Assessing the clinical utility of cancer genomic and proteomic data across tumour types

Yuan et al, 2014

Karina Isaev
Journal Club, January 22nd

Introduction

- Systematic study integrating molecular data with clinical variables
- Retrospectively predict patient survival
- Build reliable prognostic and therapeutic methods that incorporate patient molecular data

Clinical Utility of TCGA Data

- Improve accuracy of prognosis
  - Stratify patients into risk groups to provide best treatment and surveillance strategies
- Age and tumour stage are common clinical prognostic variables
  - Can we incorporate molecular data to improve prognosis?

Molecular Biomarkers

- ER, PR, HER2 protein levels and HER2 amplification in breast cancer
  - Small number of selected genes studied using limited platforms
- Patients getting selected for clinical trials based on presence of mutation
  - Current studies: catalogue alterations in clinically actionable genes

Purpose of Study

- Address how and to what degree, TCGA molecular data could impact oncology practice
- <u>Prognostic Utility</u>
  - Predict patient survival using various types of high-throughput molecular data across tumours
- <u>Therapeutic Utility</u>
  - Identify spectrum of somatic alterations in clinically actionable genes to eventually improve treatment selection

Purpose of Study

1. Evaluate performance of SCNA, DNA methylation and mRNA, microRNA and protein expression alone or in combination with clinical variables in predicting survival

2. Investigate spectrum of potentially actionable clinical alterations across 12 tumour types

Establishing Data Sets

- **4 cancer types** (KIRC, GBM, OV, LUSC)
  - TCGA datasets with survival information and enough samples characterized by multiple molecular data
- **SCNA**: ~ 100 arm of focal alterations SNP Array
- **DNA Methylation**: ~20,000 genes microarray
- **mRNA Expression** ~ 20,000 genes
- **miRNA Expression** > 500 microRNAs
- **Protein Expression** ~ 170 proteins (reverse phase protein array)

<u>Core sample set:</u>
each sample has information for survival time, clinical variables and at least 4/5 types of molecular data

Assessing Prognostic Power of Molecular Data

1. For each core set, applied Monte Carlo cross-validation to assess the predictive power of each molecular data type and clinical variables

    - Concordance Index (C-index), nonparametric measure to quantify the discriminatory power of predictive model

    - C-index = 1 indicates perfect prediction accuracy

    - C-index = 0.5 as good as random guess

2. Compiled candidate features and randomly split the core set into training and test sets 100 times

Assessing Prognostic Power of Molecular Data

1. C-index

- Rank order statistic for predictions against true outcomes
- Ratio of the concordant pairs to the total comparable pairs
- When comparing two people in a pair, one with longer survival time should have lower HR = concordance

Assessing Prognostic Power of Molecular Data

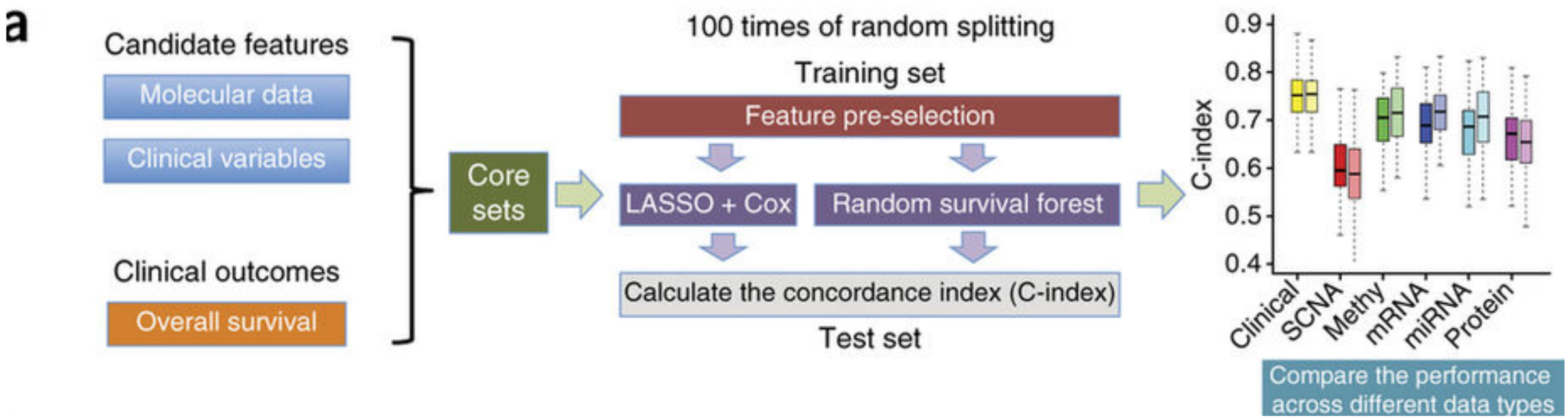


Figure 1

Assessing Prognostic Power of Molecular Data

3. Predictive model built from training set using:
    1. Cox-proportional hazards model with L1 penalized log partial likelihood (LASSO, feature selection)
    2. Random survival forest (RSF)
4. Predictive models integrating molecular data (both gene-level and molecular subtype features) and clinical data

Assessing Prognostic Power of Molecular Data

3. Predictive model built from training set using:

    1. **Cox-proportional hazards model with L1 penalized log partial likelihood (LASSO, feature selection)**

    2. Random survival forest (RSF)

4. Predictive models integrating molecular data (both gene-level and molecular subtype features) and clinical data

Assessing Prognostic Power of Molecular Data

Cox-proportional hazards model with L1 penalized log partial likelihood (LASSO, feature selection)

- Cox model is expressed by the hazard function:

$$h(t) = h_0(t) \times exp(b_1 x_1 + b_2 x_2 + \ldots + b_p x_p)$$

- h0 is the baseline hazard corresponding to the hazard if all the variable coefficients are set to 0
- *exp*(*b*i) are the hazard ratios (HR)
- A covariate with HR > 1 is called a bad prognostic factor
- A covariate with HR < 1 is called a good prognostic factor

Assessing Prognostic Power of Molecular Data

Cox-proportional hazards model with L1 penalized log partial likelihood (LASSO, feature selection)

- LASSO forces the sum of absolute value of regression coefficients to be less than a fixed value
- Forces some coefficients to be zero, allowing for a simpler model
- Performs both variable selection and regularization

Assessing Prognostic Power of Molecular Data

3. Predictive model built from training set using:
    1. Cox-proportional hazards model with L1 penalized log partial likelihood (LASSO, feature selection)
    2. Random survival forest (RSF)
4. Predictive models integrating molecular data (both gene-level and molecular subtype features) and clinical data

Assessing Prognostic Power of Molecular Data

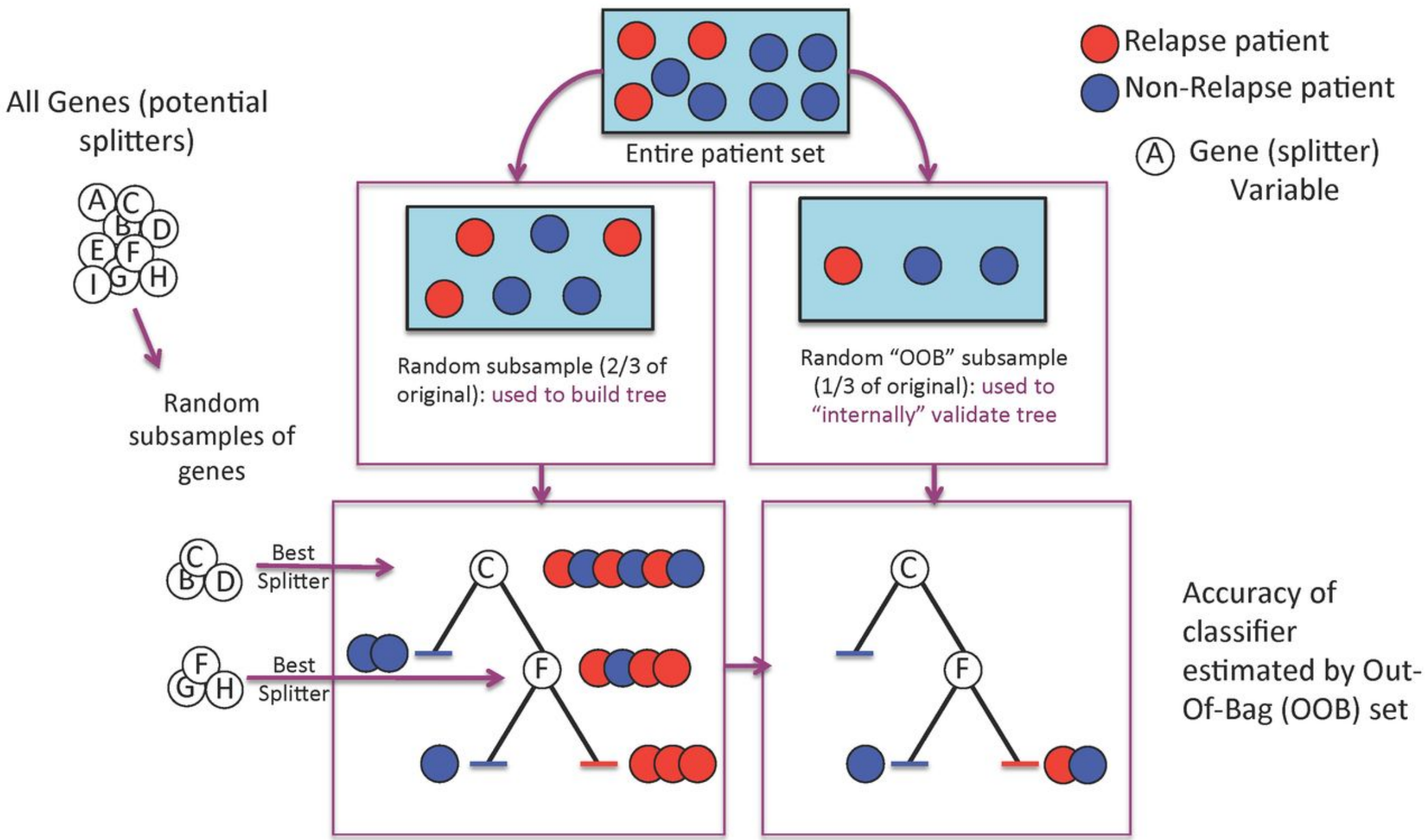
3. Predictive model built from training set using:
    1. Cox-proportional hazards model with L1 penalized log partial likelihood (LASSO, feature selection)
    2. **Random survival forest (RSF)**
4. Predictive models integrating molecular data (both gene-level and molecular subtype features) and clinical data

Assessing Prognostic Power of Molecular Data

Random survival forest (RSF)

- Randomization is introduced in two forms:
  - Randomly drawn bootstrap sample of the data is used to grow a tree
  - At each node of the tree, randomly selected subset of variables chosen as candidate variables for splitting
  - Maintains generalization

Assessing Prognostic Power of Molecular Data



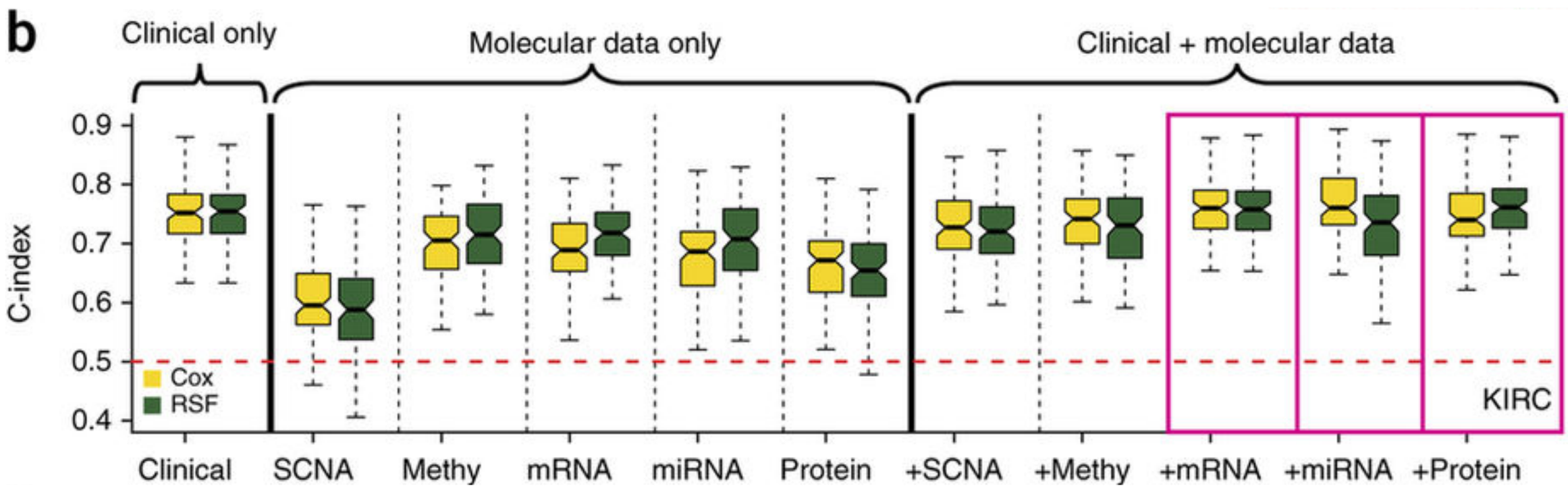
https://www.biostars.org/p/86981/

Assessing Prognostic Power of Molecular Data

3. Predictive model built from training set using:
    1. Cox-proportional hazards model with L1 penalized log partial likelihood (LASSO, feature selection)
    2. Random survival forest (RSF)
4. Predictive models integrating molecular data (both gene-level and molecular subtype features) and clinical data

Assessing Prognostic Power of Molecular Data

KIRC ($N_{\text{total}} = 243$)

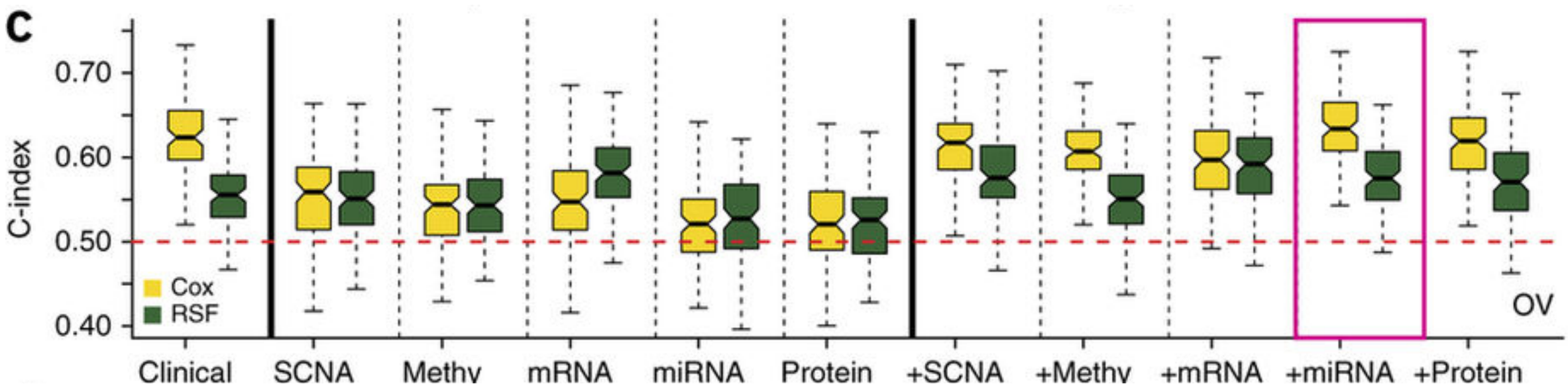


Median Somers' D = 4.0%, 7.4%, 2.2%

Figure 1

Assessing Prognostic Power of Molecular Data

OV (N_{total} = 379)



Median Somers' D =13.7%

Figure 1

Assessing Prognostic Power of Molecular Data

GBM (N_{total} = 210)

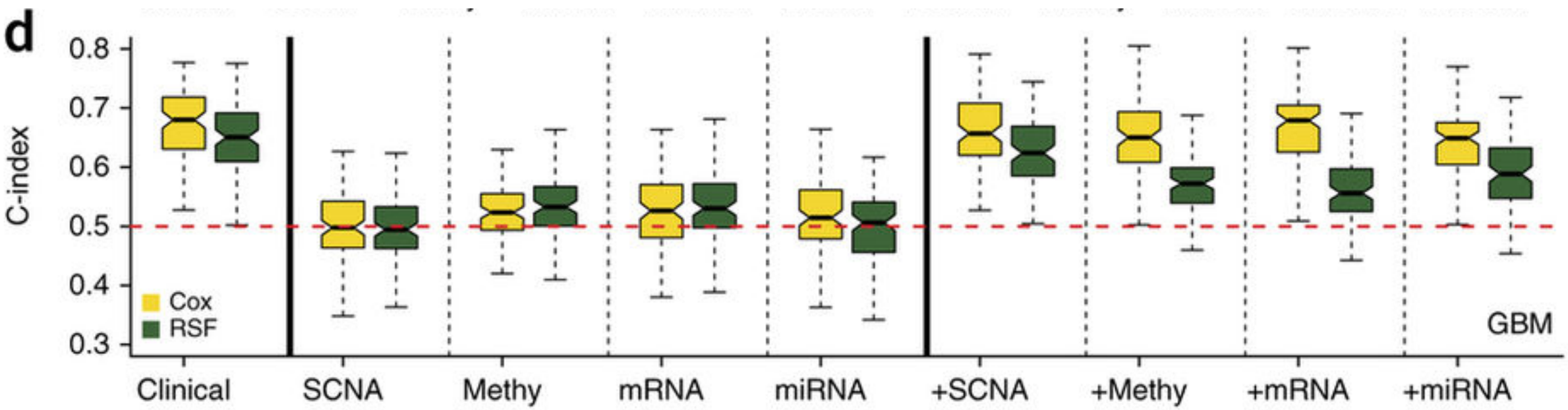
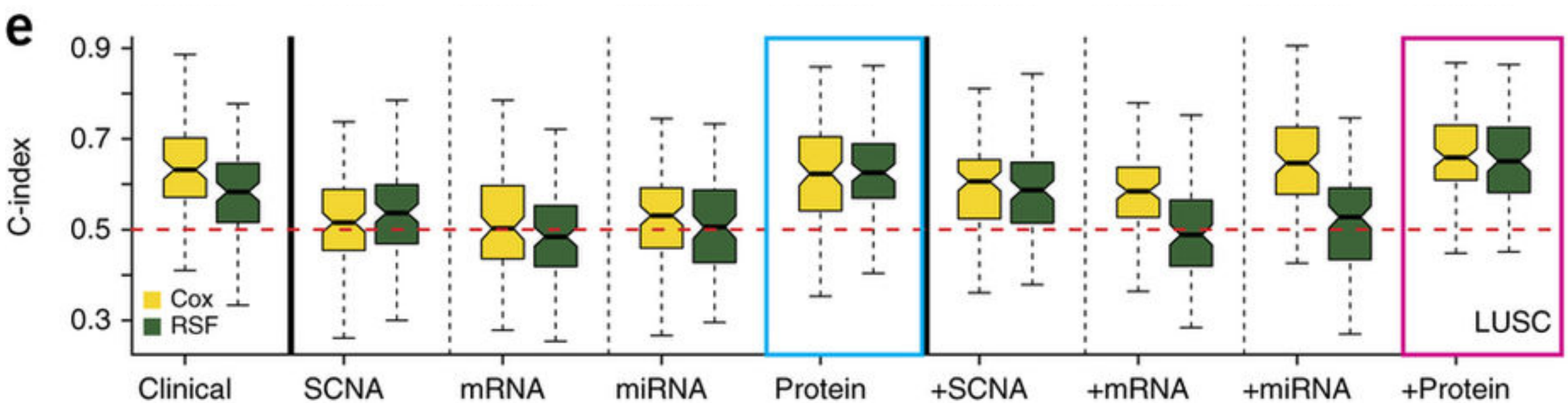


Figure 1

Assessing Prognostic Power of Molecular Data

LUSC (N_{total} = 121)



Median Somers' D =23.9%

proteins involved in DNA repair and micrositellite instability (MSH2) and metabolism (ACC1)

Figure 1

Biological Insights from top-performing prognostic models

- Molecular data in five integrative models conferred additional prognostic power
  - In 4/5, only non-clinical contributor feature was the **molecular subtype** derived from expression
    - Used consensus non-negative matrix factorization (NMF)
    - “Molecular subtypes can be regarded as higher-level assemblies of individual gene features and therefore may act as a more robust predictor than an individual marker”

Biological Insights from top-performing prognostic models

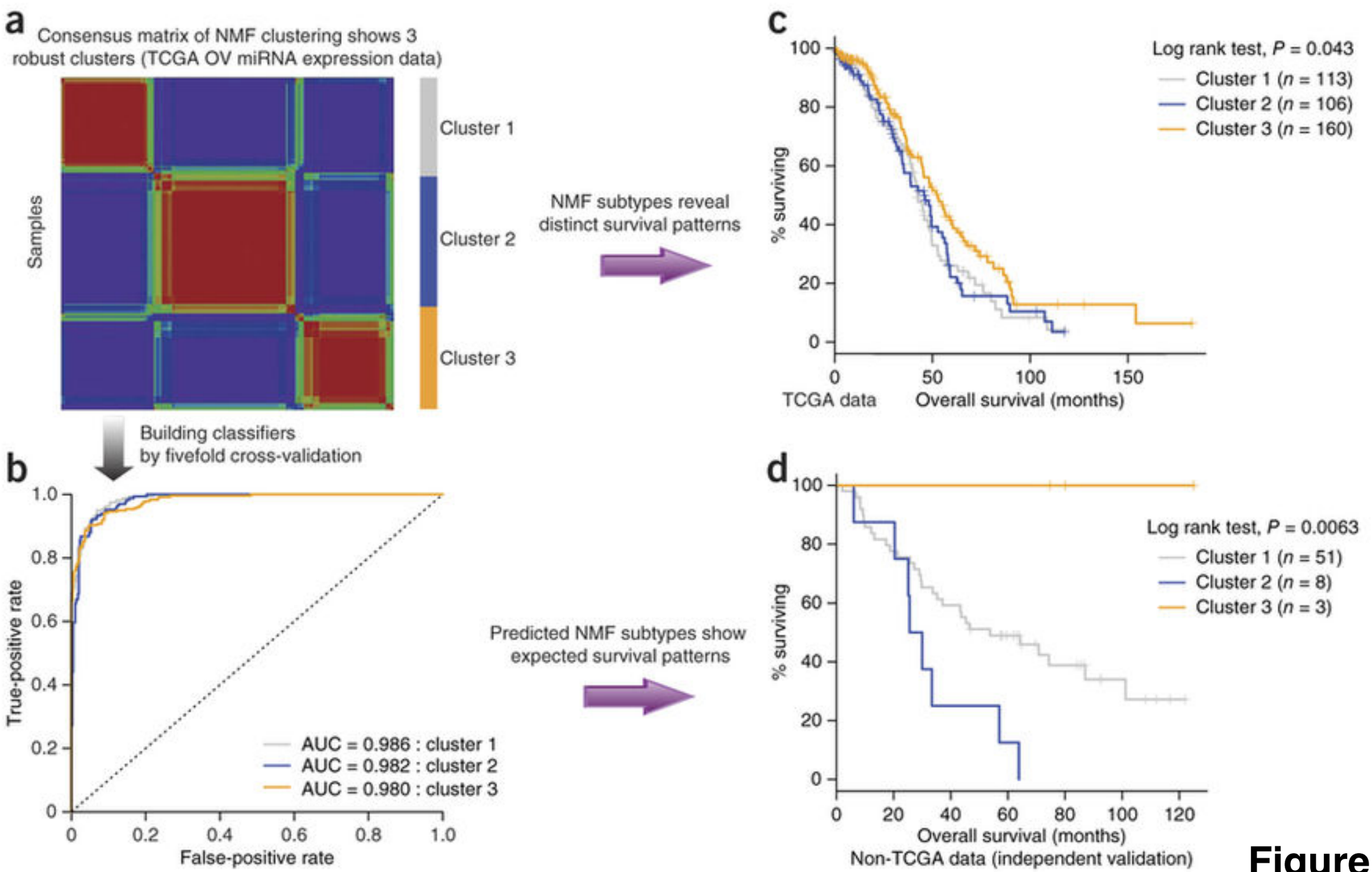
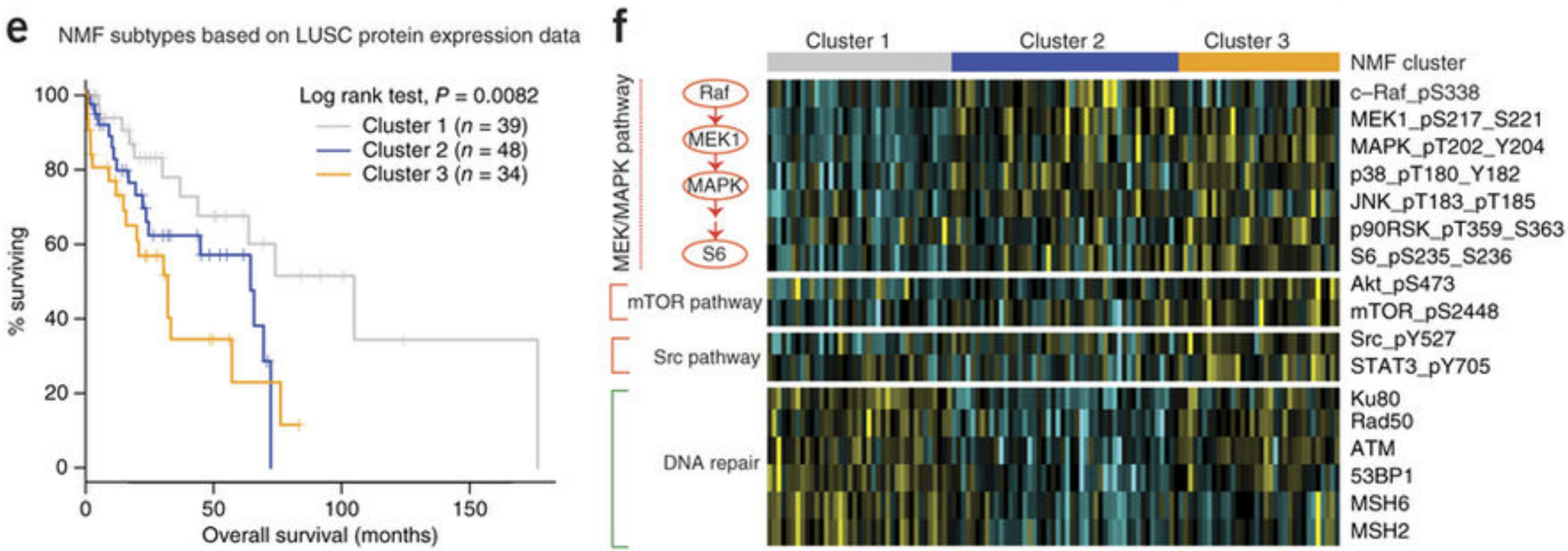


Figure 2

Biological Insights from top-performing prognostic models



pMEK1 and pMAPK top markers expressed at higher levels in patients with shorter survival (clusters 2+3)

low DNA repair in clusters 2+3

Figure 2

Patient survival prediction using cross-tumour models

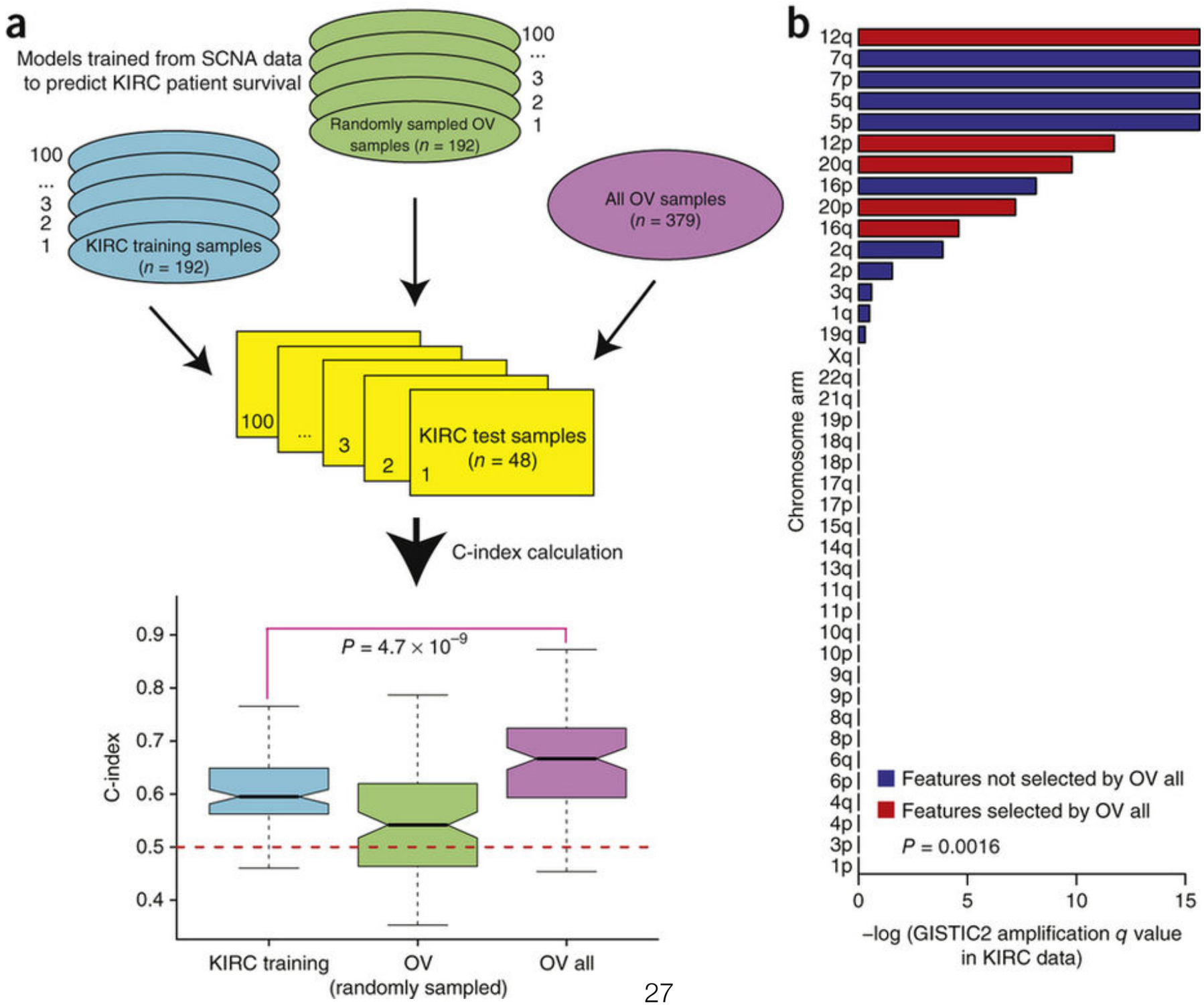
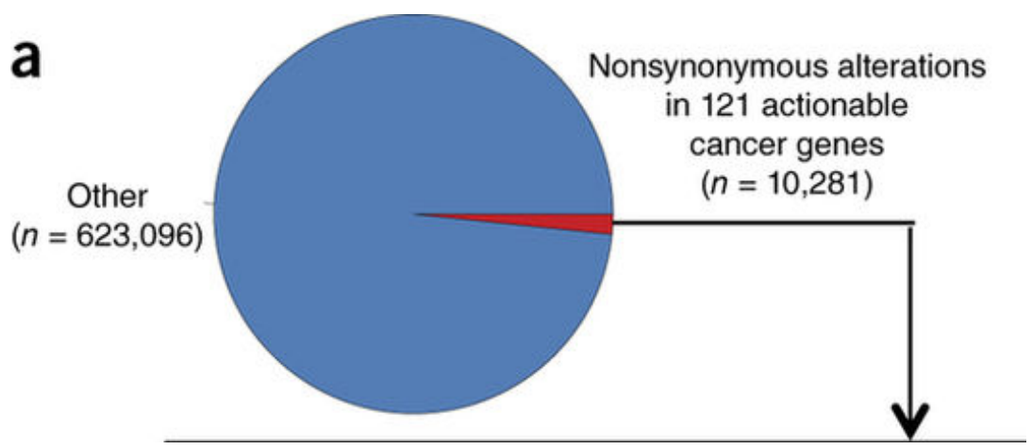


Figure 3

Somatic Alterations in Clinically Relevant Genes

- Final assessment of therapeutic utility using TCGA data
- Analyzed somatic mutations and indels in 3,277 patients across 12 tumour types
- Scored the clinical importance of each alteration in 121 clinically relevant genes
    - Somatic alteration may predict response to therapy or have diagnostic or prognostic relevance
    - Highlight that “relevant” != driver
    - Majority of these genes remain of uncertain clinical significance and require further evaluation



- Mutation —> Mutation specific therapy
- “tail” of low frequency alterations

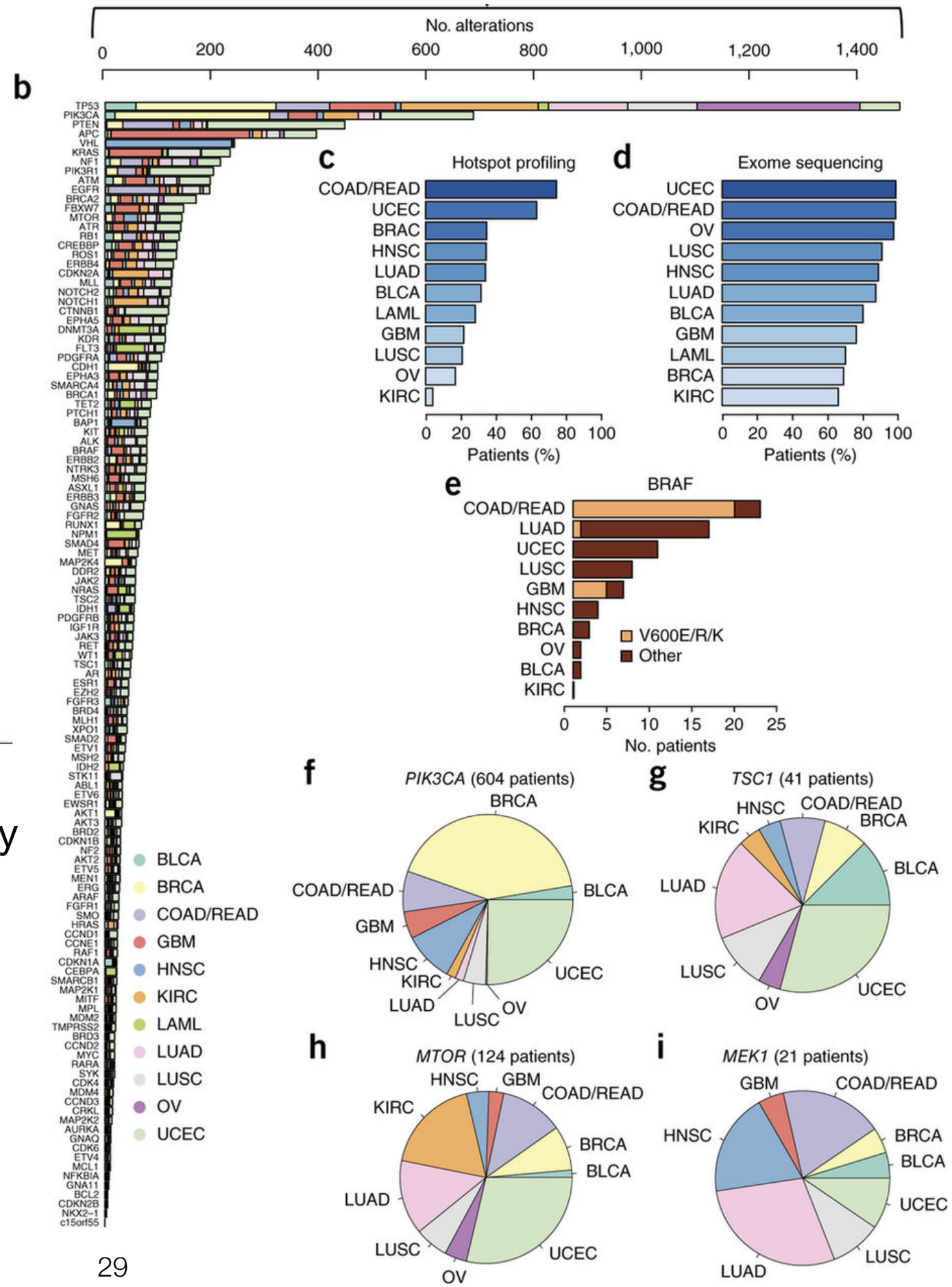


Figure 5

Discussion

- One key issue author described —> statistical significance versus magnitude difference

<u>Limitations:</u>

- Purely data-mining approaches to prognostic modelling versus candidate gene approach driven by some prior knowledge
- Did not analyze somatic mutation presence for prognostic utility since sparse across cohorts
- Combining multiple types of data —> overfitting?
- Feature selection methods

Figure 4 : Predictive performance of clinical variables, molecular data and their combination on dichotomized survival data.