0.1 Comparing bilateral models

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

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

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

Model evidence and Bayes factor

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

$$P(\mathcal{M} \mid \boldsymbol{\mathcal{D}}) = \frac{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M})\, P(\mathcal{M})}{P(\boldsymbol{\mathcal{D}})} \tag{1}$$

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

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

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

$$P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}) = \int p(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M})\, p(\theta \mid \mathcal{M}) d\theta \tag{3}$$

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

believe it is the true one. But only that among the models investigated, this one is probably the best.

Harold Jeffreys gives a scale for interpreting values of the Bayes factor [9]:

K_{1v2}	$\ln K_{1v2}$	support for $\mathcal{M}_1$
$< 10^0$	< 0	negative evidence (supports $\mathcal{M}_2$)
10^0 to $10^{1/2}$	0 to 1.15	barely worth a mention
$10^{1/2}$ to 10^1	1.15 to 2.3	substantial
10^1 to $10^{3/2}$	2.3 to 3.45	strong
$10^{3/2}$ to 10^2	3.45 to 4.6	very strong
$> 10^2$	> 4.6	decisive

We have also listed the natural logarithm $\ln K_{1v2}$ of the Bayes factor here, because what we will actually be doing is compute differences in the log-evidences.

Thermodynamic integration

Due to the integration over all model parameters, the quantity eq. (3) is usually impossible to calculate by brute force integration, even for models with only around a dozen parameters, as is the case for ours. Unless analytical solutions exist – which is rarely the case – it is often prohibitively expensive to compute the model evidence. For this reason, a large amount of approximation methods has been developed; [8] names only a few of those methods that can be used in the context of Markov-chain Monte Carlo (MCMC). Another method that is applicable in the context of MCMC is thermodynamic integration (TI), which is very well introduced in [1] and only roughly sketched out in this section.

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

$$\begin{aligned} \ln Z \coloneqq \ln p(\boldsymbol{\mathcal{D}} \mid \mathcal{M}) &= \ln \int p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right) p(\theta \mid \mathcal{M}) d\theta - \ln 1 \\ &= \ln \int p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right) p(\theta \mid \mathcal{M}) d\theta - \underbrace{\ln \int p(\theta \mid \mathcal{M}) d\theta}_{\ln Z_0} \end{aligned} \tag{4}$$

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

$$\ln Z - \ln Z_0 = \int_0^1 \frac{d}{d\beta} \ln Z_\beta d\beta \tag{5}$$

we end up with can actually be computed. Just by inspection of eq. (4) and eq. (5), one can see that on such differentiable path could be built using what we are going to call the *power posterior* $p_\beta(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M})$:

$$\begin{aligned} \ln Z_\beta &= \ln \int p_\beta(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M}) d\theta \\ &= \ln \int p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right)^\beta p(\theta \mid \mathcal{M}) d\theta \end{aligned} \tag{6}$$

with the derivative

$$\begin{aligned}\frac{d}{d\beta}\ln Z_\beta &= \int \frac{p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right)^\beta p(\theta \mid \mathcal{M})}{Z_\beta} \ln p(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}) d\theta \\ &= \mathbb{E}\left[\ln p(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M})\right]_{p_\beta(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M})} \\ &\approx \frac{1}{S}\sum_{i=1}^{S} \ln p(\boldsymbol{\mathcal{D}} \mid \hat{\theta}_{\beta i}, \mathcal{M}) = \mathcal{A}_{\mathrm{MC}}(\beta)\end{aligned} \tag{7}$$

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

$$\ln Z \approx \frac{1}{2}\sum_{j=1}^{N-1} (\beta_{j+1} - \beta_j)\left(\mathcal{A}_{\mathrm{MC}}(\beta_{j+1}) + \mathcal{A}_{\mathrm{MC}}(\beta_j)\right) \tag{8}$$

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

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

$$\ln Z = \underbrace{\int \ln p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right) p\left(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M}\right) d\theta}_{\text{accuracy } \mathcal{A}(\beta=1)} - \underbrace{\int \ln \frac{p\left(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M}\right)}{p\left(\theta \mid \mathcal{M}\right)} p\left(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M}\right) d\theta}_{\text{complexity (KL-divergence)}} \tag{9}$$

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

Implementation

To compare the introduced models $\mathcal{M}_{\mathrm{ag}}$, $\mathcal{M}_\alpha$ and $\mathcal{M}_{\mathrm{full}}$, we performed TI with a ladder of 64 β values with step sizes selected according to a fifth order power rule. For each of the steps in the ladder, we performed an ensemble sampling round using the `emcee` [7] Python package. The size of the ensemble – consisting of so-called walkers that allow sampling in parallel and mutually influence each other's

proposals – was chosen to be 20 times the number of dimensions of the parameter space. We set the sampling algorithm to propose new samples according to a mixture of two procedures: with 80% probability it selected a differential evolution move [12] and with 20% probability a snooker move, also based on differential evolution [5]. The reason for this choice was that in previous experiments, this combination of proposals yielded the fastest convergence of the chain. Every one of the 64 sampling rounds consisted of a burn-in phase lasting 1000 steps, followed by 250 steps of which every fifth step was kept for later analysis. This might seem like a relatively short chain, but since the change of the posterior we sampled from only changed very slightly from β_j to β_{j+1}, fewer steps are required to reach convergence.

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

From the samples drawn at $\beta_{64} = 1$, we also computed the Bayesian information criterion (BIC), which is in essence a first-order approximation of the log-evidence (actually, the negative one-half of the BIC is an approximation to the log-evidence) [13]. It is defined as

$$\text{BIC} \coloneqq k \ln N - 2 \max_\theta \left(\ln p \left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M} \right) \right) \tag{10}$$

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

The models were trained on the combination of two datasets: One from our institution, the University Hospital Zurich, which has been published and described in great detail in a separate publication [10]. The other was kindly provided to us by researchers of the Centre Léon Bérard in Lyon, France and was the underlying data for one of their papers [2]. Both datasets have been published in our repository `lydata`.

Different modalities were used to obtain the diagnoses for the patients in the two datasets. For the inference process, we combined all available diagnostic modalities using sensitivity and specificity values from the literature [6] using a maximum likelihood estimate. We treated this resulting "consensus diagnosis" as if it were the ground truth, i.e., we set its sensitivity and specificity to 1 respectively. Our motivation to do so was that this allows us to compare predictions of the model with data prevalences to see which of the model exhibits more flexibility in adapting to the data. If we had directly provided the models with all available diagnostic modalities and allowed it to combine them itself, as outlined in **??**, this would not have been possible. Also, in this case we are not interested in learning the exact distribution over the posterior parameters of the model, i.e. the proba-

bility rates for spread along arcs of the lymphatic graph, but rather how well the different models are able to adapt to realistic data and make use of the additional information provided via the tumor's extension over the mid-sagittal plane.

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

Reproducibility

Each of the three models investigated here, are available in `lynference`, where we have run the respective pipeline, pushed it as a tagged commit to GitHub and published it as a release alongside the produced data in the form of a data version control (DVC) remote.
The `README.md` file in this repository explains how one can reproduce an experiment and where to find documentation on the settings and configurations used.

- Model $\mathcal{M}_{\text{ag}}$: `bilateral-v1`
- Model $\mathcal{M}_{\alpha}$: `midline-with-mixing-v1`
- Model $\mathcal{M}_{\text{full}}$: `midline-without-mixing-v1`

Results

First, we wanted to make sure that all three models are still able to describe the ipsilateral spread sufficiently well. We have plotted the prevalence our trained models predict in the forms of histograms against the Beta-posterior of the observed prevalence in the data (fig. 1). These plots were created by computing the respective prevalence for samples drawn during the final 250 steps at the end of the TI process of which every fifth step was discarded.

The shown differences between the model's predictions are miniscule. For late T-stages (bottom row of fig. 1) it seems as if the model that is agnostic to the tumor's extension over the mid-sagittal plane slightly overestimates the prevalence, while the other two models seem to match them better or underestimate them by a small amount. Overall the fit of all models ipsilaterally is very good and shows no indication that one model performs better than the other.

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

The overall prevalence of contralateral involvement is plotted in fig. 2. Again, the three different models are depicted in their own column and we have distinguished between four cases for each model: The prevalence of any contralateral involvement for patients with a) early T-stage and a cealry lateralized tumor (blue histograms and curves), b) early T-stage with a tumor extending over the mid-sagittal plane (orange), c) late T-stage with, again, a lateralized tumor (green)

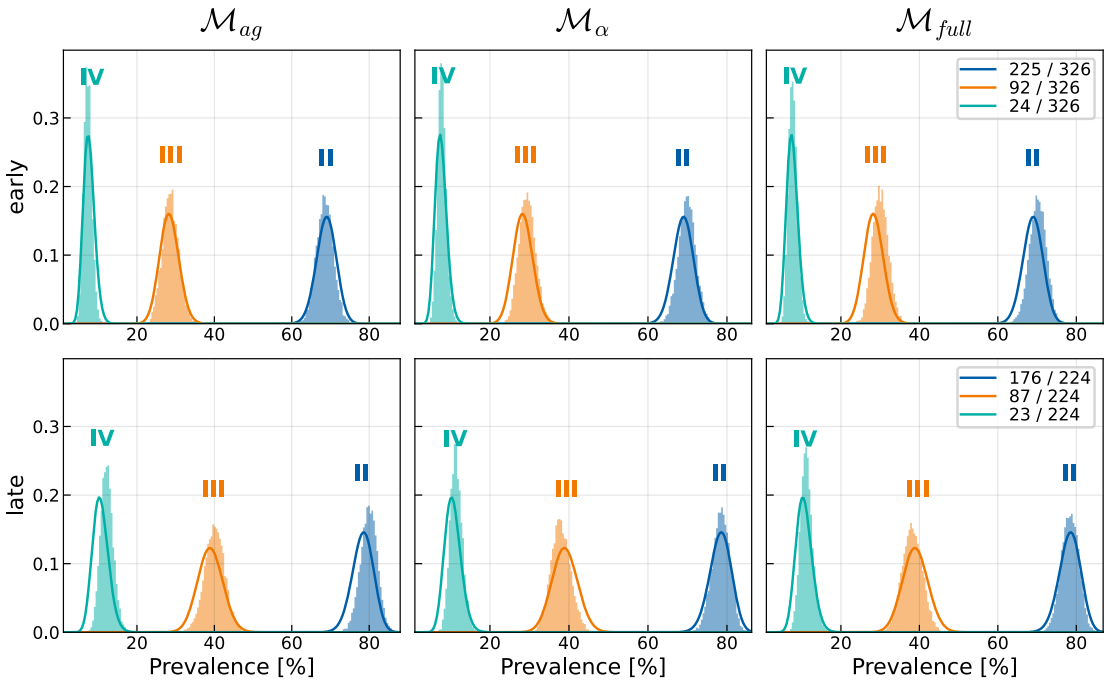

Figure 1: Predicted prevalences (shaded histograms) and posterior beta distributions of observed prevalences (solid lines) for the ipsilateral levels II (blue), III (orange) and IV (green). These prevalences have each been plotted for early T-stage patients (top row) and late T-stage (bottom row) and for the three models $\mathcal{M}_{\text{ag}}$ (left column), $\mathcal{M}_{\alpha}$ (middle column) and for $\mathcal{M}_{\text{full}}$ (right column). The differences between the models are negligible.

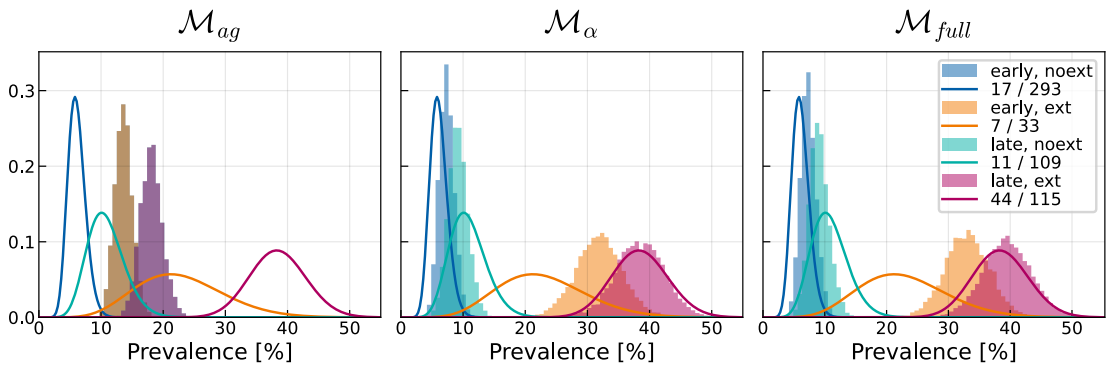

Figure 2: Predicted prevalences (shaded histograms) and posterior beta distributions of observed prevalences (solid lines) for the contralateral overall involvement (anything *not* clinically N0, on that side of the neck). Predicted and observed prevalence for early T-stage is colored blue and orange, while for late T-stage it is green and red. The prevalence for patients whose tumor does not extend over the mid-sagittal line is labelled `noext` and colored blue or green, while the same quantity for those with said extension is labelled `ext` and colored orange and red. The three models $\mathcal{M}_{\text{ag}}$, $\mathcal{M}_{\alpha}$ and $\mathcal{M}_{\text{full}}$ are depicted in the left, middle and right panel respectively.

and finally d) where the tumor is both in late T-stage and does extend over the mid plane (red).

As discussed, the agnostic model $\mathcal{M}_{\text{ag}}$ (left panel in fig. 2) cannot model midline extension, which is why the two separate histograms overlap. Its spread probability rates from the tumor to the contralateral LNLs attempt to find an average of the respective observed prevalence. Interestingly, both the model using the mixing parameter α and the full model, which has in total six parameters to model the spread from the tumor to the contralateral LNLs, perform equally well regarding the overall contralateral spread. This, in combination with fig. 1, indicates that the assumptions underlying the introduction of the mixing parameter α are feasible.

Of course one would expect that maybe modelling the correlations between involvements of the contralateral LNLs might suffer from this assumption, but this is hard to test, as cases where e.g. LNL III is involved without LNL II are very rare – in this case it is only five patients. And also clinically it is debated whether to treat or to spare the contralateral side as a whole when performing elective radiotherapy (RT) or elective bilateral neck dissection (ND), not individual LNLs [4, 11]. Therefore, a more complete model like $\mathcal{M}_{\text{full}}$, that might be able to capture correlations we cannot yet see due to insufficient data, are not worth the additional model complexity at this point.

This is supported also by the log-evidence of the three models compared, which we tabulated in . In fig. 3 we have plotted the results from computing the log-evidence for the three models in question using TI. It shows that the accuracy of the agnostic model $\mathcal{M}_{\text{ag}}$ is lower than of the other two models, which owe that to their ability to incorporate the tumor's extension over the mid plane into the prediction. However, while the mixing model $\mathcal{M}_{\alpha}$ and the full model $\mathcal{M}_{\text{full}}$ achieve the same fit to the data, the full model's accuracy rises for later β values, which results in a lower log-evidence and a higher complexity penalty (see eq. (9))

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

Table 1: Metrics computed via TI for the three bilateral models introduced in section 0.1: The BIC in the first row, the log-evidence $\ln Z$ with the respective standard deviation in the second and third row, as well as the maximum and mean likelihood of the sampling procedures in rows four and five respectively.

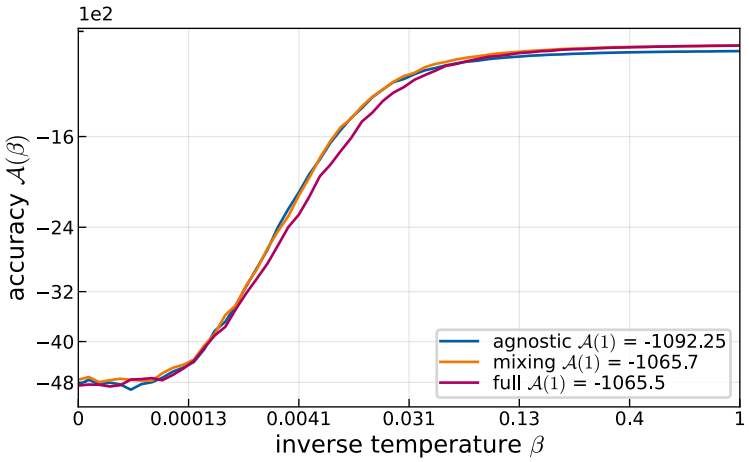

Figure 3: Expectation of the log-likelihood under the power posterior (eq. (6)) plotted against 64 inverse temperature steps β for the three models. The accuracy on the y-axis is plotted on a log-scale, while the β values, which themselves represent a fifth order annealing schedule, were plotted on an x-axis where the ticks were spaced according to a seventh order power rule. This was done to nicely visualize both the differences in accuracy and to better depict at which β values the accuracies begin to rise.

compared to the simpler mixing model.