Genomics 108 (2016) 78–83

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Comparing continuous and discrete analyses of breast cancer survival information

Vinayak Bhandari [a,b], Paul C. Boutros [a,b,c,*]

[a] Informatics and Bio-computing Program, Ontario Institute for Cancer Research, Toronto, Canada
[b] Department of Medical Biophysics, University of Toronto, Toronto, Canada
[c] Department of Pharmacology, University of Toronto, Toronto, Canada

ARTICLE INFO

Article history:
Received 29 February 2016
Received in revised form 25 May 2016
Accepted 11 June 2016
Available online 14 June 2016

Keywords:
Cox proportional hazards
Survival analysis
Continuous data
Discrete data
Machine learning
Random forest
Biomarkers

ABSTRACT

Treatment of cancer is becoming increasingly personalized and biomarkers continue to be developed to refine treatment decisions. Tumour mRNA abundance data is commonly used to develop such biomarkers, often to predict patient survival. However, survival analyses present unique challenges and it is unknown whether analysing mRNA abundance information in a discrete or continuous manner yields different results. To address this, we analysed 1988 primary breast tumour transcriptomes. When compared univariately, approximately 60% of all genes showed differences between the discrete and continuous Cox proportional hazards models with q-value differences spanning four orders of magnitude for some genes. Further, hybrid models using both continuous and discrete data used to classify poor prognosis via random forest outperformed models using a single type of information. Thus some genes appear to continuously contribute to poor prognosis while others display threshold effects, and incorporating this into biomarker development is a key unexplored avenue.

1. Background

Survival analyses associate patient outcome with one or more biologically descriptive variables. Typical goals of such studies are to evaluate the impact of a treatment or intervention on patient survival over time, relative to a control group. Alternatively, they can be used to generate models that can predict for any individual what their baseline risk of a later adverse event will be. These analyses are often much more statistically complex than simple linear models because of cases in which patient information in incomplete (*i.e.*, the data is censored). Censoring can occur for many reasons, including if the event of interest did not occur within the period of study (*e.g.* the patient is still alive at the end of the observation interval) or if a subject withdraws from the study prior to completion. In such a case, the information is right-censored and the minimum survival time is known. As a result several statistical methods have been created to handle right-censored information, including the discrete Cox proportional hazards (PH) model (also known as the log-rank test) [1] and the continuous Cox PH model [2].

Both the continuous and discrete versions of the univariate Cox PH model are routinely used to analyse data generated from survival studies [3]. Their null hypothesis is that the probability of an event occurring is not different between the populations being compared [1]. Both models make the same assumptions: that censoring is not related to prognosis, that the probability of survival is not significantly different for individuals recruited early and late in the study, and that the events happened at the specified times. Recent studies relating to gastric cancer [4], ovarian cancer [5], and lung cancer [6] provide examples of analyses utilizing the discrete Cox PH model. Other work examining gastric cancer [7], ovarian cancer [8], and breast cancer [9] have used the continuous Cox PH model.

While these models are widely used, continuous and dichotomized analyses of the same information may capture different underlying biological phenomena. For example, in some cases it is clear that a biological variable can be accurately discretized, *e.g.*, copy-number aberrations. Further, categorising quantitative information into two groups can also assist with removing batch effects and standardizing datasets. In contrast, biological processes that are sensitive to absolute values (*e.g.*, hormone levels [10]) may be better represented by analysing the data in a continuous fashion.

Fundamentally, then, these two models represent different expectations about the biology and underlying mechanism of action of the gene being studied. Consider mRNA abundance data, which is widely used to generate prognostic models for personalizing patient therapy. The continuous model assumes that each additional mRNA molecule in a cell incrementally increases or decreases the risk of an event, while the discrete model suggests that an effect is not observed until some key threshold of mRNA abundance is reached. Surprisingly, then, to

* Corresponding author at: MaRS Centre, 661 University Avenue, Suite 510, Toronto, Ontario M5G 0A3, Canada.
E-mail address: Paul.Boutros@oicr.on.ca (P.C. Boutros).

http://dx.doi.org/10.1016/j.ygeno.2016.06.002

our knowledge all biomarker-development studies using this type of data employ either the discrete ($Cox_{discrete}$) or the continuous ($Cox_{continuous}$) models for all genes. We are unaware of any systematic efforts to determine which approach better represents individual genes, nor to assess if biomarkers comprised of a mixture of continuous and discrete features will be more accurate.

To address this gap in the field, we examined the mRNA abundance information for 1988 breast cancer patients with primary breast tumours from the Metabric study [11]. In particular, we provide biologically relevant examples for which the $Cox_{discrete}$ and $Cox_{continuous}$ models produce contrasting results (*i.e.*, the q-value from one model is high while the q-value from the other model is low). In addition, we provide insight into the performance of each model independently, or in combination, in the context of predicting patient survival via a random forest classification analysis.

2. Results

2.1. Experimental design

This study used the Metabric breast cancer dataset which contains survival and mRNA abundance information from over 19,000 genes for 1988 primary breast tumours. Our study involved two major parts (Fig. 1). First, we applied the $Cox_{discrete}$ and $Cox_{continuous}$ models separately to the Metabric training dataset (996 subjects). False-discovery rate adjusted p-values (referred to as q-values henceforth), were compared to assess and quantify differences between the models. This dataset was well powered to assess such differences (Fig. S1). Next, we evaluated the impact of these models on multi-gene biomarkers by considering the top 2000 genes implicated by each model (*i.e.*, low q-values), resulting in a pool of 2759 unique genes. We assessed the null distribution of biomarker space by randomly selecting genes from this pool for modelling via a random forest classifier [12]. Each random forest model was independently validated using a validation dataset containing data from 992 subjects.

2.2. Overview of differences between models

We assessed the univariate, gene-wise differences between the q-values generated from the two models (Fig. 2). Amongst examined genes, two broad cases exist: those genes for which the two models agree and those for which they disagree. The first case includes genes that generated a high q-value via both the $Cox_{discrete}$ and $Cox_{continuous}$ models as well as genes for which the two models generated a low q-value. The second case – genes for which the models generated very different q-values – was surprisingly common. Overall, q-values from the $Cox_{discrete}$ and $Cox_{continuous}$ models were not highly correlated (Spearman's $\rho = 0.68$). Only 62% (12,318/19,877) of q-value differences were less than or equal to 0.2, demonstrating that substantial differences in q-values exist for a large subset of the dataset. Some of the differences for individual genes were very large, such as NUDT19 with $q_{continuous} = 2.5 \times 10^{-6}$ and $q_{discrete} = 0.012$ while HK3 exhibited $q_{continuous} = 0.25$ and $q_{discrete} = 0.0025$. Overall, the continuous model tended to yield smaller q-values, as might be expected from its greater statistical power.

2.3. Functional consequences of model differences

Genes demonstrating the largest discrepancies between the $Cox_{discrete}$ and $Cox_{continuous}$ models were identified for further analysis. To probe the functional roles of these genes, a pathway analysis was completed using the GoMiner software [13] (Table S1). Pathway analysis revealed that genes with the largest q-value differences between models ($n = 30$) were particularly enriched for cellular component: chromosome (GO:0005694).

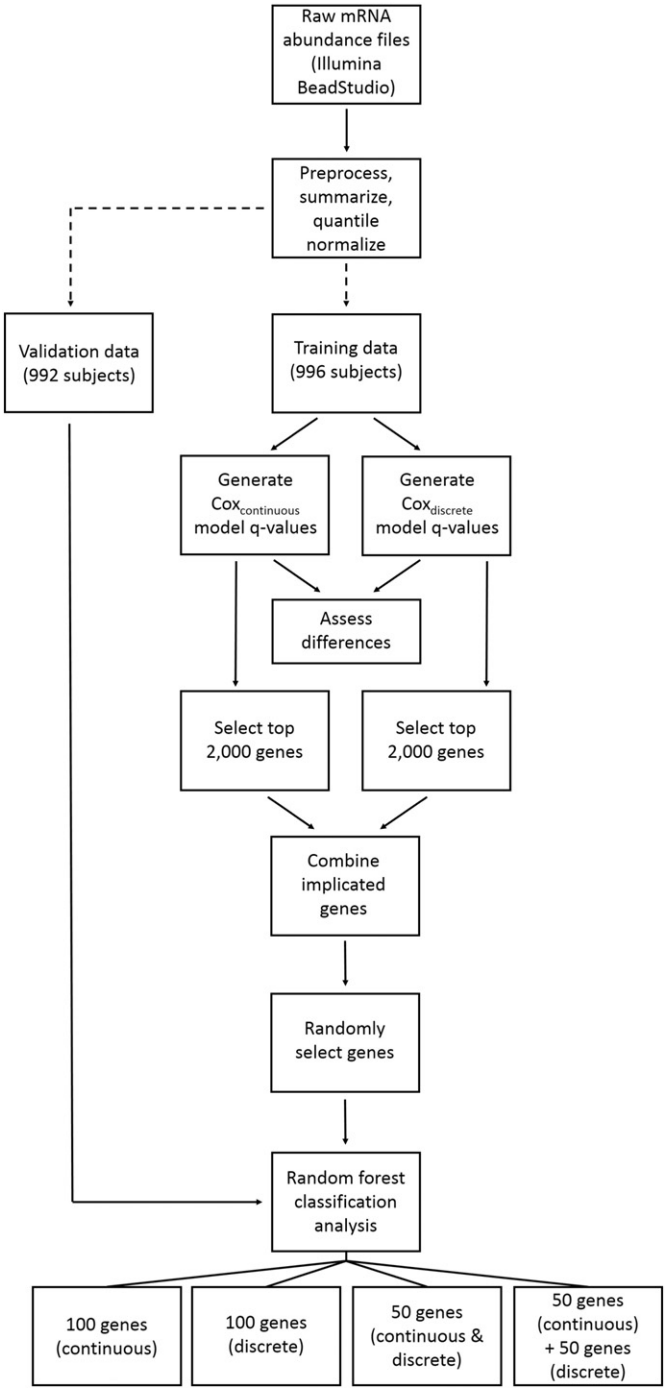

Fig. 1. Experimental design. Raw mRNA abundance files from the Metabric breast cancer dataset were preprocessed, summarized and quantile-normalized. 1988 breast cancer patients with primary breast tumours were divided into training and validation groups. The $Cox_{discrete}$ and $Cox_{continuous}$ models were applied to subjects in the training group to generate q-values and to assess differences between models. The 2000 most significant genes identified by each model were selected for further analyses. From this pool, genes were randomly selected to build random forest classification models and predict survival. The validation group was used to independently validate each of the 40 million permutations for each of four models used.

Of the genes showing the largest differences between the two models, the $Cox_{continuous}$ q-values were typically lower than those from the $Cox_{discrete}$ model (Table 1). This is of particular interest as a survival analysis using the $Cox_{discrete}$ model may not identify and investigate these genes, potentially missing biologically relevant information, despite its wide use in biomarker development studies. The mRNA abundance of those genes with large q-value differences between