"A balanced hazard ratio for risk group evaluation from survival data"

Branders, Samuel ; Dupont, Pierre

ABSTRACT

Common clinical studies assess the quality of prognostic factors, such as gene expression signatures, clinical variables or environmental factors, and cluster patients into various risk groups. Typical examples include cancer clinical trials where patients are clustered into high or low risk groups. Whenever applied to survival data analysis, such groups are intended to represent patients with similar survival odds and to select the most appropriate therapy accordingly. The relevance of such risk groups, and of the related prognostic factors, is typically assessed through the computation of a hazard ratio. We first stress three limitations of assessing risk groups through the hazard ratio: 1) it may promote the definition of arbitrarily unbalanced risk groups, 2) an apparently optimal group hazard ratio can be largely inconsistent with the p-value commonly associated to it, 3) some marginal changes between risk group proportions may lead to highly different hazard ratio values. Those...

CITE THIS VERSION


Le dépôt institutionnel DIAL est destiné au dépôt et à la diffusion de documents scientifiques émanants des membres de l'UCLouvain. Toute utilisation de ce document à des fin lucratives ou commerciales est strictement interdite. L'utilisateur s'engage à respecter les droits d'auteur lié à ce document, principalement le droit à l'intégrité de l'oeuvre et le droit à la paternité. La politique complète de copyright est disponible sur la page Copyright policy.

DIAL is an institutional repository for the deposit and dissemination of scientific documents from UCLouvain members. Usage of this document for profit or commercial purposes is strictly prohibited. User agrees to respect copyright about this document, mainly text integrity and source mention. Full content of copyright policy is available at Copyright policy.
A balanced hazard ratio for risk group evaluation from survival data: supplementary materials

BRANDERS SAMUEL and DUPONT PIERRE

Université catholique de Louvain – ICTEAM/Machine Learning Group – Belgium

1 Summary

Further experiments are reported here to extend the analyses of the main manuscript to other prognosis models and additional cancer studies. Those results are fully consistent with the proposed conclusions and illustrate that they are not specific to some clinical studies or prognostic indexes. They show that the original hazard ratio (HR) is largely inadequate to evaluate risk groups for 3 complementary reasons. HR can be highly sensitive to marginal changes in the proportions between risk groups, it can be artificially increased by considering extremely unbalanced groups. Selecting a cut-off on risk scores while maximizing HR on a training set is consequently performing poorly while using a prognosis model and its selected cut-off on independent validation samples. The proposed BHR metric fixes all the above issues while keeping an interpretation similar to the original HR.

2 Data sets

The results presented here are produced from clinical studies on breast, colon and ovarian cancers. All data sets are publicly available online from the GEO database of NCBI. We first recall the breast cancer data also used in the main manuscript. The same data will be used here with additional prognosis models. Next, we describe additional data from colon and ovarian cancer studies.

The Veridex (VDX) data sets contains 344 patients with primary breast cancer [Wang et al. 2005]. All patients were node negative and untreated. The VDX data set is available on GEO database with the accession numbers GSE2034 and GSE5327. The data set TBG contains 198 patients with a similar profile. The GEO accession numbers of this data sets is GSE7390. The third breast data set (UNT) originally contains 189 patients with primary breast cancer. To have comparable breast data sets, we choose to reduce the UNT data sets to 84 node negative and untreated patients. The duplicated patients appearing in the VDX and TBG data sets were also removed. In those three studies, distant metastasis is used as end point and gene expression data are measured on the Affymetrix HGU133a microarray platform. The three data sets were summarized according to the MAS5.0 procedure and represented in log2 scale.

Two data sets comes from colon cancer studies. The first data set contains 566 patients [Marisa et al. 2013] and is available on GEO with the accession number GSE39582. The survival end point for this data set is relapse free survival. The second data set contains 177 patients [Freeman et al. 2012]. The survival end point is distant metastasis free survival and it is available on GEO with the accession number GSE17536. For
those data sets, the gene expression data are measured on the Affymetrix HGU133plus2 microarray platform and summarized with the RMA procedure.

The ovarian data set contains a cohort of 285 patients with epithelial ovarian, primary peritoneal, or fallopian tube cancer [Tohill et al. (2008)]. The survival end point is relapse free survival. The gene expression data are measured on the Affymetrix HGU133plus2 microarray platform and summarized with the GCRMA procedure. This data set is available with the accession number GSE9899.

3 Prognostic models

In a common setting, a prognostic model aggregates the influence of several potential co-factors in a continuous risk score for each patient. Next, one or several cut-off values are chosen to cluster patients into discrete risk groups.

In the main manuscript, we discuss the influence of the cut-off choice on the performances of the Gene76 prognosis model [Wang et al. (2005)] for breast cancer prognosis. We consider first other breast cancer prognosis models. Gene70 specifically refers to the model described in [van ’t Veer et al. (2002)]. We also build two prognostic models referred as Cox76 and CoxTtest. Cox76 relies on the same set of genes as Gene76 but is estimated as a unique multivariate Cox model. In contrast, Gene76 (presented in the main manuscript) is formulated as a sum of univariate Cox models specific to sub-populations of patients according to their estrogen receptor status.

Finally, CoxTtest is a multivariate Cox model built on the 100 most differentially expressed genes in a given study. The estimation of such a gene signature requires a critical time point to be chosen in order to define 2 conditions (event observed or not before the critical time point). We chose here 5 years after treatment as commonly accepted for breast cancer studies. Since we do not have access to alternative prognosis models for colon and ovarian cancers, we will report results only with such CoxTtest models in those cases. Of course, the parameters of such models are specific to each pathology and estimated on the respective training sets considered (see below).

4 Influence of the cut-off choices to assess risk groups

In the main manuscript, we study the evolution of the various performance metrics while varying the proportions of samples within each risk group. Such variations are obtained by varying the respective cut-off values on risk scores. In this section, those results are extended to additional data sets and prognostic indexes. To ease the comparison between the different metrics, they are all rescaled between 0 and 1.

In figure 1, the comparison between HR, C-index, BCR and logrank is made on the 3 breast cancer data sets pulled together (see section 2). The four different prognostic models are used to produce the risk scores: Gene76 (a), Gene70 (b), Cox76 (c) and CoxTtest (c). Figure 2 contains the same graphs but with the balanced hazard ratio (BHR) replacing the C-index.

Figure 3 and 4 present the same kind of results but on the colon and ovarian cancer data sets. For these data sets, we only report results with CoxTtest, the other prognosis models being specific to breast cancer.

The BCR (computed at 5 years) is not appropriate for the ovarian cancer, as most patients experienced the event before this time threshold. That is why the BCR curve

\footnote{For readability, we do not report the SEP metric as it behaves essentially as the logrank.}
does not have the same shape as the logrank and the balanced hazard ratio in this case. Of course, one could consider an earlier time point after treatment but those results precisely illustrate that the BCR measure could be highly sensitive to this choice which needs to be adapted to each pathology and, possibly, to each collection of samples.

As observed in the main manuscript, the C-index behaves very similarly to the hazard ratio: one can trivially optimize them while considering artificially unbalanced groups, they exhibit many local optima and sharp fluctuations for marginally different group proportions. In contrast, the logrank test, the BCR and the BHR look more appropriate as they offer quite smoother curves with a similar global optimum observed for more balanced groups. We note that this optimum is data dependent and needs not correspond exactly to a 50%/50% balance between groups.
Figure 1: Evolution of HR, C-index, BCR and logrank on the breast data (VDX, TBG and UNT together) while varying the proportions in each risk group through adjusting the cut-off on risk scores.
Figure 2: Evolution of HR, BHR, BCR and logrank on the breast data (VDX, TBG and UNT together) while varying the proportions in each risk group through adjusting the cut-off on risk scores.
Figure 3: Evolution of HR, C-index, BHR, BCR and logrank on the colon data while varying the proportions in each risk group through adjusting the cut-off on risk scores.
Figure 4: Evolution of HR, C-index, BHR, BCR and logrank on the ovarian data while varying the proportions in each risk group through adjusting the cut-off on risk scores.
5 Cut-off choice and risk group prediction

In this section, we analyze the influence of choosing specific cut-off values associated to the risk scores produced by various prognosis models. Those cut-off values are chosen on some fraction of the data considered as a training set and the predictive performances of the associated prognosis model with the selected cut-off are evaluated on independent samples serving as validation. In the main manuscript, we show the benefits in terms of predictive performance of defining a cut-off maximizing BHR instead of HR on the training set. We validate here this result on others data sets and with additional prognostic models.

The available data for each disease is first split between a training and a validation set. For the breast cancer studies, we chose the VDX data set as training and the two others as validation since the Gene76 prognosis model was itself estimated from VDX. This splitting avoids an optimistic bias when assessing the Gene76 model by guaranteeing that the validation set is made of independent samples not used before. For the colon and ovarian cancer studies, we randomly select two thirds of the data as training and the last third as independent validation.

Prognostic models are first estimated on the respective data defined as training sets (excepted for the Gene70 and Gene76 models which are given). Next, a cut-off value is chosen on the training set to define risk groups from the estimated risk scores. We consider in particular a cut-off value maximizing either the HR or the BHR on the training set.

The same prognostic models and selected cut-off values are then used on the validation data to evaluate the predictive performances obtained. Risk group prediction is assessed through the five performance metrics discussed in the main manuscript: HR, BHR, C-index, BCR and logrank.

Figures 5, 6, 7 and 8 compare the performances in risk group prediction for the HR and BHR cut-off on the breast data. Each of those four figures is made with one of the prognostic models presented in section 3.

Figure 9 presents the prognostic performances of the two cut-offs on an independent validation set of colon cancer data. Similar results are presented on ovarian cancer data in figure 10.

The cut-off chosen to maximize BHR on the training give consistently better results on the validation samples than the cut-off maximizing the HR on the training. This is systematically true in terms of BCR, logrank and balanced hazard ratio. The BHR cut-off is also better in hazard ratio and C-index in the most cases. As expected, the predicted risk groups are much more balanced when defined according to a cut-off on risk scores maximizing the BHR.
Figure 5: Prognostic performances on an independent validation set (TBG and UNT datasets) according to two cut-off choices on the training (VDX): largest HR \(a\) and largest BHR \(b\). Prognostic model: Gene76.
Figure 6: Prognostic performances on an independent validation set (TBG and UNT datasets) according to two cut-off choices on the training (VDX): largest HR (a) and largest BHR (b). Prognostic model: Gene70.
Figure 7: Prognostic performances on an independent validation set (TBG and UNT datasets) according to two cut-off choices on the training (VDX): largest HR [a] and largest BHR [b]. Prognostic model: Cox76.
Figure 8: Prognostic performances on an independent validation set (TBG and UNT datasets) according to two cut-off choices on the training (VDX): largest HR (a) and largest BHR (b). Prognostic model: CoxTtest.
Figure 9: Prognostic performances on an independent validation set (colon data) according to two cut-off choices on the training: largest HR (a) and largest BHR (b). Prognostic model: CoxTtest.
Figure 10: Prognostic performances on an independent validation set (ovarian data) according to two cut-off choices on the training: largest HR (a) and largest BHR (b). Prognostic model: CoxTtest.
References


