Journal of Veterinary Medicine

Review Article Open Access

Treatment Effects and Risk Factors Evaluation in Longitudinal Studies: A Statistical Help for Data Analysis

Patrizia Boracchi[1*], Roberta Ferrari[2], Debora Groppetti[2] and Damiano Stefanello[2]

[1]Department of Clinical Sciences and Community Health, G. A. Maccacaro Laboratory of Medical Statistics Epidemiology and Biometry, University of Milan, Milan, Italy
[2]Department of Veterinary Medicine, University of Milan, Milan, Italy

Abstract

This paper was inspired by the experience of the Authors research group composed by oncologist veterinarians

characteristics allowed discussing some aspects which are common to survival analysis studies and, for readers who are condent with statistical packages, to give suggestions to perform the analysis by themselves.

: Survival analysis; Prognosis; Interpretation of model results

e present paper was born from the experience of cooperation among the Veterinarians and a Biostatistician (who are the Authors) in planning and analysing clinical studies. e cooperation started seven years ago and each study gave the opportunity to discuss both clinical and statistical issues. In this way the biostatistician became able to understand clinical research needs, in such a way to plan an adequate analysis, and veterinarians became able to interpret correctly statistical results, in such a way to evaluate results impact on their clinical practice. Several studies which concerned the evaluation of therapeutical strategies or the identi cation of potential risk factors, considering as end point the time elapsed from the beginning of the observation (e.g. date of the disease diagnosis, date of the surgery, starting date of pharmacological treatment) and the occurrence of an event which was related to the treatment failure or to the disease course were published and presented to meetings. Because much more debate arose around these studies than around other kinds of studies, the Authors decided to report some “critical aspects”. e Authors hope that reporting the critical aspects will be helpful to veterinarians, who have little experience on survival analysis, to evaluate and write results of prognostic studies. Since results of the statistical analysis should help clinicians in their “decision making process”, a correct methodology (by the Biostatistician) and a correct interpretation of model results (by the Veterinarian) is relevant.

To show the statistical issues two literature data sets which were standard in survival analysis books, were used:

• Dataset 1: A multicentre clinical trial on remission maintenance of

disease progression (e.g. death due to the disease). In order to perform a correct comparison among study results achieved on the same clinical condition it is important to detail which events were considered in the end points and how they were recorded.

When a small sample of individuals is evaluated, follow-up time and events for each individual can be shown and discussed, making statistical analysis not strictly necessary to understand results. Otherwise, data should be synthesized by descriptive statistics (e.g. mean, median, percentages) and inferential procedures should be considered to draw conclusions on the study results.

Follow-up data require descriptive and inferential statistical techniques which are speci c for survival analysis. e techniques take into account peculiar characteristics of follow-up data: the study end-point may not be observed for all subjects. Some subjects may be free of the event at the end of the observation period and some subjects may be lost to follow-up. e probability of being free of the event during follow-up is commonly estimated by the Kaplan-Meier method. When a putative categorical prognostic factor (e.g. lymph node metastasis) is analysed, the most frequent applied procedure is to estimate the event free survival curve for each category (e.g. lymph node metastasis present *vs.* lymph node metastasis absent) and to compare those event free survival curves by Log-Rank test. To draw conclusions on the prognostic role of the putative prognostic factor only the p-value of the test is usually considered. However, to assess the "clinical relevance" of the prognostic factor "clinically useful" measures should also be provided. ese measures could be related to the di erence between end-point probabilities at a given time (risk di erences), to the ratio between end-point probabilities at a given time (relative risks) or alternatively, to the di erences between end-point rates (rate di erences or hazard di erences) or to the ratio between end-point rates (rate ratios or hazard ratios).

When several clinical and pathological variables are analysed, Cox model is used to evaluate their joint prognostic role (multivariate analysis). Cox model is based on a speci c assumption which must be tenable for the correctness of the results (i.e.: for each variable hazard ratio should be constant over follow-up time and this is named "proportional hazard" [2]). e "optimal" approach is to include all the variables into the model to identify which variables have a "signi cant" prognostic role. Unfortunately, this approach is not always possible. Literature suggests rules on the maximum number of regressors to be considered in multivariate analysis so to obtain reliable results [4-6]. e maximum number of regressors depends on the number of observed events rather than on the number of individuals in the study. Care is also needed for the coding of quantitative and qualitative (categorical) variables in order to avoid possible biased prognostic information. For qualitative variables (e.g. Tumour Stage with categories I, II, III) a category is chosen as "reference" (e.g. Stage I) and the following two hazard ratios: Stage II/Stage I and Stage III/Stage I are obtained by the exponent of the Cox regression coe cients. If Stage II and Stage III are not distinct (considered in the same category), only one hazard ratio is estimated: Stage II or Stage III/Stage I and the clinical interpretation of model results di er from those above cited for the 3 Stage categories. To allow clinical usefulness of the model results, the categories should follow substantiated clinical criteria. For quantitative variables, a linear relationship between the logarithm of the hazard and the variable values is the simplest one. As an example, Age is a continuous variable and, under a linear relationship, the hazard ratio comparing the outcome of x years old subjects with the outcome of x+1 years old subjects is the same whatever is the subject age x. erefore, the logarithm of the hazard ratio comparing the outcome of 2 years old subjects with the outcome of 1-year old subjects is the same of the logarithm of the hazard ratio comparing the outcome of 12 years old subjects with the outcome of 11 years old subjects. However, the linear relationship could be improbable (e.g. the logarithm of the hazard ratio comparing the outcome of 2 years old subjects with the outcome of 1-year old subjects could be less or greater than the logarithm of the hazard ratio comparing the outcome of 12 years old subjects with the outcome of 11 years old subjects). In such a case, a possible complexity of the shape for the relationship between continuous variables and model response should be investigated [7]. Statistical so ware outputs are tables containing regression coe cients, the standard errors and p-values. International guidelines suggest showing regression coe cients with the corresponding 95% con dence interval, because it is simpler to evaluate than standard errors [8-11].

A "statistically signi cant" result does not imply clinical usefulness. If the aim of the study is not only exploratory but it involves clinical decisions, useful insights are provided by a measure of the predictive performance of the model [12].

D

e results of the statistical analysis retrieved for the two selected data sets were used to discuss the following issues:

i) percentages of events, mean and median time are not always appropriate, ii) Log-Rank test: p-value is not a comprehensive evaluation and a related measure of prognostic association should be given, iii) interpretation of the statistical test: a p-value >0.05 does not mean that the variables do not have a prognostic role, iv) interpretation of Cox model results: hazard ratio, risk ratio, con dence intervals, v) coding of the variables in multivariate analysis and the maximum number of regressors allowed, vi) statistical signi cance and predictive ability.

D 1: A multicentre clinical trial on remission maintenance for children with acute Lymphoblastic leukaemia was designed to test whether patients who achieved complete remission using steroid could bene t from further treatment. Forty-two patients were randomized to receive maintenance therapy whit 6-mecaptourindine (6-MP; n=21) or placebo (n=21) [1,2].

Time to relapse (in weeks) of the two groups is reported as follows:

• Placebo 1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23 (all patients in placebo group had a relapse)

• 6-MP 6,6,6,6*,7,9*,10,10*,11*,13,16,17*,19*,20,22,23,25*,32*,32*, 34*,35* (some patients in 6-MP group were still in remission when the study was stopped and were considered as censored times, indicated by*).

: Percentage of events: 100*(21/21) =100%

All patients had a relapse, but from this data presentation no information was retrievable on time when 100% was reached. Results should be referred as "the cumulative probability of relapse at 23 week was 1.0 or "the probability of remission a er 23 weeks is 0".

e probability of remaining free from relapse was 0.762 at 3 weeks, 0.571 at 6 weeks, and so on.

ese are the estimates obtained by Kaplan-Meier method. e corresponding cumulative incidence curve can be easily obtained by 1-relapse free survival probability (Figure 1).

A relapse was observed for all patients in this group. Mean time to relapse and median time to relapse can be directly calculated from follow-up observation time: mean=8.67 weeks and median=8 weeks.

6-

Nine patients with relapse were observed: 100*(9/21) =42.86% and

that obtained results are “unlikely” to arise if the null hypothesis was true. If the null hypothesis is not rejected nothing can be stated on the evidence in favour of the null hypothesis. It is worth of note that p-value is not the most relevant criterion to evaluate di erences between groups, it is also important that the observed di erences are clinically relevant. A statistical test applied to a very large case series could provide a “statistically signi cant” result for a very small di erence, which is not relevant from the clinical point of view [13]. On the other hand, a clinically relevant di erence on a small case series could result as “not statistically signi cant “because of the low power of the test (i.e. the probability to correctly conclude that in the population the survival experience of the two groups are di erent). e observed di erence could be “statistically signi cant” with a greater sample size, thus, in the situation of a clinically relevant di erence with a p-value > 0.05 it is not correct to conclude on the equivalence of survival experience of the two groups in the population. A detailed discussion on the interpretation of statistical tests is reported on the Medical Statistical books ([14] among others) and in several tutorial papers ([15] among others).

For the leukaemia trial the result of the Log-Rank test was: Chi-square= 16.8 and p-value= 4.19×10^{-5} (<0.00001). is result supported the clinical hypothesis that the relapse free survival experience of the two treatment groups was di erent. e relevance of the di erence could be evaluated from the plot of the estimated Kaplan-Meier curves (Figure 1a) but a summary measure of treatment clinical impact is not directly provided by Log-Rank test.

As the hypothesis underlying Log-Rank test is based on the ratio between hazards of events, a possible measure of clinical impact is the hazard ratio, which is assumed to be constant over follow-up. Proposed approaches to estimate hazard ratio based on Log-Rank, have been evaluated by Kitchin and Mock [16]. A simpler method was to use Cox model in which only treatment (coded 0 if 6-MP and 1 if placebo) was included as explanatory variable.

e exponent of Cox model regression coe cient is the estimated hazard ratio and for treatment in leukaemia data set it was 4.801. is means that the hazard of relapse of placebo treated patients was about 5 times the hazard of relapse of 6-MP treated patients. Relevant estimates should be reported in association with the hazard ratio: the corresponding 95% con dence interval (for treatment leukaemia data set was 2.14-10.77). Although the null hypothesis of hazard ratio equal to 1 was rejected, the 95% con dence interval was wide, thus providing the information of a low precision of the estimate.

If the cumulative probability of relapse within a follow-up time (t) is called “risk”, the relative risk was the ratio between the estimated cumulative probabilities of relapse of the two treatment groups at that time [17,18]. It could be easily shown from Figure 1b that the risk ratio was not constant over time and it was di erent from hazard ratio (4.81). For example, at 6 weeks the risk ratio was 2.30, at ten weeks was 2.50, and at twelve weeks was 2.00, thus, in this case, it cannot be reported that the risk ratio was 4.81.

C

D **2:** One-hundred and thirty-seven patients with advanced inoperable lung cancer were randomly assigned to two chemotherapy treatments: standard or experimental. Other additional variables were collected for each patient: Karnofsky Performance Score (0=bad, 100=good), Time from Diagnosis to Randomization (months), Age (years), Prior erapy (0=no, 1=yes), Cell Type (Squamous, Small, Adeno, and Large). Study primary end-point was the comparison of survival experience of the two treatment groups.

One-hundred and twenty-eight patients died (64 in both treatment groups) and nine were still alive at the end of follow-up period [3].

A rst analysis could be performed only on the variable “treatment” because randomization should “guarantee” in probability the equal distribution of other variables in the two treatment groups.

Kaplan-Meier estimates for the two treatments groups are reported in Figure 2 and similar results for the two treatments were suggested. It was worth noticing that curves crossed and this could be a “hint” for the lack of proportional hazard.

Results of the test for the proportional hazard did not provide support to the lack of proportional hazard (p-value=0.07 for Kaplan-Meier transform and p-value=0.14 for the identity transform).

Results of the Cox model including only the variable treatment (coded as 0 for control and 1 for experimental): Hazard ratio =1.018,

95% con dence interval: 0.7144 -1.45, p-value=0.922. Regardless p-value, the hazard ratio was very near to 1.0 thus a very similar result was obtained for the two treatments.

e adjustment of treatment e ect including into the model other variables retained as "clinically relevant" by previously published paper and/or previous knowledge of disease course is still debated in clinical trials literature [19-21]. However, to illustrate multivariate analysis, the other 5 variables in the dataset were included in the regression model to "adjust" treatment e ect. Some variables were quantitative (Karnofsky Performance Score, Time from Diagnosis to Randomization, Age) and others were qualitative (Prior erapy, Cell Type).

: Concerning the quantitative variables the simplest approach is assuming a linear relationship between the variable and model response (logarithm of the hazard) thus the variable can be included in the model without data transformation. is approach could be inadequate and the possible nonlinear relationship needs to be tested. It is not simple and model results are di cult to be represented. If categorisation of the variable can be performed, under clinical consideration, model results are simple to evaluate. Nevertheless it can be taken into account that categorisation may result in a loss of prognostic information.

Categorical variables must be included into the model by generating dummy variables. One of the categories is chosen as "reference" and each dummy variable allows estimating the ratio between the hazard of the category and the hazard of the reference one. For a variable with k categories k-1 dummy variables are needed. us, for Treatment and Prior erapy, having two categories, one dummy variable was generated: Standard Treatment and No Prior erapy were chosen as reference.

Cell Type was coded by 4 categories and three dummy variables were generated. Squamous was chosen as reference and the tree dummy variables allowed estimating the hazard ratio of Adeno *vs.* Squamous, Small *vs.* Squamous and Large *vs.* Squamous, respectively.categ0.054 .2 TdWhoseCoxhe mode (iusined for the multivariate analysiss ie was)Tj20095 Tw -numbape oe regreoredn mas noo badeql into thnumbape o(variablus,)0.5t it
Categorical
modeahe possible
2l
p-value= 0.3145, 0.9134, 0.8163d foKarnofskmmP perfoaren s
arrepoveuded iTriabl1e Wariol

In this case it could be preferable to show results for a "clinically meaningful" increase (e.g. 10 units increase: hazard ratio =0.720).

e predictive ability of a Cox model result can be evaluated by the area under ROC curve for censored survival data, named "Harrell's C index". is index ranges between 0.5 (lack of predictive ability) and 1 (perfect predictive ability) [12,22]. e estimated predictive ability was 0.74 for the Cox model results reported in Table 1. e model included both variable whose impact was "statistically signi cant" and variable whose impact was not "statistically signi cant". Considering a model including only statistically signi cant variables (Karnofsky Performance Score and Cell Type), the model predictive ability was 0.73, suggesting a negligible improve provided by the non-signi cant variables.

When Karnofsky Performance Score was excluded the predictive ability was 0.61 and when Cell Type was excluded the predictive ability was 0.71, suggesting a contribution of Karnofsky Performance Score greater than the contribution of Cell Type. When both the above mentioned variables were excluded the predictive ability was negligible (0.52).

C

e above considerations concern only the "most frequent discussed items". We hope this paper could stimulate clinicians to read accurately statistical analysis results and avoiding to decide only on the basis of p-values. e cooperation between clinicians and biostatisticians could help clinicians to be more con dent with statistical methods and could provide insights to evaluate the relevance of results taking into account also an adequate statistical analysis.

e attitude of some clinicians is to privilege papers where data are presented with currently adopted statistical methods because they believe this is always the best approach. is is not necessarily true. In fact some studies could require alternative statistical modelling

20. Tsiatis AA, Davidian M, Zhang MLX (2008) Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet þexible approach. Statistics in Med 27: 4658-4677.

21. Kahan BC, Jairath V, Doré CJ, Morris TP (2014) The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. Trials 15: 139.

22. Harrell FE, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Med 15: 361-87.

23. Shepherd BE, Rebeiro PF, Caribbean Central, South America network for HIV epidemiology (2017) Brief Report: Assessing and Interpreting the Association between Continuous Covariates and Outcomes in Observational Studies of HIV Using Splines. JAIDS J Acquired Immune Deýciency Syndromes. 74: e60-e63.

24. Beuscart JB, Pagniez D, Boulanger E, de Sainte FCL, Salleron J, et al. (2012). Overestimation of the probability of death on peritoneal dialysis by the Kaplan-