International Journal of Medical Informatics 147 (2021) 104373

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Customer relationship management analysis of outpatients in a Chinese infectious disease hospital using drug-proportion recency-frequency-monetary model

Min Li [a,b,1], Qunwei Wang [a,1], Yinzhong Shen [b,*], TongYu Zhu [b,*]

[a] Nanjing University of Aeronautics and Astronautics, College of Economics and Management, Nanjing, Jiangsu, 211106, China
[b] Shanghai Public Health Clinical Center, Fudan University, Shanghai, 201508, China

ARTICLE INFO

Keywords:
CRM
dRFM model
K-means
Hospitals for infectious diseases
Patient types

ABSTRACT

Background: Identifying the patient types with different economic values can be useful for hospital development.
Objective: This work uses the theory of customer relationship management (CRM) to analyze the outpatients in the hospital for infectious diseases in Shanghai, China.
Methods: A total of 2,271,020 data elements of outpatients in the research unit between August 2009 and December 2019 were extracted, analyzed and cleaned to obtain 171,107 valid data elements (1 element per person). The main diseases were viral hepatitis B (VHB) and acquired immunodeficiency syndrome (AIDS), and the average percentage of drug expenditure was 80.39 %.
We innovatively expanded the classic RFM (R: recency, F: frequency, M: monetary) model in CRM to the dRFM (d: percentage of drug expenditure) model. We selected the best clustering algorithm from the K-means, Kohonen and two-step clustering methods to find the optimal model to distinguish the types of patients with different economic values and the best decision-making algorithm from the C5.0, CART classification regression tree, CHAID and QUEST algorithms to verify the model.
Results: After performing two rounds of K-means clustering analysis on three models: RFM, RFM + dRFM and dRFM, and 97,855 data elements were retained. The RFM + dRFM model was the optimal model, clustering the patients into 3 types: potential patients (24.2 %) to be retained, with a high drug expenditure and the last visit in more than 19.06 months, high-value patients (24.5 %) to be attracted, with the last visit in about 6.66 months; basal patients (51.3 %) to be kept, with the last visit in about 3.7 months. The model was then verified using the C5.0 decision tree algorithm with an accuracy rate of 99.97 %.
Conclusion: This objective CRM analysis of the patients in the hospital for infectious diseases using the dRFM model accurately identified different types of patients, providing an objective and effective basis for hospital management.

1. Introduction

According to the World Health Organization (WHO), infectious diseases are one of the top 10 global health threats [1]. In China, the top three category A and B infectious diseases are viral hepatitis (mainly viral hepatitis type B, VHB), pulmonary tuberculosis (PTB) and syphilis in terms of the incidence rate, and AIDS, PTB and VHB in terms of the caused deaths [2]. These represent severe infectious diseases spreading around the world and should be jointly prevented [3–9].

With 167 hospitals in 2018, including 165 public ones, the hospitals for infectious diseases represent an important part of the healthcare system in China [2,3]; they accept and treat patients with a single type of infectious diseases and represent public welfare hospitals with a limited government compensation. For the public welfare hospitals, the revenue sources mainly included the government compensation and medical income. However, since the government compensation has been insufficient, the hospital development mainly relied on the medical income, which resulted in providing a high proportion of medicines [4,10–12]. In addition, for diseases that require regular long-term or life-time medical treatment, such as TB and AIDS the hospitals for infectious

* Corresponding authors.
E-mail addresses: minliji@aliyun.com (M. Li), wqw0305@126.com (Q. Wang), shenyinzhong@shphc.org.cn (Y. Shen), zhutongyu@shphc.org.cn (T. Zhu).
[1] Equal contribution, considered to be co-first authors and co-corresponding authors.

https://doi.org/10.1016/j.ijmedinf.2020.104373
Received 26 July 2020; Received in revised form 23 December 2020; Accepted 27 December 2020
Available online 29 December 2020

diseases offer fixed-point treatment, with the provision of free treatment, such that the patients regularly check up at the designated hospitals to take the medicine [4–10].

There is a need to control the medical expenses and high proportions of medicines at the hospitals for infectious diseases in China. However, the increased medical costs represent a challenge to the world health systems, and controlling this increase is mainly targeted at decreasing the drug expenditure [10–14]. In 2017, the Chinese government abolished the drug mark-up and forced the public hospitals to control the percentage of drug expenditure at 30 % [10,11,15]. In 2018, there were 59,642 health workers in the 167 hospitals for infectious diseases in China, including 15,031 licensed doctors (physician assistants) [2]. These factors added to the duties of the hospitals for infectious diseases which embody the Chinese government's public welfare nature of medical treatment and treating patients with legally infectious diseases. As a result, these hospitals had to face the challenges of insufficient government compensation, cancellation of drug additions and control of the high proportion of drugs [1,10,12].

The customer relationship management (CRM) analysis denotes the managerial efforts to manage the business processes and technologies that are designed to understand the customers of a firm. In hospital CRM, the classical recency-frequency-monetary (RFM) model and data mining are widely applied [14,16–21]. In the RMF model, recency (R) refers to the last time of shopping, frequency (F) means the frequency of shopping, and monetary (M) indicates the purchase expense. The customer value increases with the value of the RFM model [16–18]. Performing a CRM analysis of the outpatients in the hospitals to subdivide and identify the different types of patients would be of great significance to improve the further consultation rate and loyalty of the patients with infectious diseases, thus enhancing the core competitiveness of the hospitals for infectious diseases and optimizing the medical services and resources [16–18].

Lee EW (2012) applied the K-means algorithm according to the RFM model (M is the total cost of patients) to classify 1-year hospitalized patients into 2 categories: loyal patients (70 %) and general patients (30 %), and then used the decision tree algorithm to build a model for the 2 types of patients [21]. However, the study performed a very simple classification that is unsuitable for the patients' data over 2 years [16–18]. In the work of Tarokh MJ (2019) [16], CRM, the customer life cycle value (CLV), K-means clustering algorithm and decision tree algorithm were used to analyze, group and rank the patients and identify target ones. Based on the resulting types of patients, strategies can then be developed to attract potential ones, eventually increasing the loyalty of the patients and maximizing the hospital profits. Nevertheless, the negative value of the CLV model in the hospital was not considered in the study [17], and the loss analysis of the patients included their income; thus, it is not suitable for the Chinese public welfare hospitals [10].

A modified weighted RFM model (eRFM) was used in the work of Zare Hosseini Z and Mohammadzadeh M (2016) [17] along with a CLV model in an adopted CRM for an empirical research on patients to identify target and potential patients and faithful customers, thus improving the loyalty and satisfaction of the patients and maximizing the hospital profits [17]. They believed that the patients have uncertain visits, such as emergency visits, outpatient visits, hospitalization and referrals due to a sudden illness, since the patients are special consumers. As a result, the classical RMF model was modified into the weighted eRFM model which divided the patients into the lost patients (41.17 %, frequency of visits: 1 time within the last 3 years), low-value patients (time of last visit: over 1.5 years, or new patients) and high-value patients (1.45 %, time of last visit was: 0.5 year, average frequency of visits: 13.2 times within 4 years, making a great contribution to the income of the hospitals). This model can reflect the loyalty of the patients, represented in the length of their visits. However, in the Chinese infectious disease hospitals, AIDS patients need to take medicine for life, HBV patients need to take medicine for a long time, and FTB patients need to take medicine regularly [4–9]. Therefore, the length of the treatment period is not applicable to the Chinese infectious disease hospitals [17].

In a previous work of Min Li (2019) [18], 170,246 data elements were extracted from the records of the outpatients in an infectious disease hospital in Shanghai from January to July 2016. Then, data cleaning resulted in 43,448 data elements and two-step K-means clustering and two-step Kohonen clustering algorithms were used to perform an RFM model analysis. As a result, the patients were divided into 3 categories: important patients (4,786 people, 11.72 %, $R = 2.89$, $F = 11.72$, $M = 84302.95$), major patients (23,103 people, 53.2 %, $R = 5.22$, $F = 3.45$, $M = 9146.39$) and potential patients (15,559 people, 35.8 %, $R = 19.77$, $F = 1.55$, $M = 1739.09$). However, the study had the following limitations: (1) The infectious disease hospital is not considered is a non-profit hospital, and the outpatient income cannot be used to classify the types of patients; thus, the three types of patient expenses cannot truly reflect the contribution to the hospital; (2) It does not take into account that the income of the hospitals for infectious diseases from outpatients is mainly composed of the medicine expenses, i.e., the high proportion of medicines, and these expenses represent a negative indicator for public welfare hospitals in China. The model neither considers nor excludes the medicine expenses from the overall medical expenses. The income is restricted by the government and the proportion of medicines is less than 30 %, while the proportion of medicines in the hospitals run by the government is more than 50 %. The drug fees are not only confiscated but also penalized by the government [10,11,15], which means that only the patients with a high cost of drug fees or with low levels have an important contribution to the development of infectious disease hospitals. When the patient's proportion of medicine was taken into account, this changed the length of the patient's consultation cycle [16], resulting in different consultation frequency and total medical expenses; thus, the results cannot truly distinguish the type of patients in the Chinese infectious disease hospitals [18].

In summary, the study of Lee EW (2012) is too simplistic [21], the work of Tarokh MJ (2019) [16] did not take into account the negative value of the CLV model in the hospital, the length of the patient treatment cycle used in the model of Zare Hosseini Z (2016) is not applicable to the Chinese infectious disease hospitals [17], and the work of Min Li (2019) [18] did not consider the importance of the proportion of drugs to China's public hospitals and hospitals for infectious diseases. Therefore, the simple application of the RFM model cannot really distinguish the types of patients in the Chinese infectious disease hospitals [4–11, 15], and the model needs to be improved. To this end, we used the percentage of drug expenditure as a major variable in this study. From the Hospital Information System (HIS), we extracted 2,271,020 data elements of the patients visiting the Outpatient Department at the research unit from August 2009 to December 2019, and the modified drug proportion RFM (dRFM) model was applied for the patient CRM analysis to identify the basal, potential and high-value patients. This analysis provides a reference for the optimization of the medical services and resources and data support for development strategies and decision-making in the hospitals for infectious diseases in China.

2. Data and methods

2.1. Data collection and preprocessing

Using the methods from literature [16–18,21], a total of 2,271,020 data elements of the patients who visited the Outpatient Department at the research unit between August 1, 2009 and December 31, 2019 were extracted from the HIS. The data included the fields of time of visits, patient ID number, gender, age, place of residence (local or field), diagnosis, medical cost and drug expenditure. We used a unified representation of the diagnosis according to the International Classification of Diseases (ICD)-10 code of main diagnostic criteria, which was used for the disease group characteristics [21]. In Shanghai, China, real-name

Table 1
Model quality of the three clustering algorithms of RFM model and dRFM model.

Model type	Sample	K-means		Kohonen		Two step clustering	
		Model quality[1]	Importance of predictive variables	Model quality	Importance of predictive variables	Model quality	Importance of predictive variables
RFM model[2]	171,107	0.6501	R = 1 F = 1 M = 0.574	0.4858	R = 1 F = 1 M = 1	0.3557	R = 1 F = 1 M = 1
dRFM model[3]	171,107	0.6009	d = 1 R = 1 F = 1 M = 0.4805	0.4705	d = 1 R = 1 F = 1 M = 0.877	0.3516	d = 1 R = 1 F = 1 M = 1

Note: R:Recency, F:Frequency, M:Average medical cost per visit, d:Percentage of drug expenditure. Values lie in the 0–1 range, The closer to 1, the more important.
1. Model quality is an indicator that shows the goodness, importance and overall model fit of each cluster, and it is obtained as the output of the SPSS Modeler 18.0 software package when modeling is performed. Values lie in the 0–1 range, the closer to 1, the better the model quality.
2. For RFM model, the Model quality of K-means (0.6501) is better than Kohonen (0.4858) and Two step clustering (0.3557). Choose K-means clustering algorithm.
3. For dRFM mode, the Model quality of K-means (0.6009) is better than Kohonen (0.4705) and Two step clustering (0.3516). Choose K-means clustering algorithm.

treatment is performed, and the ID number is the only required field for the patient identification. The Statistical Product and Service Solutions (SPSS) 22.0 and SPSS Modeler 18.0 software package were used for the data analysis. SPSS Modeler 18.0 was used for the calculations and gave the model parameters as the output.

Next, we applied field cleaning and expansion on the data elements using the methods from literature [16–18,21] according to the following steps: (1) Add the "Recent consultation month" field and form a new variable, with December 2019 as the first month, November 2019 as the second month and so on until August 2009 as the 125th month; (2) Add 3 fields: sum up each patient's total medical costs, total drug costs and frequency of visits; (3) Generate 3 new variables as follows [2]: the average medical cost per visit = total medical cost / frequency of visits the average drug expenditure per visit = total drug expenditure / frequency of visits the percentage of drug expenditure (%) = (average drug expenditure per visit / average medical cost per visit) × 100 %. The data were checked and duplicates were deleted. As a result, an effective dataset (1 data element per person) was obtained.

2.2. Formation of variables and descriptive statistical analysis

In accordance with the literature [16–18,21], the following seven variables were created and analyzed: time of last visit (month), gender, age, place of residence (local /field), percentage of drug expenditure, frequency of visits and average medical cost per visit. The disease spectrum analysis was carried out on the diagnoses in the 2,271,020 data elements on an annual basis.

2.3. Selecting the optimal clustering model

According to the literature [16–18,21], three clustering methods were included in the RFM model and dRFM model to select the best clustering algorithm: K-means, Kohonen and two-step clustering method, and a self-defined number of clusters was applied to select an optimal clustering model. The models in this study were calculated and modeled using the SPSS Modeler 18.0 software package. The model quality value is an indicator that shows the goodness, importance and overall model fit of each cluster, and it is obtained as the output of the SPSS Modeler 18.0 software package when modeling is performed. The best decision algorithm was chosen from four decision algorithms: C5.0 algorithm, CART classification regression tree, CHAID algorithm and QUEST algorithm. Next, the best clustering algorithm and the best number of model clusters were used to construct RFM, RFM + dRFM and dRFM models.

2.4. Validation of the model

The best decision algorithm obtained in step 2.3 was applied to verify the model with reference to the literature [21,22]. We used the clustering analysis results of the optimal model as the targets, and the important predictive variables as the input variables. The data were randomly divided into a training set, covering 70 % of the data and used to build the model, and a testing set, covering 30 % of the data and employed to verify the model.

3. Results

3.1. The disease spectrum analysis

A yearly increase was observed in the number of outpatients at the research unit. VHB and other liver diseases (hepatitis) and HIV accounted for over 60 % of the disease spectrum. The detailed results are listed in Supplementary Table 1, which shows that the top 20 diseases over the years in the period between August 2009 and December 2019 were mainly various types of viral and non-viral hepatitis, AIDS and PTB.

In 2019, the visits of patients with VHB, HIV and non-viral liver disease accounted for 47.09 %, 16.65 %, and 8.57 %, of the total visits, respectively. The number of visits of the patients with viral hepatitis type C declined year by year and this disease became treatable. The progression of VHB, AIDS and FTB was prone to causing drug-induced hepatitis (increasing year by year), fatty liver, gastritis, hypertension, diabetes, pulmonary infection and sleep disorders. In addition, renal impairment and hyperuricemia were also within the top 20 in the disease spectrum. Hyperuricemia entered the spectrum due to the improvement in the living standard and the changed diets of the Chinese people. As for the renal impairment, it entered due to the introduction of China's well-known kidney transplantation team into the research unit at the end of 2016, which attracted the patients to the hospitals.

3.2. Data preprocessing results

Since the ID number is the only required item for the patient, the information of each single patient's visit was accumulated into 1 person and 1 data element. The performed data cleaning resulted in 171,107 sets of valid data (1 data element per person) for patients between August 2009 and December 2019.

3.3. Descriptive statistical analysis

The detailed results of the descriptive statistical analysis are listed in Supplementary Table 2. Among the 171,107 data elements, there were 93,028 elements for males (accounting for 54.4 %) and 78,079 for females (accounting for 45.6 %), 122,116 elements with a local place of residence (accounting for 71.4 %) and 48,991 elements with a field place (accounting for 28.6 %). The average age was 46 years. The average percentage of drug expenditure was 80.39 %, and the median was 96.13 %, with a negatively skewed distribution. The average time of last visit (month) was 32.9 months, and the median was 20 months. The average frequency of visits was 13.2 times, and the median was 2 times, with a positively skewed distribution. Finally, the average medical cost per visit was 641.67 yuan, and the median was 451.02 yuan. The

Table 2
K-means cluster number and model quality of RFM model and dRFM model.

Model type	Sample	Number of clusters (n)	Model quality[1]	Importance of predictive variables
RFM model[2]	171,107	3	0.7054	R = 1, F = 1, M = 0.3787
		4	0.5949	R = 1, F = 1, M = 0.4711
		5	0.6501	R = 1, F = 1, M = 0.574
		6	0.6539	R = 1, F = 1, M = 0.7788
dRFM model[3]	171,107	3	0.6623	d = 1, R = 1, F = 1, M = 0.3256
		4	0.4805	d = 1, R = 1, F = 1, M = 0.4653
		5	0.6009	d = 1, R = 1, F = 1, M = 0.4805
		6	0.5948	d = 1, R = 1, F = 1, M = 0.5465

Note: R:Recency, F:Frequency, M:Average medical cost per visit, d:Percentage of drug expenditure. Values lie in the 0–1 range,The closer to 1, the more important.

1. Model quality is an indicator that shows the goodness, importance and overall model fit of each cluster, and it is obtained as the output of the SPSS Modeler 18.0 software package when modeling is performed. Values lie in the 0–1 range, the closer to 1, the better the model quality.
2. For RFM model, the quality of the model with the number of clusters 3 is the best, reaching 0.7054, and the quality of other models is 0.5949, 0.6501, 0.6539.
3. For dRFM model, the quality of the model with the number of clusters of 3 categories is the best, reaching 0.6623, and the quality of other models is 0.4805, 0.6009, 0.5948.

average percentage of drug expenditure was 80.33 %, and the median was 96.13 %. The two latter aspects mainly reflected the commonweal of hospitals for infectious diseases. However, the hospitals need survival and development, through realizing a useful classification of the patients, and the patients cannot be subdivided only according to the disease spectrum and percentage of drug expenditure. Therefore, it was necessary to conduct a dRFM analysis on the patients.

3.4. The optimal model

K-means, Kohonen and a two-step clustering method were used to construct the RFM model and dRFM model. According to the model quality value obtained by the SPSS software, K-means was the best algorithm, with a model quality value of more than 0.6. The model quality using the other two clustering algorithms was less than 0.5 (Table 1).

Using K-means to model the RFM model and dRFM model, the best number of clusters was found to be 3, which represents 3 types of patients. The quality of the model exceeded 0.66, while the quality of other models was lower than 0.66 (Table 2).

Using the C5.0 algorithm, CART classification regression tree, CHAID algorithm and QUEST algorithm to predict the correctness of the RFM model and dRFM model, the C5.0 algorithm was shown to be the best decision-making algorithm, and the correctness values of the RFM and dRFM prediction models were 100 % and 99.98 %, respectively (Table 3).

The RFM, RFM + dRFM and dRFM models were separately subjected to 2 rounds of modeling, and the RFM + dRFM model was found to be the best model (Tables 4 and 5). The following points represent a detailed elaboration. 1) The RFM model: This model represented repeating the previous work of Min Li (2019) [18] in this research, without considering the public welfare issues of the Chinese hospitals run by the government or the high proportion of drugs in the Chinese hospitals for infectious diseases and the issue of negative indicators for drug costs. As a result, this model was not applicable. 2) The dRFM model: After 2 rounds of clustering analysis, the time of the last visit (in

Table 3
Four decision-making algorithms of RFM model and dRFM model.

Model type	Sample	Number of clusters (n)	K-means clustering algorithm		Decision algorithm						
			Model quality[1]	Predictor variable importanc	Predictor variable importance				Prediction model accuracy (%)		
					C5.0	CHAID	CART	QUEST	C5.0[4]	CHAID	CART	QUEST
RFM model[2]	171,107	3	0.7054	R = 1, F = 1, M = 0.3787	R = 1	R = 0.989 F = 0.0049 M = 0.006	R = 0.991 F = 0.0045 M = 0.0045	R = 0.9951 M = 0.0049	100 %	94.48 %	100 %	99.32 %
dRFM model[3]	171,107	3	0.6623	d = 1, R = 1, F = 1, M = 0.3256	d = 0.4872 R = 0.5104 F = 0.0012 M = 0.0012	d = 0.5198 R = 0.4779 F = 0.0012 M = 0.0011	d = 0.5014 R = 0.4976 M = 0.001	d = 0.4853 R = 0.5132 M = 0.0015	99.98 %	95.74 %	98.71 %	98.26 %

Note: R:Recency, F:Frequency, M:Average medical cost per visit, d:Percentage of drug expenditure. Values lie in the 0–1 range,The closer to 1, the more important.

1. Model quality is an indicator that shows the goodness, importance and overall model fit of each cluster, and it is obtained as the output of the SPSS Modeler 18.0 software package when modeling is performed. Values lie in the 0–1 range, the closer to 1, the better the model quality.
2. For RFM model, the accuracy of the prediction model of C5.0 and CART algorithm is 100 %, which is higher than CHAID (94.48 %) and QUEST (99.32 %). Choose C5.0 and CART algorithm.
3. For dRFM Model, the accuracy of the prediction model of the C5.0 algorithm is 100 %, which is higher than CHAID (95.74 %), CART (98.71 %), and QUEST (98.26 %). Choose C5.0 algorithm.
4. C5.0 algorithm is higher than CHAID, CART and QUEST algorithm in terms of the accuracy of predicting RFM model and dRFM Model, and finally choose C5.0 algorithm.

Table 4
Two rounds of K-means clustering results of RFM model and dRFM model.

Model type	Model times	Sample	Model quality[1]	Importance of predictive variables[2]	Clustering result
RFM model[3]	Round 1	171107	0.7054	R = 1, F = 1, M = 0.3787	3 clusters of patients: Cluster 1 (43,978 sets): R = 47.07, F = 7.55, M = 632.93 Cluster 2 (29,274 sets): R = 94.76, F = 6.37, M = 549.44 Cluster 3 (97,855 sets): R = 8.14, F = 17.91, M = 673.19
	Round 2	97855	0.7096	R = 1, F = 1, M = 0.1385	3 clusters of patients: Cluster 1 (20.6 %, 20,188 sets): R = 21.08, F = 7.18, M = 734.07 Cluster 2 (56.1 %, 54,927 sets): R = 2.39, F = 26.02, M = 642.42 Cluster 3 (23.2%, 22,740 sets): R = 10.52, F = 7.84, M = 693.45
RFM +dRFM model[4]	Round 1	171107	0.7054	R = 1, F = 1, M = 0.3787	3 clusters of patients: Cluster 1 (43,978 sets): R = 47.07, F = 7.55, M = 632.93 Cluster 2 (29,274 sets): R = 94.76, F = 6.37, M = 549.44 Cluster 3 (97,855 sets): R = 8.14, F = 17.91, M = 673.19
	Round 2	97855	0.6482	d = 1, R = 1, F = 1, M = 0.2866	3 clusters of patients: Cluster 1 (Potential patients, 24.2 %, 23,643 sets): d = 90.17 %, R = 19.06, F = 8.14, M = 767.70 Cluster 2 (High-value patients, 24.5 %, 24,011 sets): d = 31.83 %, R = 6.66, F = 10.70, M = 652.26 Cluster 3 (Basic patient, 51.3 %, 50,201 sets): d = 94.24 %, R = 3.70, F = 25.96, M = 638.68
dRFM model[5]	Round 1	171107	0.6623	d = 1, R = 1, F = 1, M = 0.3256	3 clusters of patients: Cluster 1 (89,591 sets): d = 94.54 %, R = 14.11, F = 18.34, M = 679.74 Cluster 2 (44,893 sets): d = 91.37 %, R = 81.40, F = 7.32, M = 583.07 Cluster 3 (36,623 sets): d = 32.31 %, R = 19.70, F = 8.18, M = 620.36
	Round 2	126214	0.6956	d = 1, R = 1, F = 1, M = 0.1838	3 clusters of patients: Cluster 1 (70.3 %, 88,724 sets): d = 94.83 %, R = 14.22, F = 18.41, M = 679.64 Cluster 2 (7%, 8,832 sets): d = 24.64 %, R = 52.59, F = 3.00, M = 536.86 Cluster 3 (22.7 %, 28,658 sets): d = 35.64 %, R = 9.07, F = 9.86, M = 648.19

Note:
1. Model quality is an indicator that shows the goodness, importance and overall model fit of each cluster, and it is obtained as the output of the SPSS Modeler 18.0 software package when modeling is performed. Values lie in the 0–1 range, the closer to 1, the better the model quality.
2. R:Recency, F:Frequency, M:Average medical cost per visit, d:Percentage of drug expenditure. Values lie in the 0–1 range,The closer to 1, the more important.
3.RFM model: For China's public welfare hospitals and China's infectious disease hospitals, the proportion of medicines is a negative value. This model does not consider the cost of medicines in the medical expenses, and does not exclude the cost of medicines. Therefore, the RFM model is not applicable to this study.
4.RFM + dRFM model: After 2 rounds of K-means clustering (the first round uses RFMmodel, the second round uses dRFM model), patients are divided into 3 categories, which are in line with the classification of patients in the Chinese Infectious Disease Hospital.
Cluster 1 (Potential patients, 24.2 %, 23,643 sets): d = 90.17 %, R = 19.06, F = 8.14, M = 767.70 Cluster 2 (High-value patients, 24.5 %, 24,011 sets): d = 31.83 %, R = 6.66, F = 10.70, M = 652.26 Cluster 3 (Basic patient, 51.3 %, 50,201 sets): d = 94.24 %, R = 3.70, F = 25.96, M = 638.68.
5. dRFM model: The most recent consultation month (Recent consultation month) of the patients was assigned to 9 months ago. All patients were lost. dRFM model is also not applicable to this study.

Table 5
Value of 3 types of patients with RFM model + dRFM model after 2 rounds of K-means clustering.

3 clusters of patients	Patient type	Proportion of patients%	Sample	d%	R	F	M	Proportion of business revenue%[1]	Average business income per time (yuan)[2]	Business income created by each patient for the hospital (yuan)[3]	Business income created by this type of patient for the hospital (ten thousand yuan)[4]
Cluster 1[5]	Potential patients	24.2	23,643	90.17	19.06	8.14	767.7	9.83	75.46	614.28	1452.35
Cluster 2[6]	High-value patients	24.5	24,011	31.83	6.66	10.7	652.26	68.17	444.65	4757.71	11423.73
Cluster 3[7]	Basic patient	51.3	50,201	94.24	3.7	25.96	638.68	5.76	36.79	955.02	4794.27

Note: d:Percentage of drug expenditure, R:Recency, F:Frequency, M:Average medical cost per visit.
1. Proportion of business revenue% = 100 % - d%.
2. Average business income per time (yuan) = M × Proportion of business revenue%.
3. Business income created by each patient for the hospital (yuan) = Average business income per time (yuan) × F = M × Proportion of business revenue% × F.
4. Business income created by this type of patient for the hospital (ten thousand yuan) = (Business income created by each patient for the hospital × Sample) / 10000.
5. Cluster 1 (Potential patients): In the past 10 years, these patients (accounting for 24.2 %) have created 1,452.35 ten thousand yuan for research units.
6. Cluster 2 (High-value patients): In the past 10 years, these patients (24.5 %) have created 11,423.73 ten thousand yuan for the research unit.
7. Cluster 3 (Basic patient): In the past 10 years, this part of patients (51.3 %) has created RMB 4,794.27 ten thousand yuan for the research unit.

months) of the outpatients was more than 9 months [13]; thus, this model was not applicable. 3) The RFM + dRFM model: After 2 rounds of clustering analysis on 97,855 data elements 3 types of patient clusters were obtained: Cluster 1 (potential patients, mainly for foreign patients), Cluster 2 (high-value patients, mainly surgical patients, such as general surgery and orthopedic patients) and Cluster 3 (basal patients, mainly for patients with VHB, PTB, AIDS). In addition, 3 major predictive variables were found: the percentage of drug expenditure, time of last visit (in months) and frequency of visits. Therefore, the RFM + dRFM model could best reflect the type of patients in the hospitals for infectious diseases according to three types: basal patients, high-value patients and potential patients.

It can be found from Tables 4 and 5 that 97,835 data elements were finally clustered into 3 types of patients after 2 rounds of K-means cluster analysis: basic patients (accounting for 51.3 %, 50,201 sets, d = 94.24 %, R = 3.70, F = 25.96, M = 638.68), high-value patients (accounting for 24.5 %, 24,011 sets, d = 31.83 %, R = 6.66, F = 10.70, M = 652.26) and potential patients (accounting for 24.2 %, 23,643 sets, d = 90.17 %, R = 19.06, F = 8.14, M = 767.70), which best reflects the type of patients in the hospital for infectious diseases.

3.5. Model verification

After constructing the RFM + dRFM model, it was verified using the

Table 6
Results of C5.0 algorithm verification of the RFM + dRFM model analysis.

Item	Dataset[1]	Node[2]/model	Importance of predictive variables	Accuracy of prediction model (%)		
Training set[5]	70 % data	d[3] < = 61.782 /[mode: 2]	d = 0.5387	Accuracy	68,840	100 %
			R = 0.4598	Inaccuracy	3	0%
		d > 61.782/[mode: 3]	F = 0.0015	Total	68,843	—
Testing set[6]	30 % data	d[4] < = 60.553 /[mode: 2]	d = 0.5474	Accuracy	29,004	99.97 %
		d > 60.553/ [mode: 3]	R = 0.4526	Inaccuracy	8	0.03%
				Total	29,012	—

Note: R:Recency, F:Frequency, d: Percentage of drug expenditure. Values lie in the 0–1 range.
1. Dataset: (97855 Sample)70 % of the data were used as the training set to construct the model, and 30 % of the data were utilized as the testing set to verify the model.
2. Nodes: d:Percentage of drug expenditure.
3. For the Training set, 61.782 was used as a node to divide d into two categories: for model 2, d≤61.782 %; for model 3, d > 61.782 %.
4. For the Testing set, 60.553 was used as a node to divide d into two categories: for model 2, d≤60.553 %; for model 3, d > 60.553 %.
5. For the Training set, 61.782 % of medicines are a sub-node; the Accuracy of prediction model is 100 %.
6. For the Testing set, 60.553 % of drugs are a sub-node; the Accuracy of prediction model is 99.97 %.

C5.0 decision tree algorithm with the optimal number of the K-means clusters in the second round as the target and 3 important predictive variables as the input variables. The data were randomly divided into two sets (training and testing sets). The training accuracy of the prediction model was 100 %, while its testing accuracy was 99.97 %. The node of prediction was the percentage of drug expenditure, and there were two important predictive variables: the percentage of drug expenditure and the time of last visit (in months) (Table 6).

4. Discussion

In this work, we used the data from a research unit that was founded in 1914 as a hospital for notifiable communicable diseases in Shanghai, China, with 660 authorized beds. From Supplementary Table 1, it can be seen that the number of outpatient visits at the research unit per year is not large; such visits are mainly intended for the treatment of various types of liver disease and AIDS patients, and the patients have a single disease type.

The clustering variables processed by the K-means algorithm are numerical, and the distance between the points is defined as the Euclidean distance. The Kohonen algorithm uses the Euclidean distance, but the numerical variables need to be converted to be between 0 and 1. Two-step clustering can handle numerical and sub-type variables, using the log-likelihood distance (log-likelihood). Since the variables of the RFM/dRFM model are all numerical, the K-means algorithm is more suitable for this research [23] and it results in the best model quality (Table 1).

The K-means clustering algorithm determines the number of clusters by adjusting the parameters to obtain the best model quality. In this study, the optimal number of clusters was found to be 3 patient types (Table 2).

The RFM model refers to the patient's personal average cost. However, the average outpatient drug proportion is as high as 80 %; thus, the RFM model should be improved and updated to the dRFM model (Table 4).

In the following points, we present an analysis of the patient RFM/dRFM models. (1) The RFM model: In the first round of clustering analysis, the data of the patients who did not visit the doctor for the past 4 years or more were excluded, and a total of 97,855 data elements were retained. In the second round of clustering analysis, Cluster 2 represented the basal patients, with the highest RFM value (the time of last visit was within 3 months, and the frequency of visits was 26 times). Cluster 1 represented the lost patients who had not visited the hospital for over 21 months, and Cluster 3 included the potential patients whose time of last visit was over 11 months. Although the RFM model had a good quality, it did not consider the influencing factor represented in the high percentage of drug expenditure. This study followed the work of Min Li (2019) [18]. Since there is a high proportion of patients in hospitals for infectious diseases in China, and the cost of medicine is not included in the hospital business income, the types of patients could not be really distinguished, according to this model and it is thus not applicable.

(2) The dRFM model: In the first round of clustering, the data of the patients who did not see the doctor for more than 81 months were excluded, and 126,214 data elements of the patients who went to the hospital in the past 19 months were retained. In the second round of clustering the times of last visit of the patients in Cluster 1, 2 and 3 were 14.22 months, over 52.59 months and 9.07 months, respectively. The patients in all 3 clusters had no visits for over 3 consultation cycles (9 months) [13], as shown in Table 1. In 2019, the patients with a liver disease accounted for 58 %, and those with AIDS accounted for 17 %. The patients need to go to designated hospitals for follow-up visits every 1–3 months. With the dRFM model, the patients' most recent visit has exceeded 9 months. These can be judged as lost patients, and the type and value of the patients could not really be distinguished. Therefore, this model was also inapplicable.

(3) The RFM + dRFM model (Table 5): In the first round, the RFM model was used to cluster the patients into 3 types. The data of the patients who visited the hospital for the past 8 months were retained (97,855 data elements), while those of the patients who did not see the doctor for the past 4 years or more were excluded. In the second round, the dRFM model was applied for clustering, resulting in the following clusters: 1) Cluster 1 (potential patients) accounting for 24.2 %, with a total of 23,643 data elements. The percentage of drug expenditure was 90.17 %, the time of the last visit was more than 19.06 months, the frequency of visits was 8.14 times, and the average medical cost per visit was 767.70 yuan (RMB). Each patient of this type created 614.28 yuan for the research unit, with a total of 1,452.35 ten thousand yuan (Table 5). Since this cluster of patients had a high percentage of drug expenditure, it was speculated that these patients may have diseases such as drug-induced hepatitis, fatty liver, gastritis, pulmonary infection or sleep disorders. It is particularly necessary to improve the comprehensive diagnosis and treatment in the hospitals for infectious diseases. Meanwhile, these patients were included in the Health Management Center (HMC) and they received annual personalized health examinations. 2) Cluster 2 (high-value patients) accounting for 24.5 %, with a total of 24,011 data elements. The percentage of drug expenditure was 31.83 %, the time of the last visit was about 6.66 months, the frequency of visits was 10.70 times, and the average medical cost per visit was 652.26 yuan. Each patient of this type created 4,757.71 yuan for the research unit, with a total of 11,142.73 ten thousand yuan (Table 5). Optimizing the medical services makes these patients and their family members important customers of the HMC. 3) Cluster 3 (basal patients) accounting for 51.3 %, with a total of 50,201 data elements. The percentage of drug expenditure was 94.24 %, the time of the last visit was about 3.70 months, the frequency of visits was 25.96 times, and the average medical cost per visit was 638.68 yuan. Each patient of this type created 955.02 yuan for the research unit, with a total of 4,794.27 ten

Summary Table:

- The hospitals for infectious diseases are facing challenges regarding funding and revenue, classifying the patients based on their potential income can help improve hospital management
- The classical recency-frequency-monetary (RFM) model is widely applied to cluster the patients, but it cannot be applied in all cases
- Applying the modified RFM model can cluster the patients in hospitals for infectious diseases in China
- Classifying the patients can help improve the decision-making and enhance the quality of the provided medical services

thousand yuan (Table 5). Treating such patients is the commonweal and social responsibility of the hospitals for infectious diseases. Therefore, it is necessary to keep the basal patients, attract high-value patients and retain the potential ones.

Regarding the decision-making algorithm, the input variables of the C5.0 algorithm can be typed variables or numerical variables, and the output variables are also typed. The decision tree branch criterion is determined based on the information gain rate to find the best grouping variable and split point. The CART classification regression tree can only build a binary tree, the CHAID algorithm needs to preprocess input variables, and the QUEST algorithm also builds a binary tree; thus, the C5.0 algorithm is more suitable for this work [23] (Table 3).

The model verification performed using the C5.0 decision tree algorithm resulted in a model accuracy rate of 99.97 % (Table 6), so the model was reliable. In this work, we first used unsupervised learning, represented by the K-means clustering algorithm to gain insights from the data. Then, we verified the obtained model using supervised learning, represented by the C5.0 algorithm, by comparing the output with the labels to obtain the correctness of the model.

Since the data of this study span over 10 years, the first round of clustering analysis retained the data of the patients for the past 8 months. In the second round of clustering analysis of the dRFM model, the drug ratio and the most recent visit time could be used as indicators to divide the patients into 3 categories: basic patients, potential patients and high-value patients. As a result, this classification truly reflects the types of patients in the infectious disease hospitals in China and is also suitable for the Chinese public welfare hospitals that need to provide services for regular dispensing and inspecting. Besides, the patients with infectious diseases that are prone to adverse drug reactions and drug resistance problems can be provided with expert outpatient services based on the treatment plan service that is adjusted for the disease development [4,10–15,24–28]. In addition, the clinical physicians and pharmacists can provide guidance regarding the adverse drug reactions in outpatient clinics to continuously improve the quality of treatment for the patients.

5. Limitations and future work

This study did not analyze the 97,855 data elements based on gender and age. In the follow-up study, an in-depth analysis can be carried out, especially on the 50,201 basic patients. Such follow-up study can investigate the treatment cycle, medical insurance type and payment amount and predict their loss based on the patient's medical behavior.

6. Conclusion

This study investigated the optimal clustering model for the patient CRM analysis using the modified RFM (dRFM) model to classify the patients from a hospital of infectious diseases in China into three clusters based on the drug expenditure and the time they spend during their visit to the hospital. This could help with the hospital management and decision making and enhance the quality of the provided medical services.

Declaration of Competing Interest

No competing interests exist.

Acknowledgments

This work was funded by the 2016 key projects in the biomedical field of Shanghai Science and Technology Commission (1641953800), hospital management research fund of Shanghai Hospital Association (201701005 and 201801092) and clinical management optimization project of Shanghai Shenkang Hospital Development Center (shdc12018618). The authors would like to thank Professor Zhou Dequn of Nanjing University of Aeronautics and Astronautics, College of Economics and Management and Professor Jiang Xiaofei from Huashan Hospital Affiliated to Fudan University, China for giving guidance on this work.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2020.104373.

References

[1] H.K. Wong, C.K. Lee, Pivotal role of convalescent plasma in managing emerging infectious diseases [published online ahead of print, 2020 Apr 2], Vox Sang (2020), https://doi.org/10.1111/vox.12927.
[2] National Health Commission of the PRC, China Health Statistics Yearbook, M]. Peking Union Medical College Press. Beijing, August 2019 (first Edition), 407 Pages with 700,000 Words (ISBN code 9787567913233), 2019.
[3] K. B.Wang, L. Xiang, L. Kang, et al., Communicable disease mortality trends and characteristics of infants in rural China, 1996-2015, BMC Public Health 20 (1) (2020) 455, https://doi.org/10.1186/s12889-020-08486-y. Published 2020 Apr 6.
[4] S. Kim, V.A.A. de Los Reyes, E. Jung, Country-specific intervention strategies for top three TB burden countries using mathematical model, PLoS One 15 (April (4)) (2020), e0230964, https://doi.org/10.1371/journal.pone.0230964 eCollection 2020.
[5] G.J. Dore, B. Cowie, Global hepatitis B virus elimination by 2030: china is pivotal and instructive, Clin. Infect. Dis. (April) (2020) 7, pii: ciaa138. doi: 10.1093/cid/ciaa138. [Epub ahead of print].
[6] J. Chen, Y. Liu, S. Liu, D. Yuan, L. Su, L. Ye, F. Gong, Y. Gao, S. Baloch, X. Pei, HIV-1 drug resistance, distribution of subtypes,and drug resistance-associated mutations in virologic failure individuals in Chengdu, SouthwestChina, 2014-2016, Biomed Res. Int. 2020 (March) (2020) 5894124, https://doi.org/10.1155/2020/5894124, eCollection 2020.
[7] Y. Wang, C. Xu, J. Ren, W. Wu, X. Zhao, L. Chao, W. Liang, S. Yao, Secular seasonality and trend forecasting of tuberculosis incidence rate in China using the advanced error-trend-Seasonal framework, Infect. Drug Resist. 13 (March) (2020) 733–747, https://doi.org/10.2147/IDR.S238225, eCollection 2020.
[8] Y. Dong, L. Wang, D.P. Burgner, J.E. Miller, Y. Song, X. Ren, Z. Li, Y. Xing, J. Ma, S. M. Sawyer, Patton GC.INfectious diseases in children and adolescents in China: analysis of national surveillance data from 2008 to 2017, BMJ 369 (April) (2020) m1043, https://doi.org/10.1136/bmj.m1043, 2.
[9] X. Yin, G. Han, H. Zhang, et al., A real-world prospective study of mother-to-child transmission of HBV in China using a mobile health application (Shield 01), J. Clin. Transl. Hepatol. 8 (1) (2020) 1–8, https://doi.org/10.14218/JCTH.2019.00057.
[10] X. Wang, F. Li, X. Wang, X. Zhang, C. Liu, D. Wang, H. Wang, Y. Chen, Effects of different mark-up drug policies on drug-related expenditures in tertiary public hospitals:an interrupted time series study in Shanghai, China, 2015-2018, Biosci. Trends 14 (March (1)) (2020) 16–22, https://doi.org/10.5582/bst.2019.01350. Epub 2020 Feb 25.
[11] X. Xie, X. Jin, L. Zhang, et al., Trends analysis for drug utilization in county public hospitals: a sample study of the pilot area of health care reform in China, BMC

Health Serv. Res. 18 (1) (2018) 812, https://doi.org/10.1186/s12913-018-3614-8. Published 2018 Oct 23.
[12] Q. Meng, G. Cheng, L. Silver, X. Sun, C. Rehnberg, G. Tomson, The impact of China's retail drug price control policy on hospital expenditures: a case study in two Shandong hospitals, Health Policy Plan. 20 (3) (2005) 185–196, https://doi.org/10.1093/heapol/czi018.
[13] M.S. Islam, M.M. Hasan, X. Wang, H.D. Germack, M. Noor-E-Alam, A systematic review on healthcare analytics: application and theoretical perspective of data mining, Healthcare (Basel) 6 (2) (2018) 54, https://doi.org/10.3390/healthcare6020054. Published 2018 May 23.
[14] N. Mehta, A. Pandit, Concurrence of big data analytics and healthcare: a systematic review, Int. J. Med. Inform. 114 (2018) 57–65, https://doi.org/10.1016/j.ijmedinf.2018.03.013IF = 2.731.
[15] X. Zang, M. Zhang, S. Wei, W. Tang, S. Jiang, Impact of public hospital pricing reform on medical expenditure structure in Jiangsu, China: a synthetic control analysis, BMC Health Serv. Res. 19 (1) (2019) 512, https://doi.org/10.1186/s12913-019-4357-x. Published 2019 Jul 23.
[16] M.J. Tarokh, M. EsmaeiliGookeh, Modeling patient's value using a stochastic approach: an empirical study in the medical industry, Comput. Methods Programs Biomed. 176 (July) (2019) 51–59, https://doi.org/10.1016/j.cmpb.2019.04.021. Epub 2019 Apr 30.
[17] Z. Zare Hosseini, M. Mohammadzadeh, Knowledge discovery from patients' behavior via clustering-classification algorithms based on weighted eRFM and CLV model: an empirical study in public health care services, Iran. J. Pharm. Res. 15 (1) (2016) 355–367.
[18] Min Li, Study on the grouping of patients with chronic infectious diseases based on data mining, J. Biosci. Med. 7 (2019) 119–135.
[19] M.K. Poku, N.A. Behkami, D.W. Bates, Patient relationship management: what the U.S. Healthcare system can learn from other industries, J. Gen. Intern. Med. 32 (1) (2017) 101–104.
[20] P. Galetsi, K. Katsaliaki, S. Kumar, Values, challenges and future directions of big data analytics in healthcare: a systematic review, Soc. Sci. Med. 241 (2019), 112533, https://doi.org/10.1016/j.socscimed.2019.112533.
[21] E.W. Lee, Data mining application in customer relationship management for hospital inpatients, Healthc. Inform. Res. 18 (3) (2012) 178–185, https://doi.org/10.4258/hir.2012.18.3.178.
[22] S.M. Lee, A.K. Lee, I.S. Park, Data mining approach to model the diagnostic service management, Stud. Health Technol. Inform. 122 (2006) 903.
[23] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Second Edition, China Machine Press, 2011, pp. 211–322 (No. 22, Baiwanzhuang Street, Xicheng District, Beijing, the 9th printing of the first edition in April 2011).
[24] S. Gong, H. Cai, Y. Ding, et al., The availability, price and affordability of antidiabetic drugs in Hubei province, China, Health Policy Plan. 33 (8) (2018) 937–947, https://doi.org/10.1093/heapol/czy076.
[25] B. Karaismailoglu, N. Saltoglu, I.I. Balkan, B. Mete, F. Tabak, R. Ozturk, A prospective pharmacovigilance study in the infectious diseases unit of a tertiary care hospital, J. Infect. Dev. 13 (7) (2019) 649–655, https://doi.org/10.3855/jidc.11503. Published 2019 Jul 31.
[26] N. Saldanha, K. Runwal, C. Ghanekar, S. Gaikwad, S. Sane, S. Pujari, High prevalence of multi drug resistant tuberculosis in people living with HIV in Western India, BMC Infect. Dis. 19 (1) (2019) 391, https://doi.org/10.1186/s12879-019-4042-z. Published 2019 May 8.
[27] L. Xu, J. Chen, A.L. Innes, L. Li, C.Y. Chiang, Prescription practice of anti-tuberculosis drugs in Yunnan, China: a clinical audit, PLoS One 12 (10) (2017), e0187076, https://doi.org/10.1371/journal.pone.0187076. Published 2017 Oct 31.
[28] Z. Lu, W. Jiang, J. Zhang, et al., Drug resistance and epidemiology characteristics of multidrug-resistant tuberculosis patients in 17 provinces of China, PLoS One 14 (11) (2019), e0225361, https://doi.org/10.1371/journal.pone.0225361. Published 2019 Nov 21.

Min Li, PhD student of Nanjing University of Aeronautics and Astronautics, the main research direction is medical data mining and application.

Qunwei Wang, Doctor of management science and engineering, doctoral supervisor, professor, the main research direction is management system engineering.

Yinzhong Shen, Doctor of Internal Medicine, Chief Physician, Associate Professor.

TongYu Zhu, Doctor of Medicine, Chief physician, Professor, Doctoral supervisor.