International Electronic Journal of Medicine

Int Electron J Med. 2019; 8(2):124-129

doi 10.34172/iejm.2019.09

Original Article

Determining the Factors Affecting the Survival of HIV Patients: Comparison of Cox Model and the Random Survival Forest Method

Nasim Karimi[1], Malihe Safari[2], Mohammad Mirzaei[3], Amir Kasaeian[4], Ghodratollah Roshanaei[5*], Tahereh Omidi[2]

[1]Clinical Research Development Unit of Shahid Beheshti Hospital, Student Research Center, Hamadan University of Medical Sciences, Hamadan, Iran.
[2]Student Research Center, Hamadan University of Medical Sciences, Hamadan, Iran.
[3]Deputy of Health, Hamadan University of Medical Sciences, Hamadan, Iran.
[4]Hematology-Oncology and Stem Cell Transplantation Research Center, Tehran University of Medical Sciences, Tehran, Iran.
[5]Modeling of Noncommunicable Diseases Research Center, School of Health, Department of Statistics, Hamadan University of Medical Sciences, Hamadan, Iran.

Abstract

Background: In recent years, sexually transmitted diseases such as AIDS have become an epidemic and are growing rapidly. Given the importance of controlling the disease in recent years, the awareness of the most important risk factors associated with patient survival is important. Therefore, this study aimed to determine the most important factors affecting the survival of HIV patients using the random survival forest (RSF) method.

Materials and Methods: In this retrospective study, medical records of 769 HIV patients in Hamadan Health Center from 1997 to 2017 were used to determine the most important factors in patient survival using Cox proportional hazards model and RSF method. The Brier score and C-index were applied to compare the Cox model and RSF method.

Results: Based on the results, 662 (86.1%) patients were male. The mean ± SD diagnosis age was 33.83 ± 9.63 years. Using Cox model, variables such as injection history, co-injection history, tuberculosis (TB) status, the first CD4 cell count, and the time of disease diagnosis until TB were determined to be variables affecting the survival of patients. According to the hazard ratio (HR), the risk of death for those with a history of injections was 12.328 times greater than that of non-injectors, and for those with TB, it was 13.565 times greater than that of non-TB patients. An increase in CD4 cell counts was associated with a decline in the risk of mortality. Based on the log-rank model, the variables such as the time until diagnosis of TB, the first CD4 cell count, ART, and history of co-injection had the highest impact on predicting the survival of HIV+ patients, respectively.

Conclusion: In case of the presence of many risk factors and the relationship between risk factors, the use of RSF offers a better performance in determining the influential survival factors as compared to Cox model which has limiting presumptions.

Keywords: HIV, AIDS, Random Survival Forest, Cox proportional hazards model

***Correspondence to**
Ghodratollah Roshanaei, Modeling of Noncommunicable Diseases Research Center, School of Health, Department of Statistics, Hamadan University of Medical Sciences, Hamadan, Iran.
Tel: +989122145528,
Email: gh.roshanaei@umsha.ac.ir

Received: October 20, 2019, Accepted: November 30, 2019, ePublished: December 15, 2019

Introduction

Currently, AIDS is the most serious threat for public health (1). This epidemic so far has devastated many individuals, families, and societies and has increasingly caused erosion of civil order and economic growth (2). More than 90% of those infected with AIDS live in developing countries, with 80% being infected through sexual relationships. This disease has been the cause of the death of more than 25 million people up to 2006. AIDS has been stated as one of the devastating pandemics in history and it is estimated that 6% of the world's population is infected with this virus (3). Currently, AIDS has no cure, but the mortality resulting from HIV has diminished thanks to administration of highly active anti-retroviral therapy (HAART) (1). Anti-retroviral therapy (ART) is useful in reducing the speed of process preventing AIDS in an HIV-positive patient and increasing their survival (4). Therefore, ART has transformed a very fatal infectious disease into a potentially chronic and controllable infection (5). Infection with HIV is associated with

gradual quantitative and qualitative reductions in CD4 cells. Therefore, the patient is at risk of catching many comorbid and opportunistic infections (6). Tuberculosis (TB) is one of the most common opportunistic infections in patients with HIV. HIV significantly increases the number of patients with TB, thus heightening the risk of mortality among the affected patients (7). Eradicating pandemic diseases such as AIDS is the third goal of the 17 objectives of sustainable development document, based on which all countries including Iran are committed to ending AIDS epidemic by 2030 (8). In this regard, the use of suitable statistical models can be effective in validating the identification of important prognostic factors and improving the accuracy of predicting patient survival.

In survival analysis, various regression models are used for predicting the probability of incidence of future events (9). Cox proportional hazards regression model is one of the most common models in identifying the potential risk factors of diseases. Several studies have used this model for determining the survival of patients with AIDS and HIV. Nevertheless, when using Cox model, some limitations such as the proportional hazards requirement, there is poor performance in complex models such as nonlinear and collinear effects of variables (10). Further, this model is not valid enough under conditions of high censor rate (11). Therefore, models should be used with fewer constraints. Random survival forest (RSF) is a nonparametric machine learning method which was developed by Ishwaran et al based on random forests (RF) (12). This model is used to address the problem of using the Cox model including concurrent assessment of complex effects and interaction effects between variables (10).

In different studies, the selection of covariates has been done in survival analysis and comparison of Cox and RSF models (11,13). Hence, RSF studies have a better performance compared to Cox model, and enjoy the ability to identify nonlinear effects automatically, while Cox model lacks this ability. On the other hand, when the number of predictors is low, RSF model underperforms compared to Cox due to sensitivity to confounding factors. Indeed, under such conditions, RSF is unusable and Cox is proposed instead (10).

Considering the efficiency and ability of different models to predict the factors affecting the survival of patients in different diseases, the aim of this model is to determine the predictive factors of AIDS patients based on Cox and RSF models and compare their accuracy.

Materials and Methods

This research has been performed as a retrospective cohort study to investigate the survival of patients with HIV in Hamadan province located in western Iran. For this purpose, 769 patients with HIV who had a medical file in the healthcare center of Hamadan province between 1997 and 2017 were studied. The information required in this study was extracted using checklists from the patient's file. The collected information included age at the time of diagnosis, gender, route of HIV transmission (injection, sexual, mother to baby, unknown), the number of CD4 cells, antiretroviral therapy, the duration of HIV diagnosis until TB, the date of diagnosing AIDS, date of death, cause of death, and date of latest news of patients. The time from disease diagnosis until death was considered as survival time. In this paper, to identify the variables affecting the survival time of HIV patients, Cox multivariate regression model and RSF method were used. RSF is a developed form of RF which is used for survival data with the right censor with the same principles of RF, possessing all of its important features (7). Random forest covers several trees based on a random sample with substitution. Generally, the RSF algorithm is as follows (12):

A number of B bootstrap samples are chosen from the main data. In every bootstrap sample, about %37 of the data are left that is known as out of bag (OOB) sample.

1. For every bootstrap sample, a survival tree is grown. In every node of the tree, q predictors (covariates) are randomly chosen for splitting. The node is divided into two daughter nodes using splitting criteria. The variable chosen for splitting is the one that creates the maximum difference in the survival of two daughter nodes.
2. The tree grows up to its maximum growth size. The last node is called the final node. The final node should not be less than $d_o>0$ (d_0 represents the number of intended events, which is death in this study).
3. For every tree, a cumulative hazard function (CHF) is calculated and then the mean of these CHFs reports the total CHF.
4. Using OOB data, the prediction error is calculated.

For splitting every node and creating the daughter nodes, log-rank splitting rule, log-rank score, and RSF have been used. The comparison of the accuracy of splitting rules is made through the prediction error, with lower values suggesting higher accuracy. The importance of every variable in the prediction is measured by VIMP index. Positive values suggest variables with predictive abilities (important value), while zero or negative values are those with no ability to predict (10, 14).

In order to compare the efficiency of the Cox proportional hazards model and RSF, two criteria called Brier score and C-index have been used. All analyses were performed by R 3.1.2 alongside random forest SRC and statistical packages for survival analysis.

Results

In this study, the survival of 769 patients with AIDS was investigated. Out of this number, 662 (86.1%) were male and 107 (13.9%) were female. The mean ± SD age of diagnosis of patients was 33.83 ± 9.63 years, with 88.4% of patients being younger than 45 years (with the range of

1-87). Further, 63.8% of patients had primary education while 36.2% had university degrees. Most of the patients were single (44.3%). Additionally, 45.6% were under ART treatment. Finally, 9.2% of patients concurrently suffered TB. Table 1 demonstrates the demographic characteristics of the patients.

First, using Cox proportional hazards regression model, the effect of influential factors for survival was determined, as presented in Table 2. Based on the results of Cox regression model, the variables of history of injection, co-injection, status of TB (Yes/No), the first CD4 cell count, and time of diagnosis until developing TB were identified as important and influential factors for the survival of patients. Based on the hazards ratio (HR), the mortality risk for those with a history of injection was 12.328 times greater than non-injection patients, and it was 13.565 greater for TB patients than non-TB individuals. The risk of mortality diminished with an increase in the CD4 cell count. Further, those with a history of co-injection had 0.122 greater risk of mortality compared to those without such history. Finally, with the increase in the time of disease diagnosis until TB, the risk of mortality diminished.

Table 1. Demographic and clinical characteristics of HIV+ patients

Variable	Value	Frequency	percent
Gender	Female	107	13.9
	Male	662	86.1
Diagnosis age	Below 45 years	680	88.4
	More than 45 years	89	11.6
Level of education	Elementary	491	63.8
	University	278	36.2
Marital status	Married	291	37.8
	Single	341	44.3
	Widowed	38	4.9
	Divorced	99	12.9
History of imprisonment	No	314	40.8
	Yes	455	59.2
Non-safe sexual behavior	No	622	80.9
	Yes	147	19.1
Relationship with opposite sex	No	666	86.6
	Yes	103	13.4
History of addiction	No	262	34.1
	Yes	507	65.9
History of injection	No	339	44.1
	Yes	430	55.9
History of co-injection	No	376	48.9
	Yes	393	51.1
ART	No	418	54.4
	Yes	351	45.6
TB status	No	698	90.8
	Yes	71	9.2
The first CD4 status	0-200	118	15.3
	201-350	111	14.4
	351-500	81	10.5
	More than 500	124	16.1
Status of mortality	No	300	39
	Yes	469	61
Time of disease diagnosis until developing TB (mean ± SD)		34.34 ± 56.44	

In order to compare Cox model with RSF models, RSF models were used based on the log-rank, RSF, and log-rank score, with the best ones being chosen based on the minimum error as the final model. Table 3 shows the error of the three models. According to the table, the log-rank method with the minimum error was identified as the best model among the three models. Figure 1 displays the important variables based on the degree of significance according to the log-rank rule. Accordingly, the variables of time of diagnosis until TB, the first CD4 cell count, ART, and history of co-injection were identified as the important variables in the survival of patients with AIDS. The error value for this rule is 16.30, which has a constant trend from 300 trees above. The important variables based on RSF and log-rank scores are presented in Figures 2 and 3.

In order to compare Cox and RSF model based on log-rank rule, Brier score and C-index were used, with the results provided in Table 4. According to this table, RSF model based on the log-rank rule was identified as the most suitable model among the models applied in this research for determining the important factors in the survival of patients with AIDS. Specifically, the variables of time of diagnosis until TB, the first CD4 cell count, ART, and history of co-injection were the important variables in predicting the survival of HIV+ patients, respectively.

Discussion

This study was done to compare Cox multivariate regression model and RSF models to determine the factors affecting the survival of HIV+ patients considering mortality as the final event. The aim of comparison was to select a model with greater accuracy and efficiency in identifying the factors affecting the survival of HIV+ patients. Accordingly, in RSF and Cox models, the effect of demographic, clinical, and laboratory factors was tested on survival. RSF model was performed using different splitting groups. Ranking of the important variables identified based on the log-rank method, log-rank score, and RSF indicated that many important variables are the same based on three rules. However, the insignificant variables are also common based on the mentioned three rules. As can be seen, the status of TB (Y/N) and the latest marital status were the least important. Based on the three rules used, the RSF model with the log-rank splitting rule had the minimum error. This finding is in line with the study by Datema et alin determining the influential factors for survival of patients with head and neck cancer. In this study, RSF models were compared with each other based on the error criterion. Accordingly, the model based on log-rank splitting rule was determined as the most suitable model in the set of RSF models (13).

Comparison between Cox multivariate model and RSF model based on the log-rank splitting rule was made using

Table 2. Estimation results of Cox proportional hazards model in determining the factors affecting mortality of HIV+ patients

Variable	Value	HR	CI for HR (95%)	p-value
Gender	Female	1	-	0.525
	Male	0.836	(0.481, 1.453)	
Diagnosis age	Below 45 years	1	-	0.496
	More than 45 years	1.216	(0.692, 2.138)	
Level of education	Elementary	1	-	0.939
	University	0.986	(0.688, 1.413)	
Marital status	Married	1	-	
	Single	1.029	(0.670, 1.581)	0.895
	Widowed	1.443	(0.722, 2.884)	0.299
	Divorced	1.562	(0.870, 2.810)	0.135
History of imprisonment	No	1	-	0.381
	Yes	1.280	(0.736, 2.225)	
Non-safe sexual behavior	No	1	-	0.116
	Yes	2.063	(0.836, 5.089)	
Relationship with opposite sex	No	1	-	0.184
	Yes	0.517	(0.196, 1.367)	
History of addiction	No	1	-	0.125
	Yes	0.542	(0.247, 1.186)	
History of injection*	No	1	-	<0.001
	Yes	12.328	(3.651, 41.621)	
History of co-injection*	No	1	-	<0.001
	Yes	0.122	(0.046, 0.322)	
ART	No	1	-	0.858
	Yes	0.858	(0.514, 1.433)	
TB status	No	1	-	<0.001
	Yes	13.565	(2.404, 28.735)	
The first CD4 status*	0-200	1	-	-
	201-350	3.487	(1.621, 7.500)	0.001
	351-500	2.802	(1.455, 5.396)	0.002
	More than 500	2.356	(1.197, 4.638)	0.013
Time of disease diagnosis until developing TB*		0.863	(0.846, 0.881)	<0.001

Table 3. Error Values Based on RSF

RSF Trees	Error Value
Log-rank	16.30
Log-rank score	24.53
Random	36.43

Table 4. The Brier Score and C-index for the RSF Models

Models	C-index	Brier Score
Log-rank	78.1	0.121
Cox	50.2	0.175

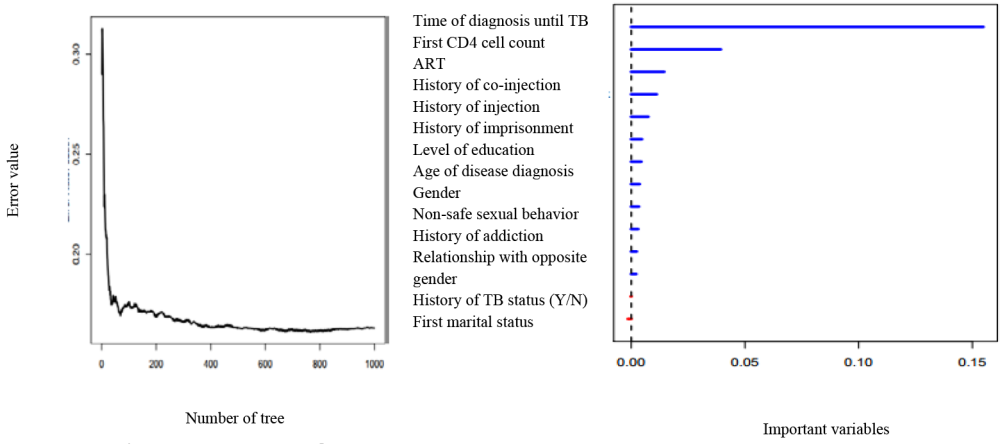

Figure 1. Important Variables Based on Long-Rank Splitting Rule.

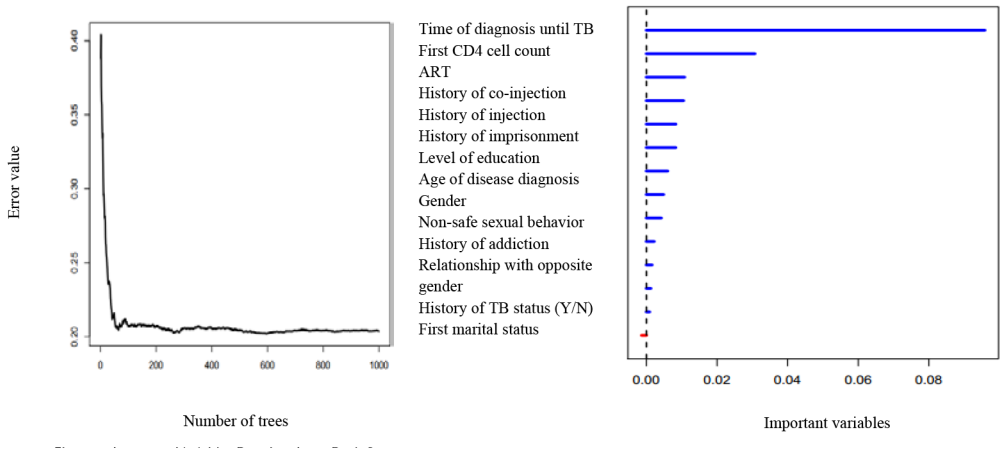

Figure 2. Important Variables Based on Long-Rank-Score.

c-index and Brier scores. Since for the log-rank and Cox, C-index was 78.1 and 50.2% and Brier score was 0.121 and 0.175, RSF model based on log-rank splitting rule was the most suitable model. Therefore, the important variables were identified based on this rule that had the maximum accuracy and efficiency in determining the survival of patients. Although many studies have not compared the efficiency of RSF and Cox models, Zhou et al investigated the accuracy of the mentioned models based on simulated data. They observed that RSF models are more accurate and flexible than Cox models, as they have no special assumption against the collinearity of variables and existence of nonlinear relations (15, 16).

According to the results of this study, the variables of time of disease diagnosis until TB, the first CD4 cell count, ART, and history of co-injection are the important variables in predicting the survival of HIV+ patients, respectively. The type of diagnosis until TB was identified as the most important factor in the mortality of patients. The World Health Organization has identified TB as the cause of death of 23% of AIDS patients. Therefore, this variable plays a significant role in the survival patients (16).

The first CD4 cell count was identified as the second most important predictive factor for the mortality of patients. Investigation of the effect of CD4 on the survival of patients indicated that the reduction in the CD4 cell count is associated with increased mortality rate, thus increasing the hazard ratio (HR) in patients. The results of many studies have suggested that the reduction in CD4 cell count plays a significant role in increasing the risk of HIV, TB, and AIDS-induced death (17). This finding is in line with previous studies (18, 19). Further, Cuong et al indicated that CD4 cell count less than 100 is an effective predictive factor for AIDS-induced death in patients (20). Based on the results of this study, the use of ART treatment was identified as one of the important variables for the survival of patients with HIV, causing increased survival of patients. Evidence has shown that ART consumption is associated with diminished mortality. On the other hand, some studies have found that older patients have a worse response compared to younger individuals (21, 22). The results of a study showed that ART leads to diminished HIV-induced mortality and increased CD4 cell count in patients with concurrent TB and HIV infection (23).

Co-injection has been one of the major causes of HIV in recent years (24), and it has been mentioned as its most common cause. Further, some injection addicts have AIDS. In this study, 65.9% of patients with HIV had a history of addiction.

This research had some limitations. Since it was a retrospective cohort study, the accuracy of the recorded information might have caused bias in the results and the reduction of validity. Further, to estimate the survival time, the precise date of developing the disease is not as clear as the date for other chronic diseases. Hence, in this study as with other survival studies, the duration of survival was considered as the time of diagnosis (i.e. the patient's referral) as the time of developing HIV infection.

Conflict of Interest Disclosure

The authors declared no conflict of interests.

Acknowledgement

Hereby, the Vice Chancellor of Research and Technology is highly appreciated.

Ethical Statement

This paper was derived from the research proposal approved by the Research Committee (9603161750) and the Ethics Committee (IR.UMSHA.REC.1396.218) of Hamadan University of Medical Sciences.

Authors' Contribution

MS and GhR designed and performed the research. Data analysis and manuscript preparation was performed by NK and AK. Data collection was done by MM and TO. All authors contributed to the final version of the manuscript and approved the final manuscript.

Funding/Support

This research was supported by the Vice-Chancellor of Research and Technology of Hamadan University of Medical Sciences.

Informed Consent

No need.

References

1. World Health Organization (WHO). HIV/AIDS. Available from: https://www.who.int/news-room/fact-sheets/detail/hiv-aids. Accessed February 2015.
2. Gayle HD, Hill GL. Global impact of human immunodeficiency virus and AIDS. Clin Microbiol Rev. 2001;14(2):327-35. doi: 10.1128/cmr.14.2.327-335.2001.
3. Albrecht H. Report from the 14th Retrovirus Conference. New data on HIV and viral hepatitis coinfection. AIDS Clin Care. 2007;19(5):41.
4. Beake S, McCourt C, Page L. Evaluation of One-to-One Midwifery Second Cohort Study: Report. London: Thames Valley University; 2001.
5. Pomerantz RJ, Horn DL. Twenty years of therapy for HIV-1 infection. Nat Med. 2003;9(7):867-73. doi: 10.1038/nm0703-867.
6. Mandell G, Dolin R, Bennett J. Mandell, Douglas, and Bennett's principles and practice of infectious diseases. Elsevier; 2009.
7. Hamidi O, Tapak M, Poorolajal J, Amini P, Tapak L. Application of random survival forest for competing risks in prediction of cumulative incidence function for progression to AIDS. Epidemiol Biostat Public Health. 2017;14(4):e12663-10. doi: 10.2427/12663.
8. United Nations General Assembly. Political Declaration on HIV and AIDS: On the Fast Track to Accelerate the Fight Against HIV and to Ending the AIDS Epidemic by 2030. New York: United Nations; 2016.
9. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. J Stat Softw. 2012;50(11):1-23. doi: 10.18637/jss.v050.i11.
10. Miao F, Cai YP, Zhang YT, Li CY. Is random survival forest an alternative to Cox proportional model on predicting cardiovascular disease? In: Lacković I, Vasic D, eds. 6th European Conference of the International Federation for Medical and Biological Engineering. Cham: Springer; 2015. p. 740-3. doi: 10.1007/978-3-319-11128-5_184.
11. Myte R. Covariate selection for colorectal cancer survival data: A Comparison case study between random survival forests and the cox proportional-hazards model. Umeå: Umeå University; 2013.
12. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2(3):841-60. doi: 10.1214/08-AOAS169.
13. Datema FR, Moya A, Krause P, Bäck T, Willmes L, Langeveld T, et al. Novel head and neck cancer survival analysis approach: random survival forests versus Cox proportional hazards regression. Head Neck. 2012;34(1):50-8. doi: 10.1002/hed.21698.
14. Hamidi O, Poorolajal J, Farhadian M, Tapak L. Identifying important risk factors for survival in kidney graft failure patients using random survival forests. Iran J Public Health. 2016;45(1):27-33.
15. Hajabdolbaqi M, Rasolinejad M, Talebi Taher M. National guidelines for the care and treatment of HIV in Iran. Tehran: Ministry of Health and Medical Education; 2011. [Persian].
16. Zhou D. Prognostic factors and predictions of survival data using cox PH models and random survival forest approaches. Biometrics & Biostatistics International Journal. 2017;5(5):165-181.doi: 10.15406/bbij.2017.05.00142.
17. Mocroft A, Reiss P, Kirk O, Mussini C, Girardi E, Morlat P, et al. Is it safe to discontinue primary Pneumocystis jiroveci pneumonia prophylaxis in patients with virologically suppressed HIV infection and a CD4 cell count <200 cells/microL? Clin Infect Dis. 2010;51(5):611-9. doi: 10.1086/655761.
18. Kilsztajn S, de Souza Lopes E, Lima LZ, da Rocha PAF, Caminhada S, Cotta IN, et al. AIDS cases and survival among injecting drug users in Sao Paulo State, Brazil. 2016. Available from: https://pdfs.semanticscholar.org/9c90/64cb948d642bbeabcfa6cb1107a04e3c76ff.pdf?_ga=2.254250184.1360747396.1583168032-1344948385.1546267048.
19. Poorolajal J, Molaeipoor L, Mohraz M, Mahjub H, Ardekani MT, Mirzapour P, et al. Predictors of progression to AIDS and mortality post-HIV infection: a long-term retrospective cohort study. AIDS Care. 2015;27(10):1205-12. doi: 10.1080/09540121.2015.1045405.
20. Cuong do D, Thorson A, Sönnerborg A, Hoa NP, Chuc NT, Phuc HD, et al. Survival and causes of death among HIV-infected patients starting antiretroviral therapy in north-eastern Vietnam. Scand J Infect Dis. 2012;44(3):201-8. doi: 10.3109/00365548.2011.631937.
21. Deeks SG, Phillips AN. HIV infection, antiretroviral treatment, ageing, and non-AIDS related morbidity. Bmj. 2009;338:a3172. doi: 10.1136/bmj.a3172.
22. Aalen OO, Farewell VT, De Angelis D, Day NE, Gill ON. A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. Stat Med. 1997;16(19):2191-210. doi: 10.1002/(sici)1097-0258(19971015)16:19<2191::aid-sim645>3.0.co;2-5.
23. Yola A, Panteleev A, Sologub T. Outcome of treatment of tuberculosis in HIV infected persons in the era of highly active antiretroviral therapy (HAART) as seen in the Second City Tuberculosis Hospital in Saint Petersburg, Russia. Retrovirology. 2005;2(1):S135. doi: 10.1186/1742-4690-2-S1-S135.
24. Moshrefi AH, Hosseini SM, Amani R, Razavimehr SV, Aghajanikhah MH, Mahmoodi P. Investigation of aids epidemiology in Mazandaran province during 1986-2014. Journal of Rafsanjan University of Medical Sciences. 2016;15(6):575-82. [Persian].