Adherence predictor variables in AIDS patients: An empirical study using the data mining-based RFM model

Min Li (minliji@aliyun.com)
Qunwei Wang
Yinzhong Shen

Research

Keywords: antiretroviral therapy, adherence prediction, RFM, AIDS

DOI: https://doi.org/10.21203/rs.3.rs-41910/v2

Abstract

Background: Highly active antiretroviral therapy (ART) is still the only effective method to stop the disease progression in acquired immunodeficiency syndrome (AIDS) patients. However, poor adherence to the therapy makes it ineffective. In this work, we construct an adherence prediction model of AIDS patients using the classical recency, frequency and monetary value (RFM) model in the data mining-based customer relationship management model to obtain adherence predictor variables.

Methods: We cleaned 257305 diagnostic data elements of AIDS outpatients in Shanghai from August 2009 to December 2019 to obtain 16440 elements. We tested the RFM and RFm (R: recent consultation month, F: consultation frequency, M/m: total/average medical costs per visit) models, three clustering methods (K-means, Kohonen and two-step clustering) and four decision algorithms (C5.0, the classification and regression tree, Chi-square Automatic Interaction Detector and Quick, Unbiased, Efficient, Statistical Tree) to select the optimal combination. The optimal model and clustering analysis were used to divide the patients into two groups (good and poor adherence), then the optimal decision algorithm was used to construct the prediction model of adherence and obtain its predictor variables.

Results: The results revealed that the RFm model, K-means clustering analysis and C5.0 algorithm were optimal. After three rounds of k-means clustering analysis, the optimal RFm clustering model quality was 0.8, 10614 elements were obtained, including 9803 and 811 from patients with good or poor adherence, respectively, and five types of patients were identified. The prediction model had an accuracy of 100% with the recent consultation month as an important adherence predictor variable.

Conclusions: This work presented a prediction model for medication adherence in AIDS patients at the designated AIDS center in Shanghai, using the RFm model and the k-means and C5.0 algorithms.

Introduction

With no cure or effective vaccine so far, acquired immunodeficiency syndrome (AIDS), caused by the human immunodeficiency virus (HIV), represents one of the most serious infectious diseases in the world [1]. Currently, the only effective method to inhibit HIV replication is the highly active antiretroviral therapy (ART), developed in 1996. This method changed AIDS from a lethal disease to a treatable chronic infectious disease [2–4]. Furthermore, it can rebuild the immune system in AIDS patients to prevent death caused by various opportunistic infections [3,5]. However, the effectiveness of this treatment depends on the patients' adherence to the therapy [3]. High adherence can reduce drug-resistant strains and further transmission of such strains and decrease the AIDS-related opportunistic infections, mortality rate, incidence, treatment costs and disease burden [3,6–8].

Adherence was defined in 2003 by the World Health Organization as the extent to which a person's medication-taking behavior, dietary compliance and/or execution of lifestyle changes correspond with the agreed recommendations from a healthcare provider in the disease treatment and control. Basically, it is the degree to which a patient sticks to a treatment plan [9–10]. Despite the positive effects of adherence to therapy on the disease progression, poor adherence is a widespread problem in the world [1].

The minimum required adherence level to achieve the disease inhibition was shown to be 90% by Chen J [11]. The same indication was also presented by Kioko MT [12], Mbengue MAS [13] and Neupane S [14]. However, according to the current studies, adherence is poor in AIDS patients. Mbengue MAS [13] found that only 26.67% of the patients exhibit high adherence, Neupane S [14] found that only 87.4% of the patients have good adherence, Souza HC (2019) [15] found that only 52.5% of the patients have good adherence, whereas 33.3% of the patients have low adherence. Sagarduy JLY [16] found that only 82% of the patients are adherent to ART. Poor adherence might be caused by the lack of an objective, simple and easily understood operational marker for determining adherence. Therefore, developing a marker to monitor the clinical treatment and evaluate adherence in AIDS patients will help to give more time and attention to the patients with low adherence to improve it and increase the therapy effectiveness.

Currently, the following methods are used in China and other countries to determine adherence in AIDS patients: (1) self-reported evaluation questionnaires [9,11,13-14,16,17-20], (2) tablet, prescription and pharmacy records for drug count [21, 22], (3) plasma drug concentration monitoring [10,22-23] and (4) a mixture of self-reported questionnaires and plasma drug concentration monitoring. However, these methods have many drawbacks. Method (1) tends to be limited by patient-related subjective factors, and its reliability and validity tend to be questionable. Method (2) requires the inspector to have high professional knowledge, and electronic pillboxes are expensive and difficult to promote. As for method (3), it is limited by shortcomings such as the difficulty of daily collecting blood, high testing costs and large differences in the metabolism of individuals.

Given the drawbacks in the currently used measurement methods, Haberer JE [24] indicated the need to improve the adherence measurement in resource-limited settings to improve the ART adherence on a large scale. Therefore, there were some attempts to build adherence prediction models. Krumme AA [25] used the classical recency, frequency and monetary value (RFM) model [26] in the customer relationship management (CRM) theory [26–28] to predict that the adherence of the patients buying cardiovascular drugs from retail pharmacies was positively correlated with the number of store visits per month and the dollar amount per visit. Zare Hosseini Z [26] employed clustering analysis and decision algorithms in an RFM model-based data mining for the prediction using the data of patients in Iran and found that the adherence in hospital patients is associated with the recent consultation time, hospital treatment cycle, consultation frequency and total amount paid. This reveals that the RFM model is effective in predicting the variables for chronic disease medication adherence. However, there is still a need to investigate the application of this theory on adherence prediction in AIDS patients.

In this study, we used the classical RFM model in the data mining-based CRM theory [26–28] to obtain predictor variables for adherence in AIDS patients [25–26,29]. The presented model shows an efficiency to guide improvements in the adherence in AIDS patients.

Materials And Methods

Research design

This is an empirical study performed on the data exported from the hospital information system (HIS) of the only designated hospital in Shanghai for treating AIDS patients from August 2009 to December 2019. The subjects included in this study were all patients who were clinically diagnosed with AIDS, excluding HIV mother-to-child transmission and individuals taking antivirals for HIV post-exposure prophylaxis.The data was used to train and test an RFM model to get predictor variables for adherence in AIDS patients.

Data extraction and preprocessing

The data of AIDS outpatients from August 1, 2009, to December 31, 2019 were exported from the HIS of the research unit using the methods from literature [25-26,29-30]. The fields included the consultation time, patient's identification card number, gender, age, place of residence (local/no-local) and medical costs, for a total of 257305 data elements (16440 patients). Public hospitals in Shanghai in China implement a system wherein the actual name of the patient is used during consultation, and the identification number of the patient is an essential field. The SPSS 22.0 and SPSS Modeler 18.0 software were used for data analysis.

We cleaned the data and used the methods from literature [25–26,30] to expand some fields as follows: (1) The consultation time field was expanded to "recent consultation month", with December 2019 as the first month, November 2019 as the second month and so on until August 2009 as the 125th month; (2) The cumulative cost field in the patient's identification card was used to calculate the "total medical costs" field; (3) The consultation time and cumulative frequency in the patient's identification card were used to obtain the "consultation frequency" field; (4) The "total medical costs" field of each patient was divided by the "consultation frequency" field to obtain the "average medical costs per visit". These constitute one data element representing one person.

Variable generation and descriptive statistical analysis

Seven variables (recent consultation month, gender, age, total medical costs, consultation frequency, average medical costs per visit and place of residence) were generated. The factors of gender, age and place of residence were statistically analyzed to investigate good or poor acceptability, then four variables (recent consultation month, total medical costs, consultation frequency and average medical costs per visit) were used to describe these 16440 data elements as mentioned in literature [25–26, 29–30].

Finding the optimal RFM or RFm model, clustering analysis and decision algorithm

In this experiment, we tested the RFM and RFm models with several clustering analysis and decision algorithms to determine the best components to construct and evaluate the adherence prediction model. We employed methods from literature [25-26,29] and used the RFM model theory as follows: (1) The three fields of recent consultation month, consultation frequency and total medical costs were used for the RFM model [26]; (2) The three fields of recent consultation month, consultation frequency and average medical costs per visit were used for the RFm model [25,29]. Three clustering methods (K-means, Kohonen and two-step clustering) were used to construct the clustering models, in which four decision algorithms (C5.0, classification and regression tree (CART), Chi-square Automatic Interaction Detector (CHAID) and Quick, Unbiased, Efficient, Statistical Tree (QUEST)) were used in each model to construct several preliminary prediction models. From these models, we determined the optimal RFM(m) model [25-26,29,31-33], clustering analysis method [25,31-33] and decision algorithm based on the quality of the model [25,31-33] and stability of important predictor variables, which were used for the adherence prediction model experiment. The models in this study were calculated and built according to the SPSS Modeler 18.0 software package. The "model quality" in the output of the model is an indicator of its quality.

Validating the adherence prediction model and obtaining the variables

In this experiment [25-26,29,31-34], we used the optimal RFM(m) model and methods found in the previous experiment to construct the best clustering model, separate patients with good adherence from those with poor adherence and identify important variables for adherence prediction. The literature methods were used as references, and the optimal decision algorithm was employed, with good and poor adherence as targets. The important predictor variables in the best clustering model were utilized as the input variables, and the data underwent randomization and binning: we used 90% of the data as the training set to construct the adherence prediction model, and the remaining 10% was used as the test set to validate the model and finally obtain the adherence predictor variables.

Results

Descriptive statistical analysis of preprocessed data

The data preprocessing resulted in 16440 data elements, such that each element represented one patient. The mean recent consultation month of the 16440 AIDS patients was 15 months, and the median was 3 months. The mean total medical cost was 11000 RMB (China's currency), and the median was 8700 RMB. The average medical cost per visit was 900 RMB, and the median was 644 RMB. The mean consultation frequency was 15.65 visits, and the median was 13 visits. These four markers conform to a positively skewed distribution. The results are shown in Table 1. Since there was no statistically significant difference between good and bad adherence according to the factors of gender, place of residence and age (see Supplemental Table 6), only the descriptive statistical analysis was conducted.

Table 1. Consultation status of AIDS patients in 2009–2019

Variable[1]	Samples	Minimum (M)	Maximum (X)	Mean (E)	Standard deviation	Median	Skewness	Kurtosis	Quartile（0）	Quartile（25）	Quartile（50）	Quartile（75）
Recent consultation month (month)[2]	16440	1	125	14.99	23.58	3	2.111	4.24	0	1	3	20
Total medical cost (RMB)[3]	16440	1	666737.98	11315.29	17503.54	8740.51	16.88	496.76	0	2807.5475	8740.51	16973.0025
Average medical cost per visit (RMB)[4]	16440	1	28270	918.01	1107.12	644.47	4.98	57.29	0	444.4842	644.4745	831.4121
Consultation frequency (visits)[5]	16440	1	397	15.65	14.90	13	3.58	52.43	0	3	13	25

Note:

1. Variable:FRM（m）Model variables.
2. Recent consultation month (month):Recency
3. Total medical cost (RMB):Monetary
4. Average medical cost per visit (RMB):Monetary
5. Consultation frequency (visits):Frequency

The optimal RFM(m) model, clustering analysis and decision algorithm

The three markers in the RFM model were used as variables to construct 13 models (including five clustering models and eight prediction models). However, the predictor variables were unstable. The three markers in the RFm model were then used as variables to construct 27 models (including 7 clustering models and 20 prediction models), the predictor variables were stable, and the model was robust. Clustering analysis was used to construct 12 models, and the k-means clustering analysis was the most robust one. The decision algorithm was used to construct 28 models, and the C5.0 algorithm was robust and had high prediction accuracy. The results showed that the RFm model, k-means clustering analysis and C5.0 algorithm were optimal. The results are shown in Table 2.

Table 2. Preliminary experiment on RFM(m) models, clustering analysis, and decision algorithms in 16440 valid datasets obtained after cleaning

Model type	Clustering analysis				Decision algorithm							
	Clustering type	Number of models constructed	Model quality5	Predictor variable importance	Predictor variable importance				Prediction model accuracy (%)			
					C5.0	CHAID	CART	QUEST	C5.0[6]	CHAID	CART	QUEST
RFM model[1]	K-Means	1st round	0.8	R：1 M：1 F：1	R：0.9865 M：0.0067 F：0.0067	R：0.7402 M：0.2547 F：0.0052	R：0.9697 M：0.0152 F：0.0152	R：0.9689 M：0.0156 F：0.0156	99.96	93.55	99.77	97.07
		2nd round	0.9	R：1 M：1 F：1	M：1	M：0.7039 F：0.2961	▯	▯	99.98	99.9	▯	▯
		3rd round	0.7	R：1 M：1 F：1	M：0.5 F：0.5	M：1	▯	▯	99.98	99.93	▯	▯
	Two-step clustering	1st round	0.4	R：1 M：1 F：1	▯	▯	▯	▯	▯	▯	▯	▯
	Kohonen	1st round	0.4	R：1 M：1 F：1	▯	▯	▯	▯	▯	▯	▯	▯
RFM model[2]	K-Means[3]	1st round	0.5	R：1 M：1 F：1	R：0.6735 M：0.017 F：0.3095	R：0.6661 M：0.0429 F：0.2910	R：0.7172 M：0.0020 F：0.2808	R：0.7479 M：0.0017 F：0.2503	99.88	90.38	97.43	95.86
		2nd round	0.8	R：1 M：1 F：1	R：0.5918 M：0.4043 F：0.0039	R：0.5938 M：0.1768 F：0.2293	R：0.5814 M：0.4129 F：0.0057	R：0.5962 M：0.3999 F：0.0039	99.96	96.66	98.06	98.33
		3rd round	0.6	R：1 M：1 F：0.37	R：0.7258 M：0.0451 F：0.1841	R：0.7708 M：0.2728 F：0.0014	R：0.971 M：0.0268 F：0.0022	R：0.6749 M：0.3245 F：0.0007	99.82	97.48	97.48	98.25
	Two-step clustering[5]	1st round	0.7	R：1 M：1 F：1	R：0.6864 M：0.2090 F：0.1046	R：0.6607 M：0.2962 F：0.0431	R：0.6283 M：0.2582 F：0.1135	R：0.5797 M：0.2624 F：0.1579	99.87	96.9	98.41	97.32
		2nd round	0.7	R：1 M：0.15 F：0.01	R：0.7058 M：0.2942	R：0.8428 M：0.1572	R：0.7351 M：0.2649	R：0.7398 M：0.2602	98.61	95.45	98.04	97.7
		3rd round	0.4	R：1 M：1 F：1	▯	▯	▯	▯	▯	▯	▯	▯
	Kohonen	1st round	0.4	R：1 M：1 F：1	▯	▯	▯	▯	▯	▯	▯	▯

Note: R：Recency, F：Frequency, M：Monetary. Values lie in the 0–1 range.

1. The predictor variables of the RFM model were either unstable, or could not be used for modeling in clustering analysis and decision tree algorithm. M:total medical costs.
2. The predictor variables of the RFm model in the decision algorithm were stable.m:average medical costs per visit.
3. The K-means clustering model was robust.
4. The accuracy of the two-step clustering in the C5.0 algorithm was lower than that of the k-means clustering model, and the quality of the model in the third round was low.
5. The "model quality" in the output of the model is an indicator of the quality of the model built.
6. The C5.0 algorithm prediction model had an accuracy of 99%.

The adherence prediction model and variables

After determining the optimal model, clustering algorithm and decision algorithm to be used, we used the R, F and m in the RFm model as the variables for three rounds of k-means clustering analysis, then used the C5.0 algorithm to construct and validate the adherence prediction model. The 16440 data elements underwent one round of k-means clustering analysis to remove the data of the patients who did visit within 24 months. The second round removed the data of the patients who did not visit within eight months. The third round resulted in the best model quality of 0.8, and 5 clusters representing 5 types of patients. Among these elements, 9803 (recent consultation month ≤ 3 months) were patients with good adherence, and 811 (recent consultation month > 3 months) were patients with poor adherence. Furthermore, two important predictor variables (recent consultation month and average medical costs per visit) were obtained. The results are shown in Table 3 and Table 4.

Table 3. Results of cleaned 16440 datasets after three rounds of k-means clustering analysis

Samples	Number of constructed models[1]	Clustering number (type)[2]	Model quality[3]	Predictor variable importance
16440	1st round	6	0.5	R□1 M□1 F□1
11585	2nd round	6	0.7	R□1 M□1 F□1
10614	3rd round	5	0.8	R□1 M□1 F□0.0563

Note: R□Recency, F□Frequency, M□Monetary. Values lie in the 0–1 range.

1. After 3 modeling times, the model quality reached the best.
2. The first round of modeling is better to cluster into 6 categories, the second round of modeling is better to cluster into 6 categories, and the third round of modeling is to cluster into 5 categories.
3. The "model quality" in the output of the model is an indicator of the quality of the model built.

Table 4 Clustering map of five types of patients (10614 datasets) after three rounds of k-means clustering analysis

cluster	cluster-4	cluster-1	cluster-3	cluster-2	cluster-5
label	good adherence	good adherence	good adherence	poor adherence	poor adherence
Sample(Patient ratio %)	4313(40.6%)	2966(27.9%)	2406(22.7%)	811(7.6%)	118(1.1%)
input	average medical costs per visit= 651.76	average medical costs per visit = 678.68	average medical costs per visit = 682.59	average medical costs per visit = 831.28	average medical costs per visit= 3733.77
	Recency = 1.00	Recency = 2.00	Recency = 3.00	Recency = 4.62	Recency =1.47
	Frequency = 20.75	Frequency = 21.63	Frequency =20.93	Frequency = 16.36	Frequency=16.67

Since the AIDS patients in the study unit can collect drugs for free from the designated hospital once every three months, and following the recommendation of Tarokh MJ [29] to designate three months as one consultation cycle, we decided to classify the patients as well adherent if the recent consultation month was between one to three months. As a result, the patients in Clusters 1, 3, 4 and 5 were classified as patients with good adherence. The patients in Cluster 5 had other underlying diseases. Therefore, the average medical costs per visit was relatively high. The patients in Cluster 2, who did not go for consultation for more than four months, were classified as poorly adherent. The C5.0 algorithm was employed, with good adherence and poor adherence as the targets, and the recent consultation month and average medical cost per visit as the input variables. Validating the adherence prediction model showed that the recent consultation month represented the adherence prediction model node. If recency ≤ 3, Model 1 has a good adherence. If recency □3, Model 2 has a poor adherence. Thus, there was only one important predictor variable: the recent consultation month. The accuracy of the prediction model was 100%. The results are shown in Table 5.

Table 5 C5.0 algorithm analysis results

Item	Data binning	Node/model[1]	Predictor variable importance	Prediction Model accuracy (%)	Amount of data (sets) indicating good/poor adherence
Training set	90% data	R ≤ 3 /[Model□1]	R□1	Correct 9582 100%	Good adherence 8846 0
		R > 3/ [Model□2]		Wrong 0 0	Poor adherence 0 736
				Total 9582 —	Total 9582
Test set	10% data	R ≤ 3 /[Model□1]	R□1	Correct 1032 100%	Good adherence 957 0
		R > 3/ [Model□2]		Wrong 0 0	Poor adherence 0 75
				Total 1032 —	Total 1074

Note: 90% of the data was used as the training set to construct the adherence prediction model, and the remaining 10% was used as the test set to validate the model

1. Nodes: Recency: R□Three months was used as a node to divide R into two categories: for Model 1, R ≤ 3 months; indicating good adherence; for Model 2, R > 3 months, indicating that poor adherence.

Discussion

In this study, we used 257305 consultation data elements that were generated for 16440 AIDS patients in 125 months (August 1, 2009 to December 31, 2019) to train and test an adherence prediction model.

First, we performed a preliminary experiment, referencing the classical RFM model [26,29] that is used in the CRM theory in various industries to determine the optimal model to be used for prediction. Zare Hosseini Z [26] used the total medical costs (RFM model), whereas Krumme AA [25] used the average cost per visit (RFm) model for the analysis. Aiming to have an overall assessment of the RFM/m models and different clustering analysis and decision algorithms, we constructed RFM and RFm models, and evaluated the predictor variable stability and clustering model robustness to find the best elements. The results showed that the RFm model was superior to the RFM model (Table 2). Furthermore, the k-means and two-step clustering models had good performance. However, in the C5.0 algorithm, the model prediction accuracy values of the first two rounds of the two-step clustering were slightly lower than that of the k-means clustering analysis (Table 2), and its clustering quality in the third round was only 0.4. Therefore, the k-means clustering was shown to be better than the other tested clustering methods. Compared with the CART, CHAID and QUEST algorithms, the C5.0 algorithm model was robust and had a prediction accuracy of 99%. Hence, it represented the best prediction model (Table 2).

Determining the suitable algorithms to be used for model selection and clustering went as follows. 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 normalized between 0 and 1. Two-step clustering can handle numerical and sub-type variables, using log-likelihood distance (log-likelihood). The variables of the RFM/m model were all numerical, so the k-means algorithm was more suitable for this research [34]. On the other hand, 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. CART classification regression tree can only build a binary tree, CHAID algorithm needs to preprocess the input variables, QUEST algorithm also builds a binary tree, so the C5.0 algorithm was more suitable for this research [34].

In the adherence prediction model experiment, k-means clustering analysis was employed for three rounds of clustering on 16440 valid data elements to obtain a good clustering model and to achieve normal data distribution, since the variables were not normally distributed, and exhibited a positively skewed distribution. In the first and second rounds, the data of the patients who did not undergo consultation for more than 24 and 8 months were removed, respectively. In the third round, the model achieved the best clustering quality of 0.8, and the patients were divided into five categories, based on the timing of their last consultation. Since the frequency at which the AIDS patients collect their medicine from the designated hospital is once every three months, the patients were classified as well adherent if their consultation was within the last three months, and poorly adherent otherwise. Two important predictor variables were obtained: the recent consultation month and average medical costs per visit. The C5.0 algorithm was used for the model construction and validation, and the accuracy was 100%. Finally, the only important predictor variable for adherence was the "recent consultation month" represented by R, and the node of the adherence prediction model was "$R \leq 3$;$R > 3$".

The obtained adherence predictor variable was reliable, which once again validates the rationality of providing free drugs to AIDS patients in the study unit once every three months during follow-up consultations, instead of the other used options of two weeks or one month [31]. This variable can be used as an important predictive marker for adherence to guide the patients towards increasing their adherence. It is simpler and has better operability compared with the methods previously used.

Since ART is a lifelong treatment, the AIDS patient services should be expanded and simplified [35]. Here we present the following recommendation in this matter. In China, the actual names are used for mobile phone, telephone and internet users; thus, we recommend that the hospitals construct an information management platform with the following tasks: (1) When the patient's mobile phone opens the information management platform, a "drug administration mobile phone check in" pop-up automatically appears. A mobile phone SMS or a fixed telephone line call reminds the patient to take their medicines if they did not check in; (2) Every time the patient comes to take the free drugs, the workstation automatically calculates the next date for drug collection and continuously pushes reminding information three working days before the next consultation date [17, 36]. If the patient does not have a fixed consultation schedule, the system automatically provides feedback to the workstation and activates manual telephone calls.

This study had some limitations. On the one hand, there is no cross-sectional study for validation. On the other hand, the data used in this work come from a single center in Shanghai. Expanding the study by including multi-center data will enrich the conclusions.

Conclusion

This study constructed and validated a prediction model for medication adherence in AIDS patients, using the RFm model and the k-means and C5.0 algorithms. We showed the recent consultation month to be an adherence predictor variable. In future studies, in-depth tracking of adherence in AIDS patients in the study unit and collaboration with other designated medical institutions for treating AIDS patients in other cities in China will be conducted.

Abbreviations

AIDS: Acquired immunodeficiency syndrome, HIV: the human immunodeficiency virus, ART: active antiretroviral therapy, RFM: recency, frequency and monetary value model, CRM: customer relationship management theory, HIS: hospital information system, CART: classification and regression tree (CART), CHAID: Chi-square Automatic Interaction Detector, QUEST: Quick, Unbiased, Efficient, Statistical Tree.

Declarations

Acknowledgments

This work was funded by the 2016 key projects in 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.

Author's contributions

Min Li: Carried out the data modeling and wrote the manuscript.

Qunwei Wang: Wrote and polished the manuscript.

Yinzhong Shen: Provided guidance and revised the manuscript.

Funding

This work was funded by the 2016 key projects in 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).

Ethics approval and consent to participate

This study has been approved by the Shanghai Public Health Clinical Center Ethics Committee (Approval Letter 2016-S024-02).

Consent for publication

Not applicable.

Competing interest

The authors declare that they have no competing interests.

Author details

[1]Nanjing University of Aeronautics and Astronautics, College of Economics and Management, Nanjing, Jiangshu, China, 211106. [2]Shanghai Public Health Clinical Center, Fudan University Shanghai, China, 201508.

References

1. Prokofjeva MM, Kochetkov SN, Prassolov VS. Therapy of HIV Infection: Current Approaches and Prospects. Acta Naturae. 2016;8(4):23-32.
2. Rout SK, Gabhale YR, Dutta A, et al. Can telemedicine initiative be an effective intervention strategy for improving treatment compliance for pediatric HIV patients: Evidences on costs and improvement in treatment compliance from Maharashtra, India. PLoS One. 2019;14(10):e0223303. Published 2019 Oct 8. doi:10.1371/journal.pone.0223303
3. Escolano A, Dosenovic P, Nussenzweig MC. Progress toward active or passive HIV-1 vaccination. J Exp Med. 2017;214(1):3-16. doi:10.1084/jem.20161765
4. Farooq T, Hameed A, Rehman K, Ibrahim M, Qadir MI, Akash MS. Antiretroviral Agents: Looking for the Best Possible Chemotherapeutic Options to Conquer HIV. Crit Rev Eukaryot Gene Expr. 2016;26(4):363-381. doi:10.1615/CritRevEukaryotGeneExpr.2016018255
5. Autran B, Carcelain G, Li TS, et al. Positive effects of combined antiretroviral therapy on CD4+ T cell homeostasis and function in advanced HIV disease. Science. 1997;277(5322):112-116. doi:10.1126/science.277.5322.112
6. Egelund EF, Dupree L, Huesgen E, Peloquin CA. The pharmacological challenges of treating tuberculosis and HIV coinfections. Expert Rev Clin Pharmacol. 2017;10(2):213-223. doi:10.1080/17512433.2017.1259066
7. Bandera A, Colella E, Rizzardini G, Gori A, Clerici M. Strategies to limit immune-activation in HIV patients. Expert Rev Anti Infect Ther. 2017;15(1):43-54. doi:10.1080/14787210.2017.1250624
8. Shah M, Risher K, Berry SA, Dowdy DW. The Epidemiologic and Economic Impact of Improving HIV Testing, Linkage, and Retention in Care in the United States. Clin Infect Dis. 2016;62(2):220-229. doi:10.1093/cid/civ801
9. Kardas P, Lewek P, Matyjaszczyk M. Determinants of patient adherence: a review of systematic reviews. Front Pharmacol. 2013;4:91. Published 2013 Jul 25. doi:10.3389/fphar.2013.00091
10. Petersen ML, LeDell E, Schwab J, et al. Super Learner Analysis of Electronic Adherence Data Improves Viral Prediction and May Provide Strategies for Selective HIV RNA Monitoring. J Acquir Immune Defic Syndr. 2015;69(1):109-118. doi:10.1097/QAI.0000000000000548
11. Chen J, Zhang M, Shang M, Yang W, Wang Z, Shang H. Research on the treatment effects and drug resistances of long-term second-line antiretroviral therapy among HIV-infected patients from Henan Province in China. BMC Infect Dis. 2018;18(1):571. Published 2018 Nov 15. doi:10.1186/s12879-

018-3489-7

12. Kioko MT, Pertet AM. Factors contributing to antiretroviral drug adherence among adults living with HIV or AIDS in a Kenyan rural community. Afr J Prim Health Care Fam Med. 2017;9(1):e1-e7. Published 2017 Jul 31. doi:10.4102/phcfm.v9i1.1343
13. Mbengue MAS, Sarr SO, Diop A, Ndour CT, Ndiaye B, Mboup S. Prevalence and determinants of adherence to antiretroviral treatment among HIV patients on first-line regimen: a cross-sectional study in Dakar, Senegal. Pan Afr Med J. 2019;33:95. Published 2019 Jun 10. doi:10.11604/pamj.2019.33.95.17248
14. Neupane S, Dhungana GP, Ghimire HC. Adherence to antiretroviral treatment and associated factors among people living with HIV and AIDS in CHITWAN, Nepal. BMC Public Health. 2019;19(1):720. Published 2019 Jun 10. doi:10.1186/s12889-019-7051-3
15. Souza HC, Mota MR, Alves AR, et al. Analysis of compliance to antiretroviral treatment among patients with HIV/AIDS. Rev Bras Enferm. 2019;72(5):1295-1303. Published 2019 Sep 16. doi:10.1590/0034-7167-2018-0115
16. Sagarduy JLY, López JAP, Ramírez MTG, Dávila LEF. Psychological model of ART adherence behaviors in persons living with HIV/AIDS in Mexico: a structural equation analysis. Rev Saude Publica. 2017;51:81. Published 2017 Sep 4. doi:10.11606/S1518-8787.2017051006926
17. Endebu T, Deksisa A, Dugasa W, Mulu E, Bogale T. Acceptability and feasibility of short message service to improve ART medication adherence among people living with HIV/AIDS receiving antiretroviral treatment at Adama hospital medical college, Central Ethiopia. BMC Public Health. 2019;19(1):1315. Published 2019 Oct 21. doi:10.1186/s12889-019-7687-z
18. Vilela Á, Bach P, Godoy P; Grupo de ITS de Lleida. Cumplimiento del estudio de contactos de personas diagnosticadas de VIH/ITS en las comarcas de Lleida [Compliance with the partner notification of HIV/STI patients in the counties of Lleida]. Rev Esp Salud Publica. 2019;93:e201912096. Published 2019 Dec 2.
19. Johnson RJ. Tactile Contact as a Marketing Tool for Improving an HIV/STD Education Program's Compliance / Retention with Crack Cocaine Users. Psychol Ment Health Care. 2020;4(1):067.
20. Dagnew AB, Tewabe T, Birhie A, et al. Factors Associated with Compliance with World Health Organization-Recommended Infant-Feeding Practices by Mothers with HIV Infection in Northwest Ethiopia. Curr Ther Res Clin Exp. 2019;91:39-44. Published 2019 Oct 18. doi:10.1016/j.curtheres.2019.100568
21. Brojan LEF, Marca LM, Dias FA, Rattmann YD. Antiretroviral drug use by individuals living with HIV/AIDS and compliance with the Clinical Protocol and Therapy Guidelines. Einstein (Sao Paulo). 2020;18:eAO4995. Published 2020 Feb 17. doi:10.31744/einstein_journal/2020AO4995
22. Dilworth TJ, Klein PW, Mercier RC, Borrego ME, Jakeman B, Pinkerton SD. Clinical and Economic Effects of a Pharmacist-Administered Antiretroviral Therapy Adherence Clinic for Patients Living with HIV. J Manag Care Spec Pharm. 2018;24(2):165-172. doi:10.18553/jmcp.2018.24.2.165
23. Gilliland WM Jr, Prince HMA, Poliseno A, Kashuba ADM, Rosen EP. Infrared Matrix-Assisted Laser Desorption Electrospray Ionization Mass Spectrometry Imaging of Human Hair to Characterize Longitudinal Profiles of the Antiretroviral Maraviroc for Adherence Monitoring. Anal Chem. 2019;91(16):10816-10822. doi:10.1021/acs.analchem.9b02464
24. Haberer JE, Sabin L, Amico KR, et al. Improving antiretroviral therapy adherence in resource-limited settings at scale: a discussion of interventions and recommendations. J Int AIDS Soc. 2017;20(1):21371. Published 2017 Mar 22. doi:10.7448/IAS.20.1.21371
25. Krumme AA, Sanfélix-Gimeno G, Franklin JM, et al. Can purchasing information be used to predict adherence to cardiovascular medications? An analysis of linked retail pharmacy and insurance claims data. BMJ Open. 2016;6(11):e011015. Published 2016 Nov 9. doi:10.1136/bmjopen-2015-011015
26. Zare Hosseini Z, Mohammadzadeh M. 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. 2016;15(1):355-367.
27. Poku MK, Behkami NA, Bates DW. Patient Relationship Management: What the U.S. Healthcare System Can Learn from Other Industries. J Gen Intern Med. 2017;32(1):101-104. doi:10.1007/s11606-016-3836-6
28. Walker DD, van Jaarsveld DD, Skarlicki DP. Sticks and stones can break my bones but words can also hurt me: The relationship between customer verbal aggression and employee incivility. J Appl Psychol. 2017;102(2):163-179. doi:10.1037/apl0000170
29. Tarokh MJ, EsmaeiliGookeh M. Modeling patient's value using a stochastic approach: An empirical study in the medical industry. Comput Methods Programs Biomed. 2019;176:51-59. doi:10.1016/j.cmpb.2019.04.021
30. Sileo KM, Wanyenze RK, Kizito W, et al. Multi-level Determinants of Clinic Attendance and Antiretroviral Treatment Adherence Among Fishermen Living with HIV/AIDS in Communities on Lake Victoria, Uganda. AIDS Behav. 2019;23(2):406-417. doi:10.1007/s10461-018-2207-1
31. Min Li. Study on the Grouping of Patients with Chronic Infectious Diseases Based on Data Mining. Journal of Biosciences and Medicines.2019,7:119-135.
32. Lee EW. Data mining application in customer relationship management for hospital inpatients. Healthc Inform Res. 2012;18(3):178-185. doi:10.4258/hir.2012.18.3.178
33. Mohammadzadeh M , Hoseini ZZ , Derafshi H .A data mining approach for modeling churn behavior via RFM model in specialized clinics Case study: A public sector hospital in Tehran. Procedia Comput Sci. 2017;120:23▯ doi: 10.1016/j.procs.2017.11.206. Epub 2017 Dec 14.
34. Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques, Second Edition [M]. China Machine Press (No. 22, Baiwanzhuang Street, Xicheng District, Beijing, the 9th printing of the first edition in April 2011): 211-322.

35. Mutasa-Apollo T, Ford N, Wiens M, et al. Effect of frequency of clinic visits and medication pick-up on antiretroviral treatment outcomes: a systematic literature review and meta-analysis. J Int AIDS Soc. 2017;20(Suppl 4):21647. doi:10.7448/IAS.20.5.21647

36. Kebede M, Zeleke A, Asemahagn M, Fritz F. Willingness to receive text message medication reminders among patients on antiretroviral treatment in North West Ethiopia: A cross-sectional study. BMC Med Inform Decis Mak. 2015;15:65. Published 2015 Aug 13. doi:10.1186/s12911-015-0193-z