A mixture of hidden Markov models to predict the lymphatic spread in head and neck cancer depending on primary tumor location

Roman Ludwig[1,2], Julian Brönnimann[1,2], Yoel Perez Haas[1,2], Esmée Lauren Looman[1,2], Sergi Benavente[11], Adrian Schubert[3,4,7], Dorothea Barbatei[8], Laurence Bauwens[8], Jean-Marc Hoffmann[2], Sandrine Werlen[4,5], Olgun Elicin[3], Matthias Dettmer[6,10], Philippe Zrounba[9], Bertrand Poumayou[2], Panagiotis Balermpas[2], Vincent Grégoire[8], Roland Giger[4,5], and Jan Unkelbach[1,2]

[1]Department of Physics, University of Zurich, Zurich, Switzerland
[2]Department of Radiation Oncology, University Hospital Zurich, Zurich, Switzerland
[3]Department of Radiation Oncology, Bern University Hospital, Bern, Switzerland
[4]Department of ENT, Head & Neck Surgery, Bern University Hospital, Bern, Switzerland
[5]Head and Neck Anticancer Center, Bern University Hospital, Bern, Switzerland
[6]Institute of Tissue Medicine and Pathology, Bern University Hospital, Bern, Switzerland
[7]Department of ENT, Head & Neck Surgery, Réseau Hospitalier Neuchâtelois, Neuchâtelois, Switzerland
[8]Department of Radiation Oncology, Centre Léon Bérard, Lyon, France
[9]Department of Head and Neck Surgery, Centre Léon Bérard, Lyon, France
[10]Institute of Pathology, Klinikum Stuttgart, Stuttgart, Germany
[11]Departement of Radiation Oncology, Hospital Vall d'Hebron, Barcelona, Spain

Abstract We previously developed a mechanistic hidden Markov model (HMM) to predict the lymphatic tumor progression in oropharyngeal squamous cell carcinomas (OPSCCs). To extend the model to other tumor subsites, defined by ICD-10 codes, in the head and neck, we employ a mixture of these HMMs and learn the cluster assignments and model parameters in an iterative, EM-like algorithm from multi-centric data. The mixture model manages to group anatomically close subsites and correctly infers the clusters' parameters. Using this mixture model allows the prediction of individual risks of occult nodal disease, given a diagnosis that includes tumor subsite.

1 Introduction

Head and neck squamous cell carcinomas (HNSCC) frequently spread through the lymphatic system [1, 2]. Current diagnostic imaging modalities are unable to detect microscopic lymph node metastases [3, 4]. To avoid nodal recurrences [5], large volumes in the neck are irradiated electively, which are at risk of harbouring occult disease. Guidelines about which lymph node levels (LNLs) to irradiate [6] are currently not based on a patient's individual risk, but only on the overall prevalence of nodal disease as reported in the literature [1, 2].
To personalize this prediction of the risk for occult disease, given a patient's individual diagnosis, we published

1. large, multi-centric data that reports per patient which LNLs where clinically and/or pathologically involved [7, 8].

And, building on this work,

2. an interpretable hidden Markov model (HMM), trained with this data, to predict the risk for occult nodal disease [9], given an individual patient's diagnosis.

Such a personalized risk prediction may allow clinicians to safely reduce the elective clinical target volume (CTV-N) and thus reduce side-effects that degrade the patient's quality of life [10].
Here, we extend the previous work by incorporating the primary tumor location (specified as ICD-10 code) into the model of lymphatic tumour progression, focusing on tumours in the oropharynx and the oral cavity. HNSCC patients with primary tumors at different subsites show different patterns of lymphatic spread [1, 2]. So far, this could be handled by training different models for broader categories of tumour locations, e.g. oropharynx and oral cavity tumours. However, this approach does not describe differences in lymphatic spread between different subsites within the oropharynx and oral cavity. To address this issue, we present an approach using mixtures of HMMs. The intuition is that the lymphatic spread of a tumor that lies anatomically at the boarder of oropharynx and oral cavity (e.g. tumours in the palate) may be described by a mixture of different models.

2 Materials and Methods

Each LNL $v \in \{1, 2, \ldots, V\}$ considered in our model is represented by a binary random variable (RV) X_v representing the true state of that level (0 for "healthy" and 1 for "involved"). A patient's state of lymph node involvement can be represented in a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_V)$. When a patient is diagnosed with HNSCC, we only observe the clinical lymph node involvement based on imaging, which we denote as another binary random variable Y_v. To compute the personalized risk of occult disease $\mathbf{X}$, given a diagnosis $\mathbf{Y}$, we apply Bayes' law:

$$P(\mathbf{X} \mid \mathbf{Y}) = \frac{P(\mathbf{Y} \mid \mathbf{X})\, P(\mathbf{X})}{\sum_{\mathbf{X}^\star} P(\mathbf{Y} \mid \mathbf{X}^\star)\, P(\mathbf{X}^\star)} \tag{1}$$

In the above equation, the term $P(\mathbf{Y} \mid \mathbf{X})$ is given by the sensitivity and specificity of the diagnostic procedure. The term

$P(\mathbf{X})$ represents the prior probability of involvement, which depends on the probability of the tumour to spread through the lymphatic system. The main task of the HMM is to model $P(\mathbf{X})$ and the main contribution of this paper is to incorporate the primary tutor subset into the model of $P(\mathbf{X})$.

2.1 Hidden Markov Model for Lymphatic Progression

A patient's state of lymph node involvement $\mathbf{X}[t]$ evolves over discrete time steps t. Let us enumerate all 2^V possible states, representing all combinations of LNLs. In this paper, we consider ipsilateral LNLs I, II, III and IV, which amounts to 16 possible states. The HMM is specified by *transition matrix* $\mathbf{A}$:

$$\mathbf{A} = (A_{ij}) = P\left(\mathbf{X}[t+1] = \boldsymbol{\xi}_j \mid \mathbf{X}[t] = \boldsymbol{\xi}_i\right) \tag{2}$$

which contains the conditional probabilities that a state $\mathbf{X}[t] = \boldsymbol{\xi}_i$ transitions to $\mathbf{X}[t+1] = \boldsymbol{\xi}_j$ over one time step. The transition matrix is specified and parameterised via the graphical model shown in figure 1. The red arcs in the graph of figure 1 are associated the probability that the primary tumor spreads directly to a LNL (parameters b_v). The blue arcs describe the spread from an upstream LNL – given it is already metastatic – to a downstream level (parameters $t_{v-1\to v}$).
Now, let $\boldsymbol{\pi}$ be the *starting distribution*

$$\boldsymbol{\pi} = (\pi_i) = P(\mathbf{X}[0] = \boldsymbol{\xi}_i) \tag{3}$$

denoting the probability to start in state $\boldsymbol{\xi}_i$ at time step 0. Assuming that every patient started their disease with all LNLs being healthy, we set π_i to zero for all states except the completly healthy state $\boldsymbol{\xi} = (0,0,0,0)$, which has probability 1.
Using the quantities introduced so far, the probability $P(\mathbf{X}[t] = \boldsymbol{\xi}_i)$ to be in state i in time step t can now be conveniently expressed as a matrix product:

$$P(\mathbf{X}[t] = \boldsymbol{\xi}_i) = (\boldsymbol{\pi} \cdot \mathbf{A}^t)_i \tag{4}$$

This evolution implicitly marginalizes over all possible paths to arrive at state $\boldsymbol{\xi}_i$ after t time-steps. Additionally, we must marginalize over the unknown time of diagnosis using a time-prior $P(t)$. This finally defines the probability distribution over all states of lymph node involvement used in equation 1.

$$P(\mathbf{X} = \boldsymbol{\xi}_i \mid \boldsymbol{\theta}) = \sum_{t=0}^{t_{\max}} P_T(t) (\boldsymbol{\pi} \cdot \mathbf{A}^t)_i \tag{5}$$

where $\boldsymbol{\theta} = \{b_v, t_{r\to v}\}$ denotes the set of all model parameters (7 in our case). Fortunately, the exact length and shape of this distribution on its own has little impact as previously shown. We set $t_{\max} = 10$ and $P_{\text{early}}(t)$ to a binomial distribution with parameter 0.3. Further details on the HMM can be found in …

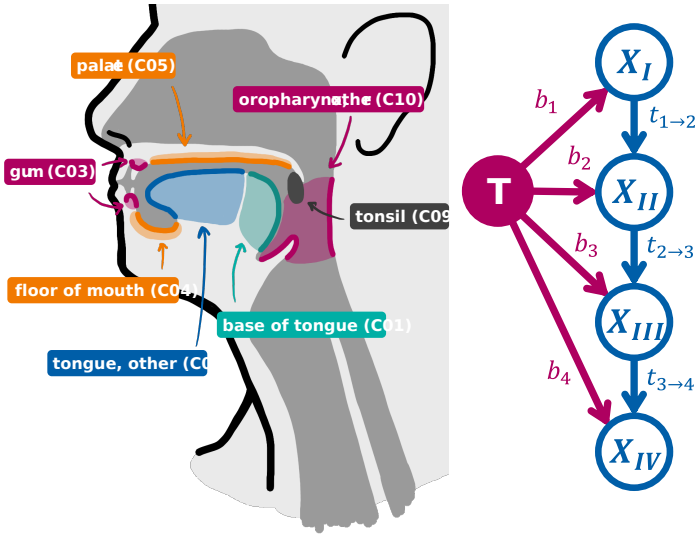

Figure 1: On the left: Rough anatomical sketch of the tumor subsites and correspsonding ICD-10 codes that are present in the used data. The subsite "other parts of mouth" (C06) was not drawn. On the right: Parametrized graph representation of the lymphatic network considering four LNLs. Blue nodes represent the hidden RVs, while the red one is the tumor. Arcs represent a conditional probability parametrized with the quantity noted next to it

2.2 Mixture of HMMs

Let us now assume that primary tumors at different subsites have different patterns of lymphatic spread, corresponding to different model parameters $\boldsymbol{\theta}$. Training a separate model for every possible subsite (ICD-10 code) would require a sufficiently large dataset for every tumor site. However, anatomically nearby locations are expected to show very similar patterns of LNL involvement. Therefore, we consider a mixture model.
Let us assume that we have a dataset $\mathbf{D}$ that is specified via the number of patients N_{is} that were diagnosed in LNL involvement state i and had a primary tutor in subsite s. Let us further assume that we want to describe this dataset using a mixture of M HMMs, each with a different set of model parameters $\boldsymbol{\theta_m}$. As the generative model of the data, we assume that a patient with subsite s is generated with probability π_{sm} from model m. The likelihood of the dataset can then be written as

$$P(\mathbf{D} \mid \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_s \prod_i \left[\sum_{m=1}^{M} \pi_{sm} P_m \left(\mathbf{X} = \boldsymbol{\xi}_i \mid \boldsymbol{\theta}_m\right) \right]^{N_{is}} \tag{6}$$

We now have two types of parameters, the probabilities of tumor spread for the different models $\boldsymbol{\theta_m}$, and the mixing coefficients π_{sm}. Assuming a uniform prior in the interval $[0,1]$ for all parameters, the posterior distribution over the parameters $P(\boldsymbol{\theta}, \boldsymbol{\pi} \mid \mathbf{D})$ is given by the likelihood equation 6 except for a normalisation constant. In this work, we use Markov chain Monte Carlo sampling (MCMC) via the `emcee` Python package [11] to sample model parame-

ters from the posterior distribution. However, $P(\boldsymbol{\theta}, \boldsymbol{\pi} \mid \mathbf{D})$ itself is a multi-model distribution because one can permute the different models. To address this problem, we revert to an *expectation-maximization (EM)* algorithm where we iteratively sample model parameters $\boldsymbol{\theta_m}$ using MCMC and then determine the most likely mixing coefficients.

2.3 Multi-Centric Data

For the analyses in this work, we used five datasets from four different institutions:

1. 287 oropharyngeal patients from the University of Zurich in Switzerland
2. 263 oropharyngeal patients from the Centre Léon Bérard in France
3. 289 oropharyngeal and oral cavity patients from the Inselspital Bern in Switzerland
4. 239 oropharyngeal and oral cavity patients from the Centre Léon Bérard in France
5. 162 oropharyngeal patients from the Hospital Vall d'Hebron in Spain

The data sets 1-4 are publicly available in the form of CSV tables [8, 12] and may be interactively explored in our **Ly**mphatic **Pro**gression e**X**plorer LyProX. Data set 5 is not yet public but of similar format. For each patient, the primary tumor subsite is reported (among other patient and tumor characteristics) and each individual LNL is reported as metastatic or healthy, according to the available diagnostic modalities (in part pathology after neck dissection, otherwise clinical involvement).
In figure 2, we have plotted the prevalence of involvement in the four ipsilateral LNLs I, II, III, and IV stratified by the primary tumor's subsite. We included patients with tumors in the oral cavity, gum (C03), floor of mouth (C04), other/unspecified parts of the mouth (C06), other/unspecified parts of tongue (C02), palate (C05), and tumors in the oropharynx (C10), tonsil (C09), base of tongue (C01), resulting in 1242 patients. The figure illustrates the variations in LNL involvement between subsites within the oral cavity and oropharynx categories.

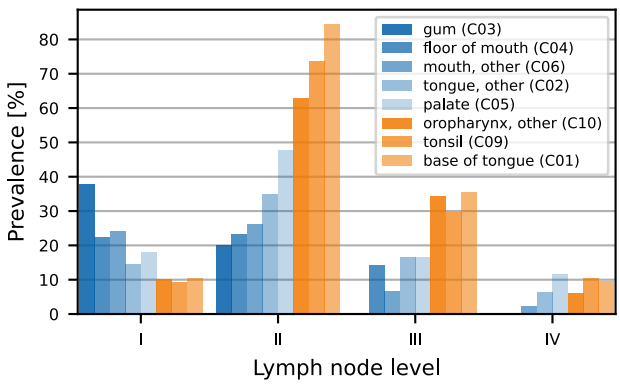

Figure 2: Prevalence of LNL involvement stratified by subsite. The subsites are sorted in ascending order by their prevalence of involvement in LNL II. Oral cavity subsites are plotted in shades of blue, oropharynx subsites in shades of orange.

3 Results

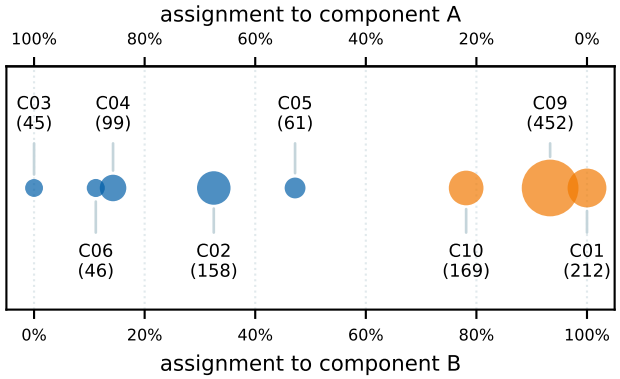

Figure 3: Assignment of each subsite to each of the two components. The further left a subsite, the more it is assigned to component A, the further right, the more to component B. The size of the marker (area) corresponds to the number of patients in the subsite.

We demonstrate the methodology for a mixture model with $M = 2$ components, considering the ipsilateral involvement of LNLs I, II, III, and IV and the primary tumor subsites shown in figure 2. In figure 3 we show the resulting mixture coefficients π_{sm}. The interpretation of this result is as follows: Tumors of the base of tongue (C01) are fully described by component A, and tumors of the gum (C03) are fully described by component B. These two subsites are the most distinct regarding the involvement of LNLs I and II, and the result is thus intuitive. Component A may be interpreted as a model for oropharynx-like tumor spread, and component B as a model for oral cavity-like tumor spread. All other subsites are described as mixtures. Tumors in the tonsil (C09) have LNL involvement similar to base of tongue tumors and are mostly assigned to component A. Instead, tumors of the palate (C05) are to similar degree assigned to components A and B, which is consistent with the anatomical location and the observation that the LNL involvement is in between oropharynx and oral cavity-type patterns.

4 Discussion

We have previously developed a model of lymphatic progression of HNSCC using HMM, which allows us to estimate the probability of occult lymph node metastases in clinically negative LNLs. Mixture models are a suitable method to incorporate the primary tumor location into the model, which allows us to account for differences in lymph node involvement for different subsites. Future work will extend the work to tumors in the hypopharynx and larynx and optimize the number of model components.

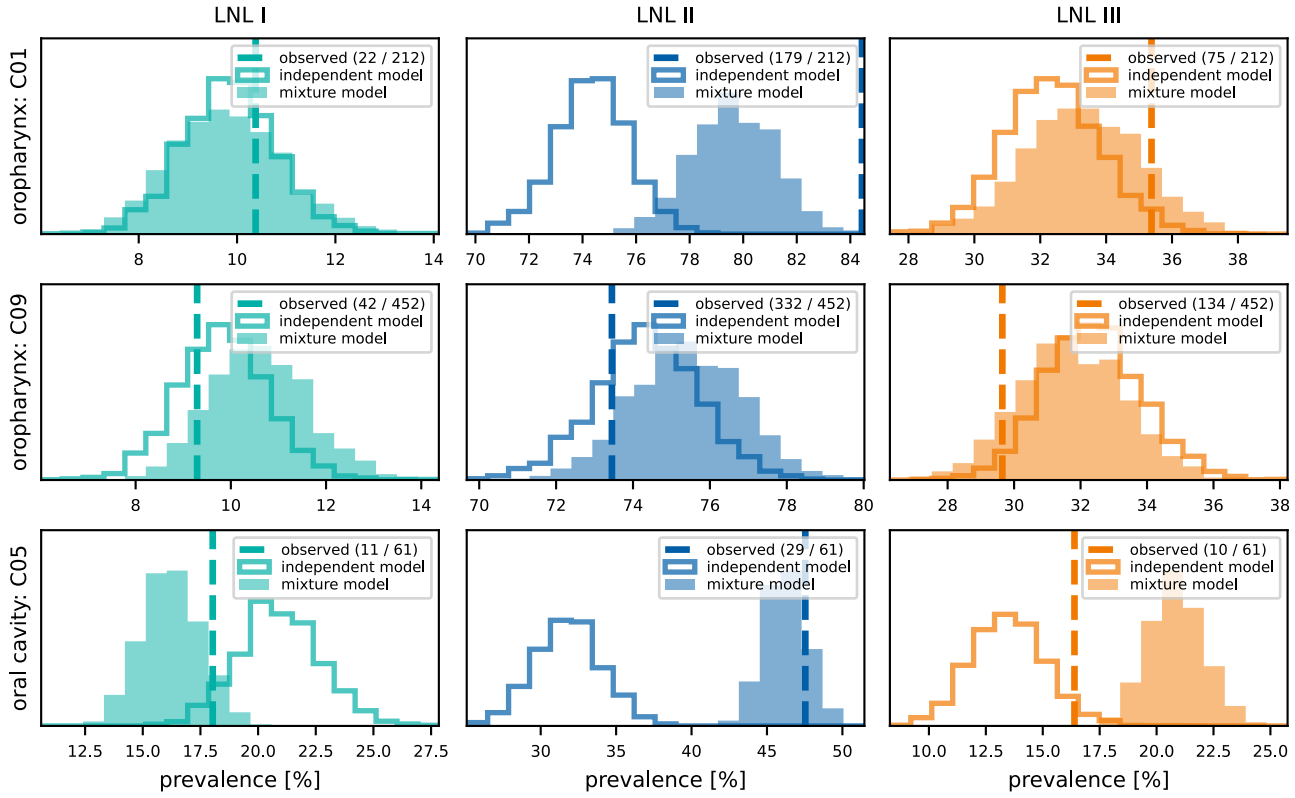

Figure 4: The prevalence of involvement as seen in the data (vertical dashed lines), predicted by an independent model for the oropharyngeal or oral cavity patients (outlined histograms), and predicted by the mixture model (filled histograms). Each row correpsonds to one subsite and each column to the predicted or observed prevalence in one LNL.

References

[1] R. Lindberg. "Distribution of Cervical Lymph Node Metastases from Squamous Cell Carcinoma of the Upper Respiratory and Digestive Tracts". *Cancer* 29.6 (June 1972), pp. 1446–1449. DOI: 10.1002/1097-0142(197206)29:6<1446::AID-CNCR2820290604>3.0.CO;2-C.

[2] J. Woolgar. "Histological Distribution of Cervical Lymph Node Metastases from Intraoral/Oropharyngeal Squamous Cell Carcinomas". *British Journal of Oral and Maxillofacial Surgery* 37.3 (June 1999), pp. 175–180. DOI: 10.1054/bjom.1999.0036.

[3] V. Snyder, L. K. Goyal, E. M. R. Bowers, et al. "PET/CT Poorly Predicts AJCC 8th Edition Pathologic Staging in HPV-Related Oropharyngeal Cancer". *The Laryngoscope* n/a.n/a (Jan. 2021). DOI: 10.1002/lary.29366.

[4] M. P. Strohl, P. K. Ha, R. R. Flavell, et al. "PET/CT in Surgical Planning for Head and Neck Cancer". *Imaging Options for Head and Neck Cancer* 51.1 (Jan. 2021), pp. 50–58. DOI: 10.1053/j.semnuclmed.2020.07.009.

[5] A. S. Ho, D. H. Kraus, I. Ganly, et al. "Decision Making in the Management of Recurrent Head and Neck Cancer". *Head & Neck* 36.1 (2014), pp. 144–151. DOI: 10.1002/hed.23227.

[6] J. Biau, M. Lapeyre, I. Troussier, et al. "Selection of Lymph Node Target Volumes for Definitive Head and Neck Radiation Therapy: A 2019 Update". *Radiotherapy and Oncology* 134 (May 2019), pp. 1–9. DOI: 10.1016/j.radonc.2019.01.018.

[7] R. Ludwig, J.-M. Hoffmann, B. Pouymayou, et al. "A Dataset on Patient-Individual Lymph Node Involvement in Oropharyngeal Squamous Cell Carcinoma". *Data in Brief* 43 (Aug. 2022), p. 108345. DOI: 10.1016/j.dib.2022.108345.

[8] R. Ludwig, A. Schubert, D. Barbatei, et al. "A Multi-Centric Dataset on Patient-Individual Pathological Lymph Node Involvement in Head and Neck Squamous Cell Carcinoma". *Data in Brief* (Dec. 2023), p. 110020. DOI: 10.1016/j.dib.2023.110020.

[9] R. Ludwig, B. Pouymayou, P. Balermpas, et al. "A Hidden Markov Model for Lymphatic Tumor Progression in the Head and Neck". *Scientific Reports* 11.1 (Dec. 2021), p. 12261. DOI: 10.1038/s41598-021-91544-1.

[10] S. S. Batth, J. J. Caudell, and A. M. Chen. "Practical Considerations in Reducing Swallowing Dysfunction Following Concurrent Chemoradiotherapy with Intensity-Modulated Radiotherapy for Head and Neck Cancer". *Head Neck* 36 (2014), pp. 291–298. DOI: 10.1002/hed.23246.

[11] D. Foreman-Mackey, D. W. Hogg, D. Lang, et al. "Emcee: The MCMC Hammer". *pasp* 125.925 (Mar. 2013), p. 306. DOI: 10.1086/670067.

[12] R. Ludwig, J.-M. Hoffmann, B. Pouymayou, et al. "Detailed Patient-Individual Reporting of Lymph Node Involvement in Oropharyngeal Squamous Cell Carcinoma with an Online Interface". *Radiotherapy and Oncology* 169 (Apr. 2022), pp. 1–7. DOI: 10.1016/j.radonc.2022.01.035.