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

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

[1]Department of Physics, University of Zurich.

[2]Radiation Oncology, University Hospital Zurich.

[3]Department of Radiation Oncology, Bern University Hospital.

[4]Department of ENT, Head & Neck Surgery, Bern University Hospital.

[5]Head and Neck Anticancer Center, Bern University Hospital.

[6]Institute of Tissue Medicine and Pathology, Bern University Hospital.

[7]Department of ENT, Head & Neck Surgery, Réseau Hospitalier Neuchâtelois.

[8]Department of Radiation Oncology, Centre Léon Bérard.

[9]Department of Head and Neck Surgery, Centre Léon Bérard.

[10]Institute of Pathology, Klinikum Stuttgart.

[11]Departement of Radiation Oncology, Hospital Vall d'Hebron.

*Corresponding author(s). E-mail(s): roman.ludwig@usz.ch;

Contributing authors: yoel.perezhaas@usz.ch; jan.unkelbach@usz.ch;

Abstract

Purpose: to be done

Methods: to be done

Results: to be done

Conclusions: to be done

Introduction

Head and neck squamous cell carcinomas (HNSCCs) are known to spread through the lymphatic system often leading to metastases in the lymph nodes [1, 2]. To minimize nodal recurrences, lymph node levels (LNLs) at risk of harboring occult metastases are typically irradiated electively. Current guidelines for different tumor locations are based on the overall prevalence of nodal disease as reported in literature [1–3].

To personalize the prediction of the risk of occult metastases, given a patient's individual diagnosis, we previously published a large, multi-centric dataset where the lymphatic involvement per LNL is available for each patient[4, 5]. Building on this dataset, we introduced an interpretable hidden Markov model (HMM), trained to predict the risk for occult nodal disease, given an individual patient's diagnosis [6].

Personalized risk predictions could enable clinicians to safely reduce the elective clinical target volume (CTV-N), potentially decreasing treatment-related side effects that impair a patient's quality of life, without compromising the efficacy of the treatment [7].

Initially, separate models were trained for distinct tumor locations, such as the oropharynx and oral cavity. These tumor locations are also used in guidelines to define the elective target volumes [3]. However, this approach did not account for variations in lymphatic spread between subsites within these tumor regions. With data from more than 2700 patients available, we can now further analyze subsite specific spread patterns. Closer analysis showed that pooling subsites into a single model led to inaccurate predictions, as it failed to capture distinct lymphatic spread patterns. To resolve this, we propose using a mixture of HMMs, which allows us to model the lymphatic spread more accurately for tumors located near anatomical borders, such as those between the oropharynx and oral cavity (e.g., tumors in the palate).

Additionally, we extend the analysis to a broader mixture model that encompasses tumors of the oral cavity, oropharynx, hypopharynx, and larynx, resulting in further personalized predictions of lymphatic spread across these regions.

Data on Lymphatic Progression Patterns

For the analyses in this work, we used seven datasets from 5 institutions resulting in 2741 patients in total.

1. 0 oropharyngeal patients from the University of Zurich in Switzerland
2. 0 oropharyngeal patients from the Centre Léon Bérard in France
3. 0 oropharyngeal, larynx and oral cavity patients from the Inselspital Bern in Switzerland
4. 0 oropharyngeal, larynx and oral cavity patients from the Centre Léon Bérard in France
5. 0 oropharyngeal, larynx and oral cavity patients from the University of Zurich in Switzerland
6. 164 oropharyngeal patients from the Hospital Vall d'Hebron in Spain (not yet public)
7. 979 hypopharynx, larynx and oral cavity patients from University Medical Center Groningen (not yet public)

The datasets 1-4 are publicly available as CSV tables [5, 8] and can be interactively explored on LyProX. For each patient the primary tumor subsite is reportd and each indicidual LNL is reported as either metastatic or healthy given the available diagnostic modalities, which include pathology after neck dissection in some patients. In this work we will stratify the tumor locations into different ICD codes which are depicted in figure 1.

The prevelance of involvement in LNLs I, II, III, IV and V is shown in figure 2. The involvement is stratified per tumor subsite and t-stage. The figure illustrates the variations in LNL involvement between subsites within oral cavity (blue), oropharynx (green), hypopharynx (red) and larynx (orange). The involvement pattern presents a continuous change over the tumor subsites. Where tumors in the oral cavity show the most prominent LNL I involvement. As the tumor location moves towards the oropharynx LNL II involvement increases. Moving the tumor location further in caudal direction towards the hypopharynx increases LNL III involvement while LNL I and II involvement decrease. Larnygeal tumors show the least LNL I involvement.

Unilateral Model for Lymphatic Progression

In this chapter we will briefly summarize unilateral model for ipsilateral lymph node involvement introduced in Ludwig et al. [6], presenting the notation which is then needed to extend the HMM to a mixture model encompassing multiple tumor subsites.

The HMM describes each LNL $v \in 1, 2, ..., V$ by a binary random variable corresponding to the status of the LNL; healthy (0) or involved (1). The entire state of a patient with V LNLs is defined by the V-dimensional vector $\mathbf{X} = [X_1, X_2, ..., X_V]$. In the HMM, a patient's involvement is modeled over time t. Thus, 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 LNL involvement. In this paper,

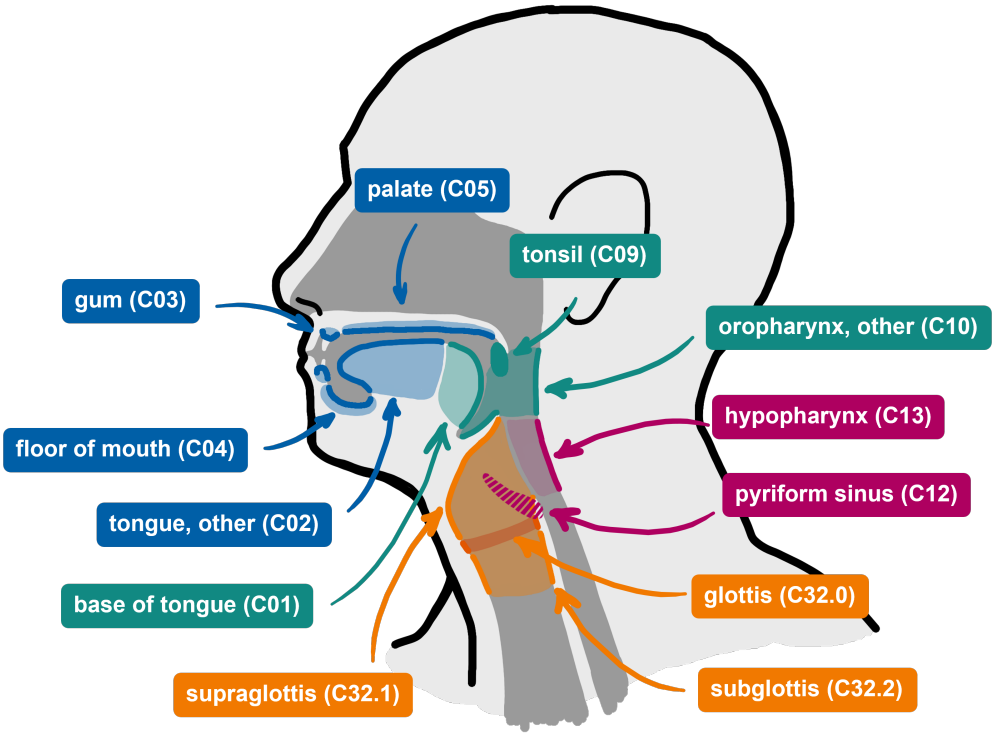

Figure 1: Anatomical sketch of the tumor subsites and their corresponding ICD-10 codes. Subsite C06 "other parts of mouth" has not been included. Further the The tumor locations are color coded in the following pattern: blue-oral cavity, green-oropharynx, red-hypopharynx, orange-larynx.

we consider ipsilateral LNLs I, II, III, IV and V, which amounts to 32 possible states. The HMM is then specified by a transition matrix $\mathbf{A}$:

$$\mathbf{A} = (A_{ij}) = P\left(\mathbf{X}[t+1] = \xi_j \mid \mathbf{X}[t] = \xi_i\right) \tag{1}$$

whose elements A_{ij} contain the conditional probabilities that a state $\mathbf{X}[t] = \xi_i$ transitions to $\mathbf{X}[t+1] = \xi_j$ over one time step. The transition matrix is specified and parameterised via the graphical model shown in figure 3. The red arcs in the graph of figure 3 are associated with 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 $t_{v \to v+1}$).

Now, let π be the *starting distribution*

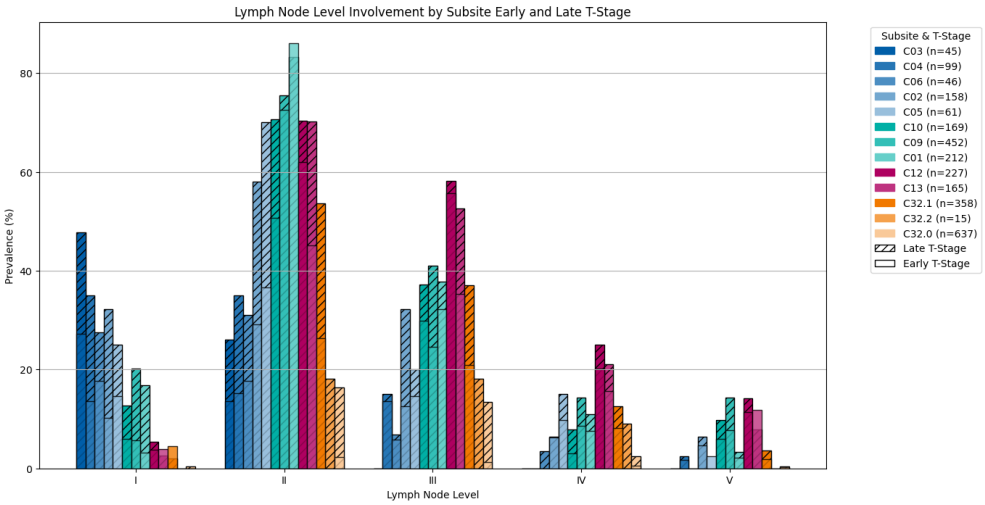

Figure 2: Prevalence of ipsilateral LNL involvement stratified by subsite. The subsites are sorted in natural order to represent the continuously changing LNL involvement. The different tumo locations are color coded, where oral cavity subsistes are depicted in blue, larynx in green, hypopharynx in red and larynx in orange. The patient data is further stratified in early t-stage (0-2) and late t-stage (3-4). The legend furter specifies the number of patients in each subsite.

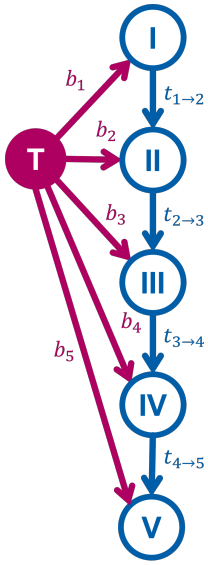

Figure 3: Parametrized graphical model of the lymphatic network considering four LNLs. Blue nodes represent the hidden states of LNLs X_v, while the red one is the tumor. Arcs represent possible routes of metastatic spread, associated with a probability.

$$\pi = (\pi_i) = P\left(\mathbf{X}[0] = \xi_i\right) \tag{2}$$

denoting the probability to start in state ξ_i at time step 0. Assuming that every patient started with all LNLs being healthy, we set π_i to zero for all states except the completly healthy state $\xi = (0, 0, 0, 0, 0)$, which has probability one.

Using the quantities introduced so far, the probability $P\left(\mathbf{X}[t] = \xi_i\right)$ to be in state ξ_i in time step t can now be conveniently expressed as a matrix product:

$$P\left(\mathbf{X}[t] = \xi_i\right) = \left(\pi \cdot \mathbf{A}^t\right)_i \tag{3}$$

This evolution implicitly marginalizes over all possible paths to arrive at state ξ_i after t time-steps. Additionally, we must marginalize over the unknown time of diagnosis using a time-prior $P_T(t)$ which is defined by a binomial distribution. The t-stage of the tumor can be included in the model by choosing different parameterizations of the binomial distribution, considering that a tumor in late t-stages was diagnosed later than a tumor in early t-stages, therefore shifting the probability of diagnosis to later time steps. This finally defines the probability distribution over all states of lymph node involvement.

$$P\left(\mathbf{X} = \xi_i \mid \theta, \mathbf{T}\right) = \sum_{t=0}^{t_{\max}} P_T(t) \left(\pi \cdot \mathbf{A}^t\right)_i \tag{4}$$

where $\theta = \{b_v, t_{v \to v+1}\}$ denotes the set of all model parameters (7 in our case). Fortunately, the exact length and shape of this distribution has little impact as previously shown [6]. 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 Ludwig et al. [6] and Ludwig [9].

With equation 4 we can compute the probability of a patient being in any state ξ_i. Therefore the likelihood of observing a the For model training we assume that the diagnoses in our data $\mathbf{D}$ we observe correspond to the hidden state $\mathbf{X}$ of the patient. Thus, learning the model parameters corresponds to maximizing the probability of observing the dataset $\mathbf{D}$:

$$P\left(\mathbf{D} \mid \theta\right) = \prod_{k}^{K} P\left(\mathbf{X}_k = \xi_i \mid \theta, T_k\right) \tag{5}$$

In equation 5 we compute the likelihood of observing patient the diagnosis of each patient k, i.e. t-stage T_k and involvement $\mathbf{X}_k = \xi_i$.

Mixture Model for Lymphatic Spread

Primary tumors at different subsites exhibit distinct lymphatic spread patterns. This presents a challenge when attempting to generalize predictive models across subsites.

One approach, as introduced in [10], uses a Hidden Markov Model (HMM) trained specifically for oropharyngeal cancer. However, extending this model to other subsites would either require generalizing over several subsites or training a separate model for each. The former approach sacrifices precision, particularly for subsites with fewer patients, while the latter approach becomes computationally intensive and introduces large uncertainties for subsites with limited patient data, such as C04 (Floor of mouth) or C05 (Palate).

To address these challenges and exploit the anatomical similarities between nearby subsites, we introduce a mixture model that combines data from all subsites into a single model. This model accounts for anatomical proximities, thereby improving predictive power while maintaining computational efficiency.

Mixture Model Formulation

The mixture model assumes that the data is generated by a set of M different lymphatic spread models. Each patient k, with their primary tumor in subsite $s \in (1, 2, \dots, S)$, is assumed to be generated by one specific model $m \in (1, 2, \dots, M)$ from this set of M models with probability π_m^s. These so-called *mixing parameters* $\pi^s = \{\pi_1^s, \pi_2^s, \dots, \pi_M^s\}$ must satisfy the condition

$$\sum_m^M \pi_m^s = 1, \quad \forall s$$

If we could record the componet m from which a patient k was drawn from, we could store this information in a binary latent vector ϵ_k. The vector ϵ_k has length M, with exactly the m-th element set to 1, indicating which model generated the patient's data, and all other elements set to 0. Thus, for patient k, the latent variable ϵ_k can be interpreted as a categorical indicator variable that encodes the assignment to one of the M lymphatic spread models. Typically in mixture models, this so-called *latent variable* is unknown. However, it can be inferred and is useful for inferring the models' parameters.

The joint probability of the observed data $\mathbf{D}$ (i.e., patient data) and the latent variables ϵ - sometimes called the *complete data likelihood* - is given by:

$$P(\mathbf{D}, \epsilon \mid \theta, \pi) = \prod_k^K \prod_m^M \left[\pi_m^{s_k} P(\mathbf{D}_k \mid \theta_m)\right]^{\epsilon_k^m} \tag{6}$$

Here:

- $\pi_m^{s_k}$ is the mixing coefficient for subsite s_k (where patient k has their tumor) and model m,
- $P(\mathbf{D}_k \mid \theta_m)$ is the likelihood of patient k's diagnosis, given that it was generated by model m,
- θ_m represents the parameters of model m,

- ϵ_k^m is the latent variable that indicates patient k was generated by model m.

In contrast, the *incomplete data likelihood* marginalizes over all possible latent assignments to reflect the uncertainty about which model generated each patient's data:

$$P(\mathbf{D}, \mid \theta, \pi) = \prod_{k}^{K} \sum_{m}^{M} \pi_m^{s_k} P(\mathbf{D}_k \mid \theta_m) \tag{7}$$

Ultimately, we want to find the parameters which maximize this likelihood function for the given data.

Note the summation inside the product. This structure of mixture models makes naive inference difficult, because the logarithm of this quantity is expensive to compute and not easy to differentiate. Thus, inferring the latent assignment is helpful, because the complete data (log-)likelihood (equation 6) does not suffer from this shortcoming.

To infer the latent assignment ϵ_k, we start with its distribution, given the observed data and model parameters:

$$\gamma(\epsilon_k^m) := \mathbb{E}[\epsilon_k^m] = \frac{\pi_m^{s_k} P\left(\mathbf{D}_k \mid \theta_m\right)}{\sum_{j \leq M} \pi_j^{s_k} P\left(\mathbf{D}_k \mid \theta_j\right)} \tag{8}$$

This expectation value - often called the *responsibility* - describes the probability that patient k was generated by model m. It can be used to compute the expected complete data likelihood:

$$\mathbb{E}_\epsilon \left[P\left(\mathbf{D}, \epsilon \mid \theta, \pi\right)\right] = \prod_{k=1}^{K} \prod_{m=1}^{M} \left[\pi_m^{s_k} P\left(\mathbf{D}_k \mid \theta_m\right)\right]^{\gamma(\epsilon_k^m)} \tag{9}$$

This expected complete data likelihood has the same tractable form as equation 6 and thus allows us to infer both the mixing coefficients π as well as each component model's parameters θ_m using straightformward inference. Also, these two sets of parameters are the only ones used for the later risk prediction. The responsiblities $\gamma(\epsilon_k^m)$ are only used during inference.

In figure 4 the mixture coefficients are illustrated. Subsites with different spread patterns, such as Gum (C03) and Base of tongue (C01), are expected to get different model assignments. Nonetheless, the latent variables for two patients with the same diagnosis, i.e. same involvement ξ and and t-stage T, but different subsites are the same.

It is important to note that this mixture model algorithm diverges from the conventional formulation of mixture models as introduced in Bishop [11]. In Bishop's work, the mixture model is defined with a single set of mixture parameters π, implying that

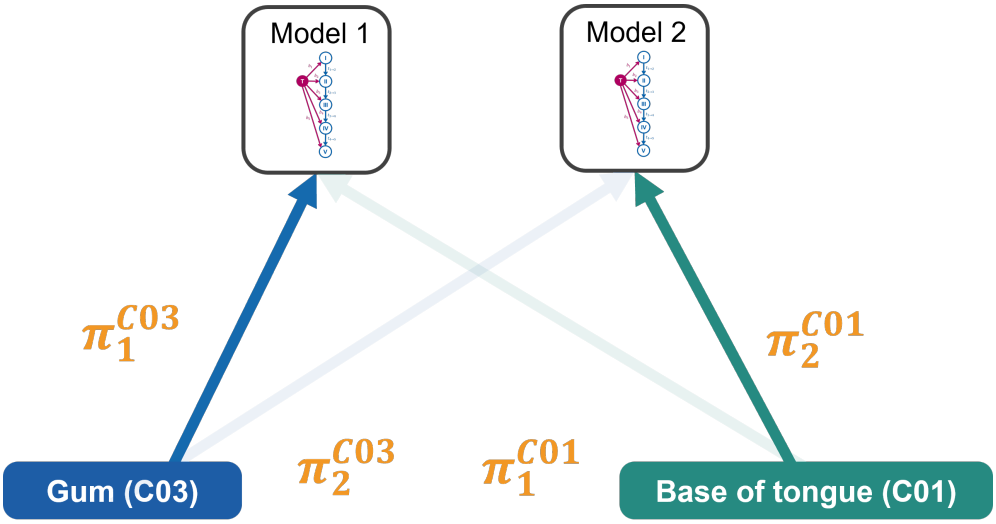

Figure 4: Illustration of mixture parameter assignment. Since Gum and Base of tongue express different spread patterns, the two models are expected to have different model assignments. The arrow visibility represents the value of the mixture parameter π, where the more visible the arrow, the larger the value for π

all data points are generated from the same mixture of models. In our approach, however, we categorize the data points (patients) into distinct cohorts (subsites), with each subsite characterized by its own set of mixture parameters. Consequently, we construct S separate mixture models, each possessing different mixture parameters while all models share the same underlying probability distributions, specifically the varying lymphatic spread models.

Expectation-Maximization (EM) Algorithm

To actually find the mixing coefficients π and all models' parameters θ_m that maximize equation 7, we follow an iterative approach called *expectation-maximization* or *EM-algorithm*. With arbitrarily initialized starting parameters, we alternate between the following two steps:

1. In the **E**xpectation step, we compute the responsibilities $\gamma(\epsilon_k^m)$, which represent the probabilities of a patient k originating from one of the models m, given the current estimates of θ and π, as given in equation 8.
2. During the **M**maximization step, we find a new set of parameters that maximize the new expected complete data likelihood (equation 9). For the mixture coefficients, we can even find an analytic solution to the new maximum:

$$\pi_m^s = \frac{1}{|K_s|} \sum_{k \in K_s} \gamma(\epsilon_k^m)$$

where we sum over the set of all patients K_s with their tumor in subsite s.
The models' new parameters are found by numerically maximizing the respective likelihood, weighted by the responsibilities:

$$\ln P(\mathbf{D}, \epsilon \mid \theta_m) = \sum_{k}^{K} \gamma(\epsilon_k^m) \left[\ln \pi_m^{s_k} + \ln P\left(\mathbf{D}_k \mid \theta_m\right) \right] \tag{10}$$

By iterating these steps, the EM algorithm is guaranteed to converge to a (local) maximum of the incomplete data likelihood (equation 7).

Three component Mixture Model

We illustrate the methodology for a mixture model with M = 3 components, considering the ipsilateral involvement of LNLs I, II, III, IV, and V. We include the ICD codes as subsites for oral cavity, hypopharynx and oropharynx. In figure 5 the convergence of the negative log-likelihood and change in model parameters is depicted. After a random inizialization, the algorithm rapidly converges. The algorithm was stopped when the difference of log-likelihood between two iterations was below 0.01.

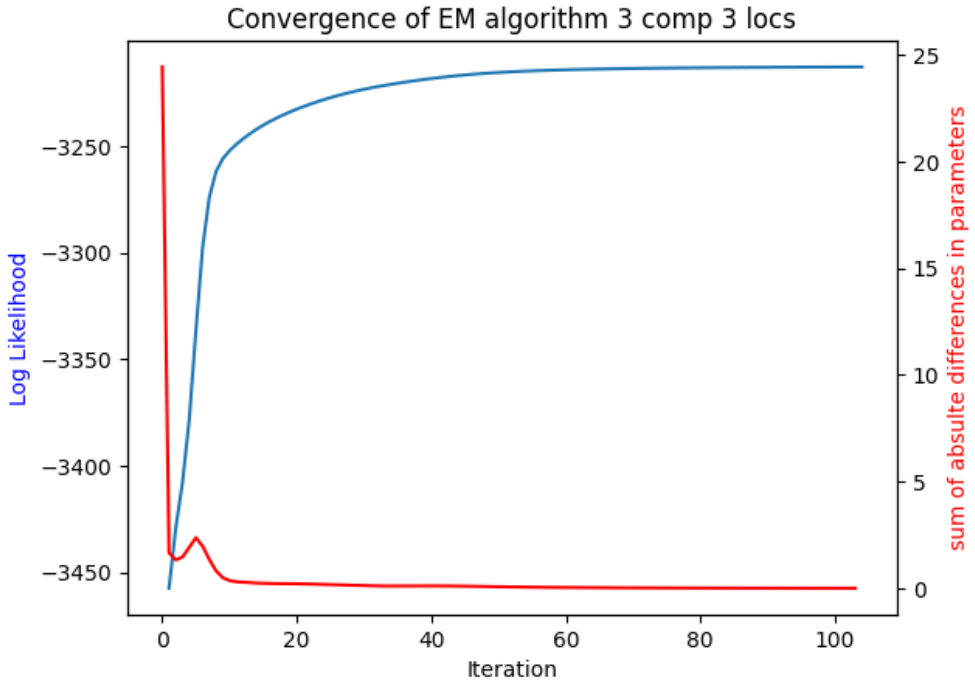

Figure 5: The y-axis on the left shows the negative likelihood convergence depicted in the blue line. The y-axis on the right shows the sum of absulte difference between all model parameters showing that the parameter values stabilize rapidly as well.

In figure 6, we visualize the resulting mixture coefficients π using a spatial representation, where the vertices of the triangle correspond to the three components. In figure 7, these mixture coefficients are presented in matrix form, with the y-axis representing the ICD codes, showing how the mixture components in each row add up to 1.

The spatial plot in figure 6 illustrates how the model assigns the three components to different tumor subsites. Component 0, located at the bottom right of the triangle, primarily characterizes oropharyngeal subsites. For instance, the base of tongue subsite (C01), which exhibits the highest involvement of LNL II, is fully assigned to this component. Similarly, subsite C10, which includes several oropharyngeal regions, is assigned roughly 50% to the oropharynx-like component, with the remaining mixture distributed across the other two components.

Hypopharyngeal subsites, on the other hand, are fully assigned to Component 2, located at the top vertex of the triangle. Meanwhile, the gum subsite (C03), with predominant LNL I involvement, is entirely assigned to Component 1, situated at the bottom left. As subsites anatomically approach the oropharynx, their mixture coefficients for the oropharynx-like component increase. This is evident in the subsites C02 (tongue) and C05 (palate), which display a higher proportion of oropharyngeal influence in their mixture. These results conform well with the involvement patterns observed in the data.

Four component Mixture Model

We can extend the mixture model to include the larynx. The larynx patients are more finely divided into ICD codes C32.0, C32.1 and C32.2 as there is a notable difference between these ICD codes in figure 2.

Simlarly to the three component model, we can analyze the convergence over the iterations of the EM-algorithm. In figure 8 we can see that in this more complex model, the likelihood space becomes more complex as at around 200 iterations, the negative log-likelihood starts to increase faster again.

The component assignment is shown in figure 9. Similarly to the 3-component model the different tumor locations are assigned to a one of the components….

Here i probably should permute the components such that we have the same ordering as in the 3-component model.

add some analysis (level specific predictions and also comparison to single tumor location models.)

References

[1] Mukherji, S.K., Armao, D., Joshi, V.M.: Cervical nodal metastases in squamous cell carcinoma of the head and neck: What to expect. Head and Neck **23**(11), 995–1005 (2001) https://doi.org/10.1002/hed.1144

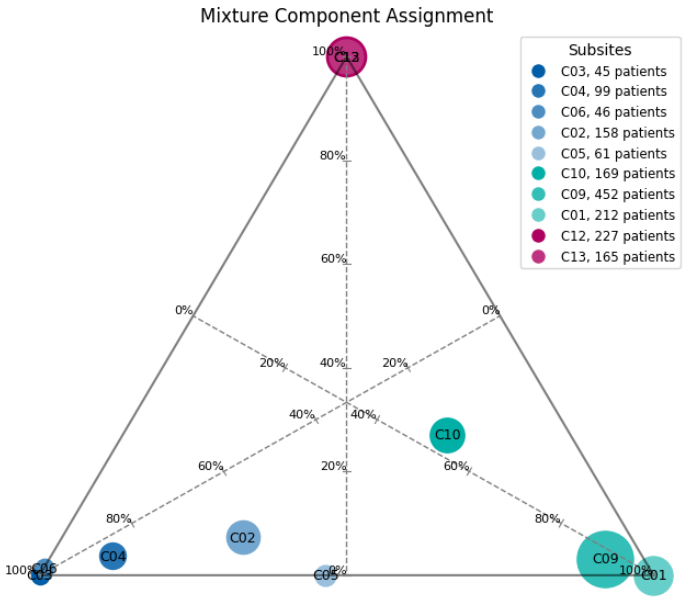

Figure 6: Assignment of each subsite to each of the three components. The closer a subsite is to a vertex, the more it is assigned to the corresponding component, with component 0 on the bottom right, 1 on the bottom left and 2 on the top. The size of the marker (area) corresponds to the number of patients in each subsite.

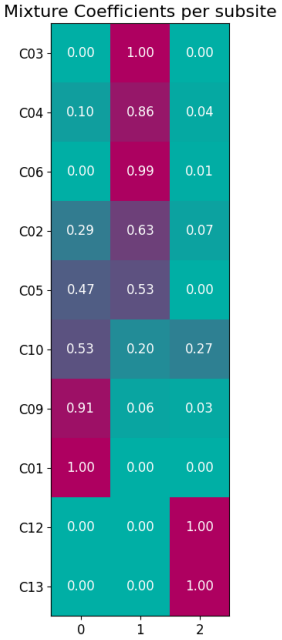

Figure 7: Matrix representation of component assignment. Each row of the matrix corresponds to each ICD code. The collums represent the three different components

[2] Shah, J.P., Candela, F.C., Poddar, A.K.: The patterns of cervical lymph node metastases from squamous carcinoma of the oral cavity. Cancer **66**(1), 109–113 (1990) https://doi.org/10.1002/1097-0142(19900701)66:1<109::AID-CNCR2820660120>3.0.CO;2-A

[3] Biau, J., Lapeyre, M., Troussier, I., Budach, W., Giralt, J., Grau, C., Kazmierska, J., Langendijk, J.A., Ozsahin, M., O'Sullivan, B., Bourhis, J., Grégoire, V.: Selection of lymph node target volumes for definitive head and neck radiation therapy: A 2019 Update. Radiotherapy and Oncology **134**, 1–9 (2019) https://doi.org/10.1016/j.radonc.2019.01.018

[4] Ludwig, R., Hoffmann, J.-M., Pouymayou, B., Morand, G., Däppen, M.B., Guckenberger, M., Grégoire, V., Balermpas, P., Unkelbach, J.: A dataset on

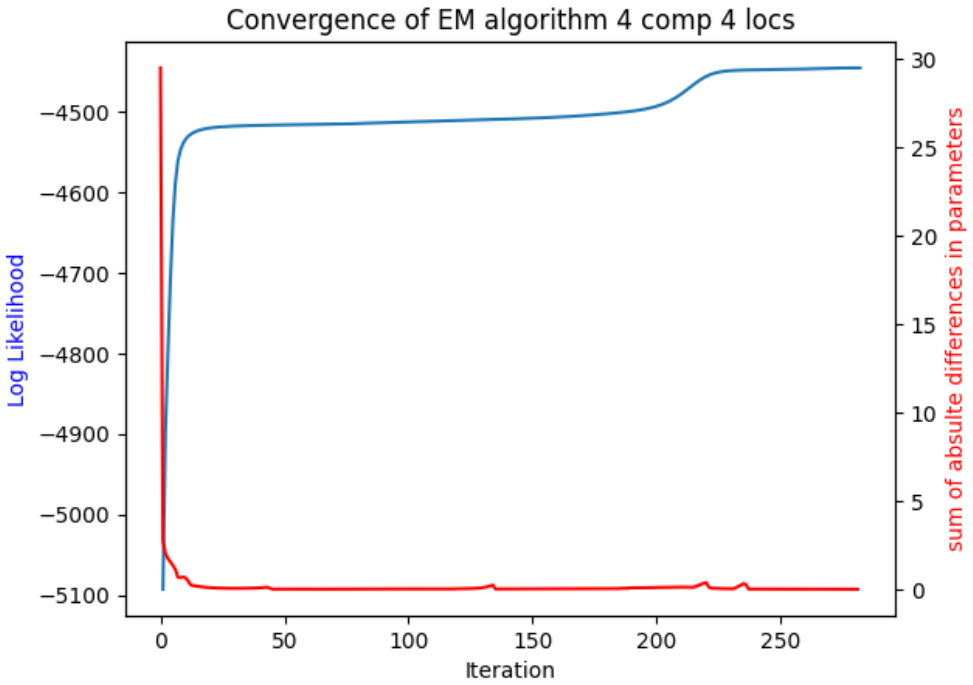

Figure 8: The y-axis on the left shows the negative likelihood convergence depicted in the blue line. The y-axis on the right shows the sum of absulte difference between all model parameters.

patient-individual lymph node involvement in oropharyngeal squamous cell carcinoma. Data in Brief **43**, 108345 (2022) https://doi.org/10.1016/j.dib.2022.108345

[5] Ludwig, R., Schubert, A., Barbatei, D., Bauwens, L., Werlen, S., Elicin, O., Dettmer, M., Zrounba, P., Balermpas, P., Pouymayou, B., Grégoire, V., Giger, R., Unkelbach, J.: A multi-centric dataset on patient-individual pathological lymph node involvement in head and neck squamous cell carcinoma. Data in Brief, 110020 (2023) https://doi.org/10.1016/j.dib.2023.110020

[6] Ludwig, R., Pouymayou, B., Balermpas, P., Unkelbach, J.: A hidden Markov model for lymphatic tumor progression in the head and neck. Scientific Reports **11**(1), 12261 (2021) https://doi.org/10.1038/s41598-021-91544-1

[7] Batth, S.S., Caudell, J.J., Chen, A.M.: Practical considerations in reducing swallowing dysfunction following concurrent chemoradiotherapy with intensity-modulated radiotherapy for head and neck cancer. Head Neck **36**, 291–298 (2014) https://doi.org/10.1002/hed.23246

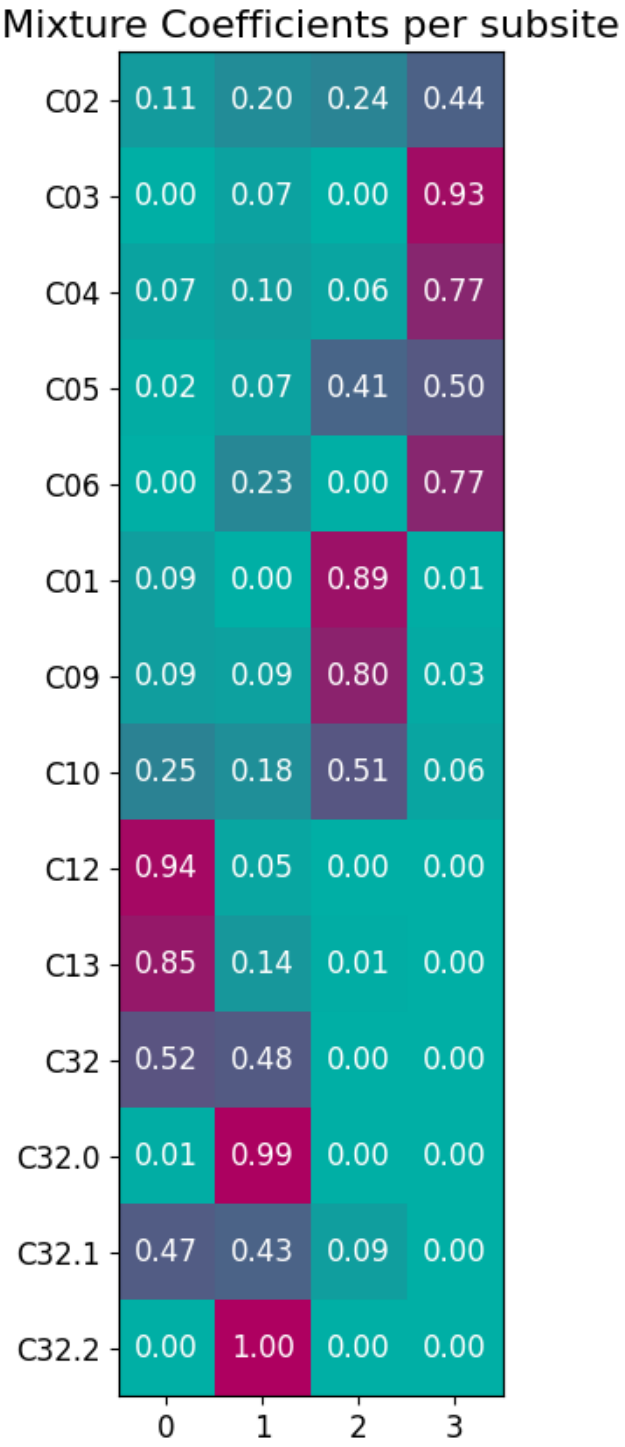

Figure 9: Matrix representation of component assignment. Each row of the matrix corresponds to each ICD code. The collums represent the three different components

[8] Ludwig, R., Hoffmann, J.-M., Pouymayou, B., Däppen, M.B., Morand, G., Guckenberger, M., Grégoire, V., Balermpas, P., Unkelbach, J.: Detailed patient-individual reporting of lymph node involvement in oropharyngeal squamous cell carcinoma with an online interface. Radiotherapy and Oncology **169**, 1–7 (2022) https://doi.org/10.1016/j.radonc.2022.01.035

[9] Ludwig, R.: Modelling lymphatic metastatic progression in head and neck cancer. PhD thesis, University of Zurich, Zurich (2023)

[10] Ludwig, R., Pouymayou, B., Balermpas, P., Unkelbach, J.: A dynamic model for lymphatic progression of cancer through the head & neck. In: ESTRO, Madrid, p. 75 (2021)

[11] Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, New York (2006)