Modelling lymphatic progression in head and neck cancer

Roman Ludwig

September 5, 2022

Abstract

Abstract goes here...

Dedication

To mum and dad

Declaration

I declare that..

Acknowledgements

I want to thank...

Contents

1 Previous Work **2**

1.1 Bayesian Network 2
1.2 Limitations 4

2 Unilateral hidden Markov model **5**

2.1 Formulating lymphatic progression as HMM 5
2.2 Parametrization of the transition matrix 6
2.3 Marginalization 8
2.4 Inference of model parameters 8
2.5 Incorporation of T-stage 9
2.5.1 * Interpretation of time-steps and time-priors 10
2.5.2 * Impact of shape and length of the time-prior 11
2.6 Sampling 14
2.7 Risk assessment of microscopic involvement 14
2.8 Inference and risk assessment for incomplete diagnoses 15
2.9 Multiple diagnostic modalities 16
2.10 Combining modalities and data 17

3 Bilateral hidden Markov model **18**

3.1 Expanding the unilateral model 18
3.2 Parameter symmetries and mid-line extension 19
3.3 Comparing bilateral models 21

4 Possible further extension **30**

4.1 Trinary hidden random variables 30

5 Detailed reporting of involvement in OPSCC **32**

5.1 Abstract 33
5.2 Introduction 33
5.3 Material & methods 35
5.4 Results 36
5.5 Discussion 41

Chapter 1

Previous Work

In this chapter I will introduce models that were previously developed to capture and predict lymphatic progression in head and neck squamous cell carcinoma (HNSCC). This also includes a Bayesian network (BN) model by [29], which served as a starting point to this work. This brief recap will also introduce the notation and formalism that will be used throughout the rest of the thesis.

After that, unrelated previous works will be mentioned and lastly, we will discuss the limitations and reasons to develop a new model and formalism.

1.1 Bayesian Network

We model the state of each lymph node level (LNL) as a hidden or unobserved binary random variable, which indicates via values 0 or 1 if an LNL is healthy or involved, respectively. This state indicates if there is truly tumor present in an LNL, including the presence of occult metastases for the involved state – motivating the term hidden or unobserved state. Every LNL can be diagnosed using one or multiple modalities. Most used for diagnosis are imaging techniques like positron emission tomography (PET), computed tomography (CT) and magnetic resonance imaging (MRI), but palpation or fine needle aspiration (FNA) are also used. The diagnosis too, is modelled as binary random variable – this time an observed one – taking on 0 for negative and 1 for positive.

For notational convenience, we collect the hidden and observed random variables in a random vector each:

$$\begin{aligned} \text{hidden} \quad \mathbf{X} &= (X_v) \to \{0,1\}^V \\ \text{observed} \quad \mathbf{Z} &= (Z_v) \to \{0,1\}^V \end{aligned} \tag{1.1}$$

where V is the number of LNLs $v \in \{1, 2, \dots, V\}$ in the graph. The conditional probabilities that link the hidden state to the observations can be written as follows:

$$\begin{aligned} P_{BN}\left(Z_v = z_v \mid X_v = x_v\right) = \left(z_v + (-1)^{z_v} \cdot s_P\right)(1 - x_v) \\ + \Big((1 - z_v) + (-1)^{1 - z_v} \cdot s_N\Big)\, x_v \end{aligned} \tag{1.2}$$

with s_N and s_P being the sensitivity and specificity of the used diagnostic method. For example, for the probability of a false negative observation, i.e. diagnostic modality misses the presence of tumor, we get

$$P_{BN}\left(Z_v = 0 \mid X_v = 1\right) = 1 - s_N \tag{1.3}$$

Spread of the tumor through the lymphatic network is represented in this model by directed arcs to and between LNLs as illustrated in **??**. We introduce an additional vertex to the graph representing the primary tumor, which we assume to be the only one. Directed arcs from the primary tumor to an LNL represent direct spread of tumor cells from the primary tumor to the LNL. These arcs are associated with parameters b_v that we call base probabilities, and which indicate the probability that the tumor spreads directly to LNL v. When LNL s receives efferent lymphatics from LNL r, this too is represented by a directed arc from LNL r to s, and $r = \mathrm{pa}\left(s\right)$ which is called a parent node of s. These arcs are associated with a transition probability t_{rs} from r to s. The resulting directed acyclic graph (DAG) is shown in **??**, comprising ipsilateral levels I, II, II, through IV, and will be used throughout this work. However, when more data of detailed LNL involvement including additional levels becomes available and/or contralateral involvement, the model can be extended. The parameters b_v and t_{rs} associated with the directed arcs represent conditional probabilities, i.e. b_v answers the question given that all parent nodes are healthy, how likely is it that the primary tumor spreads to node v? t_{rs} on the other hand, can answer the question assuming no efferent spread from the primary tumor and given that all parent nodes except r are healthy, what is the likelihood of spread to node s? The conditional probability for involvement of LNL v given the state of its parent nodes is then given by

$$\begin{aligned} P_{BN}\left(X_v = x_v \mid X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}, b_v, t_{\mathrm{pa}(v)v}\right) \\ = x_v + (-1)^{x_v}(1-b_v)(1-t_{\mathrm{pa}(v)v})^{x_{\mathrm{pa}(v)}} \end{aligned} \tag{1.4}$$

We note here that this parametrization assumes the independence of causal influences (ICI), thereby allowing us to describe the model using only a few interpretable parameters. Dropping this assumption, a BN can also be defined using conditional probability tables (CPT) that have columns for every possible combinations of parent states. However, with the increase of the number of parent nodes (causes) in the graph, the number of parameters in the respective CPT would grow exponentially.

For the graph in **??** we can write down the parametrized CPT in the following manner:

$$\begin{aligned} &P_{BN}\left(X_v = 0 \mid X_{\mathrm{pa}(v)} = 0\right) = 1 - b_v \\ &P_{BN}\left(X_v = 1 \mid X_{\mathrm{pa}(v)} = 0\right) = b_v \\ &P_{BN}\left(X_v = 0 \mid X_{\mathrm{pa}(v)} = 1\right) = \left(1 - b_v\right)\left(1 - t_{\mathrm{pa}(v)v}\right) \\ &P_{BN}\left(X_v = 1 \mid X_{\mathrm{pa}(v)} = 1\right) = 1 - \left(1 - b_v\right)\left(1 - t_{\mathrm{pa}(v)v}\right) \end{aligned} \tag{1.5}$$

In case of a more general network, in which some LNLs receive efferent lymphatics from multiple other LNLs, eq. (1.5) can be generalised and the conditional probability of the hidden state becomes

$$P_{BN}\Big(X_v = x_v \mid \big\{X_r = x_r\,,\, t_{rv}\big\}_{r\in\mathrm{pa}(v)}, b_v\Big) \\ = x_v + (-1)^{x_v}(1-b_v)\prod_{r\in\mathrm{pa}(v)}(1-t_{rv})^{x_r} \quad (1.6)$$

where we marginalized over all hidden variables X. Here we have assumed that each patient's diagnosis $\mathbf{z} = (z_1 \quad z_2 \quad \cdots \quad z_V)$ is complete, meaning that we have a diagnosis for each LNL. The likelihood can then be used to infer the model parameters via maximum likelihood inference or sampling.

1.2 Limitations

While BNs can model the probabilistic relationship between involvement in different levels, they lack an explicit way to describe the evolution of the tumor over time. The concept of dynamic Bayesian networks (DBNs) has been developed to introduce the notion of time into probabilistic models. DBNs are generalizations of hidden Markov models (HMMs) and formally similar to what we will introduce now. The metastatic spread in the lymphatic system is a dynamic system and by modeling it with a formalism that can capture this, we obtain a more intuitive model of the problem and a framework that can incorporate T-stage into estimating the risk of LNL involvement. We can do this because tumors go through the stages T1 to T4 sequentially, meaning that – for a given tumor – it is a surrogate of time.

Chapter 2

Unilateral hidden Markov model

This chapter concerns itself with modelling the unilateral lymphatic spread using the formalism of HMMs that will also be introduced below.

The content of this chapter is largely based on our publication [24] with some modifications and additions to improve continuity with the next chapter, where the presented model will be extended.

2.1 Formulating lymphatic progression as HMM

We consider discrete time-steps $t \in \{0, 1, 2, \dots, t_{\max}\}$. We will start by defining the hidden random variable for the state of the HMM at time t to be

$$\mathbf{X}[t] = (X_v[t]) \tag{2.1}$$

which represents the patient's state of LNL involvement as in the BN, but for each time-step we have an instance of it. For the diagnosis $\mathbf{Z}$ on the other hand, we do not need to differentiate between different times, since in practice we will only ever see one diagnosis. This is illustrated in **??**. The reason for this is that, if we diagnose a patient with cancer, treatment starts timely and we no longer observe the natural progression of the disease. From a modeling standpoint however, this is a problem that we will address later.

A hidden Markov model is fully described by the starting state $\mathbf{X}[0] \coloneqq \boldsymbol{\pi}$ and the two conditional probability functions that govern the progression from a state $X[t]$ at time t to a state $X[t+1]$ at the following time-step

$$P_{HMM}\left(\mathbf{X}[t+1] \mid \mathbf{X}[t]\right) \tag{2.2}$$

and the probability of a diagnostic observation given the true state of the patient

$$P_{HMM}\left(\mathbf{Z} \mid \mathbf{X}[t]\right) \tag{2.3}$$

Since both our state space and our observation space are discrete and finite, it is possible to enumerate all possible states and observations and collect them in a table or matrix. This so-called *transition matrix* would then be

$$\mathbf{A} = (A_{ij}) = \big(P_{HMM}\left(\mathbf{X}[t+1] = \boldsymbol{\xi}_i \mid \mathbf{X}[t] = \boldsymbol{\xi}_j\right)\big) \tag{2.4}$$

and the *observation matrix*

$$\mathbf{B} = (B_{ij}) = \left(P_{HMM} \left(\mathbf{Z} = \boldsymbol{\zeta}_i \mid \mathbf{X}[t] = \boldsymbol{\xi}_j \right) \right) \tag{2.5}$$

Here, $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_j$ are no new variables, but just $\mathbf{x}$ and $\mathbf{z}$ renamed and reordered. The indices i and j for one of the possible states or observations for the entire patient, not for an individual LNL. In total, there are $S = |\{0,1\}|^V$ different states and the same number of different possible observations per diagnostic modality. We order the hidden states from

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix} \tag{2.6}$$

to

$$\boldsymbol{\xi}_{16} = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \tag{2.7}$$

in this case of $V = 4$. The exact ordering does not matter, it is just a convenience for the notation. our ordering of the states can be seen in the axes of **??**. In analogy, we order the observations $\boldsymbol{\zeta}_j$ from 1 to 2^V. Note that for now we will not consider multiple diagnostic modalities and how to combine them. We will get back to that topic in **??**.

In our case, the starting state corresponds to a primary tumor being present but all LNLs are still in the healthy state. The observation matrix $\mathbf{B}$ is specified via sensitivity and specificity as described in eq. (2.7). The main task is to infer the transition matrix $\mathbf{A}$. Usually, it is inferred from a series of observations and there exist efficient algorithms for that, e.g. the sum-product algorithm, which is particularly efficient in chains. Unfortunately, these algorithms cannot be applied for our problem for two profound reasons:

1. We only have a single observation instead of a consecutive series of observations.
2. It is unclear how many time-steps it took from the starting state to the one observation we have at the time of diagnosis.

In the remainder of this section, we will detail the HMM step-by-step, starting with the parameterization of the transition matrix $\mathbf{A}$ in section 2.2. Afterwards, in section 2.3, I will tackle the aforementioned problems, followed up by explaining how we perform inference on this model (section 2.4), incorporate information about a patient's T-stage (section 2.5) and assess the risk of LNL involvement in a new patient (section 2.7). Lastly, we will introduce a way to incorporate incomplete observations in section 2.8.

2.2 Parametrization of the transition matrix

The square transition matrix $\mathbf{A}$ has $S = 2^{2V}$ entries and therefore $S(S-1) = 2^{2V} - 2^V$ degrees of freedom. Although searching the full space of viable transition matrices is possible via unparametrized sampling techniques, it is computationally challenging and hard to interpret. To achieve this reduction in degrees of freedom, and also preserve the anatomically and medically motivated structure of the Bayesian network from **??**, we can represent the transition probability from one state $\mathbf{x}[t]$ to another state $\mathbf{x}[t+1]$ using the conditional probabilities defined for the BN. The difference is that the probability of observing a certain state of LNL v now depends on the state of the patient one time-step before. Note that from here on, we will mostly drop the probabilistically correct notation $P(X = x)$ and just write $P(x)$ for brevity

$$\begin{aligned} P_{HMM}\left(\mathbf{x}[t+1] \mid \mathbf{x}[t]\right) = \prod_{v \leq V} & Q\left(x_v[t+1]; x_v[t]\right) \\ & \times \left[P_{BN}\left(x_v[t+1] \mid \left\{x_r[t]\,,\, \tilde{t}_{rv}\right\}_{r \in \mathrm{pa}(v)}, \tilde{b}_v\right)\right]^{1-x_v[t]} \end{aligned} \tag{2.8}$$

Here we have reused the conditional probability from the BN for each LNL, but we take it to the power of one minus that node's previous value. This ensures that an involved node stays involved with probability 1. The parameters $\tilde{t}_{\mathrm{pa}(v)v}$ and $\tilde{b}_v$ take the same role as in the BN, but they are now probability *rates*, since they act per time-step. Lastly, the first term Q in the product formalizes the fact that a metastatic lymph node level cannot become healthy again once it was involved. This also means that several entries in the transition matrix $\mathbf{A}$ must be zero. In a table the values of $Q\left(x_v[t+1]; x_v[t]\right)$ can be written like this:

$$\begin{aligned} Q\left(X_v[t+1] = 0; X_v[t] = 0\right) &= 1 \\ Q\left(X_v[t+1] = 0; X_v[t] = 1\right) &= 0 \\ Q\left(X_v[t+1] = 1; X_v[t] = 0\right) &= 1 \\ Q\left(X_v[t+1] = 1; X_v[t] = 1\right) &= 1 \end{aligned} \tag{2.9}$$

which gives rise to a "mask" for $\mathbf{A}$ which can be seen in **??**.

To illustrate eq. (2.8), it helps to look at a specific example. E.g., the transition probability from state $\boldsymbol{\xi}_5 = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix}$ to state $\boldsymbol{\xi}_7 = \begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix}$, which represents starting with involvement only in LNL II and asking for the probability that LNL III becomes involved as well over the next time-step:

$$\begin{aligned} & P_{HMM}\left(\mathbf{X}[t+1] = \boldsymbol{\xi}_7 \mid \mathbf{X}[t] = \boldsymbol{\xi}_5\right) \\ & \quad = Q\left(X_1[t+1] = 0; X_1[t] = 0\right) P_{BN}\left(X_1[t+1] = 0 \mid \tilde{b}_1\right)^1 \\ & \quad \times Q\left(X_2[t+1] = 1; X_2[t] = 1\right) P_{BN}\left(X_2[t+1] = 1 \mid X_1[t] = 0, \tilde{t}_{12}, \tilde{b}_2\right)^0 \\ & \quad \times Q\left(X_3[t+1] = 1; X_3[t] = 0\right) P_{BN}\left(X_3[t+1] = 1 \mid X_2[t] = 1, \tilde{t}_{23}, \tilde{b}_3\right)^1 \\ & \quad \times Q\left(X_4[t+1] = 0; X_4[t] = 0\right) P_{BN}\left(X_4[t+1] = 0 \mid X_3[t] = 0, \tilde{t}_{34}, \tilde{b}_4\right)^1 \\ & \quad = \left(1 - \tilde{b}_1\right) \cdot 1 \cdot \left(\tilde{b}_3 + \tilde{t}_{23} - \tilde{b}_3\tilde{t}_23\right) \cdot \left(1 - \tilde{b}_4\right) \end{aligned} \tag{2.10}$$

The interpretation of the last line is that this is the probability that LNL I and IV do not become involved, while LNL III gets infected through lymphatic drainage from either the main tumor or LNL II. The probability of LNL II remaining involved is 1, of course, which is why we take the respective term to the power of 0.

2.3 Marginalization

To calculate the likelihood function, we have to calculate the probability of a given diagnostic observation. To that end, we first calculate the probability of observing a given diagnosis $\mathbf{z} = \boldsymbol{\zeta}_j$ at a fixed time-step t. As depicted in **??**, we must consider every possible evolution of a patient's disease that leads to the observed diagnosis. Mathematically, this means that we need to marginalize over all such paths. And here is where the HMM-formalism comes in very useful, because this marginalization happens automatically when we multiply the transition matrix with itself and eventually with the observation matrix:

$$P\left(\mathbf{Z} = \boldsymbol{\zeta}_j \mid t\right) = \left[\boldsymbol{\pi}^\top \cdot (\mathbf{A})^t \cdot \mathbf{B}\right]_j \tag{2.11}$$

where the $\boldsymbol{\pi}$ is the column vector for the healthy starting state. $\mathbf{A}$ is multiplied with itself t times and thereby produces a matrix that describes the transition probability from the healthy state to all possible states $\mathbf{x}[t]$ in exactly t time-steps marginalized over the actual pathway of the patient's disease. The index $[\ldots]_j$ here means that from the resulting (row-)vector of probabilities we take the component that corresponds to the diagnose $\mathbf{z} = \boldsymbol{\zeta}_j$.

So, essentially, eq. (2.11) first computes the probability vector of all possible true hidden states, given a time step t

$$P\left(\mathbf{X} = \boldsymbol{\xi}_i \mid t\right) = \left[\boldsymbol{\pi}^\top \cdot (\mathbf{A})^t\right]_i \tag{2.12}$$

and then multiplies it with the respective observation probability vector, which is a column of the $\mathbf{B}$ matrix, to finally marginalize over all possible true hidden states – effectively a sum over i in eq. (2.12) – at the time t of diagnosis.

The problem that the number of time-steps until diagnosis is unknown cannot be solved in such an elegant fashion. Therefore, we must resort to brute force marginalization and introduce a prior $p(t)$, which is a discrete distribution over a finite number of time-steps. It describes the prior probability that a patient's cancer is diagnosed at a particular time-step t. To get the probability of a diagnosis $\mathbf{z}$ we must compute

$$P\left(\mathbf{Z} = \boldsymbol{\zeta}_j\right) = \sum_{t=0}^{t_{\max}} p(t) \cdot P\left(\mathbf{Z} = \boldsymbol{\zeta}_j \mid t\right) = \left[\sum_{t=0}^{t_{\max}} p(t) \cdot \boldsymbol{\pi}^\top \cdot (\mathbf{A})^t \cdot \mathbf{B}\right]_j \tag{2.13}$$

While the choice of the time-prior may seem unclear at this point, its role for including T-stage into this model will be discussed in section 2.5.

2.4 Inference of model parameters

In the formalism of the last sections, the P_{HMM} depends implicitly through P_{BN} on parameters $\theta = \left\{\tilde{b}_v, \tilde{t}_{pv} \mid v \leq V, p \in \mathrm{pa}(v)\right\}$, which – as mentioned – are now probability rates and have therefore a slightly different interpretation. Due to the marginalization over time-steps in eq. (2.13) the likelihood function additionally depends on the choice and parametrization of the prior $p(t)$. The parameters are to be inferred from a dataset of lymphatic progression patterns in a cohort of patients. We still assume that for each patient we record for every LNL v whether it is involved according to only one diagnostic modality. In other words, for each patient we observe one of the 2^V possible diagnoses. As mentioned before, we will expand this to multiple diagnostic modalities furhter down in **??**.

Formally, we can then express the dataset $\boldsymbol{\mathcal{Z}}$ of N patients as vector $\mathbf{f}$ of the number of patients f_i for which the diagnosis corresponds to the observational state $\boldsymbol{\zeta}_i$. The likelihood $P\left(\boldsymbol{\mathcal{Z}} \mid \theta\right)$ of observing this dataset, given a particular choice of parameters, is then given by

$$P\left(\boldsymbol{\mathcal{Z}} \mid \theta\right) = \prod_{i=1}^{2^V} P\left(\boldsymbol{\zeta}_i \mid \theta\right)^{f_i} \tag{2.14}$$

with the probability $P\left(\boldsymbol{\zeta}_i \mid \theta\right)$ specified by eq. (2.13). The product runs formally over all possible observational states. In reality, f_i will likely be zero for a number of rare or implausible states that are not in the dataset. Note that $\sum_i f_i = N$.

By Bayes' rule, the posterior distribution of those parameters is

$$P\left(\theta \mid \boldsymbol{\mathcal{Z}}\right) = \frac{P\left(\boldsymbol{\mathcal{Z}} \mid \theta\right) P\left(\theta\right)}{\int P\left(\boldsymbol{\mathcal{Z}} \mid \theta'\right) P\left(\theta'\right) \, d\theta'} \tag{2.15}$$

where $P(\theta)$ is the prior over these parameters. Since they are exclusively probability rates, they must all come from the interval $[0,1] \in \mathbb{R}$. In this work we will choose the most uninformative prior

$$p(\theta) = \begin{cases} 1 & \text{if } \; \theta_r \in [0,1]\,; \forall r \leq E \\ 0 & \text{otherwise} \end{cases} \tag{2.16}$$

where E is the number of edges in the DAG we use to represent the lymphatic system. While it is easy to compute the likelihood, it is not feasible to efficiently calculate the normalization constant in the denominator of eq. (2.15). Hence, we will use Markov-chain Monte Carlo (MCMC) sampling methods to estimate the parameters θ and their uncertainty.

2.5 Incorporation of T-stage

We have introduced the HMM with the promise that it could handle the concept of T-stages through its explicit modeling of dynamic processes. To keep up with that, we will now explain how this is achieved using the time-prior $p(t)$.

The core idea is to assume that early T-stage and late T-stage tumors share the same patterns of metastatic progression, except that late T-stage tumors are on average diagnosed at a later point in time, and thereby also show, on average, higher LNL involvement. Formally, this can be described by assuming a different time-prior $p_T(t)$ for every T-stage T. On the other hand, the transition matrix $\mathbf{A}$ is assumed to be the same for all T-stages.

For the inference of model parameters, the training data is split into subgroups according to T-stage. We now define a column-vector $\mathbf{f}_T$ separately for each T-stage, which counts the number of patients in the dataset that were diagnosed with one of the possible observational states and a given T-stage. The log-likelihood from which we want to sample is then simply a sum of the likelihoods as above, where the essential difference is that we equip each marginalization over time with a different time-prior $p_T(t)$, according to its T-stage:

$$\log P\left(\boldsymbol{\mathcal{Z}} \mid \theta\right) = \sum_{T=1}^{4} \log \left[\sum_{t=0}^{t_{\max}} p_T(t) \cdot \boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^t \cdot \mathbf{B} \right] \cdot \mathbf{f}_T \tag{2.17}$$

The logarithm must be taken element-wise for the resulting row-vector inside the square brackets. The only data-dependent term here is the vector $\mathbf{f}_T$ counting the occurrences of all possible observations. It is again important to note that the only difference between the part of the log-likelihood for the different T-stages is the exact shape or parametrization of the time-prior. The transition probabilities, and hence also the transition matrix $\mathbf{A}$, are the same for all T-stages. For this to work, we rely on the assumption that different typical patterns of nodal involvement for the same primary tumor location are caused mainly by different progression times

At this point, it makes sense to briefly introduce a notation of the above equation that is more suitable for the actual programmatic implementation of the inference and the extension we will discuss later. We can rewrite the term in the square brackets of eq. (2.17) by using the matrix

$$\boldsymbol{\Lambda} \coloneqq P\left(\mathbf{X} \mid \mathbf{t}\right) = \begin{pmatrix} \boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^0 \\ \boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^1 \\ \vdots \\ \boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^{t_{\max}} \end{pmatrix} \tag{2.18}$$

were row number t corresponds to the vector $\boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^t$, i.e. the probabilities for all possible hidden states, given the diagnose time. So, the element $\boldsymbol{\Lambda}_{ti}$ corresponds to the probability $P\left(\boldsymbol{\xi}_i \mid t\right)$ of a patient arriving in the ith state after t time steps. With this, we can rewrite the term in the square brackets of eq. (2.17) purely as a product of vectors and matrices:

$$\sum_{t=0}^{t_{\max}} p_T(t) \cdot \boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^t = p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda} \tag{2.19}$$

with $p_T\left(\mathbf{t}\right) = \begin{pmatrix} p_T(0) & p_T(1) & \cdots & p_T(t_{\max}) \end{pmatrix}$. The matrix $\boldsymbol{\Lambda}$ implicitly depends on the spread probabilities, while each of the $p_T(\mathbf{t})$ depends on the respective parametrization of the time prior. They are the only objects that depend on the parameters θ and they are independent of the data.

2.5.1 * Interpretation of time-steps and time-priors

To add more interpretability to the time-prior $p(t)$ introduced in section 2.3, we want to give some insights here to what we think the time-steps and the distribution over them is supposed to mean.

First, the time that passes in the real world between the abstract time-steps t and $t+1$ should not be seen as a somewhat arbitrarily chosen fixed time, measured in days or weeks. To how much real-world time that corresponds for a specific patient is irrelevant for our risk assessment, although it might prove very valuable for other research on tumor growth. Also, the time between two time-steps does not need to be constant; the model makes no assumptions about this. It merely assumes the probability of transition between states to be the same from t to $t+1$ and for all t.

The time-prior $p(t)$ is essentially the probability that a patient is diagnosed after exactly t time-steps. If we knew how long a patient had cancer before getting diagnosed and we also knew how long a typical timestep for this patient and his/her type of cancer was, then we could just fix $p(t) = 1$ for the appropriate number of time-steps t and set $p(t') = 0, \forall t' \neq t$. Since it is likely almost never known, we need to spread the probability over a range of time-steps, reflecting the fact that the diagnose of cancer happens spontaneously, e.g. during a routine checkup.

2.5.2 * Impact of shape and length of the time-prior

It turns out that length and shape of $p(t)$ have almost no effect on the risk predictions as long as we are not concerned with different T-categories. So, if we learn our parameters from a dataset that only contains T1 patients and then compute risks for T1 patients only, the result will not differ almost regardless of the time-prior that was used for learning and risk assessment. Only too few time-steps may pose a problem, since then the system might not be able to spread to all LNLs via all pathways. And too many time-steps could introduce numerical problems, because the learned probability rates $\tilde{b}_v$ and $\tilde{t}_{\mathrm{pa}(v)v}$ become smaller for longer time-priors.

To understand the impact of the number of time-steps T on the results, we looked at a simple analytical model: Assume that there is a system with only one LNL that the primary tumor can spread to that is empirically involved with probability $p^\star = 0.4$. For this situation, we can now derive how the base probability rate $\tilde{b}_v$ changes for a uniform time-prior

$$p(t) = \frac{1}{T} \qquad \text{for} \quad t \in \{1, 2, \dots, T\} \tag{2.20}$$

if we vary the total number of time-steps T. We can write $p^\star$ as

$$\begin{aligned} p^\star &= \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} 1 & 0 \end{pmatrix} \cdot \begin{bmatrix} (1-\tilde{b}_1) & \tilde{b}_1 \\ 0 & 1 \end{bmatrix}^t \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} 1 & 0 \end{pmatrix} \cdot \begin{bmatrix} (1-\tilde{b}_1)^t & 1-(1-\tilde{b}_1)^t \\ 0 & 1 \end{bmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \frac{1}{T} \sum_{t=1}^{T} \left[1 - (1-\tilde{b}_1)^t \right] = 1 - \frac{1}{T} \sum_{t=1}^{T} (1-\tilde{b}_1)^t \end{aligned} \tag{2.21}$$

The right-hand side essentially contains the partial sum of the geometric series and can easily be computed to yield

$$p^\star = 1 - \frac{\left(1-\tilde{b}_1\right)\left(1-(1-\tilde{b}_1)^T\right)}{\tilde{b}_1 T} \tag{2.22}$$

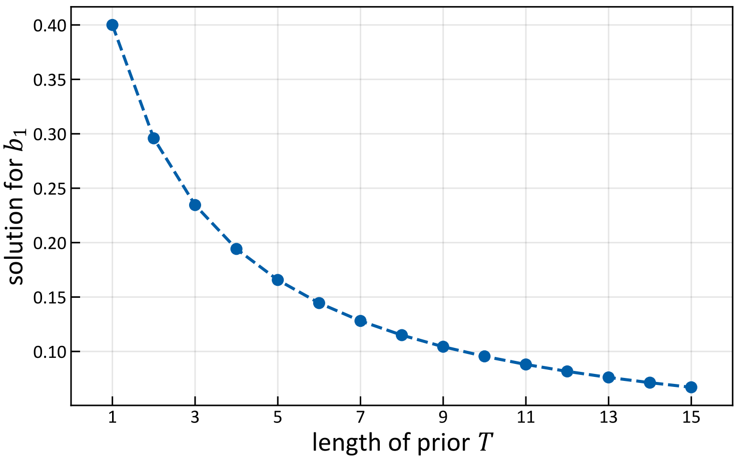

Figure 2.1: Solutions to eq. (2.22) for the base probability rate $\tilde{b}_1$ given a $p^\star$ of 0.4 and T increasing from 1 to 15.

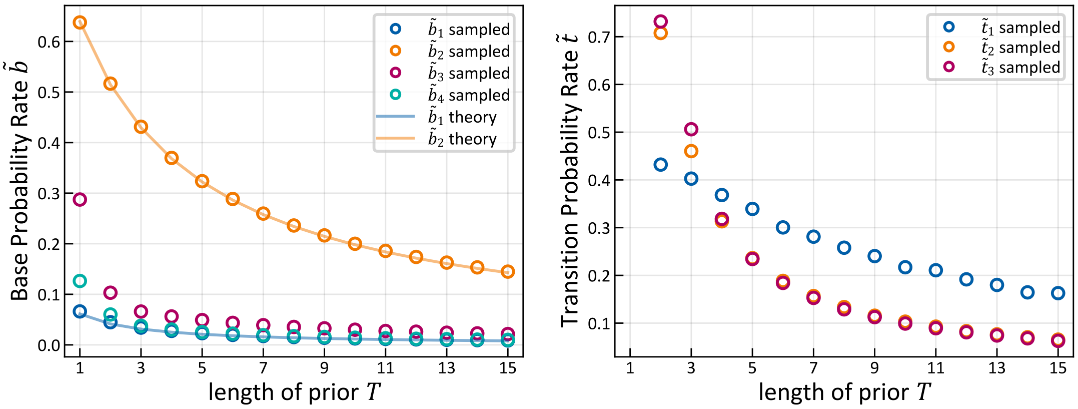

Figure 2.2: Decay of base probability rates as a function of the number of time-steps the time-prior has. Circles depict the results from learning the same dataset with different time-priors while solid lines show the analytical result starting with a $p^\star$ corresponding to the prevalence of involvement of LNL I and II respectively.

It is not possible to analytically solve for $\tilde{b}_1$ in the case of arbitrary T, but numerical solutions are very easy to find and are plotted in fig. 2.1. This confirms the intuition, that the base and transition probability rates become smaller when the total time over which the tumor spreads is divided into more but shorter time-steps.

Now we compare this idealized result to the decay of the probability rates for the full system. To that end, the model with LNLs I-IV was trained as in the same way as for the figures in the section above, but with differently long uniform time-priors instead of a Binomial prior. So, the probability for every time-step is $p(t) = 1/T$ for all $t \geq 1$, but zero for the starting state π. fig. 2.2 shows the expected value of the parameters as a function of T. It is important to stress again that the risk predicted by the models using all those different-length uniform time-priors was

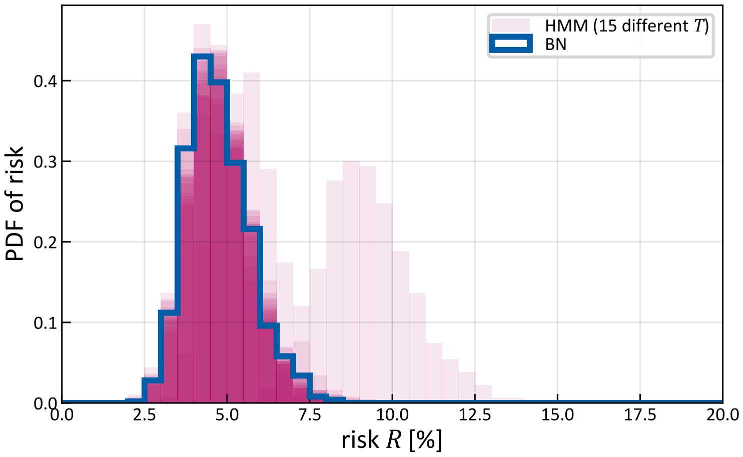

Figure 2.3: Prediction for the risk of involvement in LNL III, given that no other LNL was observed to be involved. Computed by training our hidden Markov model with 15 different-length uniform time-priors (transparent red) as well as the Bayesian network model (blue line). The one outlier is the HMM with a time-prior covering only one time-step. It is centered on the prevalence of involvement for LNL III, regardless of the given diagnose.

the same for $T \geq 2$. For the one-step model with $T = 1$ the risk prediction of a LNL does not depend on the diagnose. For example, we expect the risk in level III is higher, when level II is involved, due to the spread from LNL II to III. With such a short time-prior, however, the model cannot capture this, and all risk predictions will just yield the prevalence of involvement. This effect can be seen in Figure A3 and shows what has been stated earlier: The support of the time-prior has little effect on the model's predictions, as long as it is sufficient to capture the spread through the lymphatic system.

The theoretical result in eq. (2.22) is applicable to the parameter $\tilde{b}_1$, and approximately to $\tilde{b}_2$ since involvement of level II is driven by direct infiltration from the primary tumor rather than transition from level I. For levels III and IV, the theory is not applicable as they have two relevant parent nodes. The solid lines in fig. 2.2 show agreement of the theoretical result with the sampling based training of the full model (circles), where the probabilities $p^\star$ were set to 6.1% and 63.9%, corresponding to the prevalence of level I and II involvement in the dataset, respectively.

This again shows that, while looking at one T-category only, the time-prior's parameters overdetermine the system. For any choice of T, the base and transition probability rates can be adjusted such that the Hidden Markov Model is equivalent to the Bayesian network's performance. Only if we want to distinguish between patients of different T-category the HMM can outperform the BN.

2.6 Sampling

With a parameter set $\theta = \left(\left\{\tilde{b}_v\right\}, \left\{\tilde{t}_{rv}\right\}_{r\in\mathrm{pa}(v)}\right)$ $\forall v \leq V$, we can assess the risk of nodal involvement, given a diagnosis $\mathbf{z}$, of a new patient. Using Bayes' law, the risk for a certain LNL v being involved is given by the conditional probability

$$\begin{aligned} R\left(X_v = 1 \mid \mathbf{z}, \theta\right) &= \frac{P\left(\mathbf{Z} = \mathbf{z} \mid X_v = 1, \theta\right) P\left(X_v = 1 \mid \theta\right)}{P\left(\mathbf{Z} = \mathbf{z} \mid \theta\right)} \\ &= \sum_{i\,:\,\xi_{iv}=1} \frac{P\left(\mathbf{Z} = \mathbf{z} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right)}{P\left(\mathbf{Z} = \mathbf{z} \mid \theta\right)} \end{aligned} \tag{2.23}$$

Note that in the second line, we have explicitly written out the marginalization over all hidden states $\boldsymbol{\xi}_i$ that have LNL v involved. We have written the state of LNL v in the state $\boldsymbol{\xi}_i$ as ξ_{iv}. The denominator can be computed using eq. (2.13), which already includes the marginalization over all hidden states $\boldsymbol{\xi}_i$.

The process of sampling randomly generates L sets of parameters $\theta = \begin{pmatrix}\theta_1 & \theta_2 & \dots & \theta_L\end{pmatrix}$. They are therefore random variables and so is the risk $R\left(X_v \mid \mathbf{z}, \theta\right)$ since it is a function of θ. Using the Monte Carlo estimator, we can therefore compute the moments of the distribution over the risk, including e.g. the expectation value

$$\mathbb{E}_{\boldsymbol{\theta}}\left[R\left(X_v = 1 \mid \mathbf{z}\right)\right] = \frac{1}{L}\sum_{k=1}^{L} R\left(X_v = 1 \mid \mathbf{z}, \theta_k\right) \tag{2.24}$$

In the result sections below, we compute the individual risks for a large enough number L of sampled parameters. Thereby, we can compute histograms for the risk that will approach the real probability density of the respective risk for $L \to \infty$. This provides additional information on the uncertainty in the predicted risk resulting from uncertainty in the model parameters.

2.7 Risk assessment of microscopic involvement

With a parameter set $\theta = \left(\left\{\tilde{b}_v\right\}, \left\{\tilde{t}_{rv}\right\}_{r\in\mathrm{pa}(v)}\right)$ $\forall v \leq V$, we can assess the risk of nodal involvement, given a diagnosis $\mathbf{z}$, of a new patient. Using Bayes' law, the risk for a certain LNL v being involved is given by the conditional probability

$$\begin{aligned} R\left(X_v = 1 \mid \mathbf{z}, \theta\right) &= \frac{P\left(\mathbf{Z} = \mathbf{z} \mid X_v = 1, \theta\right) P\left(X_v = 1 \mid \theta\right)}{P\left(\mathbf{Z} = \mathbf{z} \mid \theta\right)} \\ &= \sum_{i\,:\,\xi_{iv}=1} \frac{P\left(\mathbf{Z} = \mathbf{z} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right)}{P\left(\mathbf{Z} = \mathbf{z} \mid \theta\right)} \end{aligned} \tag{2.25}$$

Note that in the second line, we have explicitly written out the marginalization over all hidden states $\boldsymbol{\xi}_i$ that have LNL v involved. We have written the state of LNL v in the state $\boldsymbol{\xi}_i$ as ξ_{iv}. The denominator can be computed using eq. (2.13), which already includes the marginalization over all hidden states $\boldsymbol{\xi}_i$.

The process of sampling randomly generates L sets of parameters $\theta = \begin{pmatrix}\theta_1 & \theta_2 & \dots & \theta_L\end{pmatrix}$. They are therefore random variables and so is the risk $R\left(X_v \mid \mathbf{z}, \theta\right)$ since it is a function of θ. Using the Monte Carlo estimator, we can therefore compute the moments of the distribution over the risk, including e.g. the expectation value

$$\mathbb{E}_{\boldsymbol{\theta}}\left[R\left(X_v = 1 \mid \mathbf{z}\right)\right] = \frac{1}{L}\sum_{k=1}^{L} R\left(X_v = 1 \mid \mathbf{z}, \theta_k\right) \tag{2.26}$$

In the result sections below, we compute the individual risks for a large enough number L of sampled parameters. Thereby, we can compute histograms for the risk that will approach the real probability density of the respective risk for $L \to \infty$. This provides additional information on the uncertainty in the predicted risk resulting from uncertainty in the model parameters.

2.8 Inference and risk assessment for incomplete diagnoses

A diagnosis is often not complete, meaning that not all LNLs might have been checked with a diagnostic modality. E.g., FNA is usually only performed in a subset of LNLs. Hence, we must be able to deal with "incomplete" observations for some LNLs. To do so, we first introduce a new observation variable

$$d_v \in \{0, 1, \emptyset\} \tag{2.27}$$

where $\emptyset$ indicates *unobserved.* Furthermore, we define a *match function*

$$\text{match}(\mathbf{d}, \mathbf{z}) := \begin{cases} \text{true} & \text{if } d_v = z_v \vee d_v = \emptyset;\ \forall v \\ \text{false} & \text{else} \end{cases} \tag{2.28}$$

which returns *true* if a - potentially incomplete - diagnosis $\mathbf{d}$ is consistent with a complete observation $\mathbf{z}$. We will use this function for conveniently marginalizing over the missing observations. In analogy to eq. (2.25), we can compute the risk for an incomplete observation as

$$\begin{aligned} R\left(X_v = 1 \mid \mathbf{d}, \theta\right) &= \frac{P\left(\mathbf{d} \mid X_v = 1, \theta\right) P\left(X_v = 1 \mid \theta\right)}{P\left(\mathbf{d} \mid \theta\right)} \\ &= \sum_{i\,:\,\xi_{iv}=1} \frac{P\left(\mathbf{d} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right)}{P\left(\mathbf{d} \mid \theta\right)} \end{aligned} \tag{2.29}$$

where the enumerator of the second line can now be rewritten using the match function:

$$\begin{aligned} P\left(\mathbf{d} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right) &= \sum_{\left\{j\,:\,\text{match}(\mathbf{d}, \boldsymbol{\zeta}_j)\right\}} P\left(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi}_i, \theta\right) P\Big(\boldsymbol{\xi}_i \mid \theta\Big) \\ &= \sum_{\left\{j\,:\,\text{match}(\mathbf{d}, \boldsymbol{\zeta}_j)\right\}} B_{ij} \Big[p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda}\Big]_i \end{aligned} \tag{2.30}$$

In this case B_{ij} denotes the element of the observation matrix that corresponds to state $\boldsymbol{\xi}_i$ and observation $\boldsymbol{\zeta}_j$. Again, the indices $\{i : \xi_{iv} = 1\}$ in eq. (2.29) correspond to all possible states with a positive involvement in lymph node level X_v.

Essentially, the whole term is the likelihood of an observation $\mathbf{d}$ where we have removed all entries that correspond to states with $X_v \neq 1$ both from the observation matrix and the resulting probability vector of the evolution. It can therefore be easily computed algebraically, too.

The evidence in the denominator of eq. (2.29) becomes a marginalization over all possible diagnoses that are not available to us or that we deem unimportant

$$P\left(\mathbf{d} \mid \theta\right) = \sum_{\left\{j \,:\, \mathrm{match}(\mathbf{d}, \boldsymbol{\zeta}_j)\right\}} \left[p_T\left(\mathbf{t}\right) \cdot \mathbf{\Lambda}\right]_j \tag{2.31}$$

We can make this summation a bit more elegant using a column vector $\mathbf{c}^{\mathbf{d}}$ that has entries corresponding to the match-function

$$c_i^{\mathbf{d}} = \mathrm{match}(\mathbf{d}, \boldsymbol{\zeta}_i) \tag{2.32}$$

where every *true* corresponds to a 1 and every *false* to a 0. This way we can rewrite eq. (2.31) in the following way:

$$P\left(\mathbf{d} \mid \theta\right) = p_T\left(\mathbf{t}\right) \cdot \mathbf{\Lambda} \cdot \mathbf{B} \cdot \mathbf{c}^{\mathbf{d}} \tag{2.33}$$

Using this notation for marginalizing over unknown or incomplete observations also allows us to encode entire datasets $\boldsymbol{\mathcal{D}} = \begin{pmatrix}\mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_N\end{pmatrix}$ of (potentially incomplete) observations in the form of a matrix

$$\mathbf{C} = \begin{pmatrix}\mathbf{c}^{\mathbf{d}_1} & \mathbf{c}^{\mathbf{d}_2} & \cdots & \mathbf{c}^{\mathbf{d}_N}\end{pmatrix} \tag{2.34}$$

so that the row-vector of likelihoods reads as

$$P\left(\boldsymbol{\mathcal{D}} \mid \theta\right) = \left(P\left(\mathbf{d}_n \mid \theta\right)\right)_{n\in[1,N]} = p_T\left(\mathbf{t}\right) \cdot \mathbf{\Lambda} \cdot \mathbf{B} \cdot \mathbf{C} \tag{2.35}$$

2.9 Multiple diagnostic modalities

Throughout the last sections, we have only dealt with diagnoses from a single modality. In practice, however, most patients undergo screening for metastases using different modalities, like CT, MRI or FNA. The sensitivities and specificities of these might vary greatly and by combining them in a probabilistically rigorous way, we may gain a additional information.

Luckily, the introduced formalism requires very little changes to be able to incorporate multiple diagnostic modalities. Let $\mathcal{O} = \{\mathrm{CT}, \mathrm{MRI}, \mathrm{FNA}, \ldots\}$ be the set of modalities. Then we can extend the collection of observed binary random variables (RVs) $\mathbf{z}$ from a single modality

$$\mathbf{z} = \left(x_v\right)_{v\in[1,V]} = \begin{pmatrix}x_1 & \cdots & x_V\end{pmatrix} \tag{2.36}$$

to multiple diagnostic modalities

$$\mathbf{z} = \left(x_v^k\right)_{\substack{v\in[1,V]\\ k\in[1,|\mathcal{O}|]}} = \begin{pmatrix}x_1^1 & \cdots & x_V^1 & x_2^2 & \cdots & x_V^{|\mathcal{O}|}\end{pmatrix} \tag{2.37}$$

where k enumerates the elements in the set $\mathcal{O}$. We can use $\boldsymbol{\zeta}_j$ again and this time the counting variable j goes from 1 to $2^{V\cdot|\mathcal{O}|}$. Notice that this means the observation matrix $\mathbf{B}$ is not square anymore. Also, it now contains the sensitivities and specificities of all the modalities in $\mathcal{O}$. If we had separate square observation matrices $\mathbf{B}^k$ for each diagnostic modality, the new total matrix' rows B_{i*} would be the outer products of the individual observation matrices:

$$B_{i*} = B^1_{i*} \otimes B^2_{i*} \otimes \cdots \otimes B^{|\mathcal{O}|}_{i*} \tag{2.38}$$

Completely analogous to how we enlarged the vector of binary RVs $\mathbf{z}$, we can also extend the vectors $\mathbf{c}$ and $\mathbf{d}$ and then immediately use the entire formalism of the section before to model lymphatic progression with potentially incomplete diagnoses from multiple modalities. However, we will drop this way of continuously enumerating the observations in the next section again, because there is a slightly more efficient and elegant way to do it. This section only served to show that it is naturally possible to extend the formalism to combine findings from different diagnostic modalities.

2.10 Combining modalities and data

Note that the matrix $\mathbf{B}$ – and also the matrix $\mathbf{C}$ – can get very large very quickly: The former is of size $2^V \times 2^{V\cdot|\mathcal{O}|}$ and the latter has dimensions $2^{V\cdot|\mathcal{O}|} \times N$, meaning both grow exponentially with the number of LNLs *and* diagnostic modalities. And although neither $\mathbf{B}$ not $\mathbf{C}$ depend on the parameters θ, meaning their product can be precomputed, we can simply iterate over all patients, possible hidden states and available diagnostic modalities to compute $\boldsymbol{\Omega} \coloneqq \mathbf{B} \cdot \mathbf{C}$ directly, which saves us building up and multiplying matrices with potentially millions of entries.

To compute this matrix $\boldsymbol{\Omega}$, we first abandon the just-introduced way of combining diagnoses for all modalities into one large vector and separate them again, so that we have complete and incomplete observations $\boldsymbol{\zeta}^k_j$ and $\mathbf{d}^k_n$ respectively for each modality, where $n \in [1, N]$ enumerates the patients in the data.

$$\begin{aligned}\Omega_{mn} = P\left(\mathbf{d}_n \mid \boldsymbol{\xi}_m\right) &= \prod_{k=1}^{|\mathcal{O}|} P\left(\mathbf{d}^k_n \mid \boldsymbol{\xi}_m\right) \\ &= \prod_{k=1}^{|\mathcal{O}|} \left[\sum_{j\,:\,\text{match}(\mathbf{d}^k_n, \boldsymbol{\zeta}^k_j)} P\left(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi}_m\right) \right] = \prod_{k=1}^{|\mathcal{O}|} \left[\sum_{j\,:\,\text{match}(\mathbf{d}^k_n, \boldsymbol{\zeta}^k_j)} B^k_{mj} \right]\end{aligned} \tag{2.39}$$

Now, the elements Ω_{mn} encode the observation likelihood of patient n's diagnose $\mathbf{d}_n$ given their true state of involvement is $\boldsymbol{\xi}_m$. Finally, with this the row-vector of likelihoods of a cohort of patients, given the model's spread parameters, becomes

$$P\left(\boldsymbol{\mathcal{D}} \mid \theta\right) = p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Omega} \tag{2.40}$$

Again, the objects $p_T(\mathbf{t})$ and $\boldsymbol{\Lambda}$ depend on the parameters and hence need to be recalculated for every sample drawn during MCMC inference. $\boldsymbol{\Omega}$ depends only on the patient data $\mathcal{D}$ and must therefore only be computed once at the beginning of the learning round.

Chapter 3

Bilateral hidden Markov model

In the previous chapter we have set up the formalism to deal with only one side of the neck. Implicitly, we have assumed that to be the ipsilateral side, i.e. the side of the sagittal plane where the primary tumor is located. This is because we assume lymphatic drainage to a process that is somewhat symmetric w.r.t. the sagittal plane, which means there can only be limited lymph flow across this plane. But depending on the tumor's location and lateralization, drainage and hence metastatic spread to the contralateral lymphatic system of the neck may also occur. In current clinical practice, a bilateral neck dissection or irradiation is often prescribed when the tumor is close to the mid-sagittal plane. So, ideally, we would like to model the risk for involvement in both sides of the neck at the same time.

The formalism of chapter 2 can easily be applied to the contralateral side and given respective training data for the sampling process, it would learn the appropriate spread probabilities to and among the contralateral LNLs just as it would learn the ones for the ipsilateral side. From clinical experience, the contralateral involvement is usually less severe than the ipsilateral one, and hence we would expect the contralateral spread to be less probable as well.

However, combining two such unilateral models naively would make the assumption that ipsi- and contralateral spread are independent, which seems unlikely: If we know a patient has advanced metastases in the contralateral neck nodes, the risk to find similarly or even more advanced disease in ipsilateral neck nodes should probably be higher than if the contralateral neck were healthy. In other words, we are now looking for the joint probability $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid \mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}\right)$, where the superscripts i and c indicate the ipsi- and contralateral side respectively.

The following section will pick up the unilateral formalism, extend and modify it to come up with a less naive bilateral model.

3.1 Expanding the unilateral model

If we start by dissecting this joint conditional probability in the following way

$$P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid \mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}\right) = \frac{P\left(\mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}} \mid \mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right) \cdot P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right)}{P\left(\mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}\right)} \tag{3.1}$$

we notice right away that the likelihood on the right factorizes: Given the true states of involvement in the two sides of the neck, their respective diagnoses must be independent. Furthermore, the two factors are already given by their corresponding observation matrices $\mathbf{B}^{\mathrm{i}}$ and $\mathbf{B}^{\mathrm{c}}$.

The joint probability of the hidden states $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right)$ does not factorize in the same manner. But if we assume the lymphatic network to be symmetric and directed, there can be no direct connection between LNLs of the two sides of the neck, which means the probability for involvement of the ipsi- and contralateral side only correlate via the diagnose time t. Hence the joint probability is a sum of factorizing terms:

$$\begin{aligned} P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right) &= \sum_{t \in \mathbb{T}} p(t) \cdot P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid t\right) \\ &= \sum_{t \in \mathbb{T}} p(t) \cdot P\big(\mathbf{X}^{\mathrm{c}} \mid t\big) \cdot P^{\top}\big(\mathbf{X}^{\mathrm{i}} \mid t\big) \end{aligned} \tag{3.2}$$

Note that the two row vectors of probabilities in the second line are multiplied using an outer product. Using the notation from the last section, We can write this in an algebraic way to effectively factorize this sum as follows

$$P\left(\mathbf{X}^{\mathrm{c}} = \boldsymbol{\xi}_n, \mathbf{X}^{\mathrm{i}} = \boldsymbol{\xi}_m\right) = \left[\boldsymbol{\Lambda}_{\mathrm{c}}^{\top} \cdot \operatorname{diag} p(\mathbf{t}) \cdot \boldsymbol{\Lambda}_{\mathrm{i}}\right]_{n,m} \tag{3.3}$$

where the $\boldsymbol{\Lambda}$ are again matrices with rows of the conditional probabilities $P\left(\mathbf{X} \mid t\right)$ which can be computed as defined in eq. (2.18). Multiplying these two matrices – one for the contralateral side from the left and one for the ipsilateral side from the right – onto a diagonal matrix containing the time prior marginalizes over the diagnose time and results in a matrix where the value in row n and column m represents the probability to find the contralateral neck in state $\mathbf{X}^{\mathrm{c}} = \boldsymbol{\xi}_n$ and the ipsilateral lymphatic system in state $\mathbf{X}^{\mathrm{i}} = \boldsymbol{\xi}_m$.

Similarily, we can now multiply the observation matrices $\mathbf{B}$ from the left and the right onto $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right)$ to compute the bilateral equivalent of **??**:

$$P\left(\mathbf{Z}^{\mathrm{c}} = \boldsymbol{\zeta}_n, \mathbf{Z}^{\mathrm{i}} = \boldsymbol{\zeta}_m\right) = \left[\mathbf{B}^{\top} \cdot \boldsymbol{\Lambda}_{\mathrm{c}}^{\top} \cdot \operatorname{diag} p(\mathbf{t}) \cdot \boldsymbol{\Lambda}_{\mathrm{i}} \cdot \mathbf{B}\right]_{n,m} \tag{3.4}$$

Formally, all necessary terms can now be computed so that both inference and the subsequent risk prediction can be performed. However, in the next section we will go into more detail regarding how this was implemented.

3.2 Parameter symmetries and mid-line extension

Although it has been omitted, eqs. (3.1) to (3.4) are still functions of the same parameters as in the unilateral model, but each side now has their own set $\boldsymbol{\theta}^{\mathrm{c}}$ and $\boldsymbol{\theta}^{\mathrm{i}}$ of spread probabilities that are used to parameterize the transition matrices $\mathbf{A}^{\mathrm{c}}$ and $\mathbf{A}^{\mathrm{i}}$ respectively.

In principle, the spread probabilities of the two sides are entirely indepent, and a lateralized primary tumor certainly spreads to a different extend to the ipsi-versus the contralateral side. But the spread probabilities among the LNLs should be equal when assuming that the lymphatic network in the head and neck region is symmetric. This means

$$\begin{aligned} \tilde{b}_v^{\mathrm{c}} &\neq \tilde{b}_v^{\mathrm{i}} \\ \tilde{t}_{rv}^{\mathrm{c}} &= \tilde{t}_{rv}^{\mathrm{i}} \end{aligned} \quad \forall\, v \leq V \,,\, r \in \mathrm{pa}(v) \tag{3.5}$$

Using this reasonable assumption of a symmetric neck anatomy, we may avoid doubling the spread parameters when we model the bilateral lymphatic spread.

However, there are cases in which the primary tumor lies almost or exactly on the mid-sagittal plane of the patient. In such cases, we cannot reasonably distinguish between the ipsi- and contralateral side. Consequently, we must assume the base probability rates as well to be symmetric: $\tilde{b}_v^{\mathrm{c}} = \tilde{b}_v^{\mathrm{i}}$
This means there must be a continuous increase in the spread probabilities from the primary tumor to the contralateral LNLs if we were to move a patient's tumor from a clearly lateralized location closer and closer to that patient's mid-sagittal plane. Ideally, we would like to factor information about the tumor's "degree of asymmetry" into our model, e.g. by considering a normalized perpendicular distance from the mid-sagittal plane to the tumor's center of mass or by considering the tumor volume on either side of this plane. Data like this, however, is rarely available. What is frequently reported and also clinically considered as a risk factor for contralateral involvement is whether or not the tumor touches or extends over the mid-sagittal plane. With this binary variable (and the information on whether the tumor is central/symmetric w.r.t. to the sagittal symmetry plane) we can now distinguish three degrees of lateralizations:

1. $\not{\mathrm{s}}, \not{\mathrm{e}}$: The tumor does not cross or touch the mid-sagittal plane and is thus clearly lateralized. The base spread probabilities are $\{\tilde{b}_v^{\mathrm{i}}\}$ and $\{\tilde{b}_v^{\mathrm{c},\not{\mathrm{e}}}\}$.
2. $\not{\mathrm{s}}, \mathrm{e}$: The tumor is lateralized, but crosses or touches the mid-sagittal plane. We will discuss how to define the spread probabilities to the contralateral side below.
3. s, e: The tumor is symmetric w.r.t. to the sagittal plane, thus $\tilde{b}_v^{\mathrm{c,s}} = \tilde{b}_v^{\mathrm{i}}$

Note that $\mathrm{s}\,(\not{\mathrm{s}})$ and $\mathrm{e}\,(\not{\mathrm{e}})$ denote the two binary variables *symmetric* (or *not symmetric*) and *extending* (or *not extending*) over the mid-sagittal plane.
We can infer that in case 2 the spread probabilities to the contralateral LNLs must be between the ones for the clearly lateralized (1) and the symmetric (3) case. Hence, we introduce a new "mixing" parameter α that defines the contralateral spread from tumor to the LNLs as a linear superposition between the two extremes:

$$\tilde{b}_v^{\mathrm{c,e}} = \alpha \cdot \tilde{b}_v^{\mathrm{i}} + (1 - \alpha) \cdot \tilde{b}_v^{\mathrm{c},\not{\mathrm{e}}} \tag{3.6}$$

This new mixing parameter must be inferred from data just like the other spread probabilities and the parametrization of the time prior.
When using the learned parameters to predict the risk of a new patient g, the set of parameters for the risk computation $\hat{\boldsymbol{\theta}}_g$ is compiled from the total set of inferred parameters $\hat{\boldsymbol{\theta}} = \{\tilde{b}_v^{\mathrm{i}}, \tilde{b}_v^{\mathrm{c},\not{\mathrm{e}}}, \alpha, \tilde{t}_{rv}, p_T\}$, depending on the risk factors the patient presents with at the time of diagnosis. As always, for $\hat{\boldsymbol{\theta}}$ we have $v \leq V$,

$r \in \mathrm{pa}(v)$ and the T-stage $T \in \{1, 2, 3, 4\}$. For example, if patient g has a T1 tumor that is clearly lateralized, their $\hat{\boldsymbol{\theta}}_g$ may be computed in the following way:

$$\hat{\boldsymbol{\theta}}_g = \left\{ \tilde{b}_v^{\mathrm{i}}, \tilde{b}_v^{\mathrm{c}} = \tilde{b}_v^{\mathrm{c},\not{\mathrm{e}}}, \tilde{t}_{rv}, p_1 \right\} \tag{3.7}$$

while another patient m with a T3 tumor that clearly crosses the mid-sagittal plane would have the following set of parameters used for their risk prediction:

$$\hat{\boldsymbol{\theta}}_m = \left\{ \tilde{b}_v^{\mathrm{i}}, \tilde{b}_v^{\mathrm{c}} = \alpha \cdot \tilde{b}_v^{\mathrm{i}} + (1 - \alpha) \cdot \tilde{b}_v^{\mathrm{c},\not{\mathrm{e}}}, \tilde{t}_{rv}, p_3 \right\} \tag{3.8}$$

In the actual computational implementation of this model, we essentially compute three different matrices $\boldsymbol{\Lambda}$ which are functions of different parameters:

$$\begin{aligned} \boldsymbol{\Lambda}_{\mathrm{i}} &= \boldsymbol{\Lambda}\left(\tilde{b}_v^{\mathrm{i}}, \tilde{t}_{rv} \right) \\ \boldsymbol{\Lambda}_{\mathrm{c},\not{\mathrm{e}}} &= \boldsymbol{\Lambda}\left(\tilde{b}_v^{\mathrm{c},\not{\mathrm{e}}}, \tilde{t}_{rv} \right) \\ \boldsymbol{\Lambda}_{\mathrm{c,e}} &= \boldsymbol{\Lambda}\left(\alpha, \tilde{b}_v^{\mathrm{c},\not{\mathrm{e}}}, \tilde{b}_v^{\mathrm{i}}, \tilde{t}_{rv} \right) \end{aligned} \tag{3.9}$$

From those, the likelihoods of all patients in the training data can be computed when used with the respective p_T – that gives rise to the corresponding $\operatorname{diag} p(\mathbf{t})$ – as in eq. (3.4).

3.3 Comparing bilateral models

Up to this point we have largely argued that the mixing parameter makes intuitive sense because of the thought experiment, where we moved the primary tumor from a clearly lateralized position closer and closer to the mid-sagittal plane, until it was perfectly symmetric w.r.t. that plane. However, we now need to actually test whether our arguments hold. For that, we decided to compare three models:

- Model $\mathcal{M}_{\mathrm{ag}}$, which is agnostic to the tumor's extension e over the mid-sagittal plane and treats the contralateral base spread in the same way for all patients.
- Model $\mathcal{M}_{\alpha}$ that uses the linear combination of the ipsilateral base probabilities and the contralateral ones for the patients without mid-plane extension to describe the spread for tumors which do extend over that plane.
- Model $\mathcal{M}_{\mathrm{full}}$, going even further by defining a completely independent set of contralateral base probabilities for the patients whose tumor extends over the mid-sagittal plane.

Essentially, we now want to know which of these three models does the best job of describing the data. Intuitively, one would argue that it must be $\mathcal{M}_{\mathrm{full}}$, but this model is also more complex than the other two. A natural choice for a metric that incorporates both the accuracy of the model and a penalty for model complexity – often also called *Occam's razor* – is the *model evidence* [1].

Model evidence and Bayes factor

In Bayesian terms, we would like to know which model $\mathcal{M}$ has the highest probability $P(\mathcal{M} \mid \boldsymbol{\mathcal{D}})$ given a dataset $\boldsymbol{\mathcal{D}}$. This probability is given by

$$P(\mathcal{M} \mid \boldsymbol{\mathcal{D}}) = \frac{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M})\, P(\mathcal{M})}{P(\boldsymbol{\mathcal{D}})} \tag{3.10}$$

If a priori all models we want to consider have the same probability $P(\mathcal{M})$ and we only make pairwise comparisons between models, then we can restrict ourselves to computing the *Bayes factor*:

$$K_{1\text{v}2} = \frac{P(\mathcal{M}_1 \mid \boldsymbol{\mathcal{D}})}{P(\mathcal{M}_2 \mid \boldsymbol{\mathcal{D}})} = \frac{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}_1)\, P(\mathcal{M}_1)}{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}_2)\, P(\mathcal{M}_2)} = \frac{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}_1)}{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}_2)} \tag{3.11}$$

On the right side in the above equation, we see the ratio of the two model's evidences, which are merely their respective likelihoods, marginalized over all parameters:

$$P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}) = \int p(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M})\, p(\theta \mid \mathcal{M}) d\theta \tag{3.12}$$

So, if we can compute this model evidence – commonly also called *marginal likelihood* or *partition function Z* from physics – for our models $\mathcal{M}_{\text{ag}}$, $\mathcal{M}_\alpha$ and $\mathcal{M}_{\text{full}}$, the respecive pairwise Bayes factors will indicate which of them is *most likely* to be the true one, given the observed data, in the probabilistic sense. Note that this does not mean it *is* the true data-generating model and not even that we should *believe* it is the true one. But only that among the models investigated, this one is probably the best.

Harold Jeffreys gives a scale for interpreting values of the Bayes factor [21]:

$K_{1\text{v}2}$	$\ln K_{1\text{v}2}$	support for $\mathcal{M}_1$
$< 10^0$	< 0	negative evidence (supports $\mathcal{M}_2$)
10^0 to $10^{1/2}$	0 to 1.15	barely worth a mention
$10^{1/2}$ to 10^1	1.15 to 2.3	substantial
10^1 to $10^{3/2}$	2.3 to 3.45	strong
$10^{3/2}$ to 10^2	3.45 to 4.6	very strong
$> 10^2$	> 4.6	decisive

We have also listed the natural logarithm $\ln K_{1\text{v}2}$ of the Bayes factor here, because what we will actually be doing is compute differences in the log-evidences.

Thermodynamic integration

Due to the integration over all model parameters, the quantity eq. (3.12) is usually impossible to calculate by brute force integration, even for models with only around a dozen parameters, as is the case for ours. Unless analytical solutions exist – which is rarely the case – it is often prohibitively expensive to compute the model evidence. For this reason, a large amount of approximation methods has been developed; [13] names only a few of those methods that can be used in the context of MCMC. Another method that is applicable in the context of MCMC is thermodynamic integration (TI), which is very well introduced in [1] and only roughly sketched out in this section.

The concept of TI originates from the field of statistical mechanics and can be motivated from that standpoint. And although this path is certainly more educational and might convey a deeper understanding w.r.t. thermodynamics and information theory, we will take a more direct approach by starting with what we want to compute and subtracting a 0 from it:

$$\begin{aligned}\ln Z := \ln p(\boldsymbol{\mathcal{D}} \mid \mathcal{M}) &= \ln \int p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right) p(\theta \mid \mathcal{M}) d\theta - \ln 1 \\ &= \ln \int p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right) p(\theta \mid \mathcal{M}) d\theta - \underbrace{\ln \int p(\theta \mid \mathcal{M}) d\theta}_{\ln Z_0}\end{aligned} \tag{3.13}$$

Writing it as this difference between two different log-evidences $\ln Z$ and $\ln Z_0$ itself does not get us far. But if we could somehow parametrize a differentiable path between the two, then maybe the integration

$$\ln Z - \ln Z_0 = \int_0^1 \frac{d}{d\beta} \ln Z_\beta d\beta \tag{3.14}$$

we end up with can actually be computed. Just by inspection of eq. (3.13) and eq. (3.14), one can see that on such differentiable path could be built using what we are going to call the *power posterior* $p_\beta(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M})$:

$$\begin{aligned}\ln Z_\beta &= \ln \int p_\beta(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M}) d\theta \\ &= \ln \int p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right)^\beta p(\theta \mid \mathcal{M}) d\theta\end{aligned} \tag{3.15}$$

with the derivative

$$\begin{aligned}\frac{d}{d\beta} \ln Z_\beta &= \int \frac{p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right)^\beta p(\theta \mid \mathcal{M})}{Z_\beta} \ln p(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}) d\theta \\ &= \mathbb{E}\left[\ln p(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M})\right]_{p_\beta(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M})} \\ &\approx \frac{1}{S} \sum_{i=1}^{S} \ln p(\boldsymbol{\mathcal{D}} \mid \hat{\theta}_{\beta i}, \mathcal{M}) = \mathcal{A}_{\mathrm{MC}}(\beta)\end{aligned} \tag{3.16}$$

The solution to computing the evidence now lies in sight: Using MCMC, we can draw samples from the power posterior p_β and use those samples to compute the expectation over the (unmodified) likelihood. Doing this for a sufficient number of steps in the interval $[0, 1]$ and integrating over the resulting $\mathcal{A}_{\mathrm{MC}}(\beta)$ will then yield an approximation to the log-evidence.

$$\ln Z \approx \frac{1}{2} \sum_{j=1}^{N-1} \left(\beta_{j+1} - \beta_j\right) \left(\mathcal{A}_{\mathrm{MC}}(\beta_{j+1}) + \mathcal{A}_{\mathrm{MC}}(\beta_j)\right) \tag{3.17}$$

This approximation gets better with larger values for S and N. But also how the β_j are chosen is crucial for computing a good estimate: Usually, the $\mathcal{A}_{\mathrm{MC}}(\beta)$ – which can be seen as accuracy terms – rise steeply for increasing β close to 0, while levelling off towards $\beta = 1$. It therefore makes sense to distribute the ladder of these values unevenly. A common choice, that we employed as well, was $\beta_j = x_j^5$ where the x_j are linearly spaced within the interval $[0, 1]$. This yields a very fine resolution for the first steps and gets successively coarser towards the end of the interval.

Lastly, we would like to give a final insight into the evidence that is quite naturally obtained when following the derivation from statistical physics, but hard to see with the brief, formal derivation we gave up to this point. Therefore, we will just state it below and point to a publication giving a nice example of how to get to this result [1]. According to this, the log-evidence can be written in the following form:

$$\ln Z = \underbrace{\int \ln p\,(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}) p\,(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M})\, d\theta}_{\text{accuracy } \mathcal{A}(\beta=1)} - \underbrace{\int \ln \frac{p\,(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M})}{p\,(\theta \mid \mathcal{M})} p\,(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M})\, d\theta}_{\text{complexity (KL-divergence)}} \tag{3.18}$$

This shows how the evidence naturally incorporates Occam's razor. The second term on the right gets larger the more the likelihood restricts the prior and the resulting penalty grows exponentially with the dimensionality of the parameter space.

Implementation

To compare the introduced models $\mathcal{M}_{\text{ag}}$, $\mathcal{M}_{\alpha}$ and $\mathcal{M}_{\text{full}}$, we performed TI with a ladder of 64 β values with step sizes selected according to a fifth order power rule. For each of the steps in the ladder, we performed an ensemble sampling round using the `emcee` [12] Python package. The size of the ensemble – consisting of so-called walkers that allow sampling in parallel and mutually influence each other's proposals – was chosen to be 20 times the number of dimensions of the parameter space. We set the sampling algorithm to propose new samples according to a mixture of two procedures: with 80% probability it selected a differential evolution move [26] and with 20% probability a snooker move, also based on differential evolution [5]. The reason for this choice was that in previous experiments, this combination of proposals yielded the fastest convergence of the chain. Every one of the 64 sampling rounds consisted of a burn-in phase lasting 1000 steps, followed by 250 steps of which every fifth step was kept for later analysis. This might seem like a relatively short chain, but since the change of the posterior we sampled from only changed very slightly from β_j to β_{j+1}, fewer steps are required to reach convergence.

In the end, we kept $S = 50 \cdot k$, where k is the dimensionality of the model, samples for each of the 64 β_j. The dimensionality k of the parameter spaces ranged from nine for the agnostic model $\mathcal{M}_{\text{ag}}$ over ten (mixing model $\mathcal{M}_{\alpha}$) to twelve in the case of the full model $\mathcal{M}_{\text{full}}$. Out of these S samples we randomly drew $M = 1000$ per β_j and integrated them over their range, yielding 1000 estimates for the log-evidence $\ln Z_l$ with $l \in [1, \dots, M]$. Using this ensemble of estimates, we could compute both the mean and the standard deviation, giving us a simple measure of uncertainty for that value.

From the samples drawn at $\beta_{64} = 1$, we also computed the Bayesian information criterion (BIC), which is in essence a first-order approximation of the log-evidence (actually, the negative one-half of the BIC is an approximation to the log-evidence) [32]. It is defined as

$$\mathrm{BIC} \coloneqq k \ln N - 2 \max_{\theta} \left(\ln p \left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M} \right) \right) \tag{3.19}$$

where – again – k is the number of parameters θ, while N is the number of patients in the dataset $\boldsymbol{\mathcal{D}}$. How reliable the BIC is for a given model, depends on whether its core assumption hold: 1) the posterior must be unimodal and decay rapidly outside its maximum, while N must be much larger than k [3]. The second assumption is not quite fulfilled, but we will see shortly that the BIC seems to be a quite good measure for model comparison in our case.

The models were trained on the combination of two datasets: One from our institution, the University Hospital Zurich, which has been published and described in great detail in a separate publication [23]. The other was kindly provided to us by researchers of the Centre Léon Bérard in Lyon, France and was the underlying data for one of their papers [2]. Both datasets have been published in our repository `lydata`.

Different modalities were used to obtain the diagnoses for the patients in the two datasets. For the inference process, we combined all available diagnostic modalities using sensitivity and specificity values from the literature [10] using a maximum likelihood estimate. We treated this resulting ”consensus diagnosis” as if it were the ground truth, i.e., we set its sensitivity and specificity to 1 respectively. Our motivation to do so was that this allows us to compare predictions of the model with data prevalences to see which of the model exhibits more flexibility in adapting to the data. If we had directly provided the models with all available diagnostic modalities and allowed it to combine them itself, as outlined in section 2.9, this would not have been possible. Also, in this case we are not interested in learning the exact distribution over the posterior parameters of the model, i.e. the probability rates for spread along arcs of the lymphatic graph, but rather how well the different models are able to adapt to realistic data and make use of the additional information provided via the tumor's extension over the mid-sagittal plane.

Lastly, we restricted ourselves to modelling the LNLs II, III and IV, because contralaterally we rarely observe involvement outside those levels and it drastically speeds up the inference process.

Reproducibility

Each of the three models investigated here are available in `lynference`, where we have run the respective pipeline, pushed it as a tagged commit to GitHub and published it as a release.
The `README.md` file in this repository explains how one can download the data our pipeline runs produced and how to reproduce an experiment yourself.

- Model $\mathcal{M}_{\text{ag}}$: `bilateral-v2`
- Model $\mathcal{M}_{\alpha}$: `midline-with-mixing-v2`
- Model $\mathcal{M}_{\text{full}}$: `midline-without-mixing-v2`

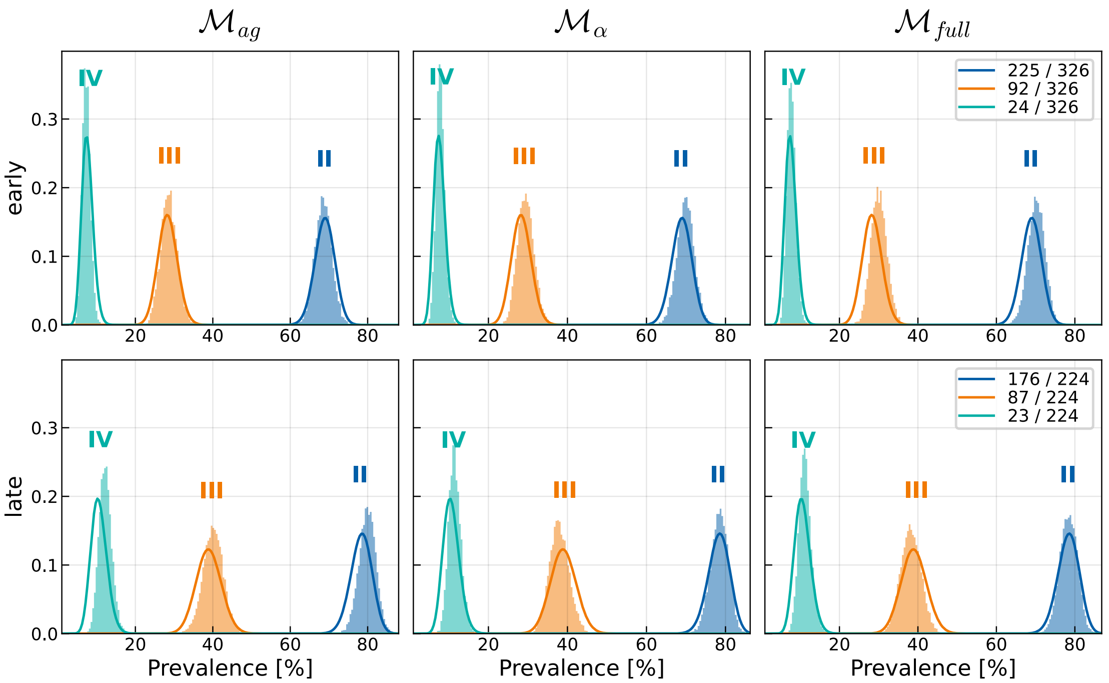

Figure 3.1: Predicted prevalences (shaded histograms) and posterior beta distributions of observed prevalences (solid lines) for the ipsilateral levels II (blue), III (orange) and IV (green). These prevalences have each been plotted for early T-stage patients (top row) and late T-stage (bottom row) and for the three models $\mathcal{M}_{\text{ag}}$ (left column), $\mathcal{M}_{\alpha}$ (middle column) and for $\mathcal{M}_{\text{full}}$ (right column). The differences between the models are negligible.

Results

First, we wanted to make sure that all three models are still able to describe the ipsilateral spread sufficiently well. We have plotted the prevalence our trained models predict in the forms of histograms against the Beta-posterior of the observed prevalence in the data (fig. 3.1). These plots were created by computing the respective prevalence for samples drawn during the final 250 steps at the end of the TI process of which every fifth step was discarded.

The shown differences between the model's predictions are miniscule. For late T-stages (bottom row of fig. 3.1) it seems as if the model that is agnostic to the tumor's extension over the mid-sagittal plane slightly overestimates the prevalence, while the other two models seem to match them better or underestimate them by a small amount. Overall the fit of all models ipsilaterally is very good and shows no indication that one model performs better than the other.

On the contralateral side, however, this does not hold anymore. Here, we do not only stratify the prevalence by T-stage, but also by midline extension. Naturally, this cannot be captured the agnostic model $\mathcal{M}_{\text{ag}}$ since it has no method of modelling this. What is of interest to us here is how the mixing model $\mathcal{M}_{\alpha}$ and the full model $\mathcal{M}_{\text{full}}$ fare against each other and whether their improvements in predicting contralateral spread are worth the additional complexity.

The overall prevalence of contralateral involvement is plotted in fig. 3.2. Again, the

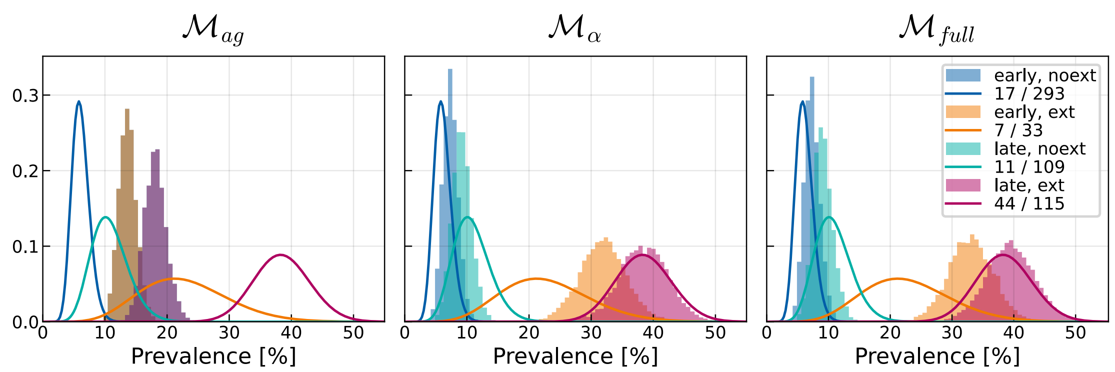

Figure 3.2: Predicted prevalences (shaded histograms) and posterior beta distributions of observed prevalences (solid lines) for the contralateral overall involvement (anything *not* clinically N0, on that side of the neck). Predicted and observed prevalence for early T-stage is colored blue and orange, while for late T-stage it is green and red. The prevalence for patients whose tumor does not extend over the mid-sagittal line is labelled `noext` and colored blue or green, while the same quantity for those with said extension is labelled `ext` and colored orange and red. The three models $\mathcal{M}_{\text{ag}}$, $\mathcal{M}_{\alpha}$ and $\mathcal{M}_{\text{full}}$ are depicted in the left, middle and right panel respectively.

three different models are depicted in their own column and we have distinguished between four cases for each model: The prevalence of any contralateral involvement for patients with a) early T-stage and a cealry lateralized tumor (blue histograms and curves), b) early T-stage with a tumor extending over the mid-sagittal plane (orange), c) late T-stage with, again, a lateralized tumor (green) and finally d) where the tumor is both in late T-stage and does extend over the mid plane (red).

As discussed, the agnostic model $\mathcal{M}_{\text{ag}}$ (left panel in fig. 3.2) cannot model midline extension, which is why the two separate histograms overlap. Its spread probability rates from the tumor to the contralateral LNLs attempt to find an average of the respective observed prevalence. Interestingly, both the model using the mixing parameter α and the full model, which has in total six parameters to model the spread from the tumor to the contralateral LNLs, overestimate the prevalence of contralateral, early T-stage involvement when the tumor extends over the mid-sagittal line. On the one hand, of the displayed cases, this is the rarest one, so it makes sense for the model to put less attention to it. On the other hand, we believe the reason also has to do with how the model treats the mid-plane extension in general: If this binary RV is observed to be true, the model assumes increased spread probabilities from the primary tumor to the contralateral LNLs from the onset. Realistically, however, this is probably not how the course of a typical patient plays out. As tumors grow over time, in many cases they will not cross the mid-sagittal plane right after they started to form, but only when they have become sufficiently large. Consequently, the spread from them to the contralateral side increases as they grow over the midline. The current model cannot capture this and therefore assume early T-stage patient's contralateral spread to be larger than observed. Despite this, the two models perform equally well regarding the overall contralateral spread. In combination with the unaffected capabilities to predict the ipsilateral prevalences, as shown in fig. 3.1, this indicates that the

Metric	**agnostic** $\mathcal{M}_{\text{ag}}$	**mixing** $\mathcal{M}_{\alpha}$	**full** $\mathcal{M}_{\text{full}}$
BIC	-1116.70	-1093.08	-1098.23
log-evidence	-1118.23 ± 1.77	-1093.33 ± 1.91	-1099.73 ± 1.99
max likelihood	-1088.31	-1061.53	-1060.37
$\mathcal{A}_{\text{MC}}(1)$	-1092.25	-1065.70	-1065.50

Table 3.1: Metrics computed via TI for the three bilateral models introduced in section 3.3: The BIC in the first row, the log-evidence $\ln Z$ with the respective standard deviation in the second and third row, as well as the maximum and mean likelihood of the sampling procedures in rows four and five respectively.

assumptions underlying the introduction of the mixing parameter α are feasible.

Of course one would expect that maybe modelling the correlations between involvements of the contralateral LNLs might suffer from this assumption, but this is hard to test, as cases where e.g. LNL III is involved without LNL II are very rare – in this case it is only five patients. And also clinically it is debated whether to treat or to spare the contralateral side as a whole when performing elective radiotherapy (RT) or elective bilateral neck dissection (ND), not individual LNLs [4, 25]. Therefore, a more complete model like $\mathcal{M}_{\text{full}}$, that might be able to capture correlations we cannot yet see due to insufficient data, are not worth the additional model complexity at this point.

This is supported also by the log-evidence of the three models compared, which we tabulated in . In **??** we have plotted the results from computing the log-evidence for the three models in question using TI. It shows that the accuracy of the agnostic model $\mathcal{M}_{\text{ag}}$ is lower than of the other two models, which owe that to their ability to incorporate the tumor's extension over the mid plane into the prediction. However, while the mixing model $\mathcal{M}_{\alpha}$ and the full model $\mathcal{M}_{\text{full}}$ achieve the same fit to the data, the full model's accuracy rises for later β values, which results in a lower log-evidence and a higher complexity penalty (see eq. (3.18)) compared to the simpler mixing model.

,,

,,,,

Chapter 4

Possible further extension

In this section we will discuss possible extensions and improvements that we considered during the development. Ultimately, we decided against implementing them into the model we presented in the previous chapters due to reasons we will layout in the following subsections.

4.1 Trinary hidden random variables

As mentioned in the beginning, the reason for developing a probabilistic model of lymphatic tumor progression in the first place is that modern imaging modalities cannot detect tumor cells directly, but only the (macroscopic) changes they exert on their region of growth, e.g. when they cause lymph nodes to swell. In contrast, a histopathological examination of a resected malignancy or biopsy sample uses various staining techniques in combination with microscopes to detect carcinoma cells directly.

That raises the question whether clinical imaging and pathological examinations can both be modelled as observations of the same hidden random variable. The current resolution of the former is magnitudes away from being able to directly detect cancer cells and yet we consider the chance of detecting a microscopic involvement to be given by the sensitivity of MRI, CT, etc., the same one even as the chance for detecting macroscopic metastases. In other words: The sensitivity for detecting lesions far smaller that the voxel resolution of imaging is zero.

To account for this issue, one could categorize lymphatic involvement a little more finely: Instead of treating the true state of a LNL as a binary random variable representing a health node and a metastaic node respectively, we could cosider micro- and macroscopic involvement separately. The hidden states were then modelled as *trinary* hidden variables (see eq. (1.1) for comparison):

$$\begin{aligned} \text{hidden} \quad \mathbf{X} &= (X_v) \to \{0, \mu, M\}^V \\ \text{observed} \quad \mathbf{Z} &= (Z_v) \to \{0, 1\}^V \end{aligned} \tag{4.1}$$

where μ and M now respectively represent *microscopic* and *macroscopic* involvement separately.

Diagnoses can still only be binary and in order to decide on a treatment, i.e. whether to irradiate the LNL in question, we also still only care about the distinction *cancerous* – meaning there are malign cells present – vs *healthy*. However, the trinary hidden state represents the underlying reality much more precisely: To an imaging modality the hidden states 0 and μ are healthy states, since it cannot detect microscopic disease, and the respective probabilities for true and false negative observations is governed by the specificity of the modality. To a pathologist however, it is the other way around: The hidden states μ and M appear as truly involved and the probability to correctly identify these states as involved is given by the sensitivity (which we usually assume to be one for pathology).

Now, it seems that with binary observations of an underlying trinary state we cannot expect to correctly infer the true state of the LNL. But imagine that we have data on diagnostic CTs and/or MRIs taken before a neck dissection. After the surgery, the resected LNLs are examined by a pathologist. The outcome is as follows:

$$\begin{aligned} Z_v^{\text{MRI}} &= 0 \\ Z_v^{\text{path}} &= 1 \end{aligned} \tag{4.2}$$

From the second observation we can immediately infer that the hidden state must be either μ or M, because it means tumor cells have been found in the LNL. The MRI diagnosis gives us the additional information that the probability for state μ is $P(X_v = \mu) = s_P$, i.e. the specificity. For state M it is $p(X_v = M) = 1 - s_N$, where s_N is the sensitivity. Obviously, we can infer the likely true hidden state by combining observations of different kinds of modalities: Imaging on the one hand and FNA and pathology on the other hand.

The inference above might seem unnecessary at first: Could the pathologist not simply distinguish the three hidden states? But the distinction between micro- and macroscopic metastases is actually done by the imaging modality. Any occult disease that cannot be detected by them is considered a microscopic disease. This ability depends on many factors, like presence of necrotic tissue and also characterisitics of the machine used to obtain the scans, that make the distinction fuzzy. Clearly, pathologists cannot routinely think about whether the disease they see would already have been visible on a scan or not and consequently, the distinction between micro and macro is not done in pathology.

Chapter 5

Detailed patient-individual reporting of lymph node involvement in oropharyngeal squamous cell carcinoma with an online interface

One critical aspect of our effort to model and predict the lymphatic tumor progression is the data we use to train the model. As previously explained, our model essentially consumes tables with rows of patients and columns involvement by LNL. Data in this relatively simple format has been extracted in the past to create studies like [7] or [33]. However, the authors then used the data to compute statistics of it – e.g. the prevalence of involvement – but stopped short of publishing that data in its raw format. From these statistics it is – with one exception [31] – usually not possible to reconstruct the correlations between the involvement of LNLs.

With almost no usable data, of course, our methodology for modelling lymphatic progression cannot be tested or applied. So, we decided to start at the University Hospital Zurich (USZ) to extract all patterns of lymphatic progression in patients with newly diagnosed oropharyngeal squamous cell carcinoma (OPSCC) between 2013 and 2019. We then not only used that data for inference on it, but also published it freely, hoping that other researchers might find it useful and that it may even motivate them to share their data in a similar fashion in the future.

In the following, I will include large parts of the publication [**ludwig˙detailed˙2022**], in which we detailed the extraction of the dataset, its characteristics and how we made it available. It is important to note that the first authorship is shared in this publication: **Jean-Marc Hoffmann**, a radiation oncologist at the USZ, extracted most of the data from digital patient and imaging records. **Bertrand Pouymayou**, a medical physicist and postdoctoral researcher at the USZ built up a complex template for easier extraction and storage of the patient information. He also created the initial interface for viewing the data. My contribution to this work was the processing of the data, creating figures and tables for the publication, host the cohort in the form of a comma separated values (CSV) table in online repositories and, lastly, develop and deploy an online interface akin to what Bertrand Pouymayou had implemented earlier (see **??** for more details).

5.1 Abstract

Purpose/Objective

Whereas the prevalence of LNL involvement in HNSCC has been reported, the details of lymphatic progression patterns are insufficiently quantified. In this study, we investigate how the risk of metastases in each LNL depends on the involvement of upstream LNLs, T-category, Human Papillomavirus (HPV) status and other risk factors.

Results

We retrospectively analyzed patients with newly diagnosed OPSCC treated at a single institution, resulting in a dataset of 287 patients. For all patients, involvement of LNLs I-VII was recorded individually based on available diagnostic modalities (PET, MRI, CT, FNA) together with clinicopathological factors. To analyze the dataset, a web-based graphical user interface (GUI) was developed, which allows querying the number of patients with a certain combination of co-involved LNLs and tumor characteristics.

Results

The full dataset and GUI is part of the publication. Selected findings are: Ipsilateral level IV was involved in 27% of patients with level II and III involvement, but only in 2% of patients with level II but not III involvement. Prevalence of involvement of ipsilateral levels II, III, IV, V was 79%, 34%, 7%, 3% for early T-category patients (T1/T2) and 85%, 50%, 17%, 9% for late T-category (T3/T4), quantifying increasing involvement with T-category. Contralateral levels II, III, IV were involved in 41%, 19%, 4% and 12%, 3%, 2% for tumors with and without midline extension, respectively. T-stage dependence of LNL involvement was more pronounced in HPV negative than positive tumors, but overall involvement was similar. Ipsilateral level VII was involved in 14% and 6% of patients with primary tumors in the tonsil and the base of tongue, respectively.

Conclusions

Detailed quantification of LNL involvement in HNSCC depending on involvement of upstream LNLs and clinicopathological factors may allow for further personalization of elective clinical target volume (CTV-N) definition in the future.

5.2 Introduction

HNSCC spread through the lymphatic system of the neck and form metastases in regional lymph nodes. Therefore, the target volume in radiotherapy of HNSCC patients includes, in addition to the primary tumor, parts of the lymph drainage volume [4], [18]. The nodal gross tumor volume nodal gross tumor volume (GTV-N) contains detectable macroscopic lymph node metastases, while the elective clinical target volume CTV-N contains parts of lymph drainage volume that is at risk of harboring microscopic tumor, i.e. occult metastases that are not yet visible with current imaging techniques.

GTV-N definition is primarily performed through imaging techniques (PET-CT/MRI, MRI or CT) as well as FNA. Imaging criteria for lymph node metastases include size, round rather than oval shape, central necrosis, and FDG uptake as summarized by Biau et al [4]. Goel et al. gives an overview over clinical practice in PET/CT for the management of HNSCC [14]. However, all imaging techniques have finite sensitivity and specificity [28], [22], [30], i.e. they fail to detect small metastases or may incorrectly identify suspicious lymph nodes as tumor.

For standardized reporting of the location of lymph node metastases as well as delineation of the CTV-N, the lymph drainage system of the neck is divided into anatomically defined regions called LNL [16]. CTV-N definition amounts to the decision which LNLs to include into the elective CTV-N and is based on international consensus guidelines. Such guidelines were first published by Grégoire et al in 2000 and have been updated in 2006, 2014 and 2019 [4], [16], [15], [17]. Current recommendations for the selection of lymph node levels in OPSCC can be found in Table 2 of the guidelines published in 2019 by Biau et al. [4]. Current guidelines are primarily based on the prevalence of LNL involvement for a given primary tumor location, i.e. the percentage of patients diagnosed with metastases in each level. It is recommended that the elective CTV-N includes all LNLs that are involved in 10–15% of patients or more. Patients are primarily stratified by primary tumor location. For example, tumors of the soft palate, the posterior pharyngeal wall and the base of tongue show lymph node metastases on both sides via crossing lymph vessels. For this reason, even for lateralized tumors of these localizations, bilateral neck treatment is recommended. However, the lymphatic drainage of the tonsil is mainly unilateral, therefore an ipsilateral irradiation is recommended for lateralized low T-category (T1/T2) tumors (at least up to lymph node stage N2a). Volume-reduced elective nodal irradiation has been or is being investigated in several trials [6], [25].

While the general patterns of lymph drainage in the neck is understood and prevalence of LNL involvement has been reported in the literature [15], [7], [20], [2], the details of progression patterns in OPSCC are poorly quantified. How much does the risk of level IV involvement increase depending on whether levels II and III harbors macroscopic metastases? How much does the risk of involvement increase for late versus early T-category? Are progression patterns different for HPV positive versus HPV negative tumors? Answering these questions quantitatively may allow for further personalizing CTV-N definition based on an individual patient's clinical presentation at the time of diagnosis.

The basis for better quantification of LNL involvement are detailed datasets of HNSCC patients for whom involvement is reported per individual LNL together with tumor and patient characteristics. For example, to answer the question of how much the risk in level IV increases depending on the involvement of upstream levels II and III, it is insufficient to only report the prevalence of LNL involvement in levels II, III, and IV. Instead, the observed frequency of certain involvement combinations must be known, e.g. how often levels II, III and IV are involved simultaneously, versus how often only the levels II and III are involved without level IV. The contributions of this work are:

- We provide a dataset of lymphatic progression patterns in 287 OPSCC patients treated at our institution in whom involvement of LNLs together with tumor characteristics are reported on a patient-individual basis.
- To visualize and explore the complex dataset, a graphical user interface is provided that allows the user to query the number of patients who were diagnosed with a specific combination of simultaneously involved LNLs and tumor characteristics.

We hope that this work provides the basis for collecting large multicenter datasets of lymphatic progression patterns, which can then inform future guidelines on further personalized CTV-N definition.

5.3 Material & methods

Data curation

We included patients diagnosed with OPSCC (primary diagnosis) between 2013 and 2019 and treated at the department of radiation oncology and/or head and neck surgery of the USZ. Patients with prior radiotherapy or surgery to the neck were excluded, resulting in a dataset of 287 patients. Specific subsites of oropharyngeal cancer included the base of tongue, the tonsils as well as the oropharyngeal side of the vallecula and the posterior or lateral wall of the oropharynx. Patient information consisted of the date of birth, gender, the date of the 1st histological confirmation of the tumor, the performed treatment (surgery with neck dissection prior to RT/RCHT vs. surgery only vs. definitive radio(chemo)therapy), risk factors such as nicotine abuse and HPV-status (p16 pos/neg), the TNM-classification (UICC 7th edition until 2017, 8th edition since 2017), the position of the primary tumor (left/right neck) as well as positive vs. negative mid-sagittal plane extension. Further details are described in the accompanying data-in-brief article [23].

The analysis of the lymphatic spread included levels Ia, Ib, IIa, IIb, III, IV, V, VII and was performed separately for the diagnostic imaging modalities available for a patient (FDG PET-CT, FDG PET-MRI, MRI, CT) as well as FNA and radiotherapy planning CT if available. This was performed by 2 experienced radiation oncologists by reviewing radiology and pathology reports together with the diagnostic images. Criteria for considering a lymph node as malignant followed the description in Biau et al [4] and are described in detail in the data-in-brief article [23].

Data base

The full dataset is available as a CSV-file via the data-in-brief article linked to this publication [23] and on GitHub at `https://github.com/rmnldwg/lydata` in a folder named `2021-usz-oropharynx`.

In addition, the dataset has been archived and given a persistent identifier: `https://doi.org/10.5281/zenodo.6024778`.

Graphical user interface

We developed an online GUI based on the Django framework [11] and provide it to explore the dataset. It allows the user to conveniently determine the number of patients that show a particular combination of co-involved LNLs and tumor characteristics. The GUI is available at `https://2021-oropharynx.lyprox.org`; its source code under MIT license is available on GitHub at `https://github.com/rmnldwg/lyprox`. Documentation is provided within the GUI; a video demonstrating the use of the GUI is available online.

5.4 Results

In this section, we summarize selected findings. To that end, LNL involvement based on CT, MRI, PET and FNA was converted into a consensus decision via a logical OR, i.e. a LNL is considered involved if it was positive for one of these 4 modalities. However, only a small fraction of the information contained in the full dataset can be summarized here in tables. Thus, the interested reader is encouraged to explore the full dataset directly.

Graphical user interface

Figure 5.1 shows the GUI for analyzing the dataset. The main patient characteristics (smoking status, HPV status, and primary treatment) are shown on the top left. In the bottom-left panel, the user can specify characteristics of the primary tumor. In the example, the user considers all subsites combined, but restricts the patient selection to late T-category tumors (T3/T4) that extend over the midsagittal plane. On the top-right, the user selected the diagnostic modalities CT, MRI, PET and FNA, which are connected via a logical OR, i.e. a lymph node level is considered as involved for a patient if it was considered positive for at least one of the modalities available for that patient. The main panel shows the involvement of LNLs. In the example, the user restricts the selection to patients with positive ipsilateral level III while all other levels are unspecified. In total, 38 out of 287 patients meet these selection criteria. The main panel now displays the number of patients with involvement of the other levels. E.g. 21 patients have contralateral level II involved, and 12 patients ipsilateral level IV.

Prevalence

Overall prevalence of lymph node level involvement is reported in table 5.1 and visualized in fig. 5.2 panel a and b.

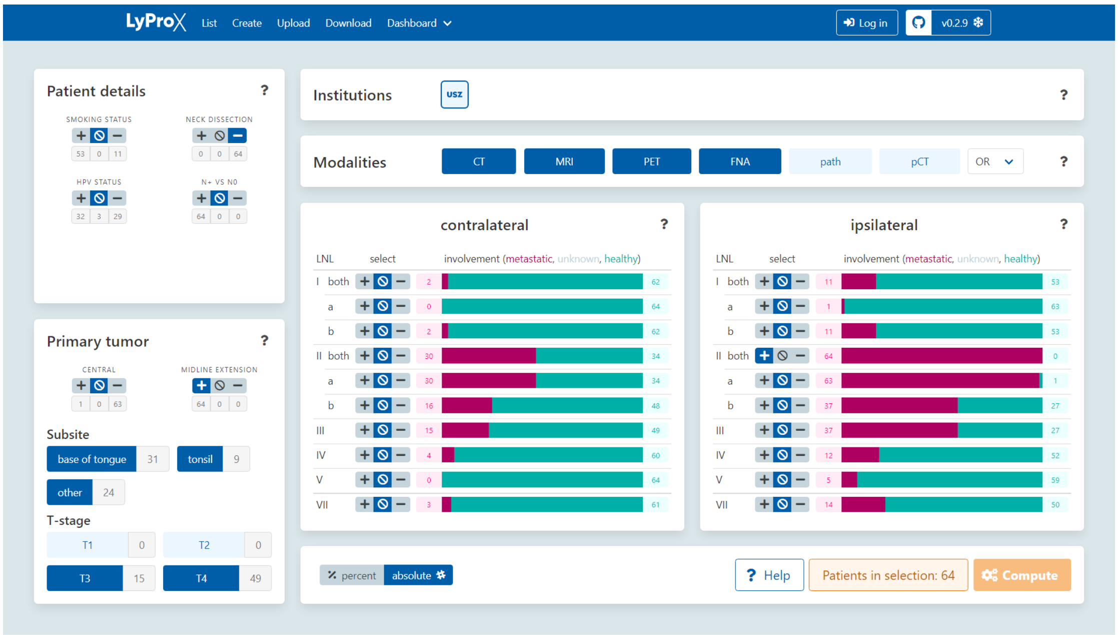

Figure 5.1: Graphical user interface to analyze lymphatic metastatic progression patterns.

Table 5.1: Prevalence of LNL involvement for the whole patient cohort (all) and stratified according to early (T1/T2) versus late (T3/T4) T-category and HPV positive (HPV +) versus HPV negative (HPV-) tumors. For each LNL, the first column indicates the number of patients showing involvement in the level, the second column the percentage of positive patients in the respective group.

		n	I		II		III		IV		V		VII	
ipsi	all	287	30	10%	232	81%	113	39%	31	11%	16	6%	30	10%
	T1/T2	150	10	7%	118	79%	48	32%	10	7%	4	3%	12	8%
	T3/T4	137	20	15%	114	83%	65	47%	21	15%	12	9%	18	13%
	HPV+	181	20	11%	155	86%	73	40%	20	11%	12	7%	18	10%
	HPV−	96	8	8%	69	72%	37	39%	11	11%	4	4%	12	13%
	HPV+, T1/T2	100	7	7%	86	86%	35	35%	8	8%	3	3%	10	10%
	HPV+, T3/T4	81	13	16%	69	85%	38	47%	12	15%	9	11%	8	10%
	HPV−, T1/T2	43	2	5%	27	63%	11	26%	2	5%	1	2%	2	5%
	HPV−, T3/T4	53	6	11%	42	79%	26	49%	9	17%	3	6%	10	19%
contra	all	287	3	1%	51	18%	21	7%	7	2%	2	1%	6	2%
	T1/T2	150	0	0%	13	9%	4	3%	2	1%	1	1%	1	1%
	T3/T4	137	3	2%	38	28%	17	12%	5	4%	1	1%	5	4%
	HPV+	181	3	2%	26	14%	13	7%	6	3%	2	1%	3	2%
	HPV−	96	0	0%	25	26%	8	8%	1	1%	0	0%	3	3%

Dependence on T-category

Table 5.1 and fig. 5.2 panels a and b compare the prevalence of LNL involvement for early (T1/T2) versus late (T3/T4) T-category-patients. Consistent with common intuition, higher involvement was observed for late T-categories. Involvement of ipsilateral level II was high also for T1/T2 & (79%) and therefore increased only moderately for T3/T4 & (83%). However, involvement of the downstream levels III, IV, V increased from 32%, 7%, 3% for early T-category patients to 47%, 15%, 9% for late T-category. On the contralateral side, involvement of levels II, III, IV, V increased from 9%, 3%, 1%, 1% for T1/T2 patients to 28%, 12%, 4%, 1% for T3/T4.

Dependence on upstream levels

Table 5.2: Simultaneous involvement in levels II, III, and IV and frequency of skip metastases, for the whole patient cohort (all) and stratified according to early (T1/T2) versus late (T3/T4) T-category. Columns 1-3 define the 8 possible combinations of involvement; subsequent columns report the number of patients with the respective combination of co-involved levels.

			ipsi						contra					
II	III	IV	all		T1/T2		T3/T4		all		T1/T2		T3/T4	
+	+	+	28	10%	8	5%	20	15%	5	2%	1	1%	4	3%
+	+	−	80	28%	37	25%	43	31%	13	5%	2	1%	11	8%
+	−	+	2	1%	2	1%	0	0%	0	0%	0	0%	0	0%
+	−	−	122	43%	71	47%	51	37%	33	11%	10	7%	23	17%
−	+	+	0	0%	0	0%	0	0%	1	0%	1	1%	0	0%
−	+	−	5	2%	3	2%	2	1%	2	1%	0	0%	2	1%
−	−	+	1	0%	0	0%	1	1%	1	0%	0	0%	1	1%
−	−	−	49	17%	29	19%	20	15%	232	81%	136	91%	96	70%
			287		150		137		287		150		137	

Table 5.2 considers the frequency of involvement in downstream levels depending on the involvement in upstream levels. On the ipsilateral side, level III harbored metastases in 47% of patients (108 out of 232) when level II was positive, but in only 9% of patients (5 out of 55) when II was negative. Analogously, level IV harbored metastases in 25% of patients (28 out of 118) when level III was positive, but in only 2% of patients (3 out of 174) when III was negative (fig. 5.3). On the contralateral side, level III harbored metastases in 35% of patients (18 out of 51) when level II was positive, but in only 1% of patients (3 out of 236) when II was negative.

Contralateral involvement

Apart from late T-category (table 5.2), extension of the primary tumor across the midsagittal plane and higher ipsilateral involvement was associated with higher contralateral involvement. table 5.3 reports the prevalence of contralateral lymph node involvement depending on three factors: T-category, midsagittal extension, and whether ipsilateral level III was involved. For all 197 patients without midline extension, contralateral involvement in levels II, III, IV, V was 10%, 3%, 2%, 1% compared to 36%, 18%, 4%, 0% with midline extension (90 patients). In addition, out of 38 patients with late T-category, midsagittal extension, and positive ipsilateral level III, 21 & (55%) showed contralateral involvement in level II and 12 & (32%) in level III. Out of 39 late T-category patients with midsagittal extension but negative ipsilateral level III, contralateral involvement was lower (24% in level II, 7% in level III). In table 5.3, we consider ipsilateral level III rather than II, because level II is involved in 81% of all patients. However, when ipsilateral level II is not involved, contralateral involvement is unlikely (1 out of 55 patients showed

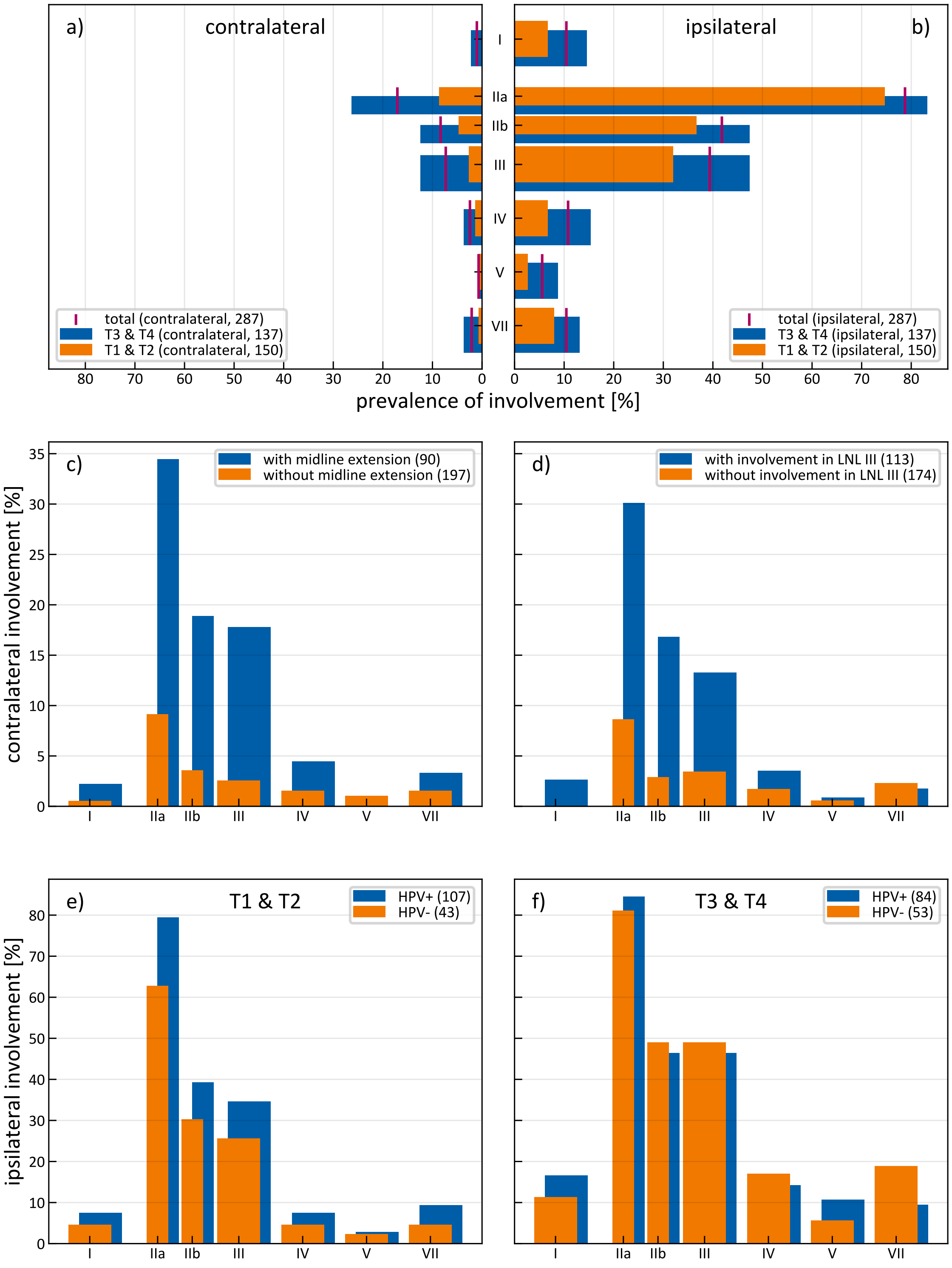

Figure 5.2: (a) Contralateral and (b) ipsilateral prevalence of LNL involvement for the whole patient cohort and stratified according to early (T1/T2) versus late (T3/T4) T-category. Contralateral LNL involvement stratified according to (c) midsagittal plan extension and (d) involvement of ipsilateral level III. Ipsilateral LNL involvement stratified according to HPV status for T1/T2 tumors (e) and T3/T4 tumors (f).

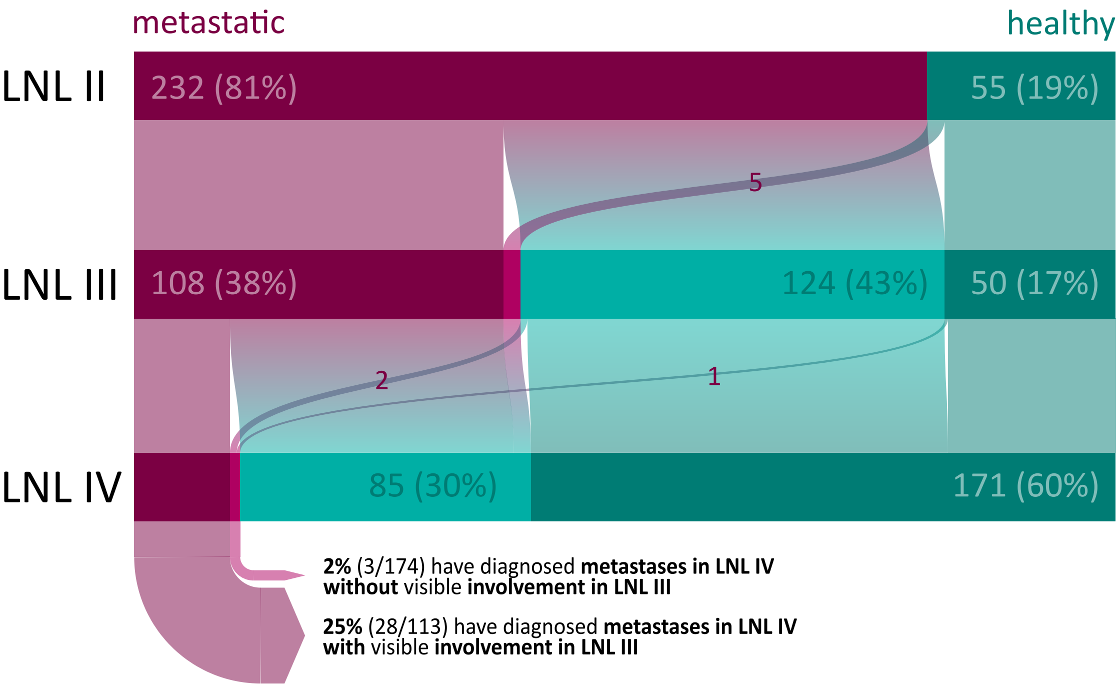

Figure 5.3: Ipsilateral involvement in levels III and IV depending on the involvement of upstream levels as flow plot.

metastases in contralateral level II). We further note that the three factors considered in table 5.3 are correlated. Out of 150 early T-category patients, 11 & (7%) showed midline extension, whereas 79 & (58%) out of 137 late T-category patients showed midline extension. As expected, contralateral involvement depended on primary tumor subsite. When considering only primary tumors strictly restricted to the tonsils (118 patients), contralateral involvement in levels II and III was 8% and 3%, respectively.

Table 5.3: Risk factors for contralateral involvement. Columns 1-3 define the 8 possible combinations of positive/negative mid-sagittal plan extension, late/early T-category, and positive/negative involvement of ipsilateral level III. Subsequent columns report the number of patients and percentages with involvement in the respective level for each combination of risk factors.

T-stage	midline	ipsi III	n	I		II		III		IV		V		VII	
early	no	–	94	0	0%	4	4%	2	2%	1	1%	0	0%	1	1%
early	no	+	45	0	0%	8	18%	1	2%	1	2%	1	2%	0	0%
early	yes	–	8	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
early	yes	+	3	0	0%	1	33%	1	33%	0	0%	0	0%	0	0%
late	no	–	31	0	0%	3	10%	1	3%	1	3%	1	3%	2	6%
late	no	+	27	1	4%	4	15%	1	4%	0	0%	0	0%	0	0%
late	yes	–	41	0	0%	10	24%	3	7%	1	2%	0	0%	1	2%
late	yes	+	38	2	5%	21	55%	12	32%	3	8%	0	0%	2	5%
			287	3		51		21		7		2		6	

Dependence on HPV status

The HPV status was positive for 181 & (63%), negative for 96 & (33%), and unknown for 10 patients & (4%). When considering LNL involvement for all T-categories and primary tumor subsites combined, the dataset provides no strong indication that LNL involvement is different for HPV positive versus HPV negative patients, neither regarding the patterns of spread to the LNLs nor in terms of prevalence of involvement. However, the data suggests that the association of higher lymph node involvement with more advanced T-category is more pronounced for HPV negative tumors than for HPV positive tumors, i.e. HPV-tumors tend to disseminate earlier to regional nodes (table 5.1). For example, for HPV positive tumors, involvement of ipsilateral level III increased from 35% for early T-category to 47% for late T-category. Instead, for HPV negative tumors, involvement was 26% versus 49%.

Involvement of levels I, V and VII

Prevalence of ipsilateral involvement was 10% (30 out of 287 patients) in level I and 6% (16 out of 287 patients) in level V. No patient had metastases in ipsilateral level I or V without involvement of ipsilateral level II. Prevalence of ipsilateral level VII involvement was 10% (30 out of 287 patients) and was more frequent for tumors of the tonsil (14%, 17 out of 118 patients) than for tumors of the base-of-tongue (6%, 5 out of 83 patients). 4 patients had metastases in ipsilateral level VII without involvement of level II. Features of more advanced disease was associated with higher involvement. For example, involvement in levels I, V, VII was 6%, 1%, 6% in early T-category patient without metastases in ipsilateral level III (102 patients) and 22%, 14%, 17% for late T-category patient with metastases in ipsilateral level III (68 patients).

5.5 Discussion

The purpose of this study was to provide detailed per-level-quantifications of cervical lymph node involvement for oropharyngeal carcinoma on a patient-individual basis, depending on T-category, involvement of other nodal levels, and various clinicopathological factors such as smoking and HPV status. To the best of our knowledge, this is the only study providing such detailed quantitative information considering multimodal diagnostic modalities, which distinguishes this study from previous publications on the overall prevalence of LNL involvement for oropharyngeal cancer. Furthermore, an elaborate user-friendly GUI is provided to visualize and explore the dataset and study the dependence of LNL involvement as a function of the above parameters. While only selected information can be presented here in the form of tables and figures, the GUI can be used to access the full information contained in the dataset and study the influence of other factors such as primary tumor subsite or smoking status.

The main limitation of this dataset is that pathological involvement for the surgically treated patients was not available because neck dissection was performed en bloc. In addition, as a single-institution dataset, the number of patients is limited. However, dataset and GUI are made publicly available. The dataset can in the future be pooled with other datasets without loss of information, and the software platform and GUI developed may serve as the basis for collecting large multi-institutional datasets.

Comparison to previous works

Prevalence of LNL involvement

Overall patterns and prevalence of LNL involvement in our study (table 5.1) is consistent with previous studies [7, 15, 20]. Our study contained a relatively low number of N0 patients (16%) compared to previous reports, which may be explained by differences in patient selection and diagnostic modalities used. Our study includes all patients treated at our institution between 2013 and 2019 irrespective of primary treatment. Hence, our patient cohort may be considered relatively unbiased. Studies reporting on surgically treated patients may introduce bias towards lower prevalence of LNL involvement when patients with advanced disease are referred to definitive chemoradiotherapy.

Dependence on upstream levels

The question on how the probability of metastases in a LNL depends on the involvement of upstream levels is poorly reported in the literature. To our knowledge, Sanguinetti et al [31] is the only publication reporting on this question for early T-category surgically treated OPSCC. For example, out of 42 patients with ipsilateral level III involved, 12 patients (29%) had also level IV involved [1], which is similar to our findings (28 out of 113 patients, 25%, for all patients combined). In agreement with previous studies, our dataset confirms that skip metastases in levels III and IV occur but are rare (table 5.2). Furthermore, we observed no case of level I or V involvement without metastases in level II, which is also confirming previous publications. Further data collection and analysis in that direction could potentially lead to treatment-de-escalation strategies by not irradiating down-stream LNLs in the absence of metastases in up-stream levels, e.g. by identifying patients in whom level IV may be excluded from the CTV-N.

Contralateral spread

A prominent example of treatment de-escalation is the sparing of the elective contralateral irradiation or the pN0, negative, neck: Chronowski et al. [8] provided data of 102 patients with tonsillar carcinoma treated with unilateral radiotherapy, of which only 2% experienced contralateral recurrence. Very similar data, with contralateral recurrence rates of only 2-3.5% were reported from the Princess Margaret Hospital [19, 27]. Moreover, similarly to the results of the large meta-analysis of Al-Mamgani [25], we could demonstrate that the probability of contralateral involvement also strongly depends on T-category and midline extension. Concerning omission of radiotherapy to the pN0 neck, a recent prospective trial, with most of the patients included suffering from oropharyngeal cancer, could demonstrate excellent tumor control rates of 97% in the unirradiated neck [9]. These results show

[1] These numbers are reconstructed from the data reported but are not directly contained in the publication.

that CTV-N reduction is possible for selected patients. However, according to current guidelines, unilateral radiotherapy is recommended in specific circumstances. In our study, incidence of contralateral involvement was 20% and the data suggests that for many patients with favorable characteristics (early T-category, no midline extension, limited ipsilateral involvement), the risk of contralateral metastases is low (**??**). If supported by further multi-institutional data, this could identify additional patients in whom the contralateral neck may be safely excluded from the CTV-N, either completely or in part. E.g., radiotherapy could be limited to level II in some patients when level II still bears a relevant risk of occult metastases but the risk in levels III and IV is minimal.

HPV-status

Consistent with the findings of Bauwens et al [2] and the general clinical observation that HPV-positive tumors seem to metastasize early despite small primaries, our data suggests that the dependence of lymph node involvement on T-category is less pronounced in HPV-positive tumors (table 5.2). Beyond that, our data does not provide evidence that lymphatic progression patterns differ between HPV-positive versus negative tumors (consistent with Bauwens et al [2]).

From macroscopic progression patterns to microscopic involvement

In this work, we consider lymphatic progression patterns assessed through imaging. Hence, the distribution of macroscopic lymph node metastases is studied. For defining the elective clinical target volume (CTV) or the extent of surgical resection, we are instead interested in the conditional probability of microscopic involvement in a LNL given that no macroscopic metastases are seen in that level. This risk depends on two aspects: (1) The sensitivity and specificity of diagnostic imaging, i.e. the probability of not diagnosing lymph node metastases that are present; and (2) The probability that tumor cells have spread to a LNL, given the observed state of tumor progression for that patient. The latter can be obtained from datasets of metastatic progression patterns in a cohort of patients as presented in this paper. Statistical methods that combine both aspects to calculate risk of microscopic involvement have been developed for ipsilateral levels I-IV [29, 24]. Future work will extend these statistical models to contralateral spread and levels V and VII, informed by the data presented here. However, this is not part of the current paper.

Summary and prospect

Detailed datasets of lymphatic progression patterns, meaning reporting the combination of simultaneously involved LNLs together with tumor characteristics on a patient-individual basis, allows for better quantification of LNL involvement. This may in turn allow for further personalization of elective CTV-N definition based on the individual patient's state of tumor progression. In this paper we publish such a dataset together with a graphical user interface to explore the dataset. The software tools are made publicly available for others to study our dataset in detail and to contribute data for building larger multi-institutional datasets. Large datasets, possibly containing thousands of patients, together with the statistical methods for analysis, may eventually inform future clinical trials and guidelines on nodal CTV definition in head & neck cancer. Potential applications are to omit irradiation of ipsilateral level IV in selected patients, or to identify additional patients in whom contralateral neck irradiation can be omitted or limited to level II.

Colophon

This thesis was made in LaTeX 2$_\varepsilon$ using my blood, sweat and tears.

List of Figures

2.1 Solutions to eq. (2.22) for the base probability rate $\tilde{b}_1$ given a $p^\star$ of 0.4 and T increasing from 1 to 15. 12

2.2 Decay of base probability rates as a function of the number of time-steps the time-prior has. Circles depict the results from learning the same dataset with different time-priors while solid lines show the analytical result starting with a $p^\star$ corresponding to the prevalence of involvement of LNL I and II respectively. 12

2.3 Prediction for the risk of involvement in LNL III, given that no other LNL was observed to be involved. Computed by training our hidden Markov model with 15 different-length uniform time-priors (transparent red) as well as the Bayesian network model (blue line). The one outlier is the HMM with a time-prior covering only one time-step. It is centered on the prevalence of involvement for LNL III, regardless of the given diagnose. 13

3.1 Predicted prevalences (shaded histograms) and posterior beta distributions of observed prevalences (solid lines) for the ipsilateral levels II (blue), III (orange) and IV (green). These prevalences have each been plotted for early T-stage patients (top row) and late T-stage (bottom row) and for the three models $\mathcal{M}_{\text{ag}}$ (left column), $\mathcal{M}_\alpha$ (middle column) and for $\mathcal{M}_{\text{full}}$ (right column). The differences between the models are negligible. 26

3.2 Predicted prevalences (shaded histograms) and posterior beta distributions of observed prevalences (solid lines) for the contralateral overall involvement (anything *not* clinically N0, on that side of the neck). Predicted and observed prevalence for early T-stage is colored blue and orange, while for late T-stage it is green and red. The prevalence for patients whose tumor does not extend over the mid-sagittal line is labelled `noext` and colored blue or green, while the same quantity for those with said extension is labelled `ext` and colored orange and red. The three models $\mathcal{M}_{\text{ag}}$, $\mathcal{M}_\alpha$ and $\mathcal{M}_{\text{full}}$ are depicted in the left, middle and right panel respectively. 27

5.1 Graphical user interface to analyze lymphatic metastatic progression patterns. 37

5.2 (a) Contralateral and (b) ipsilateral prevalence of LNL involvement for the whole patient cohort and stratified according to early (T1/T2) versus late (T3/T4) T-category. Contralateral LNL involvement stratified according to (c) midsagittal plan extension and (d) involvement of ipsilateral level III. Ipsilateral LNL involvement stratified according to HPV status for T1/T2 tumors (e) and T3/T4 tumors (f). 39
5.3 Ipsilateral involvement in levels III and IV depending on the involvement of upstream levels as flow plot. 40

List of Tables

3.1 Metrics computed via TI for the three bilateral models introduced in section 3.3: The BIC in the first row, the log-evidence $\ln Z$ with the respective standard deviation in the second and third row, as well as the maximum and mean likelihood of the sampling procedures in rows four and five respectively. 28

5.1 Prevalence of LNL involvement for the whole patient cohort (all) and stratified according to early (T1/T2) versus late (T3/T4) T-category and HPV positive (HPV +) versus HPV negative (HPV-) tumors. For each LNL, the first column indicates the number of patients showing involvement in the level, the second column the percentage of positive patients in the respective group. 37

5.2 Simultaneous involvement in levels II, III, and IV and frequency of skip metastases, for the whole patient cohort (all) and stratified according to early (T1/T2) versus late (T3/T4) T-category. Columns 1-3 define the 8 possible combinations of involvement; subsequent columns report the number of patients with the respective combination of co-involved levels. 38

5.3 Risk factors for contralateral involvement. Columns 1-3 define the 8 possible combinations of positive/negative mid-sagittal plan extension, late/early T-category, and positive/negative involvement of ipsilateral level III. Subsequent columns report the number of patients and percentages with involvement in the respective level for each combination of risk factors. 40