Continuous Biomarker Assessment by Exhaustive Survival Analysis

Dominic A. Pearce[1*], Ajit J. Nirmal[2], Tom C. Freeman[2], Andrew H. Sims[1]

[1]Applied Bioinformatics of Cancer, University of Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, Edinburgh, UK [2]Systems Immunology Group, Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG

*Corresponding author: dominic.pearce@ed.ac.uk

Abstract

Publicly available high-throughput molecular data can facilitate biomarker identification and evaluation, with common practice being to divide a dataset into good or poor prognosis groups at the median value of a quantitative marker. However, this approach ignores the biomarker's underlying distribution and potential confounding factors associated with outcome that will inevitably vary between cohorts. Here we present the *survivALL* R package (https://CRAN.R-project.org/package=survivALL) and associated web app *survivAPP* (https://survivapp.shinyapps.io/survivapp/), allowing researchers to generate visual and numerical comparisons of all possible points-of-separation and enable quantitative biomarkers to be reliably evaluated within and across datasets, independent of compositional variation. To compensate for multiple-testing a non-parametric bootstrap is applied, defining reliable confidence intervals to assess how a biomarker performs relative to chance. This approach represents a significant improvement over existing methodologies in stratifying patients using quantitative biomarker(s).

Biomarker performance is traditionally assessed by estimating survival benefit over time, often summarized as a Kaplan-Meier plot[1,2]. Whilst qualitative biomarkers can easily stratify cohorts into two or more groups, the question of how to split a cohort with a quantitative measure is considerably more challenging. The most obvious, common, yet arbitrary approach is to divide the cohort into two equally-sized groups at the median level. However, this median-split approach does not consider a marker's distribution, or any potential confounding factors that may affect a dataset's composition. Given the public availability of high-throughput molecular profiling datasets providing outcome data, the opportunity to evaluate biomarkers *in silico* is in principle becoming easier. In practice, however, these datasets, derived from cohorts of primary patient tissue, are molecularly heterogeneous in their compositions[3], and likely to exhibit varying proportions of multiple clinical factors associated with outcome (e.g. node and receptor statuses, grade, molecular subtype and age) which must be carefully accounted for.

Building upon previous best-of-split methods[4–6], we have developed *survivALL* (https://CRAN.R-project.org/package=survivALL), an R package implementation to exhaustively calculate and visualise survival statistics to assess the association between a continuous measure and outcome in a compositionally agnostic manner. Hazard ratio (HR) and p-value statistics are computed

for every point-of-separation possible, allowing significance, and the frequency of significant stratifications to be determined, regardless of confounding compositional effects. To correct for multiple testing, a non-parametric bootstrap is applied, defining reliable confidence intervals, with survival statistics calculated for repeated random sample orderings. This produces distributions of HRs expected by chance, to which observed true biological HRs can be compared. Only if a biomarker under investigation exceeds the bootstrapped thresholds is it considered significant. Complimenting *survivALL* is a companion web-based app – *survivAPP* (survivapp.shinyapps.io/survivapp/) – allowing non-programmatic, drag-and-drop *survivALL* use.

To illustrate the value of the *survivALL* package, we considered over-expression of human epidermal growth factor 2 (HER2/*ERBB2*), a well-established biomarker associated with poor prognosis in invasive breast cancer[7,8], in the frequently reanalysed NKI microarray dataset[9]. The plotALL() function enables simultaneous visualisation of HRs and p-values for all possible separations (n-1) of ordered *ERBB2* expression (Figure 1A). It establishes that 61 (19%) points-of-separation are significantly ($p<0.05$) associated with worse prognosis at higher *ERBB2* expression. Importantly, none of these 61 points are located at the median, and instead are distributed consecutively around the upper quartile/decile. This is consistent with epidemiological evidence demonstrating that, across the population, ~20-25% of breast cancers overexpress HER2[10]. Contrasted in Figure 1B are three Kaplan-Meier plots resulting from using the median ($p=0.9$), population-anticipated hypothetical upper quartile ($p = 0.04$) and *survivALL*'s data-driven optimal and most significant point-of-separation ($p=0.001$) for this dataset.

To consider the issue of dataset composition, we turned to the largest single breast cancer gene expression dataset, METABRIC[11], composed of two equally sized and composition-matched cohorts. These unusually highly similar datasets enable us to evaluate to what extent *survivALL* can derive and validate the optimal point-of-separation from one cohort to the other. For all 19,628 genes, *survivALL* was used to calculate the optimal point-of-separation in the discovery cohort (*n=980*) and the gene expression value here was then used to stratify the validation cohort (*n=991*) (Supplementary). Prediction accuracy was measured as the number of patients incorrectly classified with regard to the validation cohort's own calculated optimum (Supplementary). This procedure was similarly performed using a median approach for comparison (Figure 1C). *survivALL* significantly outperformed a median approach in predicting optimal stratification ($p=2.2\times10^{-16}$, Kolmogorov-Smirnov test), demonstrating *survivALL*'s applicability in robustly determining prognostic stratification, with no *a priori* knowledge of a gene's population level distribution.

In addition to determining the best point-of-separation within a single dataset, *survivALL* can robustly assess biomarkers across multiple datasets. Figure 1D summarises the prognostic performance of *AURKA*, a well-characterised proliferation marker prognostic in ER-positive breast cancer[12], across six ostensibly similar and identically normalised ER-positive and tamoxifen-treated invasive breast cancer Affymetrix datasets (GSE17705, GSE2990, GSE12093, GSE6532-a, GSE6532-p2, GSE9195). Due to random compositional differences, however, median ERBB2 dichotomisation failed to reveal a significant association with distant metastasis in 4/6 datasets, with no HR calculatable at all in GSE9195. In contrast, *survivALL* revealed multiple significant points-of-separation in all datasets (mean=45%, range=15-78%), confirming the expectation that highly proliferating tumours are associated with poor outcomes in ER-positive breast can-

cer and demonstrating an otherwise false-negative result by median approach. Genome-wide, cohort stratification by median gene expression levels identified an average of ~8% of genes to be prognostic in each dataset individually, while only 0.01% (3 genes) were reproducible across all 6 datasets. This contrasted with 11% being reproducible across all datasets using our exhaustive *survivALL* approach, highlighting the limitation of median methods in meta-analysis.

Researchers are increasingly using approaches such as KMplotter[13], to median stratify, or highlight the optimal point-of-separation in a set dataset. It is possible, however, that even though an optimal point-of-separation is found by these non-median methods, the extent of significant stratifications - a confidence measure of a true association and quantifiable by *survivALL* - remains unknown. Furthermore, in addition to optimal point-of-separation determination, *survivALL* can consider and visualise the effects of confounding factors in multivariate analysis, uncovering otherwise unknown factors when considering multiple datasets.

In this study, we have highlighted the issue of assessing quantitative biomarkers for survival analysis, offering *survivALL* and *survivAPP* as tools to evaluate and overcome these challenges. As an R package, *survivALL* allows greater resolution, transparency and flexibility compared to other best-of-split tools, applicable to any public or proprietary dataset and usable in larger-scale automated data operations. Furthermore, *survivALL* readily integrates with the popular visualisation tool *ggplot2*[14] for customisable output. Importantly, whilst the examples presented here relate to individual genes in breast cancer microarray datasets, *survivALL* is extensible to other diseases, proteomic data or any other quantitative measure, including multi-marker signatures. Moreover, whilst this paper has largely considered *di*chotomisation, *survivALL* also allows for additional sub-populations with varying outcomes to be visualised. Most importantly, both *survivALL* and *survivAPP* allow true biological effects and their relationship to outcome to be revealed, most notably in protecting against random compositional difference and false-negatives in a meta-analysis.

Acknowledgements

AHS is grateful for funding from Breast Cancer Now; TCF for funding from The Roslin Institute's BBSRC Institute Strategic Program Grant (BBS/E/D/20211552). DAP is recipient of a Cancer Research UK PhD studentship and ANJ recipient of an Edinburgh College of Medicine and Veterinary Medicine and Edinburgh Global Research Scholarship.

References

1. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B*. 1972;34(2):187-220. http://links.jstor.org/sici?sici=0035-9246%281972%2934%3A2%3C187%3ARMAL%3E2.0.CO%3B2-6. Accessed August 1, 2016.

2. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J Am Stat Assoc*. 1958;53(282):457-481. doi:10.1080/01621459.1958.10501452.

3. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752. doi:10.1038/35021093.

4. Camp RL, Dolled-Filhart M, Rimm DL. X-Tile: A New Bio-Informatics Tool for Biomarker Assessment and Outcome-Based Cut-Point Optimization. *Clin Cancer Res.* 2004;10(21):7252-7259. doi:10.1158/1078-0432.CCR-04-0713.

5. Motakis E, Ivshina A, Kuznetsov V. Data-driven approach to predict survival of cancer patients. *IEEE Eng Med Biol Mag.* 2009;28(4):58-66. doi:10.1109/MEMB.2009.932937.

6. Mizuno H, Kitada K, Nakai K, Sarai A. PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med Genomics.* 2009;2(1):18. doi:10.1186/1755-8794-2-18.

7. Kapp A V, Jeffrey SS, Langerød A, et al. Discovery and validation of breast cancer subtypes. *BMC Genomics.* 2006;7(1):231. doi:10.1186/1471-2164-7-231.

8. Mackay A, Weigelt B, Grigoriadis A, et al. Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *J Natl Cancer Inst.* 2011;103(8):662-673. doi:10.1093/jnci/djr071.

9. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 2002;415(6871):530-536. doi:10.1038/415530a.

10. Dean-Colomb W, Esteva FJ. Her2-positive breast cancer: Herceptin and beyond. *Eur J Cancer.* 2008;44(18):2806-2812. doi:10.1016/j.ejca.2008.09.013.

11. Curtis C, Shah SP, Chin S-F, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature.* 2012;486(7403):346-352. doi:10.1038/nature10983.

12. Ali HR, Dawson S-J, Blows FM, Provenzano E, Pharoah PD, Caldas C. Aurora kinase A outperforms Ki67 as a prognostic marker in ER-positive breast cancer. *Br J Cancer.* 2012;106(11):1798-1806. doi:10.1038/bjc.2012.167.

13. Szász AM, Lánczky A, Nagy Á, et al. Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget.* November 2014. doi:10.18632/oncotarget.10337.

14. Wickham H. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York; 2009. http://ggplot2.org.

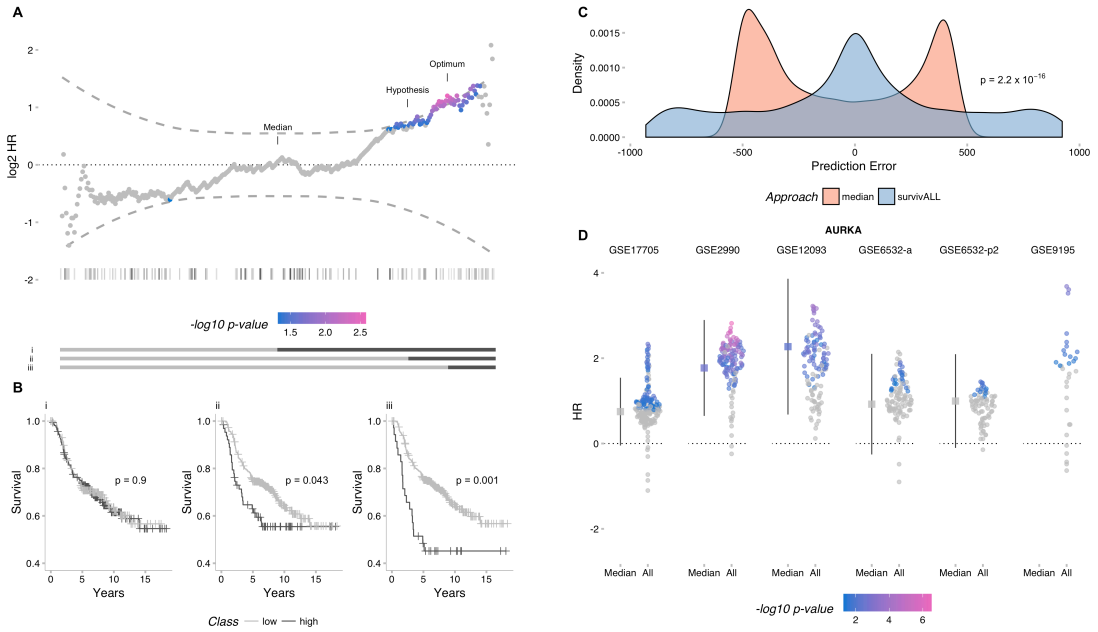

Figure 1: *Exhaustive survival analysis of continuous biomarker assessment.* **A)** The plotALL function allows hazard ratios (HRs; y-axis) and p-values (p<0.05 colour scale; p 0.05 grey) for all possible quantitative biomarker cut-points to be examined and compared against confidence intervals determined by a non-parametric bootstrap (long dashes). Patients are ordered by increasing expression of the quantitative marker – in this example, *ERBB2* expression. Events (distant metastasis-free survival) are indicated as vertical lines below the plot, with darker colours indicating chronologically earlier events (range=0-18 years). Median (p=0.9), hypothesis-driven (p=0.04) and optimal points-of-separation (p=0.001) are indicated and the associated patient stratification is shown (i, ii, & iii). **B)** Kaplan-Meier plots for *ERBB2* stratified using median (i, n = 159 vs. 160), hypothesis-driven (ii, n = 255 vs. 64) or survivALL (iii, n = 284 vs. 35) approaches comparing patients with expression lower (grey) or higher (black) than the points-of-separation determined in A. **C)** Comparison of *survivALL* and median approaches in predicting optimal cohort separation. The *survivALL*- (blue) and median-approach (red) distributions illustrate prediction error as the difference between *survivALL*- or median-determined point-of-separation between the discovery and validation datasets for all genes, where a difference of 0 indicates the same point-of-separation is identified within each cohort. The *survivALL* approach significantly (p=2.2x10^{-16}) outperforms a median approach in terms of predictive accuracy. **D)** Summary output to compare *AURKA*, a marker of proliferation across six ER-positive tamoxifen-treated breast cancer datasets (n = 196, 139, 136, 109, 87, 77, respectively, left to right). For each dataset, median approach-derived HRs (with 95% confidence intervals) and significance are displayed to the left, whilst the distribution of *survivALL*'s corresponding statistics are shown to the right. Using the *survivALL* approach, each dataset demonstrates a significant (p<0.05 colour scale; p 0.05 grey) association with prognosis, but using median separation, only two of six ER-positive datasets are significant. It is noteworthy that a HR cannot be calculated using median separation for dataset GSE9195 as all events are in patients with expression higher than median *AURKA* levels.