A Logistic Regression Model to Predict High Risk Patients to Fail in Tuberculosis Treatment Course Completion

Sharareh R. Niakan Kalhori, Mahshid Nasehi, Xiao-Jun Zeng

***Abstract*—One of the world health organization's plan priorities for tuberculosis control is improving DOTS quality. Develop a tool to predict the patients' treatment course destination to spot high risk cases for non-compliance can active DOTS more progressively. A valid logistic regression model to predict the probability of fail outcome in TB treatment course might be usable to determine the level of patients' supervision and support. Data from all registered TB patients in Iran's health system in 2005 were applied. Sex, age, weight, nationality, imprisonment and history of TB cases were identified as predictors of the given outcome (95% CL, P< 0.0001). Correlation Coefficient R^2 (X^2 (6) = 196, P < 0.0001), Wald statistic, and Hosmer-Lemeshow test showed the models` adequacy. Those TB patients who were old, male, low weight, foreign, prisoner and with history of relapse had higher probability to fail (0.5 ≤ OR ≤ 2.7, P ≤ 0.005). Optimal sensitivity and specificity were achieved for developed model. To support TB patients actively, this valid model can support health workers to realize how intensive their follow up should be in frame of DOTS as an expensive service in which it has currently become semi-passive.**

***Index Terms*— DOTS, Model, Logistic Regression, Prediction**

I. Introduction

Tuberculosis (TB) still is known as one of the principal infectious causes of disease and death worldwide, particularly in middle-and low-income countries. It is responsible for more years of healthy life lost than any other infectious disease except AIDS and malaria; only AIDS is currently responsible for more deaths [1]. Ensured cure is one of the expected aims of TB control program. The current approach is called DOTS standing for Directly-observed treatment; short-course strategy implemented in 187 countries around the world. Two main aims of DOTS as a main strategy of tuberculosis control are completion of therapy process for TB patients in order to be cured and also prevention the drug resistance development [2].

Non-completed treatment course and not entirely cured cases, not only do not remove themselves from the prevalent pool but are going to add more infected cases. Also, noncompliance has been identified as being associated with recurrence of TB [3]. Recurrent tuberculosis cause significant threats like MDR-TB. Moreover, MDR-TB and HIV-associated TB can be counted as problems resulted from TB treatment improperly [4]. MDR–TB cause higher fatality rate and compared to normal TB, it needs noticeable more expensive and prolonged treatment in addition to extensive patient supervision and support. On the other hand some studies like [5] raise this serious question that whether fifteen year of DOTS implementation is satisfactory enough to remain a major controlling strategy for TB. It has been reminded that, in spite of the impressive progress in DOTS implementing, there is currently an estimation of around 9 million people developed TB for the first time and 1.7 million people died with or from the disease globally [6]. DOTS as a core to the delivery of effective TB treatment needs some additional interventions to promote the quality of treatment and obtain the defined objectives. Promoting the quality of DOTS implementation has been considered in the second "Global Plan to Stop TB (2006–2015)" launched by WHO in 2006. Principally, in third element of DOTS` five components, it has been emphasized that all adult and paediatric TB cases affected either by pulmonary possessing sputum smear-positive, smear-negative or extra-pulmonary TB need to take standardized treatment with supervision and support. The main aim of this element of DOTS is ensuring the patients adherence to treatment and reduce the risk of the development of drug resistance TB. It has been highlighted that interrupting and stopping factors to complete treatment course completion should be recognized and depending on the local condition, support and supervision may be undertaken. Although it has been pointed that selected patient groups like prisoners, drug users, and some people with mental health disorders may need intensive support including DOT, there is no workable tool to determine other high risk TB cases requiring intensive DOTS. On the other hand, DOTS is a very expensive service; according to Floyd et al. [7] to control 80% of world's TB patients living in the 22 high-burden countries, WHO and the involved governments needed $ 1 billion annually between 2001 to 2005; plus $ 0.2

Manuscript received February 2. This work has been funded by Sharareh R. Niakan Kalhori`s scholarship from Iranian Ministry of Health and Medical Education.

Sharareh R. Niakan Kalhori is a PhD student at Machine learning group and optimization, school of Computer Science, University of Manchester, G33. Kilburn Building, Manchester, UK; Phone: 0044-161-3066993 Fax: 0044-1612756204; e-mail: shareareh.niakan@postgrad.manchester.ac.uk.

Mahshid Nasehi is PhD of Epidemiology and Head of Leprosy and Tuberculosis Office, Disease Control Department in Iranian Ministry of Health and Medical Education, No.68, Iranshahr st., Ferdowsi Sq., Tehran, Iran, email: mnasehi@yahoo.com.

Xiao-Jun Zeng is with the University of Manchester, Machine learning group and optimization, school of Computer science, Kilburn Building, Manchester, UK, email: x.zeng@manchester.ac.uk.

billion per year for TB control in low and lower-middle income countries .Thus, the resource gap is $300 million per year.

Applying predicting methods can play supportive role to reach the goal of TB treatment course completion through finding the cases with higher risk of treatment fail. Several studies have investigated to confirm the influential factors which are able of predicting non-completion of tuberculosis treatment course. Picon et al. [3] showed that HIV infection (RR= 8.04) and non-compliance treatment course (RR = 6.43) are two major factors for recurrence TB. Also, Tanguis and colleagues [8] via a comparative study among 2201 HIV positive TB patients in Barcelona found that the most significant factors were low-socio-economic level (OR = 3.65), living in neighbourhoods (OR=1.61) and history of TB (OR=1.61, CL=95%). 76.2% of cases were reported to be intravenous drug users. another study revealed that male sex, age ≥65 years, recent entry into UK for cases who were born abroad, and also recent living in suburban, pulmonary disease and drug resistance were significant failing factors in TB treatment completion [9]. In 2008, another investigation showed homelessness and old age with OR 9.91 and 1.02 respectively (95% CI) as predictors of high risk patient to fail [10]. Although by several researches the effect of several determinants on the fail result for TB treatment course completion has been exposed, predictive models to determine the probability of patients' success or fail have been lacking. Thus, the purpose of present study was to develop and validate a multiple logistic regression model to verify predictors of tuberculosis treatment completion outcome risk factors, how their relationship were with considered outcome ,and calculating the probability of successes in completing the course of treatment and getting cure. This statistical model can then be used in making decision about the level of close follow up for TB affected patients since DOTS has become a passive service practically and providing it actively for all TB patients seems ambitious.

II. Material and Methods

A. Data

A retrospective analysis was performed in 9672 subjects who were TB cases in the process of DOTS from registration to treatment in 2005. Novel professional software, 'Stop TB', developed for TB data collection was used. In present study patients' data as independent variables including Age, Sex, Weight, being in prison, Nationality, and patients' history related to tuberculosis (new, returned, imported, and abroad) were applied.

Table 1 shows the descriptive analysis of six attributes in detail. The dependent variable was failing in treatment course completion. There was no requirement of ethical approval for this retrospective study; likewise, patients' data were kept strictly anonymous and confidential.

B. Logistic Regression Model

Logistic regression has a wide range application in medical and biomedical research mainly to formulate models sorting the factors that might determine whether or not an outcome happens. The distinguishing feature of logistic regression model is that the outcome variable is binary or dichotomous. Usually, patients' data would be used to establish which attributions are influential in predicting the given outcome. They can then be measured for a new patient through placing in logistic regression model to calculate the probability of given outcome called (Y). The binary logistic regression equation is as follows when there is:

$$P(Y)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+...+\beta_n x_n+\varepsilon)}} \quad (1)$$

in which P(Y) is the probability of given outcome to be predicted and in this study was the probability of failing in treatment course completion, $ß_0$ which is a constance and x_1 to x_n symbolize as above mentioned independent variables and ß1...$ß_n$ are the models parameters play the role of constants for each attributer.

Having analyzed the prepared data in SPSS 14 and STATA 9.2 by using Enter method, the estimated coefficients and their associated standard errors were calculated in addition to the covariance and 95% Confidence Interval (CI) around the estimated probabilities of considered outcome for each case. MATLAB 7.4.0 was used to calculate cut-off probabilities.

Table 1. Descriptive analysis of predicting variable based on two considered classes (failed, not-failed)

	Sex		**Nationality**			**Prison**		**Case Type**				**Age**	**Weight**
	male	female	1	2	3	+	-	1	2	3	4		
Failed (828)	526	302	624	198	6	61	767	727	27	44	30	828	828
Not Failed (4008)	1894	2114	3323	664	21	113	3895	3672	96	160	80	4008	4008

For Case type: 1: new, 2: imported, 3: returned, 4: cure after absent.
For Nationality: 1: Iranian, 2: Afghani, 3: others including Iraqi, Pakistani and Central Asians.
For Prison'' +'': currently being in prison and ''– ''means vice versa.

III. RESULTS

C. Descriptive Analysis

The descriptive analysis of attributers has been presented in table 1 using the training set. Six applied variables were Tb patients characteristics including: sex, nationality, being in prison, type of TB (new, imported, returned and cured after being away from treatment), age (Mean: 46.6; Std: 21.1), and weight (Mean: 51.4, Std: 13.37).

B. Model Development

Logistic Regression analysis was carried out with failing in TB treatment course completion as the dependent variable and six attributions. The initial -2loglikelihood has dropped to 4231, revealing the rejection of the null hypothesis that the independent variables are not related to the dependent variable and also verifying the model improvement ($x^2(6) = 196, p < 0.0001$). Further, the model accounts for between 24.6 and 66.1% of the variance (Cox and Snell R^2) indicating a moderate association between dependent and independent variables.

In table 2, the column of coefficient list the constants of $\beta 1, \beta 2, ..., \beta 6$ for each related predictor in which it should be applied in logistic regression equation (Eq.1) to calculate P(Y). The value of coefficient for constant is $\beta 0$. Wald statistic test which is basically identical to the t-test in linear regression and the value of the regression coefficient divided by its associated standard error is presented in the fourth column. According to the value of Wald test, all given independent variables are highly significant predictors for given outcome. Besides, in the column of Exp (B), there is a list of odds ratio which presents that for each unit increase on independent variables, the odds of failing in treatment course completion would increase by a factor of the value mentioned in the column "Exp (B)" for each variable. The confidence interval for non of the variables cross 1 and it gives this confidence that the relationship between all predictors and failing found in this sample would be found in 95% of samples from the same population. Bases on the values of odd ratio for variable age, if patients get one year older, then the odds of failing would occur 1.5 times more. The risk of failing is more for male; when sex changes from male to female, the risk of failing become half. Furthermore, the value of odds ratio for weight is 1.280 which means that if the patient weight decreases by 1 kg, then the probability of failing in treatment course completion increase around 1.28 times more. The value of odds ratio for nationality is around 1.6 indicating that the risk of failing is 1.6 times more for foreign people including Central Asians, Iraqi, Pakistani, and Afghani respectively. TB cases in prison are more than twice and half as likely to quit the treatment course. Patients with history of relapse and returned TB, 1.175 more likely at risk of failing of treatment course completion rather than the new cases.

C. Model Validity

To check the model validity, Train and Test method which is composed of building a predictive model with training set (50% of whole data) and then validating it using test and validation samples (25% each). As shown in table 3, chi square for both training and testing model were significant 196 and 99 respectively when (X^2 (6) = 196 and 99, P<0.0001). The amount of Standard Error for both became less from original to final model. Also presented in table 3, Hosmer and Lemeshow`s goodness-of-fit test which tests the hypothesis that the observed data are significantly different from the predicted values by model was not significant for both training and testing sets (P=0.279).

Shown in table 4, overall percentage of correct classification were 82.05 and 81.64 for training and testing sets respectively revealing that how correct the model classified the given classes (failed or not). Furthermore, to see what proportion of true positives cases the model classified as being positive and what proportion of true negatives it classifies as being negative the sensitivity and the specificity of the model were tested; the result was satisfactory when area under curve was under 70% presented in figure 1.

Table 2. Special features of independent variables in the logistic regression equation

Variables	Beta	Standard Error	Wald	df*	Significance	Odd ratio
gender	-0.72	0.083	73.923	1	.000	0.489
age	0.02	0.002	38.193	1	.000	1.49
weight	-0.02	0.003	39.927	1	.000	1.280
nationality	0.5	0.084	36.046	1	.000	1.652
prison	0.99	0.172	32.944	1	.000	2.681
Case type	0.16	0.057	7.964	1	.005	1.175
Constant	-0.93	0.263	12.394	1	.000	0.396

* df =degree of freedom

Table 3. A comparison of adequacy measurements between training and testing models

Model	Chi-square $X^2_{df=6}$	SE (standard error)	Hosmer and Lemeshow Test	Wald statistics test
Training	196	0.038	9.80(8), P=0.279	1706(1), P < 0.0001
Testing	100	0.054	11.935(8), =0.154	862(1),P < 0.0001

To check the accuracy of model, the number of cases with correct prediction was calculated for each cut-off probabilities. As mentioned six attribution considered as X = [$x_1, x_2, x_3, ...x_6$]. And the patients` label denotes its type of outcome as following

$$r = \begin{cases} 1 & \text{failed patient in treatment course completion} \\ 0 & \text{successful patients in treatment course completion} \end{cases}$$

For each patient, there is an order pair (X,r)and the training set containing 4836 cases like:

$$[X = \{ x^t , r^t \}_{t=1}^{N=4836}]$$

Where t indexes different examples in the set, applying equation (1) we calculated P(Y) which is the estimated outcome by model developed by training set. Doing the same process for validation set, the P(Y) was calculated in each cut-off point from 0 to 1 with step of 0.01 to find the best point in which the highest number of correct classification has been conducted. As shown in Figure 2, there is a big change between cut-off 0.5 and 0.6, and the best cut-off is 0.6 and more where the most number of correct predicted outcomes was happened in these points.
For each above cut-off probability, the value of error was investigated as following:

$$E(Y \mid X) = \sum_{t=1}^{4836} (Y(x^t) \neq r^t)$$

In fact, as the value of cut-off which was along with rise in correctly classified cases increased, the value of error reduced dramatically. Shown in figure 3, at cut-off points of 0.6 and more the lowest error was observed and this can confirm the finding that the cut-off points 0.6 and more were known as the best points with highest level of correctly classification with lowest level of errors since in both plots the ideal status was happened exactly in a same points.

IV. DISCUSSION

Pursuing the idea of providing the DOTS in different levels to TB patients based on their situation is a necessary purpose requiring a tool to determine how likely they are going to fail in treatment course completion through using their data. This prediction would be carried out at commence of patient treatment in frame of DOTS. This study found out sex, age, nationality, weight, current imprisonment and type of TB cases as significant predictors for the given outcome ($P \leq 0.005$). Although former studies [8] [9] [10] [11] [12] verified nationality, age, imprisonment, and TB case type as influential factors for TB treatment course non-completion, patient's weight was new effective attributer (OR = 1.280, $P \leq 0.0001$) emphasizing that more patients` low weight, there is higher probability to be failed. In [10] females were known as high risk gender, however, present study (OR = 0.48, $P \leq 0.0001$) and [9] addressed males as more likely to be failed in treatment course completion. This study strongly confirmed the role of nationality and imprisonment since like [10] immigrants people who are mainly Afghani and Pakistani in Iran just under 2 times more likely to be failed ($P \leq 0.000$); like WHO indicates imprisonment as risk factor in [11], prisoners has been pointed out as patients were 2 and half time more likely at risk ($P \leq 0.000$). These six attributers are only the patients' demographical data plus simple TB history which can be provided easily. The easy to use gained model can straightforwardly transform the DOTS as passive to active services since the output of the model can be considered as an indicator revealing to what extent the levels of services should be supplied. For example, if there is more than 50% of failing in completing the treatment course, health workers may provide everyday home care. This service classification needs more guidelines to be defined based on the patient's status and compatible level of supplying services in frame of DOTS. Moreover, logistic regression estimates the probability of a dichotomous outcome and since there are other outcomes like treatment completion and getting cure, relapse or even dead [2], other modeling techniques [12] might be helpful .

Table4. Training and testing models classification

Model	Observed		Predicted			
			Failing		Percentage Correct	Overall Percentage correct
			N0	yes		
Training	Failing	Yes	3640	690	90.8	82.05
		No	220	604	73.3	
Testing	Failing	Yes	1809	201	90.00	81.64
		No	109	299	73.28	

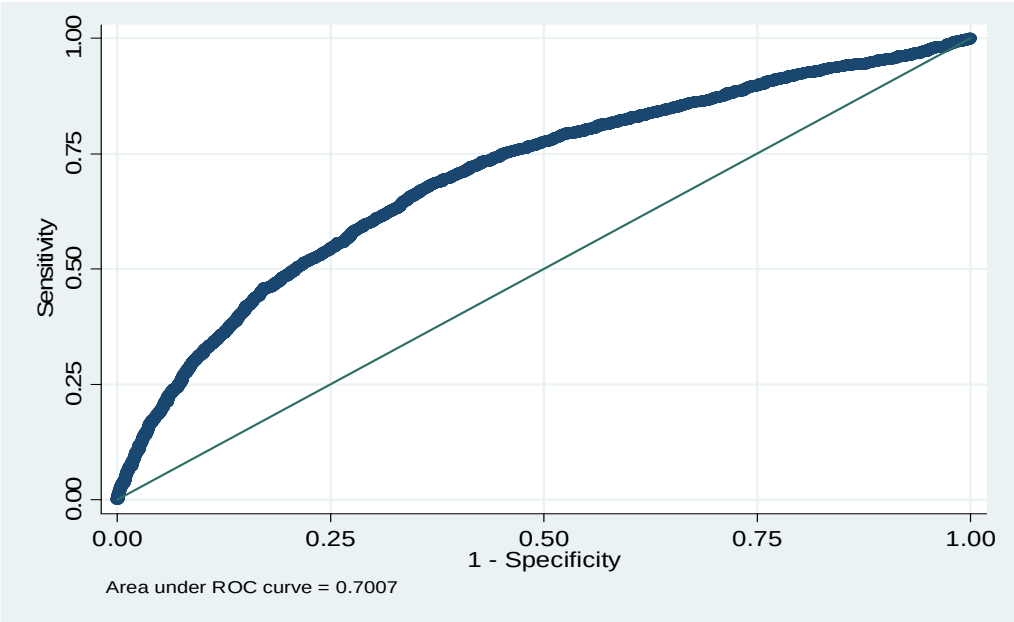

Figure1. Roc curve of sensitivity and specificity of model developed by training data set

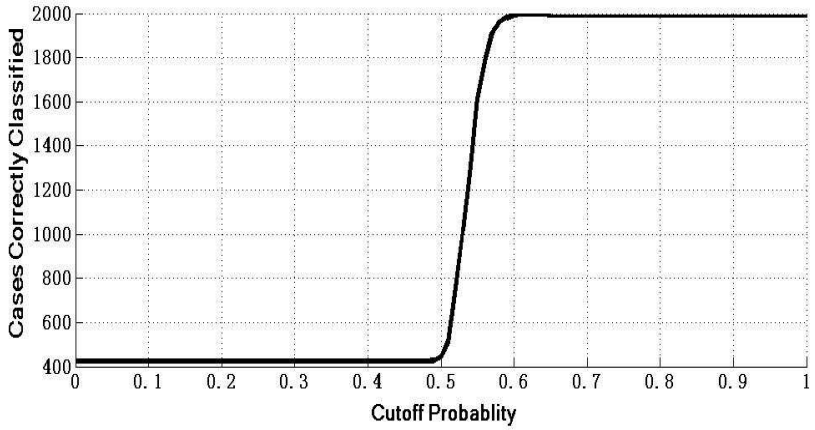

Figure2. Different cut-off probabilities with their related number of correctly classified TB cases to fail in treatment course completion

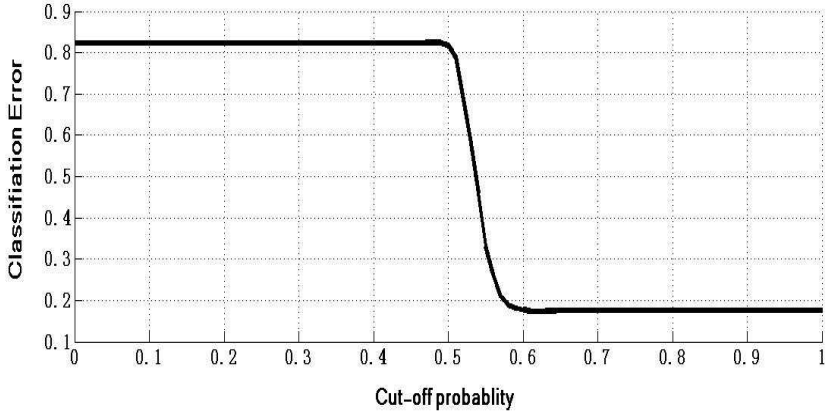

Figure3. Error for each cut-off probabilities to predict fail in TB treatment course completion

V. CONCLUSION

Since DOTS has currently become semi-passive service, there was a tool-requirement to detect high risk patient to fail in tuberculosis treatment course completion. This study developed a valid model applying patients data to define the probability of mentioned failure risk at commence of DOTS therapy in order to give patients supervision and support based on their status.

ACKNOWLEDGMENT

We are grateful to Iranian Ministry of Health and Medical Education for funding; Department of Tuberculosis and Leprosy Control of Iranian Ministry of Health for data access, helps and advice.

REFERENCES

[1] E.L. Corbett, C.J. Watt, N. Walker, D. Maher, B.G. Williams , M.C. Raviglione, C. Dye. ‘‘The growing burden of tuberculosis: global trends and interactions with the HIV epidemic’’. *Archives of Internal Medicine* , vol. 163(9), 2003, pp.1009-21.

[2] P.D.O. Davies. ‘‘The role of DOTS in tuberculosis treatment and control’’. *American journal of respiratory and critical care medicine, vol* 2(3), 2003, pp. 203-209.

[3] P.D. Picon, S.L. Bassanesi, M.L.A. Caramori, R.L.T. Ferreira, C.A. Jarczewski, P.R.B. Vieira. ‘‘Risk factors for recurrence of tuberculosis’’. *Jornal brasileiro de pneumologia*, vol 33 (5), 2007, 572-578.

[4] A. Juzar. ‘‘The Many Faces of Tuberculosis Control and the Challenges Faced’’. *Business Briefing: US Respiratory Care*, 2005, pp. 1-4.

[5] Z. Obermeyer, J. Abbott-Klafter, C.J.L. Murray. ‘‘Has the DOTS Strategy Improved Case Finding or Treatment Success? An Empirical Assessment’’. *PloS ONE*, vol 3 (3), 2008, e1721.

[6] A. D. Harries,C.Dye,‘‘Tuberculosis (Centennial review)’’. *Annals of Tropical Medicine & Parasitology*, vol. 100(5, 6), 2006, pp. 415-431.

[7] K. Floyd, L. Blanc, M. Raviglione, J. Lee. ‘‘Resource Required for Global Tuberculosis Control’’. *Science,vol* 295, 2002,pp. 2040- 2041.

[8] H.G. Tanguis, J.A. Cayla, P. Garcia de Olalla, J.M. Jansa , M.T. Brugal. ‘‘Factors predicting non-complition of tuberculosis treatment among HIV-infected patients in Barcelona (1987-1997).’’*INT J TUBERC LUNG DIS* 2000; 4(1): 55-60.

[9] D. Antoine,C. E. French, J. Jones, J.M. Watson. ‘‘Tuberculosis treatment outcome monitoring in England, Wales and Northern Ireland for cases reported in 2001’’ *J Epidemiol Community Health* 2007; 61: 302–307.

[10] I. Baussano, E. Pivetta, L.Vizzini, F. Abbona, M. Bugiani. ‘‘Predicting tuberculosis treatment outcome in a low-incidence area’’ *Int J Tuberc Lung Dis*, 2008; 12(12):1441-8.

[11] World Health Organization. ‘‘The Stop TB Strategy, Document’’ *WHO/HTM/TB/2006.35*. Geneva: WHO.

[12] S. R. Niakan Kalhori , and X. Zeng, ‘‘Fuzzy logic approach to predict the outcome of Tuberculosis treatment course destination’’ *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2009*, WCECS 2009, 20-22 October, 2009, San Francisco, USA, pp. 774-778.