Data Challenge Report

Chaitra Rao

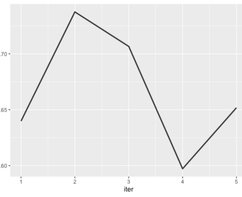
Introduction

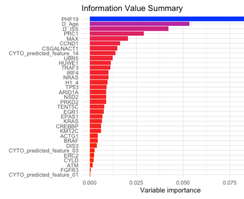
Multiple myeloma (MM) is a heterogeneous disease resulting in plasma cell malignancy. It is a challenge to cure patients suffering from this disease as they are at a high risk of relapsing and each patient responds differently to a treatment. Survival varies from patient to patient as multiple factors play a role in the disease prognosis. Hence, a better characterization of patients is required which may assist the clinical treatment of MM patients in the future. 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 (death or relapse) 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 statistically significant effect on PFS. The result from this model showed that 3 cytogenetic features had a significant effect, therefore these were selected for the further analysis.

A combination of machine learning (ML) and CoxPH algorithms will be used to evaluate the impact of risk factors on the PFS of patients newly diagnosed for MM. Following techniques were used for the prediction analysis:

- Feature selection using a random forest method and CoxPH 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) is used to build the model.





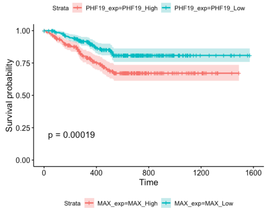
Interpretation

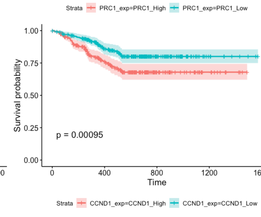
Model details:

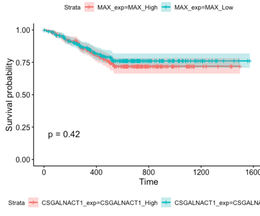
- *Type*: Random Forest SRC Model
- Evaluation metric on CV set: Concordance index
- Aggregated C-Index score = 0.66

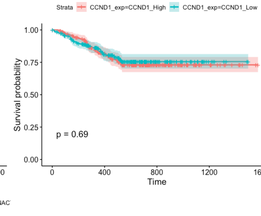
An aggregated score of 0.66 for the concordance index was achieved through cross-validation. The feature importance plot on the right shows the features ranked by their importance in risk prediction. Feature importance is the mean decrease in accuracy. The top 5 genes which are observed as most important 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 including breast cancer and lung adenocarcinoma, possibly through Wnt/beta-Catenin signalling pathway (Zhan et al. 2017). In the Kaplan–Meier curves for these two genes, it can be seen that their high expression is associated with worse PFS. CSGALNACT1 is known to be a “protective” gene in MM (Chen et al. 2021). Previous studies on predicting survival risk in MM patients have concluded that high expression of CSGALNACT1 may correlate with better prognosis in MM patients (Qi et al. 2020), and that MM cells exhibit low expression of this gene (Bret et al. 2009). Kaplan-Meier plot for this gene supports these findings as it is observed that patients belonging to CSGALNACT1_high group have better survival than patients belonging to CSGALNACT1_low group, however the difference observed in these curves is not significant. Moreover, the correlation analysis of this gene with PFS also showed a positive correlation. 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). In the context of MM, CCND1 gene amplification is associated with high percentage of plasma cell infiltration of the bone marrow resulting in bone lesions (Sewify et al. 2014). Lastly, MAX alteration is observed in MM patients (Wang et al. 2017). MAX, together with its oncogenic transcription factor MYC, can regulate gene expression by binding to DNA enhancer boxes. Low expression of both MAX and MYC increase the overall prognosis of MM.

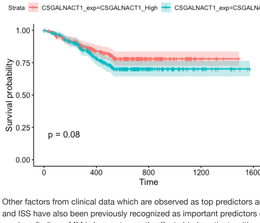
The results from prediction analysis are consistent with the MM relevant genes and PFS correlation analysis results. There, it was observed that genes such as PRC1, PHF19, CCND1 and MAX show a negative correlation with PFS and it is interesting that these features are also among the top 5 important predictors of risk and a high expression of these genes can have an adverse effect on the disease prognosis.



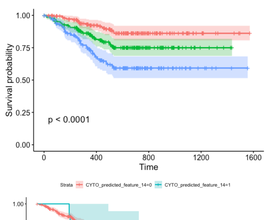


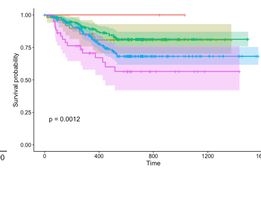


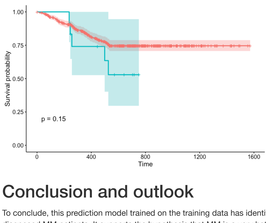




Other factors from clinical data which are observed as top predictors are age, ISS and cytogenetic feature 14 (in clinical dictionary: del(16q)). Age and ISS have also been previously recognized as important predictors of survival risk in MM patients and this analysis is consistent with the previous findings. 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 groups 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 chromosomal abnormality of del(16q) is associated with worse survival as indicated by the corresponding 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, this 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. For the purpose of this analysis, only MM relevant genes were filtered, however, for future work, it may be interesting to assess a broader spectrum of genes using an unbiased approach, 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 and results may yield superior performance when considering more granular RNA-Seq expression levels, compared to larger groups of expression level bins (high vs low). Furthermore, it might be useful to look at the correlations or associations between the expression and clinical features and draw meaningful insights from such analysis which might further increase the prediction power.

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.

Bret, Caroline, Dirk Hose, Thierry Reme, Anne-Catherine Sprynski, Karène Mahtouk, Jean-François Schved, Philippe Quittet, Jean-François Rossi, Hartmut Goldschmidt, and Bernard Klein. 2009. “Expression of Genes Encoding for Proteins Involved in Heparan Sulphate and Chondroitin Sulphate Chain Synthesis and Modification in Normal and Malignant Plasma Cells.” *British Journal of Haematology* 145 (3): 350–68. https://doi.org/https://doi.org/10.1111/j.1365-2141.2009.07633.x.

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.

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.

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.

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.

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.

Qi, Tingting, Jian Qu, Chao Tu, Qiong Lu, Guohua Li, Jiaojiao Wang, and Qiang Qu. 2020. “Super-Enhancer Associated Five-Gene Risk Score Model Predicts Overall Survival in Multiple Myeloma Patients.” *Frontiers in Cell and Developmental Biology* 8. https://doi.org/10.3389/fcell.2020.596777.

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.

Sewify, Eman M., Ola A. Afifi, Eman Mosad, Amen H. Zaki, and Sahar A. El Gammal. 2014. “Cyclin D1 Amplification in Multiple Myeloma Is Associated with Multidrug Resistance Expression.” *Clinical Lymphoma Myeloma and Leukemia* 14 (3): 215–22. https://doi.org/https://doi.org/10.1016/j.clml.2013.07.008.

Wang, Dongxue, Hideharu Hashimoto, Xing Zhang, Benjiamin 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.

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.

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.