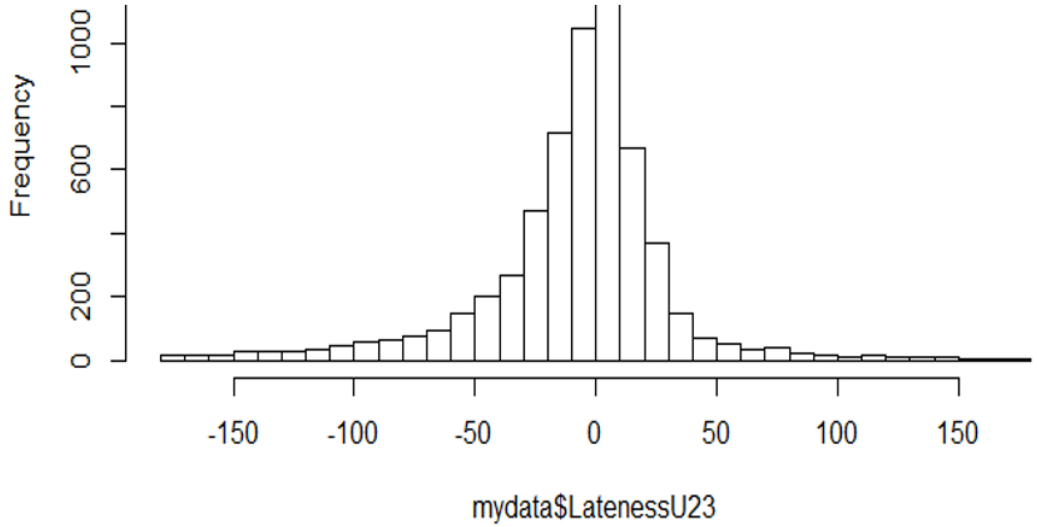


Figure 1. Histogram of **Lateness** for hospital unit **U23**.

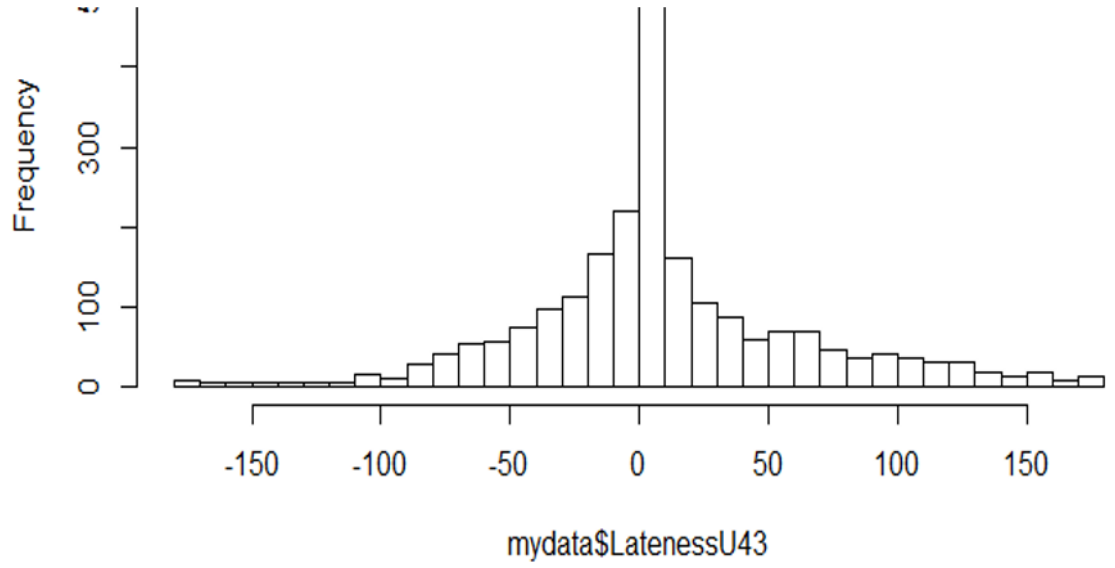


Figure 2. Histogram of **Lateness** for hospital unit **U43**

Results: Association Mining

- **“IF Antecedent THEN Consequent”**
  - minimum support of 0.1%
  - minimum confidence of 40%
  - at most 4 items in each rule

- Rules with a value of
  - **E_60_INF** (early at least 1 hour) and
  - **L_60_INF** (late at least 1 hour
- in the consequent were filtered out
- Goal: identify which attribute values are associated with
  - **very early arrivals** or
  - **very late arrivals.**

Results: Text Cloud Analysis

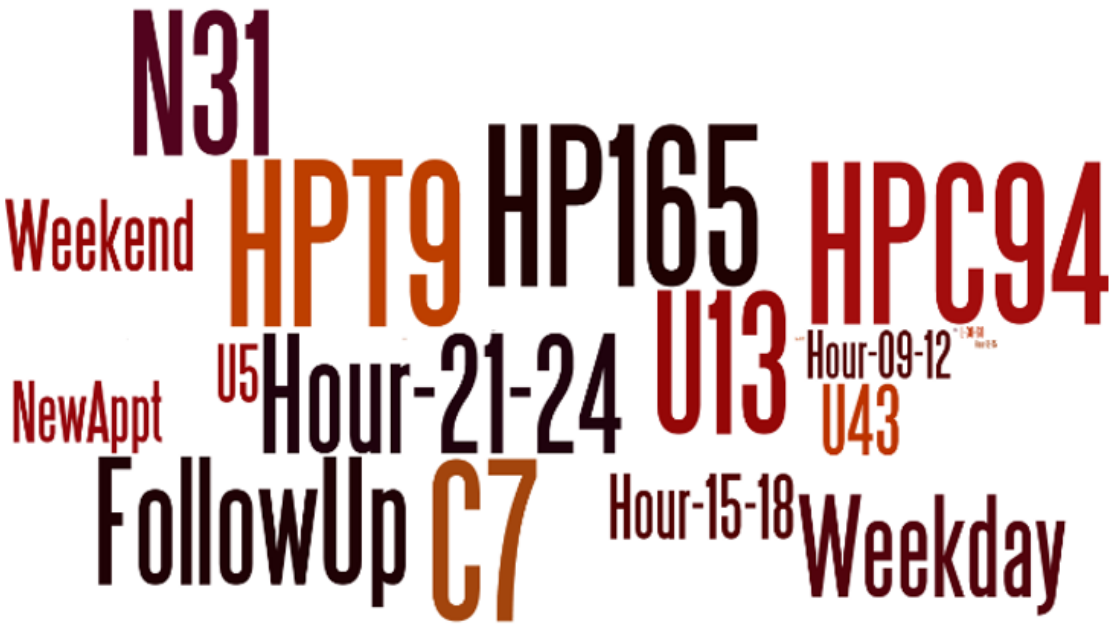


Figure 3. Attribute values when the patients are **early** at least 60 minutes, with a confidence of at least 40%.

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

Lateness < -60

Lateness > 60

Results: Pareto Analysis

- Conducted ***before*** **cross-tabular analysis.**
  - Identify the significant few values for each attribute.
  - Cross-tabular analysis will include **the few most significant attribute values.**

- Pareto analysis **threshold values ~ 90%**

- Out of the 46 hospital units, 17 account for >90% of rows.
- Out of 110 different nationalities, 9 account for > 92% of rows.
- Week days account for > 97% of rows
  - Only week days were considered in further analysis.

Results: **Cross-Tabular Analysis**

- Average lateness differs considerably
  - among **health plan categories**
  - among **different 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**.

Results: **Cross-Tabular Analysis**

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**.

Results: **Regression Analysis**

- **R** statistical language and system
- **RStudio** , an open-source integrated development environment (IDE) for R
- **glmulti** R package for automatically conducting regression analysis and systematically trying out different models.
  - genetic algorithm (GA) built into glmulti package
  - extensively searching for the best model for each hospital unit.

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
∑□	**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

• • •

Conclusions

- Data mining framework:
  - **reveals hidden patterns.**
  - **generates insights** into patient arrival patterns.
  - **identifies significant factors** (independent variables) that affect Lateness
  - **determines**
    - form and
    - parameters of the

  “best” regression function

Conclusions

- Our framework can **serve as an engine**
  - for **determining the parameters of** a subsequent **optimization** model,
  - which can **optimize appointment day and time.**

- Regression model
  - considers **Lateness as a function of various factors,**
  - including **Hour** and **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.
- **Healthcare provider can minimize the Lateness of the patient by selecting the "best" DayOfWeek-Hour combination.**

Optimization Model

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

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

Optimization Model

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

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

fitted regression function for that hospital unit

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

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

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

Research Questions

1. **How can** the **patient arrivals** into healtcare centres **be analyzed** to come up with insights into Lateness?

2. **Which factors** are **associated with Lateness**?

3. **How can patient appoinments be scheduled** to minimize Lateness?

Ertek Projects – Web Projects

https://ertekprojects.com

ERTEK PROJECTS

HOME ABOUT US PROJECTS (ENGLISH) PROJELER (TÜRKÇE) CONTACT US

Projects (English)

Dr. Gurdal Ertek's Publications

Research on Wind Turbine Accidents

Online Education

Music

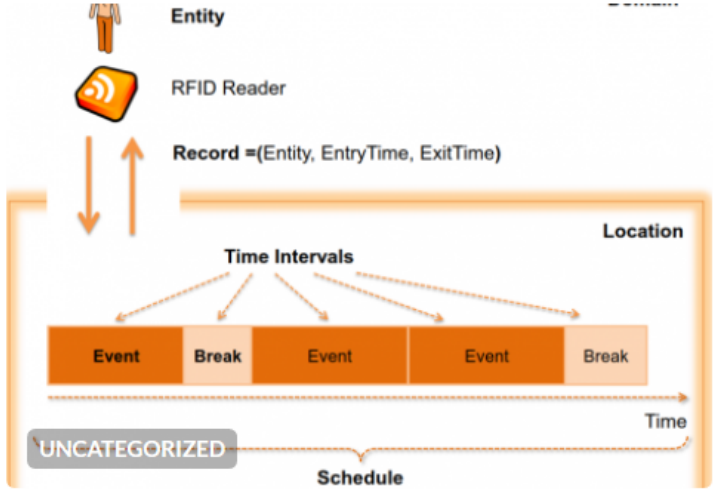
https://ertekprojects.com/gurdal-ertek-publications/

ERTEK PROJECTS

HOME ABOUT DR. GÜRDAL ERTEK DOWNLOAD

Dr. Gurdal Ertek's Publications

Learning and Personal Attributes of University Students in Predicting and Classifying the Learning Styles:



A Framework for Mining RFID Data From Schedule-Based Systems

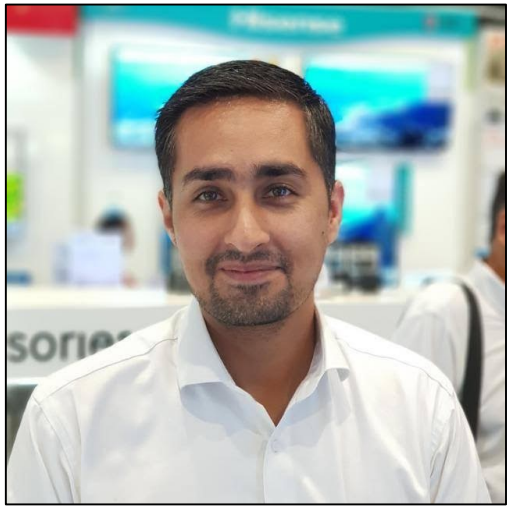
Dr. Gurdal Ertek

Wind Turbine Accidents: A Data Mining Study

Dr. Gurdal Ertek

Acknowledgement

- **Abu Dhabi Education Council (ADEC)**
  - ADEC Award for Research Excellence
  - (AARE 2015)
- **Personal Development Fund**
  - Abu Dhabi University
- **Data Cleaning & Analysis**
  - Ayaz Salman (Research Consultant)



Thank you. Your Questions?