A Mixture of Risk Prediction Models for Tailored CTV Definition in Tumor Treatment

Advancing Personalized Cancer Therapy

by
Julian Broennimann

Supervised by
Prof. Dr. Jan Unkelbach
Dr. Roman Ludwig

March 6, 2024

A document submitted in fulfillment of the requirements for the degree of
Master of Science in Physics
at the
University of Zurich

Abstract

Head and neck squamous cell carcinomas (HNSCC) frequently spread through the lymphatic system, forming metastases in various lymph nodes. Since these metastases are often too small to be detected in clinical images, large volumes of the lymphatic system in the head and neck region are prophylactically irradiated to avoid nodal recurrences. The elective volume is defined in guidelines for definitive head and neck cancer therapy. However, these guidelines neglect correlations between affected lymph nodes and rely solely on generalized tumor location categories without considering the specific anatomical origins indicated by ICD codes. In this work, we demonstrate that each ICD code exhibits unique lymph node involvement patterns. We introduce a mixture model (MM) based on Hidden Markov Models (HMMs), which enables the prediction of lymph node level (LNL) involvement by incorporating the patient diagnosis, along with T-category and the specific ICD code of the primary tumor. The MM is trained on a multicentric dataset with exact LNL involvement information for primary tumors in the Oral Cavity, Oropharynx, Hypopharynx, and Larynx. The learned parameters of the mixture model are highly interpretable. Application of the MM provide accurate predictions of LNL involvement over the ICD codes. Thus, this advanced modeling approach can be used to refine clinical guidelines, leading to more personalized and effective treatment plans for patients with HNSCC, minimizing unnecessary radiation exposure and enhancing treatment outcomes.

Acknowledges

I would like to thank Professor Jan Unkelbach for his support and logical explanations throughout the thesis. His clear way of explaining things and systematic approach of tackling new problems is something I really appreciate and want to take with me.
I also want to thank Roman for helping me get better at coding, especially with structuring code and understanding coding conventions.
Big thanks to Yoel and Esmee for the discussions and help in explaining medical topics, and to the rest of the team for the support.

Contents

1 Introduction

This thesis discusses a probabilistic machine learning model for predicting the risks of occult metastases in lymph node areas for head and neck cancer patients, based on patient-specific diagnoses and other patient specific characteristics. The core of the model was excellently developed by Ludwig et al. [1], and the concept is explained again from scratch in this thesis. Therefore, the reader does not need any prior information about neither the risk prediction model nor the biological and medical terms used in cancer therapy.
The model was initially designed for patients with head and neck squamous-cell carcinomas (HNSCC), which is why I continually refer back to this idea in this work. However, the model can also be adapted to other regions of the human body.

1.1 Current Limitations of CTV Definition Guidelines

For HNSCC patients, the personalized definition of Clinical Target Volume (CTV) remains inadequate [2]. HNSCC is known to spread through the lymphatic network [3, 4], forming metastases in lymph nodes. The detection of these metastases depends on clinical imaging techniques such as MRI and PET/CT, which, despite their technological improvements, have finite resolution limits. Consequently, these imaging methods often fail to detect microscopic metastases. To avoid nodal recurrences, large volumes of the head and neck are prophylactically included in the CTV, which delineates the target volume irradiated during radiotherapy. This uncertainty in the included volumes in the CTV, despite ensuring effective tumor control, results in a severe impact on post-treatment quality of life, with functional disturbances in speech, swallowing, hearing, and breathing [5, 6].
Currently, the regions to include in the CTV are defined in guidelines targeted for head and neck radiation therapy [2]. However, these guidelines do not consider patient specific information and have two major issues: First, they are based only on the overall prevalence of nodal metastases and thereby neglecting the influence of

the interconnections between nodal metastases [3, 4]. Correlations of involvement in nodal areas arise from the efferent flow of lymphatics between the lymph nodes. Ludwig et al. addressed this issue by initiating the collection of a multi-institutional database with detailed information on nodal involvement [7], and derived a probabilistic model to estimate the risk of nodal involvement. Thereby, the specific patient diagnose and the size of the tumor are used a predictors [1].
A second, yet unsolved issue in the guidelines is that they only consider broader tumor locations such as the Oral Cavity, Oropharynx, Hypopharynx, and Larynx. This maintains the hypothesis that all tumors within those locations exhibit the exact same lymphatic progression. In this work, we use the multi-institutional dataset to show that this one-size-fits-all hypothesis is too simple, and we introduce a solution to incorporate the exact tumor location into the prediction of lymph node involvement. This aims to refine the guidelines towards a more personalized treatment approach, by defining the CTV based on detailed patient information, such as clinical diagnosis, tumor characteristics, and the exact tumor locations. The goal is to achieve a de-escalation of the treatment volume, and thus improve the quality of life for HNSCC patients after treatment.

2 Prior Work

This chapter summarizes previous contributions to estimate the probability of nodal involvement in patients with head and neck squamous-cell carcinomas (HNSCC). It focuses on developing a probabilistic model that predicts the likelihood of nodal involvement by incorporating detailed patient diagnoses, and the T-stage of the primary tumor. The T-stage quantifies the size and extent of the primary tumor. Developing such a probabilistic model involves several steps, which are covered in this chapter:

1. Understanding HNSCC and simplify the complex anatomical lymphatic drainage region in the head and neck.
2. Compilation of a dataset with HNSCC patients, including a detailed description of the nodal involvement pattern for each patient.
3. Derive a mathematical abstraction of the lymphatic network, using a Bayesian Network, to estimate the conditional probabilities of spreading between primary tumors and nodal areas.
4. Enhance the Bayesian Network to a Hidden Markov Model (HMM) for a more interpretative and biologically plausible framework that allows additional patient characteristics in predictions.

This section provides a review and introduction of the methodologies and computational techniques which are used throughout this thesis.

2.1 Regional Lymphatic System and Lymph Node Levels

HNSCC is the sixth most prevalent type of cancer [8]. It originates in squamous cells that line the mucosal surfaces of the head and neck, and affect regions such as the oral cavity, oropharynx, larynx, and hypopharynx [9]. One characteristic of HNSCC is its ability to spread beyond its primary tumor site. The lymphatic system, a network of vessels and nodes that plays a vital role in the body's immune response, often serves as a pathway for this spread [3]. Cancer cells can detach from the primary tumor, invade the surrounding area, and enter the lymphatic vessels. Once within the lymphatic system, these cells can travel to regional lymph nodes, where they may establish secondary tumor sites, known as metastases. Understanding the lymphatic spread of HNSCC enhances the definition of Clinical Target Volume (CTV) in radiation therapy to encompass potential sites of spread while minimizing damage to healthy tissues. Simplifying the anatomical complexity of the lymphatic network is a first step in modeling the lymphatic spread and sets a common base along the oncologist and physicians for radiotherapy of HNSCC patients.
To that end, Lengelé et al. [10] divided the lymphatic drainage network into anatomically constrained regions called lymph node levels (LNLs) and labeled them I to X (Figure 2.1). Thereby, one LNL can contain up to 40 lymph nodes [11]. The lymph node levels that are included in the CTV are usually treated as a whole.
The lymphatic drainage system and the LNLs are symmetrically arranged on both sides of the head and neck area, where, in the context of metastatic spread, one distinguishes between the ipsilateral and the contralateral side. The ipsilateral side refers to the same side as the presence of the primary tumor, and the contralateral side refers to the opposite. The spreading of the primary tumor through the lymphatic network results in metastases in the LNL. Efferent flow of lymphatics between the LNLs further results in correlations between the LNL involvements[12]. For example, the probability of involvement in level IV drastically increases if level III is involved, from 8% to 23% [1].[1]
Throughout this thesis, we will only consider the ipsilateral involvement of LNL I - IV. With the simplified lymphatic network, we can go one step further and discuss the multicentric dataset, which contains detailed information on LNL involvement per patient.

[1] For clearly lateralized oropharyngeal primary tumors in early T-stage (T1 or T2).

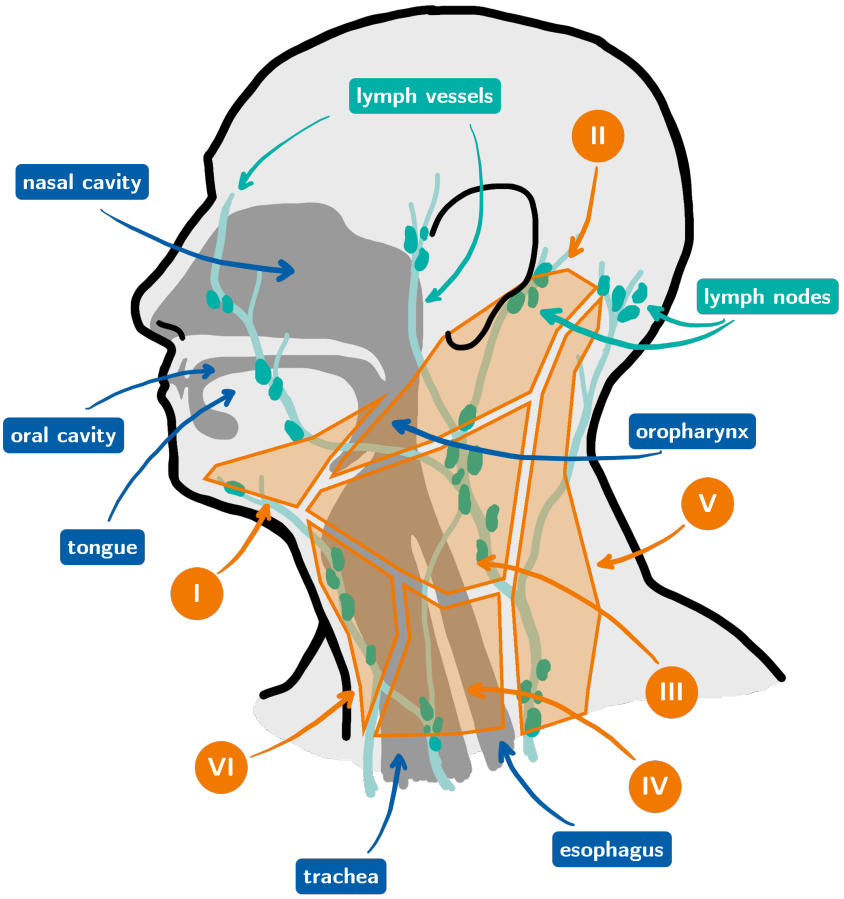


Figure 2.1: Schematic drawing of the lymph node levels and the lymphatic network. Lymphatic vessels are shown in light green, while the dark green dots represent lymph nodes. The orange shaded areas indicate the lymph node levels (LNLs), as defined by B. Lengelé et al. These broadly defined LNLs are treated in radio-oncology.

2.2 Multi-institutional Dataset and LyProX

Initiated by Ludwig et al., a multi-institutional dataset was collected, comprising 1786 HNSCC patients in whom detailed patterns of lymph node involvement have been reported [13, 14, 7]. The dataset includes patients with a primary tumor located in the oral cavity, oropharynx, hypopharynx, or larynx, and was collected from 6 institutions.

Institution	Dataset Name	Number of Patients
University Hospital Zurich	2021 USZ Oropharynx	287
Centre Léon Bérard	2021 CLB Oropharynx	263
Inselspital Bern	2023 ISB Multisite	333
Centre Léon Bérard	2023 CLB Multisite	373
University Hospital Zurich	2023 USZ Hypopharynx-Larynx	366
Vall d'Hebron Barcelona Hospital	2023 HVH Oropharynx	164

Table 2.1: Institutions and their contributions to the multicentric dataset of HNSCC patients with LNL involvement.

The data can be graphically analyzed using the previously developed web dashboard at LyProx.org. The platform allows users to analyze the data by specifying filters on various patient specifics and tumor characteristics. For example, users can filter for patients with nicotine use and a tumor in the oral cavity region, with positive involvement of LNL III. The platform then returns the number of matching patients and graphical results of the involvement of other LNLs, both in absolute numbers and percentages.

In contrast to previous datasets for HNSCC patients, where only the overall involvement of single lymph node levels was captured, this new approach, with per-patient and per-level details, allows for the training of more sophisticated statistical models that can learn the correlations between LNLs.

For some of the patients, dissected lymph node levels were pathologically analyzed. This means that experienced medical doctors surgically removed and retrospectively analyzed the tissue of the LNLs. This pathological information offers the most accurate and definitive diagnosis of whether the LNL is healthy or not. In diagnostic terms, referring to pathological information implies an assumed specificity and sensitivity of 1.

Furthermore, the dataset includes clinical involvement information based on diagnostic imaging such as CT, MRI, PET/CT, and fine needle aspiration (FNA). These methods have finite specificities and sensitivities, which are shown in Table 2.2 [15, 16]. These finite sensitivities and specificities are the reasons why we need to estimate the probability of microscopic involvement of LNL in HNSCC patients.

Sensitivity and Specificity in Diagnosis

Sensitivity: In imaging diagnostics, sensitivity reflects the ability of the imaging method to correctly identify patients with a specific disease. For instance, if an MRI scan has high sensitivity for detecting LNL involvement, it implies that the majority of patients with LNL involvement will be correctly identified by the MRI. High sensitivity ensures that cases of LNL involvement are rarely overlooked.

$$\text{sensitivity} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{2.1}$$

Specificity: Specificity evaluates the accuracy with which an imaging method can identify patients who are disease-free. An FDG-PET scan with high specificity for LNL involvement, for example, will accurately determine that most individuals without LNL involvement will have a not-involved result, thus minimizing unnecessary additional diagnostic procedures or anxiety for those individuals.

$$\text{specificity} = \frac{\text{true negative}}{\text{true negative} + \text{false positive}} \tag{2.2}$$

Modality	Specificity	Sensitivity
CT	76%	81%
FDG-PET	86%	79%
MRI	63%	81%
FNA	98%	80%
PAT	100%	100%

Table 2.2: Sensitivity and specificity for diagnostic modalities.

2.2.1 ICD-O-3 Subsites

The International Classification of Diseases for Oncology, Third Edition (ICD-O-3), categorizes tumors based on their anatomical locations (topography) and histological characteristics (morphology). For head and neck squamous cell carcinoma (HNSCC), the ICD-O-3 subsite codes enable pinpointing the exact origin of the primary tumor within the head and neck region. Our hypothesis is that each ICD-O-3 subsite is associated with a distinct metastatic pathway, resulting in different involvement patterns in the LNLs. Below is a figure outlining the anatomical location of the

ICD-O-3 codes, which are available in our dataset, together with the corresponding tumor location category.[2]

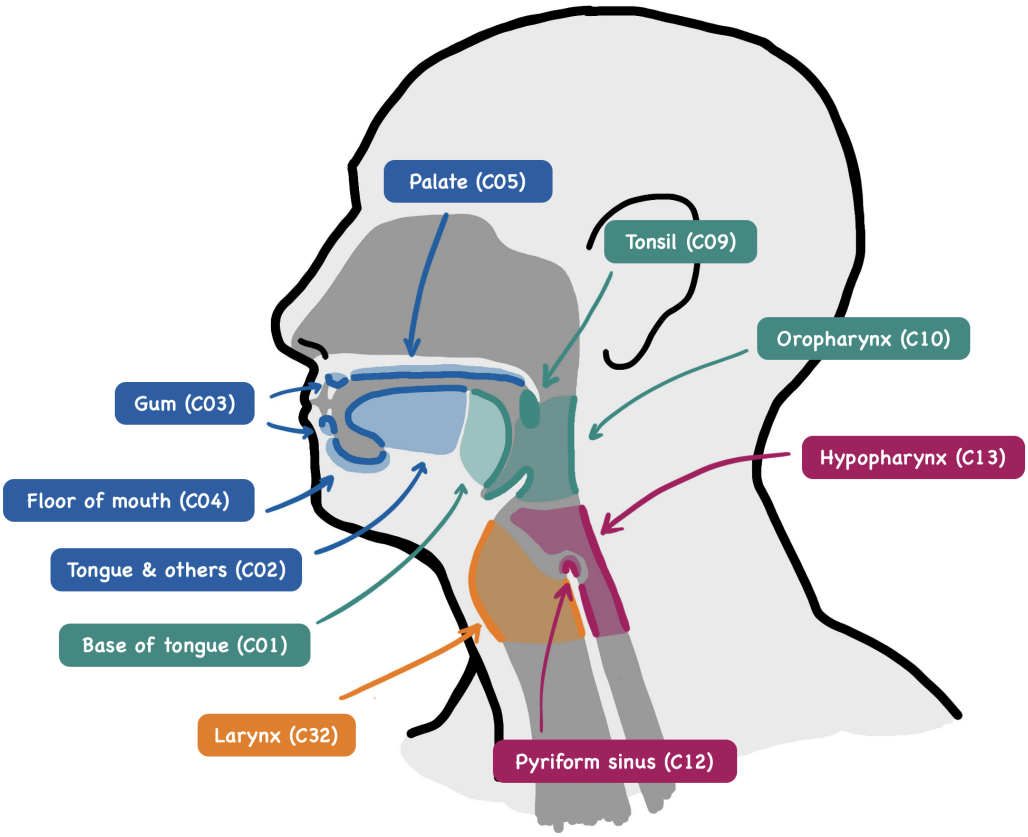


Figure 2.2: The colorized areas indicate the approximate anatomical locations of the ICD codes. The ICD codes are color-coded according to the tumor location categories, with blue representing the Oral Cavity, green representing the Oropharynx, red representing the Hypopharynx, and orange representing the Larynx. The ICD codes C00 and C08 are disregarded, due to the limited amount of data available. Further C06, other and unspecified parts of mouth, is not defined in the image.

The institutions, tumor locations, and subsites are depicted in a Sankey diagram in Figure 2.3, which also indicates the number of patients within each subsite. Note that the cumulative number of patients from subsites within a tumor location does not match up with the number of patients per tumor location due to incompleteness in the dataset.

[2]See https://training.seer.cancer.gov/head-neck/abstract-code-stage/ for more information.

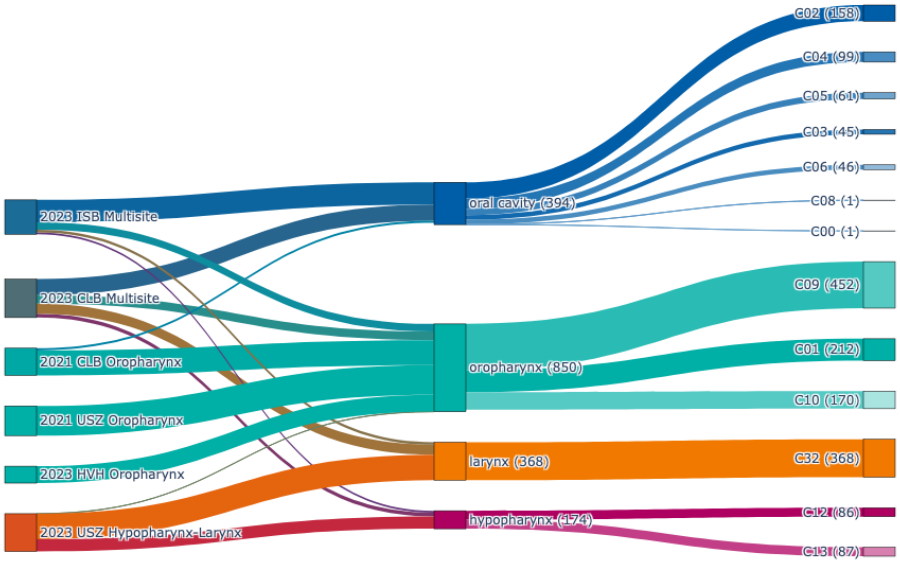


Figure 2.3: The distributions from institutions to tumor locations and from tumor locations to ICD-O-3 codes.

2.2.2 Incomplete Data and Diagnostic Consensus

The collection of lymph node level (LNL) involvement information presents its challenges. On one hand, analyzing diagnostic images to determine individual LNL involvement is time-consuming and, in some cases, impossible. This leads to incomplete datasets, where only partial LNL data may be available. On the other hand, as we accommodate LNL involvement data from multiple diagnostic modalities, a robust method to integrate these varied data sources is required.

To address these issues, we define a maximum likelihood consensus ($maxllh$) method. This method computes the consensus on LNL involvement by considering the specificity (sp) and sensitivity (sn) of various diagnostic modalities, including pathology. Consequently, this consensus is regarded as the ground truth, implying that, in diagnostic terms, both sensitivity and specificity are considered to be equal to 1.

The computation process is as follows: For each diagnostic outcome of modality $m \in \boldsymbol{M}$, if any, it updates the likelihoods of being healthy (L_{healthy}) and being involved (L_{involved}):

$$L_{\text{healthy}}* = \begin{cases} sp_m & \text{if observation is healthy} \\ 1 - sn_m & \text{if observation is involved} \end{cases}$$

$$L_{\text{involved}}* = \begin{cases} 1 - sp_m & \text{if observation is healthy} \\ sn_m & \text{if observation is involved} \end{cases}$$

Where the '$*$' sign describes the updating (multiplication) of the previous likelihoods. The subscript m in sn and sp refers to the specificity and sensitivity of this modality according to table 2.2. The final decision is made by comparing L_{healthy} and L_{involved}. The condition is considered "involved" if L_{involved} is greater than L_{healthy}, and "healthy" otherwise. This method integrates the results from multiple diagnostics into a single, more precise assessment of disease involvement.
Using this diagnostic consensus as ground truth simplifies the training and evaluation of the probabilistic model. Consequently, in this thesis, comparisons between predictions and observations are all based on this *maxllh* information.
With the dataset in place, we can leverage the data and work out a probabilistic model, which captures the correlation between LNL involvement. Therefore, the next section introduces a Bayesian Network that learns the conditional state of involvement, given some diagnosis.

2.3 Bayesian Network

Pouymayou et al. introduced a Bayesian Network (BN) to model tumor progression through the lymphatic network for HNC [12]. The main purpose of the model is to predict the risk of lymph node metastases, given the diagnosed macroscopic metastases.[3]
This section introduces the mathematical abstraction of lymphatic spread in LNLs, based on a Bayesian Network which is well-suited to model the conditional states of LNL involvement. The mathematical framework is used in more advanced models later in this work.

2.3.1 Theory

The goal is to compute the personalized risk of LNL involvement, denoted by $\mathbf{X}$, given some diagnosis $\mathbf{Z} = \mathbf{z}$. This conditional probability $P(\mathbf{X}|\mathbf{Z} = \mathbf{z})$ can be rewritten using Bayes' law as follows:

$$P(\mathbf{X}|\mathbf{Z} = \mathbf{z}) = \frac{P(\mathbf{Z} = \mathbf{z}|\mathbf{X})P(\mathbf{X})}{P(\mathbf{Z})} \tag{2.3}$$

The challenge lies in defining the terms on the right-hand side of the equation. The conditional observational probability $P(\boldsymbol{Z}|\boldsymbol{X})$ is determined by the sensitivity and specificity of the diagnostic modality, addressing the likelihood of observing $\boldsymbol{Z}$ when having the involvement $\boldsymbol{X}$. The more complex part is to ascertain the chance of involvement $P(\boldsymbol{X})$, known as the prior on $\boldsymbol{X}$. To derive this prior information of LNL involvement, the Bayesian Network proves valuable.
A graphical representation of a Bayesian Network consists of nodes and edges. In our application, each node represents either a lymph node level or the primary tumor T, and the edges represent the probability of spread from one node to another. The edges are therefore associated with the *spread probability*. These edges are directed, indicating that spreading is permitted only in one direction. Figure 2.4 presents a simple example with $V = 2$ LNLs. The arrows indicate the direction of the spread, and the arrangement illustrates the potential paths from the primary tumor to subsequent lymph nodes. This visual representation aids in understanding the probabilistic relationship between the involvement of LNLs within the lymphatic system.

[3] Macroscopic means that the metastases are visible on imaging methods.

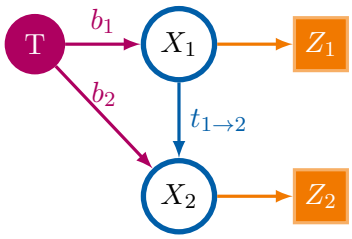


Figure 2.4: A simple example of a BN representing a primary tumor T, and $V = 2$ lymph node levels X_1 and X_2. Assuming the tumor spreads in both lymph node levels, and level I spreads to level II.

The state of a LNL is described by a random variable X, which is either healthy ($X = 0$) or involved ($X = 1$). The probability that LNL v is involved ($X_v = 1$) depends on the states of the upstream levels from which this LNL receives efferent lymphatics, $t_{pa(v)\to v}$, as indicated by the blue arcs in Figure 2.4. The subscript $pa(v)$ represents the *parental* node from which LNL v receives efferent spread. Additionally, we allow for direct tumor spread b_v from the primary tumor into LNL v, as indicated by the red arcs. This true state X is *hidden*: in reality, we do not have access to this *true* state, but only to the *observed* state, which describes the LNL as observably healthy ($Z = 0$) or observably involved ($Z = 1$). The observed states are linked to the true states by the orange arcs in the figure, representing the diagnostic imaging. We collect the states of all LNLs in a random vector of size V: $\boldsymbol{X} = \{X_1, X_2, \ldots, X_V\}$, and $\boldsymbol{Z} = \{Z_1, Z_2, \ldots, Z_V\}$ respectively. Here, V is the number of LNLs in the graph. This graphical representation of the Bayesian Network now aids in understanding the conditional state of any LNL and deriving a mathematical expression for it. For example, the state of LNL 2 (indicated by X_2 in the figure above), depends on the state of the upstream level X_1 and the tumor T. The tumor node is always in state 1 (involved).

$$\begin{aligned} P_{BN}(X_2 = 0 | X_1 = 0) &= 1 - b_2 \\ P_{BN}(X_2 = 1 | X_1 = 0) &= b_2 \\ P_{BN}(X_2 = 0 | X_1 = 1) &= (1 - b_2) * (1 - t_{12}) \\ P_{BN}(X_2 = 1 | X_1 = 1) &= 1 - (1 - b_2) * (1 - t_{12}) \end{aligned} \tag{2.4}$$

This example can be extended to a generalized formula which describes the state of LNL v, for a generic graph:

$$P_{BN}(X_v = x_v | \{X_r = x_r, t_{r\to v}\}_{r\in pa(v)}, b_v) = \\ x_v + (-1)^{x_v}(1 - b_v) \prod_{r\in pa(v)} (1 - t_{rv})^{x_r} \tag{2.5}$$

Note that we allow an LNL to have multiple parental nodes. The probability of a true state $\boldsymbol{X} = \boldsymbol{x}$ can therefore be written as the product over each LNL:

$$P_{\mathrm{BN}}(\boldsymbol{X} = \boldsymbol{x}) = \prod_{v}^{V} P_{BN}(X_v = x_v | \{X_r = x_r\}_{r\in pa(v)}) \tag{2.6}$$

As already discussed, the conditional observable probability $P(\boldsymbol{Z}|\boldsymbol{X})$ is given by the sensitivity (sn) and specificity (sp) of a diagnostic modality, and describes the orange arcs in the figure above. For a single LNL v, we can write the conditional probability as:

$$\begin{aligned} P(Z_v = 1|X_v = 1) &= sn \\ P(Z_v = 0|X_v = 1) &= 1 - sn \\ P(Z_v = 1|X_v = 0) &= 1 - sp \\ P(Z_v = 0|X_v = 0) &= sp \end{aligned} \tag{2.7}$$

For example, $P(Z_v = 1|X_v = 1) = sn$ reads as follows: Considering LNL v, given that there is a metastasis ($X_v = 1$) present, the probability of observing the metastasis ($Z_v = 1$) is given by the definition of sensitivity sn.
Also here, these four cases generalize to:

$$P(Z_v = z_v|X_v = x_v) = (z_v + (-1)^{z_v} \cdot sp)(1 - x_v) \\ + ((1 - z_v) + (-1)^{1-z_v} \cdot sn)x_v \tag{2.8}$$

The last term missing from the right-hand side of Bayes' expression in 2.3 is the denominator term $P(\boldsymbol{Z} = \boldsymbol{z}_i)$. This is commonly known as the evidence, as it is the probability of observing a diagnosis $\boldsymbol{z}_i$. This probability is given by marginalizing over all 2^V possible combinations of hidden states $\boldsymbol{X}$ and calculates the probability of observing $\boldsymbol{z}_i$ for each hidden state:

$$P(\boldsymbol{Z} = \boldsymbol{z}_i) = \sum_{j}^{2^V} P(\boldsymbol{X} = \boldsymbol{\xi}_j)\, P(\boldsymbol{Z} = \boldsymbol{z}_i | \boldsymbol{X} = \boldsymbol{\xi}_j) \tag{2.9}$$

Where we use a different notation for the true involvement: $\boldsymbol{\xi}$. This is just an ordered expression for one of the total 2^V combinations of true states. This term combines the conditional probabilities derived in 2.6 and 2.8. With that, all terms of the right-hand side of equation 2.3 are defined, which allows the prediction of true involvement $\boldsymbol{X}$, given some diagnosis $\boldsymbol{Z}$. However, before making predictions, the BN needs to learn the parameters $\theta = \{b_v, t_{v \to pa(v)}\}_{v \in V}$.

2.3.2 Model Training

To train the model, we utilize the concept of likelihood. The likelihood measures the probability of observing $\boldsymbol{z}_i$, given a set of parameters θ: $P(\boldsymbol{Z} = \boldsymbol{z}_i | \theta)$. This concept assumes that the BN is the generative model of the patient observations within the dataset $\boldsymbol{D}$. This does not imply it is the true model behind the data but assumes that if it were the true model, what would be the parameters θ that most likely create the dataset $\boldsymbol{D}$. The likelihood of observing a single observation $\boldsymbol{z}_i$ is given by 2.9. For the entire dataset $\boldsymbol{D} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$ with N patients, the likelihood becomes

$$P(\boldsymbol{D} | \theta) = \prod_{i}^{N} P(\boldsymbol{Z} = \boldsymbol{z}_i | \theta) \tag{2.10}$$

Maximizing 2.10 over the model parameters θ would then result in the set of parameters that maximize the probability of observing $\boldsymbol{D}$. In medical settings, estimating the uncertainty of predictions, such as the probability of LNL involvement, is an important feature of probabilistic models. To achieve this, we are interested in the distribution of θ. The Bayesian setting allows estimating this distribution, given the dataset $\boldsymbol{D}$. This is known as the *posterior*. Throughout this thesis, we use a sampling technique called Markov Chain Monte Carlo (MCMC) to create samples from the posterior distribution, to obtain multiple sets of θs. These θs are used for predictions of LNL involvement (called *inference*).

2.3.3 MCMC Sampling and Inference

In Bayesian inference, we can use the posterior distribution for predictions:

$$P(\boldsymbol{x}^*|\boldsymbol{z}^*, \boldsymbol{D}) = \int P(\boldsymbol{x}^*|\boldsymbol{z}^*, \theta)P(\theta|\boldsymbol{D})d\theta \tag{2.11}$$

Here, $\boldsymbol{D} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$ is the dataset, θ are the model parameters, $\boldsymbol{z}^*$ is a new diagnostic pattern, and $\boldsymbol{x}^*$ is the corresponding *true* involvement pattern of interest. The posterior distribution over the model parameters $P(\theta|\boldsymbol{D})$ can be expressed by Bayes' law:

$$P(\theta|\boldsymbol{D}) \propto P(\boldsymbol{D}|\theta)P(\theta) \tag{2.12}$$

Where $P(\theta)$ defines the prior distribution over the model parameters. Since we have no prior information on the spreading parameters, the prior is set to be uniform in $[0, 1]$, as this is the range of allowed spreading probabilities.
The Markov Chain Monte Carlo sampling technique is used to draw samples from the posterior in 2.12. Sampling is independent of any scaling, therefore we can neglect the denominator in the equation. Further, due to the uniform prior, sampling from the posterior is equivalent to sampling from the likelihood $P(\boldsymbol{D}|\theta)$ derived in 2.10, which is convenient.
The drawing of samples using MCMC results in a chain of θs, $\theta^{(1)} \ldots \theta^{(M)}$, which are used to approximate the integral in 2.11:

$$P(\boldsymbol{x} * |\boldsymbol{z}*, \boldsymbol{D}) \approx \frac{1}{M} \sum_{m=1}^{M} P(\boldsymbol{x} * |\boldsymbol{z}*, \theta^{(m)}) \tag{2.13}$$

The samples $\theta^{(m)}$ are identically and independently distributed (iid) according to $P(\theta|\boldsymbol{D})$. The consistency and sharp concentration of this estimator are guaranteed by the law of large numbers and Hoeffding's inequality, as noted by [17].
Throughout this thesis, we use the 'emcee' library to draw independent samples of θ [18]. This library is efficient and is addressed again in the application later in this thesis.
With the approximation in equation 2.13, we have derived a way to estimate the probability of LNL involvement $\boldsymbol{x}$, given some diagnosis $\boldsymbol{z}$, along with the uncertainty of the estimations. The limitation of the Bayesian Network is mainly due to the challenges of incorporating other patient features into the prediction, such that the model still remains bio plausible and interpretative. To that end, Ludwig et al.

proposed a Hidden Markov Model, which enables the estimation of LNL involvement given a diagnosis and other patient characteristics.

2.4 Hidden Markov Model

The proposed Hidden Markov Model (HMM) expands on the idea of using a Bayesian Network to model the lymphatic progression of HNC. The HMM introduces an abstract timestep $t \in \{0, 1, 2, ..., t_{\max}\}$ and describes the probability of LNL involvement at each timestep: $P_{\text{BN}}(\boldsymbol{X}) \rightarrow P_{\text{HMM}}(\boldsymbol{X}[t])$. Here, $\boldsymbol{X} = (X_1, ..., X_V)$ still represents the vector of true - but hidden - involvement of the lymph node levels, with 1 being "involved" and 0 being "healthy". We enumerate all 2^V combinations of LNL involvement by $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{2^V}$, where, for example with $V = 4$ LNLs:

$$\begin{aligned} \boldsymbol{\xi}_1 &= (0, 0, 0, 0), \\ \boldsymbol{\xi}_2 &= (0, 0, 0, 1), \\ &\vdots \\ \boldsymbol{\xi}_{16} &= (1, 1, 1, 1). \end{aligned}$$

The involvement of LNL starts at an "all healthy" starting state $\boldsymbol{X}[t = 0] = \boldsymbol{\xi}_1$. This is expressed as the *starting distribution* $\boldsymbol{\alpha}$ with

$$\boldsymbol{\alpha} = [\alpha_j] = P(\boldsymbol{X}[0] = \boldsymbol{\xi}_i) \tag{2.14}$$

which is 1 for $i = 1$ and 0 for all other states. This distribution evolves over the timesteps t according to the *transition matrix*, which is defined as:

$$\boldsymbol{A} = \Big[A_{ij}\Big] = P(\boldsymbol{X}[t+1] = \boldsymbol{\xi}_j \mid \boldsymbol{X}[t] = \boldsymbol{\xi}_i) \tag{2.15}$$

where the matrix elements A_{ij} represent the conditional probabilities of transitioning from state $\boldsymbol{X}[t] = \boldsymbol{\xi}_i$ to $\boldsymbol{X}[t+1] = \boldsymbol{\xi}_j$ in one time step. The transition probabilities are parameterized as in the BN in 2.5, where the *given* state is $\boldsymbol{X}[t]$. The size of the matrix is $2^V \times 2^V$. The probability of being in state $\boldsymbol{\xi}_i$ at timestep t is given by going through t transitions from the starting distribution. This amounts to applying the t-th power of the transition matrix $\boldsymbol{A}$ to the starting distribution $\boldsymbol{\alpha}$:

$$P(\boldsymbol{X} = \boldsymbol{\xi}_i | t) = \left[\boldsymbol{\alpha}^T \cdot (\boldsymbol{A})^t\right]_i \tag{2.16}$$

The subscript i denotes the i-th element, since we are interested in the i-th state. For further notation, we define $\mathbf{\Lambda}$ as the probability of involvement for each timestep t:

$$\mathbf{\Lambda} = \begin{bmatrix} \left[\boldsymbol{\alpha}^T \cdot (\boldsymbol{A})^0\right] \\ \left[\boldsymbol{\alpha}^T \cdot (\boldsymbol{A})^1\right] \\ \vdots \\ \left[\boldsymbol{\alpha}^T \cdot (\boldsymbol{A})^{t_{\max}}\right] \end{bmatrix} \in \mathbb{R}^{t_{\max} \times 2^V} \tag{2.17}$$

The proposal of the HMM was to include the T-category of the tumor into the model. This is where the idea of evolution over timesteps comes in handy. We do not know at which timestep the patient is diagnosed, and therefore we marginalize over all timesteps by using a different diagnose distribution (*time-priors*) for each T-category. The time-priors describe the probability of diagnosis at each timestep, implicitly assuming that patients with more advanced T-stages had - on average - a later diagnosis, thus allowing the tumor more time to develop. The time-prior differs for each T-category $\tau \in \boldsymbol{T}$ and is defined for each timestep t, parameterized by ρ_τ:

$$\boldsymbol{p}_\tau = [p_\tau(0), p_\tau(1), \ldots, p_\tau(t_{\max})] \tag{2.18}$$

The probability of *true* involvement $\boldsymbol{X} = \boldsymbol{\xi}_i$ of the HMM is the scalar product of the time prior with $\mathbf{\Lambda}$:

$$P^\tau_{\text{HMM}}(\boldsymbol{X} = \xi_i) = [\boldsymbol{p}_\tau \cdot \Lambda]_i \tag{2.19}$$

Where the subscript i indicates the i-th element of the 2^V dimensional probability vector. The superscript τ defines the time prior which is used for marginalization. To include an observation $\boldsymbol{z}_i$, we define a *diagnose vector* $\boldsymbol{d}(\boldsymbol{z}_i)$ which computes the probability of observing $\boldsymbol{z}_i$, for each hidden state $\boldsymbol{\xi}_j$, according to the conditional observation probability in 2.8:

$$\boldsymbol{d}(\boldsymbol{z}_i) = (d(\boldsymbol{z}_i)_j) = P(\boldsymbol{Z} = \boldsymbol{z}_i | \boldsymbol{X} = \xi_j) \tag{2.20}$$

The likelihood of observing a single patient with $\boldsymbol{z}_i$ and T-category τ, according to the HMM, is computed by marginalizing over all possible hidden states $\boldsymbol{\xi}$ and computing the probability of the diagnosis for this hidden state:

$$P^\tau_{\text{HMM}}(\boldsymbol{Z} = \boldsymbol{z}_i | \theta) = \prod_{j=1}^{2^V} P^\tau(\boldsymbol{Z} = \boldsymbol{z}_i | \boldsymbol{X} = \boldsymbol{\xi}_j) P^\tau(\boldsymbol{X} = \boldsymbol{\xi}_j | \tau, \theta) \tag{2.21}$$

$$= \boldsymbol{p}_\tau \cdot \Lambda^\theta \cdot \boldsymbol{d}(\boldsymbol{z}_i) \tag{2.22}$$

Where the upperscript in Λ^{θ} indicates that Λ depends on the model parameters θ. The parameters θ of the HMM are the same as in the BN, except for the parameterization of the time prior: $\theta = \{b_v, t_{pa(v)\to v}, \rho_\tau\}_{v\in V}$. Further, their interpretation differs slightly, since b_v and $t_{pa(v)\to v}$ are now spread rates, instead of spreading probabilities as in the BN, due to the incorporation of time t.
For the whole dataset, grouped into T-stages $\boldsymbol{D} = \{\boldsymbol{D}^\tau\}_{\tau\in\boldsymbol{T}}$, the likelihood is given as

$$P_{\text{HMM}}(\boldsymbol{D}|\theta) = \prod_{\tau}^{\boldsymbol{T}} \prod_{\boldsymbol{z}_i}^{\boldsymbol{D}^\tau} p_\tau \cdot \Lambda \cdot \boldsymbol{d}(\boldsymbol{z}_i) \tag{2.23}$$

Learning of the parameters θ and Bayesian inference for LNL prediction are the same as in the BN, as introduced in 2.3.3.
For the risk prediction of the HMM model, we use Bayes' theorem, as stated at the beginning of this chapter (equation 2.3):

$$P(\boldsymbol{X} = \boldsymbol{x}_j|\boldsymbol{Z} = \boldsymbol{z}_i, \tau) = \frac{P(\boldsymbol{Z} = \boldsymbol{z_i}|\boldsymbol{X} = \boldsymbol{x}_j)P^\tau(\boldsymbol{X} = \boldsymbol{x}_j)}{P(\boldsymbol{Z})} \tag{2.24}$$

Note that the subscript 'HMM' is not written in the probabilities for convenience.

2.5 Application of the HMM

This section demonstrates the application of the Hidden Markov Model (HMM) methodology to a dataset of oral cavity (OC) patients. The objective is to get familiar with the concepts and objects introduced to estimate the risk of LNL involvement.

2.5.1 Dataset and Model Configuration

The dataset consists of 394 oral cavity patients. For the ground truth data, the different modalities are collected in a *diagnostic consensus*, which combines the observations from various diagnostic modalities with a maximum likelihood approach, as outlined in 2.2.2. The HMM evolves over $t_{\text{max}} = 10$ timesteps. In this application, the T-category is neglected, grouping all patients under a single category τ_{all}, thus having the same, fixed time prior. The time prior is defined by a binomial distribution parameterized by $\rho_{\text{all}} = 0.5$.
The lymphatic network considered in this application contains 4 LNLs, I to IV, with allowed spreading of the tumor to each LNL, and efferent spread from LNL I to II, II to III, and III to IV (Figure 2.5).

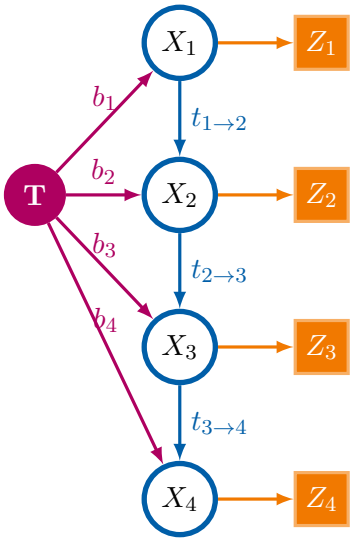


Figure 2.5: The graph structure used for the oral cavity patients

2.5.2 Parameter Learning with MCMC

The learning process for the model parameters is conducted through Markov Chain Monte Carlo (MCMC), implemented in the 'emcee' package. This package, based on the affine-invariant ensemble sampler by Goodman & Weare (2010), operates by maintaining a group of "walkers" that collectively explore the parameter space of θ. The sampling distribution is defined by the log of the likelihood function in equation 2.23.

For an intuitive understanding of the emcee package, consider this description [19]:

> "*Imagine each walker exploring a landscape where hills and valleys represent areas of high and low probability. The walkers move around, informing each other about the terrain, helping the group to understand the distribution's shape and find regions of high probability more efficiently. This collective exploration allows emcee to sample from complex likelihood distributions more effectively than traditional methods.*"

The sampler uses 200 parallel walkers, each drawing a chain of 1500 samples. The first 500 samples are discarded as burn-in, leaving 200,000 samples for the model parameters. The learned parameter distributions (posterior distributions) are visualized in a corner plot.

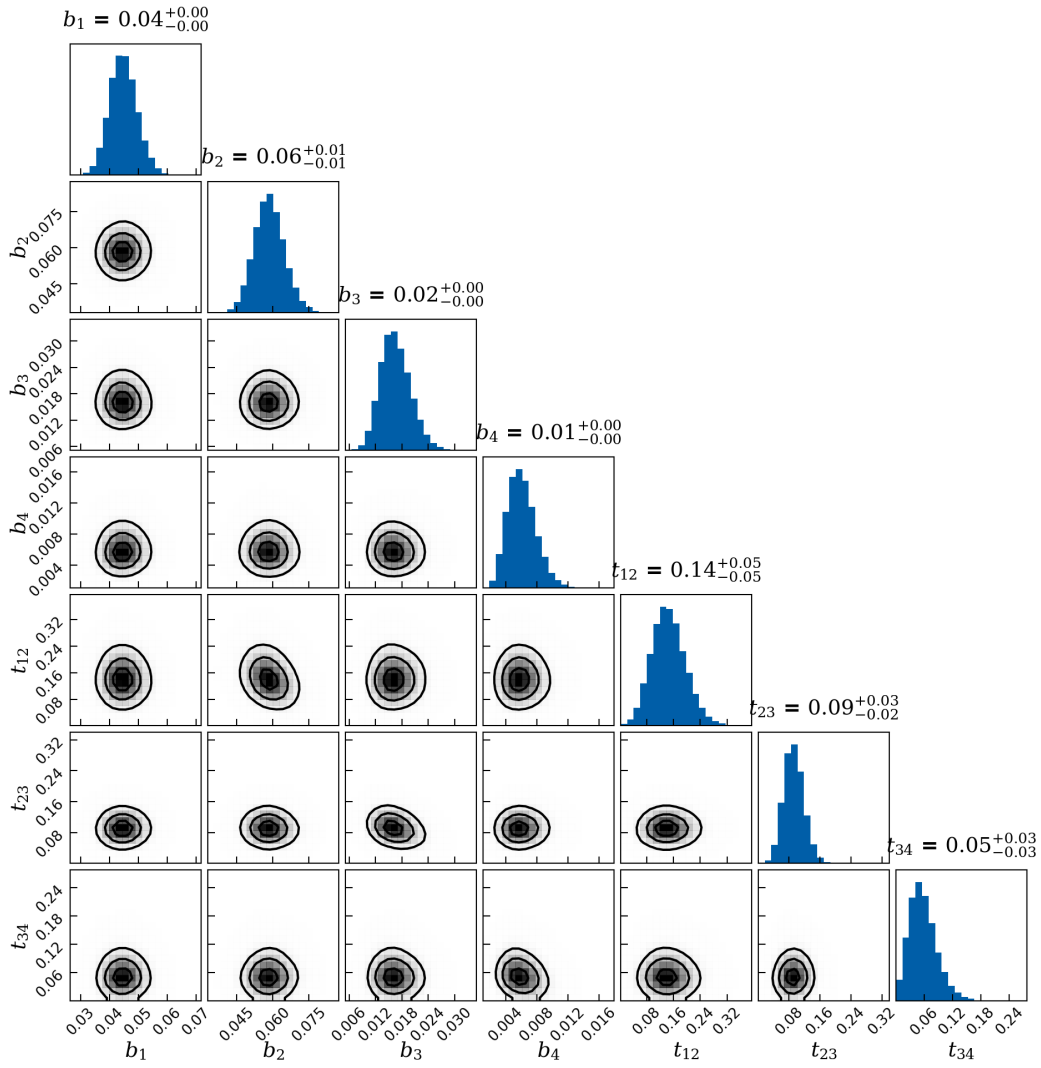


Figure 2.6: Corner plot of the sampled parameters for the HMM model, trained on oral cavity patients. The histograms on the diagonal show the 1D marginals, while the lower triangle shows all possible combinations of 2D marginals. The black lines enclose 20%, 50%, and 80% of the sampled points, respectively.

2.5.3 Transition Matrix and Evolution

With the expected values of the inferred parameters, we compute the resulting transition matrix $\boldsymbol{A}$, holding the transition probabilities from one state to any other state.

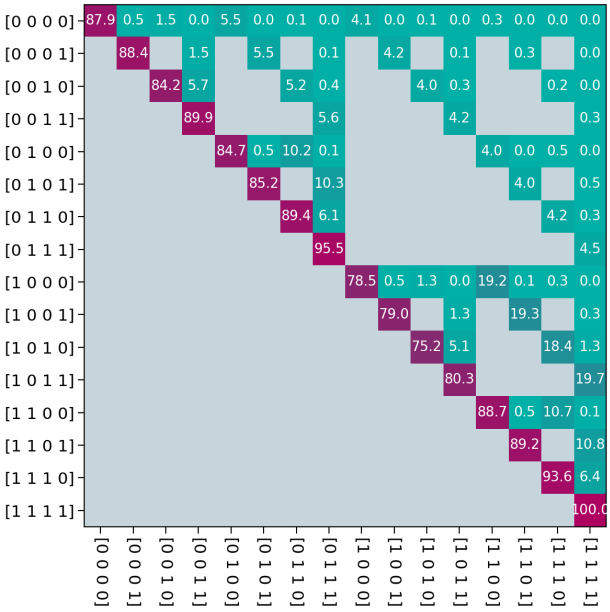


Figure 2.7: Transition matrix $\boldsymbol{A}$. The y-axis represents the state $\boldsymbol{X}[t]$, the x-axis represents $\boldsymbol{X}[t+1]$. Gray pixels indicate zero entries (impossible transitions), and colored pixels represent non-zero transition probabilities, overlaid in %.

The transition matrix defines the system's evolution over time, captured in the Evolution Matrix $\boldsymbol{\Lambda}$, by applying $\boldsymbol{A}$ t times to the starting distribution $\boldsymbol{\alpha}$.

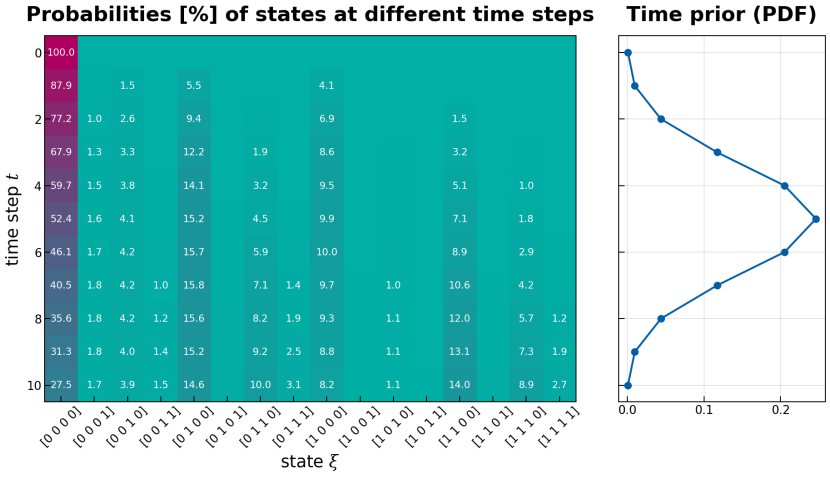


Figure 2.8: The Evolution Matrix $\boldsymbol{\Lambda}$. Probability of being in each hidden state as a function of time (left). The color indicates low (green) and high (red) probabilities, overlaid in percent if larger than 1%. On the right, the time-prior $p_{\tau_{\text{all}}}$ is plotted, which weights each column on the left. The first 'row' represents the starting distribution $\boldsymbol{\alpha}$.

Marginalized over the time prior, we obtain the prior $P(\boldsymbol{X})$ for each one of the total 2^V states.

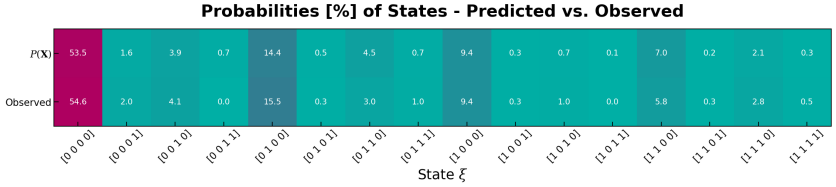


Figure 2.9: The probability of LNL involvement $P(\boldsymbol{X})$ for each $\boldsymbol{\xi}_i$, $i \in 2^V$, compared to the observed probability of this state, according to the *maxllh* diagnosis.

2.5.4 Prevalence Prediction

Prevalence describes the probability of involvement in an LNL, without any observation. Mathematically, this is calculating $P(X_v = 1)$. This is achieved by multiplying all states where LNL v is included (=1). Prevalence prediction uses Bayesian Inference, as discussed before in Section 2.3.3. To compute the prevalence in LNL v, we randomly draw 1000 samples from the posterior of θ (Figure 2.6), and compute the product of $P_{\text{HMM}}(\boldsymbol{X} = \boldsymbol{\xi}_i|\theta)$ according to 2.13 for all i where LNL v is included. The distribution in θ results in an uncertainty in the prediction. Figure 2.10 shows the prevalence prediction for LNL I to IV, and compares it to the observed prevalence from the dataset.

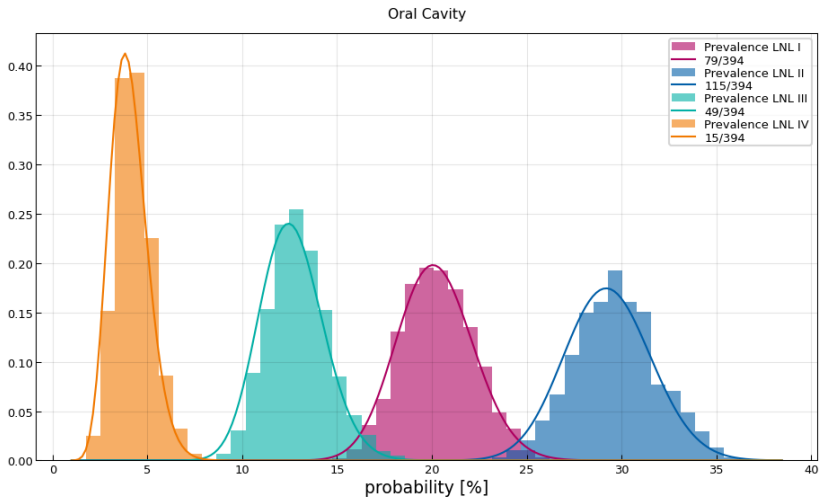


Figure 2.10: Prevalence of involvement in LNL I to IV for oral cavity patients. The shaded areas represent the predictions using 1% of the samples. The corresponding data prevalence is plotted as a Beta posterior in the same color as the prediction.

2.6 Discussion

This chapter has provided an overview of preliminary work on estimating the risk of lymph node level (LNL) involvement for patients with primary tumors in the head and neck region. The subsequent parts of the thesis will focus primarily on the Hidden Markov Model (HMM), introduced in the last section. The HMM is chosen as the foundational model due to its biologically plausible construction and its capability to accurately predict LNL involvement for a tumor location category.

3 Mixture Model Development

Abstract The Hidden Markov Model (HMM) proposed by Ludwig et al. [1] has proven successful in quantifying the risk of microscopic involvement in lymph node levels (LNLs) based on the diagnosis of macroscopic metastases and the T-category. Until now, the application of the model was limited to data from patients with tumors at a single primary location, making it applicable to only a subset of patients. However, lymphatic metastatic spread originates from all tumor locations within the head and neck region, such as the oral cavity, oropharynx, hypopharynx, and larynx. These locations are anatomically defined. However, each ICD-O-3 subsite within these tumor locations has a different spread and thus exhibits a different LNL involvement.
This chapter introduces a novel approach by incorporating the diverse and unique spread characteristics of tumor subsites into the HMM, through a mixture of HMMs (MHMM). By integrating multiple HMMs, we develop a comprehensive model that adapts its parameters to capture the distinct lymphatic spread patterns of any given tumor location. This advancement to an MHMM allows for quantifying the risk of LNL involvement for ICD subsites, alongside the discussed predictors diagnosis and T-category.

3.1 Tumor Locations and Their Unique Spread Characteristics

Each tumor location in Head and Neck Squamous Cell Carcinoma (HNSCC) presents unique characteristics in its spread to lymph node levels (LNL). The different spread can be explained by the anatomical distance between the tumor location and the LNLs.
Our dataset analysis across various tumor locations revealed that tumors in the Oral Cavity have a high involvement in LNL I, with an observed prevalence of approximately 20%, and moderate involvement in LNL II at 31% (Figure 3.1). The involvement rates decrease for LNLs III - V, with rates of 12.5%, 3.8%, and 2.8%,

respectively. This can be explained by the anatomical proximity of the oral cavity to LNL I, and the increasing lymphatic distance to III and IV. Generally, the data shows that LNL II is always disproportionate high involved.
Oropharyngeal tumors, located near LNL II, demonstrate a high prevalence of over 73% in LNL II. Hypopharyngeal tumors, anatomically adjacent to LNLs II and III, show significant involvement at these levels, with prevalences of 62% in LNL II and 48% in LNL III.
Laryngeal tumors, positioned inferiorly within the head and neck region, exhibit a lower tendency to spread into the lymphatic system. However, there is notably higher involvement in LNL V for laryngeal tumors, which is anatomically adjacent to the larynx, with almost no involvement in LNL I.
This analysis underscores that different tumor locations in HNSCC possess distinct lymphatic spread patterns, influenced by the anatomical relationships between tumor locations and LNLs.

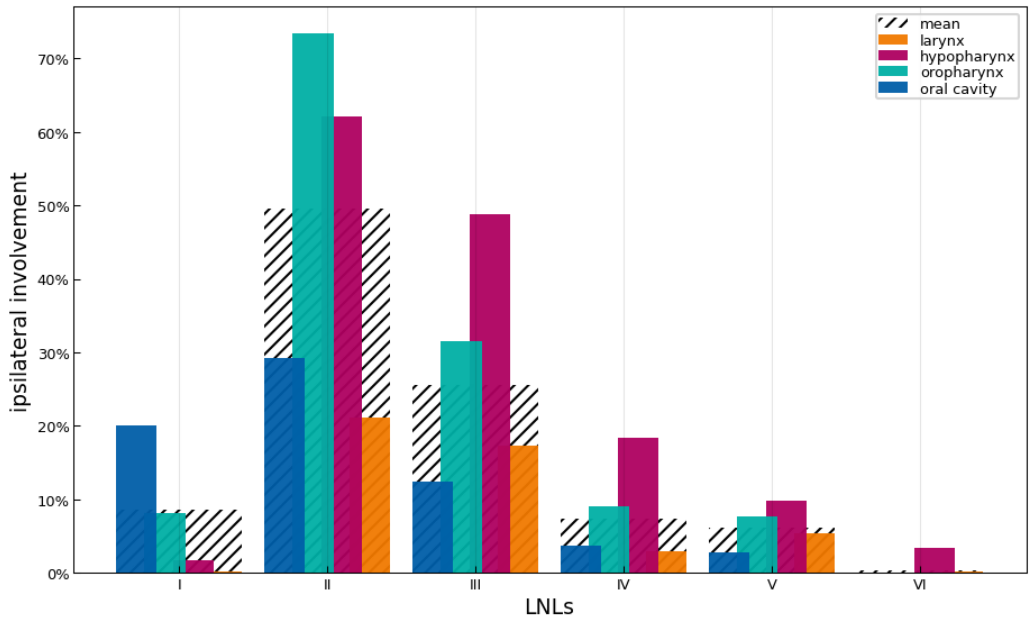


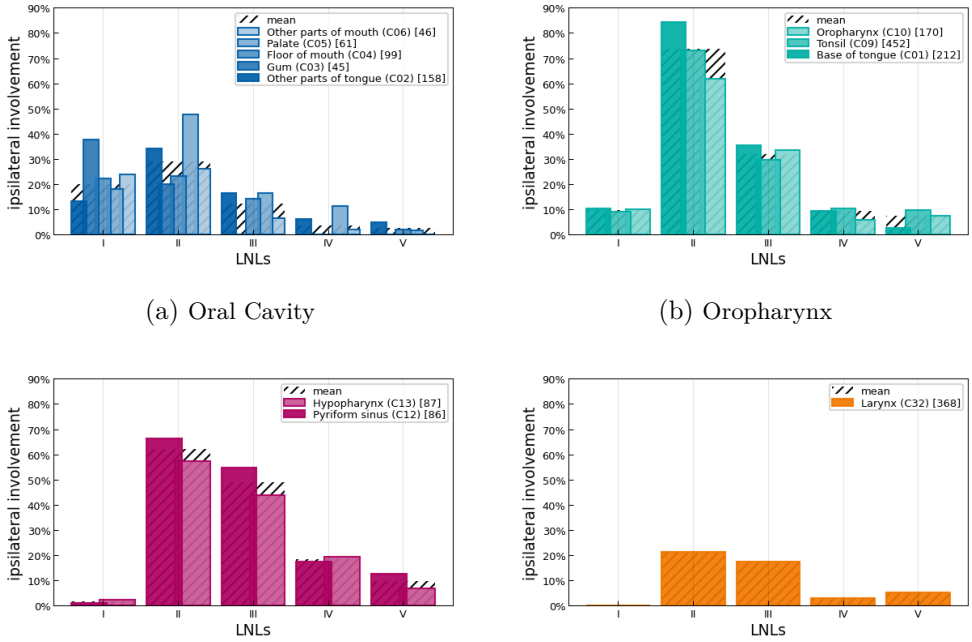
Figure 3.1: The figure shows the ipsilateral involvement of LNL I to VI for each tumor location of HNSCC. The black hashed bar in the back shows the mean over all locations.

Current guidelines for CTV-N definition rely on the presented overall prevalence rates, assuming uniform spread characteristics for all ICD-O-3 codes within a given tumor location. This simplification overlooks the differences in lymphatic spread

patterns among individual subsites, as illustrated in figure 3.2. It shows the cumulative LNL involvement for each ICD subsite, ordered and color-coded by tumor location, challenging the assumption of uniformity.

Take, for example, the ICD code C05 (Palate) displayed in the top left subplot. Although classified under Oral Cavity, its anatomical position between the Oral Cavity and Oropharynx is reflected in its LNL involvement, showing high prevalence in both LNL I, which is typical for Oral Cavity patterns, and LNL II, which is typical for Oropharyngeal patterns. Conversely, Gum (C03), situated anteriorly in the mouth, predominantly spreads to LNL I with low spread to LNL II. Among the 45 patients in this subsite available in the dataset, none exhibited involvement in LNL III to V.

This differentiation underscores the unique lymphatic spread character of each subsite within a tumor location, particularly notable within the Oral Cavity and Oropharynx categories. The following sections will discuss how these unique subsite characteristics are integrated into our probabilistic model, enabling it to predict lymph node involvement for individual patients based on the exact tumor subsite.



(a) Oral Cavity

(b) Oropharynx

(c) Hypopharynx

(d) Larynx

Figure 3.2: Observed ipsilateral prevalences in LNL I to V, for ICD subsites grouped by tumor location. The black hashed bar in the background represents the average LNL involvement for each tumor location.

3.2 Possible methods for subsite integration

Every ICD-O-3 subsite has its characteristic LNL involvement. To capture this characteristic within the HMM, we will briefly introduce some methods which are considered in the preliminary phase of the study, where only one of them is further selected for its most generic approach. One trivial method is the training of $|\boldsymbol{S}|$ independent HMMs, for each subsite $s \in \boldsymbol{S}$. This method solves the problem, however it was less suitable since it requires a large amount of data for any given ICD code. Further, this solution is not generalizeable and neglects any biological similarities between the ICD codes, even though they spread though the same lymphatic network. A second approach, which is also based on the HMM, uses the anatomical location of the subsites, and introduces a *lymphatic length* $l_{s,v}$ which measures an abstract distance between the subsite s and LNL v. The spreading rate is then given by $b_v = \frac{b_0}{l_{s,v}}$, where b_0 acts as a base transmission speed. Even though, the anatomical location of the ICD codes have a strong impact in the LNL involvement, such that an approach like this is indeed considerable, this method was excluded, due to the difficulties in determining the abstract lymphatic length.
The approach introduced in the next section is chosen due to its ability of generalization to a wide range of problems, while simultaneously minimizing the amount of parameters needed for the model. This is especially important in health care applications, where robust and interpretable models are required. Further, due to the generic approach, the framework can be used for another type of patient characteristic, which defines subcohorts in the dataset.

3.3 Mixture Model of HMMs

This section introduces the mathematical formulation of a Mixture Model (MM) that integrates Hidden Markov Models (HMMs) for predicting lymph node involvement (LNL) in HNSCC patients.

3.3.1 Model Formulation

The mixture model combines various Hidden Markov Models (HMMs) to predict lymph node involvement in head and neck cancer for a given subsite s:

$$P_{\text{MM}}(\boldsymbol{X}|s) = \sum_{k=1}^{K} \pi_{s,k} P_{\text{HMM}}(\boldsymbol{X}|\theta_k) \tag{3.1}$$

The model employs a combination (or 'mixture') of different HMMs to capture the unique patterns of tumor progression depending on the subsite. The mixture model consists of K components, where each component represents an individual HMM focusing on a distinct progression pattern of the disease. Each component has its own set of parameters (denoted as θ_k), thus predicting unique LNL involvement. Learning the model involves simultaneously determining the parameters of each component and the mixture proportions $\pi_{s,k}$ for the subsites in each component. To predict for a patient within a subsite, the model calculates the weighted sum of the component predictions according to the mixtures (equation 3.1). The model incorporates other patient-specific features such as diagnosis and T-category in the predictions from the HMM components.
In the mixture model, the component parameters have a broader interpretation compared to a single HMM. Rather than focusing solely on one tumor location, they now reflect common patterns across multiple locations. This approach enhances our understanding of common spreading patterns along the different subsites. It simplifies the challenge of integrating subsites into determining the mixture probabilities (component assignments) of the components.

3.3.2 Mathematical Foundations of the Model

The mixture model we propose aims to estimate the probability of true involvement $P_{\text{MM}}(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{F} = \boldsymbol{f}_i)$ in LNLs, where $\boldsymbol{f}_i = (\boldsymbol{z}_i, \tau, s)$ represents the patient feature vector with subsite s.
The model has K components, each a HMM defined by $P_{\text{HMM}}(\boldsymbol{X}|\boldsymbol{Z}, \tau)$ with a different set of parameters θ_k. We collect the parameter sets for each component in $\boldsymbol{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_K\}$, with each θ_k encompassing the HMM parameters $\{b_v^k, t_{r\to v}^k, \rho_\tau^k\}$, where all the elements are within [0, 1].
The mixing probabilities $\pi_{s,k}$ (component assignments) for subsite s and component k are represented in the matrix $\boldsymbol{\Pi}$ of size $\mathbb{R}^{S\times K}$:

$$\boldsymbol{\Pi} = \begin{pmatrix} \pi_{1,1} & \pi_{1,2} & \cdots & \pi_{1,K} \\ \pi_{2,1} & \pi_{2,2} & \cdots & \pi_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{S,1} & \pi_{S,2} & \cdots & \pi_{S,K} \end{pmatrix}$$

with the condition:

$$0 \leq \pi_{s,k} \leq 1, \quad \text{for all } s = 1, \ldots, S \text{ and } k = 1, \ldots, K,$$
$$\sum_{k=1}^{K} \pi_{s,k} = 1, \quad \text{for each subsite } s.$$

3.3.3 The MM Likelihood

In the mixture model, both the component parameters $\boldsymbol{\Theta}$ and the mixing parameters $\boldsymbol{\Pi}$ are initially unknown. To determine these parameters, we use the concept of likelihood, which involves finding the values of $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$ that maximize the probability of observing the dataset $\boldsymbol{D}$.
The dataset $\boldsymbol{D} = \{\boldsymbol{D}_1, ..., \boldsymbol{D}_S\}$ is grouped into S subsites, and further sub-grouped according to the T-categories within that subsite $\boldsymbol{D}_s = \{\boldsymbol{D}_{s\tau}\}_{\tau \in \boldsymbol{T}}$.
The likelihood of observing this data is expressed as:

$$P(\boldsymbol{D}|\boldsymbol{\Theta}, \boldsymbol{\Pi}) = \prod_{s} \prod_{\tau} \prod_{\boldsymbol{z}_i}^{D_{s\tau}} \left[\sum_{k=1}^{K} \pi_{s,k} P^{\tau}_{\text{HMM}}(\boldsymbol{Z} = \boldsymbol{z}_i|\theta_k) \right] \tag{3.2}$$

Where $P^{\tau}_{\text{HMM}}(\boldsymbol{Z} = \boldsymbol{z}_i|\theta_k)$ is the likelihood of observing a single patient with observation $\boldsymbol{z}_i$, with the set of parameters θ_k, according to 2.22. The superscript τ indicates that the HMM computes the likelihood only for the T-category τ of patient i.

3.3.4 Model Training and Risk Predictions

Training the MM involves identifying the model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$. In a Bayesian setting, this process includes estimating the posterior distribution $P(\boldsymbol{\Theta}, \boldsymbol{\Pi}|\boldsymbol{D})$. We employ MCMC sampling techniques for this estimation and utilize the samples for risk predictions. Assuming uniform priors for $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$, and considering no prior conditional dependency such that $P(\boldsymbol{\Theta}, \boldsymbol{\Pi}) = P(\boldsymbol{\Theta})P(\boldsymbol{\Pi})$, we can express:

$$P(\boldsymbol{\Theta}, \boldsymbol{\Pi}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\boldsymbol{\Theta}, \boldsymbol{\Pi})P(\boldsymbol{\Theta}, \boldsymbol{\Pi})}{P(\boldsymbol{D})} \overbrace{\propto}^{\substack{\boldsymbol{\Pi}, \boldsymbol{\Theta} \text{ independent} \\ P(\boldsymbol{\Pi}), P(\boldsymbol{\Theta}) \text{ uniform}}} P(\boldsymbol{D}|\boldsymbol{\Theta}, \boldsymbol{\Pi}) \tag{3.3}$$

indicating that sampling from the likelihood in 2.23 is equivalent to sampling from the posterior distribution.

The posterior $P(\boldsymbol{\Theta}, \boldsymbol{\Pi}|\boldsymbol{D})$ is referred to as the *full* posterior. Predictions based on this posterior incorporate uncertainties (distributions) in both $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$. In MM, it is common for subcohorts (in our case, subsites) to have a fixed probability of belonging to a component, without uncertainty thereof. For now, we proceed with the concept of a full posterior, derived from sampling directly from the MM likelihood, and will revisit the notion of fixed mixture probabilities later.

3.3.5 Challenges in Sampling from the Likelihood

Sampling from the likelihood in mixture models poses unique challenges. The first challenge is the interdependent nature of the mixing probabilities $\boldsymbol{\Pi}$ and component parameters $\boldsymbol{\Theta}$.[1] Essentially, assignments and component parameters mutually influence each other. Traditional MCMC techniques might suffice to navigate this complexity by efficiently exploring the entire parameter space.

A more complex issue is the permutation invariance of the likelihood concerning the mixing probabilities and component parameters, known as the *label switching problem*. For instance, consider a mixture model with $K = 2$ components applied to a dataset with subsites A and B, where subsite A shows low involvement in LNL 2 and subsite B shows high involvement. Without constraints, the mixture model does not preferentially assign any particular subsite to a specific component, as both component parameters **and** assignments are learned from scratch, making the likelihood non-uni-modal. This results in two equal maxima of the likelihood function for $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$, rendering MCMC sampling ineffective.

One possible method to address this issue is through a *Restricted Parameter Space*, which imposes constraints on $\boldsymbol{\Theta}$ for each component. A short description of the method is found in the appendix in section 8.2. While this method allows traditional MCMC sampling, its effectiveness decreases with the problem's complexity (e.g., more components, more LNLs) and depends on prior knowledge of the subsites' spreading characteristics. Due to these limitations and its lack of generalizability, this method was not selected.

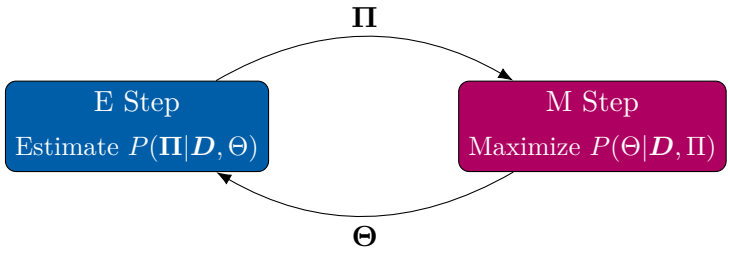
Given these limitations, we opted for an alternative approach which is commonly known as the Expectation-Maximization (EM) algorithm. This method separately solves for $\boldsymbol{\Pi}$ and $\boldsymbol{\Theta}$ in an iterative fashion, offering a general solution to the label switching problem and reducing the issues related to parameter interdependencies.

[1] This interdependence should not be confused with the assumed independent priors: $P(\boldsymbol{\Theta}, \boldsymbol{\Pi}) = P(\boldsymbol{\Theta})\, P(\boldsymbol{\Pi})$. The interdependency emerges only in our definition of the MM.

3.4 EM Algorithm

The Expectation-Maximization (EM) algorithm addresses the challenges imposed by the invariant nature of the likelihood function $P(\boldsymbol{D}|\boldsymbol{\Theta}, \boldsymbol{\Pi})$ with respect to permutations in the component assignments $\boldsymbol{\Pi}$. The EM algorithm is an iterative process that simplifies the sampling from the challenging full posterior $P(\Theta, \boldsymbol{\Pi}|\boldsymbol{D})$ by breaking it down into the conditional posterior $P(\boldsymbol{\Pi}|\boldsymbol{D}, \Theta)$ and $P(\Theta|\boldsymbol{D}, \Pi)$.
The algorithm operates in two main steps:

- **E-step (Expectation):** This step involves estimating the conditional posterior $P(\boldsymbol{\Pi}|\boldsymbol{D}, \Theta)$, for the current component parameters.
- **M-step (Maximization):** Following the E-step, the M-step updates the component parameters $\boldsymbol{\Theta}$ by maximizing the likelihood given the new assignments $\boldsymbol{\Pi}$ obtained in the E-step.



In the literature, the hidden mixing parameters $\boldsymbol{\Pi}$ are often referred to as z, while the data points are referred to as y. However, to maintain consistency in this work, we will continue using $\boldsymbol{\Pi} = (\boldsymbol{\Pi}_{s,k})$ to denote the set of mixing parameters and $\boldsymbol{D} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_S)$ to represent the set of subsites, with $\boldsymbol{D}_s = (\boldsymbol{z}_1^s, \ldots, \boldsymbol{z}_{N_s}^s)$ containing N_s patients within subsite s.

3.4.1 Mathematical Introduction to the EM Algorithm

The key function representing the iterative approach of the Expectation-Maximization (EM) is known as the Q function, which takes two sets of component parameters as inputs:

$$Q(\Theta, \Theta^{(l)}) = \int \log(P(\boldsymbol{\Theta}|\boldsymbol{\Pi}, \boldsymbol{D}))P(\boldsymbol{\Pi}|\boldsymbol{\Theta}^{(l)}, \boldsymbol{D})d\boldsymbol{\Pi}, \tag{3.4}$$

Maximizing the Q-function over $\boldsymbol{\Theta}$ results in maximizing the overall likelihood, which is shown using Jensen's inequality in the appendix 8.3. The EM-algorithm consists of iteratively maximizing 3.4.

E-step: The theoretical goal of the expectation step is to calculate the expected value of the log-likelihood function under the current estimate of parameters:

$$\mathbb{E}_{\mathbf{\Pi}\sim P(\mathbf{\Pi}|D,\Theta^{(l)})}[\log(P(\mathbf{\Theta}|\mathbf{\Pi},\boldsymbol{D}))] = Q(\Theta,\Theta^{(l)}) \tag{3.5}$$

In a Monte Carlo implementation of the EM algorithm, this involves sampling Π from $P(\mathbf{\Pi}|\mathbf{\Theta}^{(l)},\boldsymbol{D})$, where the upper-script l indicates the model parameters from the i-th iteration, and use the sample to approximate the Q function as an average:

$$Q_{l+1}(\Theta,\Theta^{(l)}) \approx \frac{1}{m}\sum_{j=1}^{m}\log(P(\Theta|\mathbf{\Pi}^{(j)},\boldsymbol{D})), \tag{3.6}$$

M-step: The maximization step maximizes 3.6 over the component parameters, to define a new proposal:

$$\mathbf{\Theta}^{(l+1)} = \arg\max_{\mathbf{\Theta}'} Q(\mathbf{\Theta},\mathbf{\Theta}^{(l)}) \tag{3.7}$$

With the new proposals, the algorithm is repeated until convergence.
Convergence in the EM algorithm is a critical aspect, usually indicated by the stabilization of parameters ($\mathbf{\Theta}$, $\mathbf{\Pi}$) and the log-likelihood function $\log(P(\boldsymbol{D}|\mathbf{\Theta},\mathbf{\Pi}))$. The algorithm is considered to have converged when changes in these parameters function become negligible between a number of successive iterations. Details on the convergence are discussed later in this thesis in Section 4.3.1.
The EM-algorithm approach depends on the conditional probabilities $P(\mathbf{\Theta}|\mathbf{\Pi},\boldsymbol{D})$ and $P(\mathbf{\Pi}|\mathbf{\Theta},\boldsymbol{D})$, and it requires sampling from both. These conditional probabilities are relatively straightforward to sample from. Using Bayes' theorem,

$$P(\mathbf{\Theta}|\mathbf{\Pi},\boldsymbol{D}) \propto P(\boldsymbol{D}|\mathbf{\Pi},\Theta)\,P(\Theta)\,P(\mathbf{\Pi}) \tag{3.8}$$

and

$$P(\mathbf{\Pi}|\Theta,\boldsymbol{D}) \propto P(\boldsymbol{D}|\mathbf{\Pi},\Theta)\,P(\Theta)\,P(\mathbf{\Pi}) \tag{3.9}$$

and due to the uniform priors, sampling from the conditional probabilities is the same as sample from the likelihood, with fixed parameters for $\mathbf{\Theta}$ or $\mathbf{\Pi}$.
We use the notation of $P^{\Pi}_{\mathrm{MM}}(\boldsymbol{D}|\mathbf{\Theta})$ for fixed mixture probabilities, and $P^{\Theta}_{\mathrm{MM}}(D|\mathbf{\Pi})$ for fixed component parameters.

4 Implementation and Validation of the Mixture Model and EM Algorithm

The preceding chapter outlines the implementation details of the MM and the MCEM algorithm. This chapter transitions from the theoretical aspects defined in the previous chapters to the practical application and validation of the MM.

4.1 From Data to Prediction with the MM

Before delving into the implementation, the workflow from having a dataset with LNL information to the risk prediction of a single LNL using an MM is outlined.

Data Preparation includes the identification of ICD subsites in the data. The raw datasets can be fetched from the publicly available LyproX database. Data preparation is done using the lyscript package [20], which includes methods such as *join* to combine multiple raw datasets from different institutions and *enhance* the dataset to infer consensus diagnosis information (*max llh*), which will be used as ground truth for model training.

Initializing the MM involves creating an instance of an MM, loading the dataset with the major subsites, and defining the number of components K the model should have. Additionally, one can determine extra configurations for the MCEM algorithm, such as *emcee* sampling parameters, an imputation function that defines the number of samples drawn in the E-step, the maximization method in the M-step, and the convergence criteria.

Once the model is initialized, model learning can begin.

Model Learning includes two steps: First, the mixture probabilities $\pi_{s,k}$ are estimated over the convergence process of the EM algorithm. After the EM has converged, the estimated mixture probabilities are fixed, and a final sampling round for the component parameters is initiated, again using MCMC sampling via *emcee*.

After model learning, the MM contains a chain of component parameters and a single set of mixture probabilities, both with the aim of maximizing the likelihood of observing the loaded dataset.

Model Prediction follows after successfully determining the component parameters and the mixture probabilities. Predictions can either be made independent of any observation, commonly referred to as *prevalence prediction*, or including a clinical observation. The latter is known as *risk predictions*.

For the predictions, we draw several component parameters from the sample chain. The mixing probabilities are fixed to a single value, and therefore, we are not using the full posterior to make predictions, but only the conditional posterior with fixed mixing probabilities.

4.2 Details of the Mixture Model Implementation

This section delves into the implementation details of the mixture model. The Hidden Markov Model (HMM) is implemented in the *lymph* package [21]. Recall that for a dataset $Z = \{\boldsymbol{z}_\tau\}_{\tau \in \boldsymbol{T}}$, where observations are grouped according to T-categories $\boldsymbol{T}$, the log-likelihood of observing this data is given by:

$$\log P(\boldsymbol{Z}|\boldsymbol{\Theta}) = \sum_{\tau}^{\boldsymbol{T}} \sum_{\boldsymbol{z}_i}^{\boldsymbol{z}_\tau} \log\Big(\boldsymbol{p}_\tau \, \boldsymbol{\Lambda}^\theta \boldsymbol{d}(\boldsymbol{z}_i)\Big)$$

Here, the elements in the formula are defined as follows:

- $\boldsymbol{p}_\tau$ is a vector of size $1 \times t_{\max}$, representing the diagnosis probability at each timestamp $t = \{1, \ldots, t_{\max}\}$, specific to the T-category τ.
- $\boldsymbol{\Lambda}^k$ is a matrix of size $t_{\max} \times 2^V$, denoting the probability of being in one of the 2^V hidden states for each timestamp. The superscript k denotes that this element is specific to this component.
- $\boldsymbol{d}(\boldsymbol{z}_i) = (d(\boldsymbol{z}_i)_j) = P(\boldsymbol{z}_i|\boldsymbol{X} = \boldsymbol{\xi}_j)$ is a vector of size 2^V, holding the probability of observing $\boldsymbol{z}_i$ for all 2^V hidden states.

The log-likelihood of the MM, with the dataset $\boldsymbol{D}$ grouped into subsites and subgrouped into T-categories within those subsites, reads:

$$\log P_{\mathrm{MM}}(\boldsymbol{D}|\boldsymbol{\Theta},\boldsymbol{\Pi}) = \sum_{s}^{S}\sum_{\tau}^{\boldsymbol{T}}\sum_{\boldsymbol{z}_i}^{\boldsymbol{D}_{s\tau}} \log\left[\sum_{k}^{K} \pi_{s,k}\, \boldsymbol{p}_\tau^k\, \boldsymbol{\Lambda}^k\, \boldsymbol{d}(\boldsymbol{z}_i)\right] \tag{4.1}$$

$$= \sum_{s}^{S}\sum_{\tau}^{\boldsymbol{T}}\sum_{\boldsymbol{z}_i}^{\boldsymbol{D}_{s\tau}} \log\left[\underbrace{\sum_{k}^{K} \pi_{s,k}\, \boldsymbol{p}_\tau^k\, \boldsymbol{\Lambda}^k}_{\mathbf{B}} \cdot \underbrace{\boldsymbol{d}(\boldsymbol{z}_i)}_{\mathbf{A}}\right] \tag{4.2}$$

The diagnosis term $\boldsymbol{d}(\boldsymbol{z}_i)$ can be extracted from the summation because it is independent of the mixture components.
Where $\mathbf{A}$ and $\mathbf{B}$ are defined as:

- **(A):** The term $\boldsymbol{d}(\boldsymbol{z}_i)$ is independent of the component and mixture parameters. We define the *diagnose matrix* for each subsite s and T-category τ to be:

$$\boldsymbol{\mathcal{D}}^{s,\tau} = (\mathcal{D}_{v,i})^{s,\tau} = [\boldsymbol{d}(\boldsymbol{z}_1), \ldots, \boldsymbol{d}(\boldsymbol{z}_{N^{s,\tau}})] \tag{4.3}$$

  The diagnose matrix $\boldsymbol{\mathcal{D}}^{s,\tau}$ has a size of $2^V \times N_{s\tau}$.

- **(B):** This term calculates the state probabilities over the mixture of components $P(\boldsymbol{X}|\boldsymbol{\Theta}, \boldsymbol{\pi}_{\boldsymbol{s},\_})$ and can be represented as a scalar product:

$$\boldsymbol{\Gamma}^{s,\tau} = (\Gamma_v)^{s,\tau} = \boldsymbol{\pi}_{\boldsymbol{s},\_} \cdot \boldsymbol{\Psi} \tag{4.4}$$

$$= \begin{bmatrix}\pi_{s,1} & \pi_{s,2} & \ldots & \pi_{s,K}\end{bmatrix} \cdot \begin{bmatrix}\boldsymbol{p}_\tau^1\, \boldsymbol{\Lambda}^1 \\ \boldsymbol{p}_\tau^2\, \boldsymbol{\Lambda}^2 \\ \vdots \\ \boldsymbol{p}_\tau^K\, \boldsymbol{\Lambda}^K\end{bmatrix} \tag{4.5}$$

  We denote $\boldsymbol{\Gamma}^{s,\tau}$ as the *State Probability*-vector of size 2^V and $\boldsymbol{\Psi} = (P(\boldsymbol{\xi})_k)_{\tau,v}$ as the *Cluster State Probability* matrix of size K, where each vector element $\boldsymbol{p}_\tau^k\, \boldsymbol{\Lambda}^k$ is of size $|\boldsymbol{T}| \times 2^V$.

With that, the log-likelihood can be rewritten as:

$$P_{\mathrm{MM}}(\boldsymbol{D}|\boldsymbol{\Theta},\boldsymbol{\Pi}) = \sum_{s}^{S}\sum_{\tau}^{\boldsymbol{T}}\sum_{i}^{N_{s\tau}} \log\Big(\boldsymbol{\pi}_{\boldsymbol{s},\_} \cdot \boldsymbol{\Psi} \cdot \boldsymbol{\mathcal{D}}^{s,\tau}_{\_,i}\Big) = \sum_{s}^{S}\sum_{\tau}^{\boldsymbol{T}}\sum_{i}^{N_{s\tau}} \log\Big(\boldsymbol{\Gamma} \cdot \boldsymbol{\mathcal{D}}^{s,\tau}_{\_,i}\Big) \tag{4.6}$$

The crucial point is to minimize the number computations during the convergence process of the EM algorithm. This emphasizes the definition of the diagnose matrix $\boldsymbol{\mathcal{D}}^{s,\tau}$. This matrix only have to computed once, when loading the dataset to the MM. Sampling for the mixture parameters requires the re-computation of only the matrix $\mathbf{\Pi}$ and $\mathbf{\Gamma}$ matrices Sampling for the component parameters requires the re-computation of only the $\mathbf{\Psi}$ and $\mathbf{\Gamma}$ matrices.

4.3 Details on the EM algorithm

The EM-algorithm sequentially finds new component parameters and mixture parameters, which maximize the log likelihood in each iteration, as discussed in section 3.4.1. Thereby the algorithms follows the following structure:

Algorithm 1 Monte Carlo Expectation-Maximization Algorithm

1: Initialize $\Theta^{(0)}$ based on the prior $P(\Theta)$.
2: **for** $i <$ max steps **do**
3: **E-step:** Draw m samples $\Pi^{(1)}, \ldots, \Pi^{(m)}$ from the likelihood $P_{\mathbf{\Theta}^{(i)}}(\boldsymbol{D}|\mathbf{\Pi})$.
4: **M-step:** Maximize 3.6 over the component parameters to propose a new $\Theta^{(i+1)}$.
5: **Check Convergence:** Exit the loop if convergence criteria are met.
6: **end for**

In the E-step, we use emcee to sample from the log-likelihood. The length of the sample chain *emcee* to approximate the posterior is configurable. Note that the number m, commonly known as the number of imputations, is the number of samples drawn from the approximated posterior. These samples are passed to the M-step to approximate the Q-function. The number of m is dynamically adjusted over the iterations:

- A single sample ($m = 1$) typically represents the mode or expected value of the posterior. In this case, the algorithm represents the standard EM algorithm. Multiple samples ($m > 1$) enhances the Monte Carlo approach and reduces variability.
- Dynamically increasing m during the convergence of the algorithm can improve performance, especially as convergence nears. This helps to reduce model variability and prevent oscillations in $\mathbf{\Theta}$ and $\mathbf{\Pi}$ while maintaining computational costs.

The maximization over the component parameters in the M-step is done via Sequential Least Squares Programming (SLSQP) optimization method from *scipy*, which returns a single set of new component parameters.

4.3.1 Convergence

Convergence in the EM algorithm is a critical aspect, usually indicated by the stabilization of parameters ($\boldsymbol{\Theta}$, $\boldsymbol{\Pi}$) and the log-likelihood function $\log(P(\boldsymbol{D}|\boldsymbol{\Theta}, \boldsymbol{\Pi}))$. The algorithm is considered to have converged when changes in these convergence features become negligible between a number of successive iterations. We define a convergence checker which takes as input the feature vector $\boldsymbol{f} = \{\boldsymbol{\Theta}, \boldsymbol{\Pi}, L\}$ where each element contains the parameters over the iterations, a predefined threshold ϵ, and a look-back period n_{iter}. The function computes the variance of each element in $\boldsymbol{f}$ over the look-back period, and returns *True* if the variance of each element is smaller than ϵ^2. This simple approach ensures that the parameters are stable within a certain range. The algorithm is built in a way to simply define other convergence measures, as for example the Kullback-Leibler Divergence from the last two posteriors in the E-step.

4.3.2 Reversed Method of the EM Algorithm

The standard procedure of the EM-algorithm in 1, as commonly found in the literature, involves sampling new mixing probabilities in the E-step and then maximizing the expected sum over the component parameters in the M-step. However, drawing on our experience with sampling component parameters from the likelihood function in the context of HMM, we propose an adjustment to this procedure. This new approach is termed the 'reversed MCEM' (rMCEM) method.
In the rMCEM method, the traditional roles of the E-step and M-step are somewhat altered. Rather than sampling the mixing probabilities $\boldsymbol{\Pi}$ from $p_{\Theta}(\boldsymbol{D}|\boldsymbol{\Pi})$ to estimate the posterior distribution $p(\boldsymbol{\Pi}|\boldsymbol{\Theta}, \boldsymbol{D})$ during the E-step, we instead sample for Θs from the likelihood $p_{\Pi}(\boldsymbol{D}|\boldsymbol{\Theta})$. The M-step then focuses on maximizing the approximation in Equation 3.6 over $\boldsymbol{\Pi}$ instead of over $\boldsymbol{\Theta}$, in contrast to how it is typically done. This reversed approach offers a novel perspective in the implementation of the MCEM algorithm. The algorithm is configurable to use either the traditional, or reversed method. Potential (dis-)advantages of either method are discussed later in this thesis.
In the next section, the implementations are validated.

4.4 Validation

The MM is expected to detect and group subsites that exhibit similar lymphatic spread patterns. Essentially, the MM uncovers common spreading characteristics across various subsites and constructs the components accordingly. The mixture probabilities of each subsite reflect the extent to which the subsite belongs to a particular component. To validate this hypothesis, consider the following toy example.

4.4.1 Toy Example: Restoring Mixing Probability

Imagine a graph structure with a primary tumor and two LNLs, I and II, where we assume no transmission occurs between the LNLs:

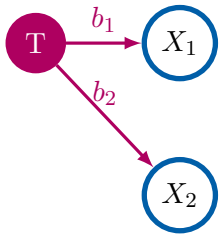


Figure 4.1: Graph structure with a primary tumor and two LNLs. The LNLs are not connected and therefore conditionally independent.

We consider 3 subsites and label them $S1$, $S2$ and $S12$. The data in subsite $S1$ and $S2$ are generated synthetically where each patient in $S1$ was draw with a 90% chance of involvement in LNL I, and 0% in LNL II, and $S2$ with 0% involvement of LNL 1 and 90% involvement of LNL 2. The subsite $S12$ samples patients randomly from $S1$ and $S2$, with 35% patients from $S1$ and 65% from $S2$. Each subsite counts 200 patients.

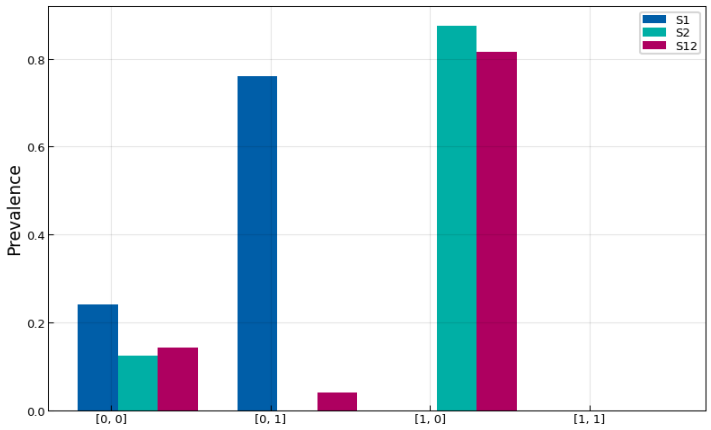
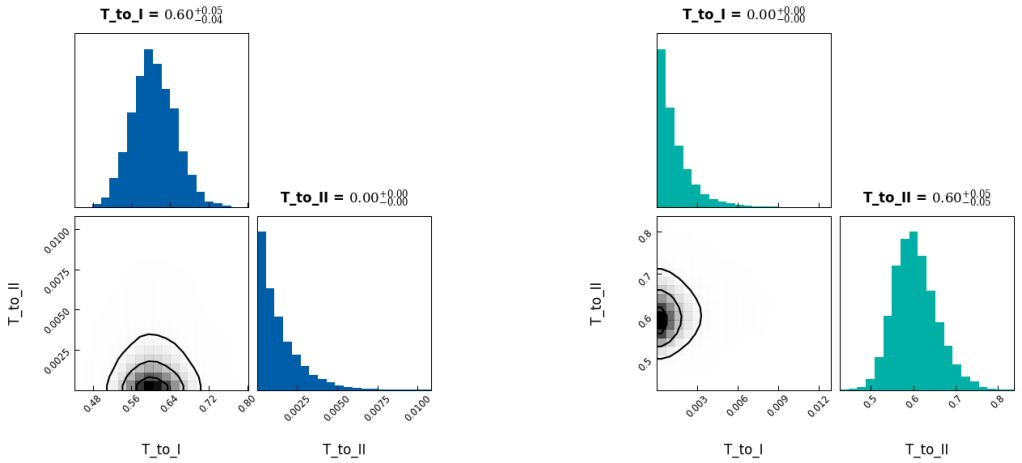


Figure 4.2: Prevalence of each of the 4 states for the synthetical subsites S1, S2 and S12. The x axis shows the 4 states, where for example [0,0] represents no involvement in LNL 1 and LNL 2 and [0,1] represents involvement of LNL 1 and no involvement of LNL2. S12 represents a mixture of S1 and S2.

When training two *independent* HMMs for $S1$ and $S2$, the base-rates are sampled to around 0.6 for the LNL exhibiting high involvement, and 0 for the other.



(a) Independent samples for $S1$.

(b) Independent samples for $S2$.

Figure 4.3: Distribution over the model parameters in a corner plot.

The mixture model is designed with $K = 2$ components. Since we know the data is generated synthetically and the underlying source is clearly two distinct subsites S1 and S2, we expect the model to rebuild those two subsites in the components. For

the subsite $S12$, we expect to get a mixture of 0.35 to the component reflecting $S1$, and 0.65 to the other.

We use the *rMCEM* method, where the component parameters are sampled during the E-step and the mixture parameters are defined during the M-step. The model is converged if the log likelihood, the component parameters and the mixture parameters change not more than $\epsilon = 0.015$ over the last 4 timesteps. Initially, the mixture weights are all set to 0.5.

The EM-algorithm runs for 6 steps until convergence, which low number can be explained by the simplicity of the problem. During convergence process, we can check how the likelihood changes over the iterations, along with the evolution of the assignment probabilities:

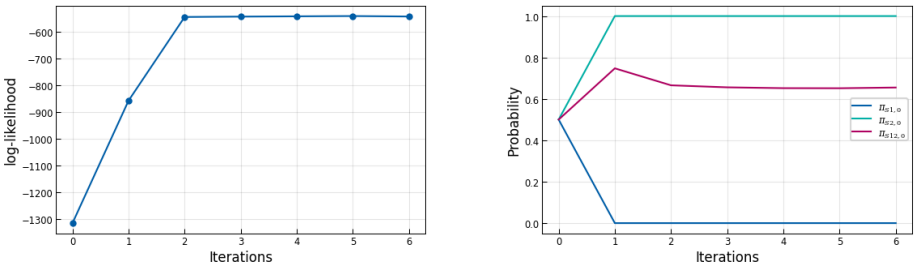


Figure 4.4: The log-likelihood (**left**) and the mixture parameters (**right**) for component 1 over the convergence. After 6 iterations, the algorithm converges and returns the found mixture parameters.

The EM-algorithm increases the log-likelihood over the iterations until convergence. The mixture model assigns subsite S1 to the component 1, and S2 to component 2, both with a probability of 100%. As expected, subsite S12 is assigned to component 1 and component 2, where the probability for component 1 is 35% and the probability for component 2 is 65%.

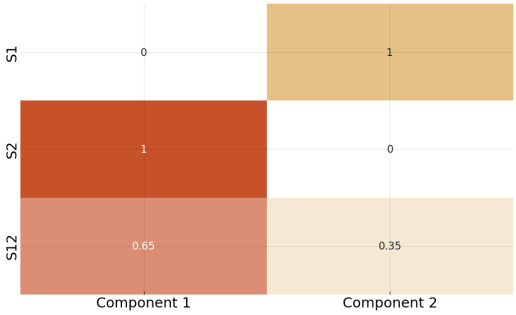
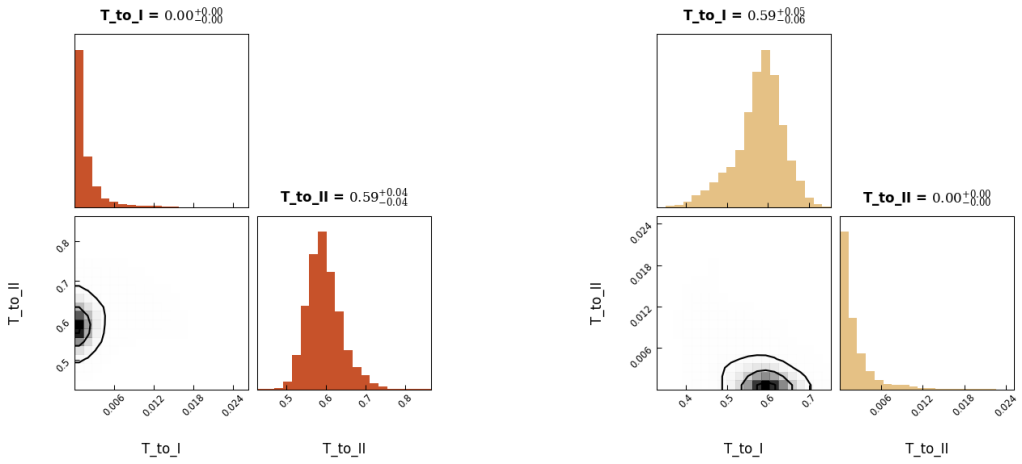


Figure 4.5: Visualization of the mixture parameter matrix ($\mathbf{\Pi}$), with annotations indicating the mixture parameters.

After estimating the mixture probabilities, the final sampling round is started which samples from the likelihood $P^{\Pi}_{\mathrm{MM}}(D|\mathbf{\Theta})$. Comparing the sampled parameters of the components with the sampled parameters from independent HMM for S1 and S2, we see that the values are the same.



(a) Component 1

(b) Component 2

Figure 4.6: Corner plots for the sampled parameters of components 1 and 2.

The mixture model was able to successfully restore the mixing parameter for the subsite S12. Further, the mixture model identifies the correct underlying generator for the different subsites. With that we can conclude, that the implementations of the mixture model and the EM-algorithm are correct.

4.4.2 Validating the MCEM method

There are 2 different methods to run the EM algorithm, the MCEM and the rMECM method. The example before demonstrates that the EM algorithm runs correctly with rMCEM. In MCEM, the algorithm samples from $P^{\Theta}_{\mathrm{MM}}(D|\mathbf{\Pi})$ during the E-step, and maximizes over $P^{\Pi}_{\mathrm{MM}}(D|\mathbf{\Theta})$ during the M-step. The final mixture parameters are then the mean value of the sample chain which results from the E-step. Using the same setup as above, with a graph as in figure 4.1, the subsites S1, S2 and S12, and $K = 2$ components, we expect similar results up on a possible switch in the mixture parameters due to their invariance in the likelihood.

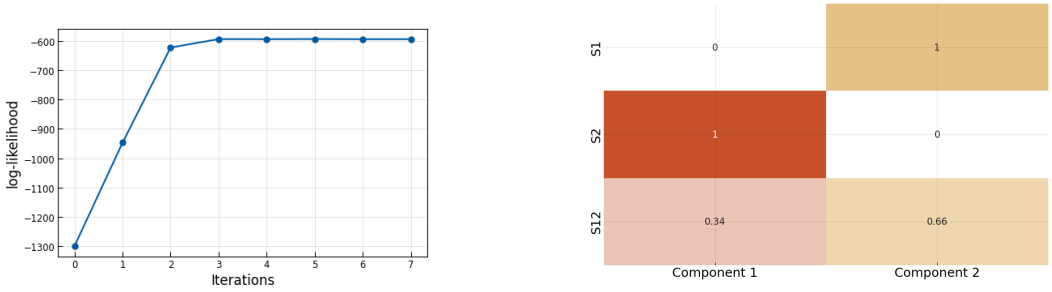


Figure 4.7: **Left:** Convergence of the log likelihood over the iterations using the MCEM method. Convergence was after 6 iterations, with a convergence threshold of 1.5%. **Right:** Final mixing parameters after 6 iterations, shown as a representation of the $\mathbf{\Pi}$ matrix.

For the mixture parameter of S12, the MCEM method results 0.34 and 0.66, which proofs that this method works too. However, one slight difference is notable, which is worth mentioning here and which is better expressed in more complex problems. The issue arises, since MCEM samples the mixture parameters during the E-step. This results in a approximation of the posterior distribution for Π. The values for the elements in Π can not exceed the range $[0, 1]$, such that, if we take the average of this distribution for the final values, the mixture parameter can not be exactly 0, or 1. This becomes clear when checking the posterior distribution after the last iteration in the figure below.

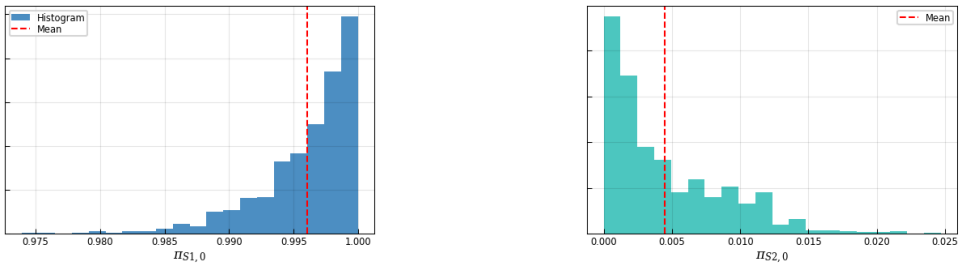


Figure 4.8: Distribution of $\pi_{S1,0}$ (left) and $\pi_{S2,0}$ (right) after the last iteration of MCEM. The red line indicates the mean value over all samples, which is the final mixing parameter.

The resulting component parameters are also affected by this. Since S1 is not fully assigned to component, the base rate to LNL 1 needs to be slightly higher (0.62 instead of 0.6), to capture the same overall involvement. And the same is true for S2, as seen in the figure below.

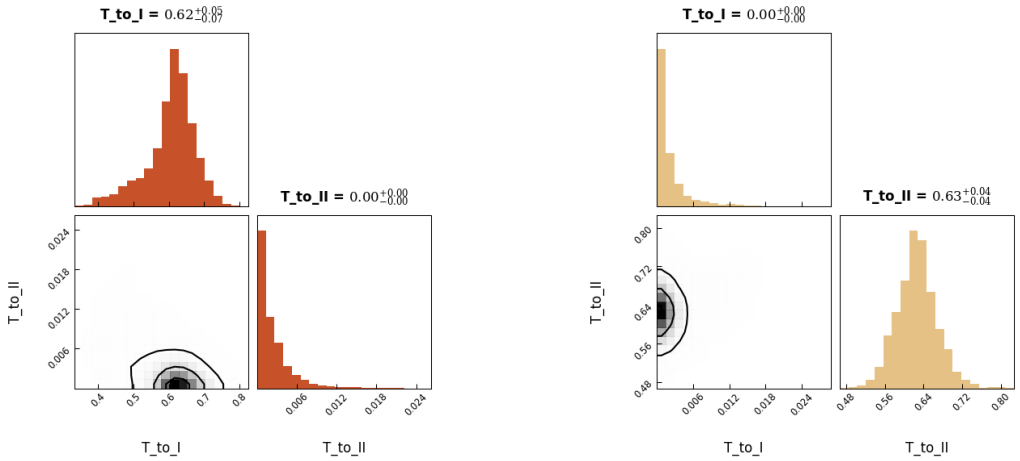


Figure 4.9: Corner plots for the sampled parameters of components 0 (Left) and 1 (Right).

This issue could be overcome by not taking the mean value from the posterior, but defining some method to find the most frequent value (as example the most frequent *bin* in the histogram and take the average value of this *bin*).
At this point, we can not determine which method works better, since the prediction results in both cases would be the same. For the application of identifying subcohorts, the rMCEM method fits our application better. With that we can conclude, that both methods work fine, however one should be aware of the slight differences in the methods where rMCEM detects underlying subcohorts with higher accuracy.

4.4.3 Invariance on Imbalance Dataset

There is a potential issue in having one subsite with fewer patients. The MM might overlook this smaller subsite and create multiple components that represent the larger subsite. This could lead to the model's inability to accurately identify the subcohorts among the patients.

To test this, we create synthetic data according to S1, S2, and S12, as described previously. S1 consists of only 50 patients, while S2 includes 1000 patients. The generators for S1 and S2 are the same as before, with S1 having a high involvement of LNL1 and S2 having a high involvement of LNL2. S12 is a 0.35/0.65 mixture of S1 and S2, also containing 50 patients.

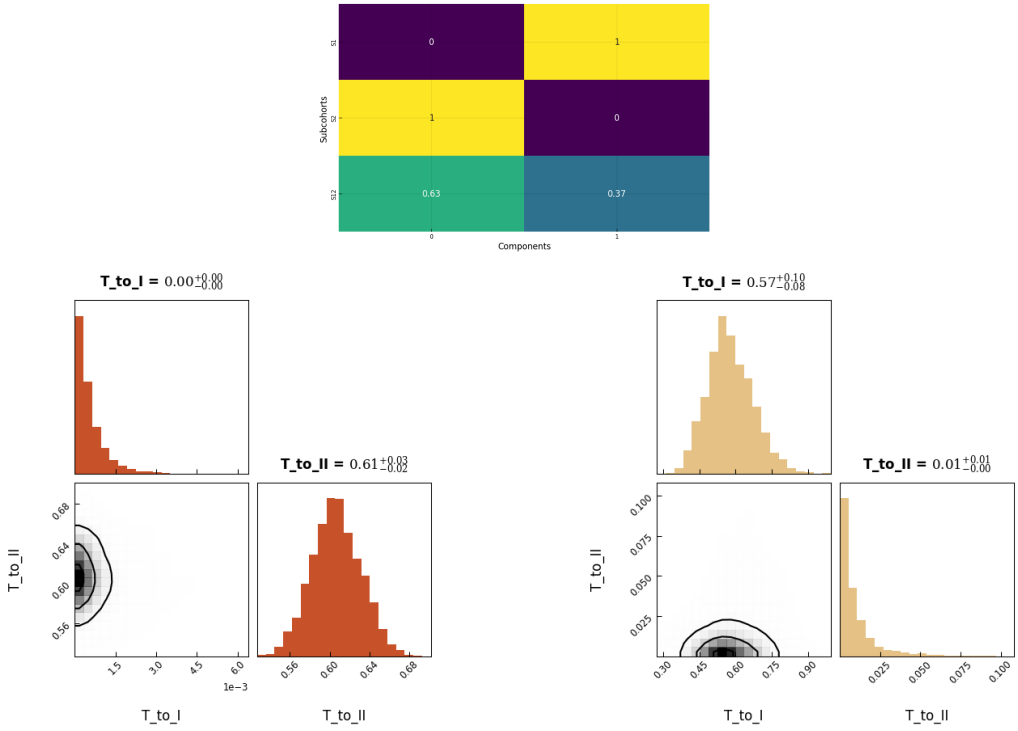


Figure 4.10: Top: Mixture probabilities of S1, S2, and S12 in matrix representation. Bottom left: Component parameters of Component 0. Bottom right: Component parameters of Component 1.

After successful training of the mixture model (MM) using the rMCEM method, we observe that (1) the model identifies two components, one representing S1 and the other representing S2, and (2) the model accurately restores the mixture probability of S12.

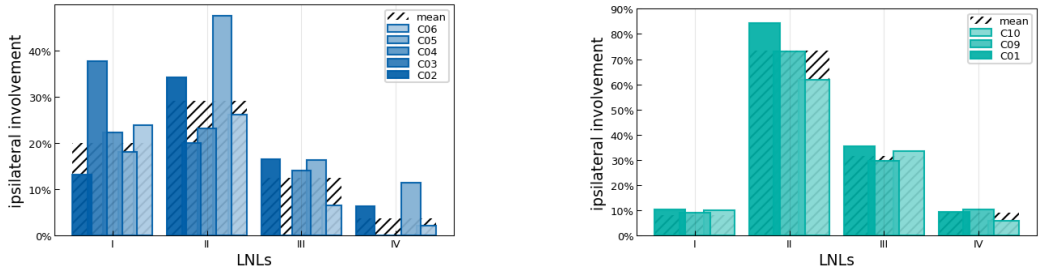
The MM demonstrates the capability to identify subsites regardless of the number of patients in each. In the final component parameters, an increase in uncertainty is observed for the component reflecting the smaller cohort (Bottom right figure in 4.10). This will be reflected as uncertainties in LNL involvement predictions. The results suggest that the model is robust regarding imbalances in the dataset.

5 Results

Finally, in this chapter, we use the MM to incorporate the primary tumor locations (according to the ICD-0-3 code) in the predictions. This results section is divided into two parts: first, only the primary tumor subsites of Oral Cavity and Oropharynx are considered. After, the entire dataset is used, including primary tumor subsites of the Oral Cavity, Oropharynx, Hypopharynx, and Larynx.
In both parts, the base HMM used for the mixture model components is the same: The base graph considers ipsilateral involvement of LNLs I, II, III, and IV. The graph structure is identical to that in figure 2.5, allowing transitions from LNL I to II, II to III, and III to IV. The diagnosis time prior is chosen to be independent of the patients' t-stage, such that all patients have the same diagnose probability at the time-steps. The time prior follows a binomial distribution parameterized by $p = 0.3$, and the model evolves over $t_{\max} = 10$ time-steps. Therefore, the model setup is equivalent to the one in the application of a HMM in section 2.5.
During the result section, the MM is compared to *independent* models. Independent models are HMM which are *independently* trained on a single tumor location, e.g. pooled patients with primary tumor in Oral Cavity. This example would be the independent Oral Cavity model.

5.1 Application of a MM to Oral Cavity and Oropharyngeal Primary Tumors

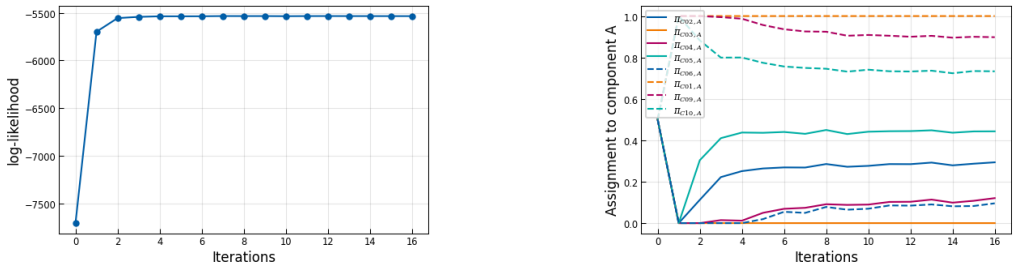
For patients with Oral Cavity (OC) and Oropharynx (OP) primary tumors, our dataset contains a total of 9 distinct ICD-0-3 subsites 2.3. For model training, the subsites C00 and C08 are excluded since they contain only 8 patients. This leaves us with a total of 8 ICD subsites. Since we are going to refer back to the observed prevalences of those subsite, the prevalence plots shown in section 3.1 are repeated:



(a) Oral Cavity (b) Oropharynx

Figure 5.1: (Repeated) Prevalence plots for subsites in Oral Cavity (left) and Oropharynx (right) for LNLs I to IV, based on the *maxllh* diagnostic.

For this application, we define a mixture model with $K = 2$ components. The MM now has for a total of 22 parameters, with 8 being the mixture parameters and $2 \cdot 7 = 14$ being the component parameters for the 2 components. Model training consists of first estimating the mixture probabilities using the rMCEM method with a convergence threshold of $\epsilon = 0.015$ and a lookback period of 4. After that, the final component parameters are sampled using the resulting mixture probabilities from step one.



(a) Log-likelihood (b) Assignment probability to component A

Figure 5.2: Results of the EM algorithm. The plots show the convergence process over the iterations, where the left is the log-likelihood and the right is the assignment probability to component A. The algorithm converges after 16 timesteps.

The rMCEM algorithm converges after 16 iterations, with the likelihood appearing constant after just 5 timesteps.

The estimated mixture probabilities π_{sk} are presented in figure 5.3. The interpretation of these results is as follows: tumors at the base of the tongue (C01) are entirely characterized by component A, and tumors of the gum (C03) by component B. These two subsites are the most distinct regarding the involvement of LNLs I

and II, rendering the results intuitive. Component A is interpreted as a model for oropharynx-like tumor spread, and component B for oral cavity-like tumor spread. All other subsites are described as mixtures. Tumors in the tonsil (C09) display LNL involvement similar to base of tongue tumors and are mostly assigned to component A. Conversely, tumors of the palate (C05) are similarly assigned to components A and B, consistent with their anatomical location and the observation that LNL involvement lies between oropharynx and oral cavity-type patterns.

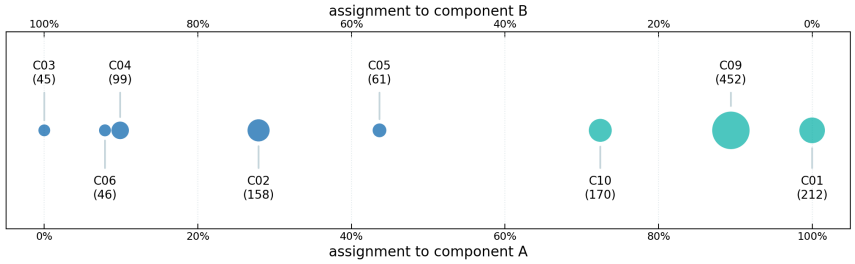


Figure 5.3: Representation of the mixture parameter matrix ($\mathbf{\Pi}$). The figure illustrates the assignment of each subsite to the two components, A and B. Subsites further to the left are more assigned to component A, and those further to the right to component B. The size of the marker (area) corresponds to the number of patients in the subsite.

With the mixture probabilities now set, we initiate the sampling round for the component parameters, choosing 20 parallel walkers and 2000 sampling steps, where the first 1000 are discarded as burn-in.

First, we plot the components of component B, which may represent more oral cavity tumor types. As a reference, the distributions are overlaid with a line representing the mean values from the HMM trained only on oral cavity patients, the independent Oral Cavity model. The interpretation is that the MM creates a component focusing on involvement of LNL I, anatomically close to the oral cavity, and decreases the involvement of LNL II. This can be seen in the base spread to LNL II, represented by b_2. In component B, b_2^B is only 0.04, in contrast to $b_2^{OC} = 0.1$ from the independent model. Additionally, the base spread to LNL III and IV is slightly lower, suggesting that typical oral cavity-type tumors spread less to these levels.

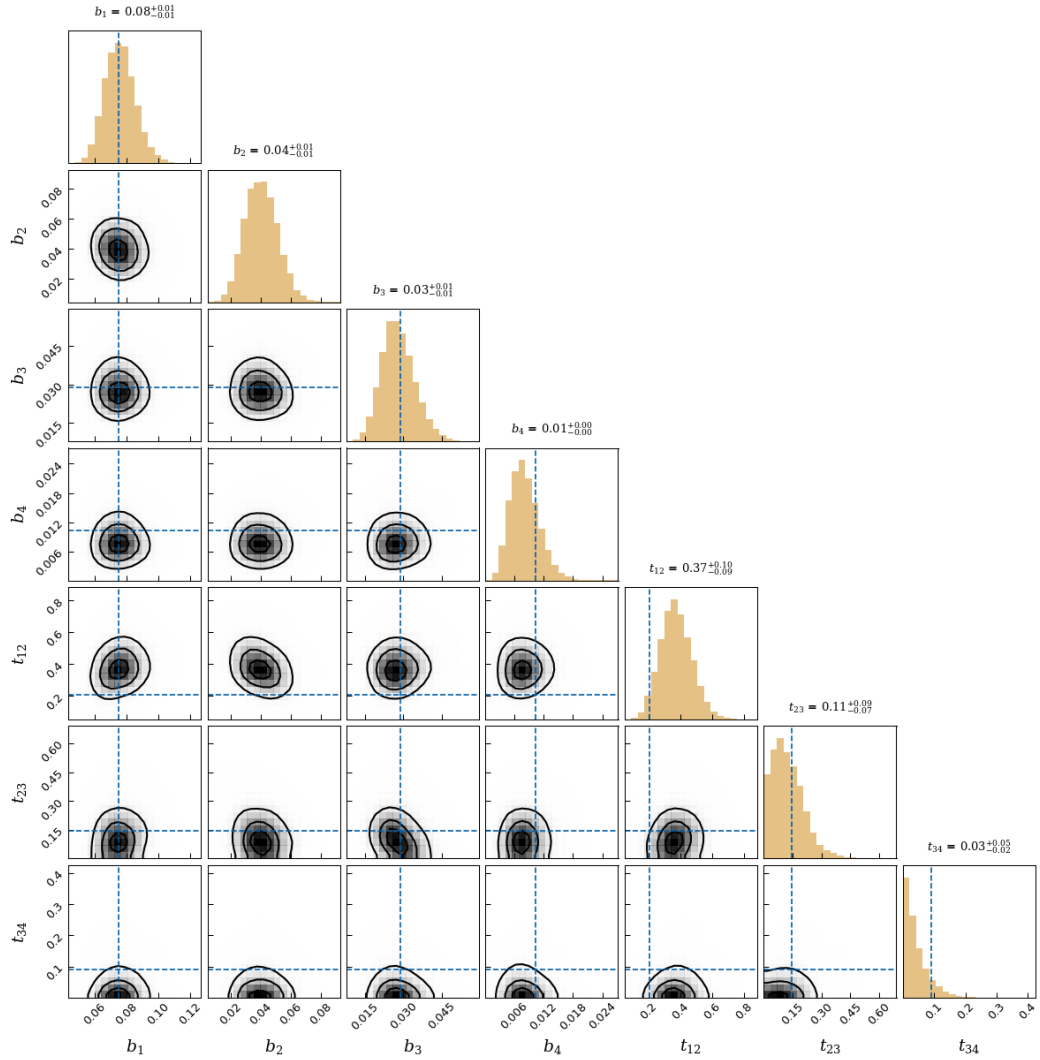


Figure 5.4: Corner plot of the sampled parameters for component B. The histograms on the diagonal show the 1D marginals, while the lower triangle shows all possible combinations of 2D marginals. The blue dashed lines represent the mean value parameters from the HMM trained on oral cavity patients only. The overline for b_2 is not visible in the shown range, since $b_2^{OC} = 0.1$.

A similar interpretation applies to component A, representing oropharyngeal tumor spread. This component focuses on high involvement in LNL II, with $b_2^A = 0.51$ compared to $b_2^{OP} = 0.41$ from the independent model. The considerable uncertainty in t_{1_2} (transition parameter from LNL I to II) results from the minimal involvement of LNL I, such that the overall likelihood is barely affected by this parameter. The model can not determine this parameter without uncertainty. What may be counter intuitive from the explanations so far, is that b_1, which accounts for involvement in

LNL I, is higher in component A than for the independent model. However, this arises due to the influence of subsites like C05, which have higher involvement in this level compared to oropharynx type tumors.

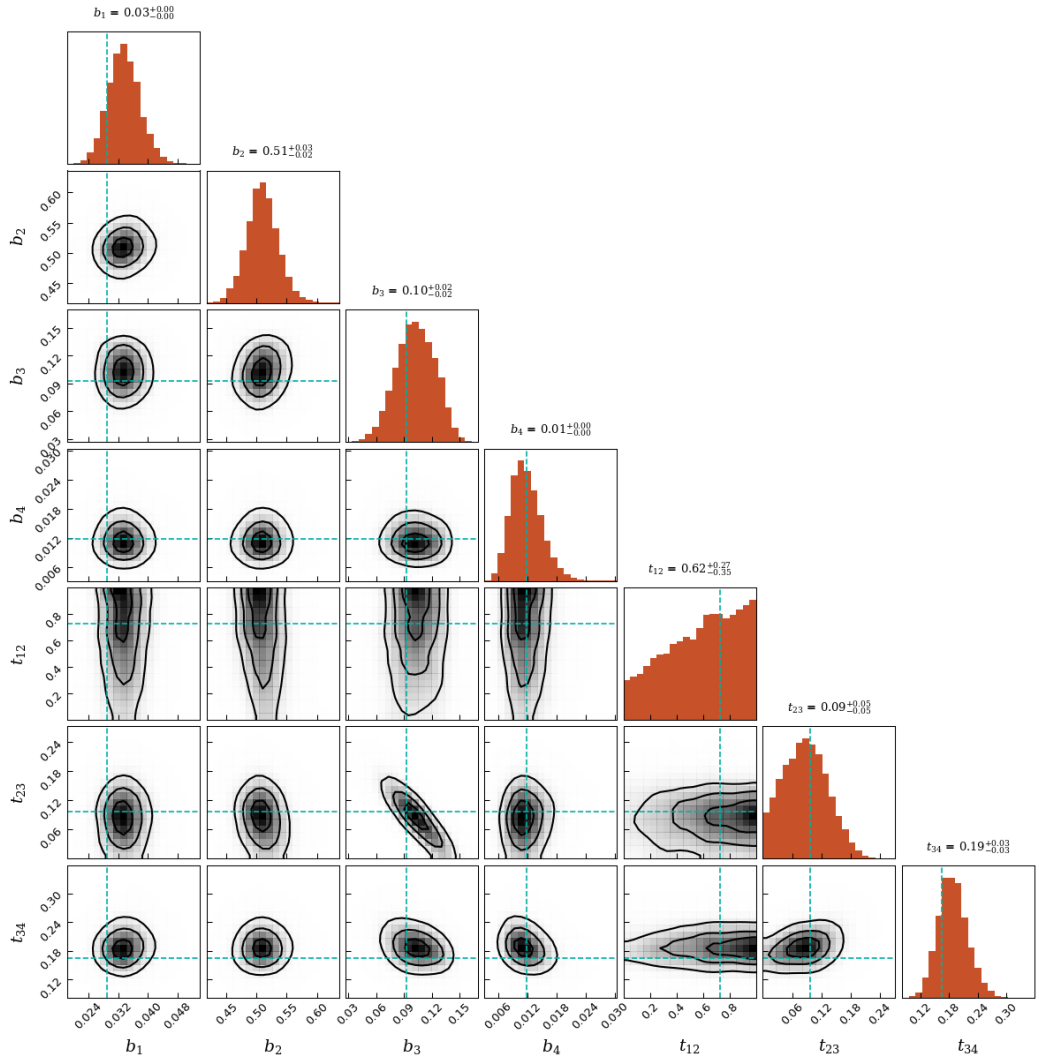


Figure 5.5: Corner plot of the sampled parameters for component A. The greenish dashed lines represent the mean value parameters from the HMM trained on oropharynx patients only. The overline for b_2 is not visible in the shown range, since $b_2^{OP} = 0.41$.

5.1.1 Prevalence predictions

The sampled parameters can be used for prevalence predictions in selected subsites, and compared to predictions from the independent model of the considered subsite.

Prevalence predictions are independent on any observation, thus representing the prior P($\boldsymbol{X}$) over the hidden states, where prevalence in a LNL is given by summation over all states where this LNL is involved.
Figure 5.6 and 5.7 show prevalence predictions for the subsites C03 (Gum) and C05 (Palate), for the levels I to IV. Both subsites are assigned to the tumor location Oral Cavity. The first figure represent C03, which is fully assigned to component B of the MM. From the sampled parameters (Figure 5.4), we see that b_1 of component B is the same as for the independent oral cavity model. This explains the similar predictions for LNL I (upper left subplot). The blue dashed line represents the observed probability of involvement, based on the *maxllh* diagnostic. In LNL II, the observed prevalence is lower than the average in oral cavity subsite. This is captured in the mixture model, as it is able to predict lower involvement of LNL II. Similarly, this is seen in LNL III and IV where in our dataset, the patients have no involvement at all. The mixture model is able to reduce the predicted prevalence in those levels, compared to the independent model.

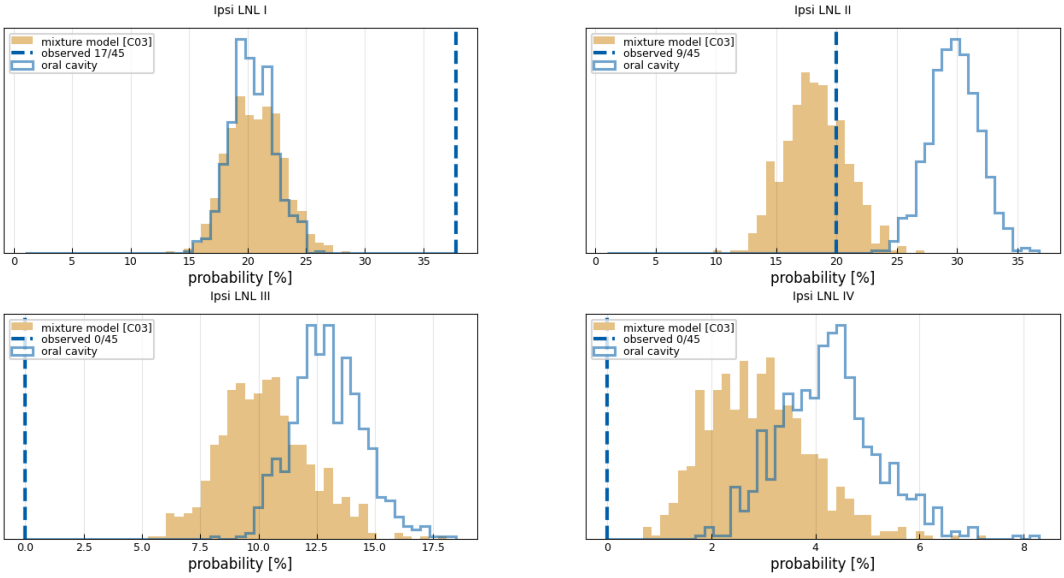


Figure 5.6: The prevalence of involvement for C03 (Gum) as seen in the data (vertical dashed lines, based on *maxllh*), predicted by an independent model for oral cavity patients (blue histograms), and predicted by the mixture model (orange histogram), for the level I to IV

Similar observations can be made for C05 (palate), which has a mixture 56% of component B and 44% of component A (Figure 5.7).

For LNL I, the MM slightly underpredicts the probability of involvement, due to the influence of component A with low base spread to LNL 1. In contrast, involvement in LNL II is accurately captured. In LNL III, one can observe the influence of component A again, pushing the predictions toward high involvement in this LNL. In LNL IV, the mixture model shows slight improvement again versus the independent model.

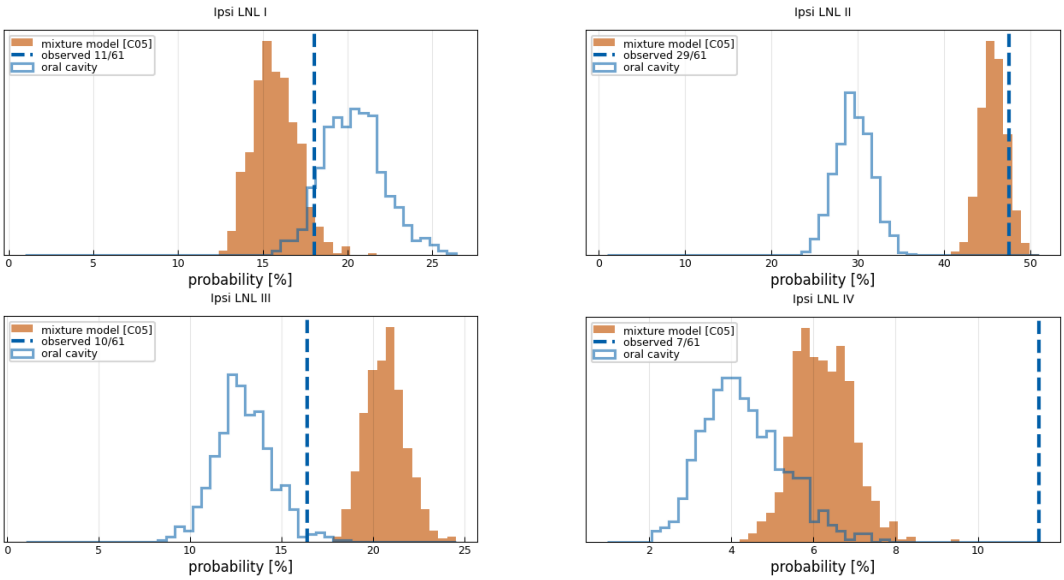


Figure 5.7: The prevalence of involvement for C05 (Palate) as seen in the data (vertical dashed lines, based on *maxllh*), predicted by an independent model for oral cavity patients (blue histograms), and predicted by the mixture model (orange histogram), for the level I to IV

The following two figures show prevalence predictions for the subsites C10 (Oropharynx, Figure 5.8) and C01 (Base of tongue, Figure 5.9), which both belong to the tumor location category oropharynx.

C10, which has a mixture of 72% to component A and 28% to component B, has no overall improvement in the predictions of the MM versus the independent oropharynx model. Generally, we see that the mixture model predicts higher involvement in LNL I than the independent model, which is explained by $b_1^A > b_1^{OP}$. An improvement is seen in LNL II. For the level III and IV, the MM has no advantage against the independent model.

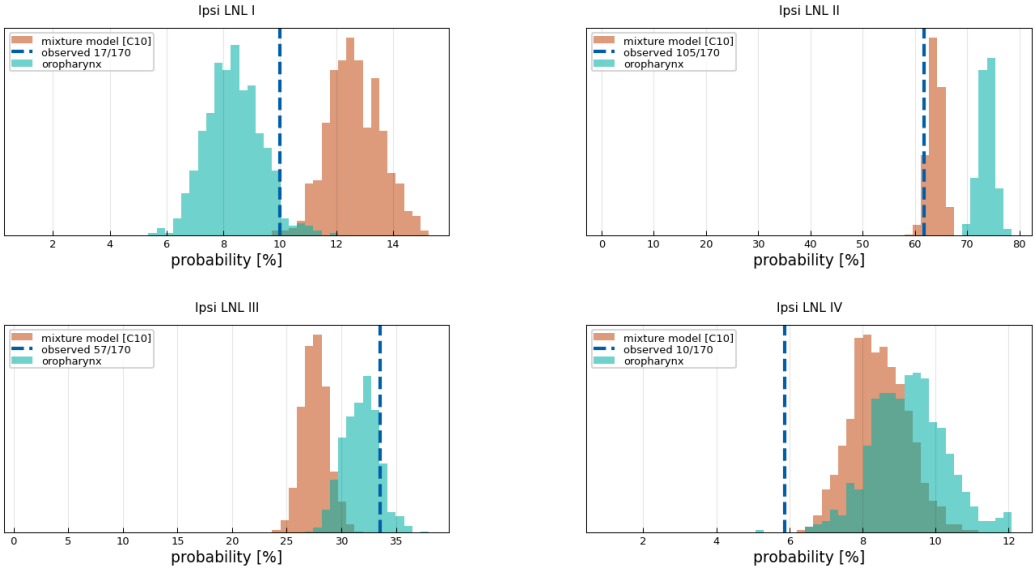


Figure 5.8: The prevalence of involvement for C10 (Oropharynx) as seen in the data (vertical dashed lines, based on *maxllh*), predicted by an independent model for oropharynx patients (green-blue histograms), and predicted by the mixture model (orange histogram), for the level I to IV

In C01, which is fully assigned to component A, the mixture model predicts better than the independent model. This improvement could be attributed due to the full assignment to one component. Especially the high involvement in LNL II can be captured, as well as the higher involvement in LNL III in contrast to the independent model. For LNL IV, both models are equally good in predicting the prevalence.
A complete figure about all the subsite predictions from oral cavity and oropharynx can be found in Figure 8.6 in the appendix.
For a general benchmark for the mixture model's performance, we compare the mean absolute errors of the mixture model versus the independent HMM model per LNL (Table 5.1). The mean absolute error (MAE) is thereby the absolute difference between observed prevalence and predicted prevalence, averaged over all subsites within a tumor location. The following table show the MAE on the prevalence prediction for each LNL, first grouped according to the tumor location categories, and then the averaged MAE.

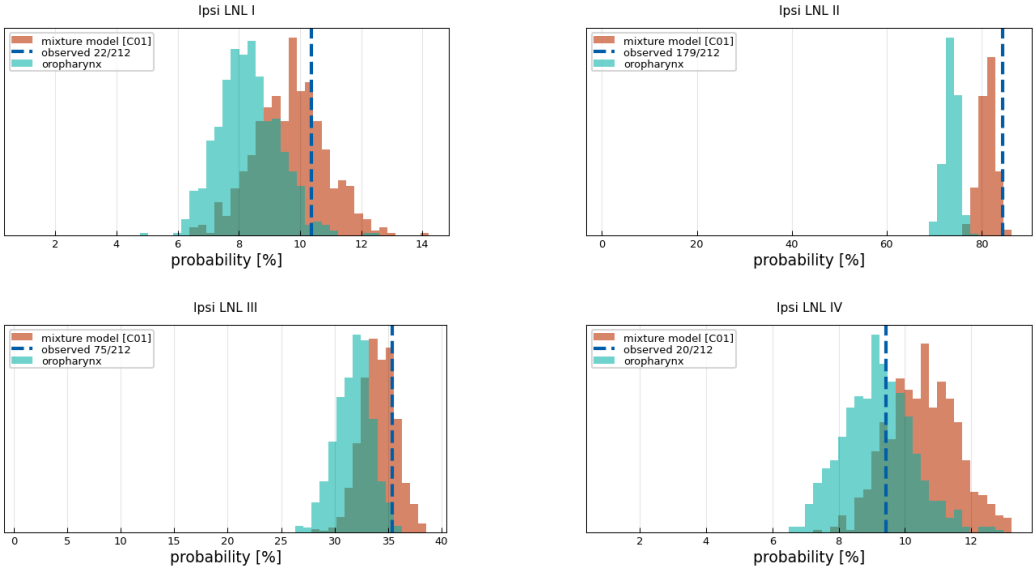


Figure 5.9: The prevalence of involvement for C01 (Base of tongue) as seen in the data (vertical dashed lines, based on *maxllh*), predicted by an independent model for oral cavity patients (blue histograms), and predicted by the mixture model (orange histogram), for the level I to IV

Location	Model	MAE LNL I	MAE LNL II	MAE LNL III	MAE LNL IV
OC	Mixture model	6.2%	1.9%	4.3%	2.7%
	Indp. model (HMM)	6.6%	8.4%	5.5%	4.0%
OP	Mixture model	1.6%	2.3%	3%	1.4%
	Indp. model (HMM)	1.6%	8%	2.4%	1.7%
Total	Mixture model	4.5%	2.0%	3.8%	2.3%
	Indp. model (HMM)	4.6%	8.2%	4.3%	3.1%

Table 5.1: Mean absolute error per LNL for the independent model vs the mixture model. for the tumor location oral cavity and oropharynx.

This table shows that for all LNL, the expected error of the mixture model is lower than for the independently trained HMM. Discussion on the results follow in the discussion section. In a second part of the result section, a mixture model is trained on all tumor subsites available in the dataset.

5.2 Application to whole dataset

In this section we define a mixture model and train it on ICD subsites of oral cavity (OC), oropharynx (OP), hypopharynx (HP) and larynx (LY). The number of subsites is 13. However, we ignore the subsite C00 and C08 due to their low number of patients, such that we are left with 11 subsites. The observed LNL prevalences of each subsite in this application are compared in the following figure:

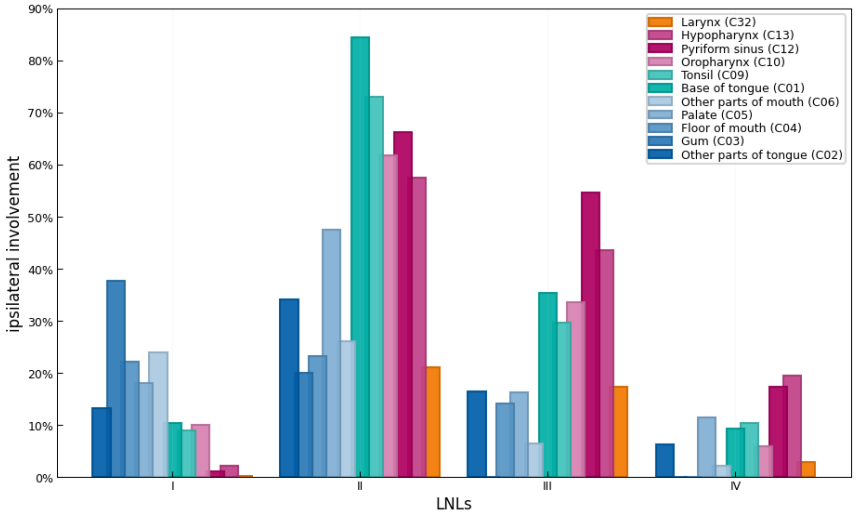


Figure 5.10: Ipsilateral observed prevalence of each subsite in LNL I to IV according to *maxllh*, colorized by the primary tumor location category. Blue is oral cavity, green is oropharynx, violet is hypopharynx and orange is larynx. The ICD code 'C00' and 'C08' are neglected since they count only low number of patients.

The training procedure is similar to the last section, where first the mixture probabilities for the subsites are estimated, and after the component parameters sampled. After that, the prevalence predictions from the mixture model is compared to the independent models. Also here, the independent models are the HMM, which are only trained on one single tumor location category.

For the application, the mixture model is set to have K = 3 components. The MCEM method is used for estimating the mixture probabilities. The MM converges after 43 iterations, where the convergence threshold is $\epsilon = 0.015$ and the look-back period is $n_{\text{iter}} = 5$. A figure about the convergence process of the EM algorithm can be found in the appendix in Figure 8.7.

Figure 5.11 presents the estimated mixture probabilities π_{sk}. The interpretation of this plot is similar to the previous results. Tumors located in the oral cavity region, as C02 to C06, are assigned to component B. Tumors located more in pharyngeal

region are assigned to component C. This includes the subsite C01 and C09 to C13. C32, which itself hold the same name as the tumor category Larynx, is mostly assigned to component A.

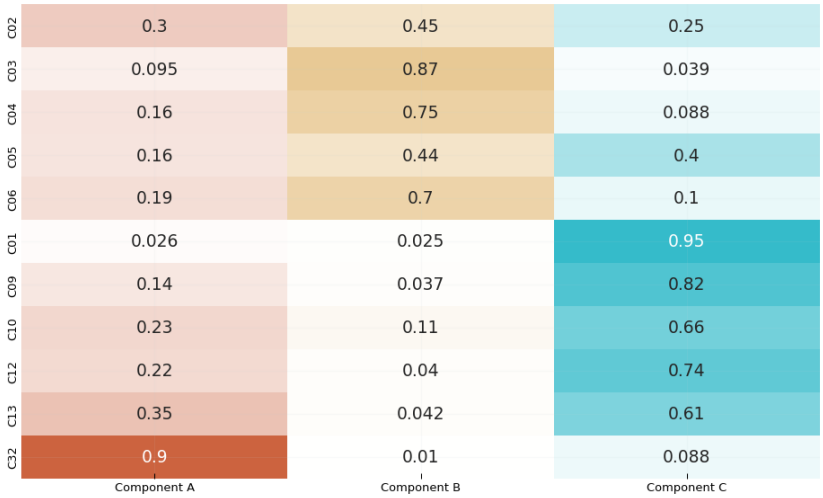


Figure 5.11: Representation of the estimated mixture parameters $\pi_{s,k}$ for all subsites s and components k.

Figure 5.12 presents a different representation of the mixture probabilities: The mixture probabilities are all located on a 2-d hyperplane of the 3-d mixture space, due to the boundary condition that the sum of the mixture probabilities $\pi_{s,\_}$ over all components has to be 1. This space is known as the standard simplex. This representation does not contain more information than the matrix representation from above, however, it holds an interesting interpretation which we will come back to in the discussion section.

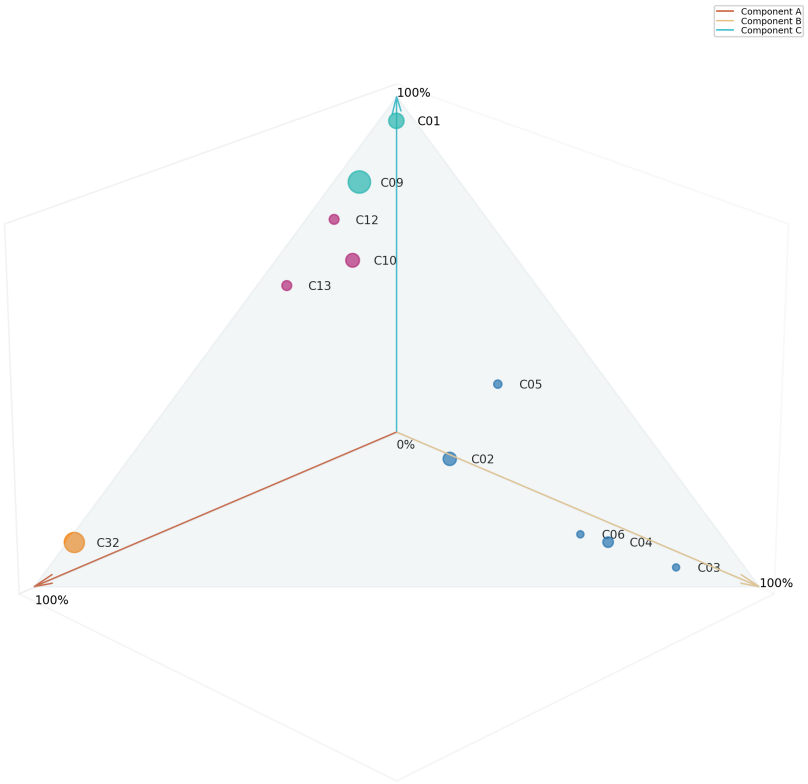


Figure 5.12: 3-D representation of the estimated mixture parameters $\pi_{s,k}$ for all subsites s and components k. The size of the dots represent the number of patients in the subsite, where the color indicates the tumor location, where blue is oral cavity, green is oropharynx, violet is hypopharynx and orange is larynx.

The component parameters for the component A, B and C is in the appendix, and only a short interpretation thereof is given here. Component A, which is dominated by the Larynx subsite C32, generally has low spreading rates overall, which is in agreement with Larynx tumors. Component B, which is dominated by Oral Cavity tumor subsites, have proportionally high spread rates in LNL I and low rates in all other LNLs. Finally component C, dominated by Oropharynx and Hypopharynx subsites, have general higher rates, especially in LNL II and III, which also is in agreement with the observed involvement.

The complete Figure of the prevalence predictions for all subsites can be found in the appendix in figure 8.11. The overall performance of the MM, compared to the four independently trained HMM is shown in the following table. This table compares the mean average error in the prediction versus the observed prevalence. The total

Location	Model	MAE LNL I	MAE LNL II	MAE LNL III	MAE LNL IV
OC	Mixture model	3.6%	4.3%	7.9%	4.5%
	Indp. model (HMM)	6.5%	8.4%	5.5%	4.0%
OP	Mixture model	1.5%	5.6%	2.9%	2.4%
	Indp. model (HMM)	1.3%	7.8%	2.4%	1.6%
HP	Mixture model	3.9%	2.4%	12.9%	6.8%
	Indp. model (HMM)	3.0%	3.1%	9.1%	5.3%
LY	Mixture model	1.7%	1.2%	1.2%	1.8%
	Indp. model (HMM)	0.3%	0.1%	0.04%	0.5%
Total	Mixture model	3.0%	3.9%	7.4%	4.4%
	Indp. model (HMM)	3.8%	6.1%	5.1%	3.4%

Table 5.2: Mean absolute error per LNL for the independent model vs the mixture model. for the tumor location oral cavity (OC), oropharynx (OP), hypopharynx (HP) and larynx (LY).

MAE shows, that especially in LNL I and II, we have improvement against the independent models. In LNL III and IV, the independent models performed better. Further, especially for tumor location categories which contain only few subsites, as example Larynx with only one subsite, the independent model performs significantly better than the mixture model over all LNLs. Further discussion thereof will be made in the next section.

The application of the mixture model on all subsites showed that the mixture probabilities are interpretable and the model performs good in predicting the prevalence. However, further adjustments have to be considered to eventually use this full model for risk predictions in clinical applications.

5.3 Risk Predictions with the K2 Mixed-HMM

Finally, we can refer back to the initial idea of computing the risk in a LNL, given some observation. This section serves more as an example, rather than using the results in some sort of clinical trial. Due to the better performance of the mixture model which is trained on the Oral Cavity and Oropharynx tumors, we use this model for this final risk prediction application:

We choose the scenario of having a patient with a tumor located in the Gum (C03). The patient has a no clinically observed involvement in any LNL, based on a CT. According to the current guidelines, LNL I, II, III and IV would all be included in the CTV-N.

For the risk prediction, we randomly sample 1000 sets of component parameters, to also estimate the uncertainty in the predictions. The result is compared to the HMM, trained on the oral cavity patients (independent model).

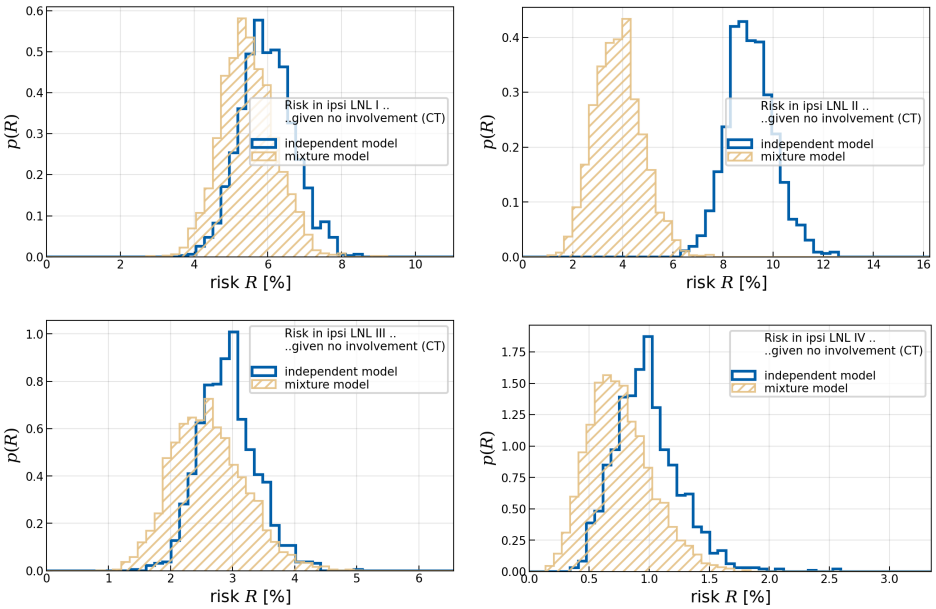


Figure 5.13: Risk prediction for C03 (Gum) of the mixture model (orange) and the independent HMM (blue) in the LNL I to IV.

In LNL I, the risk prediction of both models are almost equal, with a risk of 6%. In LNL II, the mixture model predicts a lower involvement than the independent model, with an expectation of around 4%. This result might be a correct improvement since the prevalence prediction for LNL II in C03 is closer to the observed prevalence.
The same interpretation holds true for LNL III and IV, the mixture model predicts a slightly lower risk of occult metastases of 2.6% and 0.7%.
This example proves that the mixture model is capable of predicting the risk of occult disease in the LNL, given a patients observation, and the primary tumor subsite. In this specific example of a patient with primary tumor in the Gum (C03), and no clinical involvement, the risk of involvement in LNL I to IV is low. Assuming a risk threshold of 5%, on can argue to exclude LNL II, III and IV from the CTV. However, in clinical settings, usually the cumulative risk of occult disease in all excluded LNLs should not surpass a given risk threshold, such that in this limited example with only 4 LNL the LNLs to exclude from the CTV can not be easily answered.

6 Discussion

The previous chapter demonstrated two scenarios how mixture models (MM) can be applied to a set of HNSCC patients: First is a MM with two components, trained on ICD subsite codes of Oral Cavity (OC) and Oropharynx (OP). In a second scenario we presented a MM with three components, trained on all available ICD codes from the four tumor location categories Oral Cavity, Oropharynx, Hypopharynx (HP) and Larynx (LY). Both models had the same underlying base model, namely a HMM as proposed by Ludwig et al. In both scenarios, the learned component assignments are reasonable, and their interpretations can be substantiated by either the observed prevalence patterns or their respective anatomical location. Moreover, both models have demonstrated their capability to accurately describe the spreading character expressed by individual ICD codes within the tumor locations. This was the primary goal of this thesis. Notably, ICD codes that exhibit intermediate spreading behavior, i.e. ICD codes which cannot be clearly assigned to a single tumor location category, are successfully predicted by the MM. One such example is Palate (C05), which is, according to the literature, associated with Oral Cavity, but has an atypical high involvement in LNL II. Both MM evenly distributed C05 across the components representing OC and OP spreading characters. This equitable assignment results in more accurate prevalence predictions in the LNLs compared to HMMs trained on tumor location categories, which we have referred to as the *independent* models.
The ICD code C02, which is assigned to OC, was also moderately distributed among the components, especially evident in Figure 5.12 for the MM with K=3 components. The observed prevalences for C02 underscore this: C02 does not exhibit high involvement in LNL I or II, yet LNLs 3 and 4 are disproportionately affected. This intermediate behaviour is therefore correctly identified by the model. However, I will not delve further into ICD code C02 here, as its definition, 'other and unspecified parts of the tongue,' lends itself to a broad range of interpretations.
From a prediction point of view, especially the first model with K=2 components has proven to be an effective model to predict the LNL involvement. According to table 5.1, the mean expected error on LNL prevalence prediction is smaller compared to

two independently trained HMMs for OC and OP. This is good, however, it requires further consideration: The MM, in its entirety, possesses more parameters and thus more degrees of freedom to describe the data. The additional parameters in the MM are the mixing probabilities (or component assignments), making the model more adaptable to the data. Nonetheless, the outcome of the model prediction is sufficient to depict the diverse spreading characteristics of the subsites.
The same expectations were held for the MM with K=3 components, which was used to describe the ICD Codes of all four tumor locations. The rationale for choosing three components was that OP and HP generally exhibit similar characteristics, with high involvement and thus could be described with one component. Patients with larynx tumors differ slightly from other tumor locations, as the involvement is generally lower in all LNL, but with disproportional high involvement in LNL III and IV. The idea is that a second component focuses on this spreading character. A third component was expected to be dominated by Oral cavity patients. The learned cluster assignments meet these expectations, as shown in figure 5.11. However, the prevalence predictions of the model, compared to the four independently trained HMM, did not significantly improve. The interpretation here is that three components are not sufficient to describe the 11 different ICD codes, as the prevalences of all those subsites cannot be created by mixture of only three spreading patterns. Especially the higher involvement of LNL III and IV of HP patients can no not be described with the three components. A solution is to use a MM with K=4 components, to match the number of independent HMM.
However, this reveals a disadvantage of the model: the number of parameters increases more than linearly with the number of clusters, as both the component parameters and the new mixture parameters for the new component are added. With that, the model would count $K-1$ times 11 (the number of subsite) more parameters than the combined independent models, due to the mixture components.[1]
With that, the problem of overfitting could occur. For instance, a model with K=4 components, intended to describe a graph with 4 LNLs, could distribute each LNL to a single component and simply divide the ICDs according to the prevalence of the ICD code. This would, despite the good performance of the model, not capture any correlations between the LNLs, and thus could not be used as a prediction model where we might have clinical observations of involvement of a LNL.
Another intriguing feature, which I like to point out, can be observed in the component assignments for the K=3 MM: Upon closer inspection of the 3-D representation

[1] The K minus one arises due to the normalization condition on the mixture parameters.

in figure 5.12, it can be seen that the cluster assignments accurately represent the anatomical locations of the individual ICD codes. For example, if one imagines the figure mirrored along the vertical axes. The gum (C03), located the most anterior region of the oral cavity is mostly assigned to component B. The palate (C05) is medial between OC and OP, also represented in the figure. The larynx icd code (C32) is situated significantly inferior to the oral cavity and oropharynx, also represented in the figure. Last, hypopharyngeal ICDs are situated intermediate LY and OP, but closer to OP.
Overall, what initially sounds like a very abstract concept — namely a mixture of generalized HMM components — turns out to be an interpretable and effective model in predicting LNL involvement.

7 Conclusion

In this work, we developed a Mixed Hidden Markov Model (Mixed-HMM) capable of characterizing individual ICD codes and predicting the prevalence (and thus risk) of occult metastases in lymph node levels (LNLs). This is crucial because different ICD codes, often grouped into broader tumor locations, exhibit distinct lymphatic spreading patterns. A naive approach to capture these variations would involve training individual HMMs for each ICD code. However, this faces two significant challenges: Each ICD code would require a substantial number of patients for robust model training. Fewer patients lead to increased prediction uncertainty, undesirable in clinical applications. Further, this approach would not consider any similarities between the ICD codes, even tough they spread though the same biological lymphatic network. Therefore, we opted for a Mixed-HMM capable of describing individual ICD codes without significantly increasing the model's parameter count. The mixed model demonstrated accurate predictions, even for ICD codes with limited patient data. While seemingly abstract at first, the results offered highly interpretable insights that, to some extent, reflect the anatomical arrangement of lymph nodes.

An additional, yet hypothetical, feature of the Mixed-HMM is its potential to predict previously unseen ICD codes. The challenge lies in estimating component assignments for the new code, potentially using a model capable of mapping the anatomical location of the ICD code to a mixture component within the model. Whether this feature will be utilized remains to be seen.

Finally, I return to the core research question: how can we better describe individual patients risk in the lymph node levels? This model presents a valuable method to integrate additional patient features, like the ICD code into the predictive model. Our team is actively involved in refining new guidelines to assist physicians and oncologists in developing patient-specific therapies for HNC patients. These guidelines will be built upon the predictions of the probabilistic model, whether from an independent HMM or, as introduced here, a Mixed-HMM, or a combination of both remains to be seen.

8 Appendix

8.1 Corner Plots of independent models

As a reference, the sampled parameters from the independent models (i.e. HMM trained only on one single tumor location) are shown here.
All independent models use the same graph structure as shown in Figure 2.5, where the time prior is fixed for all t-categories. The time prior is a binomial distribution parameterized by $\rho = 0.3$, where the HMM evolves over $t_{max} = 10$ timesteps.

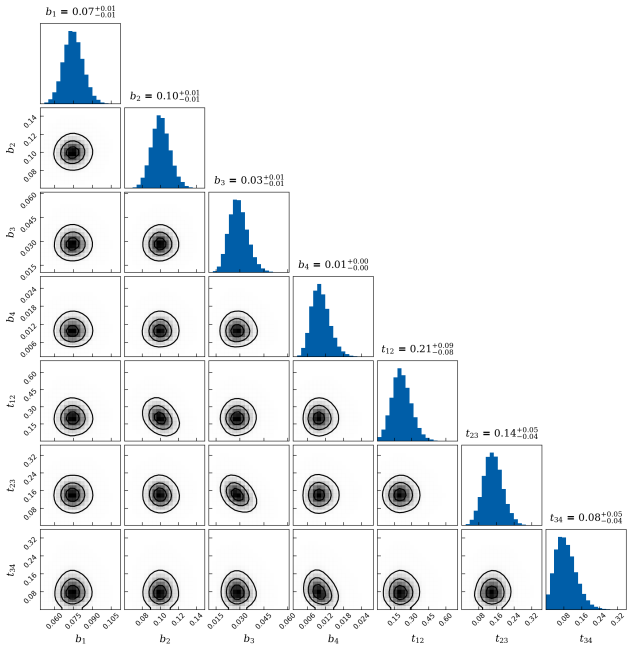


Figure 8.1: Corner Plot for independent HMM trained on pooled Oral Cavity patients

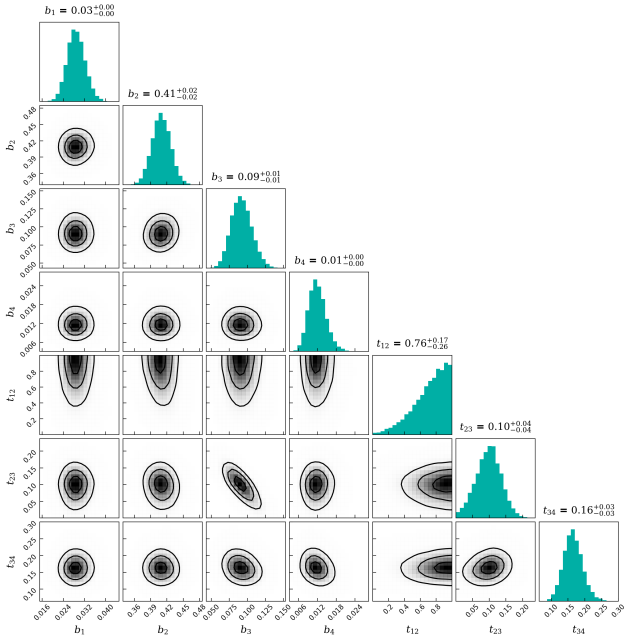


Figure 8.2: Corner Plot for independent HMM trained on pooled Oropharynx patients

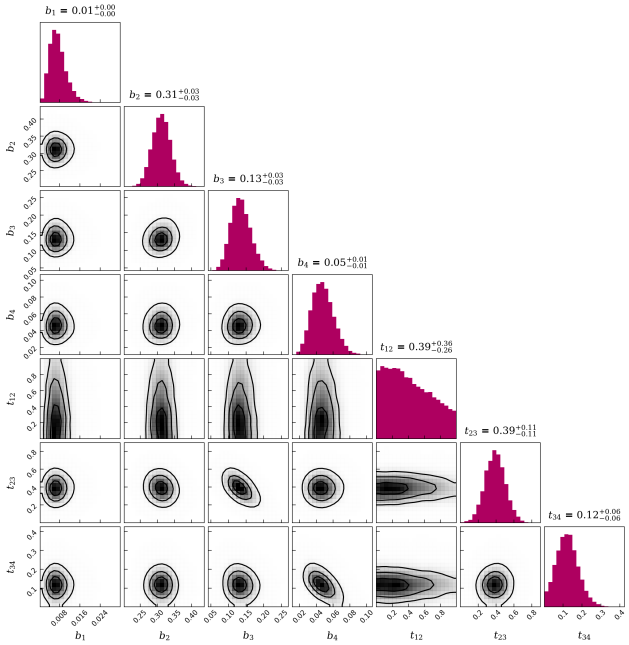


Figure 8.3: Corner Plot for independent HMM trained on pooled Hypopharynx patients

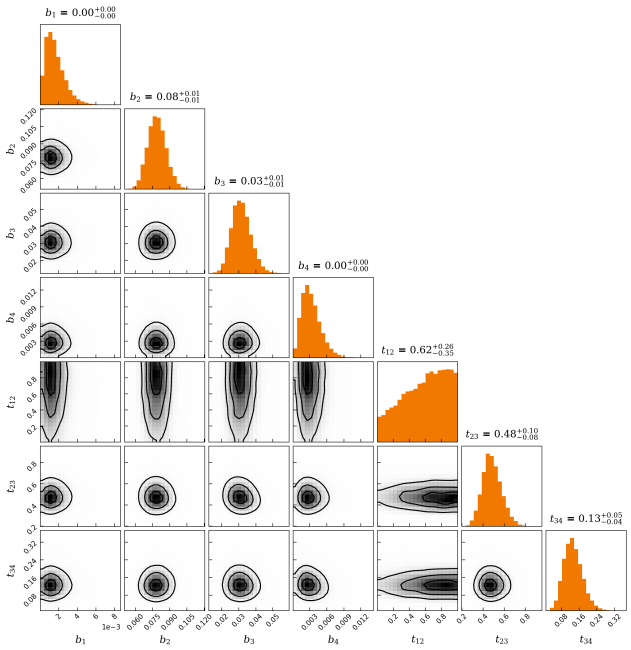


Figure 8.4: Corner Plot for independent HMM trained on pooled Larynx patients

8.2 Restricted Parameter Space for the MM likelihood

Due to the multimodal likelihood function of the Mixture Model in 3.2, traditional MCMC methods can not be used for an interpretable solution. One method to overcome this multimodality is by imposing constraint on the parameter space for the component parameters.

Imaging a simple setup with subsite A and subsite B. A has almost no involvement in LNL 2 and B has high involvement in LNL 2. By introducing constraints like $b_2^2 > b_2^1$ (where b_2^1 and b_2^2 are the base rates for component 1 and 2, respectively), we can force the model towards a unique assignment of subsites to components. Figure 8.5 illustrates this idea:

With this constraint, subsite B is forced to be assigned to component 2, and with that, the solution of the maximizers of the likelihood function would not be degenerated anymore.

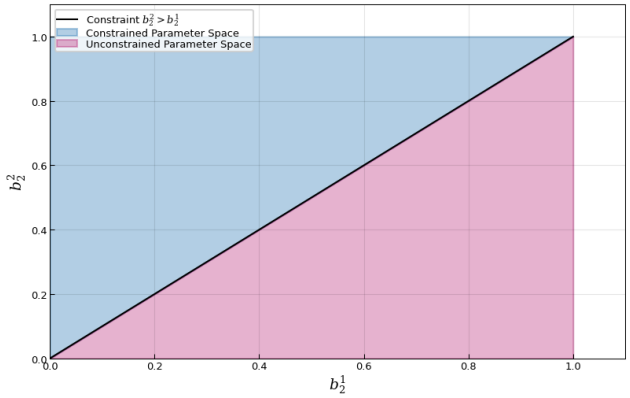


Figure 8.5: Parameter space for base rates with a constraint. The x-axis represents b_2^1, and the y-axis represents b_2^2. The shaded gray area is the unconstrained space, while the blue area above the line $y = x$ shows the constrained space where $b_2^2 > b_2^1$. This constraint helps in achieving a unique solution.

8.3 EM Algorithm: Proof of Q maximization

Given the Q function defined by:

$$Q(\Theta, \Theta_0) = \int \log(P(\boldsymbol{\Theta}|\boldsymbol{\Pi}, \boldsymbol{D}))P(\boldsymbol{\Pi}|\boldsymbol{\Theta}_0, \boldsymbol{D})d\Pi, \tag{8.1}$$

and applying Jensen's inequality to the concave function log, we have:

$$Q(\Theta, \Theta_0) \leq \log\left(\int P(\boldsymbol{\Theta}|\boldsymbol{\Pi}, \boldsymbol{D})P(\boldsymbol{\Pi}|\boldsymbol{\Theta}_0, \boldsymbol{D})d\Pi\right). \tag{8.2}$$

If we find a Θ^* such that $Q(\Theta^*, \Theta_0) > Q(\Theta_0, \Theta_0)$, then:

$$\log\left(\int P(\Theta^*|\boldsymbol{\Pi}, \boldsymbol{D})P(\boldsymbol{\Pi}|\boldsymbol{\Theta}_0, \boldsymbol{D})d\Pi\right) > \log\left(\int P(\Theta_0|\boldsymbol{\Pi}, \boldsymbol{D})P(\boldsymbol{\Pi}|\boldsymbol{\Theta}_0, \boldsymbol{D})d\Pi\right). \tag{8.3}$$

Since the logarithm is a monotonically increasing function, this implies:

$$\int P(\Theta^*|\boldsymbol{\Pi}, \boldsymbol{D})P(\boldsymbol{\Pi}|\boldsymbol{\Theta}_0, \boldsymbol{D})d\boldsymbol{\Pi} > \int P(\Theta_0|\boldsymbol{\Pi}, \boldsymbol{D})P(\boldsymbol{\Pi}|\boldsymbol{\Theta}_0, \boldsymbol{D})d\Pi. \tag{8.4}$$

By Bayes' theorem, where $P(\Theta|\boldsymbol{D}) \propto P(\boldsymbol{D}|\Theta)$, this leads to the conclusion that:

$$P(\boldsymbol{D}|\boldsymbol{\Theta}^*, \Pi) > P(\boldsymbol{D}|\boldsymbol{\Theta}_0, \Pi). \tag{8.5}$$

8.4 Additional Figures from results section

Here, we present additional information for the result section. In the result section I, we presented the application of a mixture model with $K = 2$ components, where we consider only the oral cavity and oropharynx. This leads to 7 subsites in total. The full prevalence prediction for all 7 subsites can be found in the following figure:

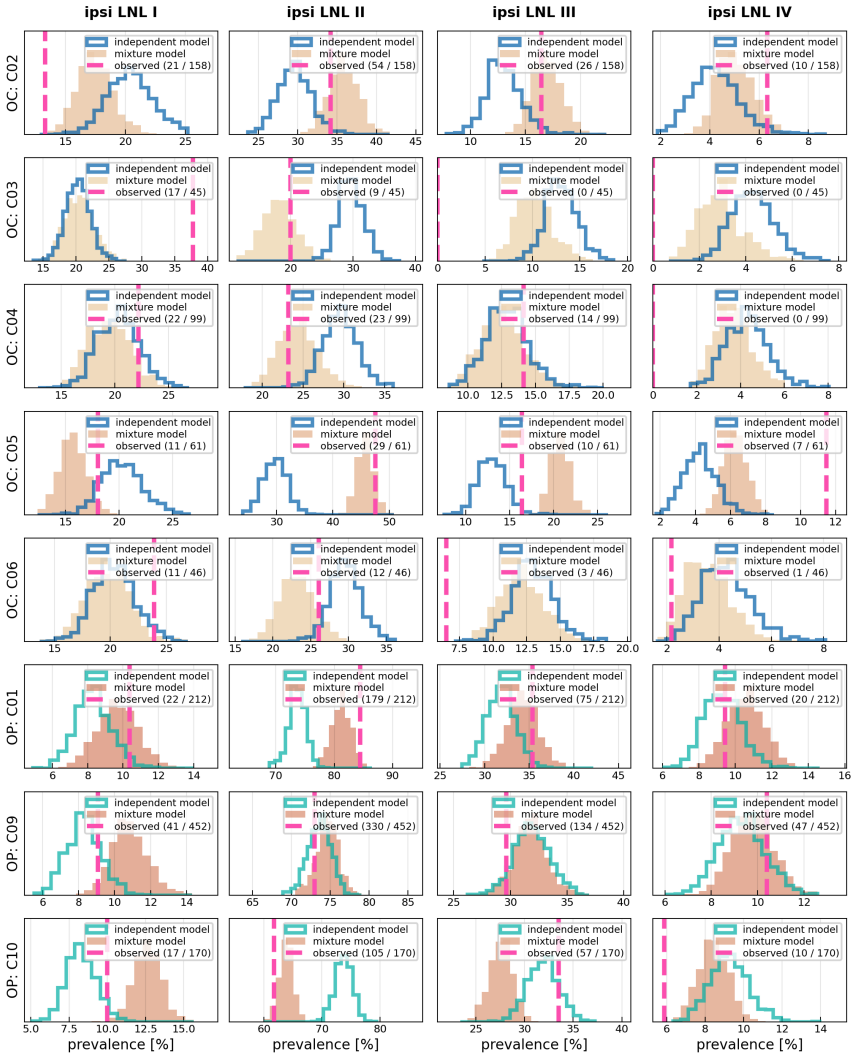


Figure 8.6: Prevalence predictions for all subsites from oral cavity and oropharynx. The outlined histograms show the predictions from the independent models. The filled histograms show the predictions from the mixture model. The blue dashed line show the observation based on the *maxllh*.

Section II

This section contains further information behind the second part of the results, where we defined a MM with K = 3 components and trained it on the dataset containing 11 distinct subsites. The convergence process of the EM algorithm is plotted in the following figure:

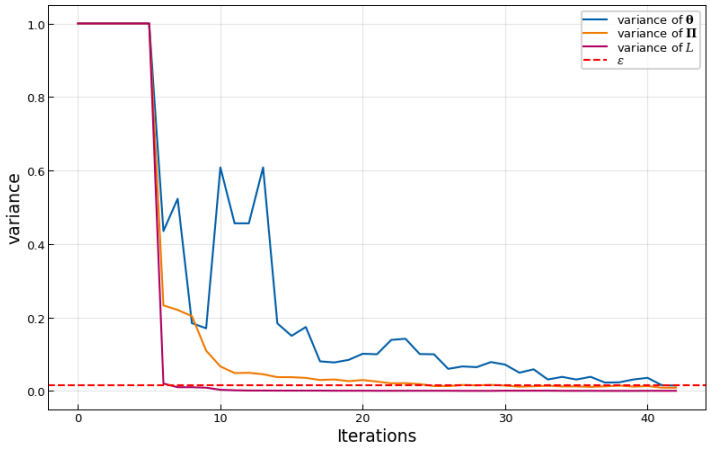


Figure 8.7: Convergence process of the EM-algorithm. The variance of the component parameters, the mixture probabilities and the log-likelihood function over the last 5 iterations versus the iterations of the EM algorithm. The red dashed line indicates the convergence threshold of 0.015.

The estimated mixture probabilities are then used to sample the component parameters for component A, B and C

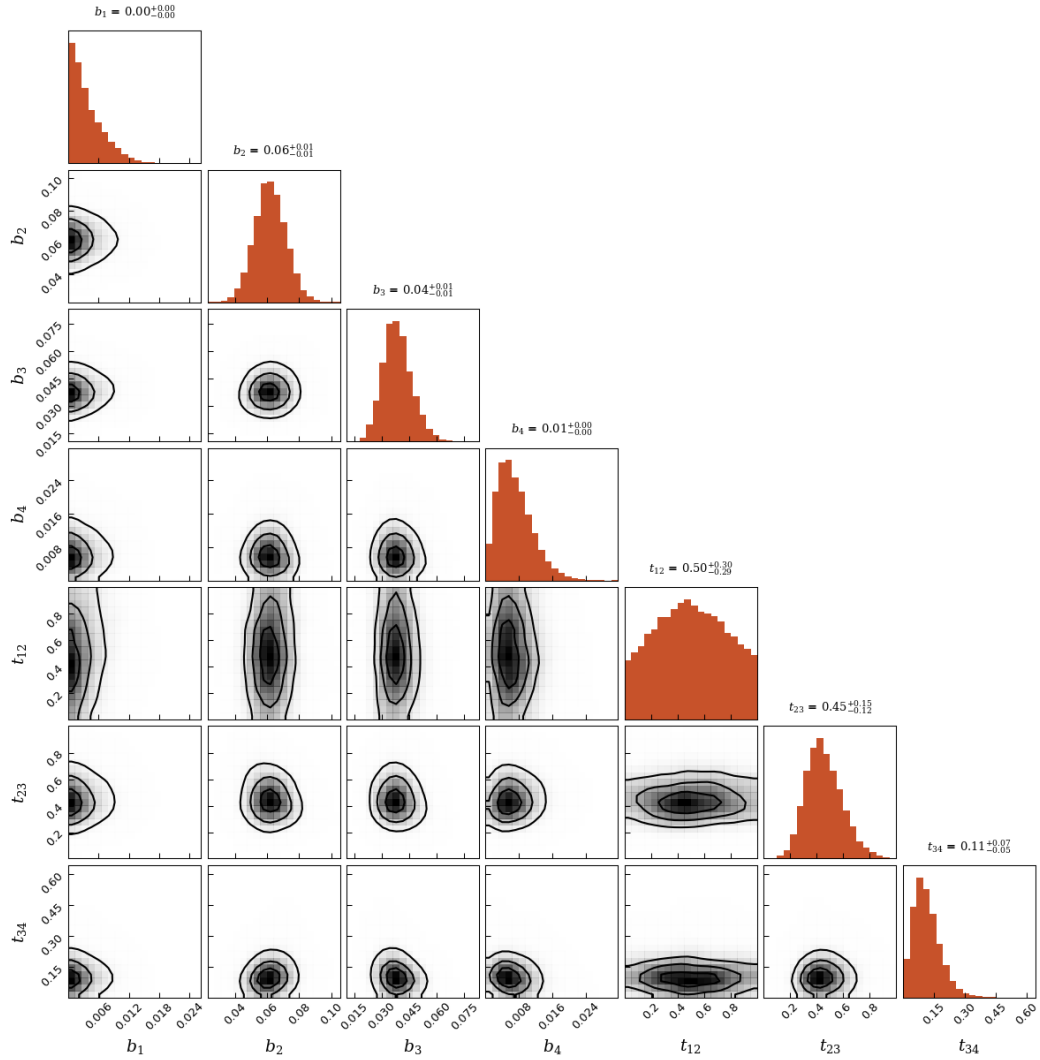


Figure 8.8: Corner plot of the sampled parameters for **component A**. The histograms on the diagonal show the 1D marginals, while the lower triangle shows all possible combinations of 2D marginals. The component is dominated by Larynx subsites.

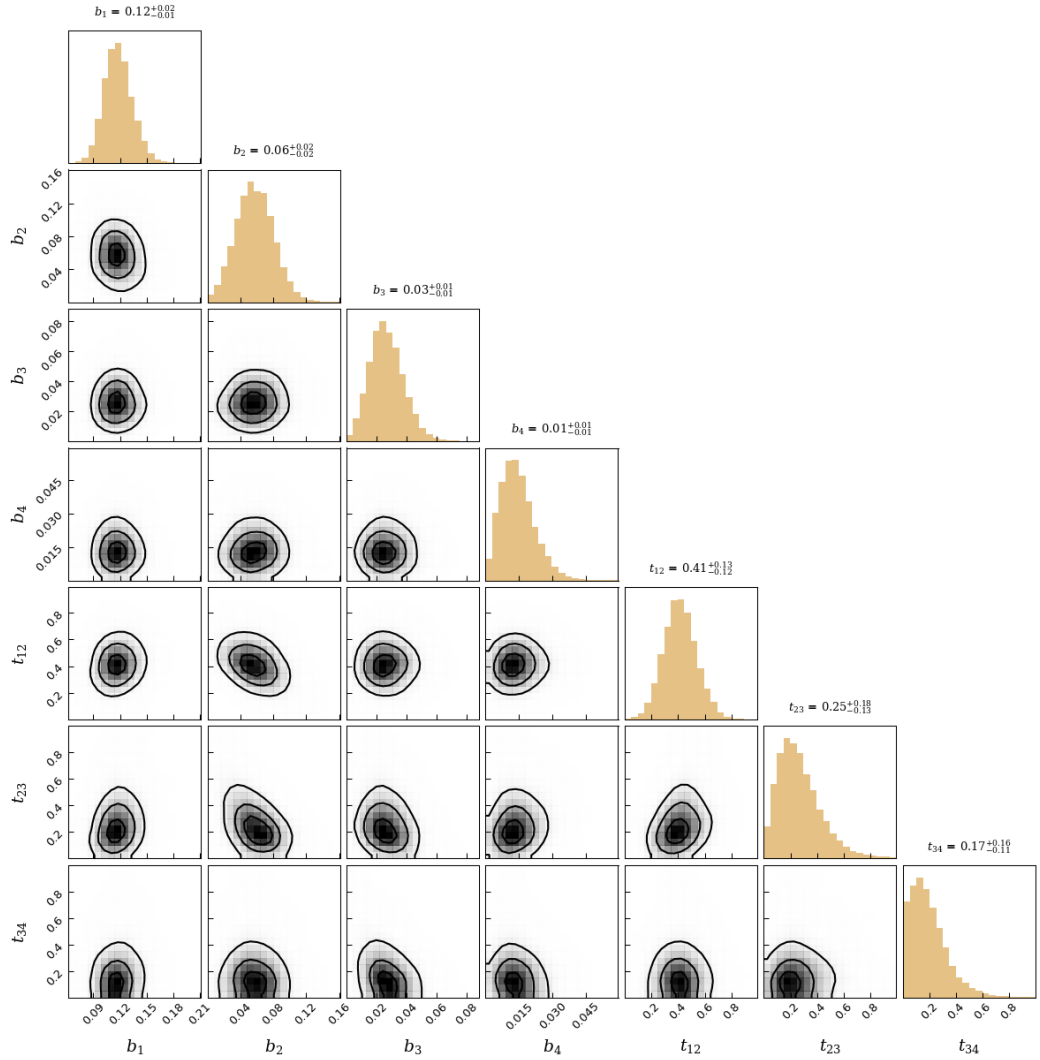


Figure 8.9: Corner plot of the sampled parameters for **component B**. The histograms on the diagonal show the 1D marginals, while the lower triangle shows all possible combinations of 2D marginals. The component is dominated by Oral Cavity subsites.

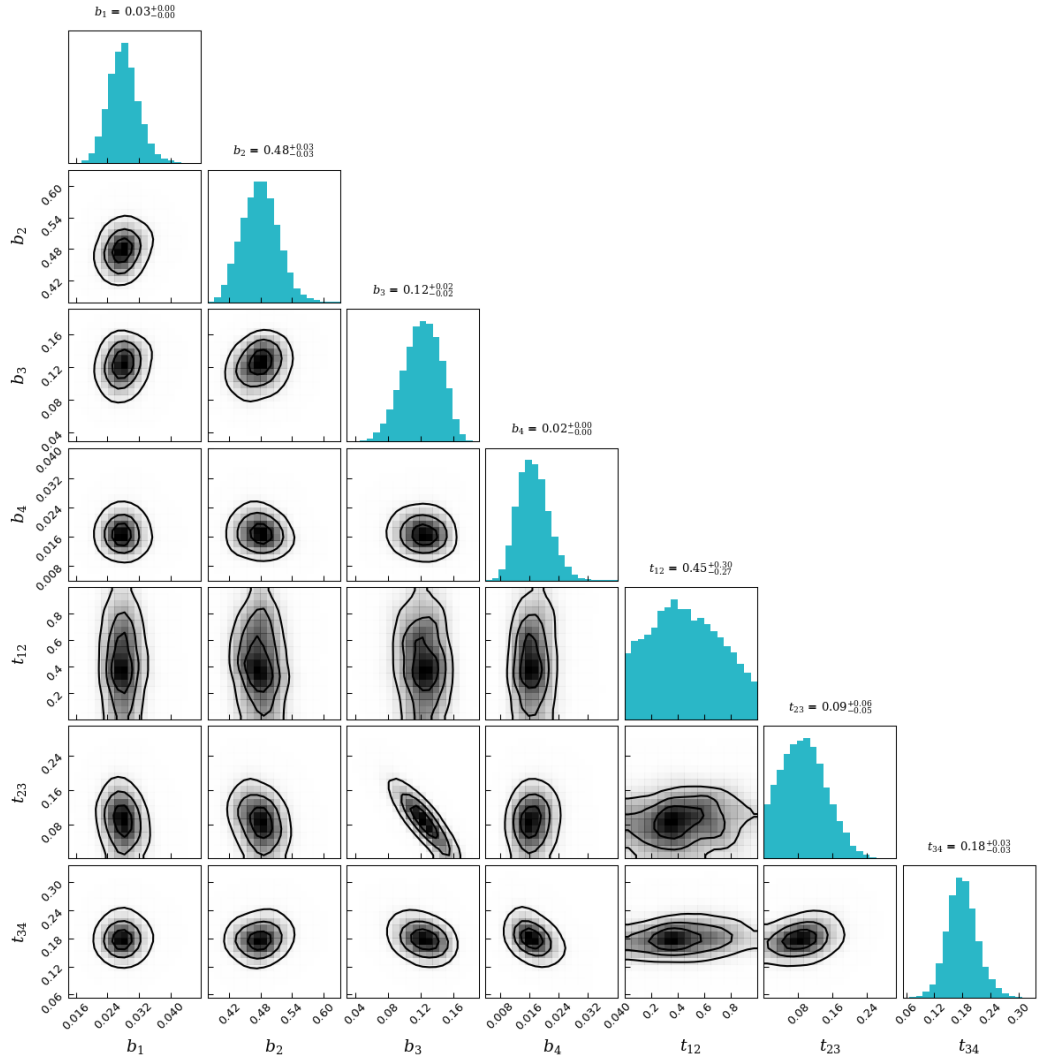


Figure 8.10: Corner plot of the sampled parameters for component B. The histograms on the diagonal show the 1D marginals, while the lower triangle shows all possible combinations of 2D marginals. The component is dominated by Oropharynx and Hypopharynx subsites.

With the component parameters and the final mixture probabilities, we can estimate the prevalence in each LNL for each subsite:

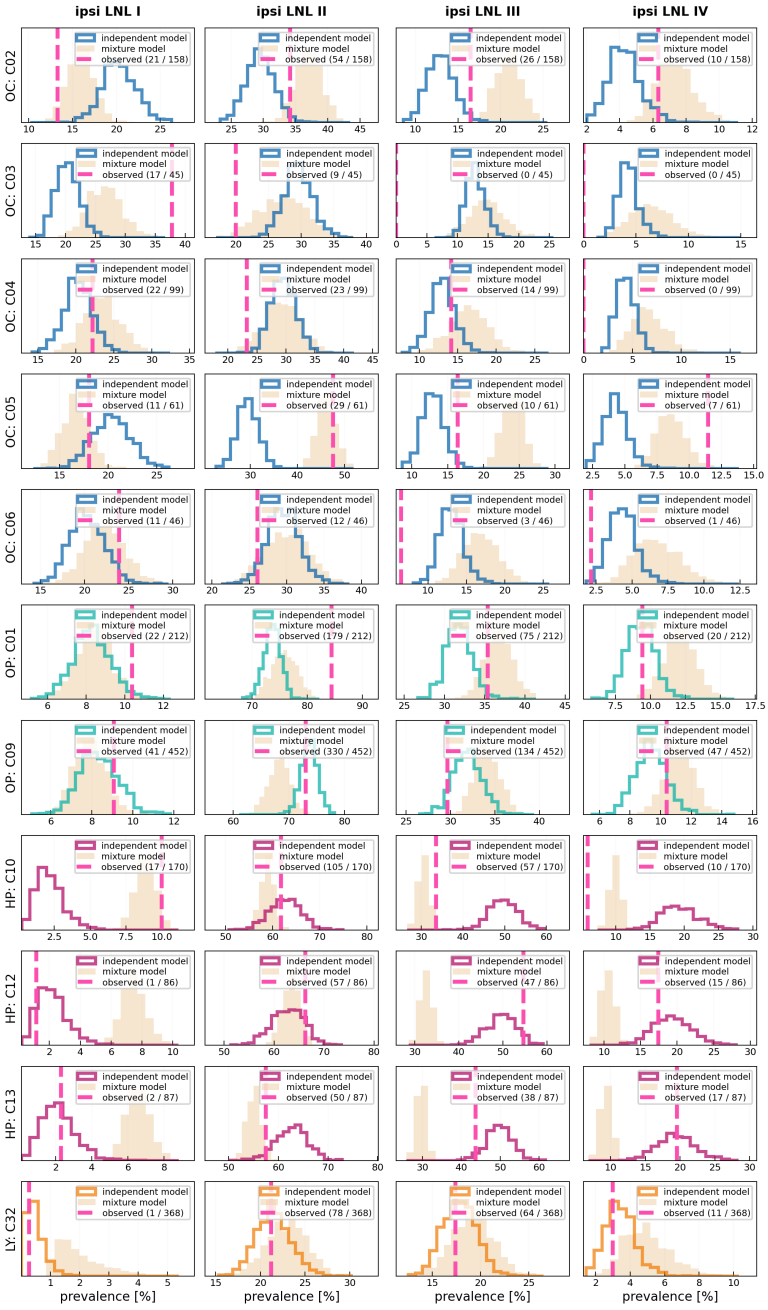


Figure 8.11: Prevalence predictions for all subsites. The outlined histograms show the predictions from the independent models. The filled histograms show the predictions from the mixture model. The blue dashed line show the observation based on the *maxllh*.

Bibliography

[1] Roman Ludwig. *Modelling Lymphatic Metastatic Progression in Head and Neck Cancer*. PhD thesis, University of Zurich, 2022-10.

[2] Julian Biau, Michel Lapeyre, Idriss Troussier, Wilfried Budach, Jordi Giralt, Cai Grau, Joanna Kazmierska, Johannes A Langendijk, Mahmut Ozsahin, Brian O'Sullivan, et al. Selection of lymph node target volumes for definitive head and neck radiation therapy: a 2019 update. *Radiotherapy and Oncology*, 134:1–9, 2019.

[3] Robert Lindberg. Distribution of cervical lymph node metastases from squamous cell carcinoma of the upper respiratory and digestive tracts. *Cancer*, 29(6):1446–1449, 1972.

[4] JA Woolgar. Histological distribution of cervical lymph node metastases from intraoral/oropharyngeal squamous cell carcinomas. *British Journal of Oral and Maxillofacial Surgery*, 37(3):175–180, 1999.

[5] Shu-Chun Chuang, Ghislaine Scelo, Jon M Tonita, Sharon Tamaro, Jon G Jonasson, Erich V Kliewer, Kari Hemminki, Elisabete Weiderpass, Eero Pukkala, Elizabeth Tracey, et al. Risk of second primary cancer among patients with head and neck cancers: a pooled analysis of 13 cancer registries. *International journal of cancer*, 123(10):2390–2396, 2008.

[6] Rosemary Martino and Jolie Ringash. Evaluation of quality of life and organ function in head and neck squamous cell carcinoma. *Hematology/oncology clinics of North America*, 22(6):1239–1256, 2008.

[7] Roman Ludwig, Adrian Schubert, Dorothea Barbatei, Laurence Bauwens, Sandrine Werlen, Olgun Elicin, Matthias Dettmer, Philippe Zrounba, Panagiotis Balermpas, Bertrand Pouymayou, et al. A multi-centric dataset on patient-individual pathological lymph node involvement in head and neck squamous cell carcinoma. *Data in Brief*, page 110020, 2023.

[8] Parkin Dm. Global cancer statistics, 2002. *Ca Cancer J Clin*, 55:74–108, 2005.

[9] Barbora Peltanova, Martina Raudenska, and Michal Masarik. Effect of tumor microenvironment on pathogenesis of the head and neck squamous cell carcinoma: a systematic review. *Molecular cancer*, 18(1):1–24, 2019.

[10] Benoit Lengelé and Pierre Scalliet. Anatomical bases for the radiological delineation of lymph node areas. part iii: Pelvis and lower limbs. *Radiotherapy and Oncology*, 92(1):22–33, 2009.

[11] Widmer, Lars and Unkelbach, Jan and Ludwig, Roman. Modelling Lymphatic Metastatic Progression in Oral Cavity Squamous Cell Carcinoma, 2023.

[12] Bertrand Pouymayou, Panagiotis Balermpas, Oliver Riesterer, Matthias Guckenberger, and Jan Unkelbach. A bayesian network model of lymphatic tumor progression for personalized elective ctv definition in head and neck cancers. *Physics in Medicine & Biology*, 64(16):165003, 2019.

[13] Roman Ludwig, Jean-Marc Hoffmann, Bertrand Pouymayou, Martina Broglie Däppen, Grégoire Morand, Matthias Guckenberger, Vincent Grégoire, Panagiotis Balermpas, and Jan Unkelbach. Detailed patient-individual reporting of lymph node involvement in oropharyngeal squamous cell carcinoma with an online interface. *Radiotherapy and Oncology*, 169:1–7, 2022.

[14] Roman Ludwig, Jean-Marc Hoffmann, Bertrand Pouymayou, Grégoire Morand, Martina Broglie Däppen, Matthias Guckenberger, Vincent Grégoire, Panagiotis Balermpas, and Jan Unkelbach. A dataset on patient-individual lymph node involvement in oropharyngeal squamous cell carcinoma. *Data in Brief*, 43:108345, 2022.

[15] Panayiotis A Kyzas, Evangelos Evangelou, Despina Denaxa-Kyza, and John PA Ioannidis. 18f-fluorodeoxyglucose positron emission tomography to evaluate cervical node metastases in patients with head and neck squamous cell carcinoma: a meta-analysis. *Journal of the National Cancer Institute*, 100(10):712–720, 2008.

[16] RBJ De Bondt, PJ Nelemans, PAM Hofman, JW Casselman, B Kremer, JMA van Engelshoven, and RGH Beets-Tan. Detection of lymph node metastases in head and neck cancer: a meta-analysis comparing us, usgfnac, ct and mr imaging. *European journal of radiology*, 64(2):266–272, 2007.

[17] C Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:35–42, 2006.

[18] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. <tt>emcee</tt>: The mcmc hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306–312, March 2013.

[19] OpenAI. ChatGPT-4: Optimizing Language Models for Dialogue. https://openai.com/, 2023. Accessed: 2024-01-28.

[20] rmnldwg. lyscripts. https://github.com/rmnldwg/lyscripts, 2023. Scripts that are used in the pipelines of the lynference repository.

[21] Ludwig, Roman and Pouymayou, Bertrand and Perez Haas, Yoel and Balermpas, Panagiotis and Unkelbach, Jan. lymph-model. https://github.com/rmnldwg/lymph, 2023. Accessed: 2024-01-28.

List of Figures