Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-101

Predicting Clinical Outcomes in Colorectal Cancer Using Machine Learning

Julian GRÜNDNER[a,1], Hans-Ulrich PROKOSCH[a], Michael STÜRZL[b], Roland CRONER[c], Jan CHRISTOPH[a] and Dennis TODDENROTH[a]
[a] *Medical Informatics, Friedrich-Alexander University, Erlangen-Nürnberg, Erlangen, Germany*
[b] *Division of Molecular and Experimental Surgery, Department of Surgery, Friedrich-Alexander University, Erlangen-Nürnberg , Erlangen, Germany*
[c] *Department of General, Visceral, Vascular and Graft Surgery, University Hospital, Magdeburg, Germany*

Abstract. Using gene markers and other patient features to predict clinical outcomes plays a vital role in enhancing clinical decision making and improving prognostic accuracy. This work uses a large set of colorectal cancer patient data to train predictive models using machine learning methods such as random forest, general linear model, and neural network for clinically relevant outcomes including disease free survival, survival, radio-chemotherapy response (RCT-R) and relapse. The most successful predictive models were created for dichotomous outcomes like relapse and RCT-R with accuracies of 0.71 and 0.70 on blinded test data respectively. The best prediction models regarding overall survival and disease-free survival had C-Index scores of 0.86 and 0.76 respectively. These models could be used in the future to aid a decision for or against chemotherapy and improve survival prognosis. We propose that future work should focus on creating reusable frameworks and infrastructure for training and delivering predictive models to physicians, so that they could be readily applied to other diseases in practice and be continuously developed integrating new data.

Keywords. Machine-learning, colorectal cancer, survival prediction, chemotherapy, relapse, predicting clinical outcomes

1. Introduction

The early prediction of clinical outcomes in cancer therapy may inform prognosis and treatment decisions. Modern machine learning (ML) techniques have improved the prognosis accuracy by 15-20% and promise to enhance diagnosis and overall prognosis of cancer [1]. Kourou et al. found a growing trend to incorporate genomic data, as well as age, weight, diet and high-risk habits into analysis. This has led to the same type of cancer having different subgroups based on genes. They further identified lack of external validation and large datasets as main problems [1]. Colorectal cancer has been studied less than more frequently diagnosed cancers, such as breast and lung cancer [2]. In order to avoid over or under treatment (e.g. chemotherapy despite it having little effect) of patients with colorectal cancer, researchers have identified subgroups to improve prognostic accuracy for stage II and III patients [3, 4]. We apply several ML algorithms to predict the following clinically relevant outcomes: disease-free survival (DFS), survival, radio chemotherapy response (RCT-R), relapse, risk group stage II (SII), risk group stage III (SIII), DFS SII, relapse SII, DFS, SIII, and relapse SIII.

[1] Corresponding Author, Julian Gruendner, Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), julian.gruendner@fau.de

2. Methods

We evaluated predictive methods based on a dataset of clinical and mRNA gene expression attributes extracted using the RT-qPCR method [5, 6] from 564 colorectal cancer patients who had a tumor resection at Erlangen University hospital after fall 2009. Of these patients, 254 (45%) suffered from rectal carcinomas and 310 (55%) from colon carcinomas. Age ranged from 24.5 to 97 years, with an average age of 67. 140 patients (24%) received neoadjuvant therapy. Of all patients 145 (26%) have experienced a relapse, and 113 patients (20%) have died.

All models were created using the same process. Starting with data preparation, we then selected the most useful features for each prediction model. Following feature selection, survival outcomes were predicted using the methods general linear model (*glmnet)*, *coxph* and random forest for survival (*rfsrc).* Non-survival outcomes or classical categorization models were trained using the methods *glmnet*, k-nearest neighbor (*knn)*, neural network (*nnet), C50* (decision tree), random forest (*rf)* and deep neural network (*dnn)*. The best models were extracted according to performance measures described below. In order to control for overfitting, the best-performing models were subsequently evaluated using separate test data deliberately withheld before the model building process. To select features we used two approaches.

Expert manual forward feature selection, where we used feedback from our clinical expert to identify feature groups that were likely to have an impact on the prediction. We then generated all predictive models using all the gene expression data, and repeated this step successively adding additional feature groups. The following feature groups were identified: Gene Expression (59 genes pre-selected as part of a prospective study on colorectal carcinoma [5, 6]), Localization (1), Epidemiology (4: gender, smoker, weight, height), Cancer Type (1: colon/rectum), Tumor Stage (1: TNM stages I-IV).

Expert automatic feature selection, which was analogous to the expert forward feature selection process, except that instead of a simple manual forward selection we used the recursive feature elimination (RFE) method or for survival outcomes and *survRandForestLearner* to train random forest models. The most successful features were then chosen according to the effect a feature has on the respective model (variable importance). These features where then used for the model building process.

To measure model performance for right-censored survival time data, we computed the so-called C-Index, while accuracy was used for the remaining models. Models with binary outcomes (Yes/No) were selected in accordance to the Youden-Index (specificity + sensitivity – 1). This was especially important for relapse predictions among stage II and III patients, as these outcomes were substantially unbalanced insofar as most patients had *"no relapse"*. To establish the validity and reliability of the risk groups identified by Merkel et al. (2001) [3, 4] on our data, we extracted them from the data using the information from the literature. We then added the risk groups to the general data set as benchmark to test our best prediction model against using a log-rank test. The whole program was written in R version 3.4.0.

3. Results

3.1 Machine learning model performance all stages

The *glmnet* method achieved the best results predicting DFS and survival with C-Index scores of 0.76 and 0.87 on test data respectively. The DFS model used Gene Expression, Localization, Epidemiology, Cancer Type and Tumor Stage and the survival model used Gene expression and Localization. The model that predicted RCT-TR (Yes/No) best was a decision tree method using Gene expression, achieving accuracy of 0.70 on test data with a specificity of 0.85 and a sensitivity of 0.53. The outcome relapse could be predicted with an accuracy of 0.71 on test data using glmnet. This model used Gene Expression, Localization, Epidemiology, Cancer Type, Tumor Stage features. The specificity was 0.73 and the sensitivity 0.63.

3.2 Machine learning model performance cancer stages II and III

The stage II (134) and III (97) patients are of special interest as they allow for the most substantial intervention by clinicians. We predicted DFS SII using the *coxph* method with Gene Expression and Localization features. The C-Index of the model was 1 on training and 0.83 on test data. The *glmnet* method using Gene Expression, Localization, Epidemiology, Cancer Type predicted Relapse SII with a Youden-index of 0.7. In the analysis of relapse of SII and SIII the focus was on the training data, as the test data only had 1 and 3 positive cases respectively, which made it difficult to interpret. The model that predicted DFS SIII with the highest C-Index score was the *rfsrc* method using Gene Expression, Localization and Epidemiology. The C-Index of the model was 0.98 on the training and 0.20 on the test data, indicating overfitting of the model. The model that predicted Relapse SIII best was the *glmnet* method using Gene Expression, Localization, Epidemiology. The accuracy of the model was 0.71 on the training and 0.54 on test data.

3.3 Comparing relapse prediction to subgroup benchmarks

The model for the prediction of relapse stage II was used to divide patients into two groups, namely relapse and no-relapse. The survival probability of the categorization by the ML model was then compared to the sub stage groupings of stage II. The Kaplan-Meyer curves in figure 1 demonstrate that the model-based subcategorization of stage II patients implies stronger separation of survival curves than the established classification from the literature. The same comparison with similar results was done for stage III patients.

4. Discussion

The *glmnet* method achieved the best results predicting DFS and survival with C-Index scores of 0.76 and 0.87 on test data respectively. These values indicate that when enough information is given, a very accurate prediction of survival and DFS can be made. Interestingly, the best model found for survival used only Gene Expression and Localization. This suggests that these two feature sets contain most of the relevant information.

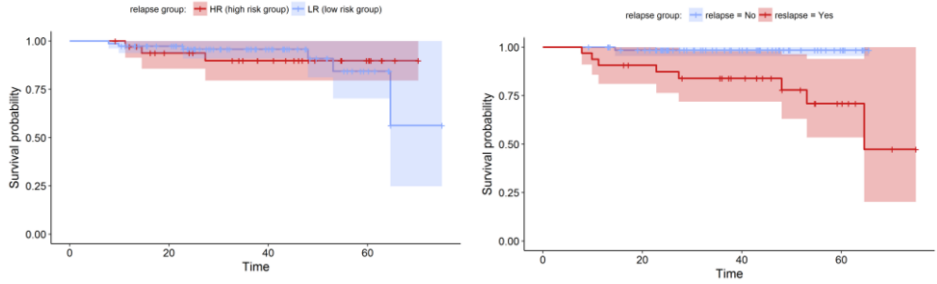

Fig. 1: survival across high-risk and low-risk groups for stage II patients identified by the literature (left): survival across relapse-no and relapse-yes groups stage II patients identified by the best fitted machine learning model (right)

Initially it was attempted to train a ML model that would predict all levels of response to radio chemotherapy as classified by the Dworak ranking, which ranks patients from 0 (no response) to 4 (very good response). Predicting all levels, the best result achieved was an accuracy of 0.40. Dimensions were then reduced to two levels by merging levels 3 and 4 into one class (good response) and the other levels (0,1,2) together into one class (no or poor response). Building a model for this reduced dimensionality a *C50 (simple tree)* model lead to a model with good *specificity* (0.85) and reasonable *sensitivity* (0.53) values. This is especially interesting as the model could be used to reliably qualify non-responders. A *glmnet* model using Gene Expression, Localization, Epidemiology, Cancer Type and Tumor Stage achieved a relapse prediction accuracy of 0.71 and was especially good at classifying whether patients would have a relapse. The fact that relapse is measured on right-censored data might have skewed these results. Yet there is a clear difference between the survival of no-relapse and relapse patients. The predictions for stage II and stage III relapse were slightly worse than relapse predictions for all stages. This might be partly because the data available for training and testing for stages II and III was significantly less than the data available when all stages were considered. However, the relapse prediction across all stages might disguise a poor fit for singular levels. The data used was unbalanced as few patients had a relapse. This could explain why the specificity for predicting a relapse was so high, while the sensitivity was significantly lower. Having established this pattern, the models trained for stage II relapse still provide a better subgroup classification than the high risk and low risk groups established in the literature [3, 4]. However, as the available data was limited, future research should study the generalizability of these models as well as improve on these results.

4.1 Clinical Relevance and potential application

Establishing good prediction models for survival and relapse promises to help physicians give a more accurate prognosis of disease progression and adjust treatment decisions. The prediction of relapse and RCT therapy response, for example, could be used to decide whether patients should undergo neoadjuvant as well as adjuvant radio-chemotherapy. The models built, despite not being perfectly accurate, are still able to predict an outcome better than chance and might be better than predictions made by physicians themselves.

4.2 Limitations and future work

The amount of available data was a limiting factor, as the feature selection process involved the data being split into three sets rather than two. A focus on subgroups of the dataset also reduced the data available for some experiments. The analysis of the most important genes is beyond the scope of this study and should be explored in further research, focusing on further narrowing down the most important genes related to colorectal cancer. Special caution should be taken evaluating the overall importance of genes as we found that importance of genes depends on the prediction problem. The program written for this study can be applied to other datasets with minor adjustments. This is a first step towards automating the ML process for future projects.

4.3 Conclusion

We created prediction models with accuracies above 0.70 using a fully automated process, which predicted relevant outcomes like chemotherapy response and survival. The main problems identified were the availability of data and choosing the right performance measure to select the best model. The outcomes that were predicted with the highest accuracies were Relapse and RCT response (Yes/No), as well as survival and disease-free survival. The models could be used in future to influence therapy decision.

Acknowledgements

The present work was performed in fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” from the Friedrich-Alexander-Universität Erlangen-Nürnberg.

This work was supported by grants from the German Research Foundation [FOR 2438 (subproject 2)], the W. Lutz Stiftung and the Forschungsstiftung Medizin and the Interdisciplinary Center for Clinical Research (IZKF) of the Clinical Center Erlangen to M.S.

References

[1] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, *Computational and structural biotechnology journal* **13** (2015), 8–17.

[2] J.A. Cruz, D.S. Wishart, Applications of Machine Learning in Cancer Prediction and Prognosis, *Cancer Informatics* **2** (2006), 59–77.

[3] S. Merkel, U. Mansmann, T. Papadopoulos, C. Wittekind, W. Hohenberger, P. Hermanek, The prognostic inhomogeneity of colorectal carcinomas Stage III, *Cancer* **92** (11) (2001), 2754–2759.

[4] S. Merkel, A. Wein, K. Guenther, T. Papadopoulos, W. Hohenberger, P. Hermanek, High-risk groups of patients with Stage II colon carcinoma, *Cancer* **92** (6) (2001), 1435–1443.

[5] Naschberger, E., A. Liebl, V.S. Schellerer, M. Schütz, N. Britzen-Laurent, P. Kölbel, ..., & M. Stürzl: Matricellular protein SPARCL1 regulates tumor microenvironment-dependent endothelial cell heterogeneity in colorectal carcinoma. *The Journal of Clinical Investigation* **126** (2016), 4187-4204

[6] Feiersinger, F., E. Nolte, S. Wach, T.T. Rau, N. Vassos, C. Geppert, A. Konrad, S. Merkel, H. Taubert, M. Stürzl & R.S. Croner: MiRNA-21 expression decreases from primary tumors to liver metastases in colorectal carcinoma. *PLoS One* **11** (2) (2016), e0148580