Published in final edited form as:

Stat Med. 2013 April 30; 32(9): 1524–1535. doi:10.1002/sim.5641.

Pathway-Index Models for Construction of Patient-Specific Risk Profiles

Kevin H. Eng^a, Sijian Wang^{a,b}, William H. Bradley^c, Janet S. Rader^c, and Christina Kendziora^{a,*}

^aDepartment of Biostatistics and Medical Informatics, University of Wisconsin-Madison

^bDepartment of Statistics, University of Wisconsin-Madison

^cDepartment of Obstetrics and Gynecology, Medical College of Wisconsin

Abstract

Statistical methods for variable selection, prediction, and classification have proven extremely useful in moving personalized genomics medicine forward, in particular leading to a number of genomic based assays now in clinical use for predicting cancer recurrence. Although invaluable in individual cases, the information provided by these assays is limited. Most often, a patient is classified into one of very few groups (e.g. recur or not), limiting the potential for truly personalized treatment. Furthermore, although these assays provide information on which individuals are at most risk (e.g. those for which recurrence is predicted), they provide no information on the aberrant biological pathways that give rise to the increased risk. We have developed an approach to address these limitations. The approach models a time-to-event outcome as a function of known biological pathways, identifies important genomic aberrations, and provides pathway-based patient-specific assessments of risk. As we demonstrate in a study of ovarian cancer from The Cancer Genome Atlas (TCGA) project, the patient-specific risk profiles are powerful and efficient characterizations useful in addressing a number of questions related to identifying informative patient subtypes and predicting survival.

Keywords

gene expression; molecular profiling; survival analysis; genomic analysis; ovarian cancer

1. Introduction

A main challenge of personalized medicine is to identify the individuating features of a patient and to specify the way in which they can be used to augment individual prognosis, follow-up and treatment. Molecular profiling holds major promise to this end. Indeed, National Institutes of Health Director Francis Collins suggests that the development of targeted therapeutics based on a molecular understanding of disease may be the “most profound consequence of the genome revolution” [1]. Unfortunately, the information available in high-throughput genomic screens has yet to be fully utilized in routine patient care.

The development of molecular based signatures to guide patient care is certainly notable with a number of high-throughput expression based assays now in clinical use (or in the

final stages of development) for predicting recurrence of breast [2, 3], colon, and prostate cancer, as well as transplant rejection [4, 5]. Although clinically useful, the information provided by these signatures is not sufficient to guide truly personalized treatment. Doing so requires approaches that provide for more refined patient classification (e.g. beyond prediction into two or three risk groups) as well as information about the specific genomic aberrations underlying each patient's disease.

We propose a statistical approach to address these limitations in studies of cancer. The approach allows for the identification of patient specific aberrant pathways which characterize the biology underlying an individual's disease. The focus on pathways is motivated by Jones et al. [6] and a number of subsequent studies of cancer [7, 8] demonstrating unexpectedly high tumor heterogeneity at the gene level, but considerable consensus at the pathway level. For example, Jones et al. [6] detail an analysis of 24 pancreatic tumors and demonstrate that the *specific genes* altered in each tumor are largely different, but that the *pathways* containing the altered genes are largely the same. They argue that "the pathway perspective helps bring order and rudimentary understanding to a very complex disease" not just for pancreatic cancers, but all solid tumors. If this is the case, statistical approaches that focus exclusively on individual genes will likely be critically underpowered.

This has been recognized by a number of investigators who have developed sophisticated statistical methods for variable selection with grouped predictors, or pathways [9, 10, 11, 12, 13, 14]. Although useful, a disadvantage of these methods is that variables are selected in an "all-in-all-out" fashion, meaning that when one variable in a group is selected, all other variables in the same group are also selected. As a result, these methods do not provide information on the relative importance of variables within an identified group. To address this shortcoming, we [15] and others [16, 17] have developed methods to not only identify groups, but also variables within groups. Unfortunately, the computational complexity is limiting in practice. In particular, these approaches result in objective functions that are non-convex, which leads to numerical instability when the number of predictors is large. The problem is exacerbated for time-to-event phenotypes observed with censoring [15].

As detailed in the next section, the approach proposed here sacrifices statistical sophistication for practical advantage. Groups of genes, referred to hereinafter as pathways, are defined *a priori*, and a survival related phenotype is modeled as a function of genes within pathways. Genes within each pathway are selected via lasso in a first step and identified as conferring increased or decreased risk. A pathway summary, or "index," measuring overall risk across the pathway is then constructed from the selected genes. As we show in Section 3, the patient-specific index scores (referred to as patient-specific risk profiles) are powerful and efficient characterizations useful in addressing a number of questions related to identifying patient subtypes, predicting survival and characterizing patient-specific risk.

An evaluation of operating characteristics via a simulation study is provided, with advantages demonstrated in an analysis of data from The Cancer Genome Atlas (TCGA) Project. Briefly, TCGA is a comprehensive and coordinated effort designed to improve the understanding and treatment of cancer through an increased understanding of the molecular basis of the disease gained by analysis and integration of multiple large-scale measurements of the genome and phenome collected on relatively large groups of patients (cancergenome.nih.gov). TCGA's first studies emphasized three disproportionately deadly cancers - lung, brain, and ovary. Analyzing the data from the ovarian cancer project, Section 3 demonstrates that the proposed method provides for improved risk prediction over competing approaches and, unlike other approaches, provides functionally relevant

information that may significantly impact ovarian cancer research and, ultimately, patient treatment.

2. Pathway Index Model

Consider survival information (Y_i, δ_i, x_i) , for $i = 1, \dots, n$ patients, where Y_i is a complete survival time when $\delta_i = 1$ and is a right censored time when $\delta_i = 0$. An individual's covariate vector $x_i = (x_{i1}, \dots, x_{ip})$ contains gene-level measurements where the number of genes, p , greatly exceeds n . Database or expert information about specific biological processes and hypotheses is represented as m sets of genes denoted by $\mathcal{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_m\}$, where each set is defined as:

$$\mathcal{P}_k = \{j: j\text{th gene is a member of the } k\text{th pathway}\}.$$

These sets of genes, called pathways, are restricted to those with n or fewer genes to reduce the estimation problem to manageable pieces and to increase the specificity and interpretability of results. No restriction is placed on the amount of pathway overlap, so some genes may be members of multiple pathways.

In the k th pathway, the hazard for the i th patient, $h_k(t, x_i)$, is described by a Cox model:

$$\log h_k(t, x_i) = \log h_{0k}(t) + \sum_{j \in \mathcal{P}_k} x_{ij} \beta_j^k,$$

where $h_{0k}(t)$ is the baseline hazard [18]. Genes are selected within each pathway via lasso penalized Cox regression which has the objective below, where $\eta_k(\beta, x_i) = \sum_{j \in \mathcal{P}_k} x_{ij} \beta_j$ for covariate vector β .

$$\max_{\beta} \sum_{i=1}^n \delta_i \left\{ \eta_k(\beta, x_i) - \log \left[\sum_{j: Y_j \geq Y_i} e^{\eta_k(\beta, x_j)} \right] \right\} \\ \text{s.t. } \sum_j |\beta_j| \leq t, t \geq 0.$$

Note that each gene is centered and scaled to have unit standard deviation per standard practice [19]; ten-fold cross validation is used to estimate tuning parameters.

The selection step here, whereby genes within pathway are selected via Cox proportional hazards lasso regression, has been proposed by Ma *et al.* [20]. Following this step, Ma *et al.* combine selected genes from each pathway into a single group and perform a second lasso across all of the selected genes. Together, this procedure amounts to a single lasso with pre-screening step. Our approach differs in that its second step retains the pathway structure and does not require a second regression model: only the signs, already estimated, are necessary to construct the pathway risk. Maintaining the pathway information allows for a direct connection between pathway and patient-specific risk, while considering the sign of selected genes helps stabilize the index as the sign consistency of lasso requires weaker conditions than the estimation consistency [21].

2.1. Patient-specific risk profiles (PSRPs)

Given selected genes, an index of pathway risk is constructed for each patient by comparing the strength of patient-specific susceptibility and resistance signals within a pathway. Specifically, selected genes are first classified by the sign of their estimated effect:

$$\begin{aligned}\mathcal{P}_k^S &= \{j: \text{expression of the } j\text{th gene negatively affects survival in the } k\text{th pathway}\} \\ \mathcal{P}_k^R &= \{j: \text{expression of the } j\text{th gene positively affects survival in the } k\text{th pathway}\}\end{aligned}$$

Equivalently, susceptibility genes ($\mathcal{P}_k^S$) refer to those genes for which increased expression is associated with higher hazard; resistance genes ($\mathcal{P}_k^R$) refer to those genes for which increased expression is associated with reduced hazard. A patient-specific pathway index, Z_{ik} , is then constructed for the i th patient in the k th pathway:

$$Z_{ik} = \bar{X}_{ik}^S - \bar{X}_{ik}^R$$

where

$$\bar{X}_{ik}^S = \begin{cases} |\mathcal{P}_k^S|^{-1} \sum_{j \in \mathcal{P}_k^S} x_{ij} & |\mathcal{P}_k^S| > 0 \\ 0 & |\mathcal{P}_k^S| = 0 \end{cases},$$

$|\mathcal{P}|$ is the cardinality of set $\mathcal{P}$, and $\bar{X}_{ik}^R$ is similarly defined for $\mathcal{P}_k^R$.

The vector of indices across all pathways for patient i forms the patient-specific risk profile (PSRP) for that patient: $\mathbf{Z}_i = (Z_{1i}, Z_{2i}, \dots, Z_{m'i})'$, where m' denotes the number of pathways for which at least one gene is selected. As we demonstrate in Section 3, the PSRPs identify biologically and survival-associated patient subtypes useful for characterizing disease. In addition, as with most of the methods previously discussed, the PSRPs can also be used for prediction as described in the next section.

2.2. Prognostic models from PSRPs

Given a new patient with expression values $\mathbf{X}^*$, their PSRP may be constructed by evaluating the pathway scores $Z_{ik}^* = \bar{X}_{ik}^{*S} - \bar{X}_{ik}^{*R}$ across each of the m' pathways. This requires only the sets of genes $\mathcal{P}_k^S$ and $\mathcal{P}_k^R$.

If a scalar measure of patient-specific risk is of primary interest, the indices in the PSRP may be combined across pathways. We define the index-count method to be a simple count of the number of positive pathway indices ($\{Z_{ik} > 0\}$). An advantage of index-count is that it naturally leads to categories of risk without having to define (often arbitrary) cut points as would be required with a continuous predictor. Furthermore, the lack of complexity of index-count attenuates the potential for overfitting and may be useful particularly when the number of selected pathways is too small to justify a second selection step. At the same time, the simplicity of index-count may be limiting in some situations.

When relative contributions of pathways are of interest, the Z_{ik} may be used jointly as features in a regression modeling framework (e.g. under parametric, accelerated failure time, proportional hazards, or additive hazards assumptions). We define the index-lasso model as

follows. Consider the pathway indices as covariates in a proportional hazards Cox model that minimizes a partial likelihood with linear predictor

$$\eta_i = \sum_{k=1}^{m'} Z_{ik} \alpha_k, \quad (1)$$

and computes the predicted relative risk $\eta_i^* = \sum_{k=1}^{m'} Z_{ik} \hat{\alpha}_k$. This second lasso step, taken because of the superior prediction properties of the lasso [22], uses 10-fold cross-validation to select its tuning parameter. Section 4 compares the selection and prediction performance of these and other lasso-type methods. While we have chosen to highlight the lasso for its superior prediction properties [22], the investigator may employ any model building method whose properties match the problem at hand: best subsets or lasso (for selection), regression trees (interactions), or ridge regression (regularization when there are too many pathways).

3. Ovarian cancer signaling pathways study

In this section, we demonstrate how the pathway index model may be used to construct PSRPs useful for predicting survival and stratifying risk. We study data from the TCGA ovarian cancer project [23, cancergenome.nih.gov], which consists of 510 serous cystadenocarcinoma patients for whom survival, clinical and gene expression data are available in the public TCGA data portal. Survival times from diagnosis to death for the 503 patients who received adjuvant chemotherapy (combined platinum and taxol) following surgery are reported here (for further detail on inclusion criteria, see the Supplement). Of these, 186/503 are censored before 5 years (37%).

We will first focus on building PSRPs for the training cohort ($n = 234$), defined in the TCGA's recent paper [23]. Then, we will validate the predictive performance in their test set and two independent studies (described below).

3.1. Building pathway index models

As described in Section 2, the pathway index model relies on pathways specified *a priori*. Shown in Table 1 are the 15 KEGG pathways [24] considered in subsequent analysis. These 15 correspond to the 12 core cancer pathways identified by Jones et al. [6]. Note that one of the 12 core pathways is DNA repair, for which the KEGG database lists four distinct gene sets (base excision repair, mismatch repair, non-homologous end-joining and nucleotide excision repair). Each mechanism is included here separately for a total of 15 pathways. The R annotation package `hthgu133a.db` was used to match probes to pathways; tuning parameters for each pathway were selected separately by 10-fold cross-validation. The model identified 104 (88 unique) genes across 9 pathways; the 6 pathways with no selected genes were excluded from further analysis. From these selected genes, the pathway indices were computed as previously described. The net effect is a vector of 9 continuous indices comprising a PSRP for each patient; the prognostic value of these PSRPs is considered below.

3.2. Index-Count Model Prognostic Results

To evaluate the prognostic capability of the PSRPs, we summarize each using the index-count approach described in Section 2.2, which simply counts the number of pathways at risk (the number of pathways for which $\{Z_k > 0\}$). The comparative value of other methods is considered in Section 4.3.

Figure 2 shows results for three patient groups derived by the index-count model: low risk (0–3 positive pathways), medium risk (4–6), and high risk (7–9). The left panel shows the restricted mean life at 60 months [25], which is the number of months a patient is expected to experience out of 60 months possible. The high risk group has an expectation of 25.6 months (standard error 0.6), the medium risk has 36.9 months (1.1), and the low risk group is expected to see 53.2 months (2.1) or nearly 90% of their next 60 months. Also shown in the left panel of Figure 1 are the predictions when patients are not classified into 3 groups, showing the restricted mean life for zero to all nine positive pathways.

Cox proportional hazards (PH) regression was used to estimate risk within each group, indicating that hazard increases by a factor of 5.13 (0.26) for medium risk over low risk patients and 11.16 (0.28) for high risk patients over low risk patients. Differences among risk groups are evident in the right panel of Figure 1, which shows Kaplan-Meier curves for each group. The differences between groups is highly significant: low vs. medium (log rank $p=2.33\text{e-}04$) and medium vs. high (log-rank $p=1.94\text{e-}12$). We note that for each analysis, the PH assumptions were satisfied (p-values not shown). In addition to its prognostic utility, the index-count model also provides for robust survival prediction, as shown below in Section 4.

3.3. Validation in independent studies

Recall that besides its training set, the TCGA data has a withheld test set ($n = 269$) which we use subsequently to evaluate the index-count model. We also consider two comparable, independent datasets available at the Gene Expression Omnibus: Tothill et al. [26] (GEO: GSE9899) with $n = 240$ arrays after dropping early stage and low malignant potential cancers; Bild et al. [27] (GEO: GSE3149) with $n = 134$ samples after averaging 8 pairs of duplicated arrays and dropping 4 identifiers with missing survival information (for further details, see the Supplement).

We compute the pathway indices for these models using the gene sets learned on the TCGA training set only. The change in hazard for each positive pathway is given in Table 2 as well as a significance test and verification that the PH assumption is met. As in Section 3.2, the count of positive pathways leads to low, medium and high risk groups whose restricted mean life estimates are also provided. Figure 2 shows Kaplan-Meier curves for each of the predicted risk groups. The curves demonstrate clear utility of the pathway-index model since the prognoses provided by index-count for the Tothill and Bild studies are not only consistent with the TCGA sets, but show improvement over the results originally published by those groups (see Figures 5B and 3C from Tothill and Bild, respectively). In the Tothill study, the authors cluster expression data into 6 groups and show survival within group. As a result, the separation observed among Kaplan-Meier plots for the 6 patient groups they identified (their Figure 5B) are ideal as they are determined by the full dataset, effectively a training dataset; Bild was similar. Validation in independent datasets was not considered in either study. Here, the pathway index model parameters are estimated in the TCGA training data, yet they perform well in Tothill and Bild (as well as in the TCGA test set). Furthermore, unlike other approaches, the pathway index model provides information connecting risk with function.

4. A comparative evaluation

The pathway index model provides a risk profile for each patient; in particular, it gives a measure of risk (present or absent) for each pathway in a biologically-motivated collection. As demonstrated in the previous section, these PSRPs are useful for identifying patient subtypes and predicting survival. When survival prediction is the main goal, there are a number of alternate methods that one may use; many are reviewed in the introduction. For

comparison with these approaches, we performed a small set of simulation studies. The simulations are not designed to evaluate each aspect of high-dimensional grouped variable selection in the context of survival analysis, but rather to provide some preliminary information on operating characteristics associated with risk prediction via the pathway index model compared with other approaches.

Each simulated dataset consists of 984 standard normal gene expression values along with survival times for each of $n = 1000$ patients. The number of genes was chosen to match the empirical data considered in Section 3; and the genes are organized into the same structure as the 15 KEGG pathways that are studied there (see Table 1). Survival times follow an exponential, proportional hazards model, $T_i = \frac{\varepsilon_i}{\exp(\Psi(X_i))}$ where ε_i are independent standard exponential random variates and $\Psi(X)$ is a scenario dependent function translating a patient's expression profile, X_i , to a log relative hazard. Censoring times were generated from an exponential distribution with rate parameter chosen to target a censoring rate of 0.35. Five scenarios are considered for generating $\Psi(X)$, with parameters determined using the TCGA training dataset, described in detail in Section 3.

- **SI: Lasso based** $\Psi(X_i) = X_i' \beta_L$ where β_L are fixed at the values estimated by a lasso-penalized Cox regression of all 984 genes in the TCGA training dataset. There were $p = 30$ genes selected.
- **SII: Average based** $\Psi(X_i) = W_i' \alpha_A$ where W_i is the average of all genes in a given pathway (as opposed to selected genes as in scenarios SIV and SV) and $\alpha_A = \alpha_I / 10$, reflecting the fact that approximately 10% of genes are selected across the KEGG pathways overall (see Table 1).

SIII: Average-Count based $\Psi(X_i) = 0.2 \times \# \{W_i > 0\}$, which specifies that the median patient has approximately double the risk of a patient for which $\# \{W_i > 0\} = 0$, as observed in the TCGA cohort.
- **SIV: Index based** $\Psi(X_i) = Z_i' \alpha_I$ where each Z_i is a pathway-index defined as in equation 1 where the susceptibility and resistance genes are identified in the TCGA training dataset; α_I is also estimated from the TCGA training data.

SV: Index-Count based $\Psi(X_i) = 0.2 \times \# \{Z_i > 0\}$, which allows all increased-risk pathways to be equally important.

For each scenario, 500 simulated datasets are considered.

4.1. Methods for comparison

Six methods were applied to each simulated dataset. Most of the methods can be classified generally into two-step procedures which select (or summarize) genes in a first step and then further select genes (or some summary) in a second step. Specifically, we consider lasso regression on the full set of genes. We also consider SIS-lasso, which uses a univariate screen before fitting a final lasso [28]. As discussed in Section 2, the method of Ma et al. [20] first screens important genes within pathway and then fits a second lasso to the selected genes without accounting for pathway structure in the second step. We refer to this as lasso-lasso. What we refer to as average-lasso averages expression within pathway and performs lasso regression on the averages [14, 13]. Two additional methods highlight different ways to use information from the pathway-index model. Index-count refers to a simple count of the derived indices $\{Z_k > 0\}$; index-lasso refers to lasso regression using the indices Z_k as covariates (see Section 2.2).

4.2. Simulation Study Results

Estimates of risk are compared to the true risks by calculating Spearman's correlation coefficients, Manhattan distance on the ranks, and misclassification error. To calculate misclassification error, we consider the top and bottom ranks (delineated by the median rank), and define a misclassification if an estimated top rank is actually in the bottom 10% of the true ranks (or a bottom rank is in the top 10% of the true ranks). We report misclassification rates for 10% tails; results (not shown) were similar for 5%, 15% and 20%. The proportion of cases where each method fails to select any features is also reported. For these cases, correlations are set to zero, and the Manhattan distance is computed setting all ranks to be tied at zero. Table 3 summarizes the results. When risk is generated by a small number of genes not contained within particular pathways (simulation SI), lasso and SIS-lasso perform best. Index-lasso shows performance comparable to SIS-lasso, suggesting that the cost incurred by pathway level estimation is minimal. For scenarios SII and SIII, risk is determined by effects from all genes and so approaches that select genes in a first step are not optimal. Not surprisingly, average-lasso, which does no gene-level selection, outperforms all other approaches in these two cases. Simulation scenarios SIV and SV generate risk through individual genes within pre-defined pathways; and the index-based methods perform best, although most selection methods show comparable performance. Taken together, these results are not surprising. The methods that best match a simulation scenario perform best for that scenario.

Some results are not as obvious. Firstly, lasso-lasso is never a leading method; and, as shown in the first rows of the table, it fails to select features in over 10% of the simulations for every simulation scenario considered. The other non-index based methods also fail fairly frequently, with the average-lasso approach failing over 40% of the time on simulations that are not average-based. This is likely due to the relatively small effect sizes that were used in the simulations. These sizes are realistic as they were chosen to match those observed in TCGA data. When the individual gene effect is small, but the genes are considered collectively within pathway, accounting for the pathway information may be advantageous. This proves to be the case here, as the pathway index based models identify features in every simulation. Furthermore, the pathway based models show prediction performance comparable to other methods for most of the simulation scenarios considered (Table 3).

4.3. Ovarian cancer application

Results from a similar evaluation using the case study data are consistent with those from the simulation. Because all methods lead to a continuous (or ordinal) value, to compare the methods' prognostic value, Figure 3 shows time-dependent receiver operating curves (ROC) [29] at 3 years, the median time for patients with at least 5 years of followup in the TCGA training set. While all methods (except average-lasso) perform similarly on training data, only the index-count method performs consistently well on all datasets. Other time points had equivalent results.

Overall, these results show improved predictive performance for methods that accommodate pathway structure. This is somewhat expected, as averaging (within a pathway) reduces variance. Thus, to determine the advantage due specifically to the defined (and biologically relevant) pathway structure, as opposed to pathways defined as arbitrary groups of genes, we considered results after permuting genes across pathways. Specifically, we permuted the 984 genes across pathways, preserving the strength of individual genes, their relationship with survival, and the size and number of pathways, while at the same time removing any relationship between gene and pathway. The pathway-index model was fit to 1000 datasets generated in this way and, in all but one, performance as assessed by the log-rank p-value was inferior to that observed in the case study data.

5. Discussion

The last decade of genome research has had a profound impact on scientific progress that has yet to be matched by a similar impact on personalized genomic medicine. Improvements in computational methods are required to move the field forward. Specific challenges include accommodating heterogeneity, predicting response, and providing for personalized assessment of risk in a way that is, ideally, clinically actionable. The pathway-index model proposed here addresses some of these challenges.

The focus on pathways was motivated by Jones et al. [6] and a number of subsequent studies which demonstrate that cancer patients show very little overlap in aberrant genes, but much consensus in aberrant pathways. The pathway-index model accommodates this heterogeneity by focusing on pathways *a priori*. As we have demonstrated, the PSRPs derived from the model may be used in a variety of ways. The main utility discussed here is in predicting survival and stratifying risk.

When survival prediction is of interest, the m' -dimensional PSRP may be summarized to provide a scalar predictor. We considered two types of summaries: index-count which simply counts the number of risky pathways and index-lasso which considers the PSRP pathway indices as covariates in a lasso regression. As we detailed, there are a number of other approaches for survival analysis with grouped predictors. The two most similar to index-count and index-lasso are referred to here as lasso-lasso and average-lasso, since lasso-lasso selects genes via lasso-regression within pre-defined pathways [20] and then considers the selected genes as covariates in a second lasso-regression; average-lasso evaluates average expression across each pathway and uses the averages in a subsequent lasso-regression [14, 13]. The key differences are that we first identify important genes within pathway (unlike average-lasso) and then keep the pathway structure in a second selection step (unlike lasso-lasso). Simulations illustrate how some of these differences affect predictive performance.

In particular, in the simulation study we evaluated the ability of the PSRPs and other approaches to predict into two groups (good vs. poor survival), as is commonly done when assessing predictive performance. Results showed poor performance of non-pathway based approaches. Not surprisingly, simulations showed that the quality of predictions obtained via the PSRPs was comparable to the average-lasso approach, the method most similar to index-lasso. The slight (but not statistically significant) advantages observed in index-lasso illustrate the modest gain obtained by selecting genes within pathway prior to construction of the index (or covariate). The simulation study results also suggest that not accounting for pathway overlap directly does not adversely affect operating characteristics to a great extent. However, for applications where there is much more pathway overlap than typically observed in the KEGG pathways considered here, PSRP components may be highly correlated, and this may affect downstream analyses. The extent of correlation should be assessed on a case-by-case basis. In summary, the simulation suggests that the pathway-index model provides an approach for survival prediction that is comparable to competing approaches, and slightly better when one considers the inability of competing approaches to identify any informative covariates in many cases (see the top of Table 3).

As demonstrated in Section 3.3, the PSRPs are also useful for stratifying patient groups. Certainly there are other approaches that could be used for this purpose, many of them highlighted in studies of ovarian cancer similar to the TCGA study considered here [30, 27, 26, 31, 32, 33, 34]. However, unlike these approaches, the PSRPs validate in independent datasets. The simplicity of the PSRPs allows them to be easily transferred to and evaluated in other settings, and the consideration of average expression underlying each PSRP index

provides for a signature that is robust to transient variation. Furthermore, as the PSRP indices are tied to biological function, the aberrant mechanism(s) underlying any given risk group may be better understood.

In addition to these advantages, there is clinical potential in the PSRPs. One may cross the PSRPs with treatment response, for example, to identify situations in which risk status of a component associates with differential response. Preliminary results toward this end (not shown) are promising. In particular, we have observed that patients at-risk for DNA repair pathways respond less well to cis-platinum, which is expected given that cis-platinum induces further DNA damage. Other associations are not as expected, and will be reported elsewhere. Taking this a step further, one could imagine refining the information provided by the PSRPs by considering pairs (or more) of risky pathways. This may prove even more useful, but is currently limited in practice, as the PSRPs derived from the TCGA ovarian cancer project are highly patient specific. Among the risk profiles derived for patients in the three datasets considered (877 patients in total), 367 are unique. In other words, there do not appear to be large subsets of patients sharing distinct sets of aberrant pathways and, as a result, increased stratification leads to sample sizes decreasing to at most a few individuals. While patient-specificity is a desirable feature of any approach that aims to facilitate truly personalized treatment, evaluating response within group becomes infeasible, at least in the way considered here, due to limited sample sizes.

In summary, the pathway-index model provides a simple and promising approach for characterizing patient-specific risk in a way that connects risk with biologically meaningful pathways. Our preliminary results demonstrate a few ways in which the PSRPs may prove useful in practice. We expect the main advantage is in characterizing the molecular components of a patient's disease, which are directly connected to function and, consequently, may prove useful in targeting treatment.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Contract/grant sponsor: National Institutes of Health: HL083806, LM007359, and GM102756; Clinical and Translational Science Institute of Southeastern Wisconsin under NIH UL1RR031973; and the Froedert Foundation.

The authors thank John Dawson, Bret Hanlon, and Michael Newton for comments that helped improve the manuscript. This work was supported by: NIH HL083806, LM007359, and GM102756; the Clinical and Translational Science Institute of Southeastern Wisconsin under NIH UL1RR031973; and the Froedert Foundation.

References

1. Collins F. Has the revolution arrived? *Nature*. 2010; 464(7289):674–675. [PubMed: 20360716]
2. Mook S, Veer L, Emiel J, Piccart-Gebhart M, Cardoso F. Individualization of Therapy Using Mammagprint®: from Development to the MINDACT Trial. *Cancer Genomics-Proteomics*. 2007; 4(3):147. [PubMed: 17878518]
3. Sparano J, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology*. 2008; 26(5):721. [PubMed: 18258979]
4. Craver M. Genetic medicine finally hitting its stride. *The Kiplinger Letter*. Jul.2010
5. Clark-Langone K, Sangli C, Krishnakumar J, Watson D. Translating tumor biology into personalized treatment planning: analytical performance characteristics of the Oncotype DX (R) Colon Cancer Assay. *BMC cancer*. 2010; 10(1):691. [PubMed: 21176237]

6. Jones S, Zhang X, Parsons DW, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*. 2008; 321:1801. [PubMed: 18772397]
7. Jones D. Pathways to cancer therapy. *Nature Reviews Drug Discovery*. 2008; 7(11):875–876.
8. Vaughan S, Coward J, Bast R, Berchuck A, Berek J, Brenton J, Coukos G, Crum C, Drapkin R, Etemadmoghadam D, et al. Rethinking ovarian cancer: recommendations for improving outcomes. *Nature Reviews Cancer*. 2011; 11(10):719–725.
9. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2006; 68(1):49–67.
10. Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*. 2009; 37(6A):3468–3497.
11. Luan Y, Li H. Group additive regression models for genomic data analysis. *Biostatistics*. 2008; 9(1):100. [PubMed: 17513311]
12. Wei Z, Li H. Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics*. 2007; 8(2):265. [PubMed: 16772399]
13. Park M, Hastie T, Tibshirani R. Averaged gene expressions for regression. *Biostatistics*. 2007; 8(2):212. [PubMed: 16698769]
14. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Staudt LM. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*. 2002; 346(25):1937–1947. [PubMed: 12075054]
15. Wang S, Nan B, Zhu N, Zhu J. Hierarchically penalized Cox regression with grouped variables. *Biometrika*. 2009; 96(2):307.
16. Huang J, Ma S, Xie H, Zhang C. A group bridge approach for variable selection. *Biometrika*. 2009; 96(2):339. [PubMed: 20037673]
17. Zhou N, Zhu J. Group variable selection via a hierarchical lasso and its oracle property. *Arxiv preprint arXiv:1006.2871*. 2010.
18. Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society, Series B*. 1972; 34:187–220.
19. Tibshirani R. Univariate shrinkage in the cox model for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*. 2009; 8(1):e21.
20. Ma S, Song X, Huang J. Supervised group Lasso with applications to microarray data analysis. *BMC bioinformatics*. 2007; 8(1):60. [PubMed: 17316436]
21. Zhao P, Yu B. On model selection consistency of lasso. *The Journal of Machine Learning Research*. 2006; 7:2541–2563.
22. Tibshirani R. The lasso method for variable selection in the cox model. *Statistics in Medicine*. 1997; 16:385–395. [PubMed: 9044528]
23. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011; 474:609–615. [PubMed: 21720365]
24. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*. 2010; 38:D355–D360. [PubMed: 19880382]
25. Karrison T. Restricted mean life with adjustment for covariates. *Journal of the American Statistical Association*. 1987; 1169–1176.
26. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, Johnson DS, Trivett MK, Etemadmoghadam D, Locandro B, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*. 2008; 14(16):5198–5208. [PubMed: 18698038]
27. Bild A, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*. 2006; 439:353–357. [PubMed: 16273092]
28. Fan J, Song R. Sure independence screening in generalized linear models with np-dimensionality. 2010

29. Heagerty P, Lumley T, Pepe M. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*. 2000; 56(2):337–344. [PubMed: 10877287]
30. Berchuck A, Iversen ES, Lancaster JM, Pittman J, Luo J, Lee P, Murphy S, Dressman HK, Febbo PG, West M, et al. Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. *Clinical Cancer Research*. 2005; 11(10):3686–3696. [PubMed: 15897565]
31. Yoshihara K, Tajima A, Yahata T, Kodama S, Fujiwara H, Suzuki M, Onishi Y, Hatae M, Sueyoshi K, Fujiwara H, et al. Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. *PLoS One*. 2010; 5(3):e9615. [PubMed: 20300634]
32. Dressman HK, Berchuck A, Chan G, Zhai J, Bild A, Sayer R, Cragun J, Clarke J, Whitaker RS, Li L, et al. An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *Journal of Clinical Oncology*. 2007; 25(5):517–525. [PubMed: 17290060]
33. Crijns APG, Fehrmann RSN, de Jong S, Gerbens F, Meersma GJ, Klip HG, Hollema H, Hofstra RMW, te Meerman GJ, de Vries EGE, et al. Survival-related profile, pathways, and transcription factors in ovarian cancer. *PLoS Medicine*. 2009; 6(2):e1000 024.
34. Denkert C, Budczies J, Darb-Esfahani S, Györfy B, Sehouli J, Köönsen D, Zeillinger R, Weichert W, Noske A, Buckendahl AC, et al. A prognostic gene expression index in ovarian cancer– validation across different independent data sets. *Journal of Pathology*. 2009; 218:273–280. [PubMed: 19294737]
35. Irizarry R, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed T. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003; 4(2):249–264. [PubMed: 12925520]

Figure 1.

(Left) Restricted mean life estimates for low (green), medium (yellow) and high (red) risk groups as defined by the number of positive pathway indices; 95% confidence intervals are indicated by the dashed lines. (Right) Estimates of the survival function confirm the magnitude of risk stratified by the fitted model.

Figure 2. Validation set Kaplan-Meier curves for risk groups defined by the number of positive pathway indices (index-count model) confirm the ordering and magnitude of effects estimated in our training set analysis. The three panels are the TCGA test set (left) and two independent studies (middle and right).

Figure 3.

Receiver operating curves (ROC) for all five methods show the superior performance of index-count in all four data sets. Note that the model parameters are first estimated in the TCGA training set and fixed. Predictions in other data sets are based on passing new information through the fitted model.

Table 1

Selection results for KEGG pathways most representative of the core cancer pathways identified in Jones et al. [6]. The pathways contain 1216 genes (984 are unique). The 104 (88 unique) selected by the pathway index model are shown for each pathway, along with the symbols for the three genes with the largest estimated effects.

KEGG Pathway Name	# Genes Selected	Top 3 genes
MAPK signaling	19 of 243	DUSP8, PLA2G2D, AKT2
Wnt signaling	18 of 133	APC, ROCK1, PPP3CA
Base excision repair	15 of 31	SMUG1, PARP3, POLD1
Mismatch repair	13 of 22	MLH3, POLD1, MSH6
mTOR signaling	11 of 48	AKT2, AKT1, RPS6KA2
TGF- β signaling	8 of 80	LTBP1, MYC, ID4
Non-homologous end-joining	8 of 13	XRCC5, XRCC4, DNTT
Apoptosis	7 of 84	AKT2, PPP3CA, CASP8
Hedgehog signaling	5 of 44	CSNK1G3, WNT5B, GAS1
Phosphatidylinositol	0 of 72	–
Notch signaling pathway	0 of 40	–
Jak-STAT signaling pathway	0 of 139	–
Nucleotide excision repair	0 of 41	–
Cell cycle	0 of 107	–
Cell adhesion molecules	0 of 119	–

Table 2

Validation statistics for the training and three test and validation sets. The change in the hazard ratio (Δ HR) reflects the % change in risk for each additional positive pathway index. The PH tests imply that model assumptions are met. The restricted means indicate the number of months expected across risk groups defined in the training data only.

Data	Δ HR per Pathway	p-value	PH test	Restricted Mean Life (s.e.)		
				Low	Med	High
TCGA Training	1.50	0.00e+00	0.8620	53.2(2.1)	36.9(1.1)	25.6(0.6)
TCGA Testing	1.15	3.89e-03	0.0702	43.5(1.5)	40.5(1.3)	33.6(1.0)
Tothill Validation	1.17	1.85e-03	0.1290	45.1(1.6)	40.0(1.3)	35.1(1.0)
Bild Validation	1.32	3.93e-05	0.5140	50.4(2.4)	39.2(1.6)	29.1(1.0)

Table 3

Proportion of cases where each method fails to select any features is shown along with averages of operating characteristics calculated across 500 simulations for each of 6 methods and 5 simulation scenarios, SI-SV (described in the text). Standard errors are order 0.001. The method with the highest average validation performance is highlighted.

Method	Screening:		SIS		Lasso		Average		Index	
	Final Model:	Lasso	Lasso	Lasso	Lasso	Lasso	Lasso	Lasso	Count	Lasso
Proportion of cases when method fails to select any features										
	SI	0.17	0.02	0.17	0.48	0.00	0.00	0.00	0.00	0.00
	SII	0.05	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00
	SIII	0.11	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00
	SIV	0.05	0.00	0.12	0.45	0.00	0.00	0.00	0.00	0.00
	SV	0.17	0.01	0.18	0.59	0.00	0.00	0.00	0.00	0.00
Correlation between true and predicted ranks										
	SI	0.26	0.21	0.18	0.00	0.13	0.20	0.13	0.20	0.20
	SII	0.26	0.33	0.43	0.99	0.32	0.46	0.32	0.46	0.46
	SIII	0.16	0.23	0.30	0.84	0.29	0.33	0.29	0.33	0.33
	SIV	0.39	0.36	0.41	0.02	0.37	0.45	0.37	0.45	0.45
	SV	0.17	0.20	0.23	0.01	0.23	0.24	0.23	0.24	0.24
Manhattan distance relative to all same rank.										
	SI	0.56	0.58	0.60	0.67	0.61	0.58	0.61	0.58	0.58
	SII	0.56	0.53	0.48	0.06	0.53	0.47	0.53	0.47	0.47
	SIII	0.60	0.57	0.54	0.25	0.54	0.53	0.54	0.53	0.53
	SIV	0.50	0.51	0.49	0.66	0.51	0.47	0.51	0.47	0.47
	SV	0.59	0.58	0.57	0.66	0.56	0.56	0.56	0.56	0.56
Misclassification in 10% tails										
	SI	0.29	0.32	0.38	0.50	0.40	0.36	0.40	0.36	0.36
	SII	0.29	0.23	0.20	0.00	0.27	0.18	0.27	0.18	0.18
	SIII	0.37	0.31	0.29	0.01	0.29	0.27	0.29	0.27	0.27
	SIV	0.19	0.21	0.21	0.48	0.23	0.18	0.23	0.18	0.18
	SV	0.35	0.33	0.34	0.50	0.34	0.33	0.34	0.33	0.33