

# Probabilistic Model of Bilateral Lymphatic Spread in Head and Neck Cancer

Roman Ludwig<sup>1,2\*</sup>, Yoel Perez Haas<sup>1,2</sup> and Jan Unkelbach<sup>1,2</sup>

<sup>1</sup>Department of Physics, University of Zurich.

<sup>2</sup>Radiation Oncology, University Hospital Zurich.

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

Contributing authors: [yoel.perezhaas@usz.ch](mailto:yoel.perezhaas@usz.ch); [jan.unkelbach@usz.ch](mailto:jan.unkelbach@usz.ch);

## Abstract

In current practice, large parts of the contralateral lymph system are included in the elective clinical target volume (CTV-N) of oropharyngeal squamous cell carcinoma (OPSCC) patients. Guidelines on how to define the CTV-N are mostly based on prevalences of nodal involvement in the different lymph node levels (LNLs) and do not consider the detailed, individual diagnosis of a patient.

In this work, we present a probabilistic model for bilateral lymphatic tumor progression in OPSCC. Its aim is to predict the personal risk for occult disease in any LNL, given the patient's clinical diagnosis, T-category, and lateralization of the primary tumor. 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 that are connected to the tumor and among each other via arcs that correspond to spread probabilities per time-step. These spread probability rates are learned via Markov chain Monte Carlo (MCMC) sampling from a dataset of 833 OPSCC patients.

The model is able to describe the data well and may in the future be used to only include LNLs in a CTV-N that show a risk above a specific threshold for that particular patient's diagnosis. E.g., it predicts the risk for occult disease in contralateral LNL II to be below 5% unless the tumor crosses the mid-sagittal line.

## 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 - lymph node levels (LNLs) - are part of the elective clinical target volume (CTV-N) [1–8]. This is to minimize the risk of regional recurrences due to missed 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).

Minimizing this 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 quantify this risk via 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 the individual 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 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 focus of this work will be to extend the formalism of these previous works to include risk predictions for the contralateral side as well. We believe this holds the potential to greatly reduce the electively irradiated volumes in the contralateral neck for patients whose individual risk for occult disease in that side is low. In turn, this may substantially reduce the toxicity from irradiation and the associated decrease in quality of life.

In the following, we begin by inspecting the dataset the model is built upon (section 2) and from it formulate some properties we require the bilateral model extension to have (section 2).

After that, in section 2, we briefly summarize the original model formalism before we extend it to the contralateral side starting in section 2.

Next, we describe how the model training and the computational experiments were set up in section 2 and present the results in section 2, followed by a closing discussion (section 2).

## Data on Lymphatic Progression Patterns

Originally, the only available data of detailed - i.e. per patient and per LNL - lymphatic involvement was a reconstructed table from the findings in Sanguineti and Others [12], which included only ipsilateral nodal involvement for the LNLs I to IV.

To be able to create models for lymphatic tumor progression that include more LNLs and the contralateral side, we have collected a detailed dataset of 833 patients with newly diagnosed oropharyngeal squamous cell carcinomas. It reports the involvement of every patient individually and per lymph node level in tabular form, in addition to other clinico-pathological information such as age, T-category, 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. Note that the data from the Inselspital Bern and the Centre Léon Bérard only consists of patients treated with some form of 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.

Source: [Article Notebook](#)

**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](https://lyprox.org).

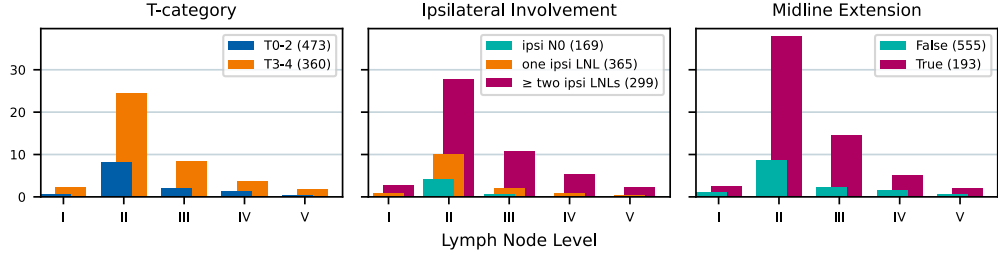
**Table 1**

Institution	Total	Age (median)	Neck Dissection	N0	Early T-Cat.	Mid. Ext.
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](#)

## Contralateral Involvement Prevalence

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 3.



**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).

The left panel in figure 1 indicates that T-category is correlated with contralateral involvement (as it is with overall 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 rare cases, bulky nodal disease ipsilaterally may also redirect lymph fluids to the contralateral side.

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.

## 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].
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 this time from T-category alone.

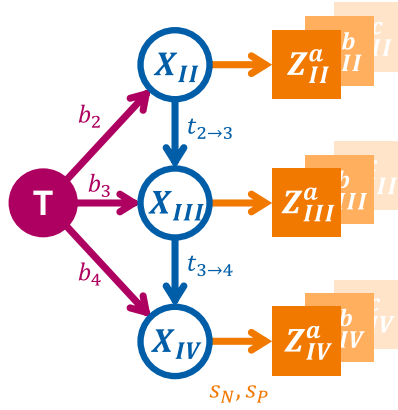
3. A tumor that extends over the mid-sagittal line should yield contralateral metastases with much higher probability.

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.

## 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. [13] and Ludwig et al. [14]. Although these publications do not include the most recent dataset addition from Vall d’Hebron Barcelona Hospital.

## Unilateral Model for Lymphatic Progression



**Figure 2:** 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, ...).

Our first model to predict the lymphatic progression of HNSCC was introduced using Bayesian networks [11]. We subsequently extended this work to a hidden Markov

model (HMM) [9] to allow an intuitive inclusion of T-category into the predictions. We will briefly summarize this HMM’s formalism before building on it to include the contralateral spread in section 2.

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

$$\mathbf{X}[t] = (X_v[t]) \quad v \in \{1, 2, \dots, V\} \quad (1)$$

Here,  $V$  is the number of LNLs the model considers. The values a LNL’s hidden binary RV 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 clinical).

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  $\xi_0 = (0 \ 0 \ \dots \ 0)$  to  $\xi_{2^V} = (1 \ 1 \ \dots \ 1)$ .

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  $\xi_0$ .
2. The *transition matrix*

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

where the value at row  $i$  and column  $j$  represents the probability to transition from state  $\xi_i$  to  $\xi_j$  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  $X_v[t + 1] = 0$ . This effectively masks large parts of the transition matrix to be zero.

3. Lastly, the *observation matrix*

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

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  $\mathbf{X}[t_D] = \xi_i$ .

Note that the transition matrix  $\mathbf{A}$  is parametrized using the spread probabilities of a directed acyclic graph (DAG) that we define as the underlying mechanistic representation of the lymphatic network. An example of such a DAG is shown in figure 2.

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  $\mathbf{A}$ . 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 chosen based on T-category. For instance, the mean of the time-prior to marginalize over the diagnosis time for early T-category patients  $P(t_D | \text{early})$  may be shifted towards earlier times than the one for advanced T-category patients  $P(t_D | \text{early})$ . This gives us for example

$$P(\mathbf{X} | T_x = \text{early}) = \sum_{t=0}^{t_{\max}} P(\mathbf{X} | t) \cdot P(t | \text{early})$$

For later use, we define at this point the matrices of the distributions over all hidden states, given all time-steps  $\Lambda$ :

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

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

In this work, we use binomial distributions  $\mathfrak{B}(t_D, p_{Tx})$  as time-priors which have one free parameter  $p_{Tx}$  for each group of patients we differentiate based on T-category. 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  $\mathbb{E}[t_D] = 10 \cdot p_{Tx}$ .

## 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  $d = (\zeta_i, Tx)$  as follows:

$$\ell = P(\mathbf{Z} = \zeta_i | Tx) = \sum_{t=0}^{t_{\max}} [\xi_0 \cdot \mathbf{A} \cdot \mathbf{B}]_i \cdot P(t | Tx) \quad (5)$$

Above, the quantity inside  $[\dots]_i$  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  $\zeta_j$  that match the provided diagnosis.

The single-patient likelihood  $\ell$  in equation 5 depends on the spread parameters shown in figure 2 via the transition matrix  $\mathbf{A}$  and on the binomial parameters  $p_{Tx}$  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:

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

And it is our goal to infer the values of these parameters for a given dataset  $\mathcal{D} = (d_1, d_2, \dots, d_N)$  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:

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

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

## Model Prediction in the Bayesian Context

Our stated goal is to compute the risk for a patient’s true nodal involvement state  $\mathbf{X}$ , *given* their individual diagnosis  $d = (\zeta_k, Tx)$ . Using Bayes’ law, this is written as:

$$P(\mathbf{X} \mid \mathbf{Z} = \zeta_k, \hat{\theta}, Tx) = \frac{P(\zeta_k \mid \mathbf{X}) P(\mathbf{X} \mid \hat{\theta}, Tx)}{\sum_{i=0}^{2^V} P(\zeta_k \mid \mathbf{X} = \xi_i) P(\mathbf{X} = \xi_i \mid \hat{\theta}, Tx)} \quad (8)$$

The term  $P(\zeta_k \mid \mathbf{X})$  is defined solely by sensitivity and specificity of the diagnostic modality. Terms like this already appeared in the definition of the observation matrix in equation 3. The *prior*  $P(\mathbf{X} \mid \hat{\theta})$  in the above equation is the crucial term that is supplied by a trained model and its parameters  $\hat{\theta}$ .

It is possible to compute this *posterior* probability of true involvement not only for one fully defined state  $\mathbf{X}$ , but also for e.g. individual LNLs: For example, the risk for involvement in level IV would be a marginalization over all states  $\xi_i$ , where  $\xi_{i4} = 1$ . Formally:

$$P(\text{IV} \mid \mathbf{Z} = \zeta_k, \hat{\theta}, Tx) = \sum_{k: \xi_{k4}=1} P(\mathbf{X} = \xi_k \mid \zeta_k, \hat{\theta}, Tx) \quad (9)$$

## 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 even enforce some shared parameters between these two models, like the parameterization of the distributions over diagnose times or the spread among the LNLs. Additionally, we could think of an approach to incorporate the primary tumor’s mid-sagittal extension as a risk factor.

However, this approach lacks a way to describe the correlation between ipsi- and contralateral involvement. This is displayed in table 3 and shows how often the contralateral LNLs I, II, III, and IV were involved, given all possible combinations of midline extension, T-category, and ipsilateral LNL III involvement. Unsurprisingly, the prevalence for contralateral involvement is consistently higher when the tumor extends over the mid-sagittal line or is of later T-category. But it is also more frequent when the ipsilateral side shows more severe involvement, which is here shown via the surrogate LNL III.

Thus, we attempt to extend the formalism in section 2 in such a way that the model’s ipsi- and contralateral side evolve synchronously. To achieve that, we start by writing down the posterior distribution of involvement an analogy to equation 8, which is now a joint probability of an involvement  $\mathbf{X}^i$  ipsilaterally *and* an involvement  $\mathbf{X}^c$  contralaterally, given a diagnosis of the ipsilateral LNLs  $\mathbf{Z}^i$  and of the contralateral ones  $\mathbf{Z}^c$ :

$$P(\mathbf{X}^i, \mathbf{X}^c | \mathbf{Z}^i, \mathbf{Z}^c) = \frac{P(\mathbf{Z}^i, \mathbf{Z}^c | \mathbf{X}^i, \mathbf{X}^c) P(\mathbf{X}^i, \mathbf{X}^c)}{P(\mathbf{Z}^i, \mathbf{Z}^c)} \quad (10)$$

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

The likelihood of the diagnoses given a hidden state simply factorise:  $P(\mathbf{Z}^i, \mathbf{Z}^c | \mathbf{X}^i, \mathbf{X}^c) = P(\mathbf{Z}^i | \mathbf{X}^i) \cdot P(\mathbf{Z}^c | \mathbf{X}^c)$ . And the two factors are contained in the observation matrices  $\mathbf{B}^i$  and  $\mathbf{B}^c$ .

The term representing the model’s prior probability of hidden involvement does not factorize. However, if we assume that lymphatic spread typically does not cross the mid-sagittal line, we can write it as a factorising sum:

$$\begin{aligned} P(\mathbf{X}^i, \mathbf{X}^c) &= \sum_{t=0}^{t_{\max}} P(t) \cdot P(\mathbf{X}^i, \mathbf{X}^c | t) \\ &= \sum_{t=0}^{t_{\max}} P(t) \cdot P(\mathbf{X}^i | t) \cdot P(\mathbf{X}^c | t) \end{aligned} \quad (11)$$

This assumption makes intuitive sense: The two sides of the lymphatic network in a typical patient are approximately mirror images of each other. Thus, no major vessels cross the mid-sagittal line. There may, however, be diffusion of lymph fluid accross this line or bulky involvement that redirects lymphatic drainage significantly.

Using this assumption along with equation 4, we can write the above distribution algebraically as a product:

$$P(\mathbf{X}^i = \xi_n, \mathbf{X}^c = \xi_m) = [\Lambda_i^\top \cdot \text{diag } P(\mathbf{t}) \cdot \Lambda_c]_{n,m} \quad (12)$$

## Parameter Symmetries

In general, the matrices  $\Lambda_i$  and  $\Lambda_c$  could be parameterized using a disjoint set of parameters. I.e., the ipsi- and contralateral spread rates are entirely different. However, using two sensible assumptions, we can reduce the parameter space by sharing some parameters between the sides:

1. 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

$$\begin{aligned} b_v^c &\neq b_v^i \\ t_{rv}^c &= t_{rv}^i \end{aligned} \quad (13)$$

for all  $v \leq V$  and  $r \in \text{pa}(v)$ .

2. 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 across the mid-sagittal plane to the same position, but on the contralateral side. In the beginning we would have  $b_v^c < b_v^i$ , 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*  $\alpha$  that captures a "degree of asymmetry":

$$b_v^{c, \epsilon=\text{True}} = \alpha \cdot b_v^i + (1 - \alpha) \cdot b_v^{c, \epsilon=\text{False}} \quad (14)$$

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,  $b_v^{c, \epsilon=\text{False}}$  for the spread to the contralateral LNLs as long as the tumor is clearly lateralized, and finally  $b_v^{c, \epsilon=\text{True}}$  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  $\alpha$ .

Our parameter space has now expanded to

$$\theta = (\{b_v^i\}, \{b_v^c\}, \alpha, \{t_{vr}\}, p_{\text{early}}, p_{\text{adv.}}) \quad \text{with} \quad \begin{matrix} v \leq V \\ r \in \text{pa}(v) \end{matrix} \quad (15)$$

which is less than double in size, compared to the unilateral model. From these parameters, we now build three different transition matrices: The unchanged  $\mathbf{A}_i$  for the ipsilateral side, one  $\mathbf{A}_c^{\epsilon=\text{False}}$  for the contralateral side that covers the progression as long as the tumor is lateralized, and  $\mathbf{A}_c^{\epsilon=\text{True}}$  in case of a tumor that has crossed the mid-sagittal line.

## Modelling Midline Extension

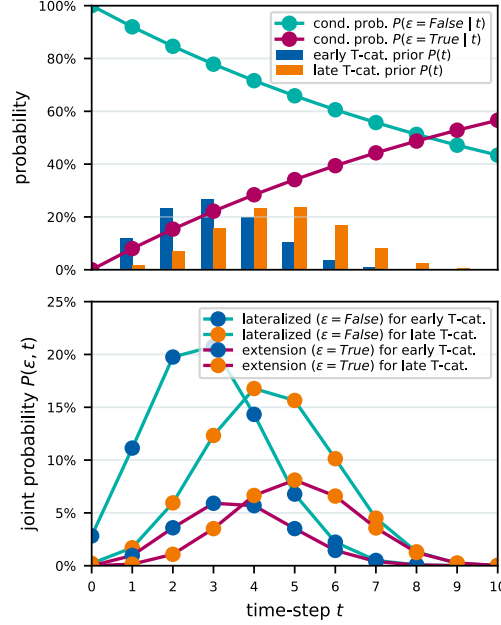
To account for the increased prevalence of involvement on the contralateral side when the tumor is not clearly lateralized anymore, we also model the tumor's extension over the mid-sagittal line as a binary random variable. It starts lateralized and at every time-step there is a finite probability  $p_\epsilon$  that the tumor grows over the symmetry plane of the patient.

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

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

$$P(\epsilon = \text{True} \mid t) = 1 - P(\epsilon = \text{False} \mid t)$$

In figure 3, we visualize how the prior distribution over diagnose times  $P(t)$ , the conditional probability of midline extension  $P(\epsilon \mid t)$ , and their joint  $P(\epsilon, t)$  evolve over the course of a patient evolution.



**Figure 3:** 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 ( $\epsilon = \text{True}$ ), 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.

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:

$$\Lambda_c^{\epsilon=\text{False}} = \begin{pmatrix} \pi^\top \cdot (\mathbf{A}_c^{\epsilon=\text{False}})^0 \\ (1 - p_\epsilon) \cdot \pi^\top \cdot (\mathbf{A}_c^{\epsilon=\text{False}})^1 \\ \vdots \\ (1 - p_\epsilon)^{t_{\max}} \cdot \pi^\top \cdot (\mathbf{A}_c^{\epsilon=\text{False}})^{t_{\max}} \end{pmatrix}$$

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

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  $b_v^{c,\epsilon=\text{True}}$  from the tumor to the contralateral LNLs, given by the linear mix in equation 14. We can achieve the correct marginalization by iteratively building the joint  $P(\mathbf{X}^c, \epsilon = \text{True} | t)$ . We start from  $t = 0$ , where we know that all LNLs are healthy - i.e., the contralateral neck is in the state  $\mathbf{X}_c = \xi_0$  - and the tumor is lateralized ( $\epsilon = \text{False}$ ):

$$P(\mathbf{X}^c, \epsilon = \text{False} | t = 0) = \xi_0$$

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  $p_\epsilon$ . 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:

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

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:

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

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 12:

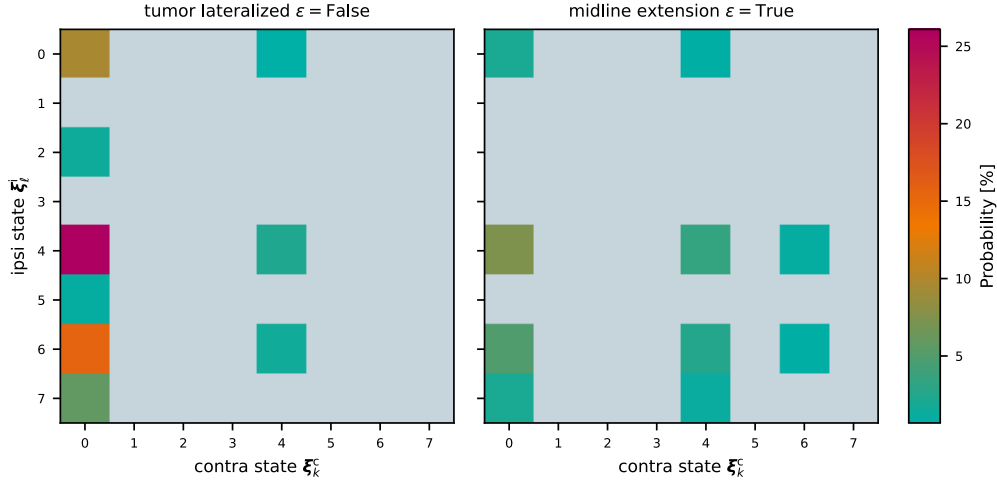
$$P(\mathbf{X}^i = \xi_n, \mathbf{X}^c = \xi_m, \epsilon) = [\Lambda_i^\top \cdot \text{diag } P(\mathbf{t}) \cdot \Lambda_c^\epsilon]_{n,m}$$

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  $\epsilon = \text{False}$  and once for  $\epsilon = \text{True}$  to marginalize over the unknown variable.

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

$$\theta = (\{b_v^i\}, \{b_v^c\}, \alpha, \{t_{vr}\}, p_{\text{early}}, p_{\text{adv.}}, p_\epsilon) \quad \text{with} \quad \begin{matrix} v \leq V \\ r \in \text{pa}(v) \end{matrix} \quad (16)$$

In figure 4 we plot the state distribution for a full midline model as two separate heatmaps. To keep it manageable to interpret, this example model only considers the LNLs II, III, and IV ipsi- and contralaterally, resulting in  $2 \times 2^3 \times 2^3 = 128$  distinct states.



**Figure 4:** 3D state distribution of a midline model with 3 LNLs (II, III, and IV) in both sides of the neck. Note the higher probability for contralateral involvement in the right subplot, where the tumor extends over the mid-sagittal line.

## Methods

In this section, we detail how the experiments were performed. Every figure, table, and result is fully reproducible via the GitHub repository [rmnlwdg/bilateral-paper](https://github.com/rmnlwdg/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 [15, 16] of imaging modalities like MRI and CT do not plausibly match some of our observations: In the USZ cohort, 78% of OPSCC patients were 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 [15, 16].

## MCMC Sampling

For parameter inference, we used the Python package `emcee` [17]. It implements efficient MCMC sampling algorithms that employ multiple parallel samplers for affine invariance and better performance on multi-core CPUs. The sample proposal algorithms used by us are based on differential evolution moves [18, 19]. 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 than  $5.0e-2$ .
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.

## 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 involvement 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  $Tx$  and whether the patient’s tumor extended over the mid-sagittal line, i.e.  $\epsilon = \text{True}$  or  $\epsilon = \text{False}$ .

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 ( $\epsilon = \text{False}$ ). 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 parameter 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 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 the 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:

$$P(\Pi^c \mid \epsilon = \text{False}, Tx = \text{early}) = \frac{P(\Pi^c, \epsilon = \text{False} \mid Tx = \text{early})}{P(\epsilon = \text{False} \mid Tx = \text{early})}$$

In the numerator, we marginalize over all ipsi- and contralateral LNLs’ involvements, except for LNL II contralaterally. This is similar to the marginalization in equation 9, although we are summing over different quantities. In the denominator, we can simply insert the joint distribution over midline extension and diagnose time  $P(\epsilon, t)$  marginalized over  $t$  using the early T-category’s time-prior.

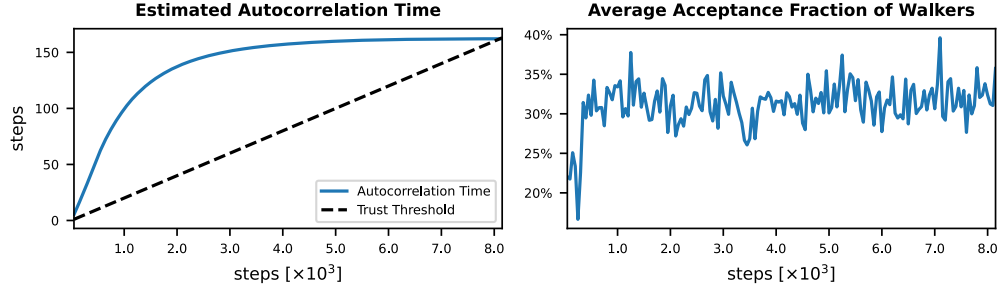
Since we compare it to the data, which does not report true but only observed involvement – although pathologically investigated LNLs may be as close as possible to the ground truth – we do not consider posteriors of the form  $P(\mathbf{X} \mid \mathbf{Z})$  here. Instead, we compute probabilities of 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

First, in figure 5, 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 5:** 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.

In table 2, we tabulate the mean and standard deviation of the sampled parameters for the full midline model.

Source: [Article Notebook](#)

**Table 2:** Mean sampled parameter estimates of the midline model and the respective standard deviation.

Table 2		
Parameter	Mean	Std. Dev.
Mid. ext. probability	8.16%	$\pm 0.48\%$
ipsi: T I	2.80%	$\pm 0.26\%$
ipsi: T II	34.89%	$\pm 1.40\%$
ipsi: T III	5.45%	$\pm 0.66\%$
ipsi: T IV	0.94%	$\pm 0.18\%$



Parameter	Mean	Std. Dev.
ipsi: T V	1.83%	$\pm 0.22\%$
ipsi: T VII	2.32%	$\pm 0.26\%$
contra: T I	0.29%	$\pm 0.09\%$
contra: T II	2.46%	$\pm 0.29\%$
contra: T III	0.14%	$\pm 0.07\%$
contra: T IV	0.19%	$\pm 0.08\%$
contra: T V	0.05%	$\pm 0.04\%$
contra: T VII	0.50%	$\pm 0.17\%$
Mixing	33.87%	$\pm 4.32\%$
I II	62.50%	$\pm 16.69\%$
II III	14.23%	$\pm 1.64\%$
III IV	15.86%	$\pm 1.93\%$
IV V	14.58%	$\pm 3.76\%$
late T-cat. binom. prob.	44.98%	$\pm 1.99\%$

Source: [Article Notebook](#)

## Prevalence Predictions

We want to investigate whether and to what extent the model can fulfill the requirements laid out in section 2. To that end, we compare its 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.

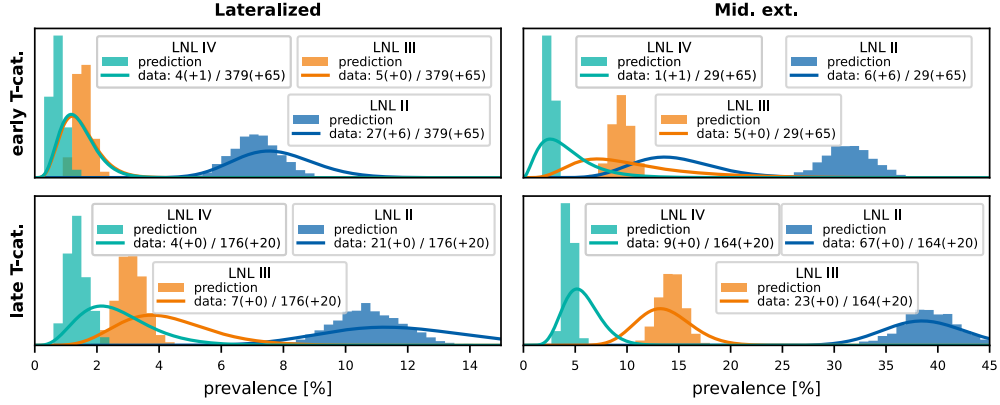
### Dependence of Contralateral Involvement on T-Category and Midline Extension

In figure 6, 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 6 shows nicely how the model is capable of accurately taking the most important risk factors, i.e. T-category and midline extension, into account. As observed in the data, the model predicts that the prevalence of contralateral LNL II involvement jumps 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.

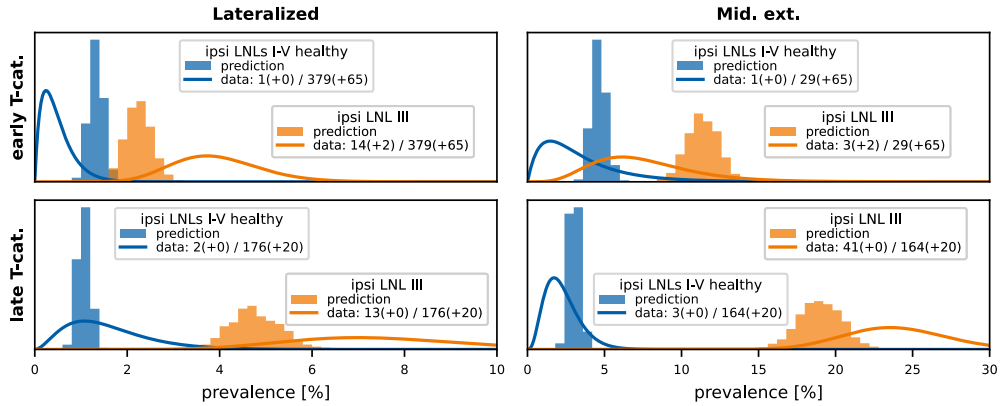
However, for early T-category scenarios with midline extension, the model does seem to overestimate contralateral LNL II and III involvement. This likely stems from the the small sample size of this relatively rare scenario as hinted at by the wide beta posteriors.

Observed vs Predicted Prevalence of Contralateral Involvement  
in the LNLs II, III, and IV



**Figure 6:** 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 different combinations of scenarios involving the risk factors T-category and midline extension.

Observed vs Predicted Prevalence of Contralateral LNL II Involvement  
Dependent on Ipsilateral Involvement

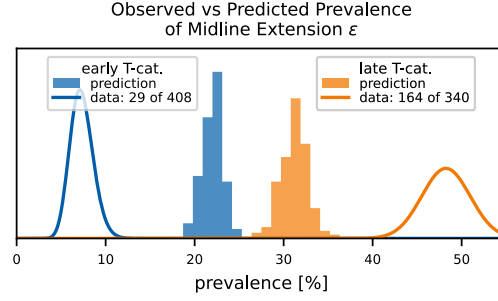


**Figure 7:** Comparison of the computed and observed prevalences for scenarios that illustrate the model's capability of accounting for the correlation between ipsi- and contralateral involvement. We show two scenarios: One where the ipsilateral neck shows no involvement (at least LNLs I to V are healthy, 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).

## Correlation between Ipsi- and Contralateral Involvement

In figure 7 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.

## Prevalence of Midline Extension



**Figure 8:** 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.

Lastly, in figure 8, 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 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  $P(\mathbf{X}^i, \mathbf{X}^c, \epsilon | \mathbf{Z}^i, \mathbf{Z}^c)$ , we typically compute  $P(\mathbf{X}^i, \mathbf{X}^c | \mathbf{Z}^i, \mathbf{Z}^c, \epsilon)$ , which does not suffer from the wrong probability of midline extension, as the distribution over hidden states is renormalized:

$$P(\mathbf{X}^i, \mathbf{X}^c | \mathbf{Z}^i, \mathbf{Z}^c, \epsilon) = \frac{P(\mathbf{Z}^i, \mathbf{Z}^c | \mathbf{X}^i, \mathbf{X}^c, \epsilon) P(\mathbf{X}^i, \mathbf{X}^c, \epsilon)}{P(\mathbf{Z}^i, \mathbf{Z}^c, \epsilon)}$$

Note that a distribution over  $\epsilon$  appears both in the numerator 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).

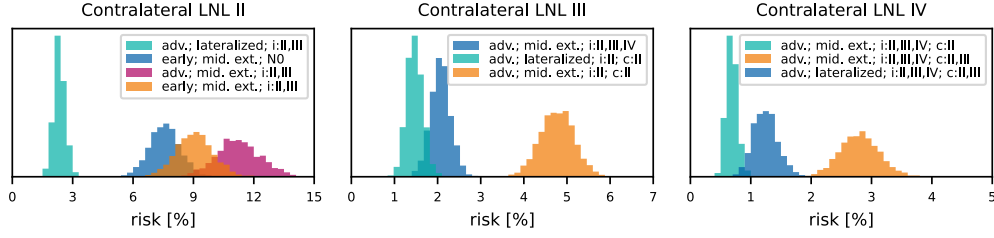
## Prediction of Risk for Occult Disease

In this section, we investigate how the model may be applied clinically: We want to estimate the risk for occult disease in some or all LNLs, given the patient’s individual diagnosis. In terms of our model, this diagnosis consists of the T-category, the lateralization of the tumor (does it extend over the mid-sagittal line?), and which LNLs are clinically involved, e.g. because some lymph nodes appear enlarged in an MRI scan or show increased glycolytic activity on an FDG PET/CT scan.

Figure 9 shows this predicted risk of occult disease in the contralateral LNLs II, III, and IV for interesting combinations of these three risk factors. This risk is computed *given* a CT diagnosis assessing the per-LNL clinical involvement for which we assumed a sensitivity of 81% and a specificity of 76% [15].

### Contralateral LNL II

The variable impacting the prediction for contralateral level II involvement in our model is the tumor’s lateralization. For example, a patient with a clearly lateralized early T-category tumor and a clinically N0 neck is assigned a 1-2% risk for occult disease in contralateral LNL II. Under the same scenario, but *with* mid-sagittal extension, the risk jumps to almost 7%.



**Figure 9:** Histograms over the predicted risk of occult contralateral LNL II (left), III (middle), and IV (right) involvement, 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, were assumed to be clinically negative.

T-category plays a lesser role: Considering the scenario of a tumor that crosses the mid-sagittal line 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 9 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).

### Contralateral LNL III

As shown in the center subplot of figure 9, the risk for occult disease in the contralateral LNL III only ever crosses a 5% threshold if the upstream level II is involved too, resulting in a risk of 5% to 6%. As long as the contralateral neck appears clinically node negative, however, even an advanced T-category and midline extension do not yield a risk higher than 2%.

### 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 does not exceed 5%

# 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 [20, 21].

The model takes the clinical 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, the tumor is of advanced T-category, crosses the midline, and the ipsilateral neck shows involvement of at least LNL II
- LNL IV only for severe ipsilateral involvement, advanced T-category with obvious midline extension and contralateral upstream involvement

## Acknowledgement

This work was supported by:

- the Clinical Research Priority Program “Artificial Intelligence in Oncological Imaging” of the University of Zurich
- the Swiss Cancer Research Foundation under grant number KFS 5645-08-2022

## Contralateral Prevalence of Involvement

Source: [Article Notebook](#)

**Table 3:** 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 3**

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

Source: [Article Notebook](#)

## Marginalizing Beta Posteriors over Unknowns

Source: [Article Notebook](#)

**Table 4:** Combinations of midline extension status and state of contralateral LNL II for early T-category patients.

Midline Extension	no	yes	unknown
contra LNL II			
healthy	352	23	59
involved	27	6	6
total	379	29	65

Source: [Article Notebook](#)

In section 2 we explained that we compare the predicted prevalence of an involvement pattern under a scenario to a beta posterior over the “true” prevalence given the observations in the data. However, our data contains incomplete observations, especially regarding the tumor’s midline extension. This information is not present for a large group of 85 patients, most of which treated at the Centre Léon Bérard in France.

To account for this missing information and plot the correct posteriors over the data prevalence, we need to marginalize over the possible combinations of matching involvement patterns and selected patients under a scenario. For example, we have only 29 early T-category patients with midline extension. But for an additional 65 patients we do not know whether the tumor crossed the mid-sagittal plane. If we now want to compute the prevalence of contralateral LNL II involvement for these early T-category patients with midline extension, then we need to account for the possibility that we are missing information.

In table 4 we show the combinations of the midline extension status and state of the contralateral LNL II. Using only available information, we would get to 6 out of 29 patients for the prevalence of contralateral LNL II involvement among the early T-category patients with midline extension. But the column with the “unknown” numbers suggests that the true prevalence could range from 6 out of 29 + 59 (7%) up to 12 out of 29 + 6 (34%). We assume that every one of the possible combinations has the same a priori probability (uniform prior) but we do need to account for permutations via the binomial factor. In general, let us say we have  $n$  patients with midline extension, of which  $k$  harbor metastases in the contralateral LNL II. And we have  $m$  patients with unknown midline extension, of which  $\ell$  presented with disease in contralateral LNL II, then the posterior distribution - marginalized over the unknown data - over the true prevalence  $x \in [0, 1]$  is given by:

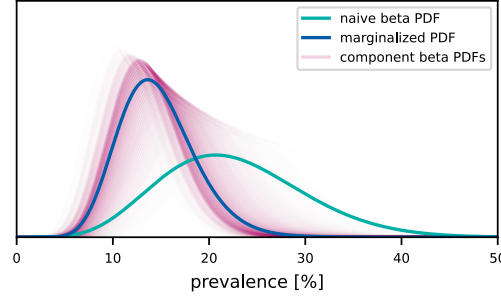
$$p(x \mid n, k, m, \ell, \alpha, \beta) = \left[ \frac{1}{m} \sum_{i=0}^m \frac{1}{\min(i, \ell)} \sum_{j=0}^{\min(i, \ell)} \frac{P(i, j \mid x)}{\binom{n}{k}} \right] \cdot P(n, k \mid x) \cdot \frac{1}{B(\alpha, \beta)} x^\alpha (1-x)^\beta$$

Note the two binomial likelihoods: one being marginalized over and one acting implicitly as a prior to the summed terms. This role of the likelihood over the data with known midline extension as a prior to the part where that data is missing corrects for the fact that there are highly unlikely combinations among those we marginalize over. E.g., expecting that exactly those 6 patients with contralateral LNL II involvement are also coincidentally the only ones of the 65 with a tumor crossing the mid-sagittal line is very unlikely. This is accounted for by the fact that this possibility has only one permutation and it is weakened by the likelihood over the data with known midline extension. With a uniform prior beta distribution ( $\alpha = 1$  and  $\beta = 1$ ), the above simplifies to:



$$p(x \mid n, k, m, \ell) = \left[ \frac{1}{m} \sum_{i=0}^m \frac{1}{\min(i, \ell)} \sum_{j=0}^{\min(i, \ell)} \binom{n+i}{k+j} x^j (1-x)^{i-j} \right] \cdot x^{k+1} (1-x)^{n-k+1}$$

We have plotted this marginal in figure 10 (blue curve) and the terms that make up the above sum (light red curves). Note that we weighted the plotting opacity with the number of permutations each of the possible combinations has. The marginal is compared against the “naive” beta posterior that only takes the data into account for which midline extension is available (green curve).



**Figure 10:** Posterior distributions over the data prevalence for contralateral LNL II involvement under the scenario of early T-category and midline extension. The green beta distribution shows this prevalence for those patients for which midline extension is known. While the blue PDF is a marginal distribution summed over all possible combinations, which are shown in shades of red. The more permutations a particular combination has, the more opaque it is plotted.

The distribution that marginalizes over the unknown midline extension cases is shifted towards a lower prevalence. This makes intuitive sense: With only 6 patients among the 65 with unknown midline extension status, the upper bound for the prevalence has moved down compared to 6 out of 29.

We chose to compare our model against such corrected prevalence posteriors, because the model, too, is capable of marginalizing over unknown midline extension in the data. It would therefore be inaccurate to compare the model trained with the full dataset to only that part of the patient cohort where the tumor lateralization was recorded.

Source: [Article Notebook](#)

## 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\*Biophysics* **53**(5), 1174–1184 (2002) [https://doi.org/10.1016/S0360-3016\(02\)02881-X](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., Balcermpas, 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., Balermipas, 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., Balermipas, 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] Sanguineti, G., Others: Defining the risk of involvement for each neck nodal level in patients with early T-stage node-positive oropharyngeal carcinoma. *Int. J. Radiat. Oncol. Biol. Phys.* **74**, 1356–1364 (2009) <https://doi.org/10.1016/j.ijrobp.2008.10.018>
  - [13] Ludwig, R., Hoffmann, J.-M., Pouymayou, B., Morand, G., Däppen, M.B., Guckenberger, M., Grégoire, V., Balermipas, 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>
  - [14] Ludwig, R., Schubert, A., Barbatei, D., Bauwens, L., Werlen, S., Elicin, O., Dettmer, M., Zrounba, P., Balermipas, 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>
  - [15] 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>
  - [16] 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>
  - [17] Foreman-Mackey, D., Hogg, D.W., Lang, D., Goodman, J.: Emcee: The MCMC Hammer. *\pas* **125**(925), 306 (2013) <https://doi.org/10.1086/670067>
  - [18] 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>

- [19] 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>
- [20] 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)
- [21] 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)