SENG 474: Data Mining - Formal Proposal

Group I - Team Members

Cameron Lindsay, Claire Champernowne, Ethan Getty, Anton Nikitenko, Paul Henderson, Elliot Kong

The Problem

Head and neck cancers (HNC) are the sixth leading cancer worldwide, with over 600 000 new cases per year [1]. While the overall incidence of HNC has decreased, oropharyngeal squamous cell carcinomas (OPSCC), a subset of HNC, has increased due to human papillomavirus (HPV)-mediated oncogenesis [2-4]. Fortunately, HPV-positive OPSCC have shown to respond better to treatment, resulting in significantly higher survival outcomes [5-7]. Chemotherapeutics are often used in combination with surgical intervention and/or radiotherapy in the treatment of HNSCC, but current chemotherapeutic agents are limited in their efficacy with certain tumors and carry a significant toxic burden on patients. Chemotherapeutic agents that hold less toxicity are in high demand and one such agent is Metformin, a first-line oral treatment of type 2 diabetes mellitus [8]. Metformin and related biguanides inhibit gluconeogenesis through the indirect inhibition of the mammalian target of rapamycin (mTOR), a kinase known to be deregulated in a variety of cancers, including OPSCC [9-12]. Metformin's chemopreventive activities are attributed through the inhibition of mTOR [13-14].

Prognostic models of survival are an integral resource for physicians in the clinical strategy and management of disease. They are evidence-based decision tools that have been supported by prior diagnostic and treatment information [15]. As computational power increases, prognostic models are becoming increasingly more accurate and complex, resulting in their progression to the forefront of clinical decision making. Artificial intelligence-based strategies including Bayesian networks, neural networks, and support vector machines are rapidly gaining popularity in the creation of these models [15-16].

Using the National Cancer Institute's (NIH) Surveillance, Epidemiology, and End Results Program (SEER) (https://seer.cancer.gov/) databases to train the potential algorithm, we propose the formulation of a prognostic model of OPSCC survival contrasting two primary factors: HPV status and metformin. Although currently granted access to data within the SEER databases, should issues of data quality arise, several other databases containing comparable data are available for academic use. The most prominent of these databases is the National Cancer Database (https://www.facs.org/quality-programs/cancer/ncdb). Authors are also in contact with known researchers of the subject matter that have access to smaller, but usable databases.

Goals

Following the processing and gross classification of patient data based on primary tumor site (isolate OPSCCs), HPV-status, and use of metformin, a regression-based neural network model in conjunction with (or modified in adherence to) contemporary survival analyses will contrast and compare these primary factors to predict patient survivability. The survival analysis methodology will be formulated during the literature review process, utilizing methods recommended by subject matter experts. Feature selection and model performance will be evaluated by root mean squared error (RMSE), confusion matrix, F1 score, and accuracy. These metrics were chosen for the following reasons:

Confusion matrix: Used to obtain a complete picture when assessing the performance of classification models, providing details on the number of misclassification between every two classes. Formal goals for this metric cannot be made due to its heavily dynamic nature and reliance on the datasets quality.

Accuracy: Provides overall performance of a model and is the dominant metric in determining how "good/bad" models are. In the context of classification problems, is used to categorize the input vector to the "correct" category. We aim to achieve a value of 70% (0.7) for this metric.

F1-score: Provides a detailed explanation of the accuracy, considering both the precision and recall in its computation. If our datasets contain imbalanced classes, the F1-score will provide more information about precision and recall. Formal goals for this metric cannot be made due to its heavily dynamic nature and reliance on the datasets quality.

RMSE: Used to measure the difference between the predicted value of our model and the observed value (ground truth). A large RMSE may indicate outliers. We aim to achieve a value of 30% (0.3) for this metric.

Regression model training on cancer datasets

Individual datasets separated on their classifications above will be used to train unique regression-based neural networks in an attempt to formulate a standardized model that will then be used on the complete datasets. This will be done in parallel with a decision tree model for comparison. The performance of the model will be measured by the criterion mentioned above and compared with individual datasets and the decision tree model. Throughout the training and testing process, metrics will be recorded to detect the time point that overfitting occurs.

Potential results and issues

Due to the potential of large amounts of samples in these datasets being incomplete (missing features), the performance of this combined model could be worse when compared to prior models. Pre-processing via pruning incomplete data could be a potential method to improve the performance (formal criterion to be

determined). Following pre-processing, the model will be trained, tested, and compared with the original model.

Plan

Our experiments will be conducted according to the outlined milestones of the project (Table 1). These milestones will help us determine any problems in our data collection process as well as provide us with feedback on the effectiveness of our experimental methods. The bulk of the experimentation process will be in the training of learning models following the processing of the SEER patient data. Our secondary research will involve the comparison of our findings with published works utilizing comparable models. This will help determine any correlations between our own findings and known trends in the field, confirming some assumptions we would make about our applied learning models.

The following milestones were preset by the project to be completed on the assigned dates. The team will create sub-milestones to be completed within the given timeframe to make sure goals are met on schedule by having set check-in times.

Table 1: Project Milestones

Event	**Date**
Milestone 1: Group formation	Monday June 15th
Milestone 2: Formal proposal	Wednesday June 24th
Milestone 3: Progress report	Monday July 6th
Milestone 4: Presentation	Early August (precise date TBD)
Milestone 5: Final report	Early August (precise date TBD)

By the time Milestone 3 (progress report) is completed, we will have conducted research into the background of the topic, collected and organized patient data for the learning algorithm(s), as well as outlined and implemented the learning methods we plan to use for training and testing our data. Our background research includes identifying specific methods that other research groups have implemented successfully (or unsuccessfully), as well as the consultation of external resources such as teaching assistants or online resources. We will have also determined a collection of learning models for potential implementation including neural networks, decision trees, or random forests that we will test and explore for applicational viability. Narrowing down the possible learning methods we will use for optimizing our model is vital, as it will help us determine the most accurate method(s) for our application. For example, we may find that discarding a certain learning method is necessary as no evidence has been found indicating its utility.

At the time of the presentation we will have our final report completed or nearing completion depending on formal deadlines. This means we will have fully implemented our learning model to successfully

analyze patient data based on our chosen learning method. At this stage we will have trained the algorithm to model OPSCC survivability given our initial analyzed factors. We will also have developed our performance measures for our algorithm with all the relevant data to be presented.

We foresee two aspects of our experiments where help from external field experts will be useful to guide our research. Firstly, we are considering consulting with this course's teaching assistant and professor about the applicability of the learning models we selected for the type of data that we shall use for the experiments. Although we are attempting to perform a classification problem with data that can be easily converted into a form that's easy to input into a learning model such a decision tree or neural network, the type of data that our input will represent may not be the most optimal choice for the learning model that we have decided to use. Furthermore, as part of our progress developing learning models that satisfy our goal, we will conduct an informal literature review of current prognostic models utilizing comparable machine learning-based methodological approaches. This type of consultation with external published resources will guide us, as similar conclusions will indicate that our experiments are conducted properly, and major deviations from expected results may be a sign that our experiments were conducted with misguided assumptions.

Task Breakdown

Cameron Lindsay:

1. Provide guidance based on prior knowledge in medical field
2. Register and becoming proficient in accessing data from SEER database using SEER*stat
3. Focus on a decision tree-based approach

Claire Champernowne:

1. Ensuring cleanliness of data obtained from SEER database
2. Focus on a decision tree-based approach

Ethan Getty:

1. Focus on a Neural Network-based approach of Survival analysis
2. Conduct research comparing our findings to similar published work

Anton Nikitenko:

1. Focus on a decision tree-based approach
2. Provide processing power from a Graphical Processing Unit (GPU) to train models

Paul Henderson:

1. Focus on a Neural Network-based approach of Survival analysis
2. Conduct research comparing our findings to similar published work

Elliot Kong

1. Focus on a Neural Network-based approach of Survival analysis
2. Conduct background research regarding further potential learning methods for training and testing

References

1. Siegel, R. L., Miller, K. D., & Jemal, A. Cancer statistics, 2015 CA Cancer J Clin 2015; 65: 5-29. *External Resources Pubmed/Medline (NLM) Crossref (DOI)*
2. Gillison, M. L., Chaturvedi, A. K., Anderson, W. F., & Fakhry, C. (2015). Epidemiology of human papillomavirus–positive head and neck squamous cell carcinoma. *Journal of Clinical Oncology*, *33*(29), 3235.
3. Forte, T., Niu, J., Lockwood, G. A., & Bryant, H. E. (2012). Incidence trends in head and neck cancers and human papillomavirus (HPV)-associated oropharyngeal cancer in Canada, 1992–2009. *Cancer Causes & Control*, *23*(8), 1343-1348.
4. Chaturvedi, A. K., Anderson, W. F., Lortet-Tieulent, J., Curado, M. P., Ferlay, J., Franceschi, S., ... & Gillison, M. L. (2013). Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers. *Journal of clinical oncology*, *31*(36), 4550.
5. Fakhry, C., Westra, W. H., Li, S., Cmelak, A., Ridge, J. A., Pinto, H., ... & Gillison, M. L. (2008). Improved survival of patients with human papillomavirus–positive head and neck squamous cell carcinoma in a prospective clinical trial. *Journal of the National Cancer Institute*, *100*(4), 261-269.
6. Ang, K. K., & Sturgis, E. M. (2012, April). Human papillomavirus as a marker of the natural history and response to therapy of head and neck squamous cell carcinoma. In *Seminars in radiation oncology* (Vol. 22, No. 2, pp. 128-142). WB Saunders.
7. Xu, C. C., Biron, V. L., Puttagunta, L., & Seikaly, H. (2013). HPV status and second primary tumours in oropharyngeal squamous cell carcinoma. *Journal of Otolaryngology-Head & Neck Surgery*, *42*(1), 36.
8. Rhee, M. K., Herrick, K., Ziemer, D. C., Vaccarino, V., Weintraub, W. S., Narayan, K. V., ... & Phillips, L. S. (2010). Many Americans have pre-diabetes and should be considered for metformin therapy. *Diabetes Care*, *33*(1), 49-54.
9. Pollak, M. (2010). Metformin and other biguanides in oncology: advancing the research agenda. *Cancer prevention research*, *3*(9), 1060-1065.
10. Pollak, M. (2008). Insulin and insulin-like growth factor signalling in neoplasia. *Nature Reviews Cancer*, *8*(12), 915-928.
11. Kourelis, T. V., & Siegel, R. D. (2012). Metformin and cancer: new applications for an old drug. *Medical Oncology*, *29*(2), 1314-1327.
12. Lindsay, C., Kostiuk, M., Conrad, D., O'Connell, D. A., Harris, J., Seikaly, H., & Biron, V. L. (2019). Antitumour effects of metformin and curcumin in human papillomavirus positive and negative head and neck cancer cells. *Molecular carcinogenesis*, *58*(11), 1946-1959.
13. Quinn, B. J., Kitagawa, H., Memmott, R. M., Gills, J. J., & Dennis, P. A. (2013). Repositioning metformin for cancer prevention and treatment. *Trends in Endocrinology & Metabolism*, *24*(9), 469-480.
14. Kasznicki, J., Sliwinska, A., & Drzewoski, J. (2014). Metformin in cancer prevention and therapy. *Annals of translational medicine*, *2*(6).
15. Abu-Hanna, A., & Lucas, P. J. (2001). Prognostic models in medicine. *Methods of information in medicine*, *40*(01), 1-5.

16. Lucas, P. J., & Abu-Hanna, A. (1998). *Prognostic methods in medicine* (Vol. 1998). Utrecht University: Information and Computing Sciences.