Author's response to reviews

Title: Classification tree analysis of second neoplasms after childhood cancer

Authors:

Janez Jazbec (janez.jazbec@mf.uni-lj.si)
Ljupco Todorovski (Ljupco.Todorovski@ijs.si)
Berta Jereb (BJereb@onko-i.si)

Version: 2 Date: 6 October 2006

Author's response to reviews: see over
Dear editor,
We are re-submitting a paper: Classification tree analysis of second neoplasms after childhood cancer (MS: 2025623405106663). Due to administrative reasons we are very late with our resubmission. We have received your comments on the 18/09, although your first e-mail was first sent 24/8. We hope that it is still not too late.

As for specific remarks from the editor:
1. We have asked prof. Giulio D'Angio, Emeritus Professor of Radiation Oncology at the University of Pennsylvania, and long time editor-in-chief of Medical and Pediatric Oncology Journal, for help on the linguistic issues of our paper. We have made changes according to his suggestions.
2. We have reformatted the abstract according to guidelines.
3. A statement on ethical issues has been inserted in Methods section, together with the number of approval of National Ethical Committee.
4. Declaration of author's interests is included.
5. We have included Author's contribution section.

Regarding reviewer's reports we would like to answer to their comments point by point.

Reviewer № 1 – Gregory Reaman:
• Difference in incidence of SMNs in our study is attributed to study design (population based study, defined time at risk from the diagnosis of primary cancer) and to the fact that our study covered the time period from 1960 –2000. In the early seventies the cure rate for childhood cancer was still low, most of the patients died of primary cancer before they had a chance to develop SMN. We have addressed this question in an added paragraph at the beginning of discussion section.
• The specific detailed data on therapy was not included in our database and attempts to expand the therapy data would result in incomplete data, which as such would not contribute much to final result. We agree that in case of complete data on therapy the results would be of great interest, but due to design of our study we were not able to recover detail on therapy of patients treated in the early period of our study.
• We agree that classification tree analysis does not adequately consider the time course, based on specific exposures for which other methods may be more suitable. Although genetic susceptibility is considered an important factor for risk for SMN, it was not possible to include it as a parameter in our analysis as the data for the whole cohort was impossible to obtain.

Reviewer № 2 – Phyllis Gimontty
The authors have presented a classification of patients based on the risk of second neoplasms. Except for having had radiation therapy for the first primary cancer, none of the other factors identified appear to define groups with significantly differential risk. One explanation for the lack of identifying significant differential risk is that the explanatory variables may not be predictive of the primary outcome; this is not discussed by the investigators. The identified risk groups need to be validated in another dataset.
We improved the explanation of how we approached the tree induction task from data. In the extended explanation, we clarified that our aim is not to accurately predict secondary neoplasm occurrence. Instead, the induced tree identifies two groups of patients with significantly higher occurrence rate for the secondary neoplasm. The two groups can be identified using a combination of values of the independent (explanatory) variables. We then discuss the significance of the result with putting it into context of existing results about secondary neoplasms occurrence.

1. Page 6; lines 20-23. What was the justification for the choice of the three misclassification costs?

We clarified this in the extended Section on Classification Tree Analysis, see esp. the newly added text on page 7.

2. Page 7 lines 10-12; misclassification costs need to be selected based on actual or estimated costs, otherwise it is a somewhat artificial classification and highly subjective. Do these costs relate to “real” misclassification costs”.

Since there is no clear basis for selecting the misclassification costs based on actual or estimated costs (this is a difficult problem in medical domains), we use a cross-validation procedure to estimate number misclassifications on test data and then based on these estimates select optimal misclassification costs.

3. Page 7 line 23-Page 8. The rates for those with and without radiation have confidence intervals that don’t overlap. However, confidence intervals for the rates for the second split into HD versus others overlap considerably. It is likely that confidence intervals for groups further down the tree will also overlap. What is the justification for statements about differential risk?

We do not claim that every split in the tree is significant. However the two identified risk groups have significantly different outcome occurrence ratio with non-overlapping confidence intervals when compared to those of the whole population. All our result interpretation is based on this

4. The authors present three scenarios that depend on misclassification costs. There was no discussion of the implication of this choice on the interpretation of the resulting risk and/or risk groups. one can always set the cost high enough so that the misclassification of "cases" is minimized, but is it meaningful.

See the reply under 2.

Sincerely,
Janez Jazbec