Author's response to reviews

Title: Variables with time-varying effects and the Cox model: Some statistical concepts illustrated with a prognostic factor study in breast cancer

Authors:
Carine A Bellera (bellera@bergonie.org)
Gaëtan MacGrogan (macgrogan@bergonie.org)
Marc Debled (debled@bergonie.org)
Christine Tunon de Lara (tunon@bergonie.org)
Véronique Brouste (brouste@bergonie.org)
Simone Mathoulin-Pélissier (mathoulin@bergonie.org)

Version: 2 Date: 21 December 2009

Author's response to reviews: see over
Dear Editor,

We thank you and the reviewers for your helpful review of our paper

Variables with time-varying effects and the Cox model:
Illustration with a prognostic factor study in cancer.

We have prepared a revised version, which takes into account the comments of the reviewers. All changes are highlighted.

Below, we describe these changes.

We thank you for considering our resubmission, and look forward to hearing from you.

Yours sincerely,

Carine A. Bellera, PhD (Corresponding author)
Gaëtan MacGrogan, MD
Marc Debled, MD
Christine Tunon de Lara, MD
Véronique Brouste
Simone Mathoulin-Pélissier, MD, PhD
Our responses to the comments of the reviewers:

Comments from Reviewer 1

1) graphical procedure: I think it should be mentioned that even though this is potentially a very useful technique that is easy to understand it is quite difficult to do in practice, because it is hard to assess if departure is problematic or within standard variation. Providing standard errors to these plots as in Martinussen and Scheike (2006) would help this problem. Unfortunately, this is difficult to do and not standard in any computer program.

We agree with the reviewer. Indeed, as we had already written in our manuscript: [Page 7-8] These visual methods are simple to implement but have limitations. When the covariate has more than two levels, Kaplan-Meier plots are not useful for discerning non-proportionality because the graphs become to cluttered (10). Similarly, although the PH assumption may not be violated, the log-minus-log curves are rarely perfectly parallel in practice, and tend to become sparse at longer time points, and thus less precise. It is not possible to quantify how close to parallel is close enough, and thus how proportional the hazards are. The decision to accept the PH hypothesis often depends on whether these curves cross each other. As a result, the decision to accept the PH hypothesis can be subjective and conservative (26), since one must have strong evidence (crossing lines) to conclude that the PH assumption is violated.

For additional clarification, we have integrated the remark of the reviewer regarding the work by Martinussen et al. (page 8):
In view of these limitations, some suggest providing standard errors to these plots (24). This approach however can be computationally intensive and is not directly available in standard computer programs.

2) Modelling the time varying effect with a fixed function is a very useful and simple approach and if this is done carefully this procedure will work very well and leads to a p-value that summarizes the seriousness of the problem. Clearly this p-value can be somewhat misleading at it depends on the choice of the specific function.

We agree with these limitations, and that the p-value can be misleading, as stated in our manuscript: [Page 9] This approach however should be used with caution. Indeed, if the function of time selected is mis-specified, the final model will not be appropriate. This is a disadvantage of this method over more flexible approach.

3) Scaled schoenfeld residuals that are smoothed as in Therneau and Grambsch is also a very useful technique, the problem here is that the amount of smoothing is underlying this procedure and that the variance estimates are somewhat hard to interpret. Therefore the procedure is not fully developed. I agree, however, that the smooths will give a good idea about the possibly time varying effects. An alternative to this are the cumulative residuals by lin, wei and ying that are considered in cortese et al. This procedure is implemented in sas and the timereg package for R. This amounts to looking at the cumulative schoenfeld residuals, or equivalently the observed score processes. These can be provided with correct p-values, but the problem here is that local departure can be hard to detect, because of the cumulative nature of things. I think this procedure is also quite useful and among the best that are available, and certainly worth mentioning.

We agree with the reviewer and now provide a discussion on cumulative residuals → See our response (part (ii)) to the third comment of the second reviewer (page 5 of this document).
I am very sceptical about the analyses that are partitioned according to the length of the survival time, clearly this is not the thing to do. The interpretation is now conditional on the length of the survival time, which makes it hard to use in any practical setting. To see how effects change over time it is much more constructive to estimate different effects for different intervals on the time-axis as the authors also advocate.

The worked example nicely illustrates the different methods in use, and I fully agree with all final remarks about the importance of checking the proportionality assumption carefully. As the authors also point out, ignoring this can lead to incorrect conclusions, and lacking understanding of the underlying subject matter.

Comments from Reviewer 2

1) It is not clear what way the manuscript relates to the existing literature promoting the same goal (3-8) and what new information it brings. The role of this manuscript in this respect should be clearly defined.

As initially mentioned in our manuscript, available articles on the Cox in the medical literature are general overviews of this model [Page 3]. Following the remark of the reviewer, we now provide additional clarification on how our manuscript relates to the existing literature:

[Page 3-4] [...] topics covered included summarizing survival data, testing for a difference between groups, presenting existing statistical models, or assessing the adequacy of a survival model. Others works focused on providing definition of specific survival endpoints (9), or on the quality of reporting of survival events (3).

Assessing whether the assumption of proportional hazards is a central theme in survival analysis, and as such is discussed in several statistical textbooks (10-13) as well as in the general statistical literature (14-16). To our knowledge however, this topic has been discussed in few medical journals. Importantly, this strong assumption does not seem to be systematically assessed. For illustration, a recent review of clinical trials with primary analyses based on survival end points showed that only one of the 64 papers that used a Cox model mentioned verifying the PH assumption (3). Our objective is to inform clinicians, as well as those who read and write manuscripts in medical journals, [...]
2) Example: the manuscript intends to teach non-statisticians how to properly analyse data, and should thus be
perfect in this sense. In particular, more care should be taken in case of age - as this is a survival study with a
long follow-up time, age is bound to have an important effect and categorizing it into two groups seems unwise.
The poor modelling of age could also be the cause of non-proportionality of some other variables.

It is unclear whether the remark of the reviewer relates to
- the dichotomization of age in our analysis,
- or the choice of our time scale.
Since, we are not sure which specific issue the reviewer was referring to, we will address both issues.

1) With regards to the dichotomization of age in our analysis, we wish to reassure the reviewer:
(i) The decision to categorize age into two subgroups was initially driven by clinical considerations. Our primary
endpoint is time to metastases. It is well known that breast cancer in young women is usually more aggressive
with a different management of their disease. Thus, we wanted to make a clear distinction between young (less
than 40) and older women.

(ii) To ensure that this dichotomization was justified from a statistical viewpoint, we first categorised age into 5
distinct age subgroups:
- 40 or younger (N= 76)
- 41 to 50 (N = 224)
- 51 to 60 (N=281)
- 61 to 70 (N= 272)
- 71 and older (N=126).

Using this classification, we first fitted a univariate Cox PH model to model time to metastases [Data not shown
in the manuscript]. Outputs from the statistical software are provided below:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Hazard Ratio</th>
<th>95% Hazard Ratio</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>age_40_50 (ref: younger than 40)</td>
<td>0.491</td>
<td>0.325 - 0.741</td>
<td>0.0007</td>
</tr>
<tr>
<td>age_50_60 (ref: younger than 40)</td>
<td>0.475</td>
<td>0.319 - 0.708</td>
<td>0.0003</td>
</tr>
<tr>
<td>age_60_70 (ref: younger than 40)</td>
<td>0.465</td>
<td>0.311 - 0.695</td>
<td>0.0002</td>
</tr>
<tr>
<td>age_70plus (ref: younger than 40)</td>
<td>0.367</td>
<td>0.217 - 0.619</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

As can be seen from the output, the hazard ratios corresponding to each age category were relatively similar.
Moreover, a test based on the scaled Schoenfeld residuals did not show any evidence of non-proportionality.
Thus, combining all women older than 40 years into a single subgroup appeared statistically justified. The result
of this updated univariate analysis is shown below:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Hazard Ratio</th>
<th>95% Hazard Ratio</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>age_40_50 (ref: younger than 40)</td>
<td>0.462</td>
<td>0.325 - 0.657</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Note that in our manuscript, we presented our results slightly differently as the reference category included
women older than 40. In such case, the HR is about 1/0.462 = 2.16 (close to the hazard ratio reported for the
multivariate analysis – See Table 2).

For these reasons, our results suggest that age was properly modelled, and as such, should not be the cause of
non-proportionality of other variables.

2) Finally, although not clearly stated, the remark of the reviewer might relate to the choice of the time scale.
In our example, time since surgery was used as the time-scale of the survival analysis. There are two choices
of time-scale for a Cox regression analysis: time-on-study, such as time since surgery in our example, or age.
Because age is a common confounder in survival analyses, researchers have suggested that age would be the
most appropriate time scale [Korn, Am J Epidemiol 1997]. Additionally, with age as the time-scale, control for
the calendar-period and/or birth cohort effects can be achieved by stratifying the model on birth cohort. As argued by Pencina et al., the question of the appropriate choice of time scale can be addressed on two levels [Pencina, Stat Med 2007]. First, which time scale is more appropriate in the intuitive sense? Second, do models reporting results from analyses based on age or time-on-study produce mathematically equivalent results? When following-up a healthy population, for example to study the development of Alzheimer disease, using age as the time scale would be preferable as one would expect the hazard to change more as a function of age than as a function of time-on-study [Korn, Am J Epidemiol 1997]. However, in a randomized study or in the natural history of disease, time since randomization or diagnosis (or surgery in our example) would be used. With respect to the second question, Pencina et al. investigated the effect of the time scale selection in a simulation study. Their results suggested that if adjustments at entry are made, there was very little evidence that there exists any practically meaningful difference in the estimated regression coefficients depending on the time scale.

The issue of selecting the appropriate time scale is particularly interesting; however, we fear that including a discussion on this theme might overload the manuscript.

3) Methods: The method of Schoenfeld residuals is rather poorly presented. In particular, it should be stressed that:

(i) graphically, this method is far more reliable and easy to interpret than the cloglog curves.
(ii) a lot more work has been done on the subject of PH testing based on Schoenfeld residuals (e.g. methods using cumulative residuals, Brownian motion ...). Presenting only a test of linear trend is an oversimplification. The test could be described to start with for pedagogical reasons, but should be expanded on. At least the methods used in the described software should be mentioned.

(i) We agree with the reviewer and have now clearly indicated that plotting the Schoenfeld residuals is far more reliable and easy to interpret than the cloglog curves. 

(ii) We are aware that there exist other tests based on residuals. However, we were initially not keen on expanding this section as 1) to our impression, among all residuals, the Schoenfeld ones appear to be the most often used for testing non-proportionality, and 2) introducing additional residuals in this manuscript would definitely require more statistical background from the reader and might make this manuscript more statistically complex while it was initially aimed at non-statisticians. 3) Graphically, the resulting plot is more difficult to interpret than plots based on Schoenfeld residuals (See comment 3 of reviewer 1: ... but the problem here is that local departure can be hard to detect)

However, following the suggestions of the two reviewers, we have now expanded this section and discuss briefly tests based on cumulative residuals:

There are other residuals of interest in the Cox model which include the martingale, deviance, score and Schoenfeld residuals. While we only presented a test based on the Schoenfeld residuals to assess proportional hazards, the cumulative sum of Schoenfeld residuals, or equivalently the observed score process can also be used to this extent (26). Under proportional hazards, the resulting curve should be a Brownian bridge (random walk) and one can then obtain confidence bands allowing testing for significance. Visualizing and interpreting the plot can however be difficult to read (10). Tests based on either the Schoenfeld or cumulative residuals can both be easily implemented in standard statistical packages. The other types of residuals can also be particularly useful as additional regression diagnostics for the Cox model. For illustration, martingale residuals are useful for determining the functional form of a covariate to be included in the model and deviance residuals can be used to examine model accuracy. Additional details can be found in (10;27).
4) Discussion: The unsatisfactory presentation of the methods (in particular the Schoenfeld residuals) makes the manuscript inconclusive and does not simplify the work of a data analyst. I believe that Schoenfeld residuals could be presented as the method to be used and the additional gain of the other methods could be described as relative to the Schoenfeld residuals based methods. For example, it should be stressed that the graphical test using Schoenfeld residuals is surely far more appropriate than the KM method (or its transform) and can be applied with all types of variables.

We partly agree with the reviewer:

1) On one hand, we agree that, as a graphical test, graphical test using Schoenfeld residuals is surely far more appropriate than the KM method (or its transform) and can be applied with all types of variables. As stated in our answer to the third comment of this reviewer, we have now added the following comment:

[Page 10] Graphically, this method is much more reliable and easy to interpret than plotting the \( \log(-\log(S(t))) \) functions presented earlier.

Moreover, we have also added the following sentence to our discussion to reinforce this idea:

[Page 13] On the other hand, the graphical test using Schoenfeld residuals is more appropriate than the Kaplan-Meier method (or its transform) and can be applied with all types of variables.

2) Contrary to the reviewer however, we do not wish to present the test based on the Schoenfeld residuals as the best test. As mentioned in our discussion, [Page 14]:

It is difficult to propose definite guidelines for the best strategy for testing for non-proportionality. Each method has its advantages and limitations, and depending on the study objective some approaches might be preferred. Before performing statistical modelling, the study objectives should be clearly stated in advance, as well as the statistical tests that will be employed.

As suggested by the simulation study by Ng’Andu (Statistics in Medicine 1997):

[Page 14] Tests requiring partitioning of the failure time have less power than other tests, while tests based on time-dependent covariates or on the Schoenfeld residuals have equally good power to detect non-proportionality in a variety of non-proportional hazards and are practically equivalent (13).

Thus, depending on the study settings, methods other than the Schoenfeld residuals, would be appropriate.

5) Page 9, line 4: the parameters associated with interactions should probably be denoted with gamma

We corrected the text as suggested by the reviewer.

6) Page 12, second paragraph: This should be rephrased. If HR is increasing over time, the estimated coefficient assuming PH is overestimating at first and underestimating later on.

We corrected the text as suggested by the reviewer.

7) Discretionary Revisions

Page 4: The introduction to survival analysis is rather poor. In particular, the presence of censoring as the defining property of the survival data is not even mentioned. I understand it is hard to describe the whole field in a paragraph, it is therefore very important to make this paragraph more focused. Another option is to skip it and only focus on the Cox model.

The introduction now includes the definition of censoring.

[Page 4] In many studies, the primary variable of interest is a delay, such as the time from cancer diagnosis to a particular event of interest. This event may be death, and for this reason the analysis of such data is often referred to as survival analysis. The event of interest may not have occurred at the time of the statistical analysis, and similarly, a subject may be lost follow-up before the event is observed. In such case, data are said to be censored at the time of the analysis or at the time the
patient was lost to follow-up. Censored data still bring some information since although we do not know the exact date of the event, we know that the event occurred later than the censoring time. Adapted statistical methods are needed to account for the characteristics of survival data.

8) Page 10, middle paragraph: it could be noted that this method is not directly testing the PH assumption, and a different parametrization would be needed to perform a test.

We have now added this remark → Page 11.

9) Discussion: It could be mentioned that other assumptions (e.g. the linearity in case of continuous variables) and methods for simultaneously checking assumptions have been developed.

We have now added the following paragraph:

[Page 10] There are other residuals of interest in the Cox model which include the martingale, deviance, score and Schoenfeld residuals. While we only presented a test based on the Schoenfeld residuals to assess proportional hazards, the cumulative sum of Schoenfeld residuals, or equivalently the observed score process can also be used to this extent (26). Under proportional hazards, the resulting curve should be a Brownian bridge (random walk) and one can then obtain confidence bands allowing testing for significance. Visualizing and interpreting the plot can however be difficult to read (10). Tests based on either the Schoenfeld or cumulative residuals can both be easily implemented in standard statistical packages. The other types of residuals can also be particularly useful as additional regression diagnostics for the Cox model. For illustration, martingale residuals are useful for determining the functional form of a covariate to be included in the model and deviance residuals can be used to examine model accuracy. Additional details can be found in (10;27).