Kidney Disease and Population Health

nephron Clinical Practice

Nephron Clin Pract 2011;119:c255–c260
DOI: 10.1159/000328916

Published online: September 14, 2011

Survival Analysis II: Cox Regression

Vianda S. Stel[a] Friedo W. Dekker[a, b] Giovanni Tripepi[c] Carmine Zoccali[c]
Kitty J. Jager[a]

[a]ERA-EDTA Registry, Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, and [b]Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands; [c]CNR-IBIM Clinical Epidemiology and Pathophysiology of Renal Diseases and Hypertension, Renal and Transplantation Unit, Ospedali Riuniti, Reggio Calabria, Italy

Key Words
Epidemiology · Survival analyses · Cox regression · Nephrology

Abstract
In contrast to the Kaplan-Meier method, Cox proportional hazards regression can provide an effect estimate by quantifying the difference in survival between patient groups and can adjust for confounding effects of other variables. The purpose of this article is to explain the basic concepts of the Cox regression method, and to provide some guidance regarding the presentation of the results.

Introduction

In a previous article in this series we presented the Kaplan-Meier (KM) method to analyze survival data [1]. The KM method is adequate to explore the survival of populations under investigation, and to test differences in the crude survival between exposure groups. Moreover, the KM method allows the presentation of survival curves. However, when investigating the relation between exposure and clinical outcomes the KM method has the important limitation of not providing an effect estimate and a related confidence interval (CI) to compare the survival in different patient groups. The purpose of this article is to explain the basic concepts of Cox regression analysis using a clinical example.

Clinical Example

For the illustration of the concepts of Cox regression we used a study from the ERA-EDTA Registry [2]. Within this example, Cox proportional hazard regression models were built to examine the association between GFR estimated by the 4-variable MDRD equation (eGFR) at the start of dialysis and all-cause mortality in 6,716 patients who started dialysis in 2003. Survival was calculated from the date at start of dialysis. Survival was censored at the date of leaving the study due to a renal transplant, recovery of renal function, loss to follow-up, or at the end of follow-up time at December 31, 2005. Using this clinical example, four main Cox regression models were built to examine the association between eGFR at the start of dialysis and mortality, each treating the determinant eGFR differently as continuous or as categorical variable (table 1).

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1660–2110/11/1193–0255$38.00/0

Accessible online at:
www.karger.com/nec

Vianda S. Stel, PhD
ERA-EDTA Registry, Department of Medical Informatics
Academic Medical Center, J1b 113.1
PO Box 22700, NL–1100 DE Amsterdam (The Netherlands)
Tel. +31 20 566 7637, E-Mail v.s.stel@amc.uva.nl

Table 1. Four Cox regression models to examine the association between eGFR at the start of dialysis in ml/min/1.73 m^2 and mortality, in which eGFR was treated differently

Model 1	eGFR divided into 2 categories, i.e. high-medium vs. low eGFR group[a]
Model 2	eGFR divided into 3 categories, i.e. high (reference), medium, low eGFR group[b]
Model 3	eGFR as continuous variable – the association of 1 ml/min/1.73 m^2 higher eGFR and survival
Model 4	eGFR as continuous variable – the association of 10 ml/min/1.73 m^2 higher eGFR and survival

[a] High-medium eGFR group (≥8 ml/min/1.73 m^2) and low eGFR group (<8.0 ml/min/1.73 m^2).

[b] High eGFR group (≥10.5 ml/min/1.73 m^2), medium eGFR group (≥8 and <10.5 ml/min/1.73 m^2) and low eGFR group (<8.0 ml/min/1.73 m^2).

Effect Estimate and Confidence Interval

Before explaining the effect estimate of Cox regression analysis, we will explain the incidence rate first. The incidence rate is the ratio of the number of subjects developing disease (or other health outcome) during the study period to the time at risk of disease [3, 4]. The incidence rate should be interpreted as an instantaneous concept, like speed, as its denominator includes a measure of time. In our clinical example, during follow-up the incidence rate of death in the high-medium eGFR group was 23.4/100 person-years, whereas that in the low eGFR group was 17.9/100 person-years. The incidence rate ratio (IRR), which is also referred to as a relative risk (RR), can be calculated by dividing the incidence rate of death in the high-medium eGFR group by the incidence rate of death in the low eGFR group. Thus we obtained an IRR of 1.31. This means that during the study period patients in the high-medium eGFR group had a 31% higher mortality rate compared to patients in the low eGFR group. Within this example, one could suspect that age may obscure the association between eGFR at the start of dialysis and mortality, because patients who start dialysis at higher eGFR levels may be older and for that reason have a higher mortality. Using the IRR in this way, however, it is not possible to adjust the association between eGFR at the start of dialysis and mortality for potential confounders like age.

The effect estimate of Cox regression analyses is a hazard ratio (HR), this is the ratio of two hazards. The hazard is the incidence rate of an event in an infinitesimal short time period (let's say the probability to die within the next second). An extensive explanation of the calculation of the HR is provided by van Dijk et al. [5]. The HR can be interpreted as an IRR or a RR. The advantage of Cox regression analysis is that one can adjust the association of interest for potential confounders [6, 7].

The reason why we need CIs is that the HR is only a point estimate of the effect and does not express the statistical variation, or random error, around the estimate. The CI helps us to quantify the precision of this sample estimate [8]. If the data collection and analyses could be replicated many times, the 95% CI will include the 'real' population HR in 95% of the replications. A value of 1 for the HR means equality of survival in high-medium and low eGFR groups. If 1 is not included in the 95% CI, we may conclude that there is a statistically significant difference between the high-medium and low eGFR group in terms of survival.

To see how this works in practice, table 2 shows the basic layout of the dataset of our clinical example, as required to perform the four unadjusted Cox regression models (table 1) in statistical packages like SPSS or SAS.

Model 1

The first model compared the mortality risk of 2 groups: the high-medium and the low eGFR groups. The output of the Cox regression analysis of model 1 is presented in table 3. The unadjusted HR of patients in the high-medium eGFR group compared to the low eGFR group was 1.30 (95% CI 1.20–1.41). This means that patients in the high-medium eGFR group had a 30% higher mortality risk than patients in the low eGFR group. This 95% CI means that if the data collection and analyses could be replicated many times, the 'real' population HR will be between 1.20 and 1.41 in 95% of the replications. As 1 is not included in the 95% CI, we may conclude that there is a statistically significant difference in survival between the high-medium and the low eGFR groups.

Please, note that if we had examined the risk of death in patients of the low eGFR group (instead of the high-medium eGFR group) compared to the high-medium eGFR group (instead of the low eGFR group) the unadjusted HR would have been 1/1.30 = 0.77 (95% CI 0.71–0.83), meaning that patients in the low eGFR group had a 23% lower chance of death compared to patients in the high-medium eGFR group.

Table 2. Basic data layout of the data set of the clinical example to perform unadjusted survival analyses

Patient No.	Survival time, years	Event (died, 0 = no, 1 = yes)	eGFR (per 1 ml/min/1.73 m^2)	eGFR (per 10 ml/min/1.73 m^2)	eGFR (ml/min/1.73 m^2) divided into 2 groups[a]	eGFR (ml/min/1.73 m^2) divided into 3 groups[b]		
						low eGFR (reference)	dummy 1 medium eGFR	dummy 2 high eGFR
1	1.23	1	6.2	0.62	0	0	0	0
2	0.56	1	14.2	1.42	1	0	0	1
3	1.11	0	8.0	0.80	1	0	1	0
4	0.30	1	10.3	1.03	1	0	1	0
5	0.90	1	9.5	0.95	1	0	1	0
6	2.00	0	8.6	0.86	1	0	1	0
7	2.00	0	7.5	0.75	0	0	0	0
8	1.75	1	9.6	0.96	1	0	1	0
9	1.46	0	11.3	1.13	0	0	0	1
10	0.10	1	11.9	1.19	1	0	0	1
…	…	…	…	…	…	…	…	…
6,716	…	…	…	…	…	…	…	…

[a] 1 = High-medium eGFR group (≥8 ml/min/1.73 m^2) and 0 = low eGFR group (<8.0 ml/min/1.73 m^2).
[b] High eGFR group (≥10.5 ml/min/1.73 m^2), medium eGFR group (≥8 and <10.5 ml/min/1.73 m^2) and low eGFR group (<8.0 ml/min/1.73 m^2).

Table 3. Output of Cox regression model 1: the association between high-medium eGFR versus low eGFR[a] (in ml/min/1.73 m^2) at the start of dialysis and mortality, unadjusted and adjusted for age at the start of dialysis

	Variable	Estimates				
		β	standard error	hazard ratio exp (β)	95% CI	
Unadjusted	eGFR: high-medium (vs. low)	0.26	0.04	1.30	1.20	1.41
Adjusted for age	eGFR: high-medium (vs. low)	0.19	0.04	1.21	1.12	1.31
	Age	0.05	0.002	1.05	1.04	1.05

[a] High-medium eGFR group (≥8 ml/min/1.73 m^2) and low eGFR group (<8.0 ml/min/1.73 m^2).

Model 2

The second model compared the mortality risk of 3 groups: the high, medium and low eGFR groups. If a categorical variable contains more than 2 groups, it is needed to create dummy variables. Dummy variables are variables that can only assume the values 0 and 1 (table 2). In this example, we needed to create two dummy variables, one to contrast medium eGFR to low eGFR and one to contrast high eGFR to low eGFR. Patient 1 was categorized into the low eGFR group, and this patient was assigned a 0 to both dummy variables. Patient 2 was assigned to the high eGFR group and in this case the first dummy variable was set at 0 while the second was set at 1. Patient 3 was categorized to the medium eGFR group and was assigned a value of 1 for the first dummy variable and a value of 0 for the second dummy variable. The two dummy variables were then used in Cox regression to estimate HRs between the groups. The HR of the low eGFR group was set at 1 as this was the reference group. The unadjusted HR of the medium eGFR group was 1.17 (95% CI 1.06–1.29), meaning that patients in the medium eGFR group had 17% higher mortality risk than patients in the low eGFR group (reference group). The unadjusted HR of the high eGFR group was 1.44 (95% CI 1.31–1.58).

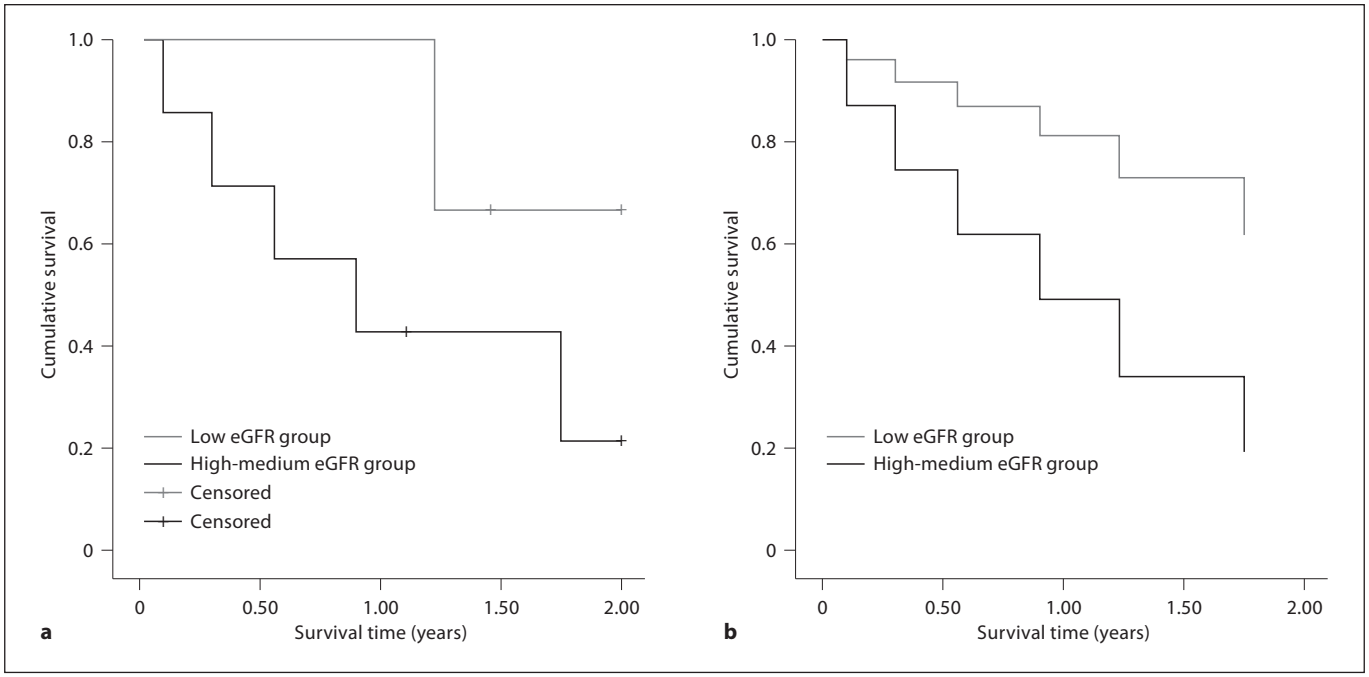

Fig. 1. Differences in survival curves using the KM method (**a**) and Cox regression method (**b**). The survival curves of both the KM method (**a**) and Cox regression method (**b**) are based on the same 10 persons of the clinical example as presented in table 2. The grey line represents the survival curve of patients of the low eGFR group (<8 ml/min/1.73 m^2) and the black line of patients of the high-medium eGFR group (≥8 ml/min/ 1.73 m^2). The KM survival curve shows the real graph, including 1 event in the low eGFR group and 5 events in the high-medium eGFR group (if the survival curve drops down, an event has occurred at that time). The Cox regression method provides a simulated survival graph, including 6 events in both the low eGFR group and high-medium eGFR group (although in reality there is 1 event in the low eGFR group and 5 events in the high-medium eGFR group). Therefore, when analyzing survival data, the survival curves should always be plotted using the KM method.

Table 4. Presentation of results of Cox regression analyses of the four different models[a] on the association of eGFR at the start of dialysis in ml/min/1.73 m^2 and mortality, unadjusted and adjusted for age at the start of dialysis

	Number of deaths	Number of deaths/ 100 person-years	Hazard ratio (95% CI)	
			unadjusted (n = 6,716)	adjusted for age at start of dialysis (n = 6,716)
Model 1				
eGFR <8 ml/min/1.73 m^2 (reference)	1,181	17.9	1.30 (1.20–1.41)	1.21 (1.12–1.31)
eGFR ≥8 ml/min/1.73 m^2	1,231	23.4		
Model 2				
eGFR <8 ml/min/1.73 m^2 (reference)	1,181	17.9	1	1
eGFR 8–10.5 ml/min/1.73 m^2	559	21.0	1.17 (1.06–1.29)	1.11 (1.00–1.22)
eGFR ≥10.5 ml/min/1.73 m^2	672	25.9	1.44 (1.31–1.58)	1.31 (1.20–1.45)
Model 3				
eGFR per 1 ml/min/1.73 m^2	2,412	20.3	1.03 (1.03–1.04)	1.02 (1.02–1.03)
Model 4				
eGFR per 10 ml/min/1.73 m^2	2,412	20.3	1.38 (1.28–1.49)	1.27 (1.17–1.38)

[a] See table 1.

Model 3

This model examined the association of a 1 ml/min/1.73 m^2 higher eGFR at the start of dialysis and mortality (treating eGFR as continuous variable). The unadjusted HR from this model was 1.03 (95% CI 1.03–1.04), meaning that a 1 ml/min/1.73 m^2 higher eGFR at the start of dialysis was associated with a 3% higher mortality risk.

Model 4

Model 4 examined the association of a 10 ml/min/1.73 m^2 higher eGFR at the start of dialysis and mortality. The unadjusted HR was 1.38 (95% CI 1.28–1.49), meaning that a 10 ml/min/1.73 m^2 higher eGFR at the start of dialysis was associated with a 38% higher risk of death. Please note that the estimate in model 3 (1.03/1 ml/min[3]) is identical to that in model 4 (1.38/10 ml/min/1.73 m^2). Indeed, a HR of eGFR per 1 ml/min/1.73 $m^{2\,10}$ = 1.033^{10} = 1.38.

Adjustment for Confounders

In order to obtain an effect estimate adjusted for confounders when analyzing survival data, one could use Cox regression analysis. The identification of potential confounders has been described extensively in a previous paper in this series [6, 7].

As mentioned before, within our clinical example, one could suspect that age may obscure the association between eGFR at the start of dialysis and mortality because patients who start dialysis at higher eGFR levels may be older and for that reason have a higher mortality. Therefore, the association between eGFR at the start of dialysis and mortality was adjusted for the variable 'age at the start of dialysis'. In this case age was entered as a second variable into the Cox regression model.

The output of the unadjusted and adjusted Cox regression analyses of model 1 is presented in table 3. In most statistical packages the output of the Cox regression analyses provides at least a HR, with its 95% CI and an estimate of the regression coefficient β. The β estimate is directly related to the HR because HR equals e^{β}. Thus the HR and β provide information on the strength of the association between eGFR and mortality. When comparing the HR or β of eGFR of the unadjusted model (HR = 1.30; β = 0.26) and adjusted model (HR = 1.21; β = 0.19) it is possible to judge how strong the confounder age affected the association between eGFR at the start of dialysis and mortality. The HR and β of high-medium eGFR in the unadjusted model are different from those in the adjusted model, meaning that age is a confounder in the association between eGFR at the start of dialysis and mortality.

Assumptions of the Cox Regression Model

Often the Cox regression model is referred to as the Cox proportional hazards model, as one of the assumptions of the model is that over the follow-up period the hazards in the groups compared should be proportional to each other and that consequently the HR should be the same during follow-up. Although in practice it is unlikely that the proportional hazards assumption is ever fully satisfied, an important violation of the proportional hazards assumption may result in wrong and misleading estimates. For example, if the survival curves of two groups cross, the HR is clearly not the same over time, and in that case the use of the Cox regression model with proportional hazards is inappropriate. Two popular approaches to test if the hazards are proportional are described elsewhere [5].

Guidance on Presentation and Interpretation of Results

When analyzing survival data, the survival curves should always be plotted using the KM method (and not using the Cox regression method) [1]. Figure 1 shows the difference in the survival curves using the KM method (providing real survival curves) and the Cox regression method (providing simulated survival curves). Cox regression analyses are needed to provide unadjusted and adjusted effect estimates with their related 95% CIs.

Table 4 shows the results of the four Cox regression models. As a main result of Cox regression analysis, one should present both the unadjusted and adjusted HRs with the corresponding 95% CIs. An advantage of presenting both the unadjusted and adjusted results is that the readers can clearly see which confounders are included in the model and to show the effect of confounders on the HR of the outcome of interest.

Observational studies (cohort, case–control, or cross-sectional designs) should report their results according to the STROBE statement [9] and randomized controlled trials according to the CONSORT statement

[10]. The STROBE statement recommends that in a cohort study, as in our clinical example, one should summarize the follow-up time, and report numbers of outcome events or summary measures over time. Therefore, when describing the results in our clinical example, it should be stated that during the average follow-up time of 1.77 years and 11,873 person-years of observation, 2,412 (35.9%) deaths occurred in the 6,716 patients who started dialysis in 2003. Some of this information is included in table 4, which presents also the number of deaths and the number of deaths per 100 person-years in each group.

Conclusion

Cox proportional hazards regression is a powerful and popular method to analyze survival data. The main advantage of Cox regression analysis is the possibility to present unadjusted and adjusted HRs with their accompanying CIs.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement number HEALTH-F2-2009-241544.

References

1 Stel VS, Dekker FW, Tripepi G, Zoccali C, Jager KJ: Survival analyses 1: the Kaplan-Meier method. Nephron Clin Pract 2011; 119:c83–c88.
2 Stel VS, Dekker FW, Ansell D, et al: Residual renal function at the start of dialysis and clinical outcomes. Nephrol Dial Transplant 2009;24:3175–3182.
3 Noordzij M, Dekker FW, Zoccali C, Jager KJ: Measures of disease frequency: prevalence and incidence. Nephron Clin Pract 2010; 115:c17–c20.
4 Jager KJ, Zoccali C, Kramar R, Dekker FW: Measuring disease occurrence. Kidney Int 2007;72:412–415.
5 van Dijk PC, Jager KJ, Zwinderman AH, et al: The analysis of survival data in nephrology: basic concepts and methods of Cox regression. Kidney Int 2008;74:705–709.
6 van Stralen KJ, Dekker FW, Zoccali C, Jager KJ: Confounding. Nephron Clin Pract 2010; 116:c143–c147.
7 Jager KJ, Zoccali C, Macleod A, Dekker FW: Confounding: what it is and how to deal with it. Kidney Int 2008;73:256–260.
8 Tripepi G, Jager KJ, Dekker FW, et al: Measures of effect: relative risks, odds ratios, risk difference, and 'number needed to treat'. Kidney Int 2007;72:789–791.
9 von EE, Altman DG, Egger M, et al: The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370:1453–1457.
10 Schulz KF, Altman DG, Moher D: CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332.