Reiterative modeling of combined transcriptomic and proteomic features refines and improves the prediction of early recurrence in squamous cell carcinoma of head and neck

Amir Salehi a,1, Lixiao Wang a,1, Philip J. Coates b, Lena Norberg Spaak a, Xiaolian Gu a, Nicola Sgaramella a, Karin Nylander a,*

a Department of Medical Biosciences/Pathology, Umeå University, Umeå, Sweden
b Research Centre for Applied Molecular Oncology, Masaryk Memorial Cancer Institute, Brno, 656 53, Czech Republic

ARTICLE INFO

Keywords:
SCCHN
Machine learning
XGBoost
Early recurrence
Multi-omics

ABSTRACT

Background: Patients with squamous cell carcinoma of the head and neck (SCCHN) have a high-risk of recurrence. We aimed to develop machine learning methods to identify transcriptomic and proteomic features that provide accurate classification models for predicting risk of early recurrence in SCCHN patients.

Methods: Clinical, genomic, transcriptomic and proteomic features distinguishing recurrence risk were examined in SCCHN patients from The Cancer Genome Atlas (TCGA). Recurrence within one year after treatment was classified as high-risk and no recurrence as low-risk.

Results: No significant differences in individual clinicopathological characteristics, mutation profiles or mRNA expression patterns were seen between the groups using conventional statistical analysis. Using the machine learning algorithm, extreme gradient boosting (XGBoost), ten proteins (RAD50, 4E-BP1, MYH11, MAP2K1, BECN1, NF2, RAB25, ERRFI1, KDR, SERPINE1) and five mRNAs (PLAUR, DKK1, AXIN2, ANG and VEGFA) made the greatest contribution to classification. These features were used to build improved models in XGBoost, achieving the best discrimination performance when combining transcriptomic and proteomic data, providing an accuracy of 0.939 and an Area Under the ROC Curve (AUC) of 0.951.

Conclusions: This study highlights machine learning to identify transcriptomic and proteomic factors that play important roles in predicting risk of recurrence in patients with SCCHN and to develop such models by iterative cycles to enhance their accuracy, thereby aiding the introduction of personalized treatment regimens.

1. Introduction

More than 550,000 people are diagnosed with squamous cell carcinoma of the head and neck (SCCHN) per year worldwide [1]. SCCHN has a high mortality rate of 40–50% [2], and patient survival has not improved significantly in the recent past [3]. Factors such as genomic complexity and field cancerization are thought to contribute to treatment resistance and recurrence in this cancer type [4–6]. In this respect, the interval to recurrence is an independent prognostic factor for patients with oral SCC [7] and prognostic models for recurrence have been developed based on clinical data [8]. Although such models have the potential to aid treatment decisions and strategies, their discrimination ability is relatively low.

With recent technological advances, high-throughput studies characterized by large multi-omics datasets including genomics, transcriptomics and proteomics have increased our knowledge of the molecular alterations related to SCCHN formation and progression and therefore have potential to identify prognostic biomarkers. Most notably, The Cancer Genome Atlas (TCGA) [5] has data on over 520 SCCHN samples using various omics platforms. Exploring the relationships of these different datasets with clinical outcomes is challenging, and artificial intelligence (AI)-based methods are a rapidly growing area in medicine and healthcare to aid understanding and interpretability of such data [9,10]. In SCCHN, support vector machines (SVM) were used to develop a model for predicting recurrence/metastasis using whole genome copy number alteration (CNA) data with an accuracy of ~80%.

* Corresponding author.
E-mail address: karin.nylander@umu.se (K. Nylander).
1 These authors contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2022.105991
Received 22 April 2022; Received in revised form 11 July 2022; Accepted 14 August 2022
Available online 18 August 2022
0010-4825/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
and combining machine-learning methods with targeted proteomics has also been used to identify a prognostic signature in oral cancer [12].

The field of AI and its applications to medical information is a rapidly expanding area, with new approaches being introduced regularly. Large scale comparison of machine learning methods indicates that gradient tree boosting algorithms outperform most methods for classification procedures [13], in which XGBoost (eXtreme Gradient Boosting) is an optimized algorithm [14]. The basic idea of gradient boosting is to transform weak learning models into strong predictive models by iteratively improving upon errors, i.e. XGBoost is a supervised machine learning system, taking many variables and assessing their ability/abilities to identify a predetermined variable, then testing and re-testing newer models based on the previous accuracy of prediction. XGBoost provides several advanced parameters for model tuning and algorithm enhancement, and improves performance by combining multiple weak learning models. In contrast to statistical methods that are designed to infer relationships between input features and outputs, machine learning methods focus on making the most accurate predictions, which usually sacrifices model interpretability. The built-in feature importance measure of XGBoost that estimates feature importance according to how valuable each feature was in constructing the model [15], provides both interpretable machine learning and high accuracy prediction modeling.

Given the availability of large genomic, transcriptomic and proteomic datasets linked to SCCHN tumor characteristics, and of different AI-based methods to analyse such data, we aimed first to identify a suitable method to predict the risk of early tumor recurrence. After an initial grid-search for appropriate methods, feature selection was performed with XGBoost and potential SCCHN prognostic biomarkers identified through the literature were systematically applied to remove non-informative or redundant features. Finally, XGBoost models were developed that accurately predict recurrence risk within a one year period.

2. Materials and methods

2.1. Patient data

Genomic and transcriptomic profiles and clinical data from 523 SCCHN tumor samples were obtained from TCGA using cBioPortal [16] (http://www.cbioportal.org). Additional clinical information was obtained via Broad GDAC Firehose (https://gdac.broadinstitute.org). Transcriptomic data were from Firehose, and reverse-phase protein array (RPPA) data were obtained from PanCanAtlas [17]. Data were downloaded from the appropriate databases in July 2020. Factors used in the various analyses are listed in Tables S1 and S2, and the methods by which they were analysed are shown in Table 1. A flow chart of the available patient datasets and their inclusion in each step of analysis is shown in Fig. 1.

2.2. Collection of prognosis-related markers

Potential prognosis-related biomarkers were acquired from two recent reviews that performed extensive literature searches of studies investigating prognostic biomarkers in SCCHN [18] and a comprehensive analysis of genomic and transcriptomic data from TCGA-HNSCC patients [19]. In addition, immune-associated biomarker genes reported in SCCHN [20] and genes associated with overall survival in oral tongue squamous cell carcinoma [21] were collected from recently published studies.

2.3. Statistical analysis

According to our criteria, patients with cancer recurrence within one year after treatment were assigned to the high-risk group, whereas patients without cancer recurrence within one year or longer were allocated as low-risk patients. Chi-square test or Mann-Whitney U test, which make no assumptions about the distribution of the population, were used for testing categorical and continuous variables, respectively. Pearson’s correlation coefficient was applied to measure the statistical relationship between continuous variables. Statistical analyses were
performed using IBM SPSS Statistics (v 26.0, IBM, Armonk, NY, USA). A p-value < 0.05 was considered significant.

2.4. Machine learning methods

We initially used a grid-search to find optimal hyperparameters for building prediction models with commonly used machine learning methods, including support vector machines, random forests and other commonly applied methods. Subsequently, the python library “Lazy Predict” (https://lazypredict.readthedocs.io/en/latest/) was used to evaluate the most suitable machine learning algorithms for prediction of SCCHN recurrence risks. Based on this analysis, XGBoost was conducted to filter redundant variables and select the most important subset of features from clinical and multi-omics data (Tables S1 and S2). XGBoost has a built-in feature importance module, which can be computed in three different ways; as ‘weight’ (the number of times a feature is used to split the data), as ‘gain’ (the average information gained across all splits where a feature was used), and as ‘cover’ (the average coverage across all splits of a feature). These features were calculated with the python machine learning library, Scikit-Learn 0.23.0 [22]. XGBoost can perform regression analysis and thus be used to predict time of recurrence. However, as most of our patients were free from recurrence at one year, we chose not to include time to event data regarding time to recurrence.

2.5. Definitions of accuracy metrics

Overall model performances were assessed via conventional definitions of sensitivity, specificity, feature impact (F) score, receiver operating characteristic (ROC) and area under the curve (AUC). The abbreviations TP (true positive), TN (true negative), FP (false positive) and FN (false negative) were used to describe correct and incorrect assignments of an unknown SCCHN patient to a risk group of recurrence. The same metrics have previously been used in epigenetic studies [23].

Sensitivity = TP/(TP + FN),
Specificity = TN/(TN + FP),
F score = TP/(TP+1/2(FP + FN)),
Accuracy = (TP + TN)/(TP + FP + TN + FN).

AUC of ROC plots was used as the main criterion for evaluating the predictive ability of the models.

3. Results

3.1. Clinical characteristics

A total of 523 SCCHN patients were included in the TCGA PanCancer Atlas database, from which information on disease-free status was available for 132 patients. Of these, 27 patients had a recurrence, and 17 of those occurred within 1 year. Of the 105 patients without recurrence, 84 with transcriptomic data were documented to remain disease free for at least 1 year (Fig. 1) and these groups were used to create a classification model to predict risk of early recurrence after treatment. To ensure that the 101 patients analysed were representative of the 523 TCGA SCCHN cohort, distributions of age, gender and TNM-stage were compared between the included and not included patients and were found to be similar (Table 2).

Kaplan-Meier survival analysis indicated that the high-risk group was associated with poor overall survival (Fig. 2), consistent with early recurrence having a poor prognosis and confirming that our patient classification scheme is biologically and clinically relevant. However, and perhaps surprisingly, there were no clinicopathological characteristics that differed significantly between patients with or without early recurrence as defined here (p > 0.05) (Table 2).

#### Table 2

<table>
<thead>
<tr>
<th>Clinical and pathologic characteristics of the study population.</th>
</tr>
</thead>
<tbody>
<tr>
<td>High risk (n = 17)</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
</tr>
<tr>
<td>Median (Range)</td>
</tr>
<tr>
<td>Tumor stage</td>
</tr>
<tr>
<td>T0</td>
</tr>
<tr>
<td>T1</td>
</tr>
<tr>
<td>T2</td>
</tr>
<tr>
<td>T3</td>
</tr>
<tr>
<td>Tx</td>
</tr>
<tr>
<td>NA</td>
</tr>
<tr>
<td><strong>Tumor margin</strong></td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Close</td>
</tr>
<tr>
<td>NA</td>
</tr>
<tr>
<td><strong>Radiation therapy</strong></td>
</tr>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>NA</td>
</tr>
<tr>
<td><strong>Neoplasm Subdivision</strong></td>
</tr>
<tr>
<td>Oral Tongue</td>
</tr>
<tr>
<td>Others</td>
</tr>
<tr>
<td><strong>HPV status</strong></td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>NA</td>
</tr>
</tbody>
</table>

NA, not available.

![Fig. 2. Kaplan-Meier curves of overall survival for high- and low-risk recurrence groups.](image)

3.2. Comparing molecular features in patients with high- and low-risk of recurrence

No statistically significant differences in genome-wide mutation burden, CNV, SNV or mRNA levels were identified between the two groups of patients (p > 0.05). Proteomic data were available only for a subset of patients (26 low-risk and 7 high-risk patients), and Mann-Whitney U tests showed that 4E-BP1 was the only protein showing a significant difference in levels between the two groups (p < 0.001) (Fig. 3A). To identify proteins with a similar expression pattern to that of 4E-BP1, The Cancer Proteome Atlas (TCPA) was also explored. The phosphorylated form of 4E-BP1 (4E-BP1_pT37T46) was the only protein that correlated with 4E-BP1 protein levels in these data (rho = 0.691, p < 0.001, Fig. 3B).
3.3. An mRNA and protein panel for classification

According to the performance of the baseline comparisons, XGBoost was one of the best algorithms for our dataset (Tables S3 and S4). XGBoost was therefore conducted to select a panel of features for group discrimination. For clinical, genomic and transcriptomic data, the training dataset included all 17 high-risk and 84 low-risk patients. Initially, all TCGA data used for statistical comparison were included for feature selection, however, no important feature was identified. Next, 29 clinical factors, 41 genomic profiles and 116 prognosis-related biomarkers were used for feature selection (Tables S1 and S2).

RPPA data for 191 proteins and their variants were available for 7 high-risk and 26 low-risk patients. Although the number of patients analysed using RPPA is small, these data indicate the potential importance of the proteins present on the array. To obtain an unbiased and compact set of features, XGBoost ranked all the features through assessment of their importance for prediction by 5-fold cross-validation.

Applying XGBoost to the selected clinical, genomic and transcriptomic data produced a prediction model with AUC 0.776, from which the mRNA levels of five genes were identified as the features that produced the largest information gain for classification. Feature importance (F-scores) of these five mRNAs (PLAUR, DKK1, AXIN2, ANG and VEGFA) are shown in Fig. 4A. Similar to the results from traditional statistical approaches, clinical and genomic features were not important for discrimination between high- and low-risk recurrence groups. The proteomic data feature selection model (AUC = 0.870) ranked ten proteins as important (RAD50, MYH11, 4E-BP1, MAP2K1, BECN1, NF2, RAB25, ERRFI1, KDR, SERPINE1), as shown in Fig. 4B.

3.4. Modeling recurrence prediction

After selecting features, we built three XGBoost models (5-mRNA, 10-protein and mRNA&Protein). The performance of each model was evaluated by leave-one-out cross-validations (LOOCV), where patients used for prediction are not the same as had been used to build the model, and each patient is considered a validation set and the remaining...
patients a training set. LOOCV is an unbiased method for estimating model performance and is widely used to evaluate model building with small training datasets. As shown in Fig. 5 and Table 3, the 10-protein-based model achieved better prediction performance (AUC = 0.907) than the 5-mRNA-based model (AUC = 0.72), and the best performance (AUC = 0.951) was achieved by the mRNA&Protein model that combines the protein and mRNA levels into a single classification model.

4. Discussion

TCGA has generated a comprehensive molecular characterization of SCCHN and has increased our knowledge of many aspects of these tumors. Recent systematic searches have identified blood- and tissue-based biomarkers associated with specific clinical outcomes for SCCHN patients [18]. However, due to the high complexity and heterogeneous nature of SCCHN at the molecular level, no predictive biomarker has proven to be effective and reached clinical practice. With advancements in analyzing high-dimensional, complex and unstructured data, AI-based methods are playing increasingly critical roles in clinical medicine. Here, by combining the power of machine learning methods and statistical techniques, we developed a robust model for accurately predicting SCCHN recurrence within a 1-year period.

Conventional statistical analysis did not identify significant differences in mRNA levels between high and low-risk patients, despite the obvious clinical differences between these two groups. However, by focusing on potential prognostic markers in SCCHN identified by others, the XGBoost model-based feature selection approach indicated that five transcripts contributed most to discrimination of recurrence, with an AUC of 0.776. Thus, even though significant differences in mRNA levels were not demonstrated by statistical methods based on q-value, interactions or interrelationships of these mRNAs are related to recurrence.

Among the five mRNAs identified, PLAU (also called uPAR), AXIN2, ANG and VEGFA are related to epithelial-mesenchymal transition (EMT) and metastasis [18], whereas DKK1 inhibits the Wnt signaling pathway that plays a key role in carcinogenesis and progression of various cancer types [24]. Previous immunohistochemical studies have indicated that PLAU [25], AXIN2 [26] and ANG [27] are upregulated in SCCHN patients with poor prognosis. In our study, the substantial improvement in performance of the classification model by including all five mRNAs as a single panel reinforces their contributions to cancer recurrence.

Although a variety of methods can be used to identify biomarkers [28,29], we selected proteins based on previous data, including some with known roles in SCCHN carcinogenesis. Of these, XGBoost indicated that RAD50 had the highest information gain, even though there was no statistically significant difference between the two groups of patients. RAD50 is crucial for sensing and repairing DNA damage [28] but has not shown any association with increase risk of SCCHN [29]. Our modeling results demonstrate that RAD50 protein levels associate with the risk of early recurrence in SCCHN patients, in keeping with previous findings that DNA repair enzyme levels inversely correlate with response to therapy in SCCHN [30].

4E-BP1 was identified as an important discriminant by both conventional statistical analysis and by machine learning (XGBoost) feature selection. 4E-BP1 is regulated by mTORC1, which controls protein synthesis according to oxygen and nutrient availability, where 4E-BP1 binds to elf4E to repress mRNA translation and inhibit protein synthesis. When phosphorylated by hormonal or growth factor signaling pathways, 4E-BP1 dissociates from elf4E and activates mRNA translation. Thus, high levels of the inactive phosphorylated form, 4E-BP1_pT37T46, are found in various human carcinomas [31], and high levels associate with SCCHN recurrence [32]. Conversely, high levels of non-phosphorylated 4E-BP1 are expected to be helpful to tumor cells by enhancing or repressing translation of specific mRNAs in the hypoxic and low nutrient conditions often seen in tumors, thereby enabling tumor cell adaptation to a hostile microenvironment [31]. In our study, patients with higher levels of 4E-BP1 showed lower risk of recurrence, consistent with the recent findings that 4E-BP1 gene loss is common in SCCHN and is associated with poor prognosis [33]. In addition, the correlation that we found between 4E-BP1 and 4E-BP1_pT37T46 indicates that this protein is constitutively phosphorylated in SCCHN, which is in keeping with the aberrant activation of mTOR in most of these tumors as a direct cause of constitutive 4E-BP1 phosphorylation (Citation [34]).

Of the other protein markers of recurrence we identified, MYH11 is a myosin molecule with an unknown role in tumors, but is reduced by DNA methylation in poor prognosis gastric cancers [34], MAP2K1 (also called ERK1) has known roles in transducing extracellular growth and mitogenic signals and is often aberrantly regulated in human cancer [35]. BECN1 (Beclin 1) is a key regulator of autophagy, a process that can be either tumor-promoting or –suppressing at different stages of oncogenesis and dependent on microenvironmental conditions [36]. NF2 (Nerofibromatosis type 2) is a well known tumor suppressor protein that regulates the anti-proliferative Hippo signaling pathway [37]. RAB25 is an epithelial-specific Rab GTPase that regulates integrin trafficking and cancer cell invasion, providing either oncogenic or tumor-suppressing activities depending on the circumstance [38]. ERRFI1 (ERBB receptor feedback inhibitor 1, also called MIG-6) negatively regulates EGFR-family receptors by inhibiting their autophosphorylation, but also acts by upregulating glucose uptake and glycolysis to aid tumor cell growth [39]. KDR (also called VEGFR2) is a receptor for VEGFA, VEGFC and VEGFD (note that VEGFA mRNA was also identified in our study) and is expressed on endothelial cells to promote angiogenesis and on immune cells to influence the immunosuppressive

| Table 3 Comparison of different models for classification. |
|-----------------|---|---|---|---|
| Model           | Sensitivity | Specificity | F-score | Accuracy | AUC  |
| mRNA&Protein    | 0.857       | 0.962       | 0.857   | 0.939    | 0.951 |
| 10-protein      | 0.714       | 0.962       | 0.769   | 0.909    | 0.907 |
| 5-mRNA          | 0.286       | 0.923       | 0.363   | 0.788    | 0.72  |

Fig. 5. ROC curve comparison of the prediction models from XGBoost. Blue curve: model built with mRNA and protein features, AUC = 0.951. Red curve: model built with 10 protein markers, AUC = 0.907. Green curve: model built with 5 mRNA markers, AUC = 0.72.
microenvironment [40]. SERPINE1 (also called PAI, plasmogen activator inhibitor 1) is a serine protease inhibitor that promotes squamous cell migration during wound repair, aids cancer cell survival, has pro-angiogenic effects and influences cancer-associated immunosuppression, and is a robust and reliable prognostic indicator for many cancer types [41]. Thus, these protein biomarkers of early recurrence reflect the many and disparate alterations to tumor cells and the tumor microenvironment that occur during cancer progression.

A strength of this study is that the one-year period was selected to classify high- and low-risk recurrence rates, as our intention was to identify transcriptomic and proteomic features to develop a model to predict early recurrence and ascertain the molecular variations associated with this event. The classification was designed to avoid pitfalls of previous studies that classified relapse and non-relapse without considering disease-free time period. Whilst the sample size was reduced by the restrictions imposed to achieve these rigorous criteria, the cohort was representative of all SCCHN samples available in the TCGA database regarding clinical information. This collection of recurrence data from TCGA is also the most comprehensive dataset available for SCCHN regarding clinical, genomic, transcriptomic and proteomic information. As such, the dataset comprises several hundred features, and XGBoost is a powerful method to reduce the dimensions of a dataset and achieve better prediction performance. The performance and interpretability of XGBoost models therefore facilitate predictive analysis and provide insights into the key biological processes involved.

In conclusion, using machine learning methods, we identified a panel of ten proteins and five mRNAs as an optimized set of features for maximum prediction accuracy. By evaluating performance in different combinations of features from the panel using XGBoost, and by assuring that the group of patients were representative of the whole TCGA cohort, we demonstrated that the best performance is obtained when both mRNA and protein features are integrated. To ensure reliable and unbiased estimation of model performance, we also used the leave-one-out cross-validation approach. The results improve our knowledge of the biological alterations associated with SCCHN recurrence and may have clinical value for prognostic determination and patient follow-up.

Authors’ contributions

AS, LW, PJC and KN designed the study. AS, LW picked out and analysed data. LNS, NS and XG participated in data analysis. AS, LW, PJC and KN wrote the manuscript. AS, LW, LNS, XG, NS, PJC and KN edited and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

All data generated or analysed during this study are included in this published article and its supplementary information files. Source codes and training data are freely available via GitHub: https://github.com/medbiomumu/SCCHN.

Funding information

This study was supported by Lion’s Cancer Research Foundation, Umeå University; The Swedish Cancer Society (contract number 20 0754 PJF 01H); Umeå University; Region Västerbotten; Ministry of Health Czech Republic, conceptual development of research organization (MMCI, 00209805).

Declaration of competing interest

The authors declare that they have no conflict of interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105991.

References


