Larynx and hypopharynx cancer

Educated choices in treatment and rehabilitation

Petersen, J.F.

Citation for published version (APA):
Development and external validation of a risk prediction model to predict 5-year overall survival in advanced larynx cancer
Development and external validation of a risk prediction model to predict 5-year overall survival in advanced larynx cancer


ABSTRACT

Objective: TNM classification inadequately estimates patient specific overall survival (OS). We aimed to improve this by developing a risk prediction model for patients with advanced larynx cancer.

Study design: Cohort study

Methods: We developed a risk prediction model to estimate the 5-year OS rate based on a cohort of 3,442 patients with T3T4N0N+M0 larynx cancer. The model was internally validated using bootstrapping samples and externally validated on patient data from five external centers (n=770). Main outcome was performance of the model as tested by discrimination, calibration and the ability to distinguish risk groups based on tertiles from the derivation dataset. The model performance was compared to a model based on T- and N classification only.

Results: We included age, gender, T and N classification, and subsite as prognostic variables in the standard model. After external validation the standard model had a significantly better fit than a model based on T- and N classification alone (C statistic 0.59 vs. 0.55, p < 0.001). The model was able to distinguish well between three risk groups based on tertiles of the risk score. Adding treatment modality to the model did not decrease the predictive power. As a post-hoc analysis, we tested the added value of comorbidity as scored by American Society of Anesthesiologists score in a subsample, which increased the C statistic to 0.68.

Conclusion: A risk prediction model for patients with advanced larynx cancer, consisting of readily available clinical variables, gives more accurate estimations of the estimated 5-year survival rate when compared to a model based on T and N classification alone.

INTRODUCTION

Larynx cancer is among the most frequently diagnosed head and neck squamous cell cancers (SCC), and approximately 40% of patients present with advanced disease. The 5-year overall survival (OS) of the advanced (T3T4) tumors varies between 34% and 49%, depending on patient-related factors, tumor-related factors and treatment. Historically, patients with advanced larynx cancer were treated with a total laryngectomy (TL) with adjuvant radiotherapy (RT). In 1991, the randomized controlled VA trial demonstrated equal OS for organ preservation (induction chemotherapy [CT] followed by chemoradiotherapy [CRT]) compared to TL plus adjuvant RT. In 2003, the results of the Radiation Therapy Oncology Group (RTOG) 91-11 study confirmed the value of CT added to RT; however, large T4N0 larynx cancer patients were excluded. Furthermore, in a later publication on the VA data, OS for T4N0 patients was significantly higher after TL. Recently, several other retrospective studies have reported significantly higher OS rates for TL when compared to organ preservation protocols.

Adequate information regarding the prognosis is crucial in communicating with patients and in clinical decision-making. The mixed results regarding the best treatment for advanced larynx cancer have made the decision process, however, a complex task. Currently, the TNM classification is often used when talking about the estimated prognosis of patients. Although the TNM classification effectively prognosticates at a population level, it works less well on the individual level. Furthermore, the influence of variables such as age and subsite on OS is difficult to assess in the individual patient. Several studies have demonstrated that OS predictions based on clinical prediction models (CPM) are superior to those made by experienced clinicians. The availability of a quantitative prediction model may therefore enhance the quality of the decisional process.

In this study we aimed to develop a CPM to aid decision making in advanced larynx cancer.
Due to changes in treatment trends over time if necessary. The predictors included in the model were chosen based on current knowledge, availability and biological plausibility, and included age (using a restricted cubic spline), gender, subsite within the larynx (International Classification of Diseases for Oncology, Third Revision), T classification and N classification.

**Model performance**

We assessed model performance using discrimination and calibration. Discrimination is the ability of a prediction model to distinguish between patients who experience an event from those who do not and can be measured by means of the C statistic. The C statistic can range from 0.5, which means equal to chance, to 1.0, which means a perfect model. In a Cox proportional hazard model, a C statistic of 0.60 implies that at any point in time, a random patient with an event has a higher risk score than a random patient without an event 60% of the time.

Calibration relates to the agreement between estimated and observed probabilities and is depicted in a calibration plot. In a perfect calibration plot the lines of the estimated and observed probabilities would follow a 45° line, which implies that the predicted probability is identical to the observed probability.

Internal validation was performed by taking 200 bootstrapping samples. Based on the results of the bootstrap validation, we applied uniform shrinkage to adjust the coefficients. We then performed external validation of the shrunken model and calculated the C statistic and calibration curves.

As a third measure of model performance we divided the validation data into 3 risk categories based on tertiles derived from the derivation data. We then plotted the observed Kaplan-Meier (KM) curve of the validation data over the expected KM curve of the derivation set based on the predicted risks, to visually inspect the agreement between observed and expected survival in each of the risk groups.

**RESULTS**

**Derivation and validation dataset**

The derivation dataset consisted of 3,442 patients. Mean age was 64 years, the majority of patients were male (79%) and the 5-year OS rates were 44% for RT, 45% for CRT and 49% for care. We hypothesized that the model would give more accurate predictions on OS than the TNM classification alone gives us now. Because of the absence of decisive evidence from randomized controlled trials on the best treatment choice for advanced larynx cancer, a secondary objective of this large observational study was to estimate the effect of treatment on expected survival.
for TL. All included variables (age, gender, subsite T- and N classification and treatment) had a significant effect on OS ($p < 0.001$ for all variables except gender: $p < 0.03$).

Patient characteristics from the derivation and validation dataset can be found in Table 1. Patients in the derivation dataset were significantly older than the validation dataset ($p < 0.001$) and had less male patients (79% vs. 85%). In the derivation data, more tumors were located in the supraglottic, and more patients were treated with primary RT (58% vs. 37%) and less with CRT (8% vs. 28%) or primary TL (34% vs. 40%). Furthermore, there were significant differences in T and N classification ($p < 0.001$) and 5-yr OS rates ($p < 0.001$).

| Table 1. Patient characteristics from the derivation and validation datasets. |
|----------------------------------|----------------|----------------|----------------|----------------|
| **Derivation**                  | **Pooled**     | **Leuven, Belgium** | **NCR Ireland** | **Baltimore, US** | **Atlanta, US** | **Lund, Sweden** |
| **dataset**                     | **validation** | **dataset**     | **datasets**    | **validation**   | **datasets**    | **datasets**    |
| Age Mean (range)                | 64 (28-100)    | 62- (68-92)     | 64 (59-89)      | 62 (35-85)       | 60 (38-82)      | 59 (63-83)      |
| Gender                          | 2705 (78.6)    | 652 (85)        | 92 (93)         | 339 (86.9)       | 72 (79%)        | 64 (74%)        |
| **TN classification**           |               |                 |                 |                 |                 |                 |
| T3N0                            | 1237 (35.9)    | 282 (36.6)      | 34 (34)         | 159 (40.8)       | 28 (30.7)       | 25 (28.1)       | 36 (36)         |
| T3N+                            | 681 (19.8)     | 145 (18.8)      | 17 (17)         | 72 (18.5)        | 21 (23.1)       | 21 (23.6)       | 14 (14)         |
| T4N0                            | 1687 (25.8)    | 174 (22.6)      | 28 (28)         | 73 (18.7)        | 18 (20.2)       | 18 (20.2)       | 38 (38)         |
| T4N+                            | 1637 (18.5)    | 169 (21.9)      | 21 (21)         | 86 (22.2)        | 25 (27.5)       | 25 (28.1)       | 12 (12)         |
| **Sublocation**                 |               |                 |                 |                 |                 |                 |                 |
| Glottis                         | 1074 (31.2)    | 335 (43.5)      | 55 (55)         | 157 (40.3)       | 43 (47.3)       | 37 (41.6)       | 43 (43)         |
| Supraglottis                    | 2172 (63.1)    | 313 (40.6)      | 38 (38)         | 147 (37.7)       | 45 (49.5)       | 36 (40.4)       | 47 (47)         |
| Subglottic                      | 88 (2.6)       | 26 (3.4)        | 3 (3)           | 16 (4.1)         | 2 (2.2)         | 2 (2.2)         | 3 (3)           |
| Larynx NNO                      | 108 (3.1)      | 96 (12.5)       | 4 (4)           | 70 (17.9)        | 1 (1)           | 14 (15.7)       | 7 (7)           |
| **Treatment**                   |               |                 |                 |                 |                 |                 |                 |
| TL                              | 1168 (33.9)    | 311 (40.4)      | 54 (54)         | 120 (30.8)       | 55 (60.4)       | 40 (44.9)       | 42 (42)         |
| RT                              | 2009 (56.4)    | 281 (36.5)      | 15 (15)         | 164 (42.3)       | 1 (1)           | 9 (10)          | 57 (57)         |
| CRT                             | 265 (7.7)      | 213 (27.7)      | 31 (31)         | 106 (27.2)       | 35 (38.5)       | 40 (44.9)       | 1 (1)           |
| Total                           | 3442           | 770             | 100             | 390             | 91              | 89              | 100             |

* After transformation to continuous variable using midpoint of given age category, median = 63
Abbreviations: Larynx NNO = Larynx not otherwise specified.
Values in parentheses are percentages.

**Model performance - Internal validation**

Our main objective was to compare the discriminative power of a multivariable prediction model with a model based on T classification and N classification alone. As a second objective, we evaluated the effect of treatment on OS, for which we added treatment modality as a prognostic variable in a third model containing the same variables as the prediction model. First, internal validation was performed taking bootstrapping samples ($n = 200$). This demonstrated that the prediction model including age, gender, T classification, N classification and subsite as predictors had significantly better discrimination (C statistic 0.65) than the model based on T- and N classification alone (C statistic 0.57) (likelihood ratio test $p < 0.001$).

**Model performance - External validation**

After external validation on the combined validation dataset ($n = 770$), discrimination proved to be significantly better for the full model (C statistic 0.59, 9% better) compared to the model based on T-and N classification alone (C statistic 0.55) (likelihood ratio test $p < 0.001$). Calibration of the two models is depicted in Figure 1A and B, which show a slight degree of miscalibration in both models as they do not exactly follow the 45° line. As a third measure of strength, the observed KM curves of the validation sets were plotted over the KM curves estimated from the derivation dataset for the two models (Fig. 2) to test whether a distinction can be made between high-, medium- and low-risk patients. The plots show that the models are able to distinguish between the three different risk categories, although OS in the medium and low risk groups of the validation set was lower compared to these risk groups in the derivation set.

**Influence of treatment modality**

Treatment modality was significantly related to OS in the validation database ($p < 0.0001$). The hazard ratio for death adjusted for age, gender, subsite, T classification and N
DISCUSSION

The results of our study confirm our hypothesis that a validated multivariable risk prediction model gives more accurate OS predictions for advanced larynx cancer compared to a model based on T and N classification alone. According to estimated and observed KM curves, the model distinguishes adequately between the three risk categories. Yet, with a C statistic of 0.59, the predictive accuracy leaves room for improvement in the context of clinical decision making for individual patients.

As a secondary objective, we aimed to investigate the effect of treatment on expected OS. Estimating the effect of treatment modality in an observational study is troublesome, since this incorporates a bias by indication. However, because a new, large, randomized controlled trial comparing TL with organ preservation strategies may never be performed, we investigated the influence of treatment modality when accounting for the other prognostic variables included in the prediction model. This analysis suggested that survival after TL is better than after CRT or RT, as was suggested by the results published by Timmermans et al. 4

As also was reported by Timmermans et al., the derivation data contained more supraglottic tumors than the validation data. Interestingly they demonstrated how this distribution was reversed in the T1T2 tumors, in which they found more glottic (78.6%) than supraglottic tumors (19.9%). 4 The RTOG 91-11 study, with mainly advanced tumors, found a similar rate of supraglottic tumors (69%). 6, 21

In recent years, several risk-prediction models have been published. In 2001, Baatenburg de Jong et al. developed a risk-prediction model for T1-T4 SCC occurring in all subsites of the head and neck except the esophagus. 22 The model was based on 1,396 patients diagnosed between 1981 and 1998, and included the prognostic predictors age, gender, tumor site, prior tumor and TNM classification. In 2013, the model was updated, and the Adult Comorbidity Evaluation-27 was added as a prognostic variable and external validation was performed. After external validation, the model showed a good C statistic of 0.69, but the validation dataset did not include hypopharynx and nasopharynx cancer. 23 In their model, the impact of severe comorbidity appeared comparable to the impact of a T4 tumor or N3 neck on OS. We were not able to include comorbidity in our original model, which might explain why our model was less accurate. The exploratory post-hoc analysis that included ASA score as an indicator of comorbidity improved the discriminative ability.

Another risk-prediction model has been developed by Egelmeer et al., who developed classification was 1.56 for RT compared to TL (p < 0.001), and 0.95 for CRT compared to TL (p = 0.71). With treatment modality as a prognostic variable added to the prediction model, the C statistic was 0.60.

Exploratory analysis

Although the prediction model was able to distinguish well between the three risk groups and performed better compared to a model based on TNM classification alone, the C statistic was still relatively low. We hypothesized that adding comorbidity as a prognostic variable might further improve model performance. However, this variable was not recorded in our derivation database since it was retrieved from a national cancer registry. We therefore performed an exploratory post-hoc analysis on a subset of the derivation dataset including 181 patients with T3T4N0+M0 SCC of the larynx, diagnosed and treated with RT, CRT or TL in the Netherlands Cancer Institute between 1999-2008, 4 for which we were able to collect American Society of Anesthesiologists (ASA) scores as a substitute measure for comorbidity. The majority of the external centers had not systematically recorded ASA scores in the patient files, thus we were unable to perform external validation on this model. After shrinkage by internal validation the C statistic was 0.68.

![Figure 2](image)

Figure 2: Kaplan Meier curve of the expected overall survival as estimated by the derivation set (dashed line) and the observed overall survival as seen in the validation set, divided in three risk groups. Green = low risk, blue = intermediate risk, red = high risk on death.
In survival predictions, comorbidity scores can be of great value. However, in our cohort comorbidity scores were missing. ASA score was available, however, for a subgroup of the derivation dataset. In this model, the authors included age, Eastern Cooperative Oncology Group (ECOG) performance status, N classification and treatment modality, but excluded variables such as T classification, subsite and smoking status using a stepwise selection procedure. Such a data-driven approach for variable selection results in a model that might not be accurate when used for new patients. External validation was not performed, and the authors note that the model needs external validation first and might not be generalizable.

In the literature, several different patient-specific and tumor-specific factors have been investigated as prognostic variables for head and neck cancer, indicating that factors such as albumin (<4 g/dL), alcohol intake, insurance status, race, tumor volume, tumor hypoxia, and several different biomarkers can have a prognostic influence on overall survival. In order to help distinguish the actual predictors for OS and creating a more accurate RPM, a large prospective cohort should be kept in which multiple parameters are collected or this data could be extracted from electronic patient files. Currently, in the Netherlands a national prospective audit is being conducted which in the future could be used to further improve our model.

Next to OS, another frequently used endpoint in clinical studies is larynx preservation. Predicting which patients benefit from organ preservation strategies and which do not could be of great value for avoiding unnecessary toxic treatment with added morbidity after salvage surgery. A well-known model to predict this is the TALK score: a prognostic model developed to facilitate the treatment decision making in larynx preservation. TALK is an acronym for T status, Albumin, Alcohol (or liquor) use and Karnofsky Performance score, which were the predictors used. In an external validation on the VA larynx cancer study dataset, a C statistic of 0.57 was obtained for predicting larynx preservation. The TALK score however does not indicate which patients suffer from a non-functioning larynx after organ preservation, such as those who have a tracheotomy or nasogastric feeding tube in situ. In our derivation cohort, larynx preservation was scored as not having had a laryngectomy after organ preservation. However, information regarding a tracheotomy or feeding tube was missing, due to the fact that it was based on a national cancer registry cohort. We therefore chose not to predict larynx preservation based on these data.

In survival predictions, comorbidity scores can be of great value. However, in our cohort comorbidity scores were missing. ASA score was available, however, for a subgroup of the derivation dataset. In the ASA score the burden of comorbidity is incorporated, thus it could potentially serve as a proxy for an actual comorbidity scale. In 2015, Young et al. compared the ASA score with the ECOG/World Health Organization performance scale as a measure of functional status in a predictive model and demonstrated equal performance in predicting length of stay after cancer surgery. In our exploratory post hoc analysis, adding ASA score as a prognostic variable increased our C statistic to 0.68. We recommend that future studies determine which comorbidity scale might be of most value for prediction of survival outcomes in head and neck cancer, and assess the added value of this scale in a multivariable prediction model.

There are certain limitations to our study. In multivariable prediction modeling, a generally accepted rule of thumb is that a minimum of \( m/10 \) predictors should be used in a model, where \( m \) is the number of uncensored event times (e.g. death). With 2,180 uncensored events times in our cohort, we could have included many more predictors without risking overfitting. However, our choice of predictors was limited to those available in the population-based database. Because the database was limited to those available in the population-based database. Because the database was limited to those available in the population-based database. Because the database was limited to those available in the population-based database. Because the database was limited to those available in the population-based database. Because the database was limited to those available in the population-based database. Because the database was limited to those available in the population-based database. Because the database was limited to those available in the population-based database.

CONCLUSION

We have developed a ready-to-use prediction model based on a large systematically coded database on advanced-stage larynx cancer. The model gives significantly more accurate predictions on OS than compared to a model based on T and N classification alone. All of the variables included in the model are readily available in clinical practice. Although it should not be used as a replacement for clinical reasoning, it may aid the decision-making process for patients with advanced larynx cancer.
REFERENCES


