A mixture of hidden Markov models to predict the lymphatic spread in head and neck cancer

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 multicentric data. The mixture model manages to cluster 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 nodal metastases, which requires pathological examination of extracted tissue [3, 4]. Since the recurrence of nodal disease is detrimental to a patient's outcome [5], large volumes in the head and neck region are often irradiated electively to minimize the risk of missing occult disease. Decision guidelines about which nodal regions – i.e., anatomically defined 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 diease, 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 clinitians to safely reduce the elective clinical target volume (CTV-N) and thus reduce side-effects that degrade the patient's quality of life [10].

HNSCC patients with primary tumors at different subsites, e.g. in the oral cavity and in the oropharynx, also show different patterns of lymphatic spread [1, 2]. Our model does so far not have the capability to naturally describe different tumor subsites. To that end, we present an approach using mixtures of these HMMs in this work. This makes intuitive sense, because if a tumor lies anatomically between two anatomically close subsites with slightly different spread patterns, we may be able to describe its lymphatic progression as a mixture of these two spread patterns.

2 Materials and Methods

To compute the personalized risk of occult disease $\mathbf{X}$, given a diagnosis $\mathbf{Y}$, we can begin by stating 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 diagnosis. The crucial term that our model attempts to compute, is the prior probability of involvement $P(\mathbf{X})$ (or rather of the hidden state of involvement, see the next section).

2.1 Hidden Markov Model for Lymphatic Progression

Each LNL $v \in \{1, 2, \dots, V\}$ considered in our model is represented by a binary random variable (RV) $X_v[t]$ taking on the true state of that level at the abstract time-step t (0 for "healthy" and 1 for "involved"). Collected in a random vector $\mathbf{X}[t] = (X_1[t], X_2[t], \ldots, X_V[t])$ they form the patient's state w.r.t. their lymphatic involvement at time t.

We model the process of tumor progression via lymphatic drainage by connecting the RVs in a graph, as shown in figure 1. The arcs in this graph represent conditional probabilities. The orange arcs correspond to observing a diagnosis Y_v, given the true state X_v. For the sake of brevity, we will not go into the details of describing how to infer the true – but technically hidden – state of LNL involvement from diagnoses with lower-than-one sensitivity and specificity. This description can be found in Ludwig [11]. Throughout this work, we instead combine diagnostic and pathologic involvement information from the data into a "maximum likelihood" diagnosis and assume its sensitivity and specificity to be one, meaning the normally hidden state X_v becomes the observed state. This simplification is reasonable for pathologic involvement and because we are at this stage more interested in testing whether our model is able to describe a realistic probability distribution over lymphatic involvement.

The red arcs in the graph of figure 1 depict the probability that the primary tumor spreads within one abstract time-step. While the blue arcs symbolize the spread from an upstream LNL – given it is already metastatic – to a downstream level. For example, the edge from X_2 to X_3 encodes the probability $P(X_3[t+1] \mid X_2[t])$, which is parametrized with b_3, and $t_{2\to 3}$ and tabulated in table 1. There is an additional restriction on any LNL's state $X_v[t+1] = 0$ to be healthy: It requires that the level was also healthy in the time-step before, meaning $X_v[t] = 0$. This is because we assume no spontaneous self-healing of metastatic levels.

Table 1: Conditional probability $P(X_3[t+1] \mid X_2[t])$ for a spread from LNL II to III during the transition from t to $t+1$. This corresponds to one of the blue arcs in figure 1. Note that the values in the row with $X_3[t+1] = 0$ is all zeros, and the row with $X_3[t+1] = 1$ all ones if $X_3[t] = 1$.

	$X_2[t] = 0$	$X_2[t] = 1$
$X_3[t+1] = 0$	$1 - b_3$	$(1 - b_3)(1 - t_{2\to 3})$
$X_3[t+1] = 1$	b_3	$1 - b_3 - t_{2\to 3} + b_3 t_{2\to 3}$

With the introduced conditional probabilities, we can now compute the joint probability of any complete state $\mathbf{X}[t] = \boldsymbol{\xi}_i$ transitioning to any other possible state $\mathbf{X}[t+1] = \boldsymbol{\xi}_j$. Here, when we use $\boldsymbol{\xi}$ instead of $\mathbf{x}$ for the values the random vector $\mathbf{X}$ can take on, the i and j enumerate all 2^V combinations of the V binary RVs. In the graph shown in figure 1, this amounts to $2^V = 8$ distinct $\boldsymbol{\xi}_i$. Because these terms are essentially products of terms like those in table 1. We can then collect these terms in a *transition matrix* $\mathbf{A}$:

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

Note that this matrix still depends on the b_v and $t_{r\to v}$ parameters, although we have dropped the explicit dependcy to keep the equations brief. Now, assuming that every patient started their disease with all LNLs being healthy, we can define the *starting distribution* $\boldsymbol{\pi}$:

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

And set every entry of this starting distribution to zero, except the first one, which we set to one. This means at $t = 0$ there is a probability of one to be in the completly healthy state $\boldsymbol{\xi}_0 = (0, 0, \ldots, 0)$.

Using the quantities introduced so far, the probability distribution vector with elements $P\left(\mathbf{X}[t] = \boldsymbol{\xi}_i\right)$ after t time-steps can now be conveniently expressed as a matrix product:

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

This evolution implicitly marginalizes over all possible paths to arrive at state $\boldsymbol{\xi}_i$ after t time-steps. Additionally, we also need to marginalize over the time of diagnosis – which is unknown – using a time-prior $P_T(t)$. Fortunately, the exact length and shape of this distribution on its own has little impact. But because we assume that early and advanced T-category patients are fundamentally the same, just on average diagnosed at different times t, we use the time-prior only to separate the respective patient's evolutions:

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

where $T \in \{\text{early}, \text{advanced}\}$ denotes the T-category. We fix $P_{\text{early}}(t)$ to a binomial distribution with parameters $n = t_{\max} = 10$ and $p_{\text{early}} = 0.3$, while the advanced T-category's time-prior is also a binomial distribution where the p_{advanced} parameter is learned.

Note that equation 5 still depends on the parametrization of the transition matrix. We collect these parameters in a vector $\boldsymbol{\theta} = \{b_v, t_{r\to v}, p_{\text{advanced}}\}$.

To train our model, we need to compute the likelihood of a dataset $\mathbf{D} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ of N patients, given the model's parameters $\boldsymbol{\theta}$. This is given by a product of the terms in equation 5:

$$P(\mathbf{D} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} P(\mathbf{X} = \mathbf{x}_i \mid T_i, \boldsymbol{\theta}) \tag{6}$$

Typically, one would compute the logarithm of this quantity for computational reasons. There is a wide array of available methods to obtain maximum likelihood estimates from this function or sample from the posterior over $\boldsymbol{\theta}$. In this work, we use Markov chain Monte Carlo sampling (MCMC) via the `emcee` Python package [12].

2.2 Mixture of HMMs

The just introduced model is capable of learning one set of or distribution over parameters $\boldsymbol{\theta}$ from a cohort of patients

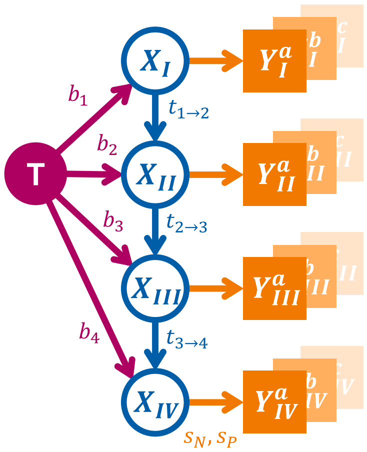

Figure 1: Parametrized graph representation of the lymphatic network considering four LNLs. Blue, round nodes represent the hidden RVs, while orange square nodes show the observed RVs. Arcs represent a conditional probability parametrized with the quantity noted next to it.

with a primary tumor in a given subsite. If we tried to train it with a cohort consisting of patients with tumors in two very different subsites, the model would likely learn parameters that represent a compromise between the two subcohort's true parameters. This compromise might describe neither of the subcohorts' lymphatic spread patterns sufficiently well.

In such cases, mixture models are often considered. They assume the data to come from a finite mixture distribution, which – in our particular case – can be written as follows:

$$P\left(\mathbf{D} \mid \boldsymbol{\Psi}\right) = \sum_{j=0}^{g} c_j P\left(\mathbf{D} \mid \boldsymbol{\theta}_j\right) \tag{7}$$

Here, the $\mathbf{c} \in [0,1]^g$ is the vector of mixing proportions with $\sum_{j=0}^{g} c_j = 1$, while the $\boldsymbol{\Psi} = \left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g\right)$ is the vector of all g models' parameters. Note that we will implement our model such that some of the parameters in each $\boldsymbol{\theta}_j$ are shared across the g components – namely the $t_{r\to v}$ corresponding to the blue arcs in figure 1.

Let now $\mathfrak{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_s)$ be a dataset consisting of s subcohorts of patients. Within a subcohort i we find N_i patients with tumors in the same subsite. We can then introduce a *latent variable* $\mathbf{Z}$ with a one-hot-encoding: Basically, it can take on values $\mathbf{z}_i \in \{0,1\}^g$ with $z_{ij} = 1$ if subcohort i belongs to component j and $z_{ij} = 0$ else.

The latent variables are helpful in resolving the invariance of the likelihood w.r.t. permutations of the component labels, which may introduce problems, e.g. for common MCMC sampling methods. The $\mathbf{Z}$ allows us to derive two sets of interdependent equations that we may solve in an iterative fashion (see e.g. Bishop [13] for a detailed derivation) that is commonly referred to as *expectation-maximization (EM)* algorithm:

The first set are the probabilities of subcohort i to belong to component j, given a set of parameters $\boldsymbol{\Psi}^\star$. These are often called the *responsibilities*:

$$\gamma(z_{ij}) = P\left(z_{ij} = 1 \mid \mathbf{D}_i, \boldsymbol{\Psi}^\star, \mathbf{c}\right) = \frac{c_j P\left(\mathbf{D}_i \mid \boldsymbol{\theta}_j^\star\right)}{\sum_{k=0}^{g} c_k P\left(\mathbf{D}_i \mid \boldsymbol{\theta}_k^\star\right)}$$

From this, we can compute new mixing proportions $c_j^\star = \sum_{i=1}^{s} \gamma(z_{ij})/s$ and then infer new parameters $\boldsymbol{\Psi}^\star$ – e.g. via MCMC sampling – from the resulting likelihood, which is the second set:

$$P\left(\mathbf{D} \mid \boldsymbol{\Psi}, \mathbf{c}^\star\right) = \sum_{j=0}^{g} c_j^\star P\left(\mathbf{D} \mid \boldsymbol{\theta}_j\right)$$

2.3 Implementation

We did this in this and that fashion…

2.4 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 Swizerland
2. 263 oropharyngeal patients from the Centre Léon Bérard in Fance
3. 289 oropharyngeal and oral cavity patients from the Inselspital Bern in Swizerland
4. 239 oropharyngeal and oral cavity patients from the Centre Léon Bérard in Fance
5. 162 oropharyngeal patients from the Hospital Vall d'Hebron in Spain

The data comes in the form of CSV tables and are – except for the last and most recent addition – publicly available [8, 14] and may be interactively explored in our **Ly**mphatic **Pro**gression e**X**plorer LyProX. Each row of these tables corresponds to one patient and details in which LNL metastatic

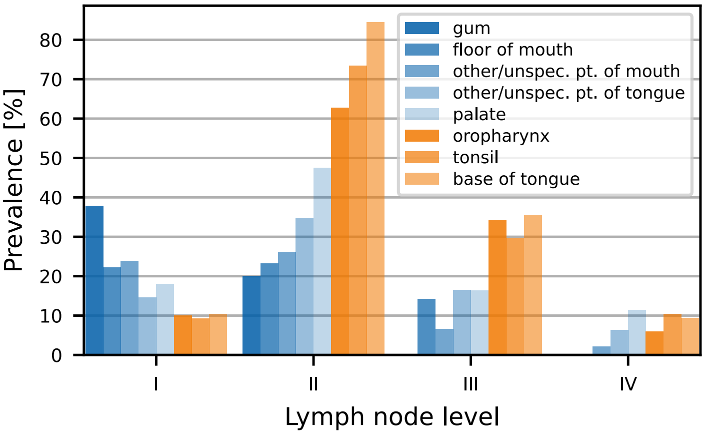

Figure 2: Prevalence of LNL involvement stratified by subsite. The subsites are sorted in ascending order by their prevalence of involvement in LNL II. Oral cavity subsites are plotted in shaed of blue, oropharynx subsites in shades of orange.

involvement was found or not, according to different diagnostic and pathologic modalities.
In figure 2, we have plotted the prevalence of involvement in the four LNLs I, II, III, and IV stratified by the primary tumor's subsite. We only included patients with tumors in the gum, floor of mouth, other/unspecified parts of the mouth, palate, oropharynx, tosil, base of tongue, and other/unspecified parts of tongue, resulting in 1242 patients.

3 Results

Works super well, of course!

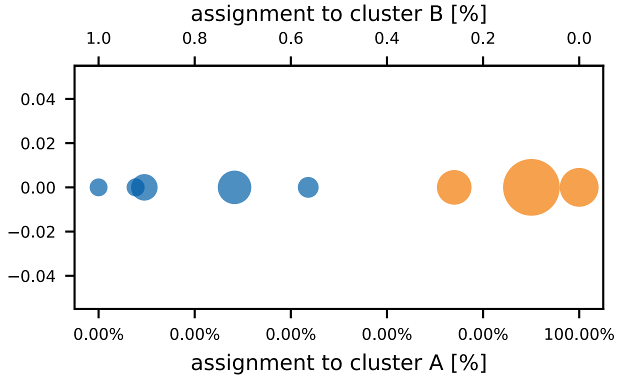

Figure 3: Assignment of each subsite to each of the two clusters. The fruther left a subsite, the more it is assigned to cluster A, the further right, the more to cluster B.

4 Discussion

All that's left is for other people to bring this into clinical practice.

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] R. Ludwig. "Modelling Lymphatic Metastatic Progression in Head and Neck Cancer". PhD thesis. Zurich: University of Zurich, 2023.

[12] 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.

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

[14] 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.