Gastroenterology and Hepatology From Bed to Bench.

ORIGINAL ARTICLE

Statistical count models for prognosis the risk factors of hepatitis C

Asma Pourhoseingholi[1], Alireza Akbarzadeh Baghban[2], Farid Zayeri[3], Seyed Moayed Alavian[4], Mohsen Vahedi[5]

[1]Student's Research Committee, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[2]Department of Basic Sciences, School of Rehabilitation, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[3]Proteomics Research Center, School of Paramedical Science Shahid Beheshti University of Medical Science

[4]Baqiyatallah University of Medical Sciences, Baqiyatallah Research Centre for Gastroenterology and Liver Disease, Tehran, Iran

[5]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

ABSTRACT

Aim: The aim of this study was to compare alternatives methods for analysis of zero inflated count data and compare them with simple count models that are used by researchers frequently for such zero inflated data.

Background: Analysis of viral load and risk factors could predict likelihood of achieving sustain virological response (SVR). This information is useful to protect a person from acquiring Hepatitis C virus (HCV) infection. The distribution of viral load contains a large proportion of excess zeros (HCV-RNA under 100), that can lead to over-dispersion.

Patients and methods: This data belonged to a longitudinal study conducted between 2005 and 2010. The response variable was the viral load of each HCV patient 6 months after the end of treatment. Poisson regression (PR), negative binomial regression (NB), zero inflated Poisson regression (ZIP) and zero inflated negative binomial regression (ZINB) models were carried out to the data respectively. Log likelihood, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to compare performance of the models.

Results: According to all criterions, ZINB was the best model for analyzing this data. Age, having risk factors genotype 3 and protocol of treatment were being significant.

Conclusion: Zero inflated negative binomial regression models fit the viral load data better than the Poisson, negative binomial and zero inflated Poisson models.

Keywords: HCV, SVR, Count models, Zero inflated models.

(Please cite as: **Pourhoseingholi A, Akbarzadeh Baghban A, Zayeri F, Alavian SM, Vahedi M. Statistical count models for prognosis the risk factors of hepatitis C. Gastroenterol Hepatol Bed Bench 2013;6(1):41-47).**

Introduction

Hepatitis C virus (HCV) infection is a major cause of liver diseases worldwide and represents a major public health problem (1-5). Both transfusion and contact with infected blood and its products, intravenous drug abuse, contamination during medical procedures and lack of attention to health precautions are different risk factors of HCV(6, 7). Between 130 and 170 million people are infected with HCV worldwide and the global prevalence of this infection is 2.2%-3% (2, 8, 9). But this prevalence varies between countries and between developed world and undeveloped countries because of difference in health policy and medical care(10). There is no exact estimation of HCV infection in Iran and estimates rely upon studies that have been performed on high-risk groups or a specific geographic location. Two Iranian studies examined the prevalence of HCV infection in the general

Received: 29 August 2012 *Accepted*: 18 October 2012
Reprint or Correspondence: Alireza Akbarzadeh Baghban, PhD. Department of Basic Sciences, School of Rehabilitation, Shahid Beheshti University of Medical Sciences, Tehran, Iran
E-mail: Akbarzad@gmail.com

population and estimated a population prevalence of less than 1% in Iran (11, 12).

Risk factor evaluation and interventions to decrease the problem in communities is one solution to protect people from acquiring the infection. In this paper viral load of HCV patient and related factors of them that can effect on low or high viral load were examined.

Viral load, like other count data needs count models to analyzing (13). PR model is one of the most established count models used by researchers. The important assumption of the PR model is that the data must not have any over-dispersion—a larger variability than expected (13). Up until recent years, the NB model has been used to describe this distribution assuming that over-dispersion is only due to unobserved heterogeneity (14). The distribution of viral load contains a large proportion of excess zeros, (HCV RNA under 100), that can lead to an over-dispersion. In this situation, alternative models may be better at accounting for over-dispersion due to excess zeros (14).

For independent counts with excessive zeroes Lambert proposed a ZIP regression model(15). Lambert showed this model had better fit than PR or NB models when data had excessive zero. Green in 1994 introduced ZINB model and showed sometimes extra over dispersion occur in zero inflated data, so the ZINB models had the best fit (16). Although the application of these models and their comparisons with other count models has also increased in medical and health fields in recent years (14), but unfortunately many researchers in Iran are not familiar with these models and they use ordinary count models such as PR and NB for analyzing zero inflated count data. Comparison between these models is needed. A review of the application and comparison of such models in health research is also reported (17). The aim of this study was in two fold; firstly, to determine the factors of SVR in HCV patients and secondly to find the best model for analyzing this data. Ordinary count models such as PR and NB, ZIP model and ZINB regression model were used and compare to identify factors related to SVR in HCV patient.

Patients and Methods

This cross-sectional study was a part of a larger longitudinal study that was conducted between 2005 and 2010. All data for this research was drawn from medical records of 186 patients with hepatitis C, who were referred to Tehran hepatitis clinic, a clinical clinic of Bagiyatallah Research Center for Gastroenterology and Liver diseases between 2005 through 2010. Patients who completed the period of treatment (duration dependent upon treatment regimen - for either 24 weeks or 48 weeks) were included in this study and patients who did not complete their recommended period of treatment were omitted. Information relating to the 186 patients included viral load (HCV-RNA) after treatment, demographic information including sex and age, genotype including genotype 1, 2 and 3, and treatment protocol including combination therapy of standard interferon (3 MU three times a week) plus Ribavirin (800-1200 mg per day) for either 24 weeks or 48 weeks (18-20) and a combination therapy of peg-interferon (Alfa 2a in a fixed dose of 180 micrograms per week) plus Ribavirin (800- 1200 mg per day) is for 24 weeks either 48 weeks (19, 21), history of blood transfusion, addiction (IV drug user) and needle stick as risk factors was extracted from their medical records.

The five covariates were age, sex, genotype, protocol of treatment and risk factors entered in this study. HCV-RNA negative (we considered zero in our analyzing) is defined as less than 100. In figure one the process of study is shown in a flow diagram.

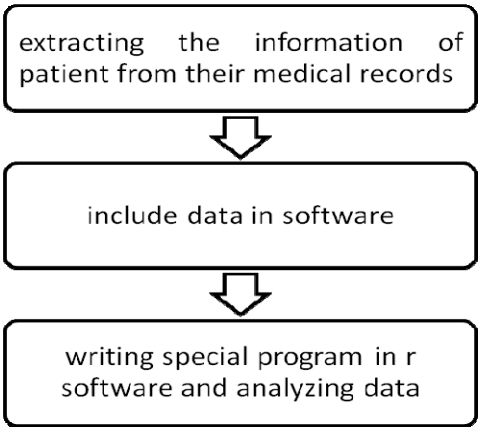

Figure 1. Diagram showing the process of study

Descriptive statistics and frequency distribution such as mean, standard deviation and percentage were calculated according to standard methods. The outcome variable was the viral load of HCV patient. 66.5% of observations were zeros in this study because of SVR. PR model is one of the models from general linear models (GLM) for describing count outcomes or proportion/rates (13). Sometimes in PR the variances are much larger than the means, whereas Poisson distributions have identical mean and variance. The phenomenon of the data having greater variability than expected for a general linear model is called over-dispersion. A common cause of over-dispersion is heterogeneity among subjects (13). NB model, is another model from GLM as an alternative to the PR model, and is a solution to account for over-dispersion due to unobserved heterogeneity (14). Sometimes the NB model may not be appropriate if the over-dispersion due to an excess of zeros in the outcome. In such a situation, alternative models such as zero inflated models are recommended (15). Alternatively, if the non-zero observation parts does not follow the Poisson model then the ZINB is used by considering count process as a negative binomial distribution (14). The ZINB model provides the possibility that account for the over-dispersion due to both types of excess zeros and unobserved heterogeneity (14, 22). The models (e.g., PR versus NB and ZIP, NB versus ZINB, ZIP versus ZINB) were compared using the Vuong test and likelihood ratio test. To compare performance of the models, there are various methods such as log likelihood, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The p-values less than 5% were considered as significant results. Stata 11 and R program were used for analyzing.

Results

A total of 186 patients were eligible and entered in this study. Of those in the study, 123 (66.5%) of patient had SVR. According to the score test that is used for checking zero inflation, these data showed significant zero inflation ($p < 0.001$). The mean age of patients was 42.88 (standard deviation, 11.17) years and range 19-76 years. The distributions of covariates between patients are shown in table 1. The significant Pearson chi square goodness of fit (gof) test ($p< 0.001$) along with other characteristics of model fit indicated that the PR model produced a poor fit for data.

Table 1. The distribution of covariance between patients

Variables	categories	N (%)
Sex	man	55(29.6)
	woman	131(70.4)
Risk factor	Positive	104(55.9)
	negative	84(44.1)
Genotype	1	142(76.3)
	2	4(2.2)
	3	40(21.5)
Protocol of treatment	Inte[¥]+ Rib[*]	100(53.8)
	Peg-inte+ Rib	86(46.2)

[¥]Inte: Interferon; [*]Rib: Ribavirin

In the NB model, the estimated dispersion statistic (α) was 3.51 (95% CI: 3.25, 3.77). A significant likelihood ratio test ($p< 0.001$) of dispersion statistic from zero favored the NB model over the PR model. Voung test was used for comparison between ZIP and PR. The significant result ($p<0.001$) showed that ZIP model was better than PR. But in comparison between ZIP and the

NB Voung test result was in favored of NB model. Between the ZINB and PR and ZINB and NB models the Voung test showed ZINB was better model too ($p<0.001$). For the significant likelihood ratio test ($p < 0.001$) the ZINB model was better than ZIP. The ZINB estimated dispersion parameter was observed different than zero as [(α = 1.87; 95% CI: (1.39, 2.52)]. Comparisons between models are shown in Table 2.

Table 2. Comparison of model fit characteristics.

	PR	NB	ZIP	ZINB
AIC	575547	21307.6	516065	21196.6
BIC	575586	21346.1	516103	21235.1
Logliklihood	-287767	-10646	-258025	-10591

The minimum AIC was observed for the ZINB model, followed by NB model. However, other validity indices of the model (maximum log likelihood, minimum BIC) favored ZINB over all other models. So ZINB model was the best model for analyzing this data. Table 3 showed the results of this model. Age, risk factor genotype 3 and protocol of treatment had significant relation with SVR of patient in ZINB model. According to these results including an increasing age (ADJ.OR=0.97; 95% CI 0.94, 0.99; P=0.03) and having one risk factor (ADJ.OR=0.47 95% CI 0.24, 0.95; P=0.03) reduces the chance of SVR. For genotype 3 (ADJ.OR=4.48; 95% CI 1.87, 12.82; P=0.001) combination therapy of Peg-interferon plus Ribavirin (ADJ.OR=2.41; 95% CI 1.22, 4.48; P=0.01) increased the chance of SVR.

Discussion

Achieving SVR is very important in treatment proceed of HCV. So in this study we examined the factors that related to SVR in HCV patient. Because of this reason that the majority of patients (66.5%) had SVR, our data set had a zero inflated form. Common approach for analyzing count data like viral load in our data are Poisson and negative binomial regression (13, 23) and there are a different method for excessive zero data such as zero inflated models that we used them in this paper and Hurdle models (24). There are lots of studies that they used these models recently (14, 25-28). Goetzel et al used Poisson, negative binomial and zero inflated Poisson To quantify the direct medical and indirect (absence and productivity) cost burden of overweight and obesity in workers in the U.S (29). Carrel et al used a zero inflated negative binomial model to examine how residence within or outside a flood protected area interacts with the probability of cholera presence and the effect of flood protection on the magnitude of cholera prevalence(28). Bergemann and Huang proposed a new method based on zero-inflated Poisson (ZIP) regression likelihood to simultaneously account for missing genotype data and genotype combinations with zero counts (26). Dwivedi et al compared zero inflated models (Poisson and negative binomial) and hurdle models to test model abilities to predict the number of involved nodes in breast cancer patients (14). In this paper, NB, ZIP and ZINB models was carried out for examining the related factor with SVR in HCV patient and according to the results improved fit of the NB model over PR and ZIP, it clearly indicates that over-dispersion is involved due to unobserved heterogeneity and/or clustering. In addition, ZIP provided evidence of over-dispersion due to excess zero in viral lode of patients in comparison to the PR model. Comparing the ZIP and ZINB models according to likelihood ratio test, the ZINB model is more appropriated than ZIP. Beside, AIC, BIC and log likelihood criterion showed that ZINB model was better than the NB regression model, indicating that the NB model may not be appropriate for describing over-dispersed data.

Young people had more SVR then older people. It seems some physiological change related to increasing the age was the reason of this results. Patient with genotypes 3 had more SVR

Table 3. Zero inflated negative binomial model for cost data

variable	Negative binomial part		Zero inflated part	
	Adj. RR* (95% CI)	*P*-value	Adj. OR** (95% I)	*P*-value
Female(reference: male)	0.98(0.95, 1.01)	0.38	0.49(0.24, 1.01)	0.05
Age	1.15(0.55, 2.40)	0.7	0.97(0.94, 0.99)	0.03
Risk factor (reference: negative)	1.04(0.45, 2.38)	0.92	0.47(0.24, 0.95)	0.03
Genotype 2 (reference: 1)	0.001(0.00, 1.01)	0.25	2.43(0.22, 10.80)	0.48
Genotype 3 (reference: 1)	0.50(0.16, 1.60)	0.24	4.48(1.87, 12.82)	0.001
Protocol of treatment (reference: interferon+ribavirin)	0.81(0.36, 1.83)	0.62	2.41(1.22, 4.48)	0.01

*Adjusted Relative Risk
**Adjusted Odds Ratio

than patient with genotype 1. These results suggest that achieving SVR in genotype 1 is more difficult than for other genetypes and this has been confirmed in other studies(30, 31). Certain patient risk factors decrease the chance of SVR. An example is that genotype 1 is associated with patient risk factors such as illegal drug use , infection by transfusion and contact with infected blood and its products (32). Two main treatment protocols were be used in this study according the genotype of patient. Accordingly the results combination therapy of peg- plus Ribavirin had better results than combination therapy of standard interferon plus ribavirin. Many studies have been conducted so far showed that peg-plus Ribavirin had the highest likelihood of a SVR response to treatment (31, 33-37), especially for genotype 1. Genotype 1 responds to treatment poorly and this difficulty is recognized in choice of treatment protocol (36, 37). Unfortunately in Iran, due to the the cost of these expensive drugs, it is not the first choice of doctors. Usually after a patient does response to initial treatments with monotherapy, doctors decide to choose Peg- plus Ribavirin (37).

In conclusion we have shown ZINB regression models is the best model for analyzing and describing viral load distribution. This confirms that the distribution of the viral load contained over-dispersion not only due to unobserved heterogeneity but also due to excessive negative HCV-RNA (zeros). As expected, the PR model had the worst model for HCV-RNA analyzing. Accounting only one source of over-dispersion, either due to excessive zeros or due to unobserved heterogeneity, the gof of models improved as indicated by ZINB, NB and ZIP models. To analyze count data with zeros it is essential to check the assumptions of different count models and then using the appropriate count model is essential to have meaningful results.

Acknowledgements

This study was supported by a grant number 1390-01-96-7664 in school of paramedical science, shahid Beheshti University of medical sciences. Also the authors would like to express their thanks to Gastroenterology and Liver Diseases Research Center Shahid Beheshti University of Medical Sciences for their valuable collaboration in this study.

References

1. Alavian SM. Are the real HCV infection features in Iranian patients the same as what is expected? Hepat Mon 2005;5:3-5.

2. Alter MJ. Epidemiology of hepatitis C virus infection. World J Gastroenterol 2007;13:2436-41.

3. Alavian SM. Hepatitis C virus infection: Epidemiology, risk factors and prevention strategies in public health in I.R.IRAN. Gastroenterol Hepatol Bed Bench 2010;3:5-14.

4. Alavian SM. New globally faces of hepatitis B and C in the world. Gastroenterol Hepatol Bed Bench 2011;4:171-74.

5. Alavian SM, Adibi P, Zali MR. Hepatitis C virus in Iran: Epidemiology of an emerging infection. Arch Iran Med 2005;8:84-90.

6. Alavian SM. We need a new national approach to control hepatitis C: It is becoming too late. Hepat Mon 2008;8:1-3.

7. Alavian SM. Optimal Therapy for Hepatitis C. Hepat Mon 2004;4:41-2.

8. Alter MJ. Epidemiology of hepatitis C. Hepatology 1997;26:62S-5S.

9. Alter MJ. Epidemiology of hepatitis C virus infection. World J Gastroenterol 2007;13:2436-41.

10. Alavian SM, Lankarani KB, Aalaei-Andabili SH, pouryasin A, Ebrahimi-Daryani N, Nassri-Toosi M, et al. Treatment of chronic hepatitis C infection: Update of the recommendation from scientific Lader's Meeting 28th July 2011-Tehran, IR Iran. Hepat Mon 2011;11:703-13.

11. Alavian SM, Ahmadzad Asl M, Lankarani KB, Shahababaie MA, Bahram Ahmadi A, Kabir A. Hepatitis C I nfection in the general population of Iran:A Systematic Review Hepat Mon 2009;9:211-23.

12. Merat S, Rezvan H, Nouraie M, Jafari E, Abolghasemi H, Radmard AR, et al. Serprevalance of hepatitis C virus: the first population-based study from Iran. International Journal of Infectious Disease 2010;145:113-6.

13. Agresti A. An Introduction to Categorical Data Analysis.John Wiley & Sons; INC,publication.; 2007.

14. Dwivedi AK, Dwivedi SN, Deo S, Shukla R, Kopras E. Statistical models for predicting number of involved nodes in breast cancer patients. Health 2010;2:641-51.

15. Lambert D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Thecnometrics 1992;34:1-14.

16. Greene w. Accounting for excess zerose and sample selection in poisson and negative binomial regression models. Working paper no,94-10,new york,university department of economics 1994.

17. Sandhu DS, Sandhu S, Karwasra RK, Marwah S. Profile of breast cancer patients at a tertiary care hospital in north India. Indian Journal of Cancer 2010;47:16-22.

18. Myers RP, Regimbeau C, Thevenot T, et al. Interferon for interferon naive patients with chronic hepatitis C (Cochrane Review). In: the Cochrane Library, Isuee 2 Oxford:Update Software 2002.

19. Poynard T, Yuen MF, Ratziu V, Lai CL. Viral hepatitis C. The Lancet 2003;362:2095-100.

20. Poynard T, Mchutchison J, Goodman Z, Ling MH, Alberch J, et al. Is an "a la carte" combination interferon alfa-2b plus ribavirin regimen possible for the first line treatment in patients with chronic hepatitis C? The ALGOVIRC Project Group. Hepatology 2000;31:211-8.

21. Fried MW, Shiffman ML, Reddy KR, et al. Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. N Engl J Med 2002;347:975-82.

22. Yau KK, Wang k, Lee AH. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biometrical Journal 2003;4:437–52.

23. Gardner W, Mulvey EP, Shaw EC. Regression analyses of counts and rates: poisson, overdispersed Poisson, and negative binomial models. Psychological Bulletin 1995;118:392-404.

24. Boucher J, Denuit M, Guillen M. Risk classification for claim counts: a comparative analysis of various zero inflated mixed Poisson and hurdle models. North American Actuarial Journal 2007;11:110–31.

25. Barondess DA, Meyer EM, Boinapally PM, Fairman B, Anthony JC. Epidemiological evidence on count processes in the formation of tobacco dependence. Nicotine Tob Res 2010;12:734-41.

26. Bergemann TL, Huang Z. A new method to account for missing data in case-parent triad studies. Human heredity 2009;68:268-77.

27. Carrel M, Escamilla V, Messina J, Giebultowicz S, Winston J, Yunus M, et al. Diarrheal disease risk in rural Bangladesh decreases as tube well density increases: a zero-inflated and geographically weighted analysis. Int J Health Geogr 2011;10:41.

28. Carrel M, Voss P, Streatfield PK, Yunus M, Emch M. Protection from annual flooding is correlated with increased cholera prevalence in Bangladesh: a zero-inflated regression analysis. Environ Health 2010;9:13.

29. Goetzel RZ, Gibson TB, Short ME, Chu BC, Waddell J, Bowen J, et al. A multi-worksite analysis of the relationships among body mass index, medical utilization, and worker productivity. J Occup Environ Med 2010; 52:S52-58.

30. Ionita-Radu F, Rascanu A, Cheiab B. IL28B polymorphism - predictive factor of HCV infected genotype 1 individuals to treatment response and management of therapy. Rom J Intern Med 2011;49:99-104.

31. Zeuzem S. Standard treatment of acute and chronic Hepatitis C. Zeitschrift Fur Gastroenterologie 2004;42:714-9.

32. Samimi-Rad K, Nategh R, Malekzadeh R, Norder H, Magnius L. Molecular epidemiology of hepatitis C virus in Iran as reflected by phylogenetic analysis of the NS5B region. J Med Virol 2004;74:246-52.

33. Snoeck E, Wade JR, Duff F, Lamb M, Jorga K. Predicting sustained virological response and anaemia in chronic hepatitis C patients treated with peginterferon alfa-2a (40KD) plus ribavirin. Br J Clin Pharmacol 2006;62:699-709.

34. Yee HS, Currie SL, Darling JM, Wright TL. Management and treatment of hepatitis C viral infection: recommendations from the Department of Veterans Affairs Hepatitis C Resource Center program and the National Hepatitis C Program office. Am J Gastroenterol 2006;101:2360-78.

35. Davis GL, Wong JB, McHutchison JG, Manns MP, Harvey J, Albrecht J. Early virologic response to treatment with peginterferon alfa-2b plus ribavirin in patients with chronic hepatitis C. Hepatology 2003;38:645-52.

36. Sanchez-Tapias JM, Diago M, Escartin P, Enriquez J, Romero-Gomez M, Barcena R, et al. Peginterferon-alfa2a plus ribavirin for 48 versus 72 weeks in patients with detectable hepatitis C virus RNA at week 4 of treatment. Gastroenterology 2006;131:451-60.

37. Kamal SM, El Kamary SS, Shardell MD, Hashem M, Ahmed IN, Muhammadi M, et al. Pegylated interferon alpha-2b plus ribavirin in patients with genotype 4 chronic hepatitis C: The role of rapid and early virologic response. Hepatology 2007;46:1732-40.