Data Challenge Report

Chaitra Rao

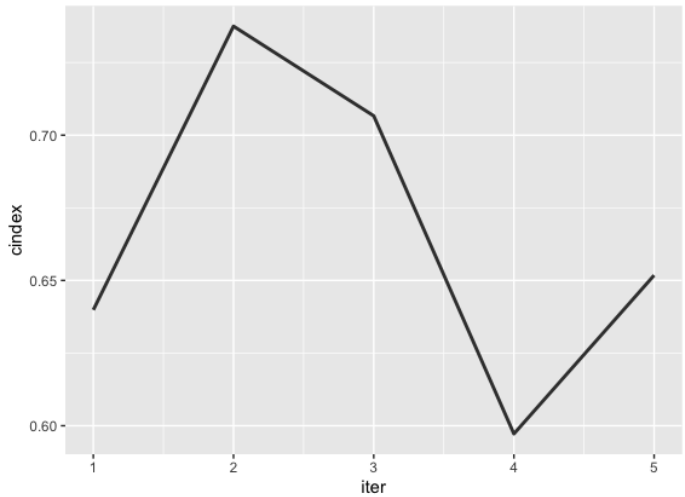
Introduction

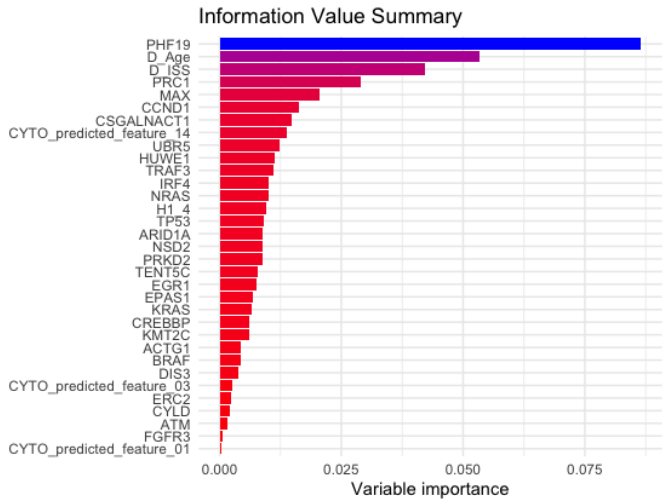
Multiple myeloma (MM) is a heterogenous disease resulting in plasma cell malignancy. It is a challenge to cure patients suffering from this disease as they are at a high risk of relapsing. Survival varies from patient to patient as multiple factors play a role. Therefore, in this analysis, the focus is on newly diagnosed MM patients and predicting their risk of dying or relapsing using expression and clinical data. The study included 583 patients aged between 20 to 95 years old. The patients are assumed to be at risk of dying or relapsing if OS or PFS < 18 months i.e. 540 days. For this analysis, PFS will be used as a surrogate for OS, assuming that more patients will progress than die, increasing the statistical power. Another important assumption made for the purpose of this analysis is that events where HR_FLAG is not TRUE (event did not occur) are assumed to be censored. Therefore, 131 out of 583 events occurred and rest are censored. From the high-dimensional expression data, MM relevant genes were filtered based on literature research (Chen et al. 2021; Mason et al. 2020; Hassan and Szalat 2021). The cytogenetic features were also filtered by building a simple 13-predictor CoxPH model to detect which out of the 13 features have a statistical significant effect on survival. The result from this model showed us that 3 cytogenetic features had a significant effect, therefore these were selected for the further analysis.

Machine learning (ML) and Proportional Hazard Regression algorithms will be used to evaluate the impact of risk factors on the survival time of patients newly diagnosed for MM. Following techniques were used to investigate the most predictive features of survival risk:

- Feature selection using a Random-forest method, Cox-PH model with univariate scores
- A benchmark study based on Cross-validation for comparing the performance of 2 different survival analysis algorithms.
- Training a Random Forest SRC model and evaluate feature importance using 5-fold CV

A total of 33 features were used to build the predictive model: expression levels of 28 MM relevant genes, 3 cytogenetic features, age and ISS. The mlr package (https://mlr-org.github.io/mlr-tutorial/release/html/index.html (https://mlr-org.github.io/mlr-tutorial/release/html/index.html)) is used to build the models.



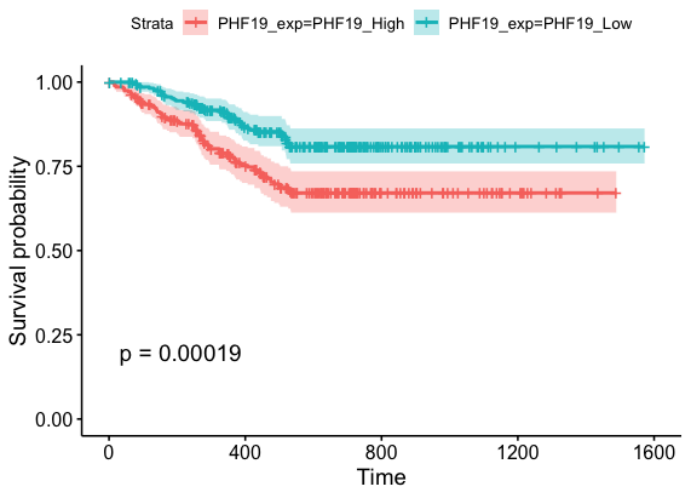


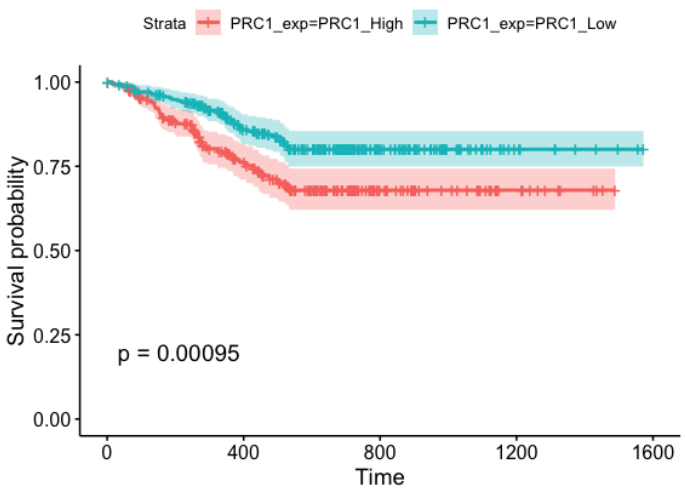
Interpretation

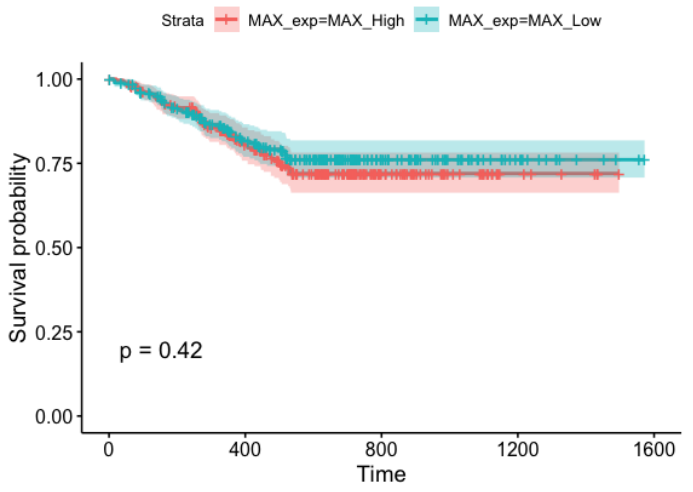
Model details:

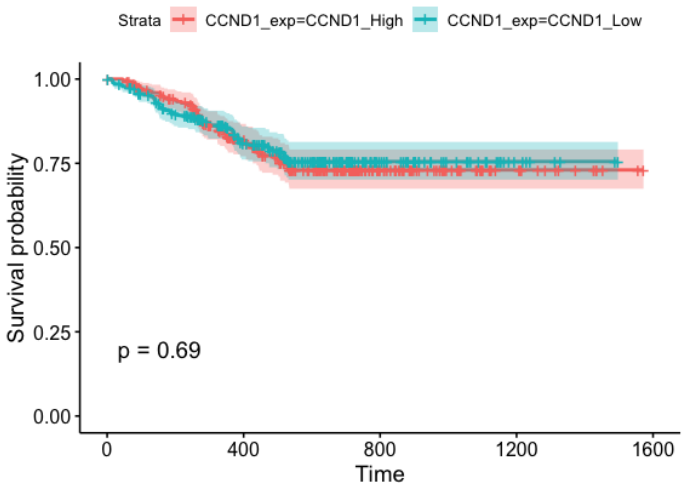
- Type: Random Forest SRC Model
- Hyperparameters: mtry=10; ntree=100 (tuned)
- Evaluation metric on CV set: Concordance index

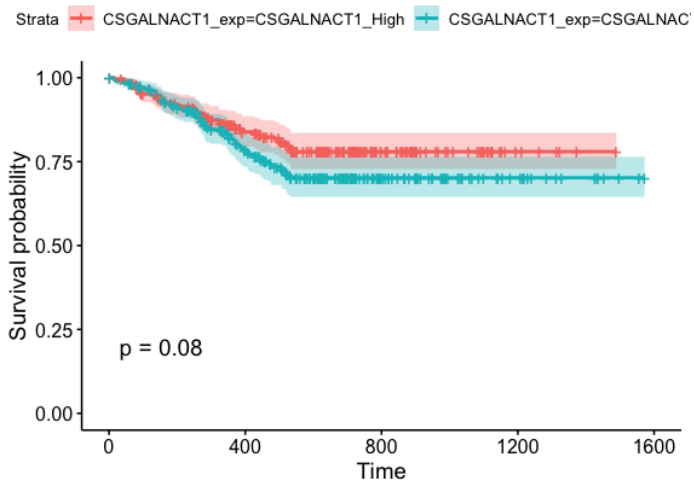
The feature importance plot on the right shows the features ranked by their importance in risk prediction. The feature importance is the mean decrease in accuracy. The top 5 genes which we observe as most important predictors are PHF19, PRC1, MAX, CCND1 and CSGALNACT1. Literature suggests that the overexpression of PHF19 is highly associated with worse clinical outcome in MM patients. It is known to increase tumorigenicity through H3K27me3 (tri-methylation) (Schinke et al. 2021). High expression of PRC1 has also been previously observed as disadvantageous for the survival outcome for MM patients (Chen et al. 2021). It is involved in cytokinesis of various other cancers as well (breast cancer and lung adenocarcinoma), possible through Wnt/beta-Catenin signaling pathway (Zhan et al. 2017). In the Kaplan-Meier curves for these two genes, we observe how high expression is associated with higher risk of survival. CSGALNACT1 is known to be a “protective” gene in MM (Chen et al. 2021). The Kaplan-Meier plot for this gene supports this hypothesis as we observe that patients belonging to CSGALNACT1_high group have better survival than patients with belonging to CSGALNACT1_low group because the survival probability of CSGALNACT1_high group is always lower than CSGALNACT1_high group over time. CCND1 is a regulatory element in cell cycle and transcriptional processes and literature suggests that the dysregulation of this gene is associated with oncogenesis (Padhi, Varghese, and Ramdas 2013). MAX alteration is observed in many MM patients (Wang et al. 2017). MAX, together with its oncogenic transcription factor MYC, can regulate gene expression by bidning to DNA enhancer boxes. Low expression of both MAX and MYC increase overall prognosis of MM.



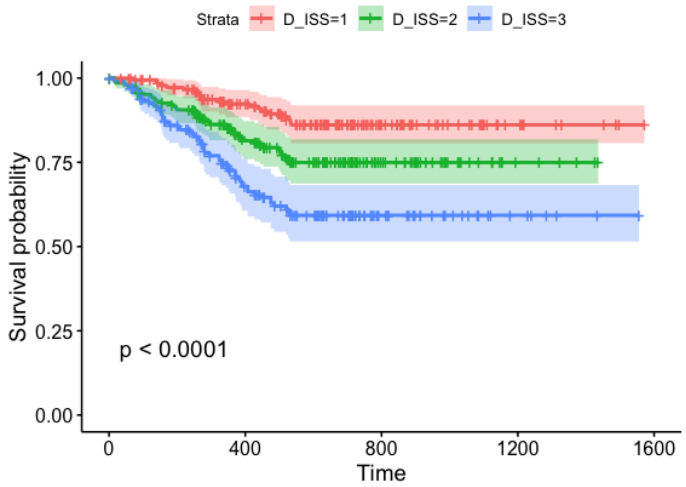


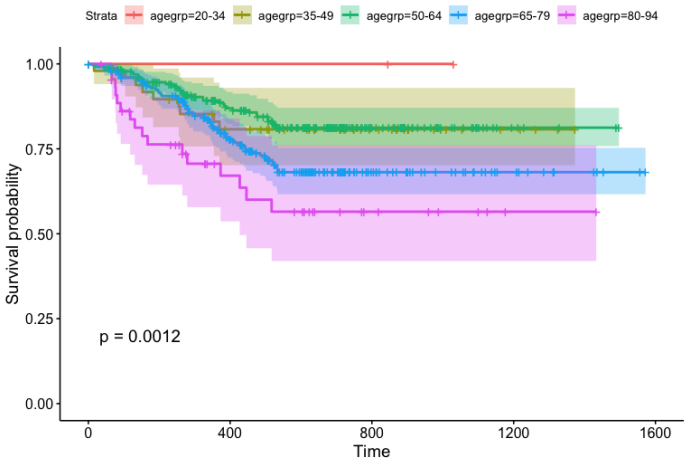


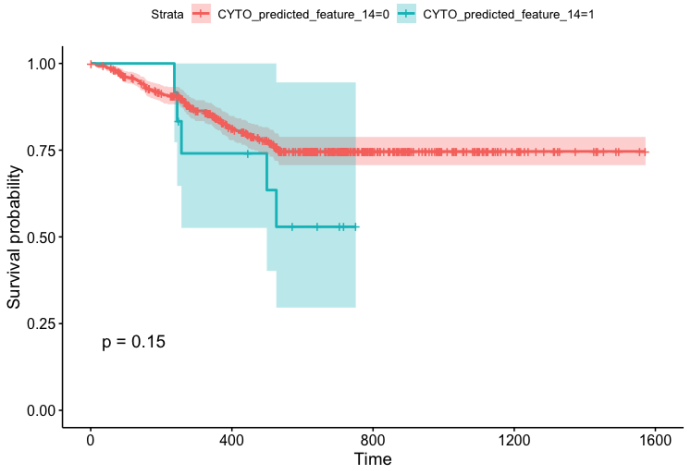




The other factors from clinical data which are observed as important predictors are age, ISS and cytogenetic feature 14 (in clinical dictionary: del(16q)). Age and ISS have also been previously recognized as important predictors of survival risk in MM patients and this analysis confirms this once again. MM is known to mostly affect elderly patients with a median age at the time of diagnosis of approximately 70 years (Zweegman et al. 2014). Kaplan-Meier plots stratified by different age groups supports this fact. Patients belonging to older age group have worse survival than those belonging to younger age groups. Plots stratified by ISS stage also show that patients from third stage group show higher risk of death or relapse indicating that ISS can increase the risk prediction accuracy. Lastly, the presence of choromosomal abnormality of del(16q) is associated with worse survival as indicated by Kaplan-Meier plot. Literature also suggests that this mutation is associated with adverse prognosis of the disease (Jenner et al. 2007) and can cause additional adverse survival impact when acting together with other cytogenetic factors (t(4;14) and del(17p)) (Avet-Loiseau et al. 2009).







Conclusion and outlook

To conclude, our prediction model trained on the training data has identified some important factors predictive of risk of death or relapse in newly diagnosed MM patients. It supports the hypothesis that MM is a very heterogeneous disease where multiple factors, including demographic and genetic, play a role in predicting clinical outcomes. However, for future work, it may be interesting to assess genes using an unbiased approach, rather than filtering only for genes with MM associations, possibly by implementing statistical, data mining, and machine learning techniques to extract patterns from the expression data. Effects of gene-gene interactions may also be worth investigating, especially interactions between the top relevant genes.

References

Avet-Loiseau, Hervé, Cheng Li, Florence Magrangeas, Wilfried Gouraud, Catherine Charbonnel, Jean-Luc Harousseau, Michel Attal, et al. 2009. “Prognostic Significance of Copy-Number Alterations in Multiple Myeloma.” *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 27 (27): 4585–90. https://doi.org/10.1200/JCO.2008.20.6136 (https://doi.org/10.1200/JCO.2008.20.6136).

Chen, Xiaotong, Lintao Liu, Mengping Chen, Jing Xiang, Yike Wan, Xin Li, Jinxing Jiang, and Jian Hou. 2021. “A Five-Gene Risk Score Model for Predicting the Prognosis of Multiple Myeloma Patients Based on Gene Expression Profiles.” *Frontiers in Genetics* 12. https://doi.org/10.3389/fgene.2021.785330 (https://doi.org/10.3389/fgene.2021.785330).

Hassan, Hamza, and Raphael Szalat. 2021. “Genetic Predictors of Mortality in Patients with Multiple Myeloma.” *The Application of Clinical Genetics* 14 (April): 241–54. https://doi.org/10.2147/TACG.S262866 (https://doi.org/10.2147/TACG.S262866).

Jenner, Matthew W., Paola E. Leone, Brian A. Walker, Fiona M. Ross, David C. Johnson, David Gonzalez, Laura Chiecchio, et al. 2007. “Gene Mapping and Expression Analysis of 16q Loss of Heterozygosity Identifies WWOX and CYLD as Being Important in Determining Clinical Outcome in Multiple Myeloma.”

Blood 110 (9): 3291–3300. https://doi.org/10.1182/blood-2007-02-075069 (https://doi.org/10.1182/blood-2007-02-075069).

Mason, Mike J., Carolina Schinke, Christine L. P. Eng, Fadi Towfic, Fred Gruber, Andrew Dervan, Brian S. White, et al. 2020. “Multiple Myeloma DREAM Challenge Reveals Epigenetic Regulator Phf19 as Marker of Aggressive Disease.” *Leukemia* 34 (7): 1866–74. https://doi.org/10.1038/s41375-020-0742-z (https://doi.org/10.1038/s41375-020-0742-z).

Padhi, Somanath, Renu G’boy Varghese, and Anita Ramdas. 2013. “Cyclin D1 Expression in Multiple Myeloma by Immunohistochemistry: Case Series of 14 Patients and Literature Review.” *Indian Journal of Medical and Paediatric Oncology : Official Journal of Indian Society of Medical & Paediatric Oncology* 34: 283–91.

Schinke, Carolina D., Jordan T. Bird, Pingping Qu, Shmuel Yaccoby, Valeriy V. Lyzogubov, Randal Shelton, Wen Ling, et al. 2021. “Phf19 Inhibition as a Therapeutic Target in Multiple Myeloma.” *Current Research in Translational Medicine* 69 (3): 103290. https://doi.org/https://doi.org/10.1016/j.retram.2021.103290 (https://doi.org/10.1016/j.retram.2021.103290).

Wang, Dongxue, Hideharu Hashimoto, Xing Zhang, Benjamin G. Barwick, Sagar Lonial, Lawrence H. Boise, Paula M. Vertino, and Xiaodong Cheng. 2017. “MAX Is an Epigenetic Sensor of 5-Carboxylcytosine and Is Altered in Multiple Myeloma.” *Nucleic Acids Research* 45 (5): 2396–2407. https://doi.org/10.1093/nar/gkw1184 (https://doi.org/10.1093/nar/gkw1184).

Zhan, Ping, Bin Zhang, Guang-min Xi, Ying Wu, Hong-bing Liu, Ya-fang Liu, Wu-jian Xu, et al. 2017. “Prc1 Contributes to Tumorigenesis of Lung Adenocarcinoma in Association with the Wnt/β-Catenin Signaling Pathway.” *Molecular Cancer* 16 (1): 108. https://doi.org/10.1186/s12943-017-0682-z (https://doi.org/10.1186/s12943-017-0682-z).

Zweegman, Sonja, Antonio Palumbo, Sara Bringhen, and Pieter Sonneveld. 2014. “Age and Aging in Blood Disorders: Multiple Myeloma.” *Haematologica* 99 (7): 1133–37. https://doi.org/10.3324/haematol.2014.110296 (https://doi.org/10.3324/haematol.2014.110296).