Original Article

Predicting appointment misses in hospitals using data analytics

Sylvester Rohan Devasahay[1,2], Sylvia Karpagam[3], Nang Laik Ma[4]

[1]Data Science, School of information Systems, Singapore Management University, Singapore; [2]Information Technology specialized in Analytics, Bangalore, India; [3]Rajiv Gandhi University of Health Sciences, Bangalore, India; [4]School of Business, Senior Lecturer, SIM University, Singapore

Contributions: (I) Conception and design: NL Ma, SR Devasahay; (II) Administrative support: NL Ma; (III) Provision of study material or patients: NL Ma; (IV) Collection and assembly of data: SR Devasahay; (V) Data analysis and interpretation: SR Devasahay; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Sylvester Rohan Devasahay. Data Science, School of information Systems, Singapore Management University, Singapore; Postgraduate in Information Technology specialized in Analytics, Bangalore, India. Email: Sylvester.rohan@gmail.com.

Background: There is growing attention over the last few years about non-attendance in hospitals and its clinical and economic consequences. There have been several studies documenting the various aspects of non-attendance in hospitals. Project Predicting Appoint Misses (PAM) was started with the intention of being able to predict the type of patients that would not come for appointments after making bookings.

Methods: Historic hospital appointment data merged with "distance from hospital" variable was used to run Logistic Regression, Support Vector Machine and Recursive Partitioning to decide the contributing variables to missed appointments.

Results: Variables that are "class", "time", "demographics" related have an effect on the target variable, however, prediction models may not perform effectively due to very subtle influence on the target variable. Previously assumed major contributors like "age", "distance" did not have a major effect on the target variable.

Conclusions: With the given data it will be very difficult to make any moderate/strong prediction of the Appointment misses. That being said with the help of the cut off we are able to capture all of the "appointment misses" in addition to also capturing the actualized appointments.

Keywords: Appointment misses; logistic regression; decision tree; prediction

Received: 23 November 2016; Accepted: 06 March 2017; Published: 17 April 2017.
doi: 10.21037/mhealth.2017.03.03
View this article at: http://dx.doi.org/10.21037/mhealth.2017.03.03

Introduction

It is essential that hospitals are able to efficiently utilize the working hours of their doctors. By maximizing doctor utilization hospitals are able to help their community better and faster, enabling all the ripple effects of an early diagnosis. Project Predicting Appoint Misses (PAM) was started with the intention of being able to predict the type of patients that would not come for appointments after making bookings. This was based on the assumption that there are certain external (independent) factors like "age" or "gender" that contributed to the missed appointments.

The hospital in study is a 590-bed multi-specialty general and acute care hospital. The hospital offers a comprehensive range of medical services and specialist care to the community in Singapore.

To visit doctors, patients are required to make an appointment beforehand. As part of the services to the patient, the contact center at the hospital helps patients with the booking/change/cancelation of appointments. Appointment reminders are sent to patients to remind them about their appointments and to confirm whether they will come for the reserved time slot or not.

Process followed for booking appointments

An appointment with the doctor is typically made when either the clinic or the patient calls the hospital phone line.

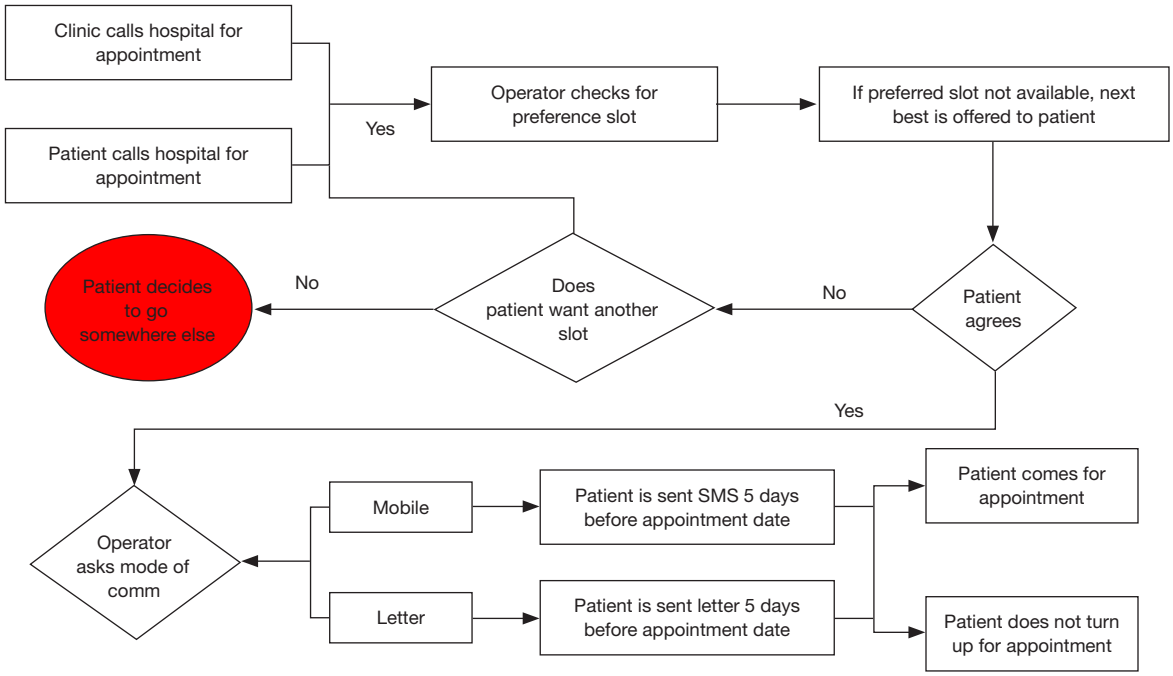

Figure 1 Appointment booking process flow.

The operator then checks if the preferred slot is available and informs the patient. The patient is reminded about their appointment 5 days before the actual appointment either by SMS or Email (*Figure 1*).

Methods

Data

The data is for the year 2013 from Jan to Dec. The overall no show rate for the year 2013 was 18.59%. There are a total of 43 columns in the data out of which there are 13 columns that either blank or not applicable and 7 are 'diagnose' related and 5 are codes that are not being considered. A Second data set containing distance of the patient's house from the hospital was added to the original data. This was the only other data that was added. The data was available in Radius distance from the hospital, i.e., we were not able to tell the exact address of the patient but the radius he would fall into so for, e.g., if the data says 2 km that means the patient could be anywhere in a 2 km radius from the hospital. Out of the overall raw data we found only 8 variables to be significant this was decided based on the "parsons test " variables that showed a chi square test of less than 0.0001 were considered for the analysis this means that there is a 0.0001 chance of the data that we have chosen to affect the target variable only by chance. Most of the 8 remaining variables had to be modified to binary variables this procedure resulted in 37 new variables. There was no further validation done. The variables were converted to binary after deciding if they were significant.

Analysis—tools and methodology

SAS 9.3 & JMP Pro11 were the primary tools used for the analysis and predictive modeling.

High level view of methodology

Below is the high level view of the process that was followed for project PAM (*Figure 2*).

High level view of predictive modelling

The first process with the data was to ensure it was clean. This included removing all the incorrect data and the missing data ensuring it was fit to run in SAS. The remaining data was processed through SAS and only 8 significant variables were considered for the study. Among the 8 variables we transformed a few to binary. This data

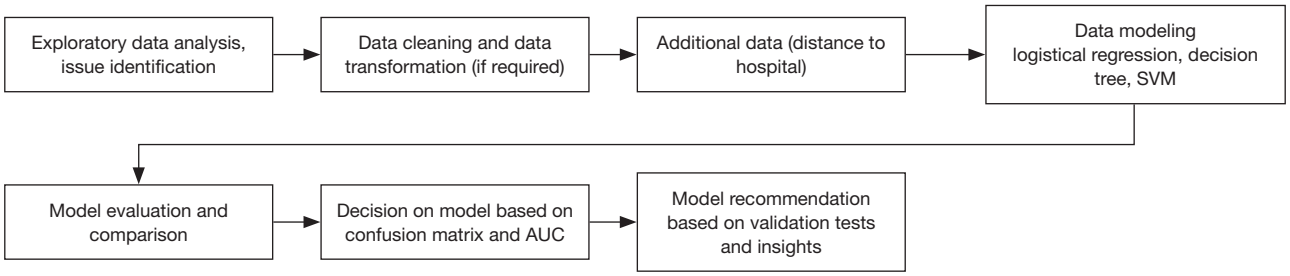

Figure 2 High level view of methodology.

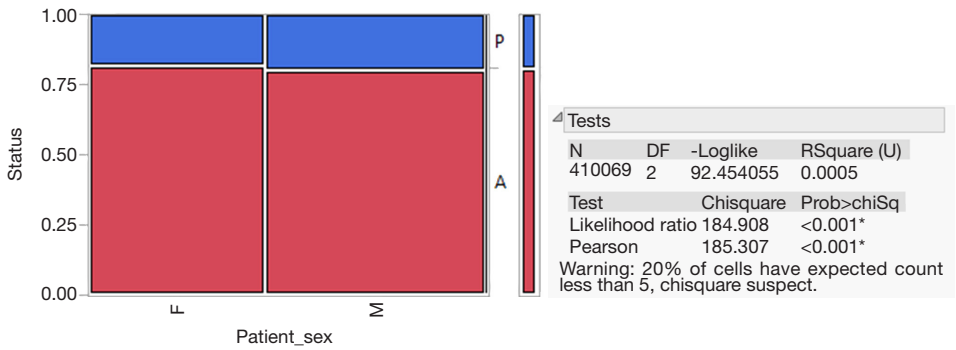

Figure 3 Significance test for gender.

was first divided into Training and Validation, with this data three models were run, logistic regression, decision trees and SVM. SVM yielded no results therefore results of the logistic regression and decision trees were compared to decide the best model. For the comparison we used the confusion matrix to see which model gave us the better sensitivity and positive predicted value.

Exploring the data

For the relevant variables the first thing done was to check the significance against the response variable (Status-) to gauge if difference was only by chance or was the difference significant. JMP was used to decide if the variables were significant or not, below are a few examples of the results from JMP (*Figure 3*).

The significance test for the variable "patient-sex" measured against the target variable "status" with the Chi-square test showing us that it is less than 0.0001 chances of the data not being significant The graph tells us the males have a slightly higher "missed appointment" rate that females. Similarly all the variables were checked the second example is for the variable "patient-age" again measured against the target variable (*Figure 4*).

When we look at mode of communication we can see that 75% of the patients usually choose updates through "SMS" while 24% chose "Letters", between them there is hardly a difference in the missed appointment percentage. However, we know from the data that if the hospital does not capture the patients details (<1%) the chances of the person not coming for hospital appointments go up (*Figure 5*).

Break down of variables contributing to appointment misses—class

In Singapore patients broadly fall under 2 class categories "Private" or "Subsidized". These classes are further broken down in to sub-classes (*Table 1*). We wanted to check if getting any sort of subsidy has an effect on whether the patient shows up for appointment. The variable was checked against each subclass (*Figure 6*) and also the broad category (private or subsidized) breakdown (*Figure 7*). Under the 2 broad categories we can see that if a person is subsidized the chances of them not turning up for an appointment is higher that a person that is private. If we do a further drill down we can see a particular subclass (PTRF-

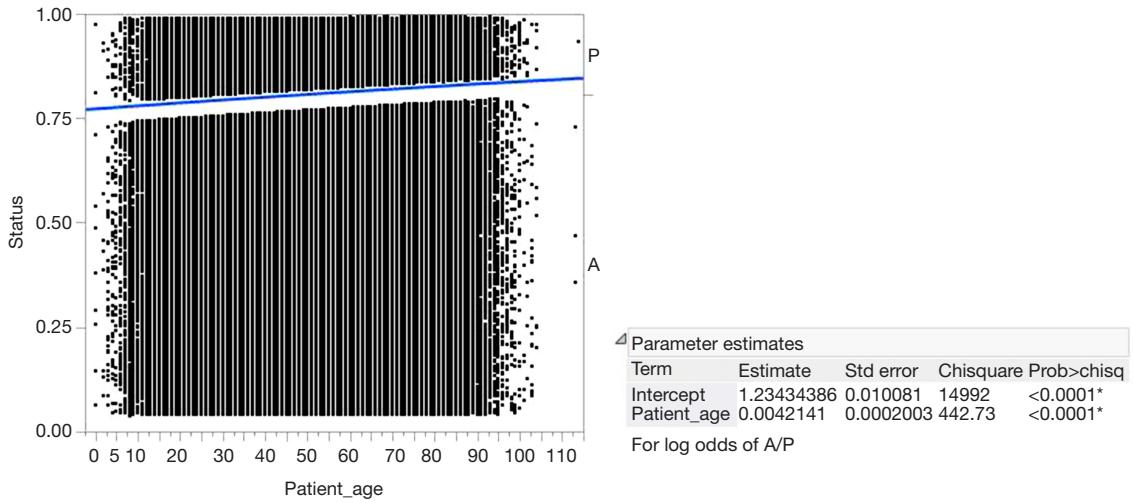

Figure 4 Significance test for gender.

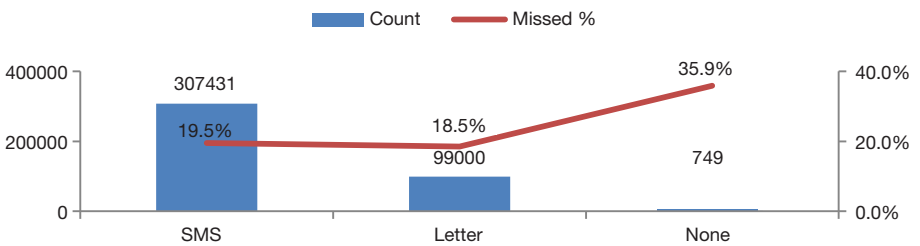

Figure 5 Distribution of missed appointments over "Mode of Communication".

Permanent Resident Foreigner) under Private that has a very high missed appoint rate. This tells that apart for the broad classes, the sub-classes also have a role to play in the patients not turning up for hospital appointments.

Break down of variables contributing to appointment misses—time related variables

With the variables labeled "Visit time" and "Visit Date" 5 sub-variables were created this was done in order to check if there was any sort of contribution from any sub-parameter that caused high no shows. The time related variables were broken down in to "month wise", "day wise", "1/3 day wise" (morning, afternoon and evening) and "hourly wise "data (*Figure 8*). While looking at the month wise data we see that there is a range of only a 2% indicating that months do not have any significant contribution towards patients missing appointments, Day wise tells us that there have been some incorrect captures by the system for Sunday apart from that there is nothing significant. The "1/3 day wise" also does not show any significant contribution to missed appointments. While the hourly break down tells us that patients in the morning and afternoon tend to be more regular. The count of the appointments in an hour are not significant against the population data but we can say that appointments that are in the morning (8 to 9 am) and appointments in the afternoon (12 to 2 pm) have a less chance of appointment misses.

Break down of variables contributing to appointment misses—demographics

Distance was not part of the original data set and had to be added. The distance data represented distance in the form of radius and did not have directionality. The distance was then grouped into bands with a change in band every 2 km. Contribution by gender and age against the target variable (status) were also looked at to see if there was a particular sub-category that had an impact on the target variable (*Figure 9*).

There are 3,263 data points without the distance parameter this is because these patients live in the country the hospital is located in. For data points that fall under the "unknown" category the missed appointment rate is around 26%. This may suggest that foreigners may have a higher missed appointment rate. Due to the small number of data points the argument cannot be sound statistically.

For the age variable we grouped patients by their age (e.g., all patients that are 30 years are one group, all that are 31 are in another group and so forth) but were not able to find any significance in relation to missed appointments, the age variable was broken down and studied because the hospital thought that patients below 21 were contributing towards "missed appointments" however this was not the case.

Table 1 Breakdown of class

Class	Count	New class
A	86	Private
AP	4	Private
ARF	25	Private
B1	167	Private
B1P	14	Private
B1RF	5	Private
B2RF	5	Private
CRF	149	Private
NR	6,589	Private
PTE	79,630	Private
PTEP	6,432	Private
PTRF	19,683	Private
B2	845	Subsidized
B2P	53	Subsidized
C	1,046	Subsidized
CP	51	Subsidized
SUB	279,190	Subsidized
SUBP	16,095	Subsidized

Break down of variables contributing to appointment misses—hospital referrals

There are a total of 796 clinics out of which the top 10 clinics with high "no shows" are displayed.

Under "Moderate" there are 95 clinics—in order to fall under the "moderate" category a clinic must have between 31 to 377 appointments per year. Similarly "others" and "rare" was created in in order to reduce the scope of the clinics. Intra-department referrals have 28% missed appointments which is way higher that the average missed appointments.

Using all the clinics in the graph will skew the data as there are several clinics that only refer patients once in a year because of this issue the data was changed to "other"—for the clinics that had less than 2 appointments for the year 2013, "Rare" for appointments between 3 and 30 for the year 2013 and "Moderate" for appointments ranging from 31 to 377 (this was the closest to 1 appointment/day and that's the reason it was taken as a cut off).

We found that internal department referrals had a high missed appointment rate (28%) with the next closest clinic

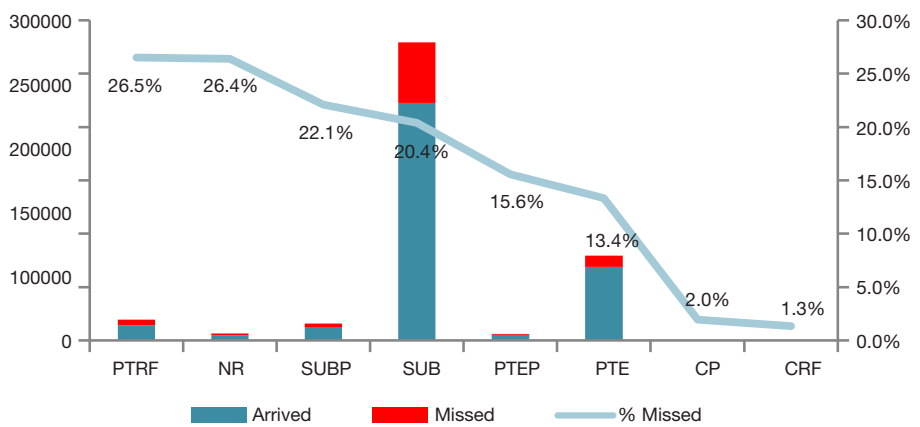

Figure 6 Distribution of missed appointments in sub-class

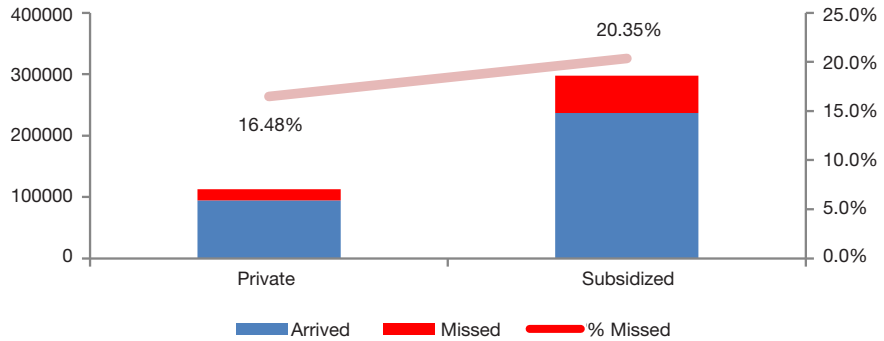

Figure 7 Distribution of missed appointments over "class" variable

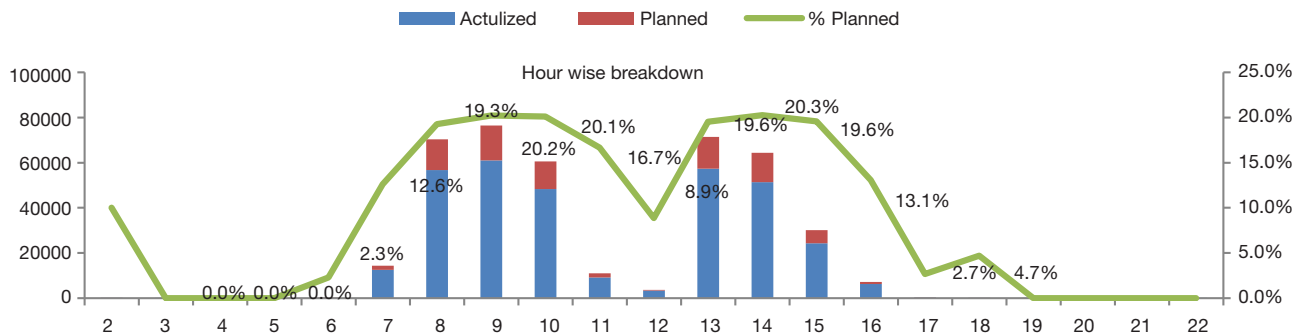

Figure 8 Distribution of missed appointments over hours of day variable.

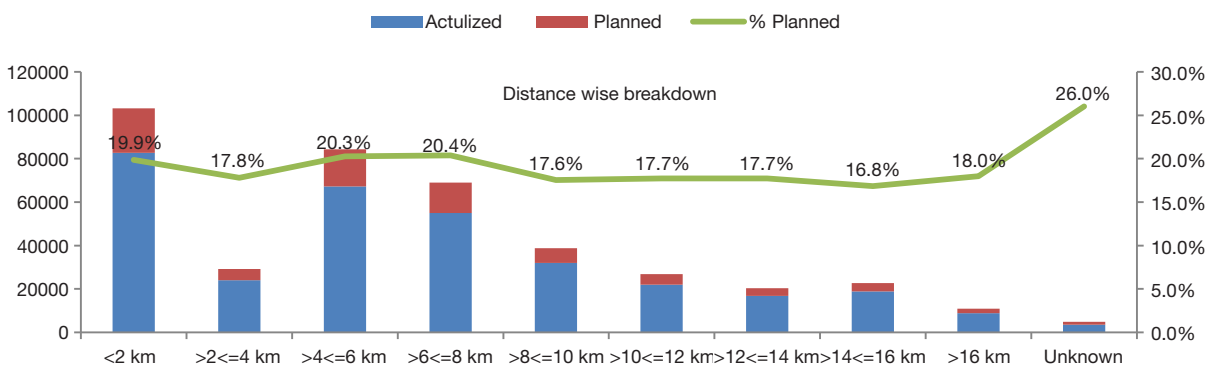

Figure 9 Distribution of Missed appointments over "Distance from Hospital" variable

at about 8% less.

Results for logistic regression & decision trees

Among the models that were tried the best fit was decision tree (*Table 2*) with a 0.15 cut off , the results were 23.22% sensitivity with a PPV of 15.58%, which tells us that even with the significant variables we will not be able to accurately make predictions on what type of patients will miss appointments. One of the primary reasons for this is because of the large difference in the proportion of the target variable (appointment misses).

Discussion

Neal *et al.* found that in the UK, missed appointments had

Table 2 Results for logistic regression and decision tree

Parameters	Decision tree results (Cutoff.20) (%)	Decision tree results (Cutoff.15) (%)	Logistic regression results (cutoff.17) (%)	Logistic regression results (cutoff.15) (%)
Sensitivity	3.52	23.22	2.19	0.11
Specificity	99.26	23.94	91.72	99.98
Positive predicted value	53.24	15.58	17.43	61.68
Negative predicted value	81.15	34.02	53.98	80.73

a prevalence of 4.5–6.5% of booked consultations (1,2). Patients who missed appointments tended to cite practice factors and their own forgetfulness as the main reasons for missing appointments with the commonest reasons being mistakes and misunderstandings (frequently by the practice) and forgetfulness (3).

Husain *et al.* [2004] have documented how health professionals hold patients responsible for missed appointments and tend to view them negatively. It is important that the reasons that patients miss appointments are analyzed so that the negative outcomes of these can be reduced (4). George *et al.* [2003] found that participants did not feel obliged to keep an appointment partly because they felt disrespected by the health care system, an effect which was compounded by participants' lack of understanding of the scheduling system (5).

Alhamad [2013] has studied the reasons for missing appointments were studied from December 2010 to March 2011 in Alwazarat health center, Riyadh, Saudi Arabia. The demographic factors associated with missed appointments were female gender, younger age group and poor socio-economic status. The important top five causes for missing appointments were difficulty booking an appointment, work commitment, long distance travel, and unavailability of transportation and visiting another healthcare facility. Knowing factors associated with missed appointments can help improve quality of care and control those variables that can be changed to reduce the percentage of missed appointments. This would have a direct impact on clinical care and the economics (6).

Alaeddini *et al.* [2011] found that the number of no-shows has a significant impact on the revenue, cost and resource utilization for almost all healthcare systems using a hybrid probabilistic model based on logistic regression and empirical Bayesian inference to predict the probability of no-shows in real time using both general patient social and demographic information and individual clinical appointments attendance records. The model also considered the effect of appointment date and clinic type. The effectiveness of the proposed approach was validated based on a patient dataset from a medical center. The felt that this prediction model could be used to enable a precise selective overbooking strategy to reduce the negative effect of no shows and to fill appointment slots while maintaining short wait times (7).

Appointment disruption occurs due to patient no-show and cancellation and can cause a disturbance in scheduling appointments in an efficient manner, leading to wastage of vital human and economic resources. Patients who require appointments will also not be able to be offered the space, usually because of extremely short notice period. Although overbooking may help in some instances, it can lead to overcrowding and dissatisfaction among patients. There is therefore a need for accurate prediction of no show and possibility of cancellation.

In article (7), they develop a hybrid probabilistic model based on multinomial logistic regression and Bayesian inference to predict accurately the probability of no-shows and cancellations in real-time. The result of the proposed method can be used to develop more effective appointment scheduling (8-12). It can also be used for developing effective strategies, such as selective overbooking for reducing the negative effects of disturbances and filling appointment slots while maintaining short waiting times (13-15). Efficacy of any scheduling system primarily depends on its ability to forecast and manage different types of disruptions and uncertainties.

Alaeddini *et al.* [2014] have developed a probabilistic model based on multinomial logistic regression and Bayesian inference to estimate individuals' probabilities of no-show, cancellation and attendance in real-time. Based on real patient data collected from a Veterans Affairs medical hospital, the team modeled the effect of the appointment date and clinic on the proposed method and state that

their approach is computationally effective and easy to implement, taking into account individual patient behavior unlike population based methods and also generating reliable probabilistic estimates. They feel that the method proposed by them could be used to develop more effective appointment scheduling systems and more precise overbooking strategies to reduce the negative effect of no-shows and fill in appointment slots while maintaining short waiting time (16).

Huang and Hanauer have developed a new approach for robust overbooking of appointments. The study has a specific department in focus (General Pediatrics) and looks at possible contributors to missed appointments in that department (17).

"Lee VJ Predictors of failed attendances" was published by *BMC Health Services* 2005. The study focuses on patients attending all outpatient clinics at Tan Tock Seng Hospital; it is a 1,400-bed hospital in Singapore. They used Stata for their analysis. The models that were used were logistic regression and the decision tree to come up with predictive models.

A few of the articles from other studies focus on the reason the patients gave for missing hospital appointments. In this study, we attempt to predict no shows using regular available hospital data to be able to make predictions on patients missing hospital appointments. A few key differences compared to the other studies are we looked at the internal and external referrals. This variable had a key part to play in missed appointments, Intra-Department referrals from A&E was 28%, which was 10% more than the average (18%) missed appointment rate; Intra-Department Referral for Subsidized was at 23%, which is 5% more that the average (18%). This was brought to the attention of the hospital administration during the Annual Management meeting. Another was the "Class" variable this variable tells us if a patient was subsidized or private and the effect on missed appointments. Analyzing this variable also gave us insights on the impact it has on missed appointments.

Challenges

The data had to be anonymized in order to meet the business requirements this led to a few complications.

- The distance variable was not very precise and only very wide assumptions could be made as we did not have the exact location and only had the radius. We could have checked if there was a certain area that had a lot of missed appointments and any causes for that.
- The data had a lot of false entries, e.g., there were appointments on Sunday when the hospital was closed.
- There was a lot of missing data and incorrect data. All these had to be removed.
- As the models that we were using was essentially for binary variables most of the data had to be modified to a binary variables this procedure was tedious as I had to create 37 variables for this.

Conclusions and major findings for the study

- Males are slightly more inclined to miss appointments, in average they have a 2% higher rate of appointment misses.
- Patients that are subsidized tend to miss more appointments than private patients however during a drill down it was found that a particular sub class of private patients (Resident Foreigners) tend to miss 8% more that the average missed appointment rate.
- The time of the day has an effect on the missed appointment rate typically mornings (8 to 9 am) and afternoon 12 to 2 pm) have a significant less appointment miss rate.
- Before the study the hospital were under the assumption that the age group (21 and below) had a significant contribution to missed appointments however there is no such trend found after analyzing the data.
- The internal referrals from the hospitals from one department to another are significantly higher (10% more) than the average missed appointment rate.
- Given the Sensitivity and the specificity of the proposed models (we need to consider the fact that the proposed cutoff point was based on optimum performance of the model). We can conclude that with the given data it will be very difficult to make any moderate/strong prediction of the Appointment misses.
- That being said with the help of the cut off we are able to capture all of the "appointment misses" in addition to also capturing the actualized appointments.

Acknowledgements

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest

to declare.

References

1. Neal RD, Lawlor DA, Allgar V, et al. Missed appointments in general practice: Retrospective data analysis from four practices. Br J Gen Pract 2001;51:830-2.
2. Waller J, Hodgkin P. Defaulters in general practice: who are they and what can be done about them? Fam Pract 2000;17:252-3.
3. Neal RD, Hussain-Gambles M, Allgar VL, et al. Reasons for and consequences of missed appointments in general practice in the UK: questionnaire survey and prospective review of medical records. BMC Fam Pract 2005;6:47.
4. Husain-Gambles M, Neal RD, Dempsey O, et al. Missed appointments in primary care: questionnaire and focus group study of health professionals. Br J Gen Pract 2004;54:108-13.
5. George A, Rubin G. Non-attendance in general practice: a systematic review and its implications for access to primary health care. Fam Pract 2003;20:178-84.
6. Alhamad Z. Reasons for missing appointments in general clinics of primary health care center in riyadh military hospital, Saudi Arabia. Int J Med Sci Public Health 2013;2:258-67.
7. Alaeddini A, Yang K, Reddy C, et al. A probabilistic model for predicting the probability of no-show in hospital appointments. Health Care Manag Sci 2011;14:146-57.
8. Chakraborty S, Muthuraman K, Lawlwey M. Sequential clinical scheduling with patient no-shows and general service time distributions. IIE Transactions 2010;42:354 66.
9. Glowacka KJ, Henry RM, May JH. A Hybrid Data Mining/Simulation Approach for Modelling Outpatient No-Shows in Clinic Scheduling. The Journal of the Operational Research Society 2009;60:1056-68.
10. Gupta D, Denton B. Appointment scheduling in health care: Challenges and opportunities. IIE Transactions 2006;40:800-19.
11. Hassin R, Mendel S. Scheduling arrivals to queues: A single-server model with no-shows. Management Science 2006;54:565-72.
12. Liu N, Ziya S, Kulkarni VG. Dynamic scheduling of outpatient appointments under patient no-Shows and cancellations. Manufacturing & Service Operations Management 2010;12:347-64.
13. LaGanga LR, Lawrence SR. Appointment scheduling with overbooking to mitigate productivity loss from no-shows. Proceedings of Decision Sciences Institute Annual Conference, Phoenix, Arizona, November 17-20, 2007.
14. Muthuraman K, Lawley M. A stochastic overbooking model for outpatient clinical scheduling with no-shows. IIE Transactions 2008;40:820-37.
15. Zeng B, Turkcan A, Lin J, et al. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Annals of Operations Research 2009;178:121-44.
16. Alaeddini A, Yang K, Reeves P, et al. A hybrid prediction model for no-shows and cancellations of outpatient appointments. IIE Trans Healthc Syst Eng 2015;5:14-32.
17. Huang Y, Hanauer DA. Patient no-show predictive model development using multiple data sources for an effective overbooking approach. Appl Clin Inform 2014;5:836-60.

doi: 10.21037/mhealth.2017.03.03
Cite this article as: Devasahay SR, Karpagam S, Ma NL. Predicting appointment misses in hospitals using data analytics. mHealth 2017;3:12.