survcomp: an R/Bioconductor package for performance assessment and comparison of survival models

Markus S. Schröder*, Aedín C. Culhane, John Quackenbush and Benjamin Haibe-Kains*
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

ABSTRACT
Summary: The survcomp package provides functions to assess and statistically compare the performance of risk prediction models. It implements state-of-the-art statistics to (i) measure the performance of risk prediction models; (ii) combine these statistical estimates from multiple datasets using a meta-analytical framework; and (iii) statistically compare the performance of competitive models.

Availability: The R/Bioconductor package survcomp is provided open source under the Artistic-2.0 License with a user manual containing installation, operating instructions and use case scenarios on real datasets. survcomp requires R version 2.13.0 or higher: http://bioconductor.org/packages/release/bioc/html/survcomp.html

Contact: bhaibe@jimmy.harvard.edu; mschroed@jimmy.harvard.edu

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on May 20, 2011; revised on August 26, 2011; accepted on September 3, 2011

1 INTRODUCTION

Building risk prediction or survival models is an important area of research, especially in cancer where gene expression signatures are used to predict risk of metastasis, response to therapy and overall survival. However, assessing the relative performance of such models is complex due to the lack of standards regarding the best criterion to use in survival analysis (Table 1; Haibe-Kains et al., 2008). Although a number of model performance estimators have been described (Cox, 1972; Graf et al., 1999; Harrell et al., 1996; Heagerty et al., 2000; Royston and Sauerbrei, 2004; Verweij and Houwelingen, 1993), they are not widely used. This is partly because these are implemented in many different R packages that use heterogeneous interfaces, which makes it difficult for the non-specialist to easily use or compare the performance of these models. Another challenge in assessing performance of expression-based prediction models is a lack of power due to small sample size. Meta-analytical methods could leverage power from the numerous microarray datasets that are now publicly available by summarizing model performance estimated in multiple-independent studies. Moreover, because they enable efficient joint analysis of multiple datasets, such an analytical framework reduces the risk of artifactual discoveries that are due to bias or confounding factors that may be present in one dataset. This is particularly important when comparing competitive risk prediction models; often authors claim better performance of a new model without properly assessing whether a model significantly outperforms its competitors. To the best of our knowledge, there is no commercial or open-source tool to enable statistical comparison of risk prediction models in a meta-analytical framework.

To address these issues, we developed a new Bioconductor package, survcomp, which implements several performance criteria for risk prediction models (Table 1), together with meta-analytical methods that enable combination of performance estimations from multiple-independent datasets [fixed- and random-effects models (Cochrane, 1954); (1999) and implemented in survcomp may be sensitive to the study-specific censoring distribution, therefore we are working to implement a modified concordance index by Uno et al. (2011) that avoids this problem and which should be available in the next release of survcomp.

Although the performance criteria that are implemented in survcomp are mostly available in other R packages (except the D index which, to the best of our knowledge, is only in survcomp), our package provides a common interface to facilitate easy use of all these estimators. Moreover, with the exception of qmplot by Broeze et al. (2009) and survJamda by Yasrebi (2011), few R packages perform meta-analysis of survival data. survcomp provides a uniform interface to simplify the use of performance assessment and statistical comparison of risk prediction models, and provides new R functions to statistically compare these in

Table 1. Functions in survcomp to measure the performance of risk prediction models

<table>
<thead>
<tr>
<th>Name</th>
<th>Function in survcomp</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concordance index</td>
<td>concordance.index</td>
<td>Harrel et al. (1996)</td>
</tr>
<tr>
<td>D index</td>
<td>d.index</td>
<td>Royston and Sauerbrei (2004)</td>
</tr>
<tr>
<td>Hazard ratio</td>
<td>hazard.ratio</td>
<td>Cox (1972)</td>
</tr>
<tr>
<td>Brier score</td>
<td>sbrier.scoreprob</td>
<td>Graf et al. (1999)</td>
</tr>
<tr>
<td>Cross-validated</td>
<td>cvpl</td>
<td>Verweij and Houwelingen (1993)</td>
</tr>
<tr>
<td>partial likelihood</td>
<td></td>
<td>Heagerty et al. (2000)</td>
</tr>
<tr>
<td>Time-dependent ROC curve</td>
<td>tdroc</td>
<td></td>
</tr>
<tr>
<td>Kaplan-Meier curve</td>
<td>km.cox.plot</td>
<td>Kaplan and Meier (1958)</td>
</tr>
<tr>
<td>Forestplot</td>
<td>forestPlot.surv</td>
<td>Lewis and Clarke (2001)</td>
</tr>
</tbody>
</table>

*To whom correspondence should be addressed.
The prognostic ability of gene expression of AURKA, a single patient with poor survival, concordance index the performance varies between datasets, all models yielded highly approach (Cochrane, 1954). As can be seen in Figure 1, although meta-estimate using the well-established random-effects model function for each model in each dataset separately and used the described in the Supplementary Material.

risk scores were calculated using the Bioconductor package GENIUS (Haibe-Kains et al., 2010). The NPI, GGI and GENIUS proliferation-related gene, was compared with the Nottingham Prognostic Index (NPI, Galea et al., 1992) clinical model for prognosis, and to risk prediction scores from two published multigene prognostic signatures; GGI (Sotiriou et al., 2006) and GENIUS (Haibe-Kains et al., 2010). The NPI, GGI and GENIUS risk scores were calculated using the Bioconductor package genefu. Each score was computed in five publicly available datasets described in the Supplementary Material.

To compare the prognostic ability of these four different risk prediction models, we estimated the concordance index for each model in each dataset separately and used the function combine.est to compute the corresponding overall meta-estimate using the well-established random-effects model approach (Cochrane, 1954). As can be seen in Figure 1, although the performance varies between datasets, all models yielded highly significant overall prognostic value (high-risk predictions represent patient with poor survival, concordance index > 0.5, one-sided $P < 0.001$, see Supplementary Material for R code). AURKA, the single proliferation gene, was the worst predictor of survival (concordance index of 0.64), whereas GENIUS, the risk prediction model taking into account the breast cancer molecular subtypes, was the best (concordance index of 0.69). The continuous risk prediction of NPI, the traditional clinical model which combines nodal status, tumor size and histological grade, yielded a relatively high concordance index (concordance index of 0.66).

To identify the best risk prediction model(s), we statistically compared their concordance indices using the function cindex.comp.meta (Table 2). Concurring with Haibe-Kains et al. (2010) we observed that GENIUS outperforms AURKA, GGI and NPI (uncorrected $P < 0.10$; Table 2). However, when $P$-values are corrected for multiple testing (Holm’s method), no $P$-value remains significant suggesting that a larger meta-analysis is required to definitively claim the superiority of one classifier over another. This also suggests that prognostic clinical models such as NPI are still extremely competitive compared with more complex gene signatures. Repeating this analysis using performance criteria in survcomp other than the concordance index, including the D index or hazard ratio (Table 1), leads to similar conclusions (examples are provided in the package user manual and documentation).

### Table 2. Statistical performance comparison for the risk prediction models used in our case study

<table>
<thead>
<tr>
<th></th>
<th>Models</th>
<th>NPI</th>
<th>AURKA</th>
<th>GGI</th>
<th>GENIUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPI</td>
<td>0.25</td>
<td>0.37</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AURKA</td>
<td>0.75</td>
<td>0.70</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GGI</td>
<td>0.63</td>
<td>0.30</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GENIUS</td>
<td>0.09</td>
<td>0.02</td>
<td>0.033</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$P$-values are computed using a one-sided paired Student’s $t$-test to test whether risk prediction models in rows are better than the ones in columns.

### 2 CASE STUDY

The prognostic ability of gene expression of AURKA, a single proliferation-related gene, was compared with the Nottingham Prognostic Index (NPI, Galea et al., 1992) clinical model for prognosis, and to risk prediction scores from two published multigene prognostic signatures; GGI (Sotiriou et al., 2006) and GENIUS (Haibe-Kains et al., 2010). The NPI, GGI and GENIUS risk scores were calculated using the Bioconductor package genefu. Each score was computed in five publicly available datasets described in the Supplementary Material.

To compare the prognostic ability of these four different risk prediction models, we estimated the concordance index for each model in each dataset separately and used the function combine.est to compute the corresponding overall meta-estimate using the well-established random-effects model approach (Cochrane, 1954). As can be seen in Figure 1, although the performance varies between datasets, all models yielded highly significant overall prognostic value (high-risk predictions represent patient with poor survival, concordance index > 0.5, one-sided $P < 0.001$, see Supplementary Material for R code). AURKA, the single proliferation gene, was the worst predictor of survival (concordance index of 0.64), whereas GENIUS, the risk prediction model taking into account the breast cancer molecular subtypes, was the best (concordance index of 0.69). The continuous risk prediction of NPI, the traditional clinical model which combines nodal status, tumor size and histological grade, yielded a relatively high concordance index (concordance index of 0.66).

To identify the best risk prediction model(s), we statistically compared their concordance indices using the function cindex.comp.meta (Table 2). Concurring with Haibe-Kains et al. (2010) we observed that GENIUS outperforms AURKA, GGI and NPI (uncorrected $P < 0.10$; Table 2). However, when $P$-values are corrected for multiple testing (Holm’s method), no $P$-value remains significant suggesting that a larger meta-analysis is required to definitively claim the superiority of one classifier over another. This also suggests that prognostic clinical models such as NPI are still extremely competitive compared with more complex gene signatures. Repeating this analysis using performance criteria in survcomp other than the concordance index, including the D index or hazard ratio (Table 1), leads to similar conclusions (examples are provided in the package user manual and documentation).

### 3 CONCLUSION

The R/Bioconductor package survcomp provides a uniform interface to an extensive set of performance assessment and statistical comparison methods for survival/risk prediction. It allows scientists
to easily implement large comparative studies integrating multiple independent datasets while providing statistical tools to identify the best model(s) as supported by the data under study.

**Funding**: National Human Genome Research Institute (1P50 HG004233, to M.S.S.); Fulbright Commission for Educational Exchange to (B.H.-K.); US National Institutes of Health (R01 LM010129-01, to B.H.-K. and J.Q.); Claudia Adams Barr Program in Innovative Basic Cancer Research (A.C.C. and J.Q.); Career Development grant (to A.C.C.) from DFCI Breast Cancer SPORE: CA 08939.

**Conflict of Interest**: none declared.

**REFERENCES**


