




OPEN

## Modelling the lymphatic metastatic progression pathways of OPSCC from multi-institutional datasets

Roman Ludwig<sup>1,2</sup>, Adrian Daniel Schubert<sup>4,5,7</sup>, Dorothea Barbatei<sup>8</sup>, Laence Bauwens<sup>8</sup>, Jean-Marc Hoffmann<sup>1</sup>, Sandrine Werlen<sup>4,5</sup>, Olgun Elicin<sup>3</sup>, Matthias Dettmer<sup>6,10</sup>, Philippe Zrounba<sup>9</sup>, Bertrand Pouymayou<sup>1</sup>, Panagiotis Balermpas<sup>1</sup>, Vincent Grégoire<sup>8</sup>, Roland Giger<sup>4,5</sup> & Jan Unkelbach<sup>1,2</sup>

The elective clinical target volume (CTV-N) in oropharyngeal squamous cell carcinoma (OPSCC) is currently based mostly on the prevalence of lymph node metastases in different lymph node levels (LNLs) for a given primary tumor location. We present a probabilistic model for ipsilateral lymphatic spread that can quantify the microscopic nodal involvement risk based on an individual patient's T-category and clinical involvement of LNLs at diagnosis. We extend a previously published hidden Markov model (HMM), which models the LNLs (I, II, III, IV, V, and VII) as hidden binary random variables (RVs). Each represents a patient's true state of lymphatic involvement. Clinical involvement at diagnosis represents the observed binary RVs linked to the true state via sensitivity and specificity. The primary tumor and the hidden RVs are connected in a graph. Each edge represents the conditional probability of metastatic spread per abstract time-step, given disease at the edge's starting node. To learn these probabilities, we draw Markov chain Monte Carlo samples from the likelihood of a dataset (686 OPSCC patients) from three institutions. We compute the model evidence using thermodynamic integration for different graphs to determine which describes the data best. The graph maximizing the model evidence connects the tumor to each LNL and the LNLs I through V in order. It predicts the risk of occult disease in level IV is below 5% if level III is clinically negative, and that the risk of occult disease in level V is below 5% except for advanced T-category (T3 and T4) patients with clinical involvement of levels II, III, and IV. The provided statistical model of nodal involvement in OPSCC patients trained on multi-institutional data may guide the design of clinical trials on volume-deescalated treatment of OPSCC and contribute to more personal guidelines on elective nodal treatment.

When treating head and neck squamous cell carcinoma (HNSCC) with radiotherapy or surgery, the aim is to irradiate or resect as much of the malignant tissue as possible. This includes the primary tumor mass and clinically detected lymph node metastases. However, to reduce the risk of locoregional failure, treatment also includes regions of the lymph drainage system of the neck with possible microscopic tumor spread, which in-vivo imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), or positron emission tomography (PET) cannot detect. This is referred to as elective nodal irradiation or prophylactic neck dissection. Treatment decisions regarding the CTV-N or the extent of neck dissection must balance the conflicting goals of treating regions at risk of occult lymph node metastases to avoid recurrences while avoiding toxicity related to unnecessary treatment of healthy tissues.

<sup>1</sup>Dep. of Radiation Oncology, University Hospital Zurich, Rämistrasse 100, 8091 Zurich, Switzerland. <sup>2</sup>Dep. of Physics, University of Zurich, Rämistrasse 71, 8006 Zurich, Switzerland. <sup>3</sup>Dep. of Radiation Oncology, Bern University Hospital, University of Bern, Freiburgstrasse 18, 3010 Bern, Switzerland. <sup>4</sup>Dep. of ENT, Head & Neck Surgery, Inselspital, Bern University Hospital, University of Bern, Freiburgstrasse 18, 3010 Bern, Switzerland. <sup>5</sup>Head and Neck Anticancer Center, Bern University Hospital, University of Bern, Freiburgstrasse 18, 3010 Bern, Switzerland. <sup>6</sup>Institute of Tissue Medicine and Pathology, Bern University Hospital, University of Bern, Murtenstrasse 31, 3008 Bern, Switzerland. <sup>7</sup>Dep. of ENT, Head & Neck Surgery, Réseau Hospitalier Neuchâtelois (RHNe), Neuchâtel, Switzerland. <sup>8</sup>Dep. of Radiation Oncology, Centre Léon Bérard, 28 Rue Laennec, 69008 Lyon, France. <sup>9</sup>Dep. of Head and Neck surgery, Centre Léon Bérard, 28 Rue Laennec, 69008 Lyon, France. <sup>10</sup>Institute of Pathology, Klinikum Stuttgart, Kriegsbergstr. 60c, 70174 Stuttgart, Germany. ✉email: roman.ludwig@usz.ch

This work concerns itself with OPSCC, where approximately 70-80% of patients present with lymph node metastases at the time of diagnosis. In clinical practice, CTV-N definition in radiotherapy is based on guidelines<sup>1-8</sup> that are mostly derived from the observed prevalence of involvement in an LNL for a given tumor location. These guidelines currently suggest extensive irradiation of both sides of the neck for most patients. In the ipsilateral neck, the CTV-N includes LNLs II, III and IV for all patients, and levels I and V for the majority of patients. These guidelines, however, do not account for the personal risk of the patients that may depend greatly on their state of tumor progression at diagnosis. E.g., a patient with macroscopic metastases detected via PET in both LNLs II and III may have a substantial risk for occult disease in LNL IV. Instead, patients who present with a clinically N0 neck or a single metastasis in LNL II may have a much smaller risk for occult disease in LNL IV.

We previously developed a model of lymphatic metastatic progression for estimating the risk of microscopic disease, given a patient's personal diagnosis. The initial model was based on the methodology of Bayesian networks (BNs)<sup>9</sup>. It was subsequently extended and formulated as an HMM to include T-category in an intuitive manner<sup>10</sup>. However, these models were introduced based only on a small dataset of approximately 100 early T-category patients available at that time<sup>11</sup>. The limited data did not allow us to quantify the probability of metastases in the rarely involved LNLs I, V, and VII, nor did the data allow us to verify that the HMM is adequate to describe the dependence on lymph node involvement on T-category. In this paper, we extend the previous work<sup>10</sup> by making the following contributions:

1. We provide an HMM of ipsilateral lymph node involvement including all relevant LNLs, namely the levels I, II, III, IV, V, and VII. To determine the optimal underlying DAG we compare different graphs by calculating the model evidence through TI.
2. We collect a multi-centric dataset consisting of 686 patients from three institutions, allowing us to train the model based on a sizable dataset<sup>12,13</sup>.
3. We use the trained model to provide personalized risk estimations for occult metastases for typical clinical states of tumor progression at diagnosis, illustrating its potential for guiding volume-deescalated treatment strategies in the future.

## HMM formalism and notation

### State of the hidden Markov model

We have introduced a probabilistic model for lymph node involvement based on Bayesian networks (BNs) in<sup>9</sup>. The model was extended using HMMs in<sup>10</sup>. We will briefly recap the hidden Markov model to introduce the notation used throughout the work.

A patient's state of (hidden) lymphatic involvement at time  $t$  is described as a collection of binary RVs, one for each of the  $V$  LNLs:

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

Each of the LNLs can be in the state  $X_v = 0$  (FALSE), meaning LNL  $v$  is healthy, or in the state  $X_v = 1$  (TRUE), indicating the LNL harbors metastases. The involved state includes occult disease.

The transition from one time-step to another is governed by the transition probability  $P(\mathbf{X}[t+1] = \xi_i | \mathbf{X}[t] = \xi_j)$ , which can conveniently be collected into a transition matrix when we enumerate all  $2^V$  distinct possible states  $\xi_i$  with  $i \in \{1, 2, \dots, 2^V\}$  of lymphatic involvement:

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

The term  $P(\xi_i | \xi_j)$  describes the probability to transition from the hidden state of lymphatic involvement  $\xi_j$  to the state  $\xi_i$  between the time  $t$  and  $t+1$ . Using a DAG as depicted in fig. 1, we can formulate this transition probability in the following way:

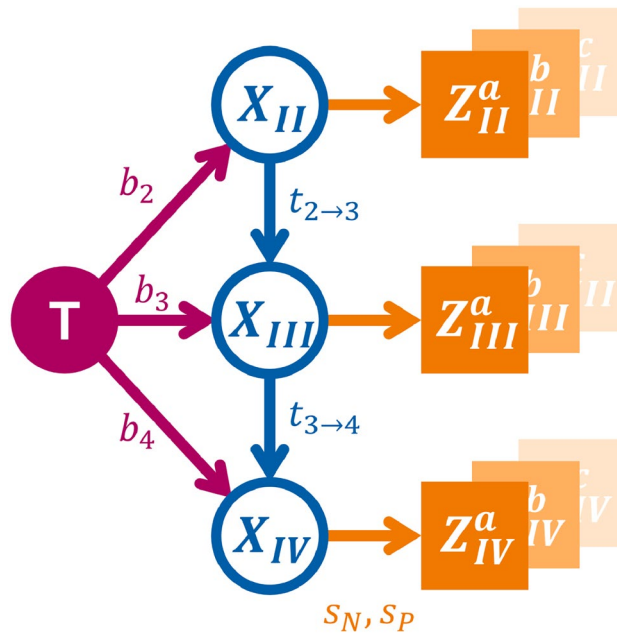
$$P(\xi_i | \xi_j) = \prod_{v \leq V} Q(\xi_{iv}; \xi_{jv}) P(\xi_{iv} | \{\xi_{jr}\}_{r \in \text{pa}(v)})^{1-\xi_{jv}} \quad (3)$$

In this equation, we have denoted LNLs that are parents of LNL  $v$  with the symbol  $r \in \text{pa}(v)$ . Also,  $\xi_{iv}$  denotes the value that LNL  $v$  takes on when the patient is in state  $\xi_i$ . The term  $Q(a; b) \in \{0, 1\}$  is there to prohibit self-healing. It is always one, except if LNL  $v$  is healthy in state  $\xi_i$ , but was metastatic in the previous state  $\xi_j$ . In that case the function becomes  $Q(0; 1) = 0$ , making the transition back to healthier states impossible.

The terms of the form  $P(\xi_{iv} | \{\xi_{jr}\}_{r \in \text{pa}(v)})$  implicitly depend on how we parameterize the arcs of fig. 1. For example, if we look at the probability of spread to LNL III ( $X_3$ ) depending on the state of that level's parent – which, in this case, is  $\text{pa}(3) = 2$  – we can write the different combinations into a conditional probability table as below.

		$X_2 = 0$	$X_2 = 1$
$X_3$	$= 0$	$1 - b_3$	$(1 - b_3)(1 - t_{23})$
	$= 1$	$b_3$	$1 - b_3 - t_{23} + b_3 t_{23}$

The variable  $b_3$  denotes the probability of lymphatic spread from the tumor to LNL III during one time-step, and  $t_{23}$  is the probability of spread from an involved level II further down the lymphatic chain into LNL III.



**Figure 1.** DAG representing a possible abstraction of the lymphatic network comprising the tumor (red shaded circle) and LNLs II through IV as hidden binary RVs (blue outlined circles). Attached to each of these is the corresponding observed RV (orange shaded squares). Lymphatic flow is depicted in the form of parameterized arrows (red and blue) that represent the probability of spread along the respective arc per time-step. Sensitivity and specificity (orange arrows) connect the hidden RVs to the diagnosis.

### Diagnostic observation

We also need to introduce a separate collection of RVs that describe the diagnostic observation of a patient’s involvement. In analogy to the hidden true state  $\mathbf{X}[t]$  at time  $t$ , we write this diagnosis as

$$\mathbf{Z} = (Z_v) \quad v \in \{1, 2, \dots, V\} \tag{4}$$

We do not need to differentiate between different times  $t$  here, since a patient is ever only diagnosed once, after which treatment usually starts timely. Diagnosis and true state of a patient are formally connected via the sensitivity  $s_N$  and specificity  $s_P$  of the used diagnostic modality. In clinical practice, these modalities are CT, MRI, or PET scan, but it may also include information from biopsies after a fine needle aspiration fine needle aspiration (FNA) or other techniques to detect lymphatic metastases. For each LNL  $v$  the conditional probability table of  $P(Z_v | X_v)$  is given by:

		$X = 0$	$X = 1$
$Z$	$= 0$	$s_P$	$1 - s_N$
	$= 1$	$1 - s_P$	$s_N$

Consequently, the conditional probability to observe a diagnosis  $\mathbf{Z} = \zeta_\ell$ , given a hidden involvement state  $\mathbf{X} = \xi_k$  is a matrix  $\mathbf{B}$  made up of products of terms from the table above:

$$\mathbf{B} = (B_{k\ell}) = \prod_{v=1}^V P(Z_v = \zeta_{\ell v} | X_v[t_D] = \xi_{kv}) \tag{5}$$

We define the time  $t = 0$  to be the moment a patient’s tumor formed, and hence  $X_v[t = 0] = 0 \quad \forall v$ . However, using this definition, we cannot know how many time-steps have passed until  $t_D$ , when the patient was diagnosed with cancer. We can only make the assumption that a patient with an earlier T-category tumor was *probably* diagnosed after fewer time-steps than a patient with an advanced T-category tumor.

We can use this assumption by marginalizing over the diagnose times  $t_D$  of patients in different T-categories using different prior distributions over the diagnose time. E.g.,  $P(t = t_D | \text{early})$  for early T-category patients (T1 & T2) and  $P(t = t_D | \text{late})$  for advanced T-category patients (T3 & T4). Throughout this work we will use binomial distributions for these probability mass functions.

$$P(t = t_D | Tx) = \mathfrak{B}(t_{\max}, p_{Tx}) \tag{6}$$

Here, the parameter  $p_{Tx}$  can be interpreted as the probability that the patient with a tumor of T-category  $x$  will be diagnosed at time-step  $t + 1$  given they are in time-step  $t$ . We will use as the latest time-step  $t_{\max} = 10$ . Binomial distributions depend only on a single parameter, which when multiplied with the number of considered

time-steps, conveniently also represents the distribution's mean (in our previous publication on the lymphatic progression model, we have shown that the shape of the time-prior and number of time-steps have no impact as long as only one T-category is considered<sup>10</sup>).  $t_{\max} = 10$  will therefore give us a distribution over the diagnosis time that has its mean at  $\mathbb{E}[t_D] = 10 \cdot p_{Tx}$ .

### The likelihood function

Using the definitions up to this point, we can compute a vector of likelihoods for every possible diagnosis:

$$\begin{aligned} \ell &= (P(\mathbf{Z} = \zeta_i)) \\ &= \sum_{t=0}^{t_{\max}} [\boldsymbol{\pi} \cdot \mathbf{A}^t \cdot \mathbf{B}] \cdot P(t | T) \end{aligned} \quad (7)$$

This likelihood implicitly depends on how we parameterize the arcs of the DAG underlying the model – see Eq. 3 – and the parameterization of the distribution over diagnosis times – e.g., as in Eq. 6.

Together with the parametrizations of the distributions over the diagnosis time, the parameters  $b_v$ , and  $t_{vr}$  that make up the transition matrix  $\mathbf{A}$  comprises the set of model parameters:

$$\boldsymbol{\theta} = (\{b_v\}, \{t_{vr}\}, p_{\text{early}}, p_{\text{late}}) \quad \text{with} \quad \begin{matrix} v \leq V \\ r \in \text{pa}(v) \end{matrix} \quad (8)$$

To infer these parameters from a dataset of  $N$  OPSCC patients  $\mathcal{D} = (d_1, d_2, \dots, d_N)$ , we compute the data log-likelihood:

$$\log \mathcal{L}(\mathcal{D} | \boldsymbol{\theta}) = \sum_{i=1}^N \log P(\mathbf{Z} = d_i) \quad (9)$$

Which effectively amounts to computing the element-wise logarithm of the likelihood vector  $\ell$  from Eq. 7 and summing up the entries that correspond to each of the patients  $d_i$  for  $i \leq N$ .

Note that it is also possible to account for incomplete diagnoses, i.e., a diagnosis where the involvement information for one or more LNLs is missing. In that case, we can sum over those elements of  $\ell$  that correspond to complete diagnoses which match the provided incomplete one. In this paper, for some patients involvement information of level VII was missing and hence marginalized over. A detailed explanation of this formalism can be found in<sup>14</sup>, section 6.2.7.

Using this log-likelihood function one may now employ a variety of inference methods to learn the parameters of the model that best describe the observed data.

### Parameter inference

We use Markov chain Monte Carlo (MCMC) sampling to draw parameter samples  $\hat{\boldsymbol{\theta}}_i$  for  $i \leq S$  from the likelihood described in Eq. 9 (i.e. the unnormalized posterior distribution over the parameters  $\boldsymbol{\theta}$ , since we used a uniform prior in this work).

More specifically, we use the Python implementation `emcee`<sup>15</sup> and two sample proposal mechanisms based on differential evolution moves<sup>16,17</sup> for sampling. Instead of proposing and then accepting or rejecting individual parameter samples one after the other (as in the classical Metropolis-Hastings algorithm), the `emcee` implementation makes use of an ensemble of  $W$  so-called “walkers”. This gives rise to  $W$  parallel chains of samples that mutually influence each others proposal such that the sampling procedure overall is *affine invariant*. This means that scaling the parameter space along any dimension has no effect on the performance of the MCMC sampling algorithm.

For the experiments in this work, we used  $W = 20 \cdot k$  walkers, where  $k$  is the dimensionality of the parameter space  $\Theta$ . After an initial “burn-in” phase, during which all drawn samples are discarded because they are not yet independent of the initial state, we continued sampling for another 200 steps of which we discarded every 10th to be left with  $S = 20 \cdot W$  samples.

These  $S$  parameter estimates are then used to compute expectation values of estimates that depend on the parameters  $\boldsymbol{\theta}$  through an integral over the parameter space  $\Theta$ :

$$\begin{aligned} \mathbb{E}_p[f] &= \int_{\Theta} p(\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\approx \frac{1}{S} \sum_{i=1}^S f(\hat{\boldsymbol{\theta}}_i) \end{aligned} \quad (10)$$

Alternatively, the individual  $\hat{f}_i = f(\hat{\boldsymbol{\theta}}_i)$  can be used to plot histograms over the distribution of  $f$ . We will do so in section 5 to show distributions over prevalence predictions and risk computations.

Another relevant model parameter that needs to be set for the inference process, is the maximum number of time-steps we used for the evolution of the system. We set this value to  $t_{\max} = 10$ , such that  $t \in \{0, 1, 2, \dots, 10\}$ . The binomial “success probability” used to fix the shape of the early T-category’s time-prior was set to  $p_{\text{early}} = 0.3$ .

### Risk estimation

The main task for personalizing the CTV-N definition is to predict the probability of the hidden possible states  $\xi_k$  given the diagnosis  $d^* = \zeta_\ell$  of a new patient at the time of diagnosis. Using Bayes’ theorem, we get

$$P(\mathbf{X} = \xi_k | \mathbf{Z} = \zeta_\ell) = \frac{P(\zeta_\ell | \xi_k)P(\xi_k | \theta)}{\sum_{r=1}^{2^V} P(\zeta_\ell | \xi_r)P(\xi_r | \theta)} \quad (11)$$

The described model along with the inferred parameters  $\hat{\theta}$  will yield an estimate (or multiple estimates) for the “prior” in the above equation  $P(\xi_k | \hat{\theta})$ .

From this probability for any possible hidden state, we can also compute the probability of, for example, involvement in LNL IV. To that end, we marginalize over all states  $\xi_k$  where  $\xi_{k4} = 1$ , meaning those states in which LNL IV harbors metastases. Formally, we can define a marginalization vector  $\mathbf{m}$  that is one for every hidden state we want to include in the marginalization and zero elsewhere. In the example of the marginalized probability for LNL IV involvement, the components would look like this:

$$m_{4k} = \text{id}(\xi_{k4} = 1) \quad (12)$$

Subsequently, we can compute the marginalization as a dot product:

$$\begin{aligned} P(\text{IV} = 1 | \mathbf{Z} = \zeta_\ell) &= \sum_{k:\xi_{k4}=1} P(\mathbf{X} = \xi_k | \mathbf{Z} = \zeta_\ell) \\ &= \mathbf{m}_4 \cdot P(\mathbf{X} = \xi_k | \mathbf{Z} = \zeta_\ell) \end{aligned} \quad (13)$$

## Complete model of ipsilateral spread in OPSCC Investigating spread graphs

The DAG shown in fig. 1 includes the LNLs II, III, and IV, which represent the most relevant lymph node levels for OPSCC. It includes the arcs from II to III and from III to IV, representing the main direction of lymphatic drainage, which is well motivated anatomically and by the data on lymph node involvement. Previous publications<sup>9,10</sup> focused on these levels because they relied on a limited reconstructed dataset of OPSCC patients<sup>11</sup>. Now, with the datasets available for this work, we can extend the graph to include random variables for all LNLs that are relevant for OPSCC: I, II, III, IV, V, and VII. The main question to answer is: which arcs between LNLs are needed to accurately model the data on lymph node involvement without increasing the model's complexity unnecessarily.

First, we notice that the direct arcs from tumor to each of the LNLs must be present, since every LNL appears metastatic in isolation at least once in the dataset. For example, some patients presented with metastases in LNL I, while the other levels appeared healthy. If no spread was allowed from the tumor to  $X_1$  (i.e.,  $b_1 = 0$ ), the likelihood of observing this patient would be zero.

We have more freedom in choosing how to connect the LNLs to each other. To investigate which connections to add we start by establishing a baseline from a model using a minimal *base graph*. It contains only the connections from LNL II to III and an arc from LNL III to IV, as motivated by the *main lymphatic pathway*<sup>18</sup>. The base graph is illustrated in fig. 2 via the red and blue arcs. The lymphatic drainage to or from levels I, V, and VII is not as clearly defined. Therefore, we define a set of candidate arcs (green arcs in fig. 2) and use the model comparison methodology described in section 3.2 to determine which graph is most supported our data.

To connect level I, we investigate two candidate arcs: from I to II and from II to I. An arc from I to II was used in<sup>9,10</sup> and is anatomically motivated. However, since LNL I is rarely involved compared to level II, the associated parameter  $t_{12}$  is mostly undetermined. Therefore, we also consider the flipped arc from LNL II to I and investigate if it helps to describe the correlations between the involvement of levels I and II.

Anatomically, the *posterior accessory pathway* that drains LNL II through LNL V motivates investigating and arc from  $X_2$  into  $X_5$ <sup>18</sup>. And, although no lymphatic pathway is described that directly drains LNL III or IV into level V, due to their proximity to each other, we will also investigate additional edges from the levels III and IV into LNL V. Finally, we look at adding an arc from LNL II to LNL VII also due to their anatomical proximity.

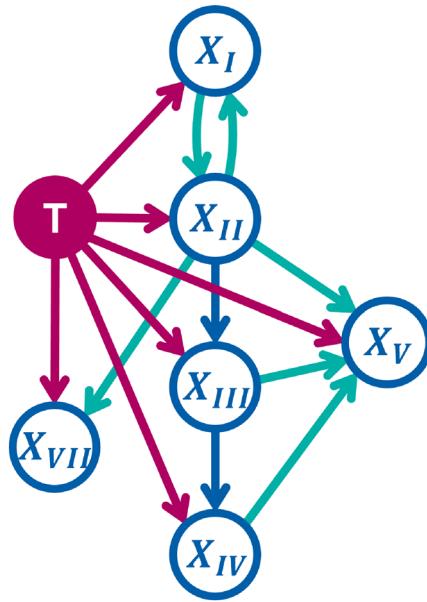
To determine the optimal graph, we first consider six models, each with one of the six green candidate arcs of Fig. 2 added to the base graph. Every one of these models was evaluated by computing an approximation to its evidence via thermodynamic integration as described below in the “Model comparison” sect.. Subsequently, graphs combining multiple arcs that individually improve the model evidence are considered. Thereby, the “winning graph” is determined, which yields the highest (i.e., the least negative) value of the logarithm of the model evidence.

### Model comparison

The aim of this work is to refine the graph structure underlying our risk model introduced in the previous section. This DAG determines the number of parameters of the model as well as how exactly the transition matrix  $\mathbf{A}$  is parameterized. To compare different models that are based on different DAGs, e.g. models  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , in a Bayesian setting, we need to compute the probabilities of these models, given the data  $\mathcal{D}$ :

$$P(\mathcal{M}_i | \mathcal{D}) = \frac{P(\mathcal{D} | \mathcal{M}_i)P(\mathcal{M}_i)}{P(\mathcal{D})} \quad (14)$$

If we assume all models  $\mathcal{M}_i$  for  $i \in \{1, 2\}$  to have the same *a priori* probability – meaning in this case  $P(\mathcal{M}_1) = P(\mathcal{M}_2)$  – then we can compute the so-called *Bayes factor* of the two models as the ratio of their likelihoods. The interpretation of the values for different Bayes factors is given in table 1. It is defined as follows:



**Figure 2.** Extended DAG representing different possible spread graphs underlying the HMM. As in fig. 1, red arcs are parametrized with probabilities of spread from the tumor (red circle) to the LNLs (blue circles). These red arcs, together with the blue arcs from LNL to LNL, make up the *base graph*. One after the other, each of the green arcs was added to the base graph. Subsequently, the performance of the resulting models in terms of its Bayesian information criterion (BIC) was compared to the base graph to assess whether the additional edge should be kept in the *winning graph* or not.

$$K_{1v2} = \frac{P(\mathcal{M}_1 | \mathcal{D})}{P(\mathcal{M}_2 | \mathcal{D})} = \frac{P(\mathcal{D} | \mathcal{M}_1)}{P(\mathcal{D} | \mathcal{M}_2)} \tag{15}$$

These likelihoods are commonly called the *model evidence* or *marginal likelihood*. The latter because computing it involves marginalizing the data likelihood over all model parameters:

$$E_{\mathcal{M}} = P(\mathcal{D} | \mathcal{M}) = \int_{\Theta} P(\mathcal{D} | \theta, \mathcal{M})p(\theta | \mathcal{M})d\theta \tag{16}$$

However, this quantity is often very hard to compute or even intractable, due to the high dimensionality of the parameter space  $\Theta$ . In our case, the number of dimensions ranges from  $k = 9$  for the *base graph* to  $k = 11$  for the *winning graph*. A brute-force integration over a unit cube with this many dimensions is inefficient and error-prone, which is why we resorted to TI for computing the (log-)evidence.

Below, we will briefly outline the main concept behind this algorithm. An intuitive and extensive derivation of TI is given by<sup>20</sup>.

We start by taking the logarithm of the model evidence  $E$  and subtract a zero from it in the form of the term  $0 = \ln \int p(\theta | \mathcal{M})d\theta$ . Further, we can multiply the distribution over the parameters  $\theta$  inside this integral by  $1 = P(\mathcal{D} | \theta, \mathcal{M})^{\beta=0}$ . Subsequently, we can write the logarithm of the evidence as an integral over a derivative:

$$\begin{aligned} \ln E &= \ln \int P(\mathcal{D} | \theta, \mathcal{M})^{\beta=1} p(\theta | \mathcal{M})d\theta - \ln E_0 \\ &= \int_0^1 \frac{d}{d\beta} \ln E_{\beta} d\beta \end{aligned} \tag{17}$$

Where we have used the (unnormalized) *power posterior*  $p_{\beta}(\theta | \mathcal{D}, \mathcal{M}) = P(\mathcal{D} | \theta, \mathcal{M})^{\beta} p(\theta | \mathcal{M})$  to compute the respective evidence  $E_{\beta} = \int p_{\beta}(\theta | \mathcal{D}, \mathcal{M})d\theta$ .

The derivatives of the log-evidences  $\ln E_{\beta}$  are essentially expectation values of the data log-likelihood under the power posteriors of the corresponding value for  $\beta$ . They can be computed using MCMC:

$$\begin{aligned} \frac{d}{d\beta} \ln E_{\beta} &= \int p_{\beta}(\theta | \mathcal{D}, \mathcal{M}) \ln P(\mathcal{D} | \theta, \mathcal{M})d\theta \\ &= \mathbb{E}[\ln P(\mathcal{D} | \theta, \mathcal{M})]_{p_{\beta}(\theta | \mathcal{D}, \mathcal{M})} \\ &\approx \frac{1}{S} \sum_{i=1}^S \ln P(\mathcal{D} | \hat{\theta}_{\beta_i}, \mathcal{M}) =: \mathcal{A}_{MC}(\beta) \end{aligned} \tag{18}$$



The integral in Eq. 17 can then be computed via a trapezoidal rule using the  $\mathcal{A}_{MC}$  to yield a numerical approximation of the model evidence:

$$\ln E \approx \frac{1}{2} \sum_{j=0}^{R-1} (\beta_{j+1} - \beta_j) \cdot (\mathcal{A}_{MC}(\beta_{j+1}) + \mathcal{A}_{MC}(\beta_j)) \quad (19)$$

This estimate gets better for more samples  $S$  per sampling from the power posterior  $p_\beta$  but more importantly it gets better for a tighter spacing of the values for  $\beta$  within the interval  $[0, 1]$ . The variable  $\beta$  is also often referred to as an *inverse temperature*, due to its origins in statistical physics. Often when performing TI, the most drastic changes in the values of the  $\mathcal{A}_{MC}$  occur at high temperatures (meaning  $\beta$  very close to zero), while the changes become smaller and smaller for lower temperatures ( $\beta$  towards one). It is therefore efficient to space the *temperature ladder* unevenly, e.g. according to a fifth order power rule:

$$\beta_j = (j/R)^5 \quad j \in \{0, 1, 2, \dots, R\} \quad (20)$$

For the TIs that were performed in this work we used such a fifth order power rule with 64 steps, meaning that  $R = 63$ .

The process of computing the log-evidence using TI was as follows: We randomly initialized the starting positions of the  $W$  samplers in the ensemble within the  $k$  dimensional unit cube  $\Theta$ . Subsequently, for each  $j \in \{0, 1, 2, \dots, R = 63\}$  we drew samples from the corresponding power posterior with the value of  $\beta_j$  set according to the power rule in eq. 20. This sampling at point  $j$  consisted of 1000 burn-in steps, followed by 200 steps, of which only every tenth was kept. The last position of the  $W$  chains for the  $j$ -th  $\beta$  value in the ladder was used to initialize the subsequent sampling round with  $\beta_{j+1}$ . Hence, after the computations are finished, we are left with  $S = 20 \cdot W$  samples  $\hat{\theta}_{i,j}$  and respective log-likelihood  $\hat{\ell}_{i,j}$  from each of the 64 power posteriors corresponding to the respective  $\beta_j$ . Subsequently, we numerically integrated the following quantity  $S$  times:

$$\ln \hat{E}_i = \frac{1}{2} \sum_{j=0}^{R-1} (\beta_{j+1} - \beta_j) \cdot (\hat{\ell}_{i,j} + \hat{\ell}_{i,j+1}) \quad (21)$$

And then computed the mean and standard deviation of all the integrated  $\ln \hat{E}_i$ . We then used this for the log-evidence and its error.

Without derivation or insight, we would like to mention that the model evidence naturally balances a model's accuracy against its complexity. The value of  $\ln E$  will generally be larger (i.e., less negative) if a model fits the data better than another while being similarly complex. On the other hand, if e.g. additional parameters are introduced without sufficiently improving how well the model explains the data, the evidence will penalize the increase in complexity.

An approximation to the evidence that also attempts to balance accuracy and complexity against each other is the heuristic called BIC. The negative one half of the BIC approximates the  $\ln E$  via Lagrange's method<sup>23</sup> and yields an easy to compute estimate that may also be used to compare models, as long as its underlying assumptions are valid:

$$-\text{BIC}/2 = \ln \hat{\mathcal{L}} - \frac{k}{2} \ln N \approx \ln E \quad (22)$$

Here,  $\hat{\mathcal{L}} = \max_{\theta} (\ln P(\mathcal{D} | \theta))$  is the maximum log-likelihood. The approximation is good, when the posterior distribution over the parameters  $p(\theta | \mathcal{D})$  is single-modal and falls quickly to zero from the maximum. Also, the number of data points  $N$  needs to be much larger than the number of parameters  $k$ . We will see that for the models we consider here, the BIC is generally a good approximation and the conclusions drawn from comparing models using this metric can be reproduced reliably using the true model evidence computed with TI.

### Multicentric dataset

The dataset  $\mathcal{D}$  that we used for inference is comprised of the detailed reports on lymph node involvement patterns in OPSCC patients treated at three different institutions in France and Switzerland: The Centre Leon Bérard (CLB) in Lyon (France), the Inselspital Bern (ISB) in Bern (Switzerland), and the University Hospital Zurich (USZ) in Zürich (Switzerland). We have previously published the patterns of nodal involvement for the USZ cohort (287 patients)<sup>12</sup> and described its characteristics in detail<sup>24</sup>. The first CLB dataset (263 patients) underlies a publication on human papilloma virus (HPV) status in OPSCC<sup>25</sup> and is made available in a separate "Data in Brief" article alongside the second dataset from France and the lymphatic progression patterns from the ISB<sup>13</sup>. All datasets may be explored online in our web-based interface [LyProX](#).

In total, the dataset contains 686 patients with newly diagnosed OPSCC. It includes patients treated with definitive (chemo)radiotherapy, adjuvant (chemo)radiotherapy following neck dissection, or neck dissection alone. Pathologically assessed LNL involvement was available for 263 surgically treated patients, while for the remainder the nodal involvement was assessed based on available diagnostic modalities (FDG-PET-CT, CT, MRI, FNA). If multiple modalities were used to diagnose a patient's lymph node involvement, the available modalities were combined into a consensus decision. When different modalities were conflicting, the conflicts were resolved by inferring the most likely state (healthy or metastatic) for each LNL separately. To do so, we used literature values for the sensitivity and specificity of the diagnostic modalities<sup>21,22</sup>, which we also tabulated in table 2. Practically, this means that, whenever pathology after neck dissection was available, the pathology result

was taken as the consensus, overruling any other clinical diagnostic modality. If, for example, PET-CT and MRI were available and conflicting, PET-CT was taken as the consensus, overruling MRI.

The dataset containing the consensus decision for the involvement of each level in every patient was then used for model parameter learning. We assumed that it represents an observation of the true hidden state  $\xi_k$ . The frequencies of some of the most important combinations of involved lymph node levels are listed in table 3.

## Results

### Involvement of levels II, III, and IV

For the base graph, we have plotted the predicted prevalence of involvement patterns in the investigated patient cohort for scenarios involving the most commonly metastatic LNLs II, III and IV in fig. 4. It shows – for each pattern of lymphatic involvement – two plots overlaid:

1. The colored histograms over the base graph model's prediction for the prevalence of the respective pattern of involvement. These histograms are obtained by computing the same prevalence with different samples from the inference process, thus providing us with a measure of uncertainty for the prediction.
2. Colored lines, depicting the beta posterior over the same involvement pattern's prevalence, given a uniform beta prior and the binomial likelihood of the observed data. The maximum of the beta distributions always coincides with the data prevalence but we additionally gain an intuition into how statistically significant the data is. E.g., Observing 3 out of 10 patients with a particular pattern of nodal metastases is less convincing than 300 out of a cohort of 1000 patients. A beta posterior over these prevalences reflects that in its variance.

fig. 4 shows that this minimal graph is already capable of describing the most important parts of the observed data very well. Notably, the model is not only accurate in its predictions, it also correctly estimates the variance stemming from the limited amount of data. The separation between involvement prevalences of early and advanced T-category tumors is also reproduced well by the model. This is remarkable because the model introduces only a single parameter to describe the differences between early and advanced T-category for all involvement patterns. This shows that expecting later diagnosis times, on average, for patients with advanced T-category tumors can explain more severe lymphatic involvement.

It is interesting to note that not all involvement patterns become more prevalent with advanced T-category. For example, a healthy LNL III together with a metastatic level II is observed slightly less often for advanced T-category tumors compared to early T-category (yellow histogram, row 1 versus 2). This is because, for advanced T-category, it is more likely the disease has already spread to LNL III (blue histogram, row 1 versus 2). Our model captures this accurately and precisely.

### Comparison of candidate graphs

The model evidences of all candidate graphs are reported in table 4. For the graph with the highest and lowest model evidence, the expected log-accuracy is plotted against the inverse temperature in fig. 3. A visual ranking is provided in fig. 5. Let us first consider the six models in which one of the candidate arcs is added to the base graph. Evidently, adding a connection from LNL I to II is strongly supported given this dataset, and is slightly superior to the reverse connection from LNL II to I. In addition, there is strong evidence for introducing an arc from LNL IV to V. Furthermore, there is substantial evidence for an arc from LNL III to V. All other investigated additions lead to improvements that are barely worth a mention or do not justify the additional complexity at all, indicated by a lower evidence.

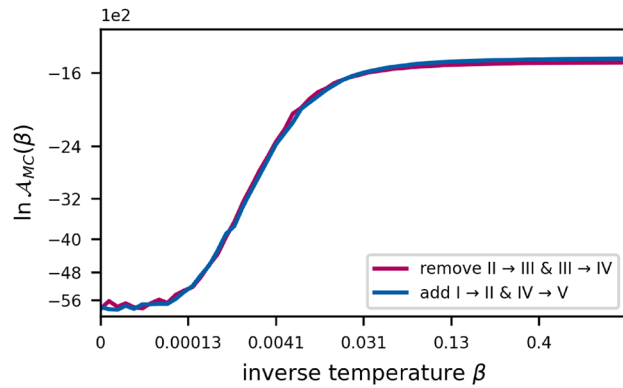
Based on these results, we consider three additional candidates for the optimal graph that combine the added arc from LNL I to II with the arc(s) III  $\rightarrow$  V and/or IV  $\rightarrow$  V. The model evidence for these three graphs is also reported in table 4. The best performing graph with decisive evidence over the base graph turned out to be the one which combines the arcs from LNL I to II and IV to V. Interestingly, the evidence gain of this “winning graph” is roughly the sum of the gains seen in the two candidates where only one of these connections was added, respectively. This indicates that the two additional parameters are largely independent of each other and manage to describe different aspects of the data.

table 4 additionally shows the evidence of graphs in which the arc from level II to III or from level III to IV is removed. The low model evidence for these graphs confirms the importance of these connections and is consistent with the anatomical motivation. The connection from level III to IV is crucial for describing the observation that metastases in level IV are extremely rare without simultaneous involvement of the upstream level III.

### The winning graph

The most likely model parameters for the winning graph, corresponding to the mean of the marginals of the sampled posterior distributions, are tabulated in table 5. We have fixed  $p_{\text{early}} = 0.3$  for early T-category tumors (i.e., T0, T1, and T2), and  $t_{\text{max}} = 10$  time steps. The result that  $t_{\text{II} \rightarrow \text{III}}$  and  $t_{\text{III} \rightarrow \text{IV}}$  are relatively large compared to  $b_{\text{III}}$  and  $b_{\text{IV}}$  reflects the observation that skip metastases in levels III and IV without involvement of the upstream level are rare. Since level II's parent node (level I) is rarely involved,  $b_{\text{II}}$  can approximately be related to the prevalence of involvement in level II. The probability for no involvement of level II when the patient is diagnosed after  $t$  time steps is  $(1 - b_{\text{II}})^t$ . The prevalence of level II involvement for advanced T-category patients is thus





**Figure 3.** The expected log-accuracy under the power posterior plotted against the value of the inverse temperature  $\beta$  for the graph with the best (blue) and worst (red) log-evidence. The area under this curve yields the respective graph's model evidence. The scaling of the x-axis is chosen such that the 64 points on the temperature ladder appear evenly spaced. This is to stress how the log-accuracy develops in the range from  $\beta = 0$  to around  $\beta = 0.1$ . Note how the winning graph's advantage begins to show already shortly before  $\beta = 0.031$ .

$K_{1v2}$	$\ln K_{1v2}$	Support for $\mathcal{M}_1$
$< 10^0$	$< 0$	Negative (supports $\mathcal{M}_2$ )
$10^0$ to $10^{\frac{1}{2}}$	0 to 1.15	Barely worth a mention
$10^{\frac{1}{2}}$ to $10^1$	1.15 to 2.3	Substantial
$10^1$ to $10^{\frac{3}{2}}$	2.3 to 3.45	Strong
$10^{\frac{3}{2}}$ to $10^2$	3.45 to 4.6	Very strong
$> 10^2$	$> 4.6$	Decisive

**Table 1.** Interpretation of Bayes factors and their natural logarithms in terms of their support for or against one of the two compared models as introduced by<sup>19</sup>.

Modality	Specificity (%)	Sensitivity (%)
CT	76	81
PET	86	79
MRI	63	81
FNA	98	80
Pathology	$\approx 100$	$\approx 100$

**Table 2.** Literature sensitivity and specificity values that we used to infer the most likely involvement for a patient when multiple diagnostic modalities reported conflicting nodal involvement<sup>21,22</sup>.

$$\text{prev}_{\text{late}}^{\text{II}} = 1 - \sum_{t=0}^{10} (1 - b_{\text{II}})^t \cdot p_{\text{late}}^t (1 - p_{\text{late}})^{(10-t)} \binom{10}{t} \approx 79\% \tag{23}$$

which agrees with the second panel from the top in fig. 4. The large value for the parameter  $t_{\text{I} \rightarrow \text{II}}$  reflects the observation that in almost all patients with level I involvement, level II is also involved. The large uncertainty in  $t_{\text{I} \rightarrow \text{II}}$  is related to the fact that level I involvement is rare compared to level II.

### Involvement of levels I and V

We can observe that the winning graph describes the involvement of levels II, III, and IV equally well as the base graph, a result that is expected and not further shown. We thus focus on the improvements w.r.t. involvement patterns that include the LNLs I and V, that more rarely harbor metastases.

*Level V:* In fig. 6 we compare the base graph's and the winning graph's estimations for prevalences of involvement patterns that include LNL V. The base graph underestimates the probability that level IV and V are

LNL involvement						T-category			
I	II	III	IV	V	VII	early		advanced	
?	●	?	?	?	●	10	(2%)	20	(7%)
?	●	?	?	?	●	4	(0%)	2	(0%)
?	?	●	?	●	?	12	(2%)	17	(6%)
?	?	●	?	●	?	16	(3%)	9	(3%)
?	●	?	?	?	?	305	(72%)	202	(76%)
?	●	●	?	?	?	100	(23%)	87	(33%)
?	●	●	?	?	?	205	(48%)	115	(43%)
?	●	●	?	?	?	18	(4%)	12	(4%)
?	?	●	?	?	?	118	(27%)	99	(37%)
?	?	●	●	?	?	25	(5%)	23	(8%)
?	?	●	●	?	?	93	(21%)	76	(28%)
?	?	●	●	?	?	6	(1%)	8	(3%)
?	?	?	●	?	?	31	(7%)	31	(11%)
?	●	?	●	?	?	29	(6%)	28	(10%)
?	●	?	●	?	?	2	(0%)	3	(1%)
?	?	?	●	●	?	7	(1%)	7	(2%)
?	?	?	●	●	?	21	(4%)	19	(7%)
●	?	?	?	?	?	18	(4%)	39	(14%)
●	●	?	?	?	?	2	(0%)	2	(0%)
●	●	?	?	?	?	16	(3%)	37	(14%)
●	●	?	?	?	?	289	(68%)	165	(62%)
total						423		263	

**Table 3.** Prevalence of involvement patterns in the multi-centric dataset. An involvement pattern is characterized by the state of the six LNLs: A red dot means the LNL was reported to be metastatic, a green dot means it was determined to be healthy and a question mark means that the prevalence was marginalized over the state of this LNL.

simultaneously involved, and overestimates the probability that level V but not IV is involved. By introducing the arc from level IV to V, the winning graph can describe the observation that level V involvement is typically associated with severe involvement of level II-IV. In the dataset, 14 patients out of 62 patients with level IV involvement have metastases in level V (22%). Instead, only 40 patients out of 624 patients without level IV involvement have metastases in level V (6%).

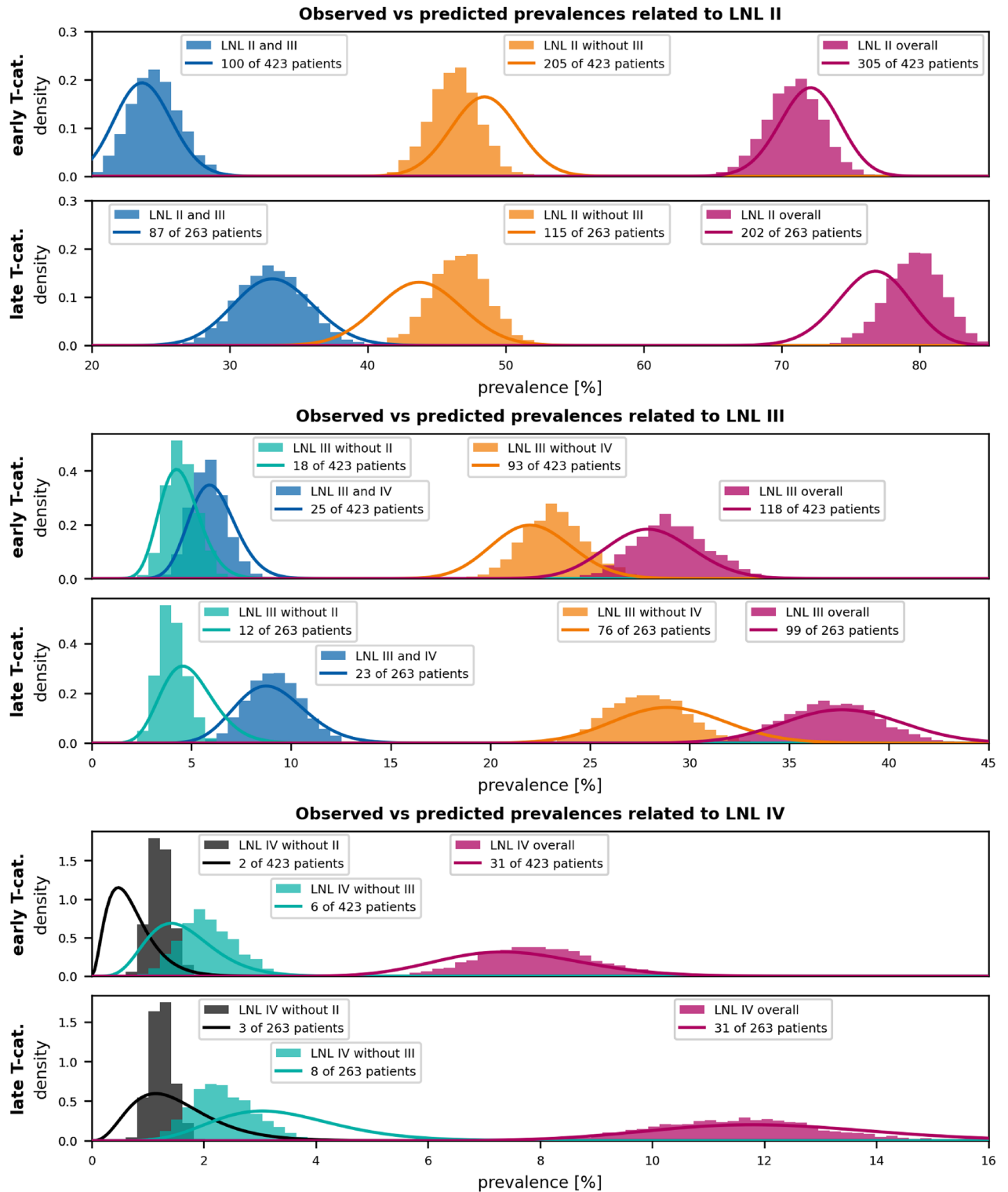
*Level I:* In fig. 7, analogous comparisons are shown for involvement patterns that include LNL I. The base graph overestimates the probability of level I involvement without simultaneous involvement of level II. By introducing the arc from level I to II, the winning graph can capture the correlations between levels I and II. It can also be noted that both models overestimate level I involvement for early T-category patients and underestimate level I involvement for advanced T-category patients. This is further described in the discussion section below.

### Risk prediction for occult disease

In this section, the model corresponding to the winning graph is applied to estimating the risk of occult metastases in clinically negative LNLs. We assume a sensitivity of 0.76 and a specificity of 0.81 for the clinical diagnosis of lymph node metastases, corresponding to CT imaging in table 2.

*Level II:* As can be seen in table 5, spread from the tumor to LNL II to be the most probable transition at any given time step. As a consequence, even for an early T-category patient that presents with a clinical N0 neck, our model predicts a  $31.13\% \pm 1.78\%$  risk for microscopic metastases in LNL II.

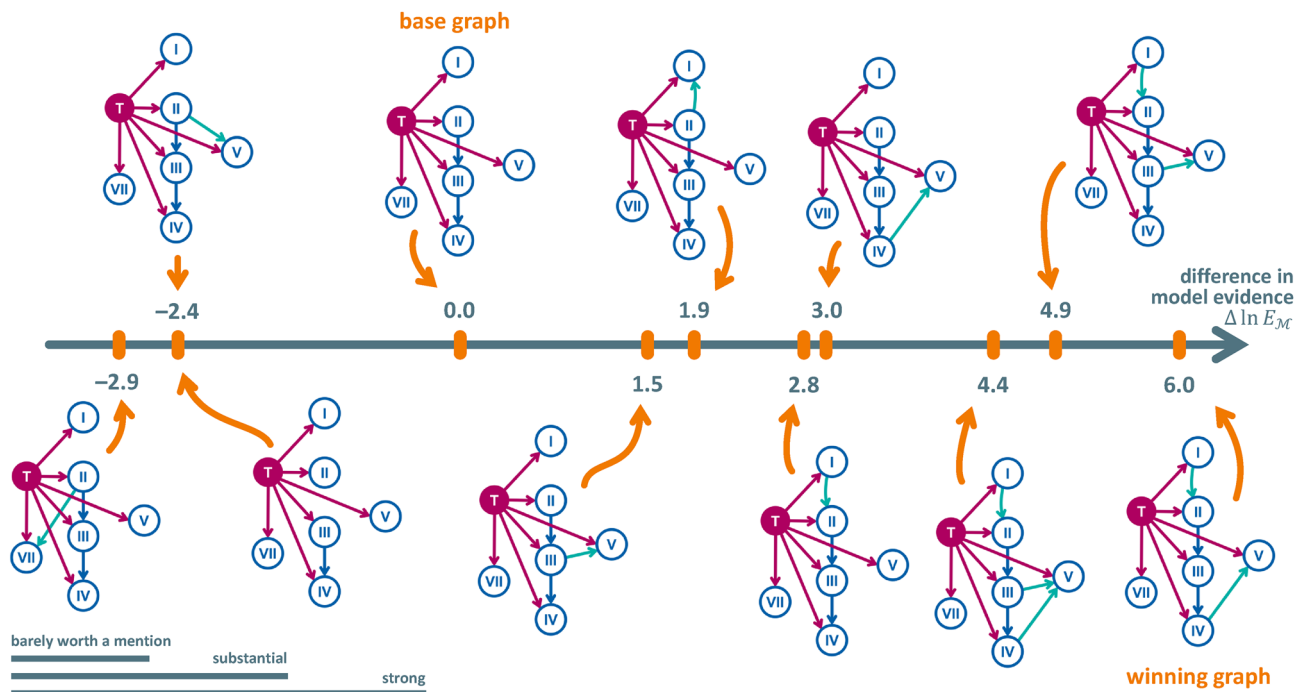
*Level III:* fig. 8 compares the risk of occult disease in level III between patients that are clinically N0 (orange) and patients with clinically diagnosed involvement of only level II (red), for early T-category (upper panel) and advanced T-category (bottom panel). The histograms represent the uncertainty in the model's risk prediction arising from the uncertainty in the model parameters and are generated by randomly drawing a tenth of the samples from the model parameter's joint posterior distribution. This amounts to  $S = 20 \cdot W$  samples, as described



**Figure 4.** Prevalence of involvement as predicted by the base graph model for different scenarios involving the most commonly metastatic LNLs II, III, and IV (shaded histograms). The model's predictions are compared to Beta posteriors over the prevalence based on the frequency of the same scenarios in the data assuming a uniform prior (solid lines). E.g., out 423 early T-category patients, 100 (23.6%) presented with involvement of the LNLs II and III, as shown by the blue line in the uppermost panel. The top panels of each of the three subfigures show some selected scenarios with early T-category tumors and the bottom panel the same scenarios for advanced T-category. Similar scenarios are color-coded: Blue for joint involvement with downstream LNL, orange for a level's involvement without the downstream LNL, red for an LNLs overall involvement prevalence, and green for involvement without direct upstream metastases. The black scenario in the bottom subfigures only appears once and did not fit into these categories. This figure shows that for the most important LNLs II, III, and IV the base graph model already fits the data well.

Graph	$\Delta$ Log-evidence
Add I $\rightarrow$ II & IV $\rightarrow$ V	5.99
Add I $\rightarrow$ II & III $\rightarrow$ V	4.96
Add I $\rightarrow$ II & III $\rightarrow$ V & IV $\rightarrow$ V	4.44
Add IV $\rightarrow$ V	3.05
Add I $\rightarrow$ II	2.86
Add II $\rightarrow$ I	1.95
Add III $\rightarrow$ V	1.56
Base graph	0.0
Remove II $\rightarrow$ III	-2.35
Add II $\rightarrow$ V	-2.35
Add II $\rightarrow$ VII	-2.84
Remove III $\rightarrow$ IV	-20.39
Remove II $\rightarrow$ III & III $\rightarrow$ IV	-22.47

**Table 4.** Model comparison results for all compared graph structures. For all DAGs we show the difference of the log-evidence to the base graph, as in fig. 5, computed via thermodynamic integration.



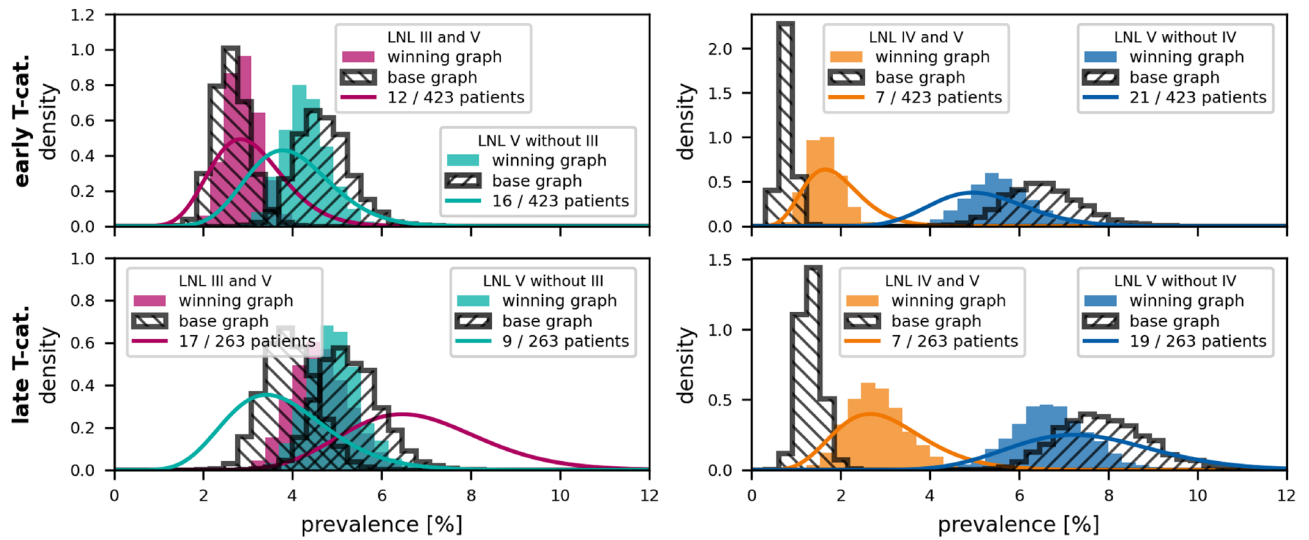
**Figure 5.** Visual ranking of the investigated graphs w.r.t. their model evidence, computed via thermodynamic integration. Not shown are the two graphs where the arc from LNL III to IV was removed. Their respective log model evidence differed from the base graph by more than  $-19$  and the two graphs would appear far left in the figure. In the bottom left corner, we provide a visual reference in analogy to table 1: E.g., any difference in the model evidence shorter than the first of the three rulers indicates that the improvement is “barely worth a mention” Anything longer than it, but shorter than the second ruler from the top indicates a “substantial” improvement.

in the “Model comparison” section. The model predicts a risk of just below 6% for early T-category tumors and 8% for T3 or T4 ones. For patients with involvement of level II, the risk in level III increases to approximately 9% and 12%, respectively.

**Level IV:** fig. 9 (left panels) compares the risk of occult disease in level IV for the typical clinical presentations: clinically N0 (green), metastases in level II (blue), and metastases in levels II and III (orange). The model predicts a low risk of 1–2% in level IV for patients with clinically healthy level III. For patients with clinical involvement of level III, the risk of occult disease in level IV increases to approximately 3% for early T-category and 5% for advanced T-category tumors.

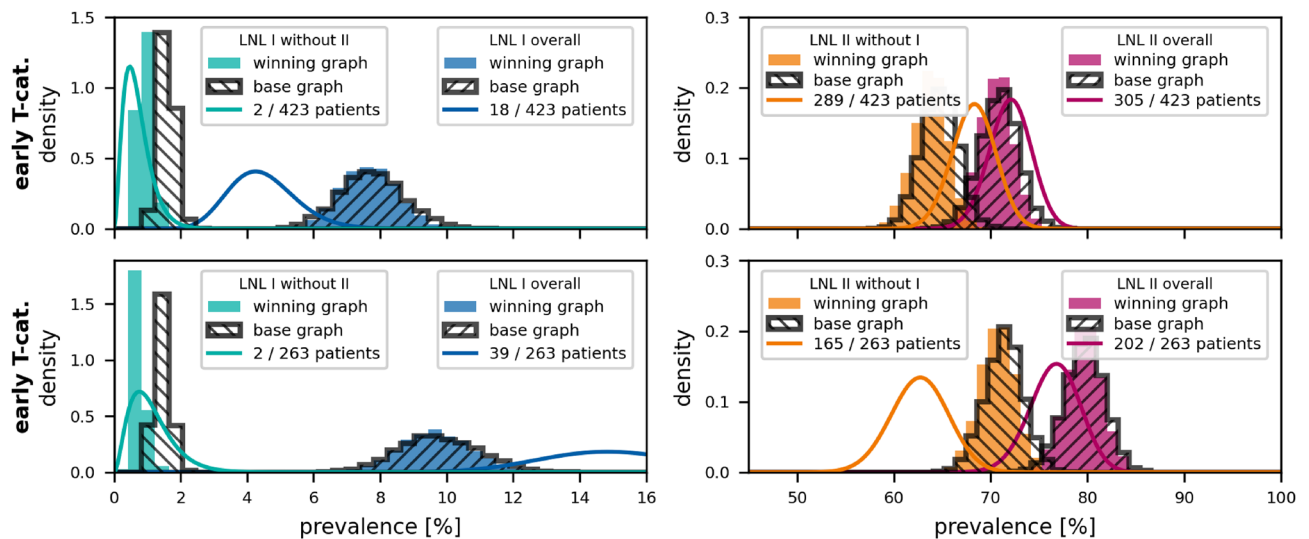
**Level V:** The right panels in fig. 9 show the risk of occult disease in level V depending on T-category and the clinical involvement of levels II–IV. For clinically N0 patients, the risk in level V is estimated to be just above

### Base and winning graph's prevalence predictions



**Figure 6.** Observed (Beta posteriors as lines) vs. predicted (histograms) prevalences of involvement combinations that include LNL V. We have plotted the predictions from the winning graph (colored histograms) and those of the base graph (black, hatched histograms). The top two panels show scenarios for early T-category patients, the bottom two panels for advanced T-category. The left two panels consider combinations of LNL III and V involvement, while the right two panels consider combinations of LNL IV and V. The colored lines show the Beta posterior over the prevalence of the respective involvement pattern, given the data. Especially the right two panels indicate the winning graph model's better fit to the data over the base graph's model.

### Base and winning graph's prevalence predictions



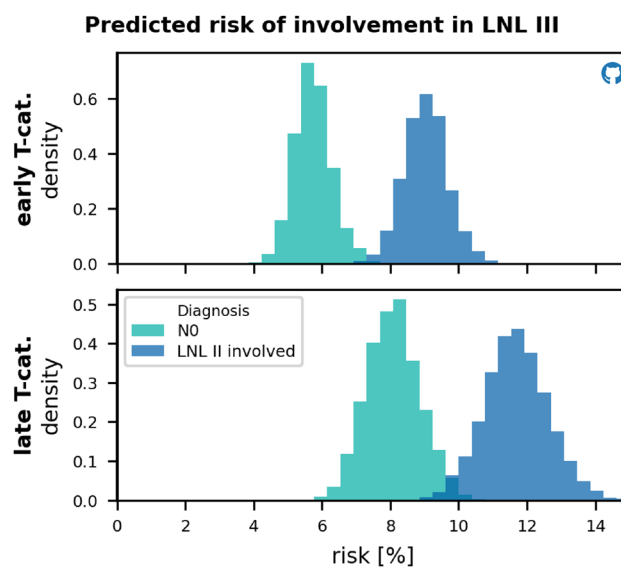
**Figure 7.** Comparison of observed and predicted prevalences of LNL I and II involvement patterns. The top and bottom panels show the prevalences for early and advanced T-category, respectively. The solid lines are Beta posteriors from the data, while the histograms are predicted prevalences (colored: winning graph, black-hatched: base graph). Blue and red plots indicate overall LNL I and II involvement, respectively. Green plots indicate LNL I involvement without level II, while orange plots indicate the opposite (LNL II without level I). The winning graph has an added edge from LNL I to II, which improves the prediction of the rare green pattern. Otherwise, the winnings graph does not meaningfully improve the model's fit to the data.

1%. Extensive nodal involvement of levels II-IV increases the risk in level V to more than 4% for advanced T-category tumors.

*Level I:* fig. 10 shows the risk of occult disease in level I depending on T-category and the clinical involvement of levels II-IV. For clinically N0 patients, the risk in level I is estimated to be in the order of 1-2%. Extensive nodal involvement of levels II-IV increases the risk in level I to just below 4% for advanced T-category tumors. It is pointed out that the winning graph does not contain arcs from levels III or IV to LNL I (and anatomically

Parameter	Mean (%)	Std. dev. (%)
$b_I$	2.65	$\pm 0.31$
$b_{II}$	37.67	$\pm 1.81$
$b_{III}$	8.1	$\pm 1.26$
$b_{IV}$	1.1	$\pm 0.24$
$b_V$	2.13	$\pm 0.28$
$b_{VII}$	2.16	$\pm 0.31$
$t_{I \rightarrow II}$	66.76	$\pm 21.37$
$t_{II \rightarrow III}$	9.49	$\pm 3.04$
$t_{III \rightarrow IV}$	14.48	$\pm 2.43$
$t_{IV \rightarrow V}$	14.57	$\pm 5.29$
$p_{late}$	38.34	$\pm 2.26$

**Table 5.** Mean and standard deviation of parameters sampled for the winning graph in percent.



**Figure 8.** Histograms over the risk for microscopic involvement in LNL III, given that a patient presents as clinically N0 (green), or given that the patient's LNL II shows clinical involvement (blue). The top panel displays these risks for early T-category, the bottom panel for advanced T-category.

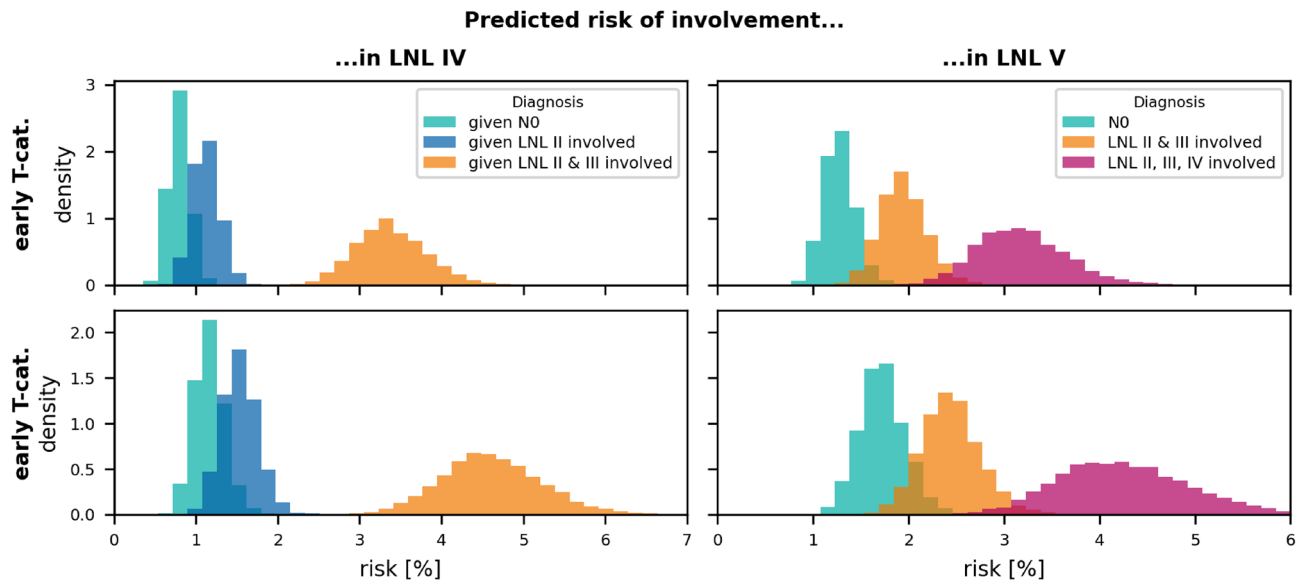
we do not assume that there is lymphatic drainage from levels III or IV to level I). Thus, the increased risk in level I is related to the time evolution: Getting diagnosed at a later time during the disease's evolution probably correlates with more advanced nodal metastasis. And involvement in the levels III and IV corresponds to a more advanced state of disease that is likely diagnosed at a later time step, such that the tumor also had more time to spread to level I. The correlation between the clinical involvement pattern, the likely time of diagnosis in the tumor's time frame based on it, and from that the risk of involvement is another benefit of the formulation of the model as an HMM.

To illustrate the flexibility of the model in predicting various risks, we have plotted the risk for occult disease in any of the LNLs I, IV, and/or V, given different clinical diagnoses in fig. 11. Similar to this, we may compute the risk for an arbitrary combination of involved levels, given a similarly arbitrary clinical diagnosis. For the base graph ([base-graph-v2](#)) and the winning graph ([win-graph-v3](#)), one may also interactively explore these risks in our web-based interface [LyProX](#), similar to how it is possible to explore the underlying data in an interactive way.

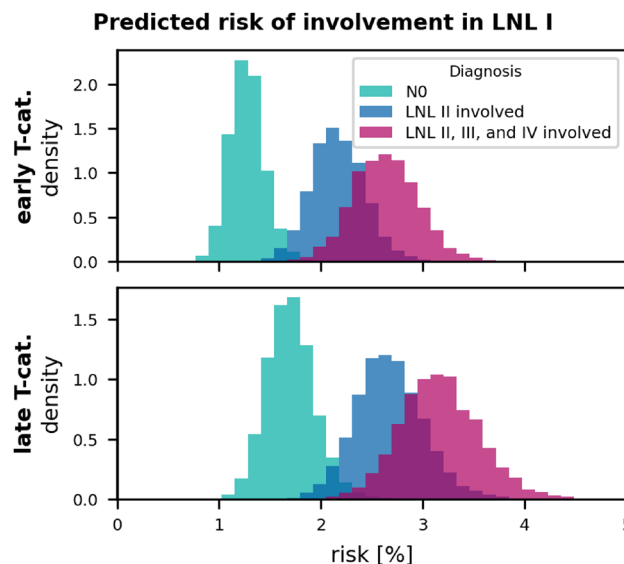
## Discussion Summary

In this publication we present a statistical model of ipsilateral lymph node involvement in oropharyngeal SCC patients. Although the basic HMM of lymphatic progression has been conceptually introduced in a previous work<sup>10</sup>, this is the first publication that evaluates the model based on a large multi-institutional dataset containing 686 patients. It is demonstrated for the first time that the model can accurately describe the patterns of lymph node involvement observed in the data, including the correlations between levels and its dependence on T-category. Furthermore, techniques from statistical physics are applied to calculate the model evidence for





**Figure 9.** Distributions over the risk for microscopic involvement in LNL IV (left panels) and in LNL V (right panels) as predicted by the winning graph model, given early (top panels) or advanced T-category (bottom panels), and different CT-based diagnoses: (1) A clinical N0 patient (green histograms), (2) visible metastases in LNL II, but otherwise healthy-looking lymph nodes (blue histograms), (3) macroscopic metastases in the LNLs II & III, with the rest of the neck still being clinically node negative (orange histograms), and finally (4) extensive clinical involvement in the levels II, III, and IV (red histograms).

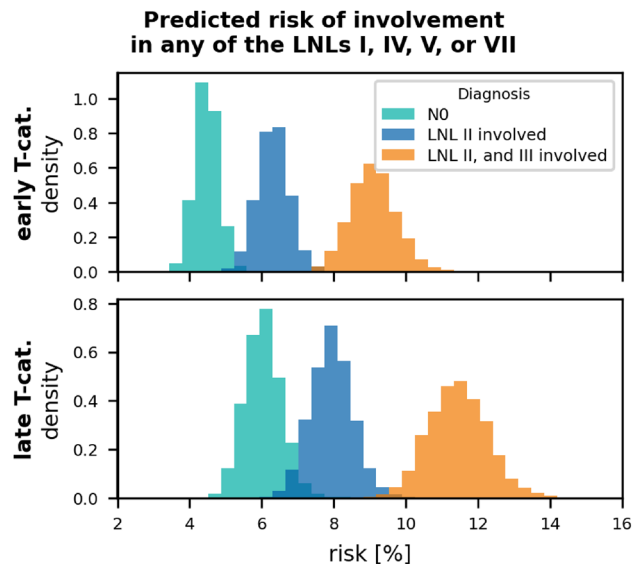


**Figure 10.** Distributions over predicted risk for involvement in LNL I, given different clinical diagnosis scenarios: For N0 patients (green), patients with macroscopic involvement in LNL II (blue), and for the case where the LNLs II, III, and IV show involvement. The top panel shows these risks for early T-category and the bottom row for advanced T-category.

Bayesian model comparison. This yields a complete model including all LNLs relevant for OPSCC: I, II, III, IV, V, and VII with a parameterization that balances accuracy and model complexity.

#### Implications for elective nodal treatment

Risk predictions obtained by the model may be used to design clinical trials on volume-deescalated treatment of OPSCC. In the context of radiotherapy, this corresponds to excluding LNLs from the CTV-N, which are irradiated according to the current guidelines. The list below should be seen as a summary of the "Risk prediction for occult disease" sect. and the limitations discussed in the "Limitations and future work" sect. should be taken into account in its interpretation. Assuming that one accepts approximately a 5% risk of occult metastases per LNL, the statistical model presented in this paper would suggest to:



**Figure 11.** Shown are the histograms over the predicted risk for involvement in any of the LNLs I, IV, V, or VII. The risk is plotted given a clinical N0 diagnosis (green), macroscopically detected metastases in LNL II (blue), and lastly given visible involvement in both LNL II and III (orange). The top row shows these risks for early T-category diagnoses and the bottom row for advanced T-category.

- Irradiate level II for all patients.
- Irradiate level III for most patients. Only for clinically N0, early T-category patients, not irradiating level III could be considered.
- Exclude level IV from the CTV-N for patients with clinically negative level III. For advanced T-category patients with involvement of level III, level IV should be irradiated.
- Exclude level V from the CTV-N for most patients. Only for patients with extensive involvement of levels II, III, and IV, irradiation of level V can be considered.
- Exclude level I from the CTV-N for early T-category patients with limited metastatic disease. For advanced T-category patients with extensive nodal involvement of levels II, III, and IV, level I should be irradiated (see also the limitations discussed in the "Limitations and future work" sect.).
- Exclude level VII in all patients, unless it is clinically involved.

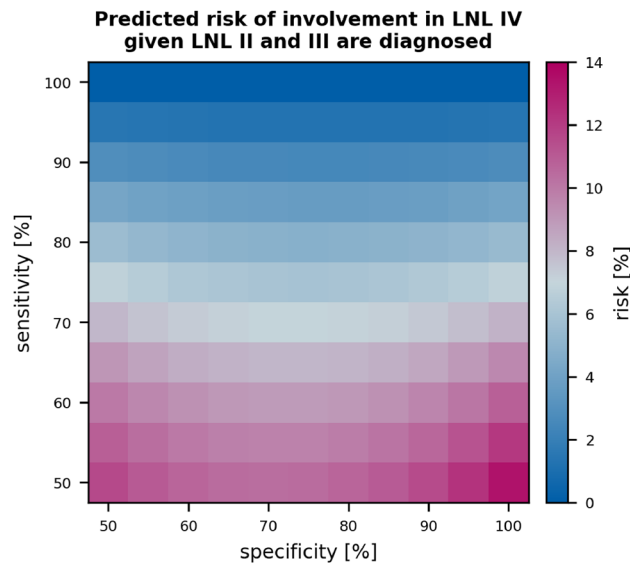
### Limitations and future work

*T-category dependence of level I involvement:* As shown in fig. 4, the model describes the involvement of LNLs II, III, and IV depending on T-category very well despite having only a single parameter related to T-category. fig. 7 shows that the model does not perfectly describe the T-category dependence of level I. It adjusts the parameters such that level I involvement is correctly described for the set of all patients combined, but it overestimates level I involvement for early T-category and underestimates it for advanced T-category. A possible explanation is that advanced T-category tumors are more likely to have grown into regions with direct lymph drainage to level I. The more severe involvement in levels II, III, IV for advanced T-category can be explained by tumors having more time to spread while keeping the spread probability rates  $b_2, b_3, b_4$  constant. Regarding level I, early versus advanced T-category tumors the model may need different spread probability rates  $b_1$  to describe their involvement correctly.

*Contralateral Spread:* The work presented in this paper considers only ipsilateral lymphatic spread. To guide the elective CTV-N definition for the contralateral neck, the model must be extended to the contralateral side. Contralateral lymph node involvement is strongly dependent on whether the primary tumor extends over the midsagittal plane, but also on T-category and the extent of ipsilateral involvement<sup>24</sup>. A comprehensive model accounting for these risk factors on contralateral spread is the subject of a follow-up publication.

*Sensitivity and Specificity:* Estimating the risk of occult metastases depends on the assumed parameter values for sensitivity and specificity of clinical detection of lymph node metastases. For this work, we adopted literature values for sensitivity and specificity. However, different authors have estimated these values using different criteria and different methods. Consequently, these values need to be considered with caution. fig. 12 illustrates for one example how the risk of occult disease depends on sensitivity and specificity. Here, we consider the risk in level IV in patients with clinically detected metastases in levels II and III. For our default parameters of 81% specificity and 76% sensitivity, the risk is 5%, but it increases to around 8% for a sensitivity of 66%.

Also, as described in the "Multicentric dataset" sect., we assumed the consensus of the data to represent the true state of nodal involvement. This was not strictly necessary: Instead of computing a consensus beforehand and providing that with sensitivity and specificity of 1 to the model, as if it were the ground truth, we could have provided multiple diagnostic modalities per patient to the model directly. In fact, for patients with a pathology report available, this would even yield the same results. But we also decided to consider the consensus as an



**Figure 12.** Dependence of risk of occult disease on sensitivity and specificity. The figure shows the risk of involvement in level IV, given the mean of the parameter samples, drawn during sampling, in patients with clinical involvement of levels II and III. This illustrates the dependence of the risk prediction on the provided sensitivity and specificity of the diagnostic modality.

observation of the true hidden state for patients without pathologically assessed involvement. We did this because the literature values for sensitivity and specificity of around 80% do not plausibly match the observation that around 78% of patients in the USZ cohort showed clinical LNL II involvement. The most likely true prevalence of involvement in LNL II would need to be close to 100%.

Discussing the possible origins for this discrepancy is beyond the scope of this work. Assuming the consensus to represent the true hidden state of a patient nonetheless allowed us to investigate if the model can describe plausible patterns of nodal involvement well. Future work may aim at developing new methods to model the difference between pathological and clinical lymph node involvement based on surgically treated patients in whom both is reported.

### Data availability

The patient data detailing lymphatic involvement is publicly available in the form of CSV tables in the GitHub repository [rmnldwg/lydata](#). In another GitHub repository, [rmnldwg/lynference](#), we define the experiments based on different graph structures. These experiments are reproducible via the tool [DVC](#) and their locally computed results are uploaded to an Azure blob storage container. Lastly, the LaTeX source code, all Python scripts to generate the figures and tables in this work, and the pipeline definition to build the document (using [show your work!](#)<sup>26</sup>) are made public in the GitHub repository [rmnldwg/graph-extension-paper](#). To rebuild

it, the [show your work!](#) tool first pulls the patient data from the [rmnldwg/lydata](#) repository and the trained

models in the form of MCMC samples from the Azure blob storage container, then runs the Python scripts, and finally compiles the LaTeX article. Detailed instructions on how to reproduce the entire pipeline, including the individual experiments, may be found in the respective repository's README .md file. We also gladly provide support for any effort to reproduce our results.

Received: 8 February 2024; Accepted: 26 June 2024

Published online: 08 July 2024

### References

- Grégoire, V. *et al.* CT-based delineation of lymph node levels and related CTVs in the node-negative neck: DAHANCA, EORTC, GORTEC, NCIC. *ROG Consensus Guidelines. Radiotherapy and Oncology* **69**, 227–236 (2003) (ISSN: 0167-8140).
- Grégoire, V. *et al.* Delineation of the Neck Node Levels for Head and Neck Tumors: A 2013 Update. DAHANCA, EORTC, HKN-PCSG, NCIC CTG, NCRI, RTOG. *TROG Consensus Guidelines. Radiotherapy and Oncology* **110**, 172–181 (2014).
- Grégoire, V. *et al.* Delineation of the Primary Tumour Clinical Target Volumes (CTV-P) in Laryngeal, Hypopharyngeal, Oropharyngeal and Oral Cavity Squamous Cell Carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO. *TROG Consensus Guidelines. Radiother. Oncol.* **126**, 3–24 (2018).
- Eisbruch, A., Foote, R. L., O'Sullivan, B., Beitler, J. J. & Vikram, B. Intensity-Modulated Radiation Therapy for Head and Neck Cancer: Emphasis on the Selection and Delineation of the Targets. *Seminars in Radiation Oncology* **12**, 238–249 (2002) (ISSN: 1053-4296).

5. Biau, J. *et al.* Selection of lymph node target volumes for definitive head and neck radiation therapy: A 2019 update. *Radiotherapy and Oncology* **134**, 1–9 (2021).
6. Chao, K., Wippold, F. J., Ozyigit, G., Tran, B. N. & Dempsey, J. F. Determination and delineation of nodal target volumes for head-and-neck cancer based on patterns of failure in patients receiving definitive and postoperative IMRT. *International Journal of Radiation Oncology Biology Physics* **53**, 1174–1184 (2002).
7. Vorwerk, H. & Hess, C. F. Guidelines for Delineation of Lymphatic Clinical Target Volumes for High Conformal Radiotherapy: Head and Neck Region. *Radiat. Oncol.* **6**, 97 (2011) (ISSN: 1748-717X).
8. Ferlito, A., Silver, C. E. & Rinaldo, A. Elective management of the neck in oral cavity squamous carcinoma: Current concepts supported by prospective studies. *British Journal of Oral and Maxillofacial Surgery* **47**, 5–9 (2021).
9. Pouymayou, B., Balermipas, P., Riesterer, O., Guckenberger, M. & Unkelbach, J. A Bayesian network model of lymphatic tumor progression for personalized Elective CTV definition in head and neck cancers. *Phys. Med. Biol.* **64**, 165003 (2019) (ISSN: 1361-6560).
10. Ludwig, R., Pouymayou, B., Balermipas, P. & Unkelbach, J. A hidden Markov model for lymphatic tumor progression in the head and neck. *Sci. Rep.* **11**, 12261 (2021).
11. Sanguineti, G. *et al.* Defining the Risk of Involvement for Each Neck Nodal Level in Patients with Early T-stage Node-Positive Oropharyngeal Carcinoma. *Int. J. Radiat. Oncol. Biol. Phys.* **74**, 1356–1364 (2009).
12. Ludwig, R. *et al.* A dataset on patient-individual lymph node involvement in oropharyngeal squamous cell carcinoma. *Data Brief* **43**, 108345 (2022) (ISSN: 2352-3409).
13. Ludwig, R. *et al.* A Multi-Centric Dataset on Patient-Individual Pathological Lymph Node Involvement in Head and Neck Squamous Cell Carcinoma SSRN Scholarly Paper. Rochester, NY, Dec. 2023. (2023).
14. Ludwig, R. Modelling Lymphatic Metastatic Progression in Head and Neck Cancer PhD thesis (University of Zurich, Zurich, 2023).
15. Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. Emcee: The MCMC Hammer. *PASP* **125**, 306 (Mar. 2013).
16. ter Braak, C. J. F. & Vrugt, J. A. Differential evolution Markov chain with snooker updater and fewer chains. *Stat. Comput.* **18**, 435–446 (2022).
17. Nelson, B., Ford, E. B. & Payne, M. J. Run DMC: an efficient, parallel code for analyzing radial velocity observations using n-body integrations and differential evolution markov chain Monte Carlo. *The Astrophysical Journal Supplement Series* **210**, 11 (2022).
18. Lengelé, B., Hamoir, M., Scalliet, P. & Grégoire, V. Anatomical Bases for the Radiological Delineation of Lymph Node Areas Major Collecting Trunks, Head and Neck. *Radiother. Oncol.* **85**, 146–155 (2022).
19. Jeffreys, H. *The Theory of Probability* (OUP Oxford, 1998).
20. Aponte, E. A. *et al.* An introduction to thermodynamic integration and application to dynamic causal models. *Cognitive Neurodynamics* **16**, 1–15 (2022).
21. De Bondt, R. *et al.* Detection of Lymph Node Metastases in Head and Neck Cancer: A Meta-Analysis Comparing US, USgFNAC, CT and MR Imaging. *Eur. J. Radiol.* **64**, 266–272 (2007).
22. Kyzas, P. A. *et al.* 18F-fluorodeoxyglucose Positron Emission Tomography to Evaluate Cervical Node Metastases in Patients with Head and Neck Squamous Cell Carcinoma: A Meta-Analysis. *J. Natl Cancer Inst.* **100**, 712–720 (2008).
23. Bhat, H. & Kumar, N. On the Derivation of the Bayesian Information Criterion (Jan. 2010).
24. Ludwig, R. *et al.* Detailed patient-individual reporting of lymph node involvement in oropharyngeal squamous cell carcinoma with an online interface. *Radiother. Oncol.* **169**, 1–7 (2022) (ISSN: 0167-8140).
25. Bauwens, L. *et al.* Prevalence and distribution of cervical lymph node metastases in HPV-positive and HPV-negative oropharyngeal squamous cell carcinoma. *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* **157**, 122–129 (2021) (ISSN: 1879-0887).
26. Luger, R. *et al.* Mapping Stellar Surfaces III: An Efficient, Scalable, and Open-Source Doppler Imaging Model Oct. 2021. [arXiv: 2110.06271](https://arxiv.org/abs/2110.06271) [astro-ph]. (2022).

## Acknowledgements

This work was supported by the Swiss Cancer Research Foundation under grant [KFS 5645-08-2022](#) and by the University Zürich under the Clinical Research Priority Program [Artificial Intelligence in Oncological Imaging](#).

## Author contributions

RL and JU wrote the main manuscript. RL created all results, figures, and tables. AS, SW, OE, MD, and RG collected the ISB patient data. DB, LB, PZ, and VG collected the CLB datasets. JMH, BP, PB, RL, and JU collected the table of USZ patients. All authors reviewed the manuscript.

## Competing Interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to R.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024