PAPER • **OPEN ACCESS**

An organ deformation model using Bayesian inference to combine population and patient-specific data

To cite this article: Øyvind Lunde Rørtveit *et al* 2023 *Phys. Med. Biol.* **68** 055009

IOP Publishing Phys. Med. Biol. 68 (2023) 055009 https://doi.org/10.1088/1361-6560/acb8fc

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
25 October 2022

REVISED
20 December 2022

ACCEPTED FOR PUBLICATION
3 February 2023

PUBLISHED
21 February 2023

PAPER

An organ deformation model using Bayesian inference to combine population and patient-specific data

Øyvind Lunde Rørtveit[1,2,*], **Liv Bolstad Hysing**[1,2], **Andreas Størksen Stordal**[3,4] and **Sara Pilskog**[1,2]

[1] Department of Oncology and Medical Physics, Haukeland University Hospital, Bergen, Norway
[2] Department of Technology and Physics, University of Bergen, Norway
[3] NORCE Norwegian Research Centre, Bergen, Norway
[4] Department of Mathematics, University of Bergen, Norway
[*] Author to whom any correspondence should be addressed.

E-mail: oyvind.rortveit@uib.no

Keywords: organ motion, Bayesian modelling, deformable registration, radiotherapy, personalized therapy

Supplementary material for this article is available online

Abstract

Objective. Organ deformation models have the potential to improve delivery and reduce toxicity of radiotherapy, but existing data-driven motion models are based on either patient-specific or population data. We propose to combine population and patient-specific data using a Bayesian framework. Our goal is to accurately predict individual motion patterns while using fewer scans than previous models. *Approach.* We have derived and evaluated two Bayesian deformation models. The models were applied retrospectively to the rectal wall from a cohort of prostate cancer patients. These patients had repeat CT scans evenly acquired throughout radiotherapy. Each model was used to create coverage probability matrices (CPMs). The spatial correlations between these estimated CPMs and the ground truth, derived from independent scans of the same patient, were calculated. *Main results.* Spatial correlation with ground truth were significantly higher for the Bayesian deformation models than both patient-specific and population-derived models with 1, 2 or 3 patient-specific scans as input. Statistical motion simulations indicate that this result will also hold for more than 3 scans. *Significance.* The improvement over previous models means that fewer scans per patient are needed to achieve accurate deformation predictions. The models have applications in robust radiotherapy planning and evaluation, among others.

1. Introduction

In radiotherapy (RT), the dose is carefully shaped to the patient anatomy as seen in the CT acquired before start of treatment (plan CT), to achieve a good compromise between disease control and risk of inducing complications. Since the variability of the organ positions and deformations is unknown before start of treatment, different measures have been adopted to safeguard against motion uncertainties through planning margins (Stroom *et al* 1999, van Herk *et al* 2000), robust optimization (Unkelbach *et al* 2018) and/or treatment plan adaptation (Yan *et al* 1997).

A statistical model for the deformation of organs of individual patients using principal component analysis (PCA) of the organ's surface shape vectors was first proposed by Söhn *et al* (2005). The main drawback of the patient-specific model is that the number of data samples (in the form of organ contours derived from 3D images) per patient is often low, which limits the robustness of the motion estimates (Thörnqvist *et al* 2013b). Budiarto *et al* (2011) proposed a population based statistical model, under the assumption that, although the size, shape and position of organs differ greatly between patients, the patterns of deformation are generally the same. The advantage is that an estimate of a patient's deformation patterns exists even when only a single observation is available. When applied to prostate target deformation, they showed that about 50% of the variation could be explained by 15 population deformation modes (i.e. principal components). Subsequent uses

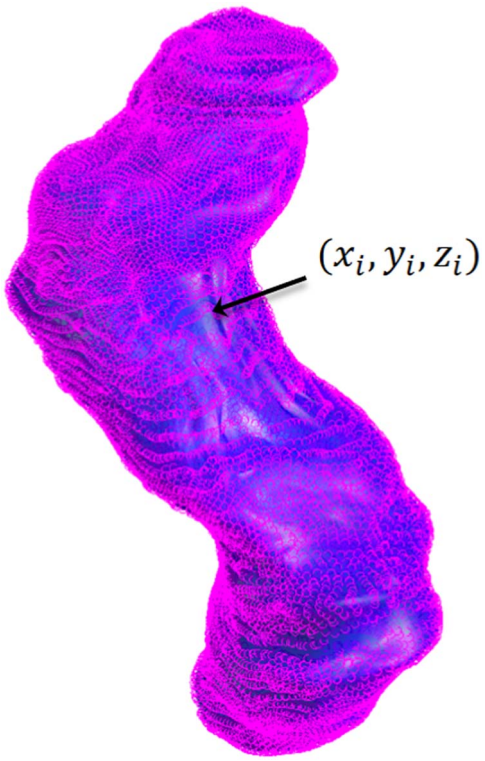

Figure 1. A rectum shape represented by a set of organ surface points.

of the population model include Bondar *et al* (2014), who used it to create margins for rectal cancer patients, Rios *et al* (2017), who modeled bladder deformation for prostate cancer RT, Szeto *et al* (2017) who modeled daily variations in the thorax, and Magallon-Baro *et al* (2019), who modeled deformation in the stomach, duodenum and bowel for pancreatic cancer RT. A weakness of the population model is its inability to model patient-specific deformation patterns, even when multiple scans are available for the patient in question. The aim of the current work is to combine the strengths of the population and patient-specific models by introducing Bayesian models that take in to account both the population deformation patterns (in terms of a *prior* distribution) and patient-specific measurements, forming an individualized *posterior* distribution. Bayesian models have previously been applied to the problem rigid shifts of the patient, termed setup errors (Lam *et al* 2005, Herschtal *et al* 2012).

In this paper, we introduce two Bayesian models, which differ in their choice of priors. The choice of model to use will be a trade-off between accuracy and simplicity. We derive necessary algorithms to efficiently calculate the approximate posterior distributions in high dimensions. We apply the introduced models to a realistic example with complex motion, in terms of the rectal wall of prostate cancer patients. We use the models to estimate coverage probability matrices (CPMs), i.e. 3D-arrays of voxels where the value in each voxel is the probability that the voxel will be covered by the rectal wall at any given time. We compare the accuracy of CPMs estimated using the two Bayesian methods, the patient-specific model by Söhn *et al* (2005) and the population model by Budiarto *et al* (2011). In addition to the presentation of new models, this is to our knowledge the first comparison between these two previous models, as well as the first time such an organ deformation model has been applied to the rectum.

2. Methods

In the class of deformation models that we study, an organ shape is represented by a set of points on the organ surface, as illustrated in figure 1. These representations are derived from organ contours segmented from 3D images. The x, y and z coordinates of all P points are gathered into a *shape vector* s:

$$s = [x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_P, y_P, z_P]^T. \quad (1)$$

With this representation, we can use standard multivariate statistical distributions.

To compare organs across scans, we need corresponding points between all shapes in the data set. This correspondence is found using deformable and rigid contour registration both within and between patients. Details are beyond the scope of the current work, but can be found in Rørtveit *et al* (2021).

Due to the random character of the organ shape, a set of shape vectors $s_1, \ldots, s_J$ derived from J scans of a patient is described as J realizations of the random variable s. For all the following methods, the shape coordinates for a specific patient are assumed to follow a multivariate Gaussian distribution:

$$s \sim \mathcal{N}(\mu, R). \tag{2}$$

The mean shape vector μ represents the patient's mean organ shape, and the covariance matrix R describes the variance of the coordinates as well as the covariance between each pair of coordinates. When μ and R are given, we can use the distribution to draw new random organ shapes for the patient. The difference between the previous patient-specific and population models and the Bayesian models introduced in section 2.3 is how μ and R are estimated. In the Bayesian methods, μ and R are considered random samples from specific prior distributions, whose parameters are calculated from the training data. *Point estimates* of μ and R are derived from the posterior distributions. Due to the high dimensions of the shape vectors, all covariance matrices are parametrized using principal component analysis (PCA), see e.g. Fujikoshi *et al* (2010, chapter 10). Under PCA, a covariance matrix is represented by a few eigenvectors and corresponding eigenvalues. These are usually found through singular value decomposition (SVD) of a *data matrix* D, whose columns are normalized mean-subtracted samples, such that $R = DD^T$.

In the following sections, we show how μ and R are estimated in the previous and the new models.

2.1. Patient-specific model

In the patient-specific model introduced by Söhn *et al* (2005), only data from the patient under consideration is used. The mean shape μ is thus set to the average of the J available shapes $s_1, s_2, \ldots, s_J$ for that patient;

$$\mu = \bar{s} = \frac{1}{J}\sum_{j=1}^{J} s_j, \tag{3}$$

while R is set to the patient-specific sample covariance matrix $\hat{R}_{\text{ps}}$:

$$\hat{R}_{\text{ps}} = \frac{1}{J-1}\sum_{j=1}^{J}(s_j - \hat{\mu})(s_j - \hat{\mu})^T. \tag{4}$$

2.2. Population model

The population model introduced by Budiarto *et al* (2011) rests on the assumption that the covariance matrix is the same for all patients, and only the mean differs. The mean is calculated as the mean shape vector for the individual patient as in (3). The covariance matrix is the average of the sample covariance matrices $\hat{R}_i$ for each patient i in the training set. Given M patients, where patient i has J_i shapes denoted $s_{i,1} \ldots s_{i,J_i}$, the estimated population covariance matrix is

$$\hat{R}_{\text{pop}} = \frac{1}{M}\sum_{i=1}^{M}\hat{R}_i = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{J_i - 1}\sum_{j=1}^{J_i}(s_{i,j} - \bar{s}_i)(s_{i,j} - \bar{s}_i)^T. \tag{5}$$

2.3. Bayesian models

In Bayesian inference, new data is combined with prior knowledge (such as population statistics) in the form of a *prior distribution*, which describes how we would expect a quantity to behave before any specific evidence is taken into account. The result of the combination of the prior and data is a *posterior distribution*.

In the following, the mean and covariance matrix for a given patient are considered random parameters that vary across the population according to a prior distribution defined by the probability density function (pdf) $f(\mu, R)$. When data for a new patient is available, we can compute the posterior pdf of μ and R given $\boldsymbol{s}$, where $\boldsymbol{s} = \{s_1, s_2, \ldots, s_J\}$, denoted $f(\mu, R|\boldsymbol{s})$, through Bayes theorem:

$$f(\mu, R|\boldsymbol{s}) = \frac{f(\boldsymbol{s}|\mu, R)f(\mu, R)}{f(\boldsymbol{s})}. \tag{6}$$

Bayes theorem gives us a distribution of the possible values of μ and R, as opposed to single values. Nevertheless, due to the complexity of the posterior distributions in our subject matter, we shall resort to looking at point estimates of μ and R, such as the expected value or mode of the posterior.

The Bayesian models we present differ in the selection of the prior distribution. We resort to priors that result in computationally feasible posterior distributions, since Markov Chain-Monte Carlo methods are computationally expensive in high dimensions. In the following sections, we present two priors which each represent a Bayesian model.

2.3.1. Normal-inverse-wishart prior

2.3.1.1. Background

We present a short background to aid the intuitive understanding of the normal-inverse-wishart (NIW) distribution. More details can be found in e.g. Bishop (2006).

A combined population and patient-specific covariance matrix $\hat{R}$ can be calculated by a simple weighted average,

$$\hat{R} = \lambda \hat{R}_{\text{pop}} + (1 - \lambda)\hat{R}_{\text{ps}}, \tag{7}$$

for some weight λ between 0 and 1. The weight should be proportional to the number J of scans used to compute the estimates. By setting $\lambda = \frac{\nu}{\nu + J}$ for some parameter ν, we obtain

$$\hat{R} = \frac{1}{\nu + J}(\nu \hat{R}_{\text{pop}} + J\hat{R}_{\text{ps}}). \tag{8}$$

We can achieve the same result by assuming an inverse Wishart (IW) prior for R and using a specific point estimate for the posterior, as shown below.

IW is a matrix distribution, and a *conjugate prior* to the multivariate Gaussian likelihood with known mean and unknown covariance matrix. This means that the posterior distribution for R is also IW, and the parameters are obtained from equations involving the prior parameters and the data. The parameters of the IW are the *scale matrix* Ψ and the *degrees of freedom* ν. Formally, if μ is given, and the prior for R is IW,

$$R \sim \mathcal{IW}(\Psi, \nu), \tag{9}$$

and the likelihood is Gaussian,

$$s|R \sim \mathcal{N}(\mu, R), \tag{10}$$

then the posterior $R|\boldsymbol{s}$, where $\boldsymbol{s} = \{s_1, s_2, \ldots, s_J\}$ is also IW,

$$R|\boldsymbol{s} \sim \mathcal{IW}(\Psi', \nu'), \tag{11}$$

with posterior parameters

$$\Psi' = \Psi + \sum_{j=1}^{J}(s_j - \mu)(s_j - \mu)^T \tag{12}$$

$$\nu' = \nu + J. \tag{13}$$

In order to obtain (8) as a point estimate for R, we define $\Psi = \nu\hat{R}_{\text{pop}}$ and set the posterior point estimate to $\hat{R} = \frac{1}{\nu'}\Psi'$. Inserting both these expressions into (12), we get

$$\hat{R} = \frac{1}{\nu + J}\left(\nu\hat{R}_{\text{pop}} + \sum_{j=1}^{J}(s_j - \mu)(s_j - \mu)^T\right). \tag{14}$$

The parameter ν determines the weight between the population covariance matrix and the sample covariance matrix of the new patient, and can be selected either by tuning or by optimization. One can think of ν as encoding the strength of our belief that $\hat{R}_{\text{pop}}$ can represent our new patient's covariance matrix.

In reality, μ is not given. One could replace μ by $\hat{\mu}$ from (3), but this will lead to bias in the covariance matrix estimate when J is small (to see this, consider equation (14) when $J = 1$ and therefore $\hat{\mu} = s_1$). Instead, we consider both μ and R random, and look for a joint prior distribution.

2.3.1.2. Normal-Inverse-Wishart distribution

The conjugate prior for the multivariate Gaussian likelihood with both unknown mean and covariance is the Normal-Inverse-Wishart (NIW) distribution. In the NIW, R is IW-distributed as in (9) , but μ and R are not independent. The conditional distribution of μ given R is Gaussian:

$$\mu|R \sim \mathcal{N}(\mu_0, \frac{1}{\kappa}R). \tag{15}$$

Here, μ_0 is the population mean, and the scalar κ represents the ratio of the variance between scans of the same patient (intra-patient) to the variance between patients (inter-patient). Thus, the NIW has the parameters μ_0, κ, Ψ and ν, and we write

$$\mu, R \sim \mathcal{NIW}(\mu_0, \kappa, \Psi, \nu). \tag{16}$$

Since this is a conjugate prior, the posterior is also NIW, and we can write

$$\mu, R|\boldsymbol{s} \sim \mathcal{NIW}(\mu_0', \kappa', \Psi', \nu') \tag{17}$$

with

$$\mu_0' = \frac{1}{\kappa + J}(\kappa\mu_0 + J\bar{s}) \tag{18}$$

$$\kappa' = \kappa + J \tag{19}$$

$$\nu' = \nu + J \tag{20}$$

$$\Psi' = \Psi + \sum_{j=1}^{J}(s_j - \bar{s})(s_j - \bar{s})^T + \frac{\kappa J}{\kappa + J}(\bar{s} - \mu_0)(\bar{s} - \mu_0)^T. \tag{21}$$

Note the similarity between (18) and (14): Both are weighted averages between population and patient-specific estimates, with the weight of the patient-specific estimate proportional to the number of patient-specific samples J. Hence, both ν and κ are parameters which determine the weight between the population and patient-specific estimates.

The final term of (21) can be considered a correction for the uncertainty of the sample mean, which makes the equation different from (12), where the mean was assumed to be known.

The maximum a-posteriori (MAP) estimate of μ is the expected value of the posterior, μ_0', so we let

$$\hat{\mu} = \frac{1}{\kappa + J}(\kappa\mu_0 + J\bar{s}). \tag{22}$$

When only a single observation for the new patient is available, i. e. $J = 1$, (22) becomes identical to the shrinkage estimation from Rørtveit *et al* (2021).

As for the IW-case, we let $\Psi = \nu\hat{R}_{\text{pop}}$ and $\hat{R} = \frac{1}{\nu'}\Psi'$. Inserting this into (21) yields

$$\hat{R} = \frac{1}{\nu + J}\left(\nu\hat{R}_{\text{pop}} + \sum_{j=1}^{J}(s_j - \bar{s})(s_j - \bar{s})^T + \frac{\kappa J}{\kappa + J}(\bar{s} - \mu_0)(\bar{s} - \mu_0)^T\right). \tag{23}$$

In practice, we never construct the full covariance matrix $\hat{R}$. Instead, it is represented by a data matrix which is augmented with extra columns, such that $D'D'^T = \hat{R}$. Given the population data matrix D, where $DD^T = \hat{R}_{\text{pop}}$, and the patient-specific data matrix S whose columns are $s_j - \bar{s}$ for $j = 1 \ldots J$, the augmented data matrix is

$$D' = \frac{1}{\sqrt{\nu + J}}\left[\sqrt{\nu}D \quad \sqrt{\frac{kJ}{k + J}}(\bar{s} - \mu_0) \quad S\right]. \tag{24}$$

2.3.2. Variational bayes model

The covariance matrix of μ describes how the individual mean varies from patient to patient, and we shall refer to it as the *inter-patient covariance matrix*. In the NIW-model, this matrix is $\frac{1}{\kappa}R$, according to (15). But the assumption that the intra-patient covariance R is proportional to the inter-patient covariance may in practice not be fulfilled. A more flexible approach is to separate the two, which motivates the following model.

Assume that the mean μ is Gaussian distributed according to

$$\mu \sim \mathcal{N}(\mu_0, \Lambda). \tag{25}$$

Here, μ_0 is the population mean, and Λ is the inter-patient covariance matrix. Assume further that R is IW distributed according to (15), and μ and R are independent (unlike in the NIW model); i.e.

$$f(\mu, R) = \mathcal{N}(\mu;\, \mu_0, \Lambda) \cdot \mathcal{IW}(R;\, \Psi, \nu). \tag{26}$$

Unfortunately, this prior is not conjugate to the Gaussian likelihood (2), and there is no simple expression for the posterior. However, both μ and R follow tractable posterior distributions *when conditioned on the other*, namely

$$\mu|R, \boldsymbol{s} = \mathcal{N}(\mu_0', \Lambda') \tag{27}$$

and

$$R|\mu, \boldsymbol{s} = \mathcal{IW}(\Psi', \nu'). \tag{28}$$

Prior distributions with this property are said to be *conditionally conjugate* to the likelihood. The conditional posterior parameters μ', Λ', Ψ' and ν' are

$$\mu_0' = (\Lambda^{-1} + JR^{-1})^{-1}(\Lambda^{-1}\mu_0 + JR^{-1}\bar{s}) \tag{29}$$

$$\Lambda' = (\Lambda^{-1} + JR^{-1})^{-1} \tag{30}$$

$$\Psi' = \Psi + \sum_{j=1}^{J}(s_j - \mu)(s_j - \mu)^T \tag{31}$$

$$\nu' = \nu + J. \tag{32}$$

The derivation of (27)–(32) is given in appendix A.

Since both μ and R are unknown, the left hand sides of (29)–(31) cannot be computed directly from the right hand sides. An alternative is to use an approximative method, known as Mean Field Variational Bayes (MFVB) (Gelman *et al* 1995). This method is applicable for conditionally conjugate priors, and is a technique used to approximate a complicated posterior distribution by a simpler distribution. The joint posterior distribution of the dependent parameters are approximated by two marginal posterior distributions by assuming independence. In our case, we are looking for densities $q_\mu()$ and $q_R()$ such that

$$q_\mu(\mu)q_R(R) \approx f(\mu, R|S). \tag{33}$$

In appendix B, we show that $q_\mu()$ is a multivariate Gaussian pdf, and $q_R()$ is an inverse Wishart pdf,

$$f(\mu, R|\boldsymbol{s}) \approx N(\mu; \mu_0^*, \Lambda^*) \cdot \mathcal{IW}(R; \Psi^*, \nu^*), \tag{34}$$

where the parameters are

$$\Lambda^* = (\Lambda^{-1} + J\nu^*\Psi^{*-1})^{-1} \tag{35}$$

$$\mu_0^* = \Lambda^*(\Lambda^{-1}\mu_0 + J\nu^*\Psi^{*-1}\bar{s}) \tag{36}$$

$$\Psi^* = \Psi + \sum_{j=1}^{J}(s_j - \mu_0^*)(s_j - \mu_0^*)^T + J\Lambda^* \tag{37}$$

$$\nu^* = \nu + J. \tag{38}$$

Equations (35)–(37) must be solved for Ψ^*, Λ^* and μ^*, but solving them analytically is not possible. We use instead a common iterative technique, where, starting at an initial guess for the parameters, the equations are iterated until convergence. If $\Psi^{*(0)}$ is the initial guess for Ψ^*, we get the following algorithm:

---

for $i = 1\ldots$ (until convergence) **do**
$\Lambda^{*(i)} = (\Lambda^{-1} + J\nu^*\Psi^{*(i-1)-1})^{-1}$
$\mu_0^{*(i)} = \Lambda^{*(i)}(\Lambda^{-1}\mu_0 + J\nu^*\Psi^{*(i-1)-1}\bar{s})$
$\Psi^{*(i)} = \Psi + \sum_{j=1}^{J}(s_j - \mu_0^{*(i)})(s_j - \mu_0^{*(i)})^T + J\Lambda^{*(i)}$
end for

The iteration is guaranteed to converge to a local optimum, but not necessarily to the global optimum. Whether we find the global optimum or not depends on the starting point. In our case, the prior and the approximate posterior have the same parameters, so the obvious choice of starting point is the corresponding parameter of the prior, i.e $\Psi^{*(0)} = \Psi$.

Finally, we extract point estimates of μ and R. We let $\hat{\mu} = \mu_0^*$. For the point estimate of $\hat{R}$, see section 2.4.4. Although we are not directly interested in Λ^*, it is needed in order to calculate the other parameters. Λ^* represents the uncertainty about the mean μ_0^*, and as such still contains information that may be valuable depending on application. Equation (35) contains the inversion of 3 matrices, all of which are of dimension $P \times P$. This is not practical; e.g. in our validation data, P is over 50000, so such an inversion would require on the order of 10^{14} floating point operations. However, these matrices are highly redundant, as they are estimated from limited data. In practice, we have found that all three update equations (35), (36) and (37) can be computed efficiently without ever constructing any $P \times P$ matrices, and with inversion of much smaller matrices only. The details of the efficient computation are given in appendix C.

2.3.3. Workflow

When new data for a patient becomes available in the form of organ contours derived from 3D-scans, the first step is to obtain point-to-point correspondence between this patient's shapes and the shapes in the training data by deformable registration to the global reference shape. Next, the resulting shape vectors $s_1, \ldots, s_J$ are used as input to one of the algorithms in this section to produce patient-specific estimates of the posterior mean and covariance matrix. How to use these further depends on the specific application.

However, the algorithms require additional parameters, specifically the hyper-parameters μ_0, Ψ and ν as well as κ or Λ depending on the model. In this section, these parameters have been assumed given. In the next section, we show how we can obtain μ_0, Ψ and Λ from training data.

2.4. Estimating model parameters from training data

Bayesian algorithms require specification of the hyperparameters of the prior. For the present models, these are μ_0, κ, Λ, Ψ and ν, with κ specific to the NIW-model and Λ specific to the variational Bayes model.The vector and matrix valued parameters μ_0, Λ and Ψ are estimated from training data. Assume that data in the form of shape vectors $s_{i,j}$ from M patients are available, where i is the patient number and j is the scan number, and patient i has J_i scans.

2.4.1. Population mean

The prior mean μ_0 is the population mean shape, which is simply calculated as the average of all the individual mean shapes in the training data:

$$\mu_0 = \frac{1}{M}\sum_{i=1}^{M} \bar{s}_i = \frac{1}{M}\sum_{i=1}^{M} \frac{1}{J_i}\sum_{j=1}^{J_i} s_{i,j}. \tag{39}$$

2.4.2. Population covariance matrix

The population covariance matrix R_{pop}, defined in (5), is in practice represented by its principal components and their variances. PCA of such a matrix has been dubbed 'simultaneous component analysis' (SCA) (Timmerman and Kiers 2003), since all patients are assumed to share the same principal components. The data matrix which is input to SCA contains all the columns from the patient-specific data matrices in the training data:

$$D_{\text{pop}} = \frac{1}{\sqrt{M}}[D_1 \quad D_2 \quad \dots \quad D_M], \tag{40}$$

where D_i is

$$D_i = \frac{1}{\sqrt{J_i - 1}}[s_{i,1} - \bar{s}_i \quad s_{i,2} - \bar{s}_i \quad \dots \quad s_{i,J_i} - \bar{s}]. \tag{41}$$

The covariance matrix $\hat{R}_{\text{pop}} = D_{\text{pop}} D_{\text{pop}}^T$ is used for both the classical population model and the NIW-model.

In the variational Bayes model, the scale matrix Ψ needs to be invertible. We will use a regularization approach for this model, where we add a constant δ_Ψ times the identity matrix, I, to the scaled sample covariance matrix:

$$\Psi = \nu\hat{R}_{\text{pop}} + \delta_\Psi I = \nu D_{\text{pop}} D_{\text{pop}}^T + \delta_\Psi I. \tag{42}$$

This structure, together with the similar structure of the inter-patient covariance matrix, makes it possible to compute the update equations(35)–(37) efficiently through the procedure detailed in appendix C.

2.4.3. Inter-patient covariance matrix

In the variational Bayes model, we also need to estimate the covariance matrix Λ of μ, the inter-patient covariance matrix. This matrix describes the uncertainty of μ. By definition,

$$\Lambda = E[(\mu - \mu_0)(\mu - \mu_0)^T], \tag{43}$$

where $E[]$ is the expected value operator. We do not have direct observations of μ, but we have estimates, $\bar{s}_i$. A natural extension of the sample covariance matrix suggests an estimator of the form

$$\hat{\Lambda}_{\text{b}} = \frac{1}{M-1}\sum_{i=1}^{M} (\bar{s}_i - \hat{\mu}_0)(\bar{s}_i - \hat{\mu}_0)^T. \tag{44}$$

This estimate of Λ is biased, since the sample mean $\bar{s}_i$ is not equal to the true mean μ. We show in appendix D that the expected value of $\hat{\Lambda}_{\text{b}}$ is

$$E[\hat{\Lambda}_{\text{b}}] = \Lambda + cE[R], \tag{45}$$

where $c = \frac{1}{M}\sum_{i=1}^{M} \frac{1}{J_i}$. The bias is therefore inversely proportional to the number of scans per patient. Since R_{pop} is an unbiased estimate of $E[R]$, we can get an unbiased estimate of Λ as

$$\hat{\Lambda} = \hat{\Lambda}_{\text{b}} - cR_{\text{pop}}. \tag{46}$$

However, since both $\hat{\Lambda}$ and $\hat{R}_{\text{pop}}$ are low rank, and they range over different subspaces, the resulting matrix is not positive semidefinite. This makes PCA a bit more complicated, but it is still possible. Details are given in appendix E. As for the intra-patient covariance matrix, the inter-patient covariance matrix must also be invertible, therefore we add a regularization factor $\delta_\Lambda I$. Additionally, since Λ expresses our uncertainly about the mean estimate, we want to have the possibility of increasing its overall size, so we introduce a constant multiplier α, which finally leads to

Table 1. Parameter values for all models. K-intra is the number of principal components used to compute the intra-patient covariance matrix, K-inter is the same for the inter-patient covariance matrix, ν and κ are scalar hyperparameters of the IW/NIW distributions, δ_Λ and δ_Ψ are regularization parameters for the matrices used in the variational Bayes iteration, and α is the weight of the inter-patient covariance matrix.

K-intra	K-inter	ν	κ	δ_Ψ	δ_Λ	α
12	20	6	0.25	240 000	80 000	4

$$\Lambda = \alpha\hat{\Lambda} + \delta_\Lambda I. \tag{47}$$

2.4.4. Probabilistic PCA

In the NIW-model, we used the point estimate $\frac{1}{\nu'}\Psi'$ for R, where Ψ' is the posterior scale matrix, and ν' is the posterior degrees of freedom. In the variational Bayes model, this is less straightforward. The posterior Ψ^* can be expressed as $D^*D^{*T} + \delta^*_\Psi I$ for some D^* and some δ^*_Ψ. The posterior δ^*_Ψ is approximately proportional to the prior δ_Ψ, and with a large δ_Ψ, the estimate $\frac{1}{\nu^*}\Psi^* = \frac{1}{\nu^*}D^*D^{*T} + \frac{1}{\nu^*}\delta^*_\Psi I$ places an unreasonable amount of variance on the shape coordinates. For this reason, we introduce a new parameter δ_R, and set the point estimate of R to

$$\hat{R} = \frac{1}{\nu^*}D^*D^{*T} + \frac{\nu}{\nu^*}\delta_R I. \tag{48}$$

For the prior distribution, the point estimate for R is found by replacing the posterior parameters values in (48) by the equivalent prior parameters. This yields

$$\hat{R}_0 = D_{\text{pop}}D_{\text{pop}}^T + \delta_R I. \tag{49}$$

When D_{pop} is found through PCA, this structure fits the description of *probabilistic PCA* (PPCA) introduced by Tipping and Bishop (1999). Their method provides a maximum likelihood estimate for δ_R given by

$$\delta_R = \frac{1}{P-K}\sum_{k=K+1}^{P}\lambda_k, \tag{50}$$

where λ_k is the kth largest eigenvalue of the population covariance matrix in (5) (i.e. the variance of the kth principal component), and K is the number of eigenpairs not discarded in PCA. In other words, δ_R is the average variance of the discarded dimensions.

3. Evaluation

3.1. Material

For evaluation, we used data from 37 patients with locally advanced prostate cancer. Each patient had 9-11 CT scans taken during treatment (typically 2 per week), including the plan CT used for RT dose planning. No laxatives were administered to the patients before or during treatment. The rectum was defined with content from the recto-sigmoid flexure to the anal verge. One single expert physicist contoured rectum on all CT scans for all patients, and all contours were reviewed and corrected by another expert physicist. This yielded a total of 373 rectum shapes, which were used in leave-one-out cross-evaluation. Details about the patients and treatment can be found in Hysing *et al* (2018). All shapes from the CT scans were converted to mesh representations with corresponding vertices, using deformable registration. Since toxicity is related to dose to the rectal wall and not its content, we evaluated the methods on the rectal wall. Since the inner wall is not seen on CT scans, we assumed 3 mm wall thickness, as in Sanguineti *et al* (2020).

3.2. Parameter values

The values of the scalar parameters were tuned manually. The values we used are shown in table 1. For the parameters K-intra and ν, which are applicable to multiple algorithms, we used the same value for all models.

3.3. Coverage probability matrices

To calculate predicted CPMs, μ_i and R_i was first estimated for each patient i using the patient-specific, population, and two Bayesian methods. For each method, 500 random rectal wall shapes per patient were then generated based on the distributions $\mathcal{N}(\mu_i, R_i)$. For each generated shape, we found which voxels (on a $1 \times 1 \times 1$ mm grid) were covered by the rectal wall using an in-house developed ray-tracing algorithm. The

coverage probability of each voxel was defined as the fraction of generated rectal walls covering that voxel. This procedure was repeated using one, two and three input scans for each method.

We used the remaining independent $J_i - 3$ scans for each patient to compute reference CPMs. Since relatively few scans (6–8) were then available, we used the bootstrapping procedure detailed in section 3.4 with this data to generate smooth CPMs. The reference CPM for each patient was computed by drawing 500 bootstrapped rectal wall shapes, and setting the coverage probability of each voxel equal to the proportion of these shapes that covered the voxel.

The predicted CPMs and reference CPMs (the ground truth) were compared in terms of their normalized cross-correlation:

$$c = \frac{\sum_{v\in V} p_{\text{predict}}(v)\, p_{\text{true}}(v)}{\sqrt{\sum_{v\in V} p^2_{\text{predict}}(v) \sum_{v\in V} p^2_{\text{true}}(v)}}, \tag{51}$$

where V is the set of all voxels, and $p_{\text{predict}}(v)$ and $p_{\text{true}}(v)$ are the predicted and true coverage probabilities at voxel v, respectively.

3.4. Convergence behaviour

To analyse convergence of the four methods without re-using structures for both training and testing, we created a virtual data set for each patient in the original data set by using a PCA-based bootstrapping procedure: For each patient, we first calculated the principal components using all the patient's available shapes. We then calculated the PCA-scores for each shape: $c_{i,j,k}$, where i is the patient number, j is the scan number and k is the component number. To generate a new random scan for patient i, a new PCA-score c_k^* was drawn for each component number k, and a new shape s_i^* was synthesized according to

$$s_i^* = \bar{s}_i + \sum_{k=1}^{J_i} c_k^* w_{i,k}, \tag{52}$$

where $w_{i,k}$ is the kth principal component vector for patient i. The c_k^* values were drawn randomly from the existing values $c_{i,j,k}$ for $j = 1 \ldots J_i$, i.e. by bootstrapping. Since the principal component scores are uncorrelated, such mixing of the scores should create realistic new shapes. The bootstrapping procedure means that no specific distribution has been assumed.

For each patient, we generated 10 shapes using this procedure. These shapes were used as input to the models to estimate CPMs. The estimated CPM for each patient was compared to the reference CPM for that patient, which was generated using all individual scans.

3.5. Impact of the uncertainty parameter δ_R

For the variational Bayes model, the parameter δ_R naturally occured from the equations and the requirement that the covariance matrix must be non-singular. The PPCA method that we used to find δ_R can also be used for the other methods. We therefore tested the effect of δ_R on the the population model, the NIW model and the variational Bayes model, and compared the result to non-probabilistic PCA, i. e. $\delta_R = 0$. PPCA is not practical for the patient-specific model with as few as 3 input scans, since it requires that some principal components are not used. For the population model, δ_R was set constant, while for the NIW and variational method, it was updated according to the update equations for Ψ, which leads to

$$\delta_R(n) = \frac{n}{n+\nu}\delta_R(0), \tag{53}$$

where n is the number of scans.

The motivation for this additional evaluation was to avoid a bias in favour of the variational Bayes model.

4. Results

Visual comparison of the four first population intra-patient modes fits with anatomical expectations (figure 2). The first mode is mainly bending of the anorectal flexure; in the bent state, the rectum is less filled than in the straight state. The second mode shows stretching and compressing of the rectum in the caudal–cranial direction. The third mode shows mainly stretching of the top of the rectum in the left-right direction, while the fourth mode shows bending left-right of the top of the rectum. A general finding is that the most caudal third of the rectum, up to slightly above the anorectal flexure, moves very little. This is corroborated by figure 3, which shows coverage probabilities of the rectum wall for two example patients, a 'small mover' and a 'large mover'.

The Bayesian models take advantage of population data also when estimating the patient-specific mean rectum $\hat{\mu}$. Figure 4 shows how the mean estimates may differ with the Bayesian models for an example patient,

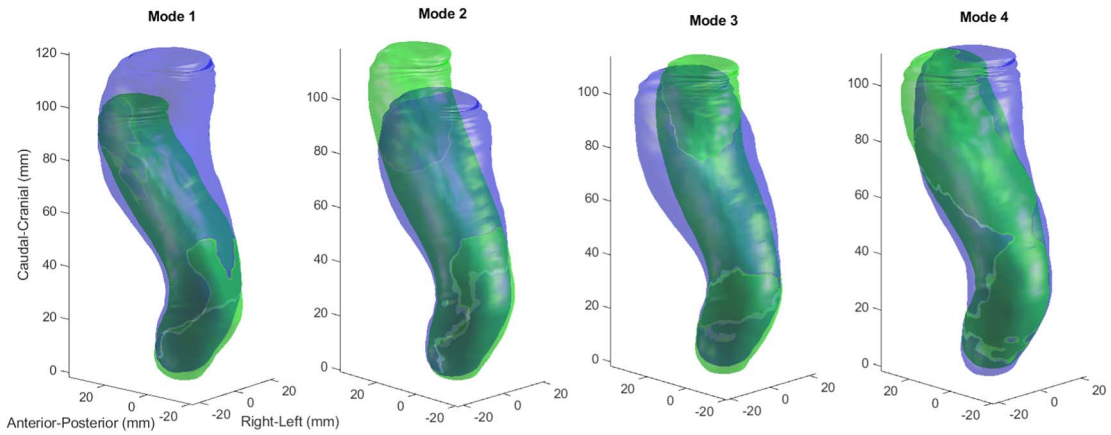

Figure 2. The four greatest population intra-patient deformation modes. The green and blue shapes represent +2 and −2 standard deviations of the deformation mode from the population mean rectum shape. A video with animation of the deformation modes is available in the supplementary material.

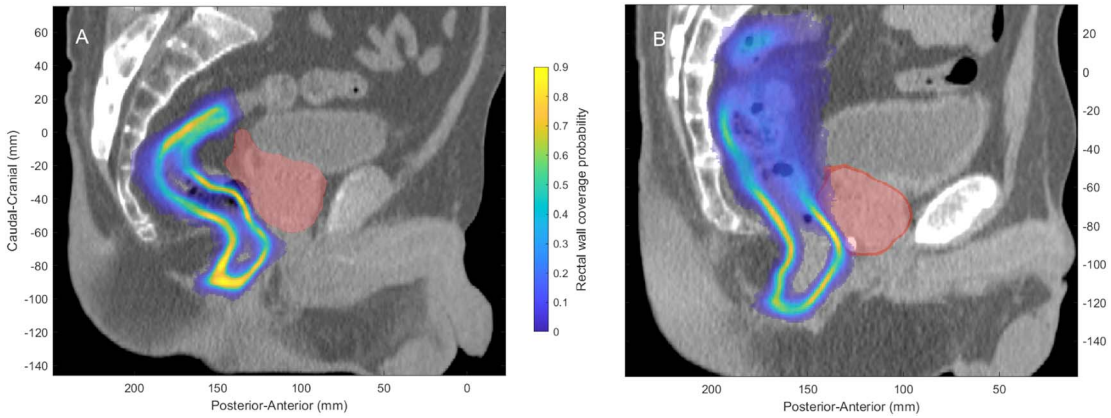

Figure 3. Coverage probability matrices for the rectal wall on a sagittal slice of the CT scan for two example patients—a 'small mover' (A) and a 'large mover' (B). The red area is the high dose volume to the prostate, that receives more than 67 Gy EQD_2.

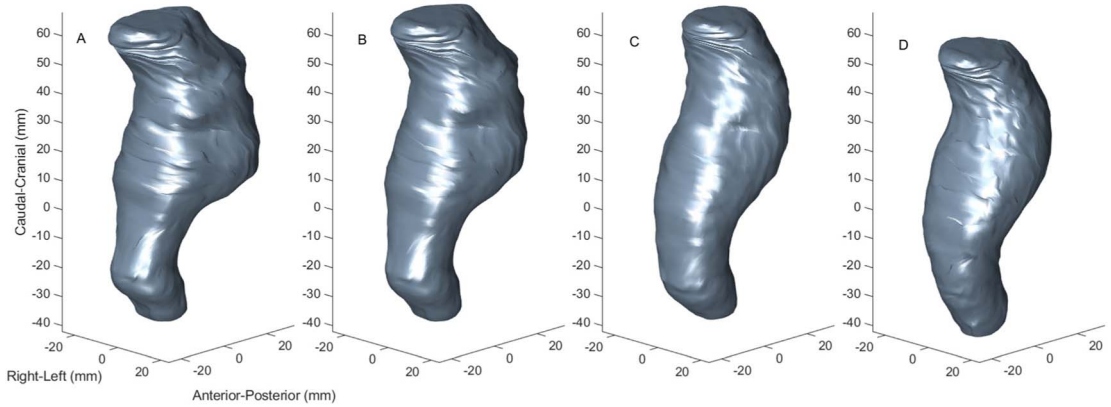

Figure 4. Estimated means and actual mean rectum shape for an example patient. A: Plan CT rectum shape, input to estimation algorithms. B: Mean rectum shape estimated by NIW model. C: Mean rectum shape estimated by variational Bayes model. D: Actual mean rectum shape over 9 scans.

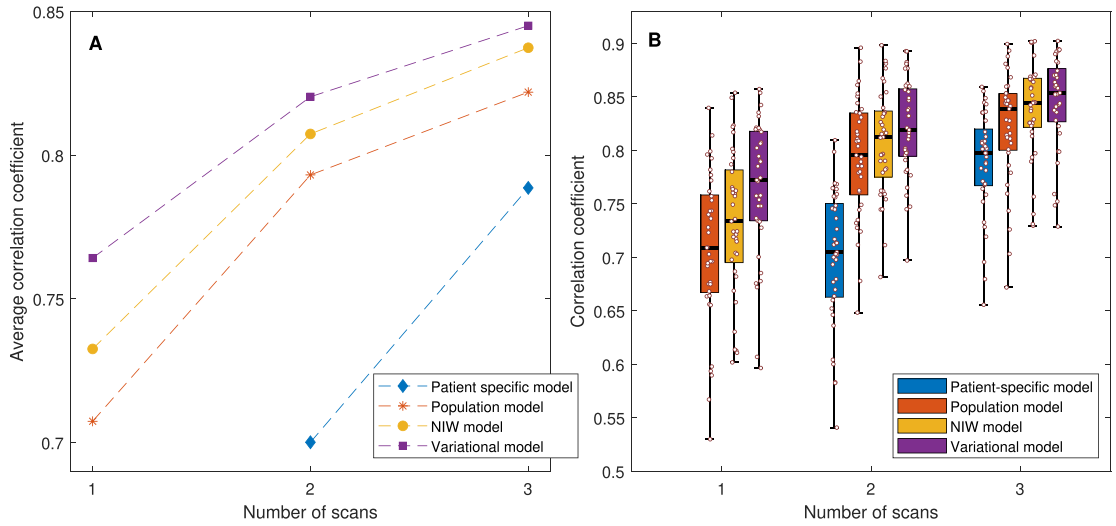

Figure 5. Correlation between the estimated CPMs and the references for the different methods using 1–3 input scans. A: Average correlation. B: Box plots showing median, 25th and 75th percentile and minimum and maximum values (whiskers). All individual values are also shown as circles over the box plot.

Table 2. Difference in CPM correlation between the population, NIW and variational models using one, two and three scans. Here, $\Delta\mu$ is the difference in average value of the CPM correlations, and %+ is the percentage of patients that saw improvement with the first method over the second.

	NIW versus pop. model			Variational versus pop. model			Variational versus NIW		
	$\Delta\mu$	p-value	%+	$\Delta\mu$	p-value	%+	$\Delta\mu$	p-value	%+
1 scan	0.026	6.2e-5	78	0.058	1.2e-8	95	0.032	2.2e-6	81
2 scans	0.014	1.8e-4	81	0.027	2.5e-6	86	0.013	1.2e-3	70
3 scans	0.015	2-2e-6	89	0.023	8.0e-7	89	0.008	0.01	62

given a single input scan. For this patient, the mean shape from variational Bayes model had the greatest similarity with the true mean shape.

The average correlation between the estimated CPMs and the references is shown in figure 5(A), while figure 5(B) shows the spread of the results among the individual patients. The two Bayesian methods outperform both the existing models, with the variational Bayes model showing superior results to the NIW-model. The results are summarized in table 2, where the patient-specific model has been left out since it performs poorly with as few as three scans. The differences between the population, NIW and variational Bayes model were consistently significant ($p < 0.05$). In comparison to the best existing model (the population model), the variation Bayes model improved correlation with the reference CPM in 35 out of 37 patients when using a single input scan (figure 6).

4.1. Convergence behaviour

The two Bayesian methods both outperform the patient-specific model with up to 6 scans, and outperforms the population model for any number of scans (figure 7). As the number of input scans increases, the patient-specific model and the two Bayesian models appear to converge toward the true CPM, while the population model improves only moderately. This is to be expected, since, in the population model, the covariance matrix representing the random error is never updated. All improvement seen in the population model is therefore from reduction of error in the mean estimate, often referred to as systematic error. The performance of the patient-specific model is comparable to that of the population model when both are given 4 scans. For more than 4 scans, the patient-specific model outperforms the population model. The variational Bayes model consistently performs slightly better than the NIW-model.

4.2. Impact of the uncertainty parameter δ_R

For all the models, PPCA through the addition of the δ_R parameter increases correlation as compared to ordinary PCA, as shown in figure 8. The difference between the models with and without the uncertainty

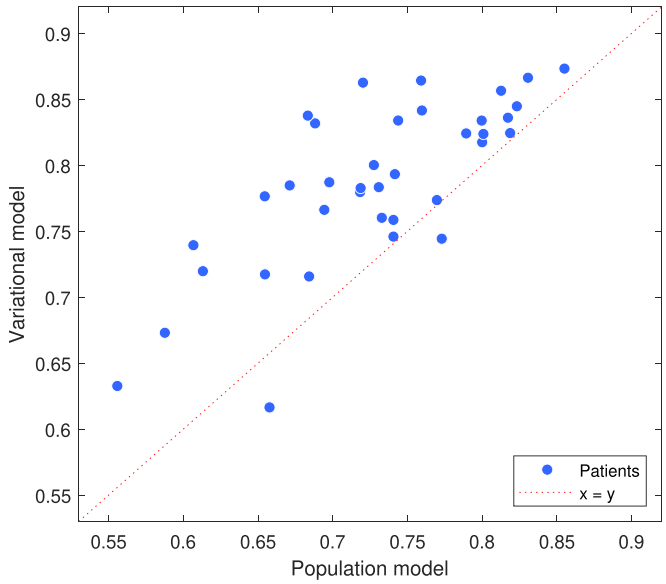

Figure 6. Correlation between the estimated CPMs and the references; comparison between the traditional population model and the proposed variational Bayes model. All points above the '$x = y$' line represent patients for which the variational Bayes method produced a better CPM estimate than the population model.

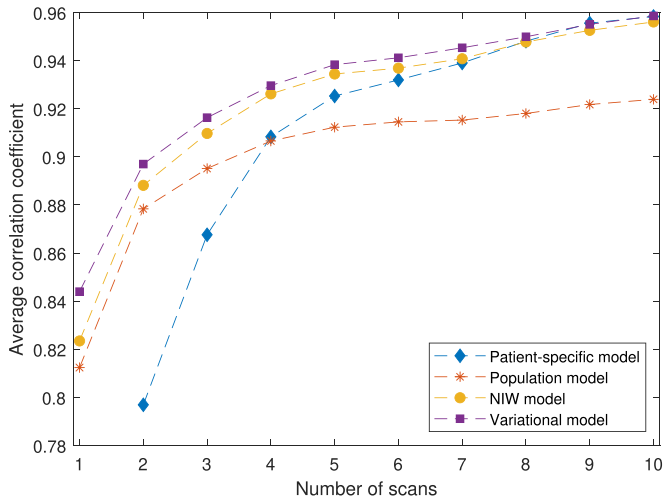

Figure 7. Average correlation between the estimated CPMs and the reference CPMs for the different methods using 1–10 scans, based on bootstrapped data.

parameter is greatest when using a single scan. Although the differences between the models decreased, both Bayesian methods with ordinary PCA still perform the same as, or better than the population model with PPCA.

5. Discussion

Both the new models outperform the existing population model significantly. Conceptually, the NIW model is only slightly more complex than the population model, so there is little rationale for rather using the population model. Additionally, figure 6 shows that the Bayesian models are robust, as evidenced by the fact that 35 out of 37 patients had improved result with the variational Bayes model over the population model (29/37 for the NIW-model without PPCA). There is therefore very little risk involved in moving to a Bayesian model.

It is to be expected that the new algorithms will perform worse for some patients due to the random nature of the data. Nevertheless, we examined the data for the two patients who performed worse with the variational Bayes than the population model using one scan to see if there were notable patterns. While no conclusion can be

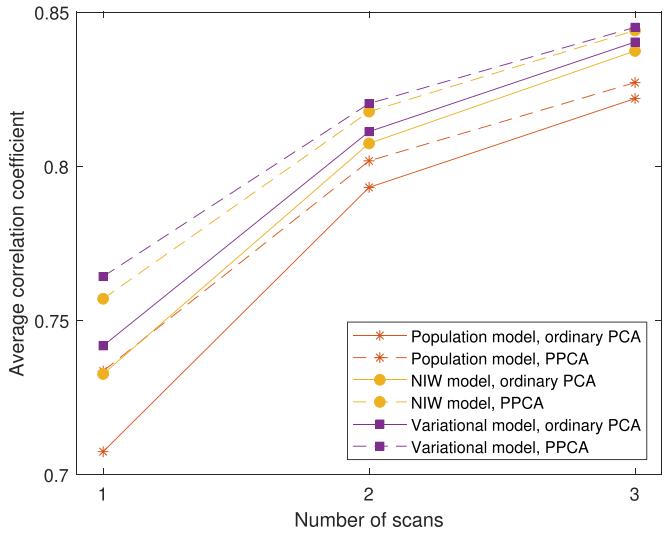

Figure 8. Comparison of ordinary and probabilistic PCA for different models. Each symbol represents the average correlation between the estimated CPMs and the references.

reached, it seems that, for these patients, the rectal shape in the pCT is coincidentally similar to the mean shape over all CTs.

The choice between the two Bayesian methods is a tradeoff between model accuracy and complexity. The main concern with the variational Bayes model is the conceptual rather than the computational complexity—it is more challenging to implement and requires more parameters than the NIW model. When using PPCA, the NIW models performance gets close to that of the variational Bayes model.

As expected, the patient-specific model cannot compete with the other models when few scans are available. This model still has an advantage in that no training data from the population is required. Additionally, deformable registration is more readily available between contours of the same patient than between contours of different patients. There are therefore applications where the patient-specific model is the only available option. However, in these cases, care should be taken that sufficient scans are available, as shown in figures 5 and 7.

The convergence analysis in figure 7 shows that we have achieved the goal of combining the advantages of both models; requiring few scans to achieve good accuracy while also improving accuracy with more scans. At around eight scans, the patient-specific model catches up with the Bayesian models. This is to be expected -at that point, the Bayesian models put very little weight on the population data since there is sufficient patient-specific data for an accurate model.

We have evaluated the model for the rectum, a highly flexible and deformable organ. The ability of the method to model other organs will depend on the amount of individual variation and the ability of the training data to replicate the variations that appear in the population. The fact that the models combine patient-specific data with the training data suggests that they should out-perform purely population based methods when there is great variability in the individual deformation. It is also possible to model multiple organs simultaneously, as done with the individual model in Söhn *et al* (2005). This may be advantageous, as correlations between the deformations of the different organs and their relative positions are taken into account.

As far as our experience goes, the variational Bayes iteration is not sensitive to the selected starting guess of the scale matrix Ψ^*, it appears to converge to the same solution regardless of starting point. The iteration takes less than a second to run for a single patient. Generating a CPM with a resoluton of 1 mm (about 3 million points) from 500 generated shapes took about 5 seconds on a standard PC. In practice, the main computational effort will be spent on deformable registration, which takes about 2 minutes for a single registration in our setup[5]

5.1. Applications

The calculation of CPMs play a key role in many applications of organ deformation models (Price and Moore 2007). The CPMs can be used for robust RT planning (Baum *et al* 2006), or to calculate margins based on the formula of Stroom *et al* (1999), as in Hysing *et al* (2011), Thörnqvist *et al* (2013a), Magallon-Baro *et al* (2019). In Ramlov *et al* (2017), Lindegaard *et al* (2017), CPMs were used clinically to reduce toxicity in nodal boosting of cervical cancer RT. Applications besides CPMs include robust evaluation through treatment course simulation (Söhn *et al* 2012, Hysing *et al* 2018), generation of plan libraries for RT personalized to motion (Rigaud *et al* 2019)

[5] Matterhorn software from Erasmus MC (Rotterdam), running on an Intel i7-4600U 2.1 GHz CPU.

and motion-robust optimization (Sobotta *et al* 2010, Unkelbach *et al* 2018). Recently, Owens *et al* (2022) used a pure inter-patient model to reconstruct colorectal dose in childhood cancer survivors who had received RT with no CT simulation. Thus, applications also extends to improving evaluation of complications from RT.

The Bayesian approach offers additional advantages because it quantifies the model uncertainty. Consider for example the robust evaluation in Söhn *et al* (2012), Hysing *et al* (2018): predicting dose-volume histograms (DVHs) with uncertainties (such as 5th and 95th percentiles). When using a non-Bayesian deformation model, the correctness of the predicted values rely on the correctness of the model's parameters. With a Bayesian model, the uncertainty of the parameters will translate to additional uncertainty regarding the dose-volume histogram, thus increasing the difference between the expected value and the 5/95 percentiles.

Interfractional geometrical errors in RT are often divided into systematic and random errors. The random error is the motion around the mean shape and position at each fraction, while the systematic error is the difference between the actual mean and the estimated mean, usually the shape and position at the plan CT. In terms of the deformation models, the systematic error is the difference between the estimated and the true patient mean, $\hat{\mu} - \mu$. The presented Bayesian models reduces the systematic error as compared to the previous methods by utilizing population data when estimating $\hat{\mu}$ (see figure 4). In addition, the new models provide a personalized *distribution* for the systematic error in terms of the posterior inter-patient distribution. The widely applied margin recipe by van Herk *et al* (2000) uses the formula $2.5\Sigma + 0.7\sigma$, where Σ and σ are the standard deviations of the systematic and random errors, respectively. Because the distribution of both the systematic and random errors are modeled under the Bayesian framework, it is in principle possible to use similar recipes for margins due to deformation.

5.2. Choice of evaluation metric

The cross-correlation metric puts proportionally higher weight on voxels that have a high coverage probability. Since a large portion of the organ tends to overlap in most or all shapes for one patient, all methods will tend to produce relatively high correlation values. Therefore, the differences between the methods may seem small. We still choose to use this metric because of its simplicity and ease of reproduction.

5.3. Gaussian likelihood

Both the Bayesian models and the models we compare to make the assumption that the data for a given patient is multivariate Gaussian distributed. This has been a standard assumption in applications of deformation models (e.g. Söhn *et al* 2012, Rios *et al* 2017). In the high dimensions that we operate in, it would require unrealistically many individual scans to disprove Gaussianness. Nevertheless, this assumption is a possible source of error, which showcases the need to evaluate the model against real data.

It should be possible to adapt the patient-specific and population models to use a nonparametric distribution of the PCA-scores as in Fontenla *et al* (2001), but this has not yet been demonstrated. In a Bayesian model, a non-Gaussian likelihood would make calculating the posterior mathematically intractable.

5.4. Parameter values

The values of the scalar parameters in table 1 were hand tuned with the objective to maximize the CPM correlations. Since it is not possible to evaluate the accuracy of the estimated distribution for a new patient without having many individual scans, one must in practice trust that parameter values that worked well for the training data also works well for new patients. If new data source is in some way different from the training data (e.g. a different image modality or IGRT routine, a different diagnosis or otherwise different type of patient), the parameters should at least be evaluated for this kind of data. However, in such cases Bayesian inference should perform better than a pure population approach, as it tailors the distribution to the data at hand.

The parameter κ for the NIW-model was set to 0.25. Using equation (18), we find that, given one input scan, this represents an *shrinkage factor* of 0.2; i.e the estimated mean is 'shrunk' by a factor 0.2 towards the population mean (Rørtveit *et al* 2021). The parameter ν, the number of degrees of freedom of the Wishart distribution, was set to 6 for both the NIW and the variational model. Normally, ν represents the number of samples from which Ψ was computed. However, this is under the assumption that these samples were all drawn from the same multivariate Gaussian distribution. In our case, the samples were drawn from M different Gaussian distributions with covariance matrices R_i, none of which match a future patient's covariance matrix. Therefore, we are much less certain about R, and we need to choose a value for ν that is much smaller than the total number of observations in the training data.

When tuning the values of δ_Ψ and δ_Λ, we found that these needed to be set surprisingly large to achieve satisfactory results. Possibly, some assumptions or parts of the model do not actually fit the data well, and increasing the regularization values then compensates for the poor fit. This underscores the importance of evaluating the models with realistic data, and tailoring the parameters to the case at hand.

5.5. Degenerate inverse Wishart distribution

The inverse Wishart distribution is usually defined in terms of the (forward) Wishart distribution: If a random $n \times n$ matrix G is Wishart distributed with $G \sim \mathcal{W}(\Psi, \nu)$, then its inverse G^{-1} is inverse Wishart distributed with $G^{-1} \sim \mathcal{IW}(\Psi^{-1}, \nu)$. However, when $\nu < n$, the Wishart distribution is degenerate, as any matrix G with a non-zero probability density has rank ν and is therefore singular. Then this definition of the IW distribution does not work. A singular inverse Wishart distribution is defined through the pseudo-inverse of W (Cook and Forzani 2011, Bodnar *et al* 2016). Unfortunately, this distribution is not well behaved, and does not have a finite expected value. Since we do not explicitly use the distribution, but rather a point estimate, this does not make a difference when using the models as described in this paper. However, care must be taken if using the full Bayesian model as described in section 5.6, as individual realizations of G can have very large eigenvalues.

5.6. Extensions

We have applied the models to the rectum alone, however, for use in e.g. robust optimization, it would be advantageous to model several structures simultaneously so that the correlation between structures are taken into account.

In the evaluation of the algorithms, we used point estimates for μ and R as opposed to a full distribution. We have thus ignored the uncertainty in the model itself, and therefore sinned against the Bayesian philosophy. We chose to do this for the sake of computational complexity. However, it is possible to account for the additional uncertainty: When performing Monte-Carlo sampling, one would first sample μ and R from the posterior distribution every time before sampling s from $\mathcal{N}(\mu, R)$. The resulting distribution of s is called the *posterior predictive* distribution. Particularly the sampling of R is computationally intensive. An alternative approach might therefore be to use a point estimate for R while sampling μ, as systematic errors are often of greater importance than random errors.

The presented models have been applied to deformably registered organ surfaces. A more common form of deformable registration is the deformation of 3D-images with image intensities. Since both types of registration produce deformation vector fields, it is possible, with some adaptions, to apply these models to deformed images as well.

6. Conclusions

We have implemented and evaluated two Bayesian methods for modelling organ deformation occuring during RT treatment. The NIW and the variational Bayes models both outperformed previous organ deformation models when applied to the rectal wall of prostate cancer patients.

Acknowledgments

The authors would like to thank Markus Alber for his contribution with ideas and discussions in the early phase of this project. We also thank Andras Zolnay at Erasmus MC Cancer center for valuable insight and discussion.

Funding

This work was funded by Trond Mohn Foundation [grant number BFS2017TMT07].

Ethical statement

All patients gave their consent before being enrolled in a phase II dose-escalation trial delivered with moderately hypo-fractionated pelvic IMRT at Haukeland University Hospital, Bergen, Norway. The study protocol was approved by the Regional Ethical Committee (REK 2006/15727) before enrollment starting in 2007. The research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local statutory requirements.

Appendix A. Derivation of the conditional posteriors

The pdf for the multivariate Gaussian distribution for a vector x of dimension p is

$$\mathcal{N}(x;\,\mu,\,R) = \frac{1}{\sqrt{(2\pi)^p|R|}}\exp\left(-\frac{1}{2}(x-\mu)^T R^{-1}(x-\mu)\right). \tag{A.1}$$

The pdf of the inverse Wishart distribution of a $p \times p$ matrix Q is

$$\mathcal{IW}(Q;\,\Psi,\,\nu) = \frac{|\Psi|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\frac{\nu}{2})}|Q|^{-(\nu+p+1)/2}\exp\left(-\frac{1}{2}\mathrm{tr}(\Psi Q^{-1})\right). \tag{A.2}$$

The joint pdf of μ, R and the samples $S = \{s_1, s_2,\ldots,s_n\}$, based on our prior and our likelihood is

$$f(\mu,\,R,\,S) = \mathcal{N}(\mu;\,\mu_0,\,\Lambda)\mathcal{IW}(R;\,\Psi,\,\nu)\prod_{i=1}^{n}\mathcal{N}(s_i;\,\mu,\,R). \tag{A.3}$$

Writing this out using (A.1) and (A.2), and leaving out any constant factors (factors that do not contain μ, R or S), we find

$$f(\mu,\,R,\,S) \propto \frac{|R|^{-(\nu+p+1)/2}}{|R|^{n/2}}\exp\left(-\frac{1}{2}(\mu-\mu_0)^T\Lambda^{-1}(\mu-\mu_0)\right.$$
$$\left.-\frac{1}{2}\mathrm{tr}(\Psi R^{-1}) - \frac{1}{2}\sum_{i=1}^{n}(s_i-\mu)^T R^{-1}(s_i-\mu)\right). \tag{A.4}$$

Using the property of the trace $\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$ and the fact that a scalar is its own trace, the sum within the exponential can be written as

$$\sum_{i=1}^{n}(s_i-\mu)^T R^{-1}(s_i-\mu) = \mathrm{tr}\left([\sum_{i=1}^{n}(s_i-\mu)(s_i-\mu)^T]R^{-1}\right). \tag{A.5}$$

Furthermore, since $\mathrm{tr}(A) + \mathrm{tr}(B) = \mathrm{tr}(A+B)$, we can write

$$\mathrm{tr}(\Psi R^{-1}) + \sum_{i=1}^{n}(s_i-\mu)^T R^{-1}(s_i-\mu)$$
$$=\mathrm{tr}([\Psi + \sum_{i=1}^{n}(s_i-\mu)(s_i-\mu)^T]R^{-1}) \tag{A.6}$$

To condition (A.4) on μ and S, we can leave out any factors not containing R - that is, the first term in the exponenial. Using (A.6), we find

$$f(R|\mu,\,S)$$
$$\propto |R|^{-(\nu+p+1+n)/2}\exp\left(-\frac{1}{2}\mathrm{tr}([\Psi + \sum_{i=1}^{n}(s_i-\mu)(s_i-\mu)^T]R^{-1})\right)$$
$$\propto \mathcal{IW}(R;\,\Psi',\,\nu'), \tag{A.7}$$

with

$$\Psi' = \Psi + \sum_{i=1}^{n}(s_i-\mu)(s_i-\mu)^T \tag{A.8}$$

and

$$\nu' = \nu + n, \tag{A.9}$$

which concludes the derivation of the conditional posterior for R.

Next, we condition (A.4) on R and S to find

$$f(\mu|R,\,S)$$
$$\propto \exp\left(-\frac{1}{2}(\mu-\mu_0)^T\Lambda^{-1}(\mu-\mu_0) - \frac{1}{2}\sum_{i=1}^{n}(s_i-\mu)^T R^{-1}(s_i-\mu)\right) \tag{A.10}$$

Looking actively for a Gaussian distribution, we want to find that the terms inside the exponential are equal to

$$-\frac{1}{2}(\mu-\mu_0')^T\Lambda'^{-1}(\mu-\mu_0') + c, \tag{A.11}$$

for some Λ' and μ', with any constant term c. Grouping the terms that are quadratic in μ, we find

$$-\frac{1}{2}\mu^T\Lambda^{-1}\mu - \frac{1}{2}\sum_{i=1}^{n}\mu^T R^{-1}\mu = -\frac{1}{2}\mu^T(\Lambda^{-1} + nR^{-1})\mu, \tag{A.12}$$

therefore, if this is a Gaussian distribution, we must have

$$\Lambda'^{-1} = \Lambda^{-1} + nR^{-1} \rightarrow \Lambda' = (\Lambda^{-1} + nR^{-1})^{-1}. \tag{A.13}$$

Grouping the linear terms, we find

$$-\mu^T\Lambda^{-1}\mu_0 - \sum_{i=1}^{n}\mu^T R^{-1}s_i = -\mu^T(\Lambda^{-1}\mu_0 + nR^{-1}\bar{s}). \tag{A.14}$$

Setting this equal to the linear terms in (A.11), we have

$$\mu^T\Lambda'^{-1}\mu_0' = \mu^T(\Lambda^{-1}\mu_0 + nR^{-1}\bar{s}), \tag{A.15}$$

which is true for any μ if and only if

$$\mu_0' = \Lambda'(\Lambda^{-1}\mu_0 + nR^{-1}\bar{s}). \tag{A.16}$$

The constant terms can be ignored, as they will be absorbed by the normalization. Finally, this gives us

$$f(\mu|R, S) \propto \exp\left(-\frac{1}{2}(\mu - \mu_0')^T\Lambda'^{-1}(\mu - \mu_0')\right) \tag{A.17}$$

$$\propto \mathcal{N}(\mu; \mu_0', \Lambda'), \tag{A.18}$$

with μ_0' as in (A.16) and Λ' as in (A.13). □

Appendix B. Variational approximation

To find the functions q_μ and q_R, we follow the procedure presented in Gelman *et al* (1995). The minimizing functions are given by

$$\log q_\mu(\mu) = E_R[\log f(\mu|R, S)] + \text{const} \tag{B.1}$$

and

$$\log q_R(R) = E_\mu[\log f(R|\mu, S)] + \text{const}, \tag{B.2}$$

where E_R and E_μ indicate an average over R only or μ only, respectively.

Inserting (A.10) into (B.1), we get

$$\log q_\mu(\mu) = E_R\Bigg[-\frac{1}{2}(\mu - \mu_0)^T\Lambda^{-1}(\mu - \mu_0) - \frac{1}{2}\sum_{i=1}^{n}(s_i - \mu)^T R^{-1}(s_i - \mu)\Bigg] + \text{const} \tag{B.3}$$

$$= -\frac{1}{2}(\mu - \mu_0)^T\Lambda^{-1}(\mu - \mu_0) - \frac{1}{2}\sum_{i=1}^{n}(s_i - \mu)^T E[R^{-1}](s_i - \mu) + \text{const}. \tag{B.4}$$

Following the lines of the derivation in appendix A, we find

$$q_\mu(\mu) = \mathcal{N}(\mu; \mu_0^*, \Lambda^*), \tag{B.5}$$

with

$$\mu_0^* = (\Lambda^{-1} + nE[R^{-1}])^{-1}(\Lambda^{-1}\mu_0 + nE[R^{-1}]\bar{s}). \tag{B.6}$$

and

$$\Lambda^* = (\Lambda^{-1} + nE[R^{-1}])^{-1}. \tag{B.7}$$

Similarly, we insert (A.7) into (B.2) to find

$$\log q_R(R) = E_\mu\Bigg[\log(|R|^{-(\nu+p+1+n)/2}) - \frac{1}{2}\text{tr}([\Psi + \sum_{i=1}^{n}(s_i - \mu)(s_i - \mu)^T]R^{-1})\Bigg] + \text{const} \tag{B.8}$$

$$= \log(|R|^{-(\nu+p+1+n)/2}) - \frac{1}{2}\mathrm{tr}([\Psi + \sum_{i=1}^{n} E[(s_i - \mu)(s_i - \mu)^T]]R^{-1}) + \text{const} \tag{B.9}$$

The term within the expectation operator is

$$\sum_{i=1}^{n} E[(s_i - \mu)(s_i - \mu)^T] = \sum_{i=1}^{n}(s_i s_i^T + E[\mu\mu^T] - E[\mu]s_i^T - s_i E[\mu]^T) \tag{B.10}$$

$$= \sum_{i=1}^{n}(s_i - E[\mu])(s_i - E[\mu])^T + n(E[\mu\mu^T] - E[\mu]E[\mu]^T) \tag{B.11}$$

$$= \sum_{i=1}^{n}(s_i - E[\mu])(s_i - E[\mu])^T + n \cdot \mathrm{cov}(\mu). \tag{B.12}$$

This leads to

$$q_R(R) = \mathcal{IW}(R; \Psi^*, \nu^*), \tag{B.13}$$

with

$$\nu^* = \nu + n \tag{B.14}$$

and

$$\Psi^* = \Psi + \sum_{i=1}^{n}(s_i - E[\mu])(s_i - E[\mu])^T + n \cdot \mathrm{cov}(\mu). \tag{B.15}$$

Finally, we replace the moments in (B.6), (B.7) and (B.15) by the moments from the approximate distributions q_μ and q_R. Since R, according to (B.13), is inverse-Wishart distributed with scale matrix Ψ^* and $\nu^* = \nu + n$ degrees of freedom, its inverse R^{-1} is Wishart-distributed with scale matrix Ψ^{*-1} and $\nu + n$ degrees of freedom. Its expectation is $E[R^{-1}] = \nu^*\Psi^{*-1}$. Therefore we find

$$\mu_0^* = (\Lambda^{-1} + n(\nu + n)\Psi^{*-1})^{-1}(\Lambda^{-1}\mu_0 + n(\nu + n)\Psi^{*-1}\bar{s}) \tag{B.16}$$

and

$$\Lambda^* = (\Lambda^{-1} + n(\nu + n)\Psi^{*-1})^{-1}. \tag{B.17}$$

By (B.5), the mean and covariace of μ is μ_0^* and Λ^*, therefore (B.15) becomes

$$\Psi^* = \Psi + \sum_{i=1}^{n}(s_i - \mu_0^*)(s_i - \mu_0^*)^T + n\Lambda^*. \tag{B.18}$$

Appendix C. Efficient computation of the update iteration

The key to finding the estimated mean and covariance matrix for a patient is iteration over the update equations, repeated here for convenience:

$$\Lambda^* = (\Lambda^{-1} + n(\nu + n)\Psi^{*-1})^{-1} \tag{C.1}$$

$$\mu_0^* = \Lambda^*(\Lambda^{-1}\mu_0 + n(\nu + n)\Psi^{*-1}\bar{s}) \tag{C.2}$$

$$\Psi^* = \Psi + \sum_{j=1}^{n}(s_j - \mu_0^*)(s_j - \mu_0^*)^T + n\Lambda^* \tag{C.3}$$

$$\nu^* = \nu + n \tag{C.4}$$

Only ν^* can be calculated directly. The other parameters rely on each other, and therefore require an iteration to converge to the correct values.

Putting the iteration number i in a superscript (replacing $\cdot^*$), we can write the iteration as

$$\Lambda^{(i)} = (\Lambda^{-1} + n(\nu + n)\Psi^{(i-1)-1})^{-1} \tag{C.5}$$

$$\mu_0^{(i)} = \Lambda^{(i)}(\Lambda^{-1}\mu_0 + n(\nu + n)\Psi^{(i-1)-1}\bar{s}) \tag{C.6}$$

$$\Psi^{(i)} = \Psi + \sum_{j=1}^{n}(s_j - \mu_0^{(i)})(s_j - \mu_0^{(i)})^T + n\Lambda^{(i)} \tag{C.7}$$

We can see that we need to supply a starting guess for the first value $\Psi^{(0)}$.[5] A natural starting guess is $\Psi^{(0)} = \Psi$.

In theory, the iteration represented by equations (C.5)–(C.7) can be implemented directly in any numerically oriented programming language. However, this would require storing and inverting very large $P \times P$ matrices, which is not attainable in practice. However, due to the structure of Ψ and Λ (when estimated as in sections 2.4.2 and 2.4.3), memory and computation requirements can be drastically reduced.

Both matrices Λ and Ψ can be represented as an outer product of a data matrix with itself plus a scalar multiple of the identity matrix:

$$\Lambda = D_\Lambda D_\Lambda^T + \delta_\Lambda I \tag{C.8}$$

$$\Psi = D_\Psi D_\Psi^T + \delta_\Psi I. \tag{C.9}$$

Here, D_Ψ and D_Λ are $P \times N_\Psi$ and $P \times N_\Lambda$ matrices, with $N_\Lambda, N_\Psi \ll P$. Multiplying a vector a by such a matrix is much faster than the general $O(P^2)$ figure, since e. g.

$$\Lambda a = (D_\Lambda D_\Lambda^T + \delta_\Lambda I)a = D_\Lambda(D_\Lambda^T a) + \delta_\Lambda a, \tag{C.10}$$

which is easily computed in $O(N_\Lambda P)$ time. Furthermore, it is also fast to solve an equation such as $\Lambda x = b$.

Throughout this derivation we shall make heavy use of the following special case of the *Woodbury matrix identity*, which holds for any matrices A and B and scalar δ as long as the involved inversions are possible:

$$(\delta I + ABA^T)^{-1} = \delta^{-1}I - \delta^{-1}A(\delta B^{-1} + A^T A)^{-1}A^T. \tag{C.11}$$

This means that the inverses of Λ and Ψ can also be written in the form $DCD^T + \delta I$ for some D, C and δ.

C.1. Computing $\Lambda^{(i)}$

We shall show later that $\Psi^{(i)}$ can be written for any i as

$$\Psi^{(i)} = D^{(i)}G^{(i)}D^{(i)T} + \delta_\Psi^{(i)} I, \tag{C.12}$$

for some $\delta_\Psi^{(i)}$ and $G^{(i)}$, and where

$$D^{(i)} = [D_\Lambda \quad D_\Psi^{(i)}] \tag{C.13}$$

for some $D_\Psi^{(i)}$ of dimension $P \times (N_\Psi + n)$. Inserting (C.8) and (C.12) into (C.5), we get

$$\Lambda^{(i)} = [(D_\Lambda D_\Lambda^T + \delta_\Lambda I)^{-1} + n(\nu + n)(D^{(i-1)}G^{(i-1)}D^{(i-1)T} + \delta_\Psi^{(i-1)}I)^{-1}]^{-1}. \tag{C.14}$$

Using the matrix inversion lemma (C.11) on both the inner inverses of (C.14), we get

$$\begin{aligned}\Lambda^{(i)} = [&\delta_\Lambda^{-1}I - \delta_\Lambda^{-1}D_\Lambda(\delta_\Lambda I + D_\Lambda^T D_\Lambda)^{-1}D_\Lambda^T + n(\nu + n)\delta_\Psi^{(i-1)-1}I \\ &- n(\nu + n)\delta_\Psi^{(i-1)-1}D^{(i-1)}(\delta_\Psi^{(i-1)}G^{(i-1)-1} + D^{(i-1)T}D^{(i-1)})^{-1}D^{(i-1)T}]^{-1}\end{aligned} \tag{C.15}$$

In order to group the terms, note that

$$-\delta_\Lambda^{-1}D_\Lambda(\delta_\Lambda I + D_\Lambda^T D_\Lambda)^{-1}D_\Lambda^T = D^{(i-1)}QD^{(i-1)T}, \tag{C.16}$$

where Q is a block-diagonal matrix

$$Q = \begin{bmatrix} -\delta_\Lambda^{-1}(\delta_\Lambda I + D_\Lambda^T D_\Lambda)^{-1} & \\ & 0_{N_\Psi + n \times N_\Psi + n} \end{bmatrix}. \tag{C.17}$$

We also define

$$L^{(i)} = -\delta_\Psi^{(i)-1}(\delta_\Psi^{(i)}G^{(i)-1} + D^{(i)T}D^{(i)})^{-1} \tag{C.18}$$

and

$$F^{(i)} = Q + n(\nu + n)L^{(i-1)} \tag{C.19}$$

Now, we can write

$$\Lambda^{(i)} = (\delta_\Lambda^{-1} + n(\nu + n)\delta_\Psi^{(i-1)-1})I + D^{(i-1)}F^{(i)}D^{(i-1)T})^{-1}, \tag{C.20}$$

Applying the matrix inversion lemma again, we find

$$\Lambda^{(i)} = \delta^{(i)}I - \delta^{(i)}D^{(i-1)}H^{(i)}D^{(i-1)T}, \tag{C.21}$$

[5] Given that the iteration starts with the equation for $\Lambda^{(1)}$. If we had started with one of the other equations, a starting guess for at least one other parameter would need to be provided.

where

$$H^{(i)} = [D^{(i-1)T}D^{(i-1)} + \delta^{(i)-1}F^{(i)-1}]^{-1} \tag{C.22}$$

and

$$\delta^{(i)} = (\delta_\Lambda^{-1} + n(\nu + n)\delta_\Psi^{(i-1)-1})^{-1}. \tag{C.23}$$

equation (C.21) gives us an expression for $\Lambda^{(i)}$ using only lower dimensional matrices and scalars. In practice, we never construct $\Lambda^{(i)}$—it is represented implicitly by $D^{(i)}$, $H^{(i)}$ and $\delta^{(i)}$ through (C.21).

C.2. Computing $\mu_0^{(i)}$

Through the derivation of $\Lambda^{(i)}$, we have already come a long way towards computing $\mu_0^{(i)}$. We can write (C.6) as

$$\mu_0^{(i)} = \Lambda^{(i)}r^{(i)}, \tag{C.24}$$

with

$$r^{(i)} = \Lambda^{-1}\mu_0 + n(\nu + n)\Psi^{(i-1)-1}\bar{s}. \tag{C.25}$$

The first term of (C.25) is constant, and can be computed once. Using the matrix inversion lemma on (C.8), we find

$$\Lambda^{-1}\mu_0 = \delta_\Lambda^{-1}\mu_0 - \delta_\Lambda^{-1}D_\Lambda(\delta_\Lambda I + D_\Lambda^T D_\Lambda)^{-1}(D_\Lambda\mu_0) \tag{C.26}$$

The last term needs to be computed for each iteration. We find it by using the matrix inversion lemma on (C.12):

$$\Psi^{(i)-1}\bar{s} = (\delta_\Psi^{(i)-1}I - \delta_\Psi^{(i)-1}D^{(i)}(\delta_\Psi^{(i)}G^{(i)-1} + D^{(i)T}D^{(i)})^{-1}D^{(i)T})\bar{s} \tag{C.27}$$

$$= \delta_\Psi^{(i)-1}\bar{s} + D^{(i)}L^{(i)}(D^{(i)T}\bar{s}) \tag{C.28}$$

Finally, inserting (C.21) into (C.24), we find

$$\mu_0^{(i)} = \delta^{(i)}r^{(i)} - \delta^{(i)}D^{(i-1)}H^{(i)}(D^{(i-1)T}r^{(i)}). \tag{C.29}$$

C.3. Computing $\Psi^{(i)}$

The update equation for Ψ is

$$\begin{aligned}\Psi^{(i)} &= \Psi + \sum_{j=1}^{n}(s_j - \mu_0^{(i)})(s_j - \mu_0^{(i)})^T + n\Lambda^{(i)} \\ &= D_\Psi D_\Psi^T + \delta_\Psi I + \sum_{j=1}^{n}(s_j - \mu_0^{(i)})(s_j - \mu_0^{(i)})^T + n\Lambda^{(i)}.\end{aligned} \tag{C.30}$$

We can augment the data matrix D_Ψ by inserting new columns which are the mean-subtracted data vectors;

$$D_\Psi^{(i)} = [D_\Psi \quad s_1 - \mu_0^{(i)} \quad s_2 - \mu_0^{(i)} \quad \dots \quad s_n - \mu_0^{(i)}], \tag{C.31}$$

and we find

$$\Psi^{(i)} = D_\Psi^{(i)}D_\Psi^{(i)T} + \delta_\Psi I + n\Lambda^{(i)}. \tag{C.32}$$

Inserting (C.21), we get

$$\Psi^{(i)} = D_\Psi^{(i)}D_\Psi^{(i)T} + \delta_\Psi I + n(\delta^{(i)}I - \delta^{(i)}D^{(i-1)}H^{(i)}D^{(i-1)T}) \tag{C.33}$$

We want to group the terms of this equation, but run into a slight problem: One term contains $D_\Psi^{(i)}$, while another term contains $D^{(i-1)}$ (which contains $D_\Psi^{(i-1)}$). In practice, this can easily be resolved by replacing $D^{(i-1)}$ by $D^{(i)}$; this is in line with the algorithm philosophy of always using the most recent guess of each parameter, and also guarantees that the equations (C.1)–(C.3) hold at convergence (at convergence, we have $D^{(i)} = D^{(i-1)}$). Now, to group the terms, first note that

$$D_\Psi^{(i)}D_\Psi^{(i)T} = D^{(i)}KD^{(i)T}, \tag{C.34}$$

where

$$K = \begin{bmatrix} 0_{N_\Lambda \times N_\Lambda} & \\ & I_{N_\Psi + n} \end{bmatrix}. \tag{C.35}$$

Thus, we can write

$$\Psi^{(i)} = D^{(i)}(K - n\delta^{(i)}H^{(i)})D^{(i)T} + (\delta_\Psi + n\delta^{(i)})I. \tag{C.36}$$

We now see that we must have

$$G^{(i)} = K - n\delta^{(i)}H^{(i)} \tag{C.37}$$

and

$$\delta_{\Psi}^{(i)} = \delta_{\Psi} + n\delta^{(i)} \tag{C.38}$$

in order for $\Psi^{(i)}$ to be written as

$$\Psi^{(i)} = D^{(i)}G^{(i)}D^{(i)T} + \delta_{\Psi}^{(i)}I. \tag{C.39}$$

C.4. Initial values

Initially, we want to get $\Psi^{(0)} = \Psi$, i. e. $D^{(0)}G^{(0)}D^{(0)T} + \delta_{\Psi}^{(0)}I = D_{\Psi}D_{\Psi}^{T} + \delta_{\Psi}I$ which achieve by setting

$$\delta_{\Psi}^{(0)} = \delta_{\Psi} \tag{C.40}$$

$$D_{\Psi}^{(0)} = [D_{\Psi} \quad 0_{P\times n}] \tag{C.41}$$

$$G^{(0)} = K. \tag{C.42}$$

However, $G^{(0)}$ is not invertible, which makes it impossible to compute $L^{(0)}$ as in (C.18). Instead, $L^{(0)}$ must be initialized to

$$L^{(0)} = \begin{bmatrix} 0_{N_\Lambda\times N_\Lambda} & \\ & -\delta_{\Psi}^{-1}(\delta_{\Psi}I + D_{\Psi}^{(i)T}D_{\Psi}^{(i)})^{-1} \end{bmatrix}. \tag{C.43}$$

C.5. Algorithm summary

Input: $\mu_0, D_\Lambda, D_\Psi, \delta_\Psi, \delta_\Lambda, s_1 \ldots s_n, \nu$
Output: $\mu_0^*, D^*, G^*, \delta_\Psi^*, \delta^*, H^*$

$K = \begin{bmatrix} 0_{N_\Lambda\times N_\Lambda} & \\ & I_{N_\Psi+n} \end{bmatrix}$.

$Q = \begin{bmatrix} -\delta_\Lambda^{-1}(\delta_\Lambda I + D_\Lambda^T D_\Lambda)^{-1} & \\ & 0_{N_\Psi+n\times N_\Psi+n} \end{bmatrix}$

$q \leftarrow \delta_\Lambda^{-1}\mu_0 - \delta_\Lambda^{-1}D_\Lambda(\delta_\Lambda I + D_\Lambda^T D_\Lambda)^{-1}(D_\Lambda^T\mu_0)$ /*q is $\Lambda^{-1}\mu_0$ */

$D_\Psi^{(0)} = [D_\Psi \quad 0_{P\times n}]$

$D^{(0)} \leftarrow [D_\Lambda \quad D_\Psi^{(0)}]$

$\delta_\Psi^{(0)} \leftarrow \delta_\Psi$

$\mu_0^{(0)} \leftarrow \mu_0$

$L^{(0)} \leftarrow \begin{bmatrix} 0_{N_\Lambda\times N_\Lambda} & \\ & -\delta_\Psi^{-1}(\delta_\Psi I + D_\Psi^{(0)T}D_\Psi^{(0)})^{-1} \end{bmatrix}$

$i \leftarrow 0$

repeat

$i \leftarrow i + 1$

$\delta^{(i)} \leftarrow (\delta_\Lambda^{-1} + n(\nu+n)\delta_\Psi^{(i-1)-1})^{-1}$

$F^{(i)} \leftarrow Q + n(\nu+n)L^{(i-1)}$

$H^{(i)} \leftarrow (D^{(i-1)T}D^{(i-1)} + \delta^{(i)-1}F^{(i)-1})^{-1}$

$r^{(i)} \leftarrow q + n(\nu+n)(\delta_\Psi^{(i-1)-1}I + D^{(i-1)}L^{(i-1)}D^{(i-1)T})\bar{s}$

$\mu_0^{(i)} \leftarrow \delta^{(i)}r^{(i)} - \delta^{(i)}D^{(i-1)}H^{(i)}(D^{(i-1)T}r^{(i)})$

$D_\Psi^{(i)} \leftarrow [D_\Psi \quad s_1 - \mu_0^{(i)} \quad s_2 - \mu_0^{(i)} \quad \ldots \quad s_n - \mu_0^{(i)}]$

$D^{(i)} \leftarrow [D_\Lambda \quad D_\Psi^{(i)}]$

$\delta_\Psi^{(i)} \leftarrow \delta_\Psi + n\delta^{(i)}$

$G^{(i)} \leftarrow K - n\delta^{(i)}H^{(i)}$

$L^{(i)} \leftarrow -\delta_\Psi^{(i)-1}(\delta_\Psi^{(i)}G^{(i)-1} + D^{(i)T}D^{(i)})^{-1}$

until $\|\mu_0^{(i)} - \mu_0^{(i-1)}\| < \epsilon$

$\mu_0^* \leftarrow \mu^{(i)}, D^* \leftarrow D^{(i)}, G^* \leftarrow G^{(i)}, \delta_\Psi^* \leftarrow \delta_\Psi^{(i)}, \delta^* \leftarrow \delta^{(i)}, H^* \leftarrow H^{(i)}$

/* Implicit, not computed: $\Lambda^* = \delta^* I - \delta^* D^* H^* D^{*T}$ */

/* Implicit, not computed: $\Psi^* = D^* G^* D^{*T} + \delta_\Psi^* I^*$ */

Appendix D. Bias of the inter-patient covariance matrix estimate

We estimate the inter-patient covariance matrix as

$$\hat{\Lambda} = \frac{1}{M-1}\sum_{i=1}^{M}(\bar{s}_i - \hat{\mu}_0)(\bar{s}_i - \hat{\mu}_0)^T. \tag{D.1}$$

This is the sample covariance matrix of $\bar{s}_i$, as opposed to μ which we are interested in. But $\bar{s}_i$ are not identically distributed if J_i varies. We can show that

$$E[\hat{\Lambda}] = \frac{1}{M}\sum_{i=1}^{M}\mathrm{cov}(\bar{s}_i). \tag{D.2}$$

To avoid clutter, the proof of this result is given at the end of the appendix.

The covariance matrix of a sample mean based on n i.i.d. samples is always given by $1/n$ times the covariance matrix of one sample. In other words,

$$\mathrm{cov}(\bar{s}_i|\mu, R) = \frac{1}{J_i}R. \tag{D.3}$$

Now we can use the *law of total covariance*, which states, for two scalar random variables a and b,

$$\mathrm{cov}(a, b) = E[\mathrm{cov}(a, b|c)] + \mathrm{cov}(E[a|c], E[b|c]). \tag{D.4}$$

In our case, we get

$$\mathrm{cov}(\bar{s}_i) = E[\mathrm{cov}(\bar{s}_i|\mu, R)] + \mathrm{cov}(E[\bar{s}_i|\mu, R]) \tag{D.5}$$

$$= E\left[\frac{1}{J_i}R\right] + \mathrm{cov}(\mu) \tag{D.6}$$

$$= \frac{1}{J_i}E[R] + \Lambda \tag{D.7}$$

since, by definition, $\mathrm{cov}(\mu) = \Lambda$. Inserting (D.7) into (D.2) yields

$$E[\hat{\Lambda}] = \frac{1}{M}\sum_{i=1}^{M}\left(\Lambda + \frac{1}{J_i}E[R]\right) = \Lambda + cE[R], \tag{D.8}$$

where

$$c = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{J_i}. \tag{D.9}$$

□

Proof of (D.2):

We start by manipulating (D.1):

$$\hat{\Lambda} = \frac{1}{M-1}\sum_{i=1}^{M}(\bar{s}_i - \hat{\mu}_0)(\bar{s}_i - \hat{\mu}_0)^T \tag{D.10}$$

$$= \frac{1}{M-1}\left(\sum_{i=1}^{M}\bar{s}_i\bar{s}_i^T + \sum_{i=1}^{M}\hat{\mu}_0\hat{\mu}_0^T - \sum_{i=1}^{M}\bar{s}_i\hat{\mu}_0^T - \sum_{i=1}^{M}\hat{\mu}_0\bar{s}_i^T\right) \tag{D.11}$$

$$= \frac{1}{M-1}\left(\sum_{i=1}^{M}\bar{s}_i\bar{s}_i^T + \sum_{i=1}^{M}\hat{\mu}_0\hat{\mu}_0^T - (\sum_{i=1}^{M}\bar{s}_i)\hat{\mu}_0^T - \hat{\mu}_0(\sum_{i=1}^{M}\bar{s}_i^T)\right) \tag{D.12}$$

$$= \frac{1}{M-1}\left(\sum_{i=1}^{M}\bar{s}_i\bar{s}_i^T + M\hat{\mu}_0\hat{\mu}_0^T - M\hat{\mu}_0\hat{\mu}_0^T - M\hat{\mu}_0\hat{\mu}_0^T\right) \tag{D.13}$$

$$= \frac{1}{M-1}\left(\sum_{i=1}^{M}\bar{s}_i\bar{s}_i^T - M\hat{\mu}_0\hat{\mu}_0^T\right), \tag{D.14}$$

where we used $\hat{\mu}_0 = \frac{1}{M}\sum_{i=1}^{M}\bar{s}_i$. Taking the expectation, and using the general formula $E[xx^T] = \mathrm{cov}(x) + E[x]E[x]^T$, we find

$$E[\hat{\Lambda}] = \frac{1}{M-1}\left(\sum_{i=1}^{M}E[\bar{s}_i\bar{s}_i^T] - ME[\hat{\mu}_0\hat{\mu}_0^T]\right) \tag{D.15}$$

$$= \frac{1}{M-1}\sum_{i=1}^{M}(\mathrm{cov}(\bar{s}_i) + E[\bar{s}_i]E[\bar{s}_i]^T) - \frac{M}{M-1}(\mathrm{cov}(\hat{\mu}_0) + E[\hat{\mu}_0]E[\hat{\mu}_0]^T) \tag{D.16}$$

$$= \frac{1}{M-1}\left(\sum_{i=1}^{M}\mathrm{cov}(\bar{s}_i) - M\mathrm{cov}(\hat{\mu}_0)\right), \tag{D.17}$$

since $E[\bar{s}_i] = E[E[\bar{s}_i|\mu]] = E[\mu] = \mu_0 = E[\hat{\mu}_0]$. Looking at $\mathrm{cov}(\hat{\mu}_0)$, we find

$$\mathrm{cov}(\hat{\mu}_0) = \mathrm{cov}\left(\frac{1}{M}\sum_{i=1}^{M}\bar{s}_i\right) \tag{D.18}$$

$$= \frac{1}{M^2}\sum_{i=1}^{M}\mathrm{cov}(\bar{s}_i), \tag{D.19}$$

since $\bar{s}_i$ are independent (though not identically distributed). Inserting (D.19) into (D.17) yields

$$E[\hat{\Lambda}] = \frac{1}{M-1}\left(\sum_{i=1}^{M}\mathrm{cov}(\bar{s}_i) - \frac{M}{M^2}\sum_{i=1}^{M}\mathrm{cov}(\bar{s}_i)\right) \tag{D.20}$$

$$= \frac{1}{M-1}\left(1 - \frac{1}{M}\right)\sum_{i=1}^{M}\mathrm{cov}(\bar{s}_i) \tag{D.21}$$

$$= \frac{1}{M}\sum_{i=1}^{M}\mathrm{cov}(\bar{s}_i). \tag{D.22}$$

□

Appendix E. PCA for the bias-corrected inter-patient covariance matrix

The bias-corrected inter-patient covariance matrix estimate is given by

$$\tilde{\Lambda} = \hat{\Lambda} - c\hat{R}_{\mathrm{pop}}, \tag{E.1}$$

where $c = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{J_i}$. This matrix is not positive semidefinite, and cannot be expressed with a real-valued data matrix D as $\tilde{\Lambda} = DD^T$. It can, however, be expressed as

$$\tilde{\Lambda} = AB^T, \tag{E.2}$$

where $A = [D_\Lambda \quad \sqrt{c}D_{\mathrm{pop}}]$ and $B = [D_\Lambda \quad -\sqrt{c}D_{\mathrm{pop}}]^T$.

As usual $\tilde{\Lambda}$ is too big to practically perform eigenvalue decompostion on. However, there is a relation between the eigenvalue decomposition of AB^T and that of B^{TA}. The latter is a small matrix, and its eigenvalue decomposition can easily computed using any numerical software package. Given the kth eigenvalue λ_k and the kth eigenvector v_k of B^{TA}, the kth eigenvalue of $\tilde{\Lambda}$ is λ_k, and the kth eigenvector is

$$w_k = Av_k. \tag{E.3}$$

A proof of this result is given at the end of the appendix. The scale of w_k is arbitrary, so we want to normalize it as

$$w_k' = \frac{w_k}{\|w_k\|}. \tag{E.4}$$

As usual in PCA, we discard the eigenpairs corresponding to the smallest eigenvalues. In this case, since the matrix is not positive semidefinite, several of the eigenvalues will be negative. We need to discard all eigenpairs corresponding to negative eigenvalues, since we cannot have negative variance for any of the modes (which would lead to a complex data matrix). The PCA-reduced covariance matrix can now be represented by a data matrix $\tilde{D}_{PCA}$ as

$$\tilde{\Lambda}_{PCA} = \tilde{D}_{PCA}\tilde{D}_{PCA}^T, \tag{E.5}$$

with

$$\tilde{D}_{PCA} = [\sqrt{\lambda_1}w_1' \quad \sqrt{\lambda_2}w_2' \quad \ldots \quad \sqrt{\lambda_K}w_K']. \tag{E.6}$$

Proof of (E.3):

Let w_k be an eigenvector of AB^T, and λ_k be the corresponding eigenvalue, i. e.

$$AB^T w_k = \lambda_k w_k. \tag{E.7}$$

We can transform AB^T into B^{TA} by what we may call a pseudo-similarity transformation:

$$A^+(AB^T)A = (A^T A)^{-1}A^T AB^T A = B^T A, \tag{E.8}$$

where A^+ denotes the pseudo-inverse of A. Also note that AA^+ is a projection matrix onto the subspace spanned by A. Since w_k, as an eigenvector of AB^T, is already in this subspace, we have

$$AA^+ w_k = w_k. \tag{E.9}$$

Using the three previous equations, we can now write

$$B^T A(A^+ w_k) = A^+ AB^T AA^+ w_k = A^+ AB^T w_k = \lambda_k A^+ w_k. \tag{E.10}$$

This shows that λ_k is an eigenvalue of B^{TA}, with corresponding eigenvector $v_k = A^+ w_k$. However, we want to find w_k given v_k. Using (E.9) again, we find

$$v_k = A^+ w_k \tag{E.11}$$

$$\rightarrow Av_k = AA^+ w_k = w_k. \tag{E.12}$$

□

ORCID iDs

Øyvind Lunde Rørtveit https://orcid.org/0000-0001-6545-663X
Liv Bolstad Hysing https://orcid.org/0000-0002-7593-7549
Sara Pilskog https://orcid.org/0000-0002-3475-7939

References

Baum C, Alber M, Birkner M and Nüsslin F 2006 Robust treatment planning for intensity modulated radiotherapy of prostate cancer based on coverage probabilities *Radiother. Oncol.* **78** 27–35

Bishop C M 2006 Pattern recognition and machine learning *Information Science and Statistics* (New York: Springer)

Bodnar T, Mazur S and Podgórski K 2016 Singular inverse Wishart distribution and its application to portfolio theory *J. Multivariate Anal.* **143** 314–26

Bondar L, Intven M, Burbach J P M, Budiarto E, Kleijnen J P, Philippens M, van Asselen B, Seravalli E, Reerink O and Raaymakers B 2014 Statistical modeling of CTV motion and deformation for IMRT of early-stage rectal cancer *Int. J. Radiat. Oncol.***Biol.***Phys.* **90** 664–72

Budiarto E, Keijzer M, Storchi P R, Hoogeman M S, Bondar L, Mutanga T F, de Boer H C J and Heemink A W 2011 A population-based model to describe geometrical uncertainties in radiotherapy: applied to prostate cases *Phys. Med. Biol.* **56** 1045–61

Cook R D and Forzani L 2011 On the mean and variance of the generalized inverse of a singular Wishart matrix *Electron. J. Stat.* **5** 146–58

Fontenla E, Pelizzari C A, Roeske J C and Chen G T Y 2001 Using serial imaging data to model variabilities in organ position and shape during radiotherapy *Phys. Med. Biol.* **46** 2317–36

Fujikoshi Y, Ulyanov V V and Shimizu R 2010 *Multivariate Statistics : High-Dimensional and Large-Sample Approximations* (Hoboken, UNITED STATES: John Wiley & Sons, Incorporated)

Gelman A, Carlin J B, Stern H S and Rubin D B 1995 *Bayesian Data Analysis* (Chapman and Hall/CRC)

Herschtal A, Foroudi F, Greer P B, Eade T N, Hindson B R and Kron T 2012 Finding the optimal statistical model to describe target motion during radiotherapy deliverya Bayesian approach *Phys. Med. Biol.* **57** 2743–55

Hysing L B, Ekanger C, Zolnay A, Helle S I, Rasi M, Heijmen B J M, Sikora M, Söhn M, Muren L P and Thörnqvist S 2018 Statistical motion modelling for robust evaluation of clinically delivered accumulated dose distributions after curative radiotherapy of locally advanced prostate cancer *Radiother. Oncol.: J. Eur. Soc. Therapeutic Radiol. Oncol.* **128** 327–35

Hysing L B, Söhn M, Muren L P and Alber M 2011 A coverage probability based method to estimate patient-specific small bowel planning volumes for use in radiotherapy *Radiother. Oncol: J. Eur. Soc. Therapeutic Radiol. Oncol.* **100** 407–11

Lam K L, Ten Haken R K, Litzenberg D, Balter J M and Pollock S M 2005 An application of Bayesian statistical methods to adaptive radiotherapy *Phys. Med. Biol.* **50** 3849–58

Lindegaard J C, Assenholt M, Ramlov A, Fokdal L U, Alber M and Tanderup K 2017 Early clinical outcome of coverage probability based treatment planning for simultaneous integrated boost of nodes in locally advanced cervical cancer *Acta Oncol. (Stockholm, Sweden)* **56** 1479–86

Magallon-Baro A, Loi M, Milder M T W, Granton P V, Zolnay A G, Nuyttens J J and Hoogeman M S 2019 Modeling daily changes in organ-at-risk anatomy in a cohort of pancreatic cancer patients *Radiother. Oncol.* **134** 127–34

Owens C A *et al* 2022 Development and validation of a population-based anatomical colorectal model for radiation dosimetry in late effects studies of survivors of childhood cancer *Radiother. Oncol.* **176** 118–126

Price G J and Moore C J 2007 A method to calculate coverage probability from uncertainties in radiotherapy via a statistical shape model *Phys. Med. Biol.* **52** 1947–65

Ramlov A, Assenholt M S, Jensen M F, Grønborg C, Nout R, Alber M, Fokdal L, Tanderup K and Lindegaard J C 2017 Clinical implementation of coverage probability planning for nodal boosting in locally advanced cervical cancer *Radiother. Oncol.* **123** 158–63

Rigaud B *et al* 2019 Statistical shape model to generate a planning library for cervical adaptive radiotherapy *IEEE Trans. Med. Imaging* **38** 406–16

Rios R, De Crevoisier R, Ospina J D, Commandeur F, Lafond C, Simon A, Haigron P, Espinosa J and Acosta O 2017 Population model of bladder motion and deformation based on dominant eigenmodes and mixed-effects models in prostate cancer radiotherapy *Med. Image Anal.* **38** 133–49

Rørtveit Ø L, Hysing L B, Stordal A S and Pilskog S 2021 Reducing systematic errors due to deformation of organs at risk in radiotherapy *Med. Phys.* **48** 6578–87

Sanguineti G, Faiella A, Farneti A, D'Urso P, Fuga V, Olivieri M, Giannarelli D, Marzi S, Iaccarino G and Landoni V 2020 Refinement & validation of rectal wall dose volume objectives for prostate hypofractionation in 20 fractions *Clin. Transl. Radiat. Oncol.* **21** 91–7

Sobotta B, Söhn M and Alber M 2010 Robust optimization based upon statistical theory *Med. Phys.* **37** 4019–28

Söhn M, Birkner M, Yan D and Alber M 2005 Modelling individual geometric variation based on dominant eigenmodes of organ deformation: implementation and evaluation *Phys. Med. Biol.* **50** 5893–908

Söhn M, Sobotta B and Alber M 2012 Dosimetric treatment course simulation based on a statistical model of deformable organ motion *Phys. Med. Biol.* **57** 3693–709

Stroom J C, de Boer H C, Huizenga H and Visser A G 1999 Inclusion of geometrical uncertainties in radiotherapy treatment planning by means of coverage probability *Int. J. Radiat. Oncol. Biol. Phys.* **43** 905–19

Szeto Y Z, Witte M G, van Herk M and Sonke J J 2017 A population based statistical model for daily geometric variations in the thorax *Radiother. Oncol.* **123** 99–105

Thörnqvist S, Hysing L B, Zolnay A G, Söhn M, Hoogeman M S, Muren L P, Bentzen L and Heijmen B J M 2013a Treatment simulations with a statistical deformable motion model to evaluate margins for multiple targets in radiotherapy for high-risk prostate cancer *Radiother. Oncol.* **109** 344–9

Thörnqvist S, Hysing L B, Zolnay A G, Söhn M, Hoogeman M S, Muren L P and Heijmen B J M 2013b Adaptive radiotherapy in locally advanced prostate cancer using a statistical deformable motion model *Acta Oncol. (Stockholm, Sweden)* **52** 1423–9

Timmerman M E and Kiers H A L 2003 Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences *Psychometrika* **68** 105–21

Tipping M E and Bishop C M 1999 Probabilistic principal component analysis *J. R. Stat. Soc.* B **61** 611–22

Unkelbach J *et al* 2018 Robust radiotherapy planning *Phys. Med. Biol.* **63** 22TR02

van Herk M, Remeijer P, Rasch C and Lebesque J V 2000 The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy *Int. J. Radiat. Oncol. *Biol. *Phys.* **47** 1121–35

Yan D, Vicini F, Wong J and Martinez A 1997 Adaptive radiation therapy *Phys. Med. Biol.* **42** 123–32