Probabilistic Model of Bilateral Lymphatic Spread in Head and Neck Cancer

Roman Ludwig[1,2*], Yoel Perez Haas[1,2] and Jan Unkelbach[1,2]

[1]Department of Physics, University of Zurich.

[2]Radiation Oncology, University Hospital Zurich.

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

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

Abstract

Purpose: According to current guidelines for elective nodal irradiation of oropharyngeal squamous cell carcinoma (OPSCC) patients, large parts of the contralateral lymphatic system are included in the elective clinical target volume (CTV-N), even for lateralized tumors without clinical lymph node involvement in the contralateral neck. In this work, we present a probabilistic model for bilateral lymphatic tumor progression in OPSCC to predict the personal risk for occult disease in any lymph node level (LNL), given the patient's clinical lymph node involvement, T-stage, and lateralization of the primary tumor.

Methods: We extend a hidden markov model for lymphatic tumor progression, which was previously developed for ipsilateral lymph node involvement, to the contralateral neck. The model represents each of the LNLs I, II, III, IV, V, and VII of both sides of the neck as a hidden binary random variable with states healthy and involved. LNLs are connected to the tumor and among each other via arcs that correspond to spread probabilities. These spread probability rates are learned via Markov chain Monte Carlo (MCMC) sampling from a dataset of 833 OPSCC patients.

Results: The model is able to describe the data on lymph node involvement well with a small number of interpretable parameters. Midline extension of the primary tumor is the main risk factor for contralateral involvement. In addition, the risk of occult metastases in contralateral lymph node level increases with more advanced T-stage and more severe ipsilateral involvement. The probability

of involvement in contralateral level III is very low if the upstream level II is clinically negative. Similarly, the probability of involvement in contralateral level IV is very low if the upstream level III is clinically negative.

Conclusions: The model may guide personalized volume reduction of the elective CTV-N. It suggests that the contralateral neck may be spared from elective irradiation for lateralized tumors not crossing the midline. In patients with primary tumors crossing the midline but clinically negative contralateral neck, the contralateral elective irradiation may be limited to level II.

Source: Article Notebook

Source: Article Notebook

Introduction

When treating head and neck squamous cell carcinomas (HNSCC) with radiotherapy or surgery, not only the primary tumor and the clinically detected lymph node metastases are targeted. Per current guidelines, large volumes of the neck are part of the the elective clinical target volume (CTV-N) [1–8]. This is to minimize the risk of regional recurrences due to untreated microscopic disease which cannot be detected by current in-vivo imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), or positron emission tomography (PET). However, minimizing the risk of missing occult disease in the lymph drainage region must be balanced against the toxicity related to unnecessary treatment of healthy tissue.

The aforementioned guidelines for the definition of the CTV-N are based on anatomically defined lymph node levels (LNL) [2] and the overall prevalence of lymph node metastases in the LNLs. Based on this, they recommend extensive irradiation of both sides of the neck for a majority of patients. However, this prevalence the guidelines are based on does not take into account an individual patient's state of tumor progression, which may have a great impact on a patient's personal risk for occult disease in a given LNL. For example, a patient with substantial nodal involvement in the ipsilateral side and an advanced primary tumor that crosses the mid-sagittal plane would receive extensive elective irradiation of the contralateral LNLs II, III, and IVa. However, if the patient presented with no clinically detectable nodal disease (cN0) and only a small, clearly lateralized T1 tumor, the same contralateral CTV-N would be recommended [5].

To better quantify this personalized risk of occult disease we have previously developed an intuitive, probabilistic hidden Markov model [9, 10] which was originally based on a conceptually similar Bayesian network model [11]. So far, these models were limited to the ipsilateral side of the neck. The main contribution of this work will be to extend the formalism to include risk predictions for the contralateral side as well. We believe this holds the potential to reduce the electively irradiated volumes in the

contralateral neck for patients whose individual risk for occult disease is low. In turn, this may substantially reduce the toxicity from irradiation and the associated decrease in quality of life.

The main contributions of this paper are as follows:

1. In section section 2 we present a multi-centric dataset on lymph node involvement in 833 OPSCC patients. Based on the dataset the main risk factors for contralateral lymph node involvement are identified and the requirements for the bilateral model extension are described (section 2).
2. In section 2 we present a bilateral HMM of lymphatic progression that accounts for the T-stage lateralization of the primary tumor as well as clinical involvement as risk factors for occult contralateral involvement. In section 2 we describe how the model training and the computational experiments were set up.
3. In the results in section 2, we demonstrate that the model is able to describe the patterns of contralateral lymph node involvement observed in the dataset, and we apply the model to estimating risk of occult disease for typical patients. In section 2, implications for volume-deescalated radiotherapy are discussed.

Data on Lymphatic Progression Patterns

To be able to create models for lymphatic tumor progression that include all relevant LNLs including the contralateral side, we have collected a detailed dataset of 833 patients with newly diagnosed oropharyngeal squamous cell carcinomas [12, 13]. It reports the lymph node involvement per LNL of every patient individually in tabular form, in addition to other primary tumor and patient characteristics such as T-category, subside, and lateralization of the primary tumor, and HPV p16 status. Their patient records have been collected at four different institutions and a brief overview over some of their patients' characteristics are shown in table 1. The data from the Inselspital Bern and the Centre Léon Bérard only consist of patients who received a neck dissection. Since this treatment is more commonly chosen for early T-category patients, they also make up a larger portion of the respective dataset. The majority of patients from the University Hospital Zürich and the Vall d'Hebron Barcelona Hospital was treated with definitive radiotherapy. For a subset of 83 patients in the CLB dataset, the lateralization of the primary tumor was not reported.

Source: Article Notebook

3

Table 1: Overview over the five datasets from four different institutions used to train and evaluate our model. Here, we briefly characterize the total number of OPSCC patients from the respective institution, their median age, what proportion received some form of neck dissection, the N0 portion of patients, what percentage presented with early T-category, and the prevalence of primary tumor midline extension. For a much more detailed look at the data, visit lyprox.org.

Table 1

Institution	Total	Age (median)	Neck Dissection	N0	Early T-Cat.	Mid.
Centre Léon Bérard	325	60	100%	19%	69%	18%
Inselspital Bern	74	61	100%	18%	66%	14%
University Hospital Zurich	287	66	26%	18%	52%	31%
Vall d'Hebron Barcelona Hospital	147	58	5%	21%	34%	34%

Source: Article Notebook

Consensus on Involvement Status

Pathological involvement is thus available only for a subset of the patients and only for the dissected levels. For the remaining patients, only clinical involvement is available. For the analysis performed in this paper, all available diagnostic information has been combined into a consensus decision representing the most likely state of involvement for each patient and LNL, using literature values for sensitivity and specificity of different modalities [14, 15]. These values are tabulated in table 2.

This consensus was formed as follows: Suppose the involvement status X_v of LNL v was assessed using different diagnostic modalities $\mathcal{O} = \{\mathrm{MRI}, \mathrm{CT}, \mathrm{pathology}, ...\}$, each characterized by their own pair of sensitivity and specificity values s_N^o and s_P^o, with $o \in \mathcal{O}$. Then we have $|\mathcal{O}|$ observations $z_v^o \in [0, 1]$, where 0 stands for "healthy" and 1 for "involved". We can then compute the most likely true involvement X_v using the likelihood function

$$\ell\left(X_v \mid \{z_v^o\}_{o \in \mathcal{O}}\right) = \prod_{o \in \mathcal{O}} (1 - X_v) \cdot [z_v^o \cdot (1 - s_P^o) + (1 - z_v^o) \cdot s_P^o]$$
$$+ X_v \cdot [z_v^o \cdot s_N^o + (1 - z_v^o) \cdot (1 - s_N^o)]$$

We now assume the true state X_v to take on the value 1 if $\ell\left(X_v = 1 \mid ...\right) > \ell\left(X_v = 0 \mid ...\right)$ and 0 otherwise. For example, if we have $z_{\mathrm{II}}^{\mathrm{CT}} = 0$ and $z_{\mathrm{II}}^{\mathrm{MRI}} = 1$ we would compute the following likelihoods:

$$\ell\left(X_{\mathrm{II}} = 1 \mid z_{\mathrm{II}}^{\mathrm{CT}} = 0, z_{\mathrm{II}}^{\mathrm{MRI}} = 1\right) = (1 - s_N^{\mathrm{CT}}) \cdot s_N^{\mathrm{MRI}} = 15.39\%$$
$$\ell\left(X_{\mathrm{II}} = 0 \mid z_{\mathrm{II}}^{\mathrm{CT}} = 0, z_{\mathrm{II}}^{\mathrm{MRI}} = 1\right) = s_P^{\mathrm{CT}} \cdot (1 - s_N^{\mathrm{MRI}}) = 14.44\%$$

4

In this example, we would thus assume the true state to be involved ($X_{\text{II}} = 1$).

This method of computing a consensus also ensures that the pathology reports always override any conflicting clinical diagnosis, due to pathology's high sensitivity and specificity.

Table 2: Specificity and sensitivity values from the literature [14, 15].

| Modality	Specificity	Sensitivity
CT	76%	81%
PET	86%	79%
MRI	63%	81%
FNA	98%	80%
pathology	100%	100%

Data Availability

The entire data, including additional patients with tumors in other primary locations than the oropharynx, is publicly available: It may be downloaded from LyProX where it can be interactively explored too, from GitHub, from zenodo, or via the *Data-in-Brief* publications Ludwig et al. [12] and Ludwig et al. [13]. The Data-in-Brief publications contain a detailed description of the datasets and the data format. However, these publications do not include the most recent dataset addition from Vall d'Hebron Barcelona Hospital.

Patterns of Contralateral Involvement

These datasets allow us to investigate correlations between the involvement of individual LNLs, or between risk factors and patterns of involvement. In figure 1, we have plotted the prevalence of each contralateral LNL's involvement, stratified by T-category, ipsilateral number of involved LNLs, and whether the tumor extended over the mid-sagittal line. A similar but more complete stratification is also tabulated in the appendix in table 4.

The left panel in figure 1 indicates that T-category is correlated with contralateral involvement (as it is with overal involvement). This is simply because T-category may on average be considered a surrogate for the time between onset of disease and diagnosis. I.e., a patient with a T4 tumor was – on average – diagnosed later than a patient with a T1 tumor. Thus, the former did have more time to develop metastases.

Similarly, ipsilateral involvement correlates with contralateral metastasis. The tumor of a patients with many metastases in ipsilateral LNLs was probably able to spread for longer (or faster) compared to a tumor in a patient with no nodal disease. This, too, may therefore be considered a surrogate for the duration of the disease. In addition, it has been hypothesized that bulky nodal disease ipsilaterally may also redirect lymph fluids to the contralateral side.
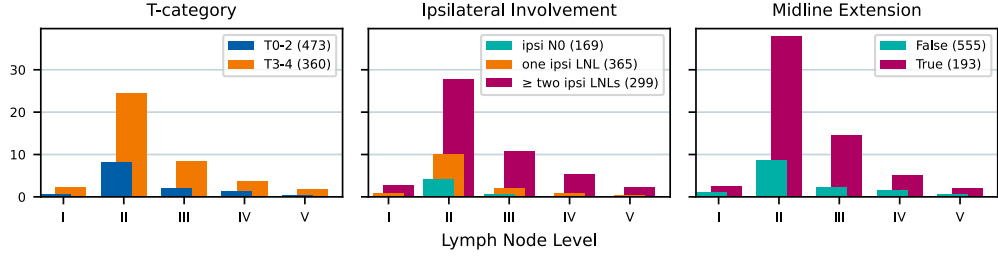
Figure 1: Contralateral involvement stratified by T-category (left panel), the number of metastatic LNLs ipsilaterally (center panel), and whether the primary tumor extended over the mid-sagittal line or was clearly lateralized (right panel).

Lastly, the right panel in figure 1 shows that patients with a tumor crossing the mid-sagittal line show contralateral involvement vastly more often compared to patients with clearly lateralized tumors. This makes intuitive sense, because the lymphatic system in the head and neck region is typically symmetric and thus no major vessels cross the midline. Therefore, interstitial fluids from the primary tumor – which we assume to carry living malignant cells – may only reach the blind-ended lymphatic vessels in the contralateral neck via short-ranged diffusion. Which in turn is only possible when the primary tumor is close enough to the mid-sagittal line or crosses it.

The three risk factors for contralateral involvement, midline extension, T-category, and ipsilateral involvement are correlated. For example, midline extension was present in 48% of advanced T-category tumors but only in 7% of early T-category tumors. The higher prevalence of contralateral metastases in advanced T-category tumors may thus be partially explained by a higher rate of midline extension in these tumors. However, figure 2 indicates that advanced T-category and severe ipsilateral involvement do represent additional risk factors . The figure considers only patients with lateralized tumors not extending over the midline. For early T-category and healthy ipsilateral levels I-V, only 1% (1 out of 84 patients) shows involvement of contralateral level II. This increases to 5% (9 out of 165 patients) if ipsilateral level II is involved, to 15% (9 out of 60 patients) if ipsilateral levels II and III are involved, and to 23% (6 out of 26 patients) for advanced T-category with ipsilateral levels II and III involved.

Requirements for a Bilateral Model

Based on the observations of the previous section, any potential model that aims to also predict the risk for contralateral nodal involvement, should be able to take the following into account:

1. More advanced T-category should lead to higher risk for nodal disease. One approach to achieve this via the expected time of diagnosis has already been developed in the form of a hidden Markov model [9].
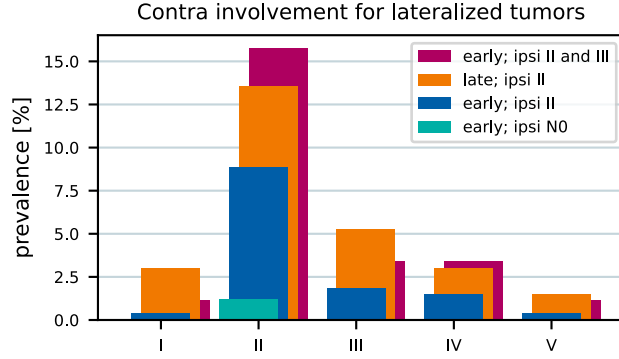
Figure 2: Contralateral involvement prevalence by LNL for lateralized tumors. Shown are four different scenarios: 1) early T-category with an ipsilateral N0 neck (green), 2) early T-category with ipsilaterally LNL II involved (blue), 3) late T-category with ipsilateral LNL II metastatic (orange), and 4) early T-category with ipsilateral both levels II and III involved (red). This illustrates that even when controlling for the tumor's lateralization, T-category and ipsilaeral involvement are predictive of contralateral metastases.

2. The degree of ipsilateral involvement should give the model information on the time that may have passed between onset and diagnosis of the disease. This should come in addition to what can be inferred about the time from T-category alone.
3. A tumor that extends over the mid-sagittal line should yield contralateral metastases with much higher probability.

Over the course of this work, we will first briefly recap the mentioned HMM in section 2, which was so far used to model ipsilateral lymphatic progression only. Then, we intuitively extend it to include the contralateral side as well in section 2. In this section, we also introduce a way of modelling the tumor's midline extension as a random variable (section 2) and lastly talk about how it may affect the contralateral spread in section 2.

Unilateral Model for Lymphatic Progression

This paper builds on the previously developed unilateral model for ipsilateral lymph node involvement presented in [cite 2024 graph extension paper]. In this section we provide a brief recap of the unilateral model to introduce the notation needed to extend the framework to a bilateral model describing both ipsilateral and contralateral lymph node involvement in section 2. For further details in the ipsilateral model, the reader is referred to the earlier publications [9, cite 2024 graph extension paper]

We model a patient's state of involvement at an abstract time-step t as a vector of hidden binary random variables:

$$\mathbf{X}[t] = (X_v[t]) \qquad v \in \{1, 2, ..., V\} \tag{1}$$

Here, V is the number of LNLs the model considers. The values a LNL's hidden binary random variables may take on are $X_v[t] = 0$ (False), meaning the LNL v is healthy or free of metastatic disease, or $X_v[t] = 1$ (True), corresponding to some form of tumor presence (i.e., occult or clinically detected). Since the state vector $\mathbf{X}[t]$ is V-dimensional and binary, there are 2^V distinct possible lymphatic involvement patterns, which we enumerate from $\xi_0 = (0\ 0\ \cdots\ 0)$ to $\xi_{2^V} = (1\ 1\ \cdots\ 1)$. In addition, each LNL is associated with an observed binary random variable Z_v that describes the clinical involvement of a LNL based on imaging: $Z_v = 0$ (False) indicates that the LNL v is healthy based on clinical diagnosis, and $Z_v = 1$ (True) indicates that suspicious lymph nodes were detected that are deemed metastatic.

Any hidden Markov model is fully described by three quantities:

1. A starting state $\mathbf{X}[t = 0]$ at time $t = 0$ just before the patient's tumor formed. In our case, this is always the state where all LNLs are still healthy ξ_0.
2. The *transition matrix*

$$\mathbf{A} = \left(A_{ij} \right) = \left(P\left(\mathbf{X}[t+1] = \xi_j \mid \mathbf{X}[t] = \xi_i \right) \right) \tag{2}$$

   where the value at row i and column j represents the probability to transition from state ξ_i to ξ_j during the time-step from t to $t+1$. Note that we prohibit self-healing, meaning that during a transition, no LNL may change their state from $X_v[t] = 1$ to $X_v[t+1] = 0$. Consequently, many elements of the transition matrix are zero.
3. Lastly, the *observation matrix*

$$\mathbf{B} = \left(B_{ij} \right) = \left(P\left(\mathbf{Z} = \zeta_j \mid \mathbf{X}[t_D] = \xi_i \right) \right) \tag{3}$$

   where in row i and at column j we find the probability to *observe* a lymphatic involvement pattern $\mathbf{Z} = \zeta_j$, given that the true (but hidden) state of involvement at the time of diagnosis t_D is $\mathbf{X}[t_D] = \xi_i$.

The transition matrix $\mathbf{A}$ is parametrized using a directed acyclic graph (DAG) as an abstract representation of the underlying lymphatic network. Directed arcs from the primary tumor to a LNL are associated with a probability b_v for direct spread to LNL v during one time step. Directed arcs from a LNL v to a LNL r are associated with a probability t_{vr} for the tumor to progress from one LNL to the next. In this paper, we build on the DAG shown in figure 3 which was obtained by maximizing the model evidence as described in [cite 2024 graph extension paper].

Using the introduced quantities, we can evolve the distribution of all possible hidden states from $\mathbf{X}[t = 0] = \xi_0$ step by step, by successively multiplying this vector with the transition matrix $\mathbf{A}$. For later use, we define at this point the matrices of the dsitributions over all hidden states, given all time-steps Λ:
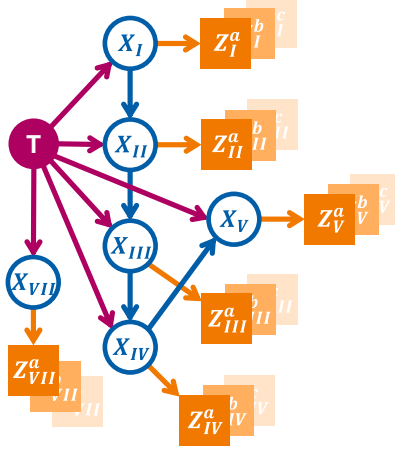
Figure 3: Directed acyclic graph (DAG) representing the abstract lymphatic network in the head and neck region. Blue nodes are the LNLs' hidden random variables, the red node represents the tumor, and the orange square nodes depict the binary observed variables. Red and blue arcs symbolize the probability of lymphatic spread along that edge during one time-step. The orange arcs represent the sensitivity and specificity of the observational modality (e.g. CT, MRI, pathology, …).

$$\Lambda = P\left(\mathbf{X} \mid \mathbf{t}\right) = \begin{pmatrix} \pi^\top \cdot \mathbf{A}^0 \\ \pi^\top \cdot \mathbf{A}^1 \\ \vdots \\ \pi^\top \cdot \mathbf{A}^{t_{\max}} \end{pmatrix} \tag{4}$$

Where the k-th row in this matrix corresponds to the probability distribution over hidden states after $t = k - 1$ time-steps.

At the time of the diagnosis t_D, we multiply the result with the observation matrix $\mathbf{B}$. We may then look up the likelihood of a patient presenting with the diagnosis $\mathbf{Z} = \zeta_i$ in the i-th entry of the final result. However, the remaining issue is that the value of t_D is unknown, i.e. over how many time-steps the HMM should be evolved. We solve this problem by marginalizing over the time of diagnosis. Different distributions over the diagnosis times can then be choosen based on T-category. For instance, the mean of the time-prior to marginalize over the diagnosis time for early T-category patients $P\left(t_D \mid \text{early}\right)$ may be shifted towards earlier times than the one for advanced T-category patients $P\left(t_D \mid \text{early}\right)$. This gives us for example

$$P\left(\mathbf{X} \mid \mathrm{T}x = \text{early}\right) = \sum_{t=0}^{t_{\max}} P\left(\mathbf{X} \mid t\right) \cdot P(t \mid \text{early})$$

In this work, we use binomial distributions $\mathfrak{B}\left(t_D, p_{\mathrm{T}x}\right)$ as time-priors which have one free parameter $p_{\mathrm{T}x}$ for each group of patients we differentiate based on T-category.

9

Also, we fix $t_{\max} = 10$, which means that the expected number of time-steps from the onset of a patient's disease to their diagnosis is $\mathbb{E}[t_D] = 10 \cdot p_{\mathrm{T}x}$.

Likelihood Function

With the formalism introduced above, we can write the likelihood function for a patient to present with a diagnosis consisting of an observed state and a T-category $d = (\zeta_i, \mathrm{T}x)$ as follows:

$$\ell = P(\mathbf{Z} = \zeta_i \mid \mathrm{T}x) = \sum_{t=0}^{t_{\max}} \left[\xi_0 \cdot \mathbf{A}^t \cdot \mathbf{B}\right]_i \cdot P(t \mid \mathrm{T}x) \tag{5}$$

Above, the quantity inside $[...]_i$ denotes the i-th component of the vector that is the result of the vector and matrix multiplications in the square brackets. Note that it is also possible to account for missing involvement information: If a diagnosis (like fine needle aspiration (FNA)) is only available for a subset of all LNLs, we can sum over all those possible complete observed states ζ_j that match the provided diagnosis.

The single-patient likelihood ℓ in equation 5 depends on the spread parameters shown in figure 3 via the transition matrix $\mathbf{A}$ and on the binomial parameters $p_{\mathrm{T}x}$ via time-priors. In this work, we will only differentiate between "early" (T1 & T2) and "advanced" (T3 & T4) T-categories. Therefore, the parameter space of the unilateral model is:

$$\theta = \left(\{b_v\}, \{t_{vr}\}, p_{\text{early}}, p_{\text{adv.}}\right) \quad \text{with} \quad \begin{smallmatrix} v \leq V \\ r \in \mathrm{pa}(v) \end{smallmatrix} \tag{6}$$

And it is our goal to infer the values of these parameters for a given dataset $\mathcal{D} = (d_1, d_2, \ldots, d_N)$ of OPSCC patients. The likelihood of these N diagnoses is simply the product of their individual likelihoods as defined in equation 5. For numerical reasons, we typically compute the data likelihood in log space:

$$\log \mathcal{L}(\mathcal{D} \mid \theta) = \sum_{i=1}^{N} \log \ell_i \tag{7}$$

The methodology we use to infer the model's parameters is detailed in section 2.

Extension to a Bilateral Model

A naive approach to model the contralateral lymphatic spread would be to simply employ two independent unilateral models as introduced in section 2. During training, one could enforce that some parameters are shared between these two models, e.g. the parameterization of the distributions over diagnose times or the spread among the LNLs ($t_{vr}). However, this approach lacks a way to describe the

10

correlation between ipsi- and contralateral involvement discussed in section {#sec-data-strat}. Two independent unilateral models would not describe the observation that contralateral involvement becomes more likely with more severe ipsilateral involvement.

Thus, we extend the formalism in section 2 in such a way that the model's ipsi- and contralateral side evolve synchronously over time. To achieve that, we start by writing down the posterior distribution of involvement an analogy to equation 15, which is now a joint probability of an involvement $\mathbf{X}^\text{i}$ ipsilaterally *and* an involvement $\mathbf{X}^\text{c}$ contralaterally, given a diagnosis of the ipsilateral LNLs $\mathbf{Z}^\text{i}$ and of the contralateral ones $\mathbf{Z}^\text{c}$:

$$P\left(\mathbf{X}^\text{i}, \mathbf{X}^\text{c} \mid \mathbf{Z}^\text{i}, \mathbf{Z}^\text{c}\right) = \frac{P\left(\mathbf{Z}^\text{i}, \mathbf{Z}^\text{c} \mid \mathbf{X}^\text{i}, \mathbf{X}^\text{c}\right) P\left(\mathbf{X}^\text{i}, \mathbf{X}^\text{c}\right)}{P\left(\mathbf{Z}^\text{i}, \mathbf{Z}^\text{c}\right)} \tag{8}$$

For the sake of brevity, we omit the dependency on the parameters and the T-category here.

The probability of the diagnoses given a hidden state factorises: $P\left(\mathbf{Z}^\text{i}, \mathbf{Z}^\text{c} \mid \mathbf{X}^\text{i}, \mathbf{X}^\text{c}\right) = P\left(\mathbf{Z}^\text{i} \mid \mathbf{X}^\text{i}\right) \cdot P\left(\mathbf{Z}^\text{c} \mid \mathbf{X}^\text{c}\right)$, and the two factors are described through observation matrices $\mathbf{B}^\text{i}$ and $\mathbf{B}^\text{c}$.

The term representing the model's prior probability of hidden involvement does not factorize. However, we assume that there is no direct lymph drainage from an ipsilateral LNL to a contralateral LNL. No major lymph vessels cross the mid-sagittal plane. In the graphical model, this means that we assume no directed arcs between ipsilateral and contralateral LNLs, i.e. tumor spread to the contralateral side is due to the primary tumor alone. We can thus write the joint probability $P\left(\mathbf{X}^\text{i}, \mathbf{X}^\text{c}\right)$ as a factorising sum:

$$\begin{aligned} P\left(\mathbf{X}^\text{i}, \mathbf{X}^\text{c}\right) &= \sum_{t=0}^{t_\text{max}} P(t) \cdot P\left(\mathbf{X}^\text{i}, \mathbf{X}^\text{c} \mid t\right) \\ &= \sum_{t=0}^{t_\text{max}} P(t) \cdot P\left(\mathbf{X}^\text{i} \mid t\right) \cdot P\left(\mathbf{X}^\text{c} \mid t\right) \end{aligned} \tag{9}$$

This assumption makes intuitive sense: The state of the ipsilateral and contralateral sides of the lymphatic network evolves independently over time as there are no major lymph vessels crossing the midline. However, both sides are coupled via time. Qualitatively speaking, according to the model in {#eq-bilateral-marginal}, a joint state with severe contralateral involvement and limited ipsilateral involvement would be unlikely, because severe contralateral involvement becomes likely only at later time steps when limited ipsilateral involvement becomes unlikely.

Using {#eq-bilateral-marginal} along with equation 4, we can write the above distribution algebraically as a product:

11

$$P\left(\mathbf{X}^{\mathrm{i}} = \xi_n, \mathbf{X}^{\mathrm{c}} = \xi_m\right) = \left[\Lambda_{\mathrm{i}}^{\top} \cdot \operatorname{diag} P(\mathbf{t}) \cdot \Lambda_{\mathrm{c}}\right]_{n,m} \tag{10}$$

Parameter Symmetries

In general, the matrices Λ_{i} and Λ_{c} could be parameterized using a disjoint set of parameters. I.e., the ipsi- and contralateral spread rates are entirely different. However, using three sensible assumptions, we can reduce the parameter space by sharing some parameters between the sides:

1. We assume that both ipsilateral and contralateral spread is described through the same graph shown in {#fig-full-graph}.
2. We assume the spread *among* the LNLs to be same on both sides. It is reasonable to assume the lymphatic system is symmetric. Thus, the spread rates from one LNL to the other should be symmetric, too. Formally, this means

$$\begin{aligned} b_v^{\mathrm{c}} &\neq b_v^{\mathrm{i}} \\ t_{rv}^{\mathrm{c}} &= t_{rv}^{\mathrm{i}} \end{aligned} \tag{11}$$

   for all $v \leq V$ and $r \in \mathrm{pa}(v)$.
3. The tumor's spread to the contralateral side in case of an extension over the midline is larger than if it was clearly lateralized, but smaller than its spread to the ipsilateral side. This assumption stems from a simple thought experiment: Consider moving the tumor from a clearly lateralized position accross the mid-sagittal plane to the same position, but on the contralateral side. In the beginning we would have $b_v^{\mathrm{c}} < b_v^{\mathrm{i}}$, while in the end, the situation is reversed. If a tumor extends over the mid-sagittal line, its contralateral spread rate can be expected to be in between these two extremes. We encode this in a *mixing parameter* $\alpha \in [0,1]$ that captures a "degree of asymmetry":

$$b_v^{\mathrm{c},\epsilon=\texttt{True}} = \alpha \cdot b_v^{\mathrm{i}} + (1 - \alpha) \cdot b_v^{\mathrm{c},\epsilon=\texttt{False}} \tag{12}$$

   This means the model now uses three different sets of parameters to describe the spread from the tumor to the LNLs: b_v^{i} for the spread to the ipsilateral LNLs, $b_v^{\mathrm{c},\epsilon=\texttt{False}}$ for the spread to the contralateral LNLs as long as the tumor is clearly lateralized, and finally $b_v^{\mathrm{c},\epsilon=\texttt{True}}$ when it crosses the midline. Note, however, that these three sets of spread rates only account for $2 \cdot 2^V + 1$ parameters, since they are coupled via the mixing parameter α.

Our parameter space has now expanded to

$$\theta = \left(\{b_v^{\mathrm{i}}\}, \{b_v^{\mathrm{c}}\}, \alpha, \{t_{vr}\}, p_{\mathrm{early}}, p_{\mathrm{adv.}}\right) \quad \text{with} \quad \substack{v \leq V \\ r \in \mathrm{pa}(v)} \tag{13}$$

which is less than double in size, compared to the unilateral model. From these parameters, we now build three different transition matrices: The unchanged matrix $\mathbf{A}_{\mathrm{i}}$ for the ipsilateral side, one $\mathbf{A}_{\mathrm{c}}^{\epsilon=\texttt{False}}$ for the contralateral side that covers the progression

as long as the tumor is lateralized, and $\mathbf{A}_{\mathrm{c}}^{\epsilon=\mathrm{True}}$ in case of a tumor that has crossed the mid-sagittal line.

Modelling Midline Extension

It can be assumed that most tumors that cross the midline at the time of diagnose have started as lateralized tumors and grew over the midline at a later time step. Hence, the transition matrix $\mathbf{A}_{\mathrm{c}}^{\epsilon=\mathrm{True}}$ only applies for a subset of the time steps. Therefore, we also model the tumor's extension over the mid-sagittal line as a binary random variable. A tumor starts lateralized and at every time-step there is a finite probability p_ϵ that the tumor grows over the midsaggital plane. The overall probabilities to find a patient with a clearly lateralized tumor or one that extends over the mid-sagittal line after t time-steps are then respectively given by

$$P(\epsilon = \mathtt{False} \mid t) = (1 - p_\epsilon)^t$$
$$P(\epsilon = \mathtt{True} \mid t) = 1 - P(\epsilon = \mathtt{False} \mid t)$$

Using this, it is straightforward to write down the matrix of state distributions for all time-steps, as in equation 4 covering the contralateral hidden state evolution:

$$\Lambda_{\mathrm{c}}^{\epsilon=\mathtt{False}} = \begin{pmatrix} \pi^\top \cdot \left(\mathbf{A}_{\mathrm{c}}^{\epsilon=\mathtt{False}}\right)^0 \\ (1 - p_\epsilon) \cdot \pi^\top \cdot \left(\mathbf{A}_{\mathrm{c}}^{\epsilon=\mathtt{False}}\right)^1 \\ \vdots \\ (1 - p_\epsilon)^{t_{\max}} \cdot \pi^\top \cdot \left(\mathbf{A}_{\mathrm{c}}^{\epsilon=\mathtt{False}}\right)^{t_{\max}} \end{pmatrix}$$

where we used the transition matrix $\mathbf{A}_{\mathrm{c}}^{\epsilon=\mathtt{False}}$ that depends on the base spread parameters $b_v^{\mathrm{c},\epsilon=\mathtt{False}}$.

The case when midline extension is eventually present is more complicated: We already marginalize over the exact time-step when the tumor grows over the mid-sagittal line. But whenever that happens, we also need to change the contralateral transition matrix to use the increased spread rates $b_v^{\mathrm{c},\epsilon=\mathtt{True}}$ from the tumor to the contralateral LNLs, given by the linear mixing in equation 12. We can achieve the correct marginalization by iteratively building the joint distribution $P\left(\mathbf{X}^{\mathrm{c}}, \epsilon = \mathtt{True} \mid t\right)$. We start from $t = 0$, where we know that all LNLs are healthy (the contralateral neck is in the state $\mathbf{X}_{\mathrm{c}} = \xi_0$) and the tumor is lateralized ($\epsilon = \mathtt{False}$):

$$P\left(\mathbf{X}^{\mathrm{c}} = \xi_0, \epsilon = \mathtt{False} \mid t = 0\right) = 1$$

whereas all other states have zero probability. Then we consider an arbitrary later time-step $t = \tau + 1$. There are two possible scenarios we need to marginalize over:

1. The tumor was still lateralized at $t = \tau$ and just grew over the midline. Under this scenario, the probability for having a midline extension at $t = \tau + 1$ is given by p_ϵ.

We use this to weight the contralateral state distribution that was so far evolved without increased contralateral spread.

2. The mid-sagittal line was already crossed by the tumor. In that case, the probability is 1 to remain in that lateralization state. Thus, for this case we simply add the same distribution we are computing, but from one time-step earlier, arriving at a recursion relation:

$$P\left(\mathbf{X}^c, \epsilon = \texttt{True} \mid \tau + 1\right) = \left[p_\epsilon P\left(\mathbf{X}^c, \epsilon = \texttt{False} \mid \tau\right) + P\left(\mathbf{X}^c, \epsilon = \texttt{True} \mid \tau\right)\right]^\top \cdot \mathbf{A}_c^{\epsilon=\texttt{True}}$$

We can collect the iteratively computed distributions for the midline extension case to define the matrix over the states given all time-steps, in analogy to equation 4:

$$\Lambda_c^{\epsilon=\texttt{True}} = \begin{pmatrix} P\left(\mathbf{X}^c, \epsilon = \texttt{True} \mid 0\right) \\ P\left(\mathbf{X}^c, \epsilon = \texttt{True} \mid 1\right) \\ \vdots \\ P\left(\mathbf{X}^c, \epsilon = \texttt{True} \mid t_{\max}\right) \end{pmatrix}$$

Using this, we can again write the joint of ipsi- and contralateral involvement - now also for the case of mid-sagittal extension - algebraically as before in equation 10:

$$P\left(\mathbf{X}^i = \xi_n, \mathbf{X}^c = \xi_m, \epsilon\right) = \left[\Lambda_i^\top \cdot \text{diag}\, P(\mathbf{t}) \cdot \Lambda_c^\epsilon\right]_{n,m}$$

We can use this term to compute the likelihood of all patients with and without midline extension separately. And if for some patients the information of tumor lateralization is not available, we can simply sum the above term once for $\epsilon = \texttt{False}$ and once for $\epsilon = \texttt{True}$ to marginalize over the unknown variable.

The final parameter space of our extended model has now reached this size:

$$\theta = \left(\{b_v^i\}, \{b_v^c\}, \alpha, \{t_{vr}\}, p_{\text{early}}, p_{\text{adv.}}, p_\epsilon\right) \quad \text{with} \quad \begin{smallmatrix} v \leq V \\ r \in \text{pa}(v) \end{smallmatrix} \tag{14}$$

Model Prediction in the Bayesian Context

Our stated goal is to compute the risk for a patient's true ipsi- and contralateral nodal involvement states $\mathbf{X}^i$ and $\mathbf{X}^c$, *given* their individual diagnosis $d = \left(\zeta_k^i, \zeta_\ell^c, \epsilon, \text{T}x\right)$. Here, this diagnosis consists of the observed ipsi- and contralateral nodal involvements, the patient's midline extension ϵ, and their tumor's T-category $\text{T}x$. Using Bayes' law, we can write this risk as:

$$P(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid d, \widehat{\theta}) = \frac{P\left(\zeta_k^{\mathrm{i}} \mid \mathbf{X}^{\mathrm{i}}\right) P\left(\zeta_\ell^{\mathrm{c}} \mid \mathbf{X}^{\mathrm{c}}\right) P(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}, \epsilon \mid \widehat{\theta}, \mathrm{T}x)}{\sum_{i=0}^{2^V} \sum_{j=0}^{2^V} P\left(\zeta_k^{\mathrm{i}} \mid \mathbf{X}^{\mathrm{i}} = \xi_i^{\mathrm{i}}\right) P(\zeta_\ell^{\mathrm{c}} \mid \mathbf{X}^{\mathrm{c}} = \xi_j^{\mathrm{c}}) P(\mathbf{X}^{\mathrm{i}} = \xi_i^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} = \xi_j^{\mathrm{c}}, \epsilon \mid \widehat{\theta}, \mathrm{T}x)}$$

(15)

The terms $P\left(\zeta_k^{\mathrm{i}} \mid \mathbf{X}^{\mathrm{i}}\right)$ and $P\left(\zeta_\ell^{\mathrm{c}} \mid \mathbf{X}^{\mathrm{c}}\right)$ are defined solely by sensitivity and specificity of the diagnostic modality. Terms like these already appeared in the definition of the observation matrx in equation 3. The *prior* $P(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}, \epsilon \mid \widehat{\theta}, \mathrm{T}x)$ in the above equation is the crucial term that is supplied by a trained model and its parameters $\widehat{\theta}$.

It is possible to compute this *posterior* probability of true involvement not only for one fully defined state $(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}})$, but also for e.g. individual LNLs: For example, the risk for involvement in the contralateral level IV would be a marginalization over all ipsilateral states ξ_i^{i} and all contralateral states ξ_j^{c} where $\xi_{j4}^{\mathrm{c}} = 1$. Formally:

$$P(\mathrm{cIV} \mid \mathbf{Z}^{\mathrm{i}} = \zeta_k^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}} = \zeta_\ell^{\mathrm{c}}, \widehat{\theta}, \mathrm{T}x) = \sum_k \sum_{\ell \,:\, \xi_{\ell 4} = 1} P(\mathbf{X}^{\mathrm{i}} = \xi_k^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} = \xi_\ell^{\mathrm{c}} \mid \zeta_k^{\mathrm{i}}, \zeta_\ell^{\mathrm{c}}, \epsilon, \widehat{\theta}, \mathrm{T}x)$$

(16)

Computational methods

In this section, we detail how the experiments were performed. Every figure, table, and result is fully reproducible via the GitHub repository `rmnldwg/bilateral-paper`. It also contains the raw manuscript and instructions on how to recreate all figures, tables, and the final document.

Involvement Data Consensus

It is possible to provide our model with multiple different diagnostic modalities, each being characterized by different pairs of sensitivity and specificity. However, we instead chose to combine them into a single "consensus" diagnosis before parameter inference. We opted for this because the literature values of sensitivity and specificity [14, 15] of imaging modalities like MRI and CT do not plausibly match some of our observations: In the USZ cohort, 78% of OPSCC patients where diagnosed with ipsilateral LNL II involvement via diagnostic imaging. This is virtually impossible with sensitivities around 80% and specificities lower than 100%.

Our pre-training consensus was formed by considering all reported diagnostic information for a particular patient and LNL. When conflicts arose, we computed the *most likely* true state of involvement using the literature sensitivity and specificity values [14, 15].

MCMC Sampling

For parameter inference, we used the Python package `emcee` [16]. It implements efficient MCMC sampling algorithms that employ multiple parallel samplers for affine invariance and better performance on multi-core CPUs. The sample proposal algirothms used by us are based on differential evolution moves [17, 18]. The `emcee` library was provided with the likelihood implemented by our `lymph-model` Python package.

For each dimension in the parameter space of the model, we initialized 12 of these parallel samplers, called "walkers", with random values in the unit cube. Every time all of these walkers advanced 50 steps, the autocorrelation time of the chains was estimated. For short chains, this estimate is not trustworthy, but stabilizes for longer chains. We therefore considered a sampling to be converged when two criteria were met:

1. The change in the autocorrelation time was less then 5.0e-2.
2. The estimate of the autocorrelation dropped below $n / 50$ where n is the length of the chain up to that point.

All samples up to this convergence - called the *burn-in phase* - were discarded. We only kept another 10 samples after that, which were spaced 10 steps apart.

First, in figure 4, we verify the sampling converged successfully by inspecting two monitoring quantities: The autocorrelation time of the MCMC chain and the acceptance fractions of the parallel walkers.
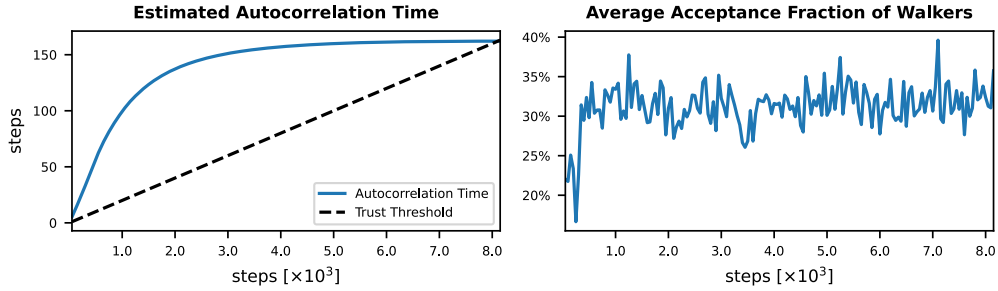
Figure 4: Monitoring quantities during the burn-in phase of the parameter sampling. Left: The autocorrelation time of the sampling chain estimated at different sampling steps. We consider the chain converged when the estimate of the autocorrelation time is stable and drops below the trust threshold of $n/50$ where n is the number of steps. Right: Fraction of accepted MCMC proposals averaged over all parallel walkers. Values around 30% indicate good mixing of the walkers.

Computing the Observed and Predicted Prevalence of Involvement Patterns

We want to assess the model's capability to approximate the distribution of lymphatic involvement patterns seen in the data. To that end, we compare the prevalence of some invovlement patterns under selected scenarios with the model's prediction for how often these involvements it expects to see, given these scenarios.

In this context, a "scenario" includes the patient's T-category $\mathrm{T}x$ and whether the patient's tumor extended over the mid-sagittal line, i.e. $\epsilon = \texttt{True}$ or $\epsilon = \texttt{False}$.

An involvement pattern specifies for each ipsi- and contralateral LNL whether it is "healthy", "involved", or "masked". If it is "masked", we essentially state that we are not interested in the involvement of that LNL and the prevalence will be marginalized over this LNL's involvement.

For example, we may be interested in the prevalence of contralateral LNL II involvement (i.e., contra LNL II "involved" and all other LNLs "masked") under the scenario of early T-category (T0-T2) and no midline extension ($\epsilon = \texttt{False}$). To compute this prevalence in the data, we select all patients of this scenario (in our data, this amounts to 379 patients). Of those, 28 were found to harbor metastases in their contralateral LNL II. Therefore, the prevalence is 7%.

When displaying this data prevalence, we often choose to draw a *beta posterior* over the "true" prevalence, hinting at the fact that our data merely represents a limited sample. The beta posterior follows from a uniform beta distribution as prior and a binomial likelihood for the number of patients with the involvement of interest, given the parameer for the "true" prevalence. The resulting distribution has its maximum at the observed prevalence, but in addition gives a visual intuition for the variance of of the observed quantity. I.e., when we observe 3 out of 10 events, the beta posterior is much wider than if we observe 300 out of 1000 for he same prevalence. It also allows us to check not only if the model is accurate, but also whether it reflects the uncertainty contained in the data.

Predicting the prevalence using our model amounts to computing the following probability:

$$P\left(\mathrm{II}^{\mathrm{c}} \mid \epsilon = \texttt{False}, \mathrm{T}x = \text{early}\right) = \frac{P\left(\mathrm{II}^{\mathrm{c}}, \epsilon = \texttt{False} \mid \mathrm{T}x = \text{early}\right)}{P\left(\epsilon = \texttt{False} \mid \mathrm{T}x = \text{early}\right)}$$

In the enumator, we marginalize over all ipsi- and contralateral LNLs' involvements, except for LNL II contralaterally. This is similar to the marginalization in equation 16, although we are summing over different quantities. In the denominator, we can simply insert the joint distribution over midline extension and diagnose time $P\left(\epsilon, t\right)$ marginalized over t using the early T-category's time-prior.

Since we compare it to the data, which does not report true but only observed involvement – although pathologically investigated LNLs may be as close as possible to the

ground truth – we do not consider posteriors of the form $P(\mathbf{X} \mid \mathbf{Z})$ here. Instead, we compute probabilities of observed involvement $P(\mathbf{Z})$, as in the likelihood equation 5.

When plotted, we usually display histograms over the model's predictions. Each of their values was computed from a different parameter set drawn during MCMC sampling, effectively giving us a distribution over the prevalences. Ideally, the histograms approximate the location and width of the Beta posteriors when attempting to describe the data they were trained on.

Note that we decided to omit the y-axis ticks and labels in these figures over prevalences and risks. The y-axis in these plots measures the probability density and its numerical values are not intuitively interpretable. Instead, we occasionally use the freed space to label e.g. rows of subplots.

Results: Model evaluation

In table 3, we tabulate the mean and standard deviation of the sampled parameters for the full midline model. Considering the parameters of the bilateral model that describe the ipsilateral spread (b_v^i, t_{vr} and $p_{\text{adv.}}$), we note that the bilateral model mostly reproduces the parameter values reported in the earlier publication [10] for the unilateral model. The small differences may be mostly due to the differences in training dataset. Therefore, the analysis of the models capability to adequately describe ipsilateral lymph node involvement is not repeated here and we focus instead on the analysis of contralateral involvement. We can further note that the contralateral spread parameters b_v^c are small compared to the ipsilateral parameters b_v^i, reflecting the low prevalence of contralateral lymph node involvement for lateralized tumors.

Source: Article Notebook

Table 3: Mean sampled parameter estimates of the midline model and the respective standard deviation.

Table 3

Parameter	Mean	Std. Dev.
Mid. ext. probability	8.16%	± 0.48%
ipsi: T I	2.80%	± 0.26%
ipsi: T II	34.89%	± 1.40%
ipsi: T III	5.45%	± 0.66%
ipsi: T IV	0.94%	± 0.18%
ipsi: T V	1.83%	± 0.22%
ipsi: T VII	2.32%	± 0.26%
contra: T I	0.29%	± 0.09%
contra: T II	2.46%	± 0.29%
contra: T III	0.14%	± 0.07%
contra: T IV	0.19%	± 0.08%

| Parameter	Mean	Std. Dev.
contra: T V	0.05%	± 0.04%
contra: T VII	0.50%	± 0.17%
Mixing	33.87%	± 4.32%
I II	62.50%	± 16.69%
II III	14.23%	± 1.64%
III IV	15.86%	± 1.93%
IV V	14.58%	± 3.76%
late T-cat. binom. prob.	44.98%	± 1.99%

Source:

Illustration of the model

In this subsection we illustrate parts of the mathematical framework defined in the previous section. In figure 5 (top panel), we visualize the prior distribution over diagnose times $P(t)$. By parameter choice, early T-category tumors are diagnosed on average after 3 time steps. Advanced T-category tumors are diagnosed on average after 4.5 time steps ($p_{\mathrm{adv.}} = 0.45$) , a parameter learned from the data to describe the higher involvement for advanced T-category tumors. The probability per time step for the tumor to grow over the midline, $p_\epsilon = 0.0816$ yields the probability of midline extension $P(\epsilon \mid t)$ at a given time step t (red line). The parameter is adjusted such that the model yields the correct proportion of patients with tumors extending over the midline seen in the dataset. The bottom panel in figure 5 shows the joint probability $P(\epsilon, t)$ to be diagnosed at time step t with a given state of midline extension and T-category.

The model introduced in the previous section defines a model of the joint probability distribution over midline extension and ipsi- and contralateral lymph node involvement, $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}, \epsilon\right)$. How this quantity is computed is illustrated in figure 6, visually representing equation 10. To keep it manageable to interpret, this example only considers the LNLs II, III, and IV ipsi- and contralaterally, resulting in $2^3 = 16$ distinct states per side and thus $2 \times 8 \times 8 = 128$ total states. LNLs I, V, and VII and the associated spread parameters have been removed from the model while the remaining parameters are set to the mean values in table xxx. figure 6 shows the joint distribution as two separate heatmaps for the two midline extension states for early/advanced??? T-category tumors. The most likely state is a lateralized tumor with involvement of ipsilateral level II and no contralateral involvement, which has a probability of approximately 25%. The second most probable state is a lateralized tumor with ipsilateral involvement of levels II and III and no contralateral involvement. The most likely state with contralateral involvement are tumors with midline extension with involvement of contralateral level II and ipsilateral level II and III.
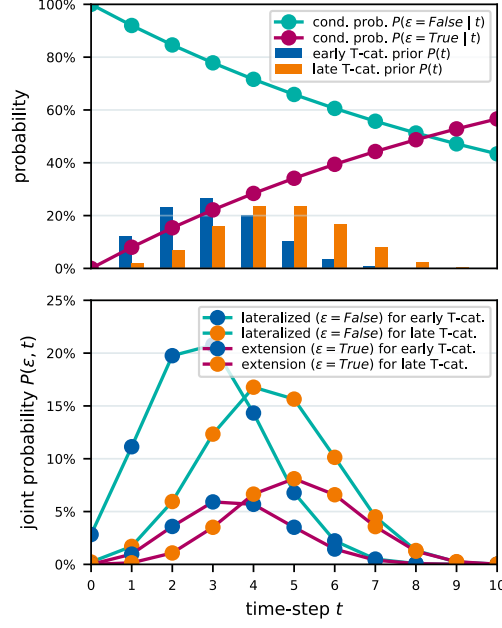
Figure 5: The top panel shows the prior probability to get diagnosed at time-step t for early and late T-category tumors as bars. Also in the top panel, we plot the conditional probability of the tumor's midline extension ($\epsilon = \texttt{True}$), given the time-step t as a line plot. In the bottom panel, we show the joint probability of getting diagnosed in time-step t *and* having a tumor that crosses the midline.
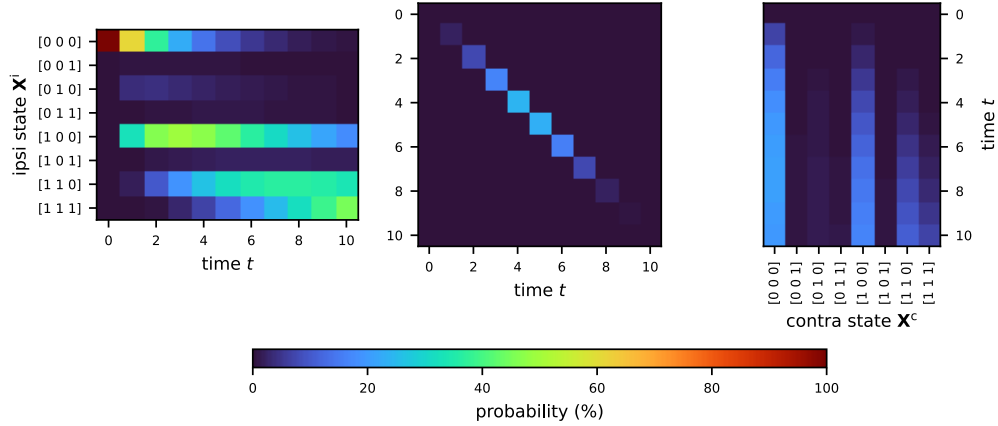

Figure 6: Visual representation of equation 10 for the case of midline extention. The left and right matrix represent the evolution of the possible hidden states for the ipsi- and contralateral neck respectively. In particular, the right matrix describes the evolution of the contralateral hidden states *and* midline extension. In the center, the time-prior is plotted as a diagonal matrix. The result of this matrix computation is the joint distribution $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}, \epsilon = \texttt{True}\right)$

20

Prevalence predictions for contralateral involvement

The bilateral model was designed to fulfil the requirements laid out in section 2. Now, we investigate to what extent the model can quantitatively describe the patterns of lymph node involvement observed in the dataset. To that end, we compare the model's predictions for contralateral involvement against observations in the data. This is done given scenarios that differ in T-category and/or midline extension and/or ipsilateral involvement.

Dependence of Contralateral Involvement on T-Category and Midline Extension

In figure 7, we plot the prevalence of contralateral involvement of the LNLs II, III, and IV for the four scenarios made up of the possible combinations of early and late T-category, as well as lateralized and midline extending tumors.
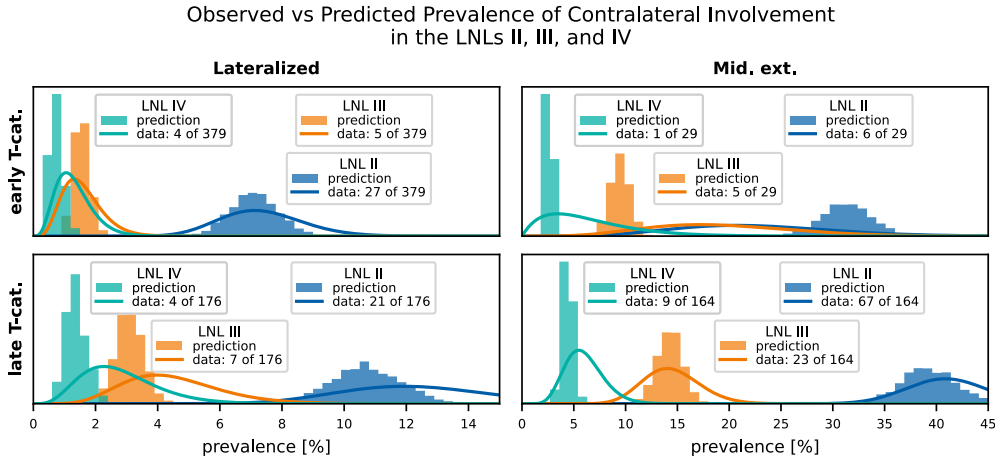

Figure 7: Comparison of predicted (histograms) vs observed (beta posteriors) prevalences. Shown for the contralateral LNLs II (blue), III (orange), and IV (green). The top row shows scenarios with early T-category tumors, the bottom row for late T-category ones. The left column depicts scenarios where the primary tumor is clearly lateralized, the right column scenarios of tumors extending over the mid-sagittal line. This figure illustrates the model's ability to describe the prevalence of involvement for different combinations of the risk factors T-category and midline extension.

Figure 7 shows nicely that the model is capable of accurately accounting for the most important risk factors, i.e. T-category and midline extension. As observed in the data, the model predicts that the prevalence of contralateral LNL II involvement increases from below 8% for early T-category lateralized tumors to almost 40% when the tumor is of advanced T-category and crosses the mid-sagittal line. Similarly, the prevalence
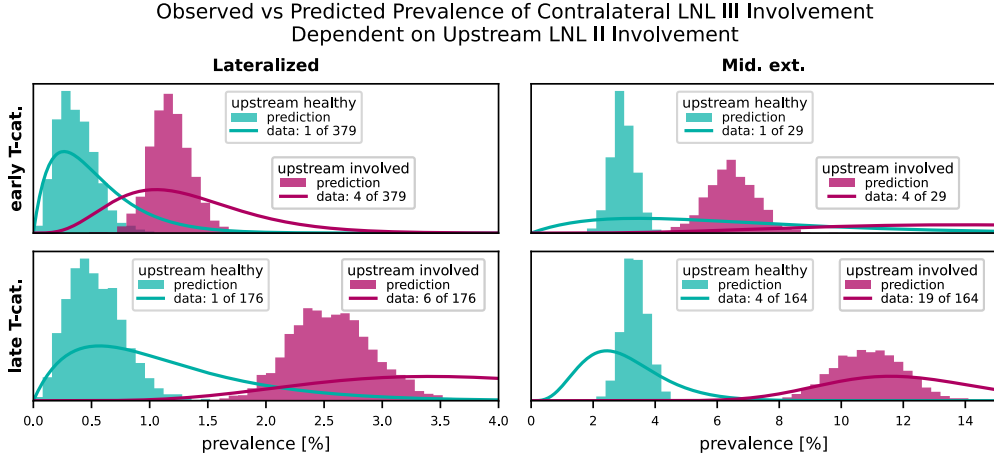
of contralateral LNL III involvement increases from around 2% for early T-category lateralized tumors to almost 15% when the tumor is of advanced T-category and crosses the mid-sagittal line.

Correlation between Ipsi- and Contralateral Involvement
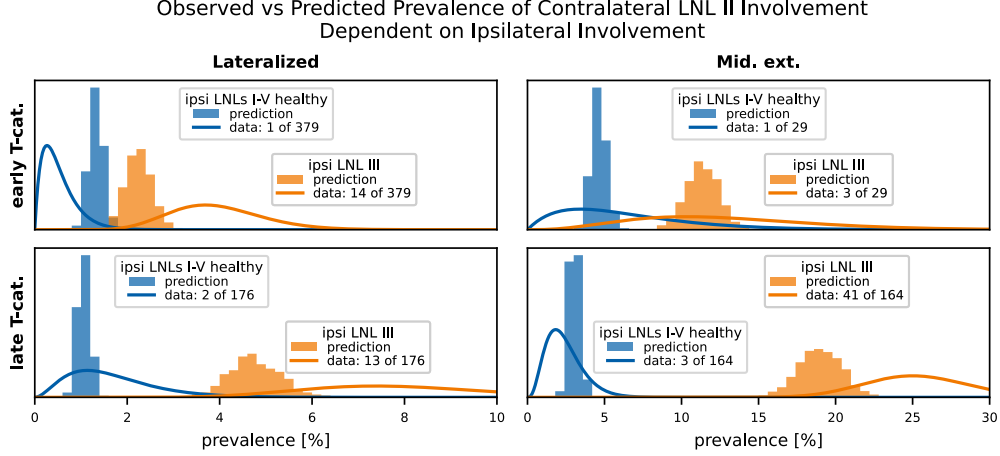
Figure 8: Comparison of the computed and observed prevalences for scenarios that illustrate the model's capability of accounting for the correlation between ipsi- and contralateral involvement. We show two scenarios: One where the ipsilateral neck shows no involvement (at least LNLs I to V are healthy, LNL VII was ignored because data on it is missing for some patients) in blue and one where at least LNL III was involved in orange. These two scenarios are plotted for all combinations of T-category (early in top row, advanced in bottom row) and tumor lateralization (lateralized in left column, extending over mid-sagittal line in right column).

In figure 8 we display the model's ability to capture the correlation between ipsi- and contralateral involvement. It shows that the prevalence of metastases in the two sides of the neck is correlated via the time of diagnosis, despite the model not having any direct connections between the two side. However, there are some discrepancies in the model's prediction: It cannot quite capture the correlation of ipsi- and contralateral involvement that is seen in the data. This is notable for example in the bottom right subplot where the prevalence is overestimated for the scenario of an (almost) healthy ipsilateral neck and underestimated when at least LNL III shows metastases. One reason for this may be the width of the time-prior: If it was wider, the posterior over the diagnosis time given the ipsilateral involvement could more flexibly shift to reflect a more severe contralateral involvement.

Figure 9: The influence of the upstream LNL II's involvement on the prevalence of contralateral level III for the four combinations of tumor lateralization (lateralized or extending over midline) and T-ctageory (early or advanced). Our model predictions (histograms) are plotted against the observations in the data (beta posteriors).

Influence of Upstream Involvement on Contralateral Metastasis

Figure 9 illustrates how rarely the contralateral LNL III harbors metastases when its upstream level II is healthy. Also, the model can capture this correlation rather well, with the exception of early T-category tumors that extend over the mid-sagittal line. As in figure 7 and figure 8, these cases are relatively rare resulting in a broad distribution over the true prevalence.

Prevalence of Midline Extension

Lastly, in figure 10, we plot the prevalence of midline extension in the data versus our model's prediction. It is obvious the model cannot match the large spread between early and advanced T-category seen in the data. This is because to achieve that, it would need to increase the advanced T-category patient's prior distribution over diagnosis times and at the same time reduce the probability of the tumor to cross the midline during a time-step. But since the time-priors parameter is also coupled with the spread probabilities among the LNLs, the model does not have that freedom.

However, we do not consider this discrepancy a major limitation of the model: We will not realistically be interested in the probability of midline extension, as it is always possible to assess it with high certainty. That is also the reason why we initially modelled the midline extension *not* as a random variable, but as a global risk factor that would have been turned on or off from the onset of a patient's disease evolution. This, however, lead to overly high risks for contralateral involvement in advanced T-category patients with midline extension, because then the model assumes an increased spread to the contralateral side from the onset of the disease. Which is probably not true in
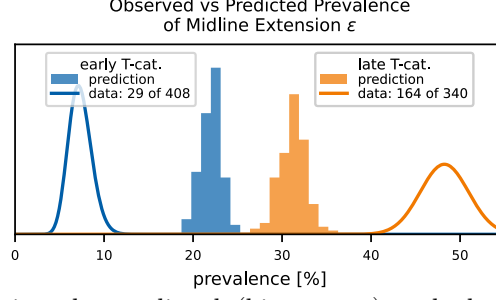
Figure 10: Comparing the predicted (histograms) and observed (lines depicting beta posteriors) prevalence of midline extension for early (blue) and late (orange) T-category. While the prevalence is predicted correctly when marginalizing over T-category, the model cannot capture the degree of separation observed in the data. Since the tumor's midline extension is virtually always part of the diagnosis and hence *given* when predicting a patient's risk, we do not consider this discrepancy a major issue.

a majority of those cases. Thus, treating it as a random variable that only becomes true during a patient's disease evolution resulted in a better description of the data.

Formally, the wrong prediction of midline extension prevalence makes little difference, since it is always given: Instead of $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}, \epsilon \mid \mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}\right)$, we typically compute $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid \mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}, \epsilon\right)$, which does not suffer from the wrong probability of midline extension, as the distribution over hidden states is renormalized:

$$P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid \mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}, \epsilon\right) = \frac{P\left(\mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}} \mid \mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}, \epsilon\right) P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}, \epsilon\right)}{P\left(\mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}, \epsilon\right)}$$

Note that a distribution over ϵ appears both in the enumerator and the denominator, which largely cancel each other, leaving only the midline extension's effect on the distribution over hidden states in the prediction.

Also, the discrepancy in midline extension prevalence between early and advanced T-category is particularly pronounced in oropharyngeal SCC patients. For example, in oral cavity SCC, the midline extension only increases from 15.4% (20 out of 130) to 33.3% (13 out of 39).

Results: Prediction of Risk for Occult Disease

For a clinical application of the model we are interested in the risk of occult metastases rather than the observed patterns of lymph node involvement. We want to estimate the risk for occult disease in clinically negative LNLs, given the patient's individual diagnosis. In terms of our model, the diagnosis consists of the T-category, the lateralization of the tumor (if it extend over the mid-sagittal plane), and which LNLs are clinically involved based on imaging and possibly fine needle aspiration (FNA). For

the results shown in this section, we assume that involvement of a LNL is clinically diagnosed through imaging with a sensitivity of 81% and a specificity of 76% [14]. For FNA we assume a specificity of 98% and a sensitivity of 80%. This amounts to the assumption that lymph node involvement confirmed by FNA is almost certainly true involvement, i.e. there are almost no false positive diagnoses.
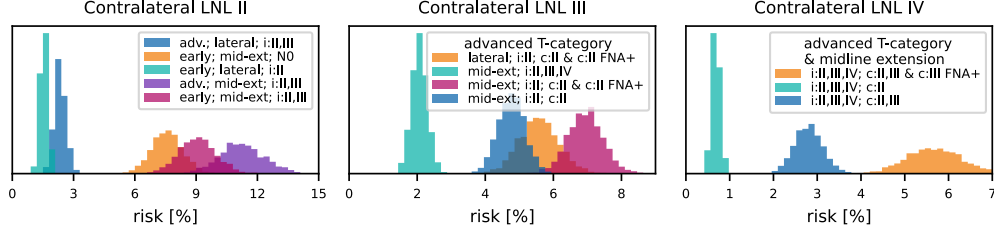
Figure 11: Histograms over the predicted risk of occult contralateral LNL II (left), III (middle), and IV (right) invovlement, shown for some combinations of T-category, tumor lateralization, and clinical LNL diagnoses. All LNLs not explicitly mentioned in the legend, including the LNL for which the risk of occult disease was computed, where assumed to be clinically negative, based on CT imaging (specificity 76%, sensitivity 81%).

Contralateral LNL II

Figure 11 (left panel) shows the predicted risk of occult disease in contralateral LNL II. The most important variable impacting the prediction for contralateral level II involvement in our model is the tumor's lateralization. A patient with a clearly lateralized early T-category tumor and a clinically N0 neck is predicted to have a 1-2% risk for occult disease in contralateral LNL II. For a patient with a clearly lateralized early T-category tumor but *with* mid-sagittal extension, the risk increases to almost 7%.

Also advanced T-category increases the risk of occult disease but plays a lesser role: Considering the scenario of a tumor that crosses the midline and an ipsilateral neck where at least LNL III is clinically involved, the risk for occult contralateral LNL II disease is around 8.5%. For the same scenario and an advanced T-category tumor, the risk increases to 11%.

Lastly, the predicted risk is also correlated via the time-steps to the degree of ipsilateral involvement. Changing the aforementioned scenario (advanced T-category, midline extension, ipsilateral LNL III clinically involved) to one where the patient presents with a clinically N0 neck, the risk for occult disease in the contralateral LNL II falls from 11% to 9.5%.

Taken together, T-category and ipsilateral involvement may still considerably impact the risk prediction for contralateral involvement: In figure 11 the scenarios underlying the orange (6.5%) and the red (11%) histograms differ in T-category (early vs advanced) and clinical diagnosis (N0 vs ipsilateral at least LNL III involved).

Contralalteral LNL III

As shown in the center subplot of figure 11, the risk for occult disease in the contralateral LNL III may only cross a 5% threshold if the upstream level II is clinically involved, too. If the clinical diagnosis of the contralateral upstream level II is pathologically confirmed by FNA the risk increases by another percentage point. When the tumor additionally crosses the mid-sagittal line, the risk may reach around 7%.

Contralateral LNL IV

Our model's prediction for the contralateral LNL IV show that even in the most extreme case, with an advanced T-category tumor extending over the mid-sagittal line as well as metastases in every ipsilateral LNL and every contralateral upstream level, the risk stays just below 3%. We only observe a risk of 5-6% when the involvement of the upstream level III is pathologically confirmed after taking a biopsy using FNA. This is because the model's prior probability for contralateral LNL IV involvement is so low that even given a clinical diagnosis in that level, a false positive in LNL III is a more likely explanation. This possibility, however, drops virtually to zero when tumor cells are pathologically confirmed in level III.

Discussion

Summary

In this work we present a formalism to model the ipsi- and contralateral lymphatic involvement of oropharyngeal SCC patients. An ipsilateral model has been previously developed and published [9, 10]. Based on this, we introduce an extension that leaves the ipsilateral model untouched, but extends it to the contralateral side in an intuitive and comprehensible manner.

The model's performance w.r.t. its ability to describe the data on contralateral nodal involvement was evaluated based on a dataset of 833 patients from four institutions. To the best of our knowledge, this is the first time lymphatic tumor progression has been modelled in such detail and to such an extent. While there was work on regional tumor spread, these models were conceptually different, more limited in the LNLs for which they make predictions, and not trained with real patient data [19, 20].

The model takes the clincal diagnosis of the lymph node levels, the primary tumor's T-category, as well as lateralization into account when predicting the personalized risk for occult disease in any LNL of interest. Owing to the relatively few parameters, the model is highly interpretable and every parameter can be intuitively explained, while still offering good accuracy in its predictions.

Implications for Contralateral Elective Nodal Treatment

The predictions of this model are being used to inform a clinical trial on volume deescalation at the University Hospital Zurich [citation needed]. When accepting a 5% risk of occult disease in any given LNL, the model suggests the following contralateral elective irradiation (assuming the respective LNL appears clinically healthy):

- LNL II when the tumor extends over the mid-sagittal line
- LNL III when LNL II is clinically involved, may need confirmation via FNA
- LNL IV only when upstream level III pathologically confirmed

Acknowledgement

Contralateral Prevalence of Involvement

Source: Article Notebook

Table 4: Contralateral involvement depending on whether the primary tumor extends over the mid-sagittal line, the T-category, and whether the ipsilateral LNL III was involved or healthy.

Table 4

T-cat.	ipsi	Mid. ext.	I		II		III		IV		total
			n	%	n	%	n	%	n	%	n
early	0	False	0	0.00	1	1.16	0	0.00	0	0.00	86
early	0	True	0	0.00	1	10.00	1	10.00	0	0.00	10
early	1	False	1	0.53	11	5.82	2	1.06	1	0.53	189
early	1	True	1	11.11	2	22.22	0	0.00	0	0.00	9
early	2	False	1	0.96	15	14.42	3	2.88	3	2.88	104
early	2	True	0	0.00	3	30.00	4	40.00	1	10.00	10
advanced	0	False	0	0.00	2	5.88	0	0.00	0	0.00	34
advanced	0	True	0	0.00	3	12.50	0	0.00	0	0.00	24
advanced	1	False	0	0.00	3	4.55	0	0.00	0	0.00	66
advanced	1	True	1	1.64	18	29.51	5	8.20	1	1.64	61
advanced	2	False	4	5.26	16	21.05	7	9.21	4	5.26	76
advanced	2	True	3	3.80	46	58.23	18	22.78	8	10.13	79

References

[1] Grégoire, V., Levendag, P., Ang, K.K., Bernier, J., Braaksma, M., Budach, V., Chao, C., Coche, E., Cooper, J.S., Cosnard, G., Eisbruch, A., El-Sayed, S., Emami, B., Grau, C., Hamoir, M., Lee, N., Maingon, P., Muller, K., Reychler, H.: CT-based delineation of lymph node levels and related CTVs in the node-negative neck: DAHANCA, EORTC, GORTEC, NCIC,RTOG consensus guidelines. Radiotherapy and Oncology **69**(3), 227–236 (2003) https://doi.org/10.1016/j.radonc.2003.09.011

[2] Grégoire, V., Ang, K., Budach, W., Grau, C., Hamoir, M., Langendijk, J.A., Lee, A., Le, Q.-T., Maingon, P., Nutting, C., O'Sullivan, B., Porceddu, S.V., Lengele, B.: Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. Radiotherapy and Oncology **110**(1), 172–181 (2014) https://doi.org/10.1016/j.radonc.2013.10.010

[3] Grégoire, V., Others: Delineation of the primary tumour clinical target volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines. Radiother. Oncol. **126**, 3–24 (2018) https://doi.org/10.1016/j.radonc.2017.10.016

[4] Eisbruch, A., Foote, R.L., O'Sullivan, B., Beitler, J.J., Vikram, B.: Intensity-modulated radiation therapy for head and neck cancer: Emphasis on the selection and delineation of the targets. Seminars in Radiation Oncology **12**(3), 238–249 (2002) https://doi.org/10.1053/srao.2002.32435

[5] 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

[6] Chao, K.S.C., Wippold, F.J., Ozyigit, G., Tran, B.N., Dempsey, J.F.: Determination and delineation of nodal target volumes for head-and-neck cancer based on patterns of failure in patients receiving definitive and postoperative IMRT. International Journal of Radiation Oncology*Biology*Physics **53**(5), 1174–1184 (2002) https://doi.org/10.1016/S0360-3016(02)02881-X

[7] Vorwerk, H., Hess, C.F.: Guidelines for delineation of lymphatic clinical target volumes for high conformal radiotherapy: Head and neck region. Radiation Oncology **6**(1), 97 (2011) https://doi.org/10.1186/1748-717X-6-97

[8] Ferlito, A., Silver, C.E., Rinaldo, A.: Elective management of the neck in oral cavity squamous carcinoma: Current concepts supported by prospective studies. British Journal of Oral and Maxillofacial Surgery **47**(1), 5–9 (2009) https://doi.org/10.1016/j.bjoms.2008.06.001

[9] 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

[10] Ludwig, R., Schubert, A., Barbatei, D., Bauwens, L., Hoffmann, J.-M., Werlen, S., Elicin, O., Dettmer, M., Zrounba, P., Pouymayou, B., Balermpas, P., Grégoire, V., Giger, R., Unkelbach, J.: Modelling the Lymphatic Metastatic Progression Pathways of OPSCC from Multi-Institutional Datasets. arXiv (2023). https://doi.org/10.48550/arXiv.2312.11270

[11] Pouymayou, B., Balermpas, P., Riesterer, O., Guckenberger, M., Unkelbach, J.: A Bayesian network model of lymphatic tumor progression for personalized elective CTV definition in head and neck cancers. Physics in Medicine & Biology **64**(16), 165003 (2019) https://doi.org/10.1088/1361-6560/ab2a18

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

[13] 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

[14] De Bondt, R., Others: Detection of lymph node metastases in head and neck cancer: A meta-analysis comparing US, USgFNAC, CT and MR imaging. Eur. J. Radiol. **64**, 266–272 (2007) https://doi.org/10.1016/j.ejrad.2007.02.037

[15] Kyzas, P.A., Others: 18F-fluorodeoxyglucose positron emission tomography to evaluate cervical node metastases in patients with head and neck squamous cell carcinoma: A meta-analysis. J. Natl Cancer Inst. **100**, 712–720 (2008) https://doi.org/10.1093/jnci/djn125

[16] Foreman-Mackey, D., Hogg, D.W., Lang, D., Goodman, J.: Emcee: The MCMC Hammer. \pasp **125**(925), 306 (2013) https://doi.org/10.1086/670067

[17] ter Braak, C.J.F., Vrugt, J.A.: Differential Evolution Markov Chain with snooker updater and fewer chains. Statistics and Computing **18**(4), 435–446 (2008) https://doi.org/10.1007/s11222-008-9104-9

[18] Nelson, B., Ford, E.B., Payne, M.J.: RUN DMC: AN EFFICIENT, PARALLEL CODE FOR ANALYZING RADIAL VELOCITY OBSERVATIONS USING N-BODY INTEGRATIONS AND DIFFERENTIAL EVOLUTION MARKOV CHAIN MONTE CARLO. The Astrophysical Journal Supplement Series **210**(1), 11 (2013) https://doi.org/10.1088/0067-0049/210/1/11

[19] Benson, N., Whipple, M., Kalet, I.J.: A Markov model approach to predicting regional tumor spread in the lymphatic system of the head and neck. AMIA ... Annual Symposium proceedings. AMIA Symposium, 31–35 (2006)

[20] Jung, H., Law, A., Grunblatt, E., Wang, L.L., Kusano, A., Mejino, J.L.V., Whipple, M.E.: Development of a Novel Markov Chain Model for the Prediction of Head and Neck Squamous Cell Carcinoma Dissemination. AMIA Annual Symposium Proceedings **2016**, 1832–1839 (2017)