Enhancing the Prediction of Prostate Cancer Recurrence with Multi-omic Molecular and Clinical Data

UCLA Health
Jonsson Comprehensive Cancer Center

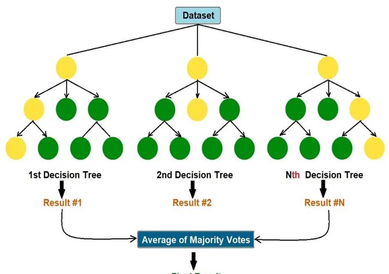
Adriana Wiggins[1], Jaron Arbet[2-4], Paul C. Boutros[2-4]

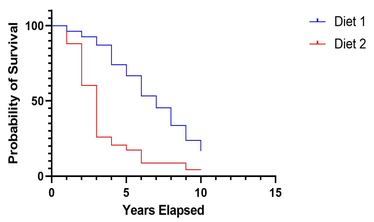
Affiliations listed below

Abstract

- **Clinical measurements** such as age, PSA, ISUP grade, tumour stage are used to inform prostate cancer treatment decisions; profound heterogeneity remains.
- To address the need to develop more robust biomarkers, we evaluated the predictive power of **Random Forest** and **CoxPH** models using various combinations of clinical and **multi-omic tumour features** including DNA methylation, CNA, RNA, and driver mutation data.
- Our findings suggest that **combining molecular and clinical information improves accuracy** of predicting disease prognosis and personalized cancer treatment.

Background





Random Forest[1]

a. Draw B number of bootstrap samples from a dataset (with replacement)
b. Randomly select M number of features to fit a 'tree' to each of the B bootstrap samples
c. Predictions or prediction accuracy across all B number of trees/models are averaged

- This method produces more stable estimates (reduced variance) compared to a single tree.

Cox Proportional Hazards Regression[2]

- A semi-parametric model of the survival curve (hazard/event rate as a function of time)
  - Makes no assumptions of the baseline event rate
  - Assumes event rates are proportional across patient groups
- Regularized: performs variable selection by attempting to remove unimportant variables.
- Supports any time-to-event outcome, not just death.

Methods

- Data was collected from 3 cohorts[3-5] and included clinical variables and multi-omic variables: DNA methylation, CNA, RNA, and driver mutations.
- Features with >30% missing values were removed. Remaining missing values were imputed using KNN[6].
- The final, imputed dataset has 774 patients, which was split randomly into 70% Training and 30% Test data.
- A RF and CoxPH model, both predicting **time-until-biochemical recurrence (BCR)**, were trained and tested on clinical data and five other combinations of clinical and multi-omic tumour data.
- C-index was used to measure predictive performance (0 to 1, higher is better).
- **Feature Screening**:
  - ~50,000 features were tested for association with time-until-BCR (CoxPH
  - adjusting for pre-treat clinical features), only features with $p < 0.005$ were kept.
  - Several p-value cutoffs were considered, with $\alpha = 0.005$ yielding the best results.

References

[1] Arya, N. (2023, November 9). A guide to random forest in machine learning. EJable. https://www.ejable.com/tech-corner/ai-machine-learning-and-deep-learning/random-forest-in-machine-learning/

[2] The Ultimate Guide to Survival Analysis. Graphpad. (2024). https://www.graphpad.com/guides/survival-analysis

[3] Fraser M, Sabelnykova VY, Yamaguchi TN, Heisler LE, Livingstone J, Huang V, et al. Genomic hallmarks of localized, non-indolent prostate cancer. Nature 2017 doi 10.1038/nature20788.

[4] Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, et al. The molecular taxonomy of primary prostate cancer. Cell 2015;163(4):1011-25.

[5] Kirby MK, Ramaker RC, Roberts BS, Lasseigne BN, Gunther DS, Burwell TC, et al. Genome-wide DNA methylation measurements in prostate tissues uncovers novel prostate cancer diagnostic biomarkers and transcription factor binding patterns. BMC cancer 2017;17:1-10.

[6] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, Russ B. Altman, Missing value estimation methods for DNA microarrays , Bioinformatics, Volume 17, Issue 6, June 2001, Pages 520–525, https://doi.org/10.1093/bioinformatics/17.6.520

Affiliations

[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA.
[2] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, California.
[3] Institute for Precision Health, University of California, Los Angeles, Los Angeles, California.
[4] Department of Human Genetics, University of California, Los Angeles, Los Angeles, California.

Results

- The model trained on only clinical data yielded a test C-Index of **0.807** for RF and 0.641 for CoxPH.
- The clinical, methylation, and RNA model yielded a test C-Index of **0.852** for RF and 0.631 for CoxPH.
- **# features by data combination after feature screening:** Pre-treatment clinical data (P) = 5 features. Clinical and methylation (PM) = 12; Clinical, methylation, and RNA (PMR) = 19; Clinical, methylation, and CNA (PMC) = 32; Clinical, methylation, and driver mutation (PMD) = 12; Clinical, methylation, RNA, CNA, and driver mutation (PMRCD) = 39 screened features.
- **Random Forest was more robust** to overfitting than CoxPH model (test and training error were much closer in Fig. 3-4).
- Feature screening (Fig 5) was necessary to improve upon the clinical only model.

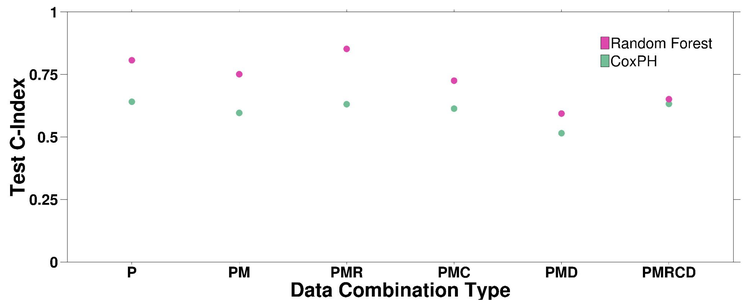


Fig.1 Test C-Index by Data Combination and Model Type. Data combination types are P (pre-treatment clinical), PM (clinical and methylation), PMR (clinical, methylation, RNA), PMC(clinical, methylation, CNA), PMD (clinical, methylation, driver mutation), PMRCD (clinical, methylation, RNA, CNA, driver mutation).

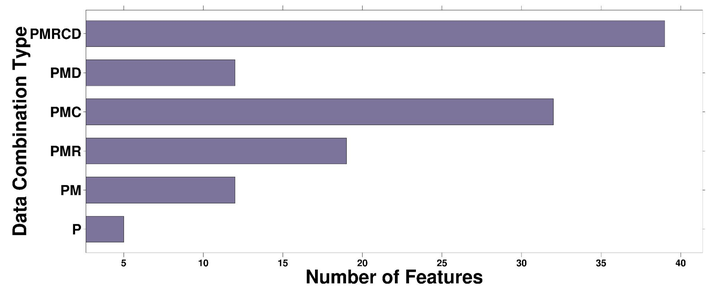


Fig.2 Number of Features by Data Combination Type. ~50,000 features were tested for association with time-until-BCR, and only features with $p < 0.005$ were kept. Data combination types are P (pre-treatment clinical), PM (clinical and methylation), PMR (clinical, methylation, RNA), PMC(clinical, methylation, CNA), PMD (clinical, methylation, driver mutation), PMRCD (clinical, methylation, RNA, CNA, driver mutation).

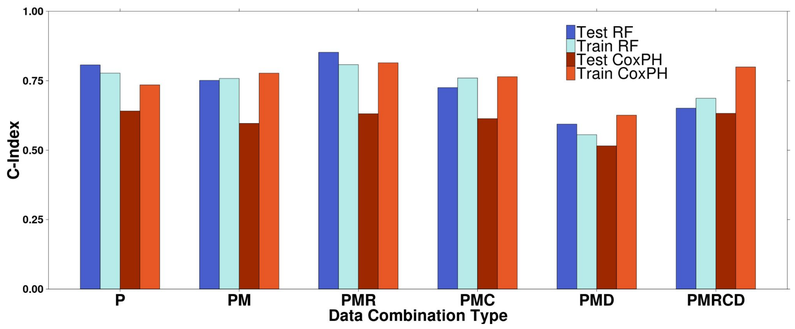


Fig.3 Train and Test C-Indexes by Data Combination Type and Model Type. Red indicates random forest (RF) and blue is CoxPH; darker values give results on Test data, while lighter values for Train data. Data combination types are P (pre-treatment clinical), PM (clinical and methylation), PMR (clinical, methylation, RNA), PMC(clinical, methylation, CNA), PMD (clinical, methylation, driver mutation), PMRCD (clinical, methylation, RNA, CNA, driver mutation).

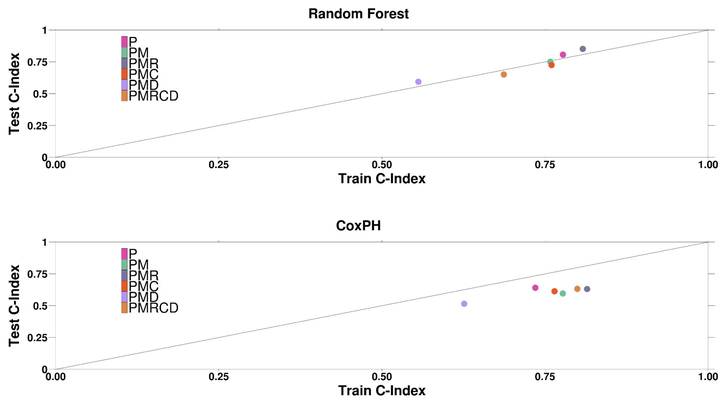


Fig.4 Test vs. Train C-Index by Data Combination Type, for RF and CoxPH. Data combination types are P (pre-treatment clinical), PM (clinical and methylation), PMR (clinical, methylation, RNA), PMC(clinical, methylation, CNA), PMD (clinical, methylation, driver mutation), PMRCD (clinical, methylation, RNA, CNA, driver mutation).

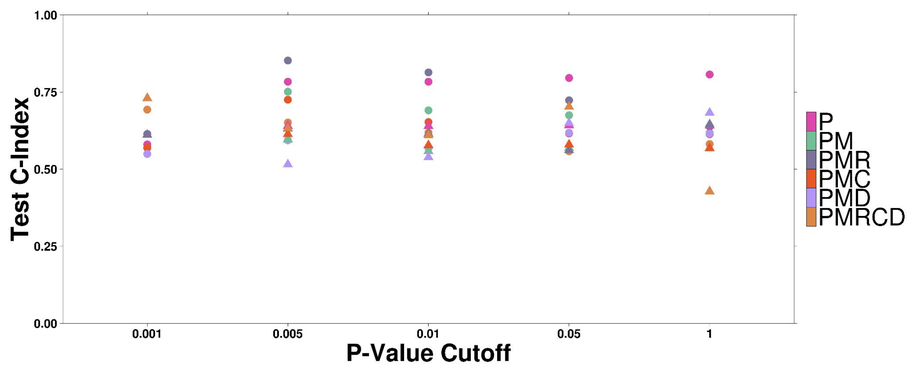


Fig.5 Test C-Index by Screening P-Value, Data Combination, and Model Type. Data combination types are P (pre-treatment clinical), PM (clinical and methylation), PMR (clinical, methylation, RNA), PMC(clinical, methylation, CNA), PMD (clinical, methylation, driver mutation), PMRCD (clinical, methylation, RNA, CNA, driver mutation). **Circles** are RF and **triangles** are CoxPH models. P-value cutoff of 1 indicates no screening. The p-value cutoff of 0.005 yielded the highest C-index.

Conclusion

- Random Forest modelling outperformed CoxPH regression for each data combination, suggesting that **RF could be more appropriate for predicting time-until-BCR using multi-omic tumour data**.
- Our findings suggest that **molecular data could add to the predictive power of clinical data**, so its combination could improve prognosis and better inform the treatment a patient decides to take – surgery or radiation.
- Personalization of cancer treatment is instrumental in its success and the comfort of the patient.

BOUTROS LAB