Am J Cardiovasc Dis 2021;11(4):484-488
www.AJCD.us /ISSN:2160-200X/AJCD0134215

Original Article

Bayesian statistics versus classical statistics in survival analysis: an applicable example

Moslem Taheri Soodejani[1], Seyyed Mohammad Tabatabaei[2,3], Marzieh Mahmoudimanesh[4]

[1]Center for Healthcare Data Modeling, Department of Biostatistics and Epidemiology, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran; [2]Medical Informatics Department, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran; [3]Clinical Research Unit, Imam Reza Hospital, Mashhad University of Medical Sciences, Mashhad, Iran; [4]PhD Student in Biostatistics, Department of Biostatistics and Epidemiology, School of Health, Kerman University of Medical Sciences, Kerman, Iran

Received April 13, 2021; Accepted June 29, 2021; Epub August 15, 2021; Published August 30, 2021

Abstract: Background: Heart disease is the leading cause of death in the world and 17 million people die from cardiovascular diseases around the world each year, so finding factors that affect the survival of these patients is of particular importance. Therefore, finding the best model to analyze patient survival can help to find more accurate results. Methods: There are different methods to survival analysis that assess one or more risk factors; the classic Kaplan-Meier method, Cox regression, parametric survival models, and newer models such as Bayesian survival. Cox regression is most common and is generally used for time-dependent data, and the main difference between cox regression and Bayesian models is that the prior distribution in Bayesian models can affect the values of the parameters. Some survival analysis models have certain conditions that need to be considered before analyzing the data. In this paper, we use a dataset from Kaggle and discuss these conditions. This dataset contains medical records of 299 patients with heart failure collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) from April to December 2015. Results: This paper discusses that if the effective sample size is not sufficient, Bayesian survival models can be used to achieve more accurate results because this model is not affected by the sample size. The results of both methods are shown on a sample of cardiac data and based on the results of Bayesian Cox regression model, it was observed that Age, Anemia, Ejection fraction, High blood pressure and Serum creatinine were effective on patient survival. Conclusion: The Bayesian models are much more accurate to determine survival and determine risk factors when dealing with data on rare diseases or diseases with low mortality, including heart patients whose survival probability is higher than that of cancer patients.

Keywords: Heart failure, survival, cox regression, Bayesian

Introduction

Heart failure diseases and cancers are known to be the leading causes of death, since approximately 17 million people die from cardiovascular diseases around the world each year; Therefore, identifying factors affecting the survival of people with heart disease may lead to better care and reduce their risk of early death and thus increase their survival [1, 2].

To investigate one or more risk factors, classical methods such as Kaplan-Meier method, Cox regression, and parametric survival models could be used to determine survival rate and factors affecting survival. But newer, and more accurate methods based on Bayesian statistics have also been proposed for survival analysis [3].

While using classical models such as Cox regression, a sufficient sample size (usually 10 samples per parameter) would lead to more accurate results, so the achievement of correct results depends on the sample size; But sometimes a large sample is not accessible, or even the number of cases observed is small compared to the number of investigating parameters, or even collecting a larger sample is not possible due to ethical concerns. Bayesian analysis, however, does not require large samples and can typically be used in smaller datasets without losing power while maintaining accuracy [4, 5].

During a classical analysis, low sample size leads to unreliable estimation, and this a weakness. In survival analysis, when the number of events is not sufficient or in other words, the effective sample size is small, the results of these models may not be accurate enough; For example, consider the data in Article [6, 7] which is available on Kaggle [8]. This dataset contains medical records of 299 patients with heart failure collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) from April to December 2015. These patients consist of 105 women and 194 men with the age of 40 to 95 years. Of these, only 96 patients died (about 30 percent), which is not enough, given the number of variables, to do survival analysis through classical models. So, in such circumstances, Bayesian analysis can provide more reliable results. Therefore, the factors affecting the survival of these patients have been investigated using Bayesian and classical analysis. First, an explanation of the models is provided and then the importance of Bayesian analysis while dealing with low sample size is discussed.

Materials and methods

Dataset

Dataset which is available on Kaggle site [8], contains medical records of 299 patients with heart failure collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) from April to December 2015.

Bayesian analysis

Bayesian analysis works based on the fact that combining everything that is known about a parameter before the observation with the information obtained from the data itself (likelihood) would result in updated knowledge about the parameter (posterior). Prior information can be obtained from meta-analyzes, previous studies on comparable research population, a pilot study, specialists, or a wide range of other sources. If such information is available, it is called informative prior; otherwise, it is non-informative prior [9, 10].

Cox regression

Cox regression is generally used for time-dependent data; Such as the time period from the beginning of the treatment to the death or the time between the first and second heart attack and so on.

In Cox regression, the purpose is to examine the relationship between one or more independent variables during the time until the occurrence of an event, and thus the effect of independent variables on disease risk is investigated. Modeling is done as follows:

$$h(t,X) = h_0(t)\exp(\sum_{i=1}^{p} \beta_i X_i);\ X = (X_1,\ldots,X_p)$$

In which the parameters to be estimated are the regression coefficients of the variables (β_i) [11, 12].

Bayesian Cox regression

The purpose of using Bayesian Cox regression is similar to classical one but they use different methods to estimate parameters. In this model, in addition to the present data, the prior information under the heading of the prior distribution can affect the values of the parameters and consequently, the significance or non-significance of the variable [4, 13].

Ethical

The dataset used in this study was obtained from the Kaggle site (https://www.kaggle.com/andrewmvd/heart-failure-clinical-data) and has been referenced to the article of the owners of that data [6].

Results

Survival analysis of patients with heart failure

In the mentioned data, the effect of 11 variables including Age, Anemia, Creatinine phosphokinase, Ejection fraction, Platelets, Serum sodium, Smoking, Diabetes, High blood pressure, Serum creatinine and Sex on the survival of patients has been investigated using two models including Bayesian Cox regression and classical Cox regression. Generally, as a rule of thumb, at least 10 samples are needed for each variable. In survival analysis, the sample size is actually the effective sample size or the

Table 1. Results of Bayesian Cox regression

Variable	Mean	Std. Dev	95% CI-Low	95% CI-Up
Age	0.0467	9.495×10^{-3}	0.0285	0.0654
Anaemia	0.4399	0.2136	0.0076	0.8529
Creatinine phosphokinase	2.017×10^{-4}	1.039×10^{-4}	-1.976×10^{-5}	3.875×10^{-4}
Ejection fraction	-0.0498	0.0106	-0.0705	-0.0288
Platelets	-5.671×10^{-7}	1.135×10^{-6}	-2.813×10^{-6}	1.601×10^{-6}
Serum sodium	-0.0422	0.0229	-0.0877	0.0025
Smoking	0.1179	0.2465	-0.3837	0.5861
Diabetes	0.1384	0.2213	-0.2953	0.5724
High blood pressure	0.4651	0.2159	0.0414	0.8893
Serum creatinine	0.3105	0.0708	0.1668	0.4403
Sex	-0.2415	0.2532	-0.7328	0.2484

Table 2. Results of Cox regression

Variable	β	SE (β)	Exp (β)	*P*-value
Age	0.0464	0.0093	1.0480	6.45×10^{-7}
Anaemia	0.4601	0.2168	1.5840	0.0338
Creatinine phosphokinase	2.207×10^{-4}	9.919×10^{-5}	1	0.0260
Ejection fraction	-0.0489	0.0104	0.9522	2.98×10^{-6}
Platelets	-4.635×10^{-7}	1.126×10^{-6}	1	0.6806
Serum sodium	-0.0441	0.0232	0.9568	0.0575
Smoking	0.1289	0.2512	1.1380	0.6078
Diabetes	0.1399	0.2231	1.1500	0.5307
High blood pressure	0.4757	0.2162	1.6090	0.0278
Serum creatinine	0.3210	0.0701	1.3790	4.76×10^{-6}
Sex	-0.2375	0.2510	0.7886	0.3452

number of occurred events; in this data, the desired event is the death of the patient. Now, if there were at least 110 deaths in this dataset, the results of the classical analysis could be trusted, but only 96 deaths were reported. Although the difference is not large and the results of the two models are expected to be close to each other, the number of deaths is still less than the minimum required.

The results of both Bayesian and classical models are stated in **Tables 1** and **2**. The results of Bayesian Cox model demonstrated that the 95% probability interval for variables including Age, Anemia, Ejection fraction, High blood pressure and Serum creatinine does not contain zero, so it can be said that these variables would have a significant effect on the survival.

Comparing the results of classical and Bayesian Cox regression give us a lead that when using classical Cox regression, Creatinine phosphokinase was identified to be significant in addition to Age, Anemia, Ejection fraction, High blood pressure and Serum creatinine, which were not confirmed by using Bayesian Cox regression.

According to the results of both classical and Bayesian models, in patients with heart failure, the risk of death increases by approximately 50% for every 10 years of age (exp (10×0.046) = 1.49). Also, the risk of death in patients with anemia is about 55% more than the others.

According to the results of classical regression, the variable Creatinine phosphokinase was identified as a significant variable on the risk of death, but due to the value which is close to zero, its effect on that risk can be negligible. So, the results of Bayesian Cox regression which showed this variable not to be significant can be trusted.

Discussion and conclusion

Survival analysis and risk factors analysis have always been among the topics of interest for health domain researchers. Nowadays, various methods such as Cox regression, parametric models, Bayesian analysis, machine learning and data mining techniques has been provided to analyze Survival data. While using machine learning techniques, data needs to be divided into two partitions including train and test datasets, and sometimes three partitions including train, validation, and test datasets. Therefore, a larger sample size could provide better results because more data results in better learning of the model or network [14, 15].

Classical models are also affected by the sample size due to the fact that they use the maximum likelihood method to estimate the parameters, especially in survival analysis because what matters is the number of events, that is when it is small compared to number of people in the study may produce unreliable results. In such circumstances, considering that Bayesian methods also use the prior information and, most importantly, are not affected by the sample size and the accuracy of the results does not depend on the sample size, they can be a good choice [4, 16].

Davide Chicco et al. used machine learning techniques on collected data from patients with heart failure disease. It was resulted that the two most important factors in predicting patient survival are Ejection fraction and Serum creatinine [6]. The results of both classical and Bayesian Cox regression also confirmed the importance of these two variables in patient survival. In addition, Bayesian regression showed that two other variables including High blood pressure and Anemia could be considered as important as those two mentioned variables.

In this study, the mean value of the parameters in Bayesian method was close to the ones in Cox regression, but on the other hand, in Cox regression, the variable Creatinine phosphokinase was detected to be significant. On the contrary, this significance was not confirmed by Bayesian analysis, which could be due to the fact that prior information about the parameters was not available, and in fact, the normal distribution with high variance was considered as a non-informative prior; However, if a meta-analysis is performed to produce informative prior, the value of the parameters and even their significance or non-significance may be confirmed with greater accuracy and specificity. But what is important is that using Bayesian methods can come up with good results without worrying about insufficient sample size [4, 9].

Even though the effective sample size (number of events) in this data was less than the minimum it should be, due to the number of variables, it was very close to this number and therefore the results of the two models are expected to be similar, but if the number of events occurrences was much lower than this or more parameters on patient survival was investigated, using Bayesian Cox regression could provide more accurate results [13].

Therefore, in many cases, such as while dealing with data on rare diseases or containing low mortality, considering that Bayesian models uses a combination of prior information and collected data to predict survival and determine risk factors, and considering that high iteration-based simulation methods are used in it to estimate the parameters while its test power and results accuracy do not depend on the sample size, they can provide more reliable results in survival analysis.

Disclosure of conflict of interest

None.

Address correspondence to: Marzieh Mahmoudimanesh, Biostatistics, Department of Biostatistics and Epidemiology, School of Health, Kerman University of Medical Sciences, Medical University Campus, Haft-Bagh Highway, Kerman 7616913555, Iran. Tel: +98-9364121827; E-mail: m_mahmudi69@yahoo.com

References

[1] Taheri Soodejani M, Lotfi MH, Tabatabaei SM, Mohammadzadeh M and Dolatian M. Application of population attributable fraction in prevention of cardiovascular diseases. 2015; 7-13.

[2] Triposkiadis F, Xanthopoulos A, Parissis J, Butler J and Farmakis D. Pathogenesis of chronic

heart failure: cardiovascular aging, risk factors, comorbidities, and disease modifiers. Heart Fail Rev 2020; [Epub ahead of print].

[3] Rafati S, Baneshi MR and Bahrampour A. Factors affecting long-survival of patients with breast cancer by non-mixture and mixture cure models using the Weibull, log-logistic and Dagum distributions: a Bayesian approach. Asian Pac J Cancer Prev 2020; 21: 485-490.

[4] Van De Schoot R, Broere JJ, Perryck KH, Zondervan-Zwijnenburg M and Van Loey NE. Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. Eur J Psychotraumatol 2015; 6: 25216.

[5] Gannon MA, de Bragança Pereira CA and Polpo A. Blending Bayesian and classical tools to define optimal sample-size-dependent significance levels. Am Stat 2019; 73: 213-222.

[6] Chicco D and Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 2020; 20: 16.

[7] Joseph L, du Berger R and Bélisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. Stat Med 1997; 16: 769-781.

[8] https://www.kaggle.com/andrewmvd/heart-failure-clinical-data.

[9] Brard C, Le Teuff G, Le Deley MC and Hampson LV. Bayesian survival analysis in clinical trials: what methods are used in practice? Clin Trials 2017; 14: 78-87.

[10] Hosseinnataj A, RezaBaneshi M and Bahrampour A. Mortality risk factors in patients with gastric cancer using Bayesian and ordinary Lasso logistic models: a study in the Southeast of Iran. Gastroenterol Hepatol Bed Bench 2020; 13: 31-36.

[11] Hu Y. Survival analysis of cardiovascular diseases. 2013.

[12] Kleinbaum DG and Klein M. Survival analysis. Springer 2010.

[13] Omurlu IK, Ozdamar K and Ture M. Comparison of Bayesian survival analysis and Cox regression analysis in simulated and breast cancer data sets. Expert Syst Appl 2009; 36: 11341-11346.

[14] Dekker FW, De Mutsert R, Van Dijk PC, Zoccali C and Jager KJ. Survival analysis: time-dependent effects and time-varying risk factors. Kidney Int 2008; 74: 994-997.

[15] Vayena E, Blasimme A and Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med 2018; 15: e1002689.

[16] Vakili M, Taheri M and Sartipzadeh N. Study of risk factors for acute myocardial infarction in patients registered at shahid Sadooghi hospital in Yazd: a case-control study. Quarterly J Sabzevar Univ Medical Sci 2015; 22: 144-122.