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;

Source: Article Notebook

Source: Article Notebook

Introduction

- head and neck cancer spreads through the lymphatic network
- may sometimes spread to contralateral side
- spreads more frequently and to larger extent contralaterally when tumor extends the mid-sagittal line
- we describe a model based on previously published hidden Markov model
- we extend it to cover the contralateral side, too
- naive approach: make two independent models for each side, but that is not supported by the data
- ipsi- and contralateral side are correlated via time of diagnosis, which is correlated with T-category
- tumor extension over mid-sagittal line is modelled as random variable
- spread probabilities from tumor to contralateral LNLs in case of midline extesnion are linear combinations of these probabilities in case of ipsilateral spread and contralateral spread when tumor is clearly lateralized

Data on Lymphatic Progression Patterns

We have collected a detailed dataset of 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.

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

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.

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

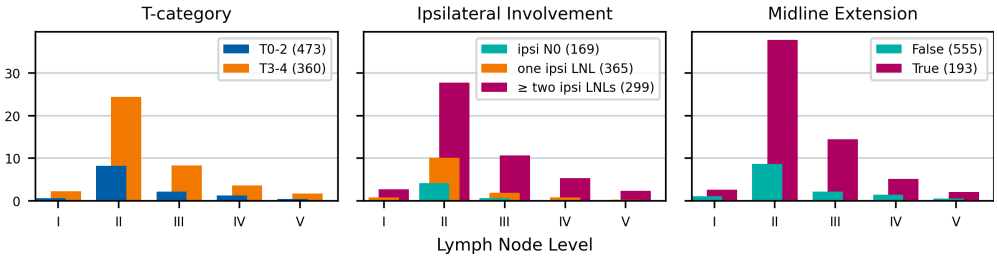

Figure 1: Contralateral involvement stratified by T-category (left panel), the number of metastatic LNLs ipsilaterally (center panel), and whether the primary tumor extended over the mid-sagittal line or was clearly lateralized (right panel).

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 intutitive 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 [1].
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. [2] and Ludwig et al. [3]. Although these publications do not include the most recent dataset addition from Vall d'Hebron Barcelona Hospital.

Unilateral Model for Lymphatic Progression

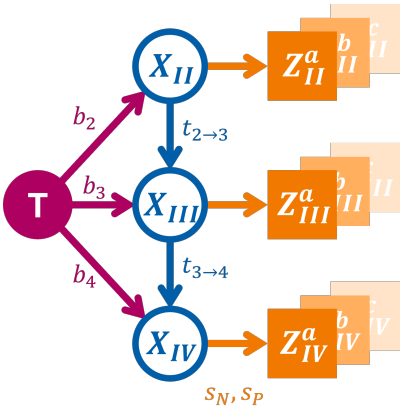

Figure 2: Directed acyclic graph (DAG) representing the abstract lymphatic network in the head and neck region. Blue nodes are the LNLs' hidden random variables, the red node represents the tumor, and the orange square nodes depict the binary observed variables. Red and blue arcs symbolize the probability 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 [4]. We subsequently extended this work to a hidden Markov model (HMM) [1] 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:

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

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

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 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 $X_v[t+1] = 0$. This effectively masks large parts of the transition matrix to be 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 $\mathbf{X}[t_D] = \xi_i$.

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

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

For later use, we define at this point the matrices Λ:

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

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 $\mathfrak{B}(t_D, p_{\mathrm{T}x})$ as time-priors which have one free parameter $p_{\mathrm{T}x}$ 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 $\mathbb{E}[t_D] = 10 \cdot p_{\mathrm{T}x}$.

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:

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

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 2 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, our parameter space is:

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

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:

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

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 $d = (\zeta_k, \mathrm{T}x)$. Using Bayes' law, this is written as:

$$P(\mathbf{X} \mid \mathbf{Z} = \zeta_k, \hat{\theta}, \mathrm{T}x) = \frac{P\left(\zeta_k \mid \mathbf{X}\right) P(\mathbf{X} \mid \hat{\theta}, \mathrm{T}x)}{\sum_{i=0}^{2^V} P\left(\zeta_k \mid \mathbf{X} = \xi_i\right) P(\mathbf{X} = \xi_i \mid \hat{\theta}, \mathrm{T}x)} \tag{8}$$

The term $P\left(\zeta_k \mid \mathbf{X}\right)$ is defined solely by sensitivity and specificity of the diagnostic modality. Terms like this already appeared in the definition of the observation matrx 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 ξ_i, where $\xi_{i4} = 1$. Formally:

$$P(\mathrm{IV} \mid \mathbf{Z} = \zeta_k, \hat{\theta}, \mathrm{T}x) = \sum_{k\,:\,\xi_{k4}=1} P(\mathbf{X} = \xi_k \mid \zeta_k, \hat{\theta}, \mathrm{T}x) \tag{9}$$

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 $\mathbf{X}^{\mathrm{i}}$ ipsilaterally *and* an involvement $\mathbf{X}^{\mathrm{c}}$ contralaterally, given a diagnosis of the ipsilateral LNLs $\mathbf{Z}^{\mathrm{i}}$ and of the contralateral ones $\mathbf{Z}^{\mathrm{c}}$:

$$P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid \mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}\right) = \frac{P\left(\mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}} \mid \mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right) P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right)}{P\left(\mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}\right)} \tag{10}$$

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: $P\left(\mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}} \mid \mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right) = P\left(\mathbf{Z}^{\mathrm{i}} \mid \mathbf{X}^{\mathrm{i}}\right) \cdot P\left(\mathbf{Z}^{\mathrm{c}} \mid \mathbf{X}^{\mathrm{c}}\right)$. And the two factors are contained in the observation matrices $\mathbf{B}^{\mathrm{i}}$ and $\mathbf{B}^{\mathrm{c}}$.

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:

$$\begin{aligned} P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right) &= \sum_{t=0}^{t_{\max}} P(t) \cdot P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid t\right) \\ &= \sum_{t=0}^{t_{\max}} P(t) \cdot P\left(\mathbf{X}^{\mathrm{i}} \mid t\right) \cdot P\left(\mathbf{X}^{\mathrm{c}} \mid t\right) \end{aligned} \tag{11}$$

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:

$$P\left(\mathbf{X}^{\mathrm{i}} = \xi_n, \mathbf{X}^{\mathrm{c}} = \xi_m\right) = \left[\Lambda_{\mathrm{i}}^{\top} \cdot \operatorname{diag} P(\mathbf{t}) \cdot \Lambda_{\mathrm{c}}\right]_{n,m} \tag{12}$$

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_ϵ that the tumor grows over the symmetry plane of the patient.

Technically, the introduction of this additional random variable doubles the space of the hidden states and therefore quadruples the size of the transition matrix $\mathbf{A}$. However, since we assume no correlation between the tumor's lateralization and the metastases in the LNLs, we can evolve the two parts separately.

The 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 given by

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

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.

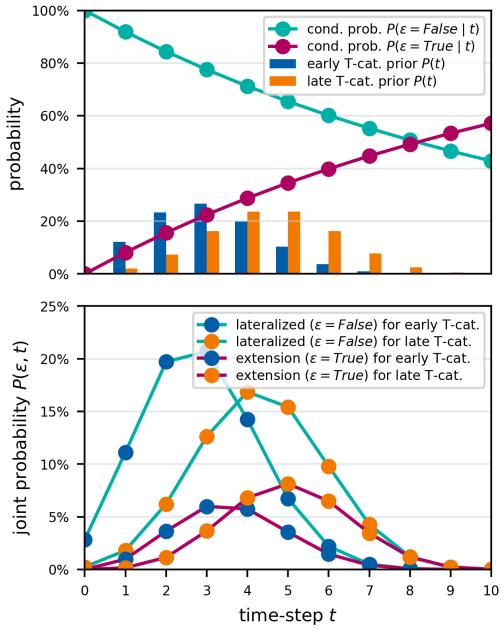

Figure 3: The top panel shows the prior probability to get diagnosed at time-step t for early and late T-category tumors as bars. Also in the top panel, we plot the conditional probability of the tumor's midline extension ($\epsilon = \texttt{True}$), given the time-step t as a line plot. In the bottom panel, we show the joint probability of getting diagnosed in time-step t *and* having a tumor that crosses the midline.

To get the joint probability over the ipsi- and contralateral hidden states, as well as the state of the tumor's midline extension $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}, \epsilon\right)$, we simply add the above terms to the marginalization in equation 11:

$$P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}, \epsilon\right) = \sum_{t=0}^{t_{\max}} P(t) \cdot P(\epsilon \mid t) \cdot P\left(\mathbf{X}^{\mathrm{i}} \mid t\right) \cdot P\left(\mathbf{X}^{\mathrm{c}} \mid t\right)$$

Again, this can be written algebraically, by defining the vector $P(\mathbf{t}, \epsilon) = P(\mathbf{t}) \cdot P(\epsilon \mid \mathbf{t})$, to achieve the same form as in equation 12:

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

In figure 4 we plot the state distribution for a full midline model as two separate heatmaps. To keep it readable, this example model only considers the LNLs II, III, and IV ipsi- and contralaterally.

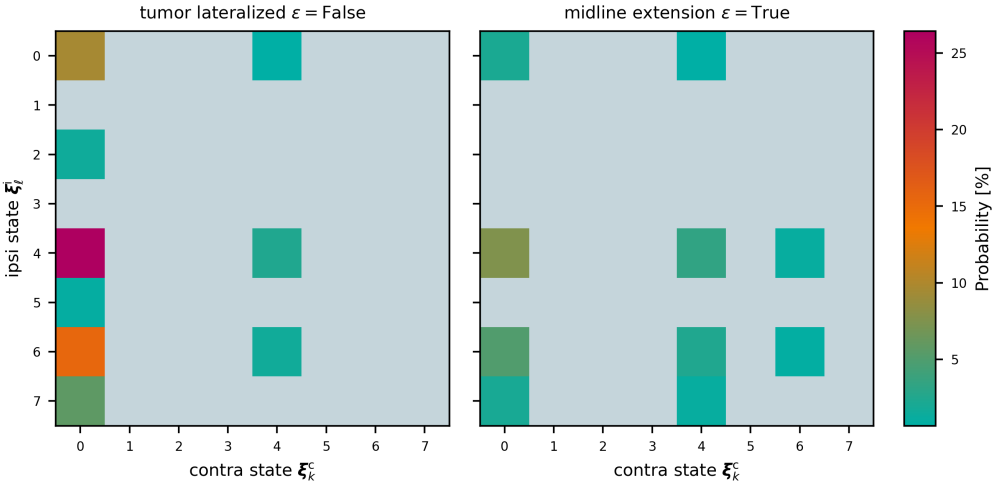

Figure 4: 3D state distribution of a midline model with 3 LNLs (II, III, and IV) in both sides of the neck.

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 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

$$\begin{aligned} b_v^{\mathrm{c}} &\neq b_v^{\mathrm{i}} \\ t_{rv}^{\mathrm{c}} &= t_{rv}^{\mathrm{i}} \end{aligned} \tag{13}$$

   for all $v \leq V$ and $r \in \mathrm{pa}(v)$.
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 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* α that captures a "degree of asymmetry":

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

   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, $b_v^{\mathrm{c},\epsilon=\texttt{False}}$ 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.

Together with the explicit modelling of the tumor's midline extension ϵ from section 2, we now have a model that may be capable of capturing the higher prevalence of contralateral involvment that comes with tumors extending 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 `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 [5] 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 [5].

MCMC Sampling

For parameter inference, we used the Python package `emcee` [6]. It implements efficient MCMC sampling algorithms that employ multiple parallel samplers for affine invariance and better performance on multi-core CPUs. 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 where 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.

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 Tx and whether the patient's tumor extended over the mid-sagittal line, i.e. $\epsilon =$ `True` or $\epsilon =$ `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 9, although we are summing over different quantities. In the denominator, we can simply insert the joint distribution over midline extension and diagnose time $P\left(\epsilon, t\right)$ 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 $P\left(\mathbf{X} \mid \mathbf{Z}\right)$ here. Instead, we compute probabilities of observed involvement $P\left(\mathbf{Z}\right)$, 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.

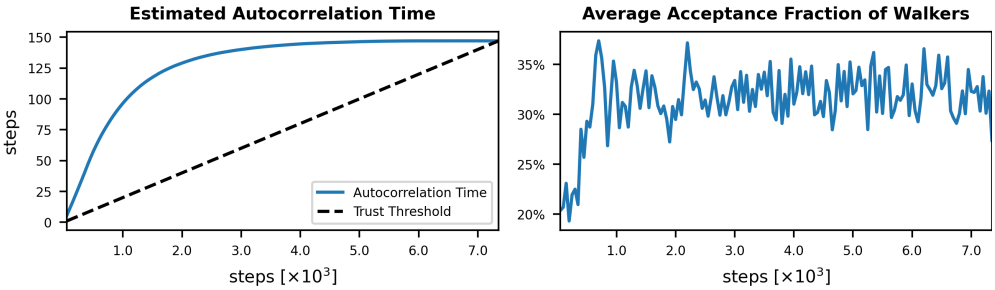

Figure 5: Monitoring quantities during the burn-in phase of the parameter sampling. Left: The autocorrelation time of the sampling chain estimated at different sampling steps. We consider the chain converged when the estimate of the autocorrelation time is stable and drops below the trust threshold of $n/50$ 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	8.13%	± 0.60%
ipsi: T II	35.45%	± 1.75%
ipsi: T III	5.59%	± 0.82%
ipsi: T IV	0.91%	± 0.21%
contra: T II	2.64%	± 0.35%
contra: T III	0.16%	± 0.09%
contra: T IV	0.21%	± 0.10%
Mixing	21.01%	± 3.24%
II III	13.46%	± 1.95%
III IV	15.74%	± 2.24%
late T-cat. binom. prob.	45.35%	± 2.57%

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.

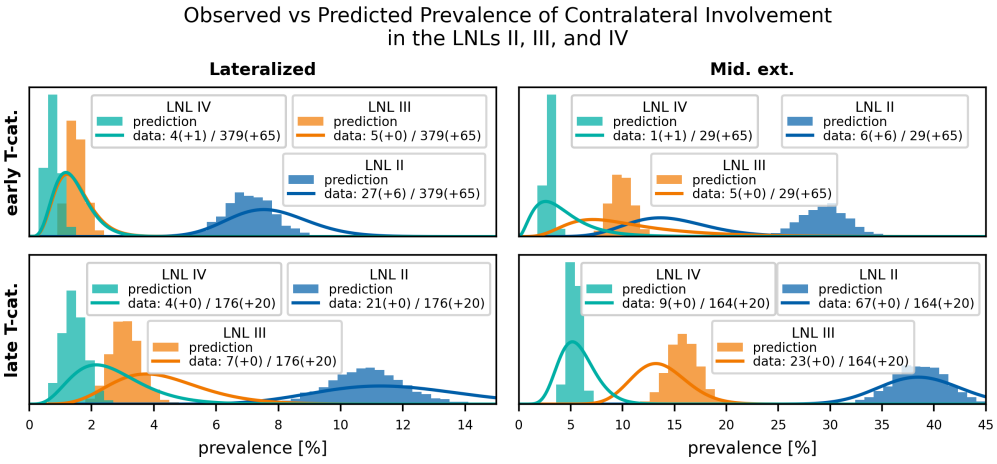

Figure 6: Comparison of predicted (histograms) vs observed (beta posteriors) prevalences. Shown for the contralateral LNLs II (blue), III (orange), and IV (green). The top row shows scenarios with early T-category tumors, the bottom row for late T-category ones. The left column depicts scenarios where 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.

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 invovlement. This likely stems from the the small sample size of this relatively rare scenario as hinted at by the wide beta posteriors.

Correlation between Ipsi- and Contralateral Involvement

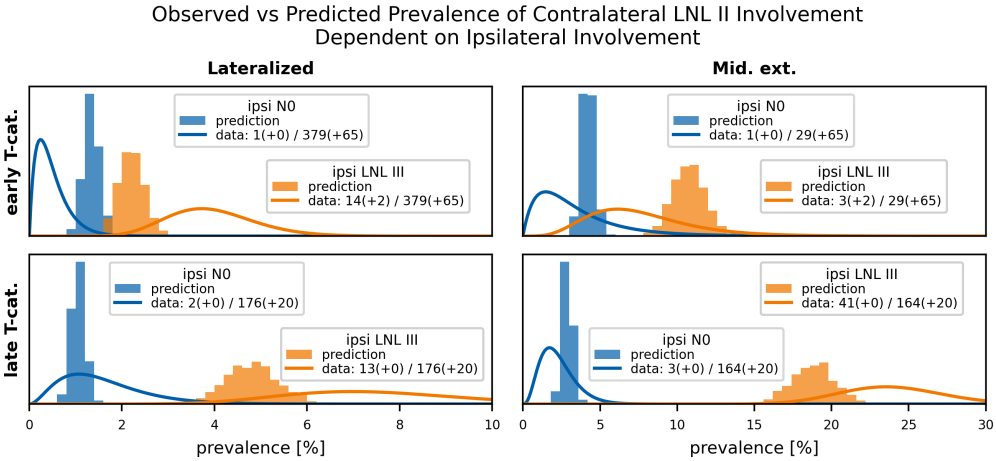

Figure 7: Comparison of predicted (histograms) vs observed (beta posteriors) prevalences. Shown are four scenarios, all including contralateral LNL II involvement: Early T-category and an ipsilateral N0 neck (green), early T-category and ipsilateral LNL II involvement (blue), as well as the same two scenarios but for advanced T-category (orange and red). The figure shows that the model is capable of describing the correlation between ipsi- and contralateral involvement. Although for the scenario of LNL II involvement in both sides, the prediction's split between early and advanced T-category is not large enough.

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 small discrepancies in the model's prediction: When considering the scenario of LNL II involvement in the ipsi- *and* contralateral side, it cannot quite capture the split between early and late T-category. The model slightly overestimates the prevalence of this scenario for early T-category patients and underestimates it for advanced T-category.

Prevalence of Midline Extension

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

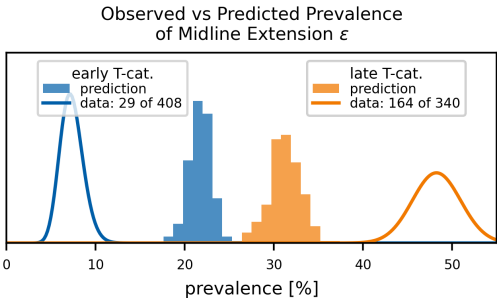

Figure 8: Comparing the predicted (histograms) and observed (lines depicting beta posteriors) prevalence of midline extension for early (blue) and late (orange) T-category. While the prevalence is predicted correctly when marginalizing over T-category, the model cannot capture the degree of separation 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.

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

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.

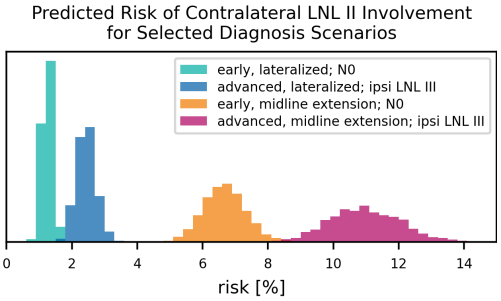

Figure 9: Histograms over the predicted risk of occult contralateral LNL II invovlement, shown for some combinations of T-category, tumor lateralization, and ipsilateral clinical involvement. The contralateral side was always assumed to be clinically negative.

Figure 9 shows this predicted risk of occult disease in contralateral LNL II for four 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% [5].

The variable impacting the prediction for contralateral 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%.

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).

Discussion

- conceptually extended model to cover both sides of neck
- explicitly models midline extension (also allows for easy marginalization)
- midline extension intuitively causes contralateral spread to become more similar to ipsilateral spread

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

T-cat.	ipsi	Mid. ext.	I		II		III		IV		total
			n	%	n	%	n	%	n	%	n
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 contra LNL II	no	yes	unknown
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 invovlement 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 invovlement 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%). Assuming that every one of the possible combinations has

the same a priori probability (uniform prior) and accouting for permutations via the binomial factor, the posterior marginalized over the unknowns could be written as

$$
\begin{aligned}
p(x; \alpha = 1, \beta = 1) &= \sum_{m=0}^{65} \sum_{\ell=0}^{\min(m,6)} \binom{n+m}{k+\ell} x^{k+\ell} (1-x)^{n+m-(k+\ell)} \cdot \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha} (1-x)^{\beta} \\
&= \sum_{m=0}^{65} \sum_{\ell=0}^{\min(m,6)} \binom{n+m}{k+\ell} x^{k+\ell+1} (1-x)^{n+m-(k+\ell)+1}
\end{aligned}
$$

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).

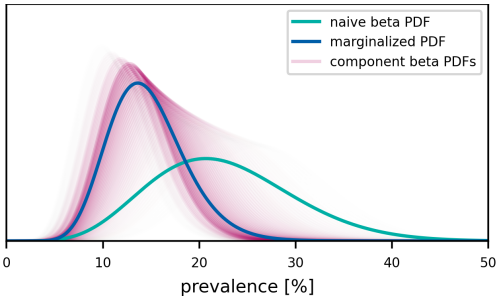

Figure 10: Posterior distributions over the data prevalence for contralateral LNL II involvement under the scenario of early T-category and midline extension. The green beta distribution shows this prevalence for those patients for which midline extension is known. While the blue PDF is a marginal distribution 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 six patients among the 65 with unknown midline extension status, the upper bound for the prevalence has moved down compared to six 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] Ludwig, R., Pouymayou, B., Balermpas, P., Unkelbach, J.: A hidden Markov model for lymphatic tumor progression in the head and neck. Sci Rep **11**(1), 12261 (2021) https://doi.org/10.1038/s41598-021-91544-1

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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