Published in final edited form as:

Biometrics. 2014 June ; 70(2): 366–377. doi:10.1111/biom.12146.

Evaluation of Image Registration Spatial Accuracy Using a Bayesian Hierarchical Model

Suyu Liu^{1,*}, Ying Yuan^{1,**}, Richard Castillo², Thomas Guerrero², and Valen E. Johnson^{3,***}

¹Department of Biostatistics, MD Anderson Cancer Center, Houston, Texas, U.S.A.

²Department of Radiation Oncology, MD Anderson Cancer Center, Houston, Texas, U.S.A.

³ Department of Statistics, Texas A&M University, Texas, U.S.A.

Summary

To evaluate the utility of automated deformable image registration (DIR) algorithms, it is necessary to evaluate both the registration accuracy of the DIR algorithm itself, as well as the registration accuracy of the human readers from whom the "gold standard" is obtained. We propose a Bayesian hierarchical model to evaluate the spatial accuracy of human readers and automatic DIR methods based on multiple image registration data generated by human readers and automatic DIR methods. To fully account for the locations of landmarks in all images, we treat the true locations of landmarks as latent variables and impose a hierarchical structure on the magnitude of registration errors observed across image pairs. DIR registration errors are modeled using Gaussian processes with reference prior densities on prior parameters that determine the associated covariance matrices. We develop a Gibbs sampling algorithm to efficiently fit our models to high-dimensional data, and apply the proposed method to analyze an image dataset obtained from a 4D thoracic CT study.

Keywords

Image processing; latent variable; Bayesian analysis; spatial correlation

1. Introduction

Image registration is a common and necessary task in medical diagnostic imaging. The goal of image registration is to determine the spatial correspondence between anatomical structures in a source image and similar structures in one or more target images. Image pairs may correspond to images of the same structures collected at distinct points in time, for instance of a tumor volume imaged before and after therapeutic treatment, or of the same anatomical structure imaged at different points in, say, the cardiac or respiratory cycles. In cancer clinical trials, the Response Evaluation Criteria in Solid Tumors (RECIST) evaluation of tumor response requires an estimate of changes in tumor volume, which can

*syliu@mdanderson.org. **yyuan@mdanderson.org. ***vjohnson@stat.tamu.edu.

8. Supplementary materials

Web Appendix referenced in Section 3 and the program code used to analyze the data are available with this paper at the Biometrics website on Wiley Online Library.

(in theory) be automated using deformable image registration (DIR) algorithms (Padhani and Ollivier, 2001). Image registration is also commonly used in the administration and management of radiation therapy for thoracic malignancies. It allows radiation oncologists to account for the cyclical changes in thoracic anatomy that are associated with the respiratory cycle (Hof et al. 2003; Keall 2004). A recent review of DIR methods used to register thoracic CT images in radiotherapy applications is provided in Sarrut (2006).

The fundamental difficulty associated with the objective evaluation of DIR algorithms for use in clinical settings arises from the lack of a ground truth. Although biomechanical computer simulations and mechanical phantoms provide a means of assessing DIR algorithms in which the spatial transformation between images is known precisely (Schnabel et al. 2003; Guerrero et al. 2004), such simulations and mechanical phantoms lack sufficient realism to provide a credible measure of DIR performance in actual clinical applications. As a consequence of this fact, recent publications on the evaluation of DIR methods reflect a preference for the use of landmark point pairs to evaluate spatial accuracy (Kaus et al. 2007; Al-Mayah et al. 2008; Boldea et al. 2008; Li et al. 2008; Wolthaus et al. 2008; Wu et al. 2008; Castillo et al. 2009; Castillo et al. 2010; Gu et al. 2010). Using the approach of landmark point pairs, expert readers visually identify pairs of “identical” landmarks in both a source and target image. The accuracy of a DIR algorithm is then determined by comparing the mapped locations of each landmark in the source image to the target location identified by the expert reader. Vandemeulebroucke et al. (2007) reported on the public availability of a reference standard dataset, the POPI model, which consists of 41 centrally located lung landmarks identified on each component phase image of a clinically acquired 4D CT image set. For each reference landmark feature, they defined “ground truth” to be the average spatial coordinate pair defined by two independent observers. A drawback of the POPI dataset is the limited number of landmarks and their limited spatial distribution. Kabus et al. (2009) argued that these limitations could result in systematic errors in the assessment of the spatial accuracy of DIR algorithms, concluding that well-distributed landmarks are necessary for reliable registration accuracy assessment. To address this issue, Castillo et al. (2009) proposed a framework for systematically identifying and selecting landmark features and utilizing the reference landmark point pairs for objective evaluation of DIR algorithms.

The work of Castillo et al. (2009) focuses on developing a practical framework and operative procedures for radiologists to rapidly identify, process, and manage large numbers of landmark point pairs in multiple clinical image sets, without providing a formal statistical method to analyze the resulting data. The purpose of this article is to fill that gap and formally develop statistical methodology for characterizing the accuracy and uncertainty associated with image registration methods for the landmark point pairs approach. We apply our method to registration data collected both from human readers and automated DIR algorithms.

The data upon which we base our methodology arises from a 4D thoracic CT study of five patients conducted at MD Anderson Cancer Center to evaluate the spatial accuracy of registration by human observers and DIR algorithms. In that study, an expert in thoracic imaging manually delineated 200 anatomical landmark pairs in maximum inhale and exhale component phase images for each of the five patient image sets included in the study. The

landmarks in the inhale phase images (i.e., source images) were regarded as source landmarks and were typically defined using easily identifiable anatomical structures, like the bifurcations of blood vessels. Subsequent to the identification of the landmarks in the source images, on two separate occasions the same expert identified the location of these landmarks in the exhale images (i.e., target images). Finally, two additional readers also identified the source landmark locations in the target images, so that a total of four sets of 200 registered landmarks were available for each of the five clinical image sets. In addition to the expert registration data, two automatic DIR methods, one based on an optical flow method (OFM; Horn and Schunck, 1981) and the other on a moving least-squares algorithm (MLS; Schaefer, McPhail and Warrent, 2006), were also used to identify the landmarks in the five patient studies.

The model through which we propose to assess the accuracy of human experts and automated DIR algorithms in performing image registration must account for several features of the process used to collect the landmark data. We address four procedural features in our model. First, in order to obtain efficient estimates, we propose to incorporate information from all expert human readers, as well as the DIR methods, in estimating the true (but unknown) locations of landmarks in each target (exhale) image set. That is, we propose to model human readers and DIR methods jointly.

Second, the human experts used a cursor on a computer display to identify the locations of landmarks in the target image, clicking a mouse when the cursor was properly located (Figure 1). The software program that translated the cursor location to a physical location in the target image truncated the position of the cursor to an integer value. In order to account for the impossibility that the experts were able to identify the locations of landmarks with sub-pixel accuracy, we model the recorded landmark locations as being interval-censored data.

Third, because the DIR methods provided continuous mappings between the source image volume and the target image volume, the registration errors of the DIR methods were spatially correlated. To account for this correlation, we model the DIR registration errors using Gaussian process models.

Finally, the performance of the expert readers and DIR algorithms varied across image sets. To account for this variation, we propose a Bayesian hierarchical model to borrow strength across readers, DIRs, and patient images. Related Bayesian hierarchical spatial models have been proposed by Xu, Johnson, Nichols and Nee (2009) to analyze inter-subject variability in functional MRI image data, and Zhang, Johnson, Little and Yao (2008) to identify pathological alternations of gliomas using quantitative MRI image data. We also describe an efficient Markov chain Monte Carlo algorithm to fit the proposed model.

The remainder of this paper is organized as follows. In Section 2, we propose a Bayesian hierarchical model for data collected from expert readers and DIR methods. In Section 3, we describe a Gibbs sampler to fit these data and discuss concomitant numerical issues. In Section 4, we describe simulation studies to evaluate the performance of the proposed

method. We present the results of the analysis in Section 5. We conduct model assessment in Section 6, and conclude with a brief discussion in Section 7.

2. Models for paired landmark registration data

2.1 A model for expert reader data

Image registration data from the expert readers were obtained using a custom software package called APRIL (Assisted Point Registration of Internal Landmarks). As illustrated in Figure 1, this program displays highlighted landmarks in a source image adjacent to a target image on a computer monitor. The program allows readers to navigate through the three-dimensional target image space to locate the source image landmark. Once located with the cursor, the expert clicks a mouse to record the three-dimensional location of the landmark in the target image. It is important to note that the landmark identification process differs according to the coordinate axis. Although readers are able to locate landmarks in a continuum for in-slice dimensions (i.e., $j = 1, 2$, the two dimensions displayed on the monitor), they are required to perform the discretization associated with the localization of the landmark in the third dimension, or out-of-slice image plane, by toggling between image slices until they identify the image slice most likely to contain the landmark displayed in the source image. Consequently, the program discretizes and recorded the landmark locations on an integer scale.

To model the errors associated with the localization of landmark points in the target image, let μ_{pij} denote the true value of coordinate j for landmark i in source image p , and let s_{pijk} denote the corresponding discretized reading obtained from expert reader k . We assume s_{pijk} is generated according to the following latent reading process. First, we assume that reader k visually identifies a point in the target image volume that corresponds to a landmark displayed in the source image volume and moves the mouse-controlled pointer to this location. Denote the location of this point in the target image volume by $\{x_{pijk}; j = 1, 2, 3\}$. We assume that the coordinates of this point are independently distributed around the true landmark location according to

$$x_{pijk} \sim N(\mu_{pij}, \tau_{pj}^2). \quad (1)$$

Next, the location of the pointer is discretized by either the software program ($j = 1, 2$) or the reader ($j = 3$) and recorded as $s_{pijk} = \text{round}(x_{pijk})$, the integer nearest to x_{pijk} , satisfying

$$s_{pijk} - 0.5 \leq x_{pijk} < s_{pijk} + 0.5. \quad (2)$$

In our study, three readers identified 200 (three-dimensional) landmark locations in five target images. The first reader performed this task twice, resulting in four sets of correlated landmarks for the five target images. Because the first reader identified the landmarks in two independent occasions separated by several weeks, the two sets of landmark locations identified by this reader can be (approximately) viewed as mutually independent.

Conditionally on μ_{pij} and τ_{pj}^2 , we modeled the likelihood function for the reader landmark identifications $s = \{s_{pijk}\}$ according to

$$L(s|\mu_{pij}, \tau_{pj}^2) \propto \prod_{p=1}^5 \prod_{i=1}^{200} \prod_{k=1}^4 \frac{1}{\tau_{pj}^2} \exp \left\{ -\frac{1}{2} \left(\frac{x_{pijk} - \mu_{pij}}{\tau_{pj}} \right)^2 \right\} I(s_{pijk} - 0.5 \leq x_{pijk} < s_{pijk} + 0.5).$$

For simplicity in notation, note that we have indexed k between 1 and 4 even though two values of k , $k = 1$ and 4, correspond to independent readings from the first reader, so that there are only three values of τ_{pj} with $\tau_{pj1} \equiv \tau_{pj4}$.

To characterize the variance properties of readers across images, we assumed that the reader error variances τ_{pj}^2 were generated from inverse gamma densities according to

$$\tau_{pj}^2 \sim IG(\alpha_{jk}, \lambda_{jk}),$$

where α_{jk} and λ_{jk} are shape and rate (or inverse-scale) parameters, respectively. Because navigation through the target image volume was more difficult in the out-of-slice plane (the direction perpendicular to the computer display) and may have differed along horizontal and vertical axes, we fit distinct variance parameters and hyperparameters for each image dimension.

We assigned independent Jeffreys' priors to the hyperparameters $(\alpha_{jk}, \lambda_{jk})$, and independent uniform priors on the target image volume for each value of μ_{pij} . That is, we assumed

$$\begin{aligned} \pi(\alpha_{jk}, \lambda_{jk}) &\propto \lambda_{jk}^{-1} \sqrt{\alpha_{jk} PG(1, \alpha_{jk}) - 1} \\ \pi(\mu_{pij}) &\propto 1, \end{aligned}$$

where $PG(\cdot, \cdot)$ denotes the polygamma function.

2.2 Model for DIR algorithm data

Let y_{pijl} denote voxel coordinate j for landmark i in source image p identified by DIR algorithm l . In image p , let $\mathbf{y}_{p:j} = (y_{p1j}, \dots, y_{p200j})$ and $\boldsymbol{\mu}_{p:j} = (\mu_{p1j}, \dots, \mu_{p200j})$, and define the distance between two landmarks i and i' to be

$$d'_{pi} = \|\boldsymbol{\mu}_{pi} - \boldsymbol{\mu}'_{pi}\| = \sqrt{\sum_{j=1}^3 (\mu_{pij} - \mu'_{pij})^2}, \quad (3)$$

where $\boldsymbol{\mu}_{pi} = (\mu_{pi1}, \mu_{pi2}, \mu_{pi3})$. In principle, the mathematical definitions of the DIR algorithms considered in this article suggest that DIR localization errors are rotationally invariant. That is, rotating the source and target volumes should not affect the registration error. Based on these considerations, we modeled the DIR errors as being independent across dimensions. This assumption implies (quite reasonably) that knowing the registration error in the axial direction (i.e., in-slice dimensions) at landmark i does not provide information regarding the registration error in the coronal direction (i.e., out-of-slice dimension or vertical direction) at either the given landmark i or another landmark i' .

All commonly-used DIR algorithms provide continuous mappings between the source image volume and the target image volume, which means that errors must be correlated within each dimension. To model this correlation structure, we assume that for each image, coordinate readings from the DIR algorithm within each dimension follow a Gaussian process (GP) of the form

$$\mathbf{y}_{p \cdot j} | \mathbf{I} \sim N(\boldsymbol{\mu}_{p \cdot j}, \tilde{\Sigma}_{pjl}).$$

Here, $\tilde{\Sigma}_{pjl}$ is an exponential covariance matrix with (i, i') entry given by

$$\tilde{\Sigma}_{pjl}(i, i') = \sigma_{pjl}^2 \exp(-\gamma_{pjl} d'_{pii}),$$

where γ_{pjl} is an unknown decay parameter controlling correlations among coordinates at different locations in an image. Small values of γ_{pjl} induce strong correlations.

Another feature we should account for in our model is that landmark localization based on DIR algorithms is affected by the accuracy of the human reader in identifying landmarks in the source image. This is because DIR algorithms map a point in the target image to a point in the source image, and the latter must be identified *a priori* by a human reader using methodology similar to that described previously (for reader identification of landmark points in the target image). Any uncertainty or error associated with the identification of the coordinates of the landmark in the source image by the human reader will affect the identification of the corresponding landmark in the target image by DIR algorithms. Note that this issue does not apply to landmark localizations in the target image made by expert readers because human readers identify the landmarks in the source image visually, and match these points to the corresponding points in the target image. Thus, the coordinates of the landmark in the source image are not explicitly used in identifying the landmark in the target image for human readers.

To be more specific, let $\boldsymbol{\zeta}_p = (\zeta_{p1}, \zeta_{p2}, \zeta_{p3})'$ denote the true location of a chosen landmark in source image p , let $\mathbf{w}_p = (w_{p1}, w_{p2}, w_{p3})'$ denote expert k 's (latent) continuous estimate of $\boldsymbol{\zeta}_p$, and let $\mathbf{t}_p = (t_{p1}, t_{p2}, t_{p3})'$ denote the discretized, voxel coordinates of $\mathbf{w}_p$ in the source image, i.e., $t_{pj} = \text{round}(w_{pj})$, $j = 1, 2, 3$. Define $\mathbf{f}(\cdot)$ to be the vector-valued function through which a DIR algorithm maps a point in the source image to a point in the target image. That is, if the true value $\boldsymbol{\zeta}_p$ of a landmark location in the source image were known, the DIR algorithm's prediction of its location in the target image would be $\mathbf{f}(\boldsymbol{\zeta}_p)$.

For each coordinate of $\mathbf{w}_p$ and $\mathbf{t}_p$ in paired image set p , we assume that the latent reader location and DIR transformation of its corresponding voxel, say y_{pj} , can be expressed as

$$w_{pj} = \zeta_{pj} + \epsilon_{pj}, \quad t_{pj} = w_{pj} - u_{pj}, \quad \text{and} \quad y_{pj} = \mathbf{f}_j(t_{pj}) - v_{pj},$$

where u_{pj} and v_{pj} denote independent uniform (0,1) random deviations accounting for the discretization process, and ϵ_{pj} denotes independent, mean-zero normal random deviates

with the reader variance τ_{pjk}^2 . We further assume that $\mathbf{f}_j$ acts locally as a simple translation, at least to the first order, in the sense that the displacement of a landmark and a point very close to a landmark are approximately the same, i.e., $\mathbf{f}_j(\mathbf{a} + \boldsymbol{\varepsilon}) \approx \mathbf{f}_j(\mathbf{a}) + \boldsymbol{\varepsilon}$ for small $\boldsymbol{\varepsilon}$. This assumption is accurate for typical DIR algorithms, which use continuity restrictions to insure that adjacent volumes of tissue remain adjacent to each other in the target image. Then, the difference between the predicted location of the landmark in the target image obtained at the true landmark location and the location obtained from the discretized estimate of the source-image landmark location provided by the reader can be expressed as

$$\mathbf{f}_j(\zeta_{pj}) - y_{pj} \approx u_{pj} + v_{pj} - \epsilon_{pjk}.$$

As u_{pj} and v_{pj} are independent and follow uniform (0,1), $Var(u_{pj}) = Var(v_{pj}) = 1/12$. Thus, the variance of the error resulting from discretization to voxel units and reader errors in the source image is $1/6 + \tau_{pjk}^2$. Figure 2 illustrates the mis-registration effect caused by uncertainty in the landmark location in the source image.

To account for these errors, we note that reader 1 identified all of the landmarks in all of the source images; hence, we add a “nugget” variance component of $1/6 + \tau_{pjk}^2$ to the diagonal elements of $\tilde{\Sigma}_{pjl}$. This leads to a GP covariance matrix Σ_{pjl} with elements

$$\Sigma_{pjl}(i, i') = \begin{cases} \sigma_{pjl}^2 \exp(-\gamma_{pjl} d_{pii'}) & i \neq i' \\ \sigma_{pjl}^2 + 1/6 + \tau_{pjl}^2 & i = i'. \end{cases} \quad (4)$$

We assume that the variance parameters σ_{pjk}^2 are drawn independently from an inverse gamma distribution, i.e.,

$$\sigma_{pjl}^2 \sim IG(\omega_{jl}, \beta_{jl}),$$

and we assign Jeffreys’ prior densities to the hyperparameters ω_{jl} and β_{jl} , which in our model are proportional to

$$\pi(\omega_{jl}, \beta_{jl}) \propto \beta_{jl}^{-1} \sqrt{\omega_{jl} PG(1, \omega_{jl}) - 1}.$$

The specification of the prior for the range parameter γ_{pjl} requires more care. Berger, Oliveira and Sansó (2001) showed that several commonly-used prior densities for this parameter lead to improper posterior distributions. Fortunately, Paulo (2005) provided expressions for independent Jeffreys priors for general GP covariance parameters that yield proper posterior distributions on these parameters. To apply Paulo’s results to our case, let $\mathbf{D}_p$ denote the distance matrix between landmarks with elements defined in (3), and let $\mathbf{D}_p^*$ Σ_{pjl} denote the element-wise product of $\mathbf{D}_p$ and Σ_{pjl} . Then Jeffreys’ prior for γ_{pjl} is proportional to

$$\pi(\gamma_{pjl}) \propto \sqrt{n_p \text{tr}(\mathbf{U}^2) - (\text{tr}(\mathbf{U}))^2}, \quad (5)$$

where $\mathbf{U} = (\mathbf{D}_p * \Sigma_{pjl}) \Sigma_{pjl}^{-1}$ and $\text{tr}(\mathbf{U})$ is the trace of the matrix $\mathbf{U}$.

For clarify, we briefly summarize our notations and model as follows. Let subscripts p, i, j, k, l index image set, landmark, coordinate, reader and DIR algorithm, respectively. There are four primary (latent or observed) measures for landmark locations: (I) μ_{pijk} , the unobserved true location of the landmark; (II) x_{pijk} , the latent continuous landmark locations visually identified by readers; (III) s_{pijk} , the rounded integer values of x_{pijk} that are actually available for data analysis; (IV) y_{pijl} , the landmark location data generated from DIR algorithms. We jointly model (the actually observed) reader data $\{s_{pijk}\}$ and DIR algorithm data $\{y_{pijl}\}$ by assuming that x_{pijk} (i.e., the latent continuous version of s_{pijk}) follows a Gaussian distribution centering around the unobserved true landmark location μ_{pijk} , and y_{pijl} follows a Gaussian process centering around μ_{pijl} . That is, through the latent landmark location, we link the model for the reader data with that for the DIR algorithm data.

3. Bayesian inference for model parameters

From a practical perspective, implementing the joint statistical model described in the previous section is computationally difficult because the covariance matrices used to define the distribution of registration errors for the DIR algorithms depends on the unknown landmark locations $\boldsymbol{\mu}$ through the distance matrix $\mathbf{D}$. Because updating the value of any component of $\boldsymbol{\mu}$ changes the value of the covariance matrix Σ_{pjl} , any sequential updating of landmark locations thus incurs a substantial computational burden.

To understand these computational costs more thoroughly, let $D = \{s_{pijk}, y_{pijl}\}$ denote the observed data, and let $\boldsymbol{\theta}$ denote a generic symbol that represents the values of all other model parameters, including the latent locations μ_{pij} 's. Then the full conditional distribution of landmark μ_{pij} can be expressed as

$$\begin{aligned} f(\mu_{pij} | D, \boldsymbol{\theta}) &\propto \prod_{k=1}^4 f(x_{pijk} | \mu_{pij}, \tau_{pijk}^2) \prod_{j=1}^3 \prod_{l=1}^2 f(y_{pijl} | \mu_{pijl}, \Sigma_{pjl}) \\ &= \prod_{k=1}^4 \exp \left\{ -\frac{(x_{pijk} - \mu_{pij})^2}{2\tau_{pijk}^2} \right\} \prod_{j=1}^3 \prod_{l=1}^2 \exp \left\{ -\frac{(\mathbf{y}_{p \cdot jl} - \boldsymbol{\mu}_{p \cdot j})^T \sum_{pjl}^{-1} (\mathbf{y}_{p \cdot jl} - \boldsymbol{\mu}_{p \cdot j})}{2} \right\}. \end{aligned}$$

Note that Σ_{pjl} depends on the landmark locations $\{\mu_{pij}\}$, so updating an element of $\{\mu_{pij}\}$ from its conditional distribution requires a hybrid Gibbs/Metropolis-Hastings step and the evaluation Σ_{pjl}^{-1} at each new value of $\{\mu_{pij}\}$. With five image sets, each of which contains 200 three-dimensional landmarks, one iteration of a naïve Gibbs sampler that updates $\boldsymbol{\mu}$ on a component-wise basis thus requires inverting 3,000 200×200 covariance matrices. Although a full matrix inversion and related complications can be streamlined, even efficient sampling from the full posterior distribution is computationally expensive.

We alleviate this computational burden associated with posterior inference by exploiting the fact that an update of μ_{pjl} changes only the i th row and column of the covariance matrix Σ_{pjl} , leaving the remaining elements in Σ_{pjl} unaltered. This means that the following matrix property, known as the Sherman-Morrison formula (Sherman and Morrison, 1950), can be used to reduce a matrix inverse operation into a simpler matrix multiplication. Let Σ_{pjl}^* denote an updated covariance matrix obtained by an update of μ_{pjl} , and let $\Sigma_{pjl}(, i)$ and $\Sigma_{pjl}(i, i)$ denote the i th column and the (i, i) element in Σ_{pjl} , respectively. Define $e_i = (0, \dots, 0, 1, 0, \dots, 0)'$ with the i th element of 1, and

$$v = \Sigma_{pjl}^*(, i) - \Sigma_{pjl}(, i) + e_i \left(\Sigma_{pjl}^*(i, i) - \Sigma_{pjl}(i, i) \right) / 2, \text{ so that}$$

$$\Sigma_{pjl}^* = \Sigma_{pjl} + v e_i' + e_i v'.$$

That is, Σ_{pjl}^* is obtained by adding the vector v to both row i and column i of Σ_{pjl} . Given the value of Σ_{pjl}^{-1} , the Sherman-Morrison formula states that

$$\Sigma_{pjl}^{*-1} = \left(\Sigma_{pjl} + v e_i' + e_i v' \right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} e_i v' \mathbf{A}^{-1}}{1 + v' \mathbf{A}^{-1} e_i},$$

where $\mathbf{A}^{-1}$ is the inverse of the matrix obtained by adding v to the i th column of Σ_{pjl} ; that is,

$$\mathbf{A}^{-1} = \left(\Sigma_{pjl} + v e_i' \right)^{-1} = \sum_{pjl}^{-1} - \frac{\sum_{pjl}^{-1} v e_i' \sum_{pjl}^{-1}}{1 + e_i' \sum_{pjl}^{-1} v}.$$

Using this formula, the covariance matrix Σ_{pjl} has to be inverted only one time for each update of all components of $\boldsymbol{\mu}$, which reduces the computational burden that would be incurred by having to invert Σ_{pjl} after the update of each component of $\boldsymbol{\mu}$. In our application, because the dimension of the matrix is not that high (200×200), this computational strategy reduces the computational burden by about 20%, cutting the computational time from 25 hours to 20 hours. For modern big data, which often involve inverting matrixes of thousands by thousands or larger, the save of the computational time by using the above strategy can be more significant. The details of the MCMC algorithm for fitting the proposed model are provided in the Web Appendix.

4. Simulation Study

We conducted a simulation study to evaluate the performance of the proposed method. We assumed that 3 expert readers and 2 DIR algorithms identified 50 landmarks from 5 pairs of 3D images. We generated the true landmark locations in the p th image as $\mu_{pjl} = 2p + i + 4j$, for $p = 1, \dots, 5$, $i = 1, \dots, 50$ and $j = 1, 2$ and 3. We set $(\omega_{j1}, \beta_{j1}) = (18, 68)$ and $(\omega_{j2}, \beta_{j2}) = (38, 222)$, such that the expected error variances for two DIR algorithms were

$$E(\sigma_{pj1}^2) = 4 \quad \text{and} \quad E(\sigma_{pj2}^2) = 6, \text{ respectively, with } Var(\sigma_{pj1}^2) = Var(\sigma_{pj2}^2) = 1. \text{ We}$$

considered two scenarios for the error variances of readers. In scenario 1, the error variances of readers were comparable to those of the DIR algorithms. We set $(\alpha_{j1}, \lambda_{j1}) = (27, 130)$, $(\alpha_{j2}, \lambda_{j2}) = (11, 30)$ and $(\alpha_{j3}, \lambda_{j3}) = (51, 350)$, such that the expected error variances for three readers were $E(\tau_{pjk}^2) = 0.5, 0.3$ and 0.7 , for $k = 1, 2$ and 3 , with $E(\tau_{pjk}^2) = 0.1$. In scenario 2, the error variances of readers were substantially smaller than those of the DIR algorithms. We set $(\alpha_{j1}, \lambda_{j1}) = (4.5, 1.75)$, $(\alpha_{j2}, \lambda_{j2}) = (2.9, 0.57)$ and $(\alpha_{j3}, \lambda_{j3}) = (6.9, 4.13)$, such that the expected error variances for three readers were $E(\tau_{pjk}^2) = 5, 3$ and 7 , for $k = 1, 2$ and 3 , with $Var(\tau_{pjk}^2) = 1$. In scenario 2, the error variances of readers were substantially smaller than those of the DIR algorithms. We set $(\alpha_{j1}, \lambda_{j1}) = (4.5, 1.75)$ and $(\alpha_{j2}, \lambda_{j2}) = (2.9, 0.57)$ and $(\alpha_{j3}, \lambda_{j3}) = (6.9, 4.13)$, such that the expected error variances for three readers were $E(\tau_{pjk}^2) = 0.5, 0.3$ and 0.7 , for $k = 1, 2$ and 3 , with $Var(\tau_{pjk}^2) = 0.1$. We took decay parameters $(\gamma_{p11}, \gamma_{p21}, \gamma_{p31}) = (0.05, 0.1, 0.15)$ for the first DIR algorithm and $(\gamma_{p12}, \gamma_{p22}, \gamma_{p32}) = (0.2, 0.3, 0.4)$ for the second DIR algorithm. Under each scenario, we conducted 1,000 simulations based on the model described in Section 2. We compared the proposed method (i.e., joint modeling) to a two-stage modeling approach, in which at the first stage we fitted only the reader's data and obtained the estimates of landmark locations $\{\hat{\mu}_{pjk}\}$, and then at the second stage, we treated $\{\hat{\mu}_{pjk}\}$ as the true landmark locations and fitted the DIR algorithm data. That is, in the two-stage modeling, we ignored the first-stage estimation uncertainty of $\{\hat{\mu}_{pjk}\}$ at the second-stage estimation.

Table 1 shows the simulation results, including the empirical bias, mean square error and the coverage rate of 95% credible interval (CI) for the estimates of reader error variances τ_{p1k}^2 DIR algorithm error variance σ_{p1l}^2 for $p = 1, \dots, 5, k = 1, 2, 3$, and $l = 1, 2$. The results for the estimates of the variances in the other two dimensions (i.e., $\tau_{p2k}^2, \tau_{p3k}^2, \sigma_{p2l}^2, \sigma_{p3l}^2$) were similar and thus omitted. When the expected error variances were comparable between readers and DIR algorithms (i.e., scenario 1), the joint modeling and two-stage modeling yielded similar results for estimating reader error variances τ_{p1k}^2 , with minimal biases and reasonable coverage rates. However, in terms of estimating DIR algorithm error variances σ_{p1l}^2 , the joint modeling outperformed the two-stage modeling. The coverage rates of the 95% CI under the joint modeling were generally close to the nominal value of 95%, whereas those of the two-stage approach were often lower than 70%. In addition, the empirical bias and MSE of $\hat{\sigma}_{p1l}^2$ under the joint modeling were smaller than those of the two-stage modeling. These results show that failure to propagate the first-stage estimation uncertainty into the second-stage estimation, as in the two-stage modeling, could lead to poor coverage rates and biased estimates. In scenario 2, where the reader's error variances were much smaller than the DIR algorithm's error variances, the joint modeling and two-stage modeling performed similarly, both yielding minimal bias and sound coverage rates. This result was somewhat expected because when the reader variances were small, the first-stage estimation uncertainty became negligible, compared to the second-stage estimation uncertainty, and thus ignoring it had little impact on the second-stage estimation.

We also conducted a sensitivity analysis to evaluate the effect of the misspecification of the reader's model on the estimation of the DIR algorithm registration errors. Instead of assuming that expert readers identified landmark locations unbiasedly as model (1), we simulated the reader's data based on the following model

$$x_{pijk} \sim N\left(\mu_{pij} + c\tau_{pjk}, \tau_{pjk}^2\right),$$

where c is a constant controlling the bias of the reader in the unit of standard deviation (SD) of x_{pijk} (i.e., τ_{pjk}). The values of other simulation parameters were set as those in scenario 2.

Figure 3 depicts the empirical bias of $\hat{\sigma}_{p1l}^2$ (i.e., the estimates of the error variances for the second DIR algorithm in the first image) when c increases from 0 to 2. The results for other σ_{p1l}^2 's, $p = 2, \dots, 5$, $l = 1$, followed similar patterns and were not shown. We can see that when the reader's bias was smaller than one SD (i.e., $c \leq 1$), the increase of the bias was not substantial; however, if the reader's bias was greater than one SD, the bias of $\hat{\sigma}_{p1l}^2$ increased rapidly and was more of a concern.

5. Application

We applied the models and estimation procedures described in Sections 2 and 3 to the registration data by running the Gibbs sampler with 1,000 burn-in iterations, followed by 10,000 updates from the supposed stationary distribution. Graphical checks based on trace plots suggest that the proposed MCMC algorithm had converged by the end of the burn-in phase.

Figure 4 displays boxplot estimates of the posterior distributions of the τ_{pjk}^2 's (i.e., the registration errors for expert readers). This plot suggests that reader 2 had the smallest registration errors and that reader 3 had the largest registration errors. The registration errors for all three readers varied substantially across dimensions, but less noticeably across images. For expert reader 1, the average variances of the reading error on three dimensions were 0.12, 0.14 and 0.05, respectively, while the average variances of the reading error for five images were 0.10, 0.08, 0.13, 0.08 and 0.13, respectively. Compared to those of readers 1 and 2, the variance of the reading errors of reader 3 was substantially larger. For example, for dimension 2 of image 5, the variance of reading errors for reader 3 was 0.62, while that for readers 1 and 2 were only 0.20 and 0.16.

Table 2 shows posterior estimates of the registration error σ_{pjl}^2 and the range parameter γ_{pjl} for the DIR algorithms. Across three dimensions and five images, the registration error of the moving least squares (MLS) algorithm was substantially smaller than that of the optical flow method (OFM), in particular in the second dimension. For example, in the second dimension of image set 1, the posterior mean of the registration error of the MLS algorithm ($\hat{\sigma}_{111}^2$) was 2.3, compared to 107.8 for the OFM ($\hat{\sigma}_{112}^2$). For both the MLS and OFM algorithms, the registration errors differed across dimensions, with the smallest errors occurring in the third dimension. In image set 1, for example, the registration error of the

MLS algorithm in the third dimension was 0.7, approximately one third of what it was in the second dimension.

The decay parameter γ_{pjl} characterizes how strongly the DIR registration errors are spatially correlated, with larger values of γ_{pjl} indicating faster decay of the correlation. Across images and dimensions, the OFM algorithm tends to have a larger decay parameter (or smaller correlations) than the MLS algorithm. For instance in the first dimension of image 1, the decay parameters were 0.009 and 0.029 for the MLS and OFM algorithms, respectively. The correlation among the DIR registration errors for both MLS and OFM algorithms varied significantly across images and dimensions. In particular, the DIR registration errors for images 1 and 2 were more correlated than they were in the other image sets, as indicated by the smaller values of γ_{pjl} reported in Table 2. To summarize the performance of expert readers and DIR algorithms in the three-dimensional space, we defined the mean registration error for the k th expert reader to be

$$e_{pk} = E \left[\sqrt{(x_{pi1k} - \mu_{pi1})^2 + (x_{pi2k} - \mu_{pi2})^2 + (x_{pi3k} - \mu_{pi3})^2} \right]$$

where expectations were computed with respect to the posterior distribution on the image parameters. The expected registration error e_{pk} measures the average Euclidean distance between the landmark location identified by the reader and the true location. Similarly, we define the expected registration error for the l th DIR algorithm to be

$$e_{pl} = E \left[\sqrt{(y_{pi1l} - \mu_{pi1})^2 + (y_{pi2l} - \mu_{pi2})^2 + (y_{pi3l} - \mu_{pi3})^2} \right].$$

Estimates of the expected registration errors for expert readers and DIR algorithms are summarized in Table 3. Among the experts, the second reader had the smallest expected registration error and the third reader the largest. The expected registration errors were relatively stable across images for the first two expert readers, with more variation observed for the third reader.

The MLS algorithm appeared to be more accurate than the OFM algorithm, and their relative performance varied across images. For images 1, 2, 4 and 5, the 95% CIs of the registration error difference between the OFM and MLS algorithms contained 0, but highly skewed toward positive values, suggesting that the MLS algorithm tended to yield smaller registration errors than the OFM algorithm. For image 3, the MLS algorithm was more accurate than the OFM algorithm with a significantly smaller registration error, and the 95% CI for the registration error difference between the OFM and MLS algorithms was (0.24, 30.98). When we averaged registration errors across five images, the registration errors for the MLS and OFM algorithms were 2.14 and 8.68, respectively, and the 95% CI for the difference between the two algorithms was (2.18, 13.4), suggesting that overall the MLS algorithm was more accurate than the OFM algorithm. As a rule of thumb, a DIR algorithm that has an expected registration error that is less than twice that of an average human rater might be viewed as a good algorithm, while a DIR algorithm that has an expected

registration error that is comparable to that of the average human rater is a rater-equivalent algorithm. Based on this rule, the MLS algorithm seems promising, with the expected registration error about 4 times that of human raters. With further improvement, the MLS algorithm could be potentially used as a replacement for human raters to automate image registration.

For the purpose of comparison, in Table 3, we also reported the results yielded by the two-stage modeling. The results were rather similar to those obtained by the joint modeling described above. This is not surprising because the registration errors of expert readers were substantially smaller than those of DIR algorithms, and thus ignoring the first-stage estimation uncertainty had little impact on the second-stage estimation of the parameters for the DIR algorithms. For this specific dataset, the joint modeling did not demonstrate its advantages over the two-stage modeling. However, with the continuous improvement on DIR algorithms, we expect that the accuracy of DIR algorithms will eventually approach that of human readers. The ultimate goal of developing DIR algorithms is to use them to replace human readers to automate image registration. In that case, as shown in the simulation, the joint modeling will have clear edge over the two-stage modeling.

6. Model assessment

We evaluated the adequacy of both the Gaussian process (GP) assumptions for the DIR algorithms and the independent normality assumptions for the reader registration errors using Bayesian goodness-of-fit diagnostics derived from pivotal quantities (Johnson, 2007; Yuan and Johnson, 2012). This approach to model assessment is based on the fact that posterior draws of pivotal quantities follow the same distributions as the corresponding pivotal quantities drawn from the true parameter values. The pivotal quantities employed as goodness-of-fit diagnostics can be tailored to detect particular aspects of a lack of fit, and can be used to assess the adequacy of a specific structural component in a hierarchical model. In this application, the parameters of primary interest are the registration errors of expert readers and the DIR algorithms. We therefore assessed the goodness-of-fit of each individual reader based on the following χ^2 statistics:

$$T_1 = \sum_{p=1}^5 \sum_{i=1}^{200} \sum_{j=1}^3 \frac{(x_{pij1} - \mu_{pij})^2 + (x_{pij4} - \mu_{pij})^2}{\tau_{pij1}^2},$$

$$T_k = \sum_{p=1}^5 \sum_{i=1}^{200} \sum_{j=1}^3 \frac{(x_{pijk} - \mu_{pij})^2}{\tau_{pijk}^2}, k=2, 3.$$

Each value of T_k was obtained for a single draw of $\{x_{pijk}\}$, $\{\mu_{pij}\}$, and $\{\tau_{pijk}\}$ from the posterior distribution. Although T_k 's may not be the most powerful test statistics, they are intuitive and simply to calculate. The GP assumptions imposed on the registration errors for the DIR algorithms were tested using chi-squared statistics defined according to

$$S_l = \sum_{p=1}^5 \sum_{j=1}^3 (\mathbf{y}_{p \cdot jl} - \mu_{p \cdot j})^T \Sigma_{pjl}^{-1} (\mathbf{y}_{p \cdot jl} - \mu_{p \cdot j}), \quad l=1, 2,$$

again evaluated at parameter values drawn from the posterior distribution. Provided that the proposed models have been correctly specified, the marginal distribution of each posterior draw of T_1 follows a χ^2_{6000} distribution, whereas the marginal distributions of posterior draws of T_2 , T_3 , S_1 and S_2 follow χ^2_{3000} distributions. As shown in Figures A1 and A2 in the Web Appendix, the posterior distributions of the χ^2 statistics for the expert readers and DIR algorithms are generally very close to the corresponding reference distributions. The posterior probabilities that a value of the χ^2 statistic exceeded the 0.95 quantile from their reference χ^2 distributions were approximately 0.04, 0.05 and 0.05 for primary and secondary readers, respectively, and were 0.05 and 0.10 for the MLS and OFM, suggesting that the proposed model generally fits the data well. To explore why the OFM demonstrated a slightly higher deviation from the reference distribution than expected, we conducted more detailed model assessments on each image and dimension for the OFM using “lower-level” χ^2 statistics of the form

$$S_{pj2} = (\mathbf{y}_{p \cdot j2} - \mu_{p \cdot j})^T \Sigma_{pj2}^{-1} (\mathbf{y}_{p \cdot j2} - \mu_{p \cdot j}), \quad p=1, \dots, 5, j=1, 2, 3, \quad (6)$$

which are nominally distributed as χ^2_{200} distributions.

Figure A3 in the Web Appendix shows a quantile-quantile plot of the posterior distributions of $\{S_{pj2}\}$ against the expected quantiles of the χ^2_{200} distribution, suggesting that the lower level of the model fit the data generally well except on the third dimension of image 2. It appears that the OFM algorithm shifted some voxels on the out-of-plane axis in image 2, resulting in the suggested minor lack-of-fit. As a sensitivity analysis, we removed the lack-of-fit data and refitted the model. The estimates of the parameters associated with other images remained largely unchanged.

7. Discussion

We have proposed a Bayesian hierarchical model to simultaneously evaluate the spatial accuracy for human expert readers and automatic DIR methods. In our approach, true landmark locations are treated as unknown latent variables. Our model explicitly accounts for the variation among multiple experts and the discretization process of landmark identification. When evaluating the spatial accuracy of the automatic DIR methods, our model accounts for the random errors associated with the experts' registration errors and utilizes a hierarchical model to borrow information across multiple images. A Gibbs sampling algorithm was developed to fit the data efficiently. Our work provides a framework for objective evaluation of DIR spatial accuracy using large sets of expert-determined landmark feature pairs.

Acknowledgments

We are grateful to Professor Russell Millar, the associate editor, and two anonymous referees for their many constructive and insightful comments that have resulted in significant improvements in the article. This work was partially supported by NCI grants R01 CA154591 (Y.Y.), CA016672 (Y.Y.), and R01 CA158113 (V.J), and NIH grants 1K01CA181292 (R.C.), and DP2OD007044 (T.G.).

REFERENCES

- Al-Mayah A, Moseley J, Brock KK. Contact surface and material nonlinearity modeling of human lungs. *Physics in Medicine and Biology*. 2008; 53(1):305–317. [PubMed: 18182705]
- Berger JO, De Oliveira V, Sanso B. Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*. 2001; 96:1361–1374.
- Boldea V, Sharp GC, Jiang SB, Sarrut D. 4D-CT lung motion estimation with deformable registration: quantification of motion nonlinearity and hysteresis. *Medical Physics*. 2008; 35:1008–1018. [PubMed: 18404936]
- Castillo E, Castillo R, Martinez J, Shenoy M, Guerrero T. Four-dimensional deformable image registration using trajectory modelling. *Physics in Medicine and Biology*. 2010; 55:305–327. [PubMed: 20009196]
- Castillo R, Castillo E, Guerra R, Johnson VE, McPhail T, Garg AK, Guerrero T. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Physics in Medicine and Biology*. 2009; 54:1849–1870. [PubMed: 19265208]
- Gu X, Pan H, Liang Y, Castillo R, Yang D, Choi D, Castillo E, Majumdar A, Guerrero T, Jiang SB. Implementation and evaluation of various demons deformable image registration algorithms on a GPU. *Physics in Medicine and Biology*. 2010; 55:207–219. [PubMed: 20009197]
- Guerrero T, Zhang G, Huang TC, Lin KP. Intrathoracic tumour motion estimation from CT imaging using the 3D optical flow method. *Physics in Medicine and Biology*. 2004; 49:4147–4161. [PubMed: 15470929]
- Hof H, Herfarth KK, Munter M, Essig M, Wannenmacher M, Debus J. The use of the multislice CT for the determination of respiratory lung tumor movement in stereotactic single-dose irradiation. *Strahlentherapie und Onkologie*. 2003; 179(8):542–7. [PubMed: 14509953]
- Horn BKP, Schunck BG. Determining optical flow. *Artificial Intelligence*. 1981; 17:185–203.
- Johnson VE. Bayesian model assessment using pivotal quantities. *Bayesian Analysis*. 2007; 2:719–734.
- Kabus S, Klinder T, Murphy K, van Ginneken B, Lorenz C, Pluim J. Evaluation of 4D-CT lung registration. *Medical Image Computing and Computer-Assisted Intervention MICCAI*. 2009; 2009:747–754.
- Kaus MR, Brock KK, Pekar V, Dawson LA, Nichol AM, Jaffray DA. Assessment of a model-based deformable image registration approach for radiation therapy planning. *International Journal of Radiation Oncology Biology Physics*. 2007; 68:572–580.
- Keall P. 4-dimensional computed tomography imaging and treatment planning. *Seminars in Radiation Oncology*. 2004; 14(1):81–90. [PubMed: 14752736]
- Li P, Malsch U, Bendl R. Combination of intensity-based image registration with 3D simulation in radiation therapy. *Physics in Medicine and Biology*. 2008; 53:4621–4637. [PubMed: 18695293]
- Padhani AR, Ollivier L. The RECIST (Response Evaluation Criteria in Solid Tumors) criteria: implications for diagnostic radiologists. *British Journal of Radiology*. 2001; 74(887):983–6. [PubMed: 11709461]
- Paulo R. Default priors for Gaussian process. *The Annals of Statistics*. 2005; 33:556–582.
- Sarrut D. Deformable registration for image-guided radiation therapy. *Zeitschrift für Medizinische Physik*. 2006; 16(4):285–297. [PubMed: 17216754]
- Schaefer, S.; McPhail, T.; Warren, J. *ACM SIGGRAPH 2006 Papers*. ACM Press; Boston, MA: 2006. Image deformation using moving least squares.
- Sherman J, Morrison WJ. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*. 1950; 21(1):124–127.
- Schnabel JA, Tanner C, Castellano Smith AD, Degenhard A, Leach MO, Hose R, Hill DLG, Hawkes DJ. Validation of non-rigid image registration using finite element methods: application to breast MR images. *IEEE Transactions on Medical Imaging*. 2003; 22:238–247. [PubMed: 12716000]
- Vandemeulebroucke, J.; Sarrut, D.; Clarysse, P. The POPI-Model, a point validated pixel-based breathing thorax model. *International Conference on the Use of Computers in Radiation Therapy (ICCR)*; Toronto, Canada. 2007.

- Wolthaus JWH, Sonke JJ, van Herk M, Damen MF. Reconstruction of a time-averaged midposition CT scan for radiotherapy planning of lung cancer patients using deformable registration. *Medical Physics*. 2008; 35(9):3998–4011. [PubMed: 18841851]
- Wu Z, Rietzel E, Boldea V, Sarrut D, Sharp GC. Evaluation of deformable registration of patient lung 4D CT with subanatomical region segmentations. *Medical Physics*. 2008; 35(2):775–781. [PubMed: 18383700]
- Xu L, Johnson TD, Nichols TE, Nee DE. Modeling inter-subject variability in fMRI activation location: a Bayesian hierarchical spatial model. *Biometrics*. 2009; 65:1041–1051. [PubMed: 19210732]
- Yuan Y, Johnson VE. Goodness-of-fit diagnostics for Bayesian hierarchical models. *Biometrics*. 2012; 68:156–164. [PubMed: 22050079]
- Zhang X, Johnson TD, Little RJA, Cao Y. Quantitative magnetic resonance imaging via the EM algorithm with stochastic variation. *Annals of Applied Statistics*. 2008; 2:736–755. [PubMed: 20046819]

Figure 1.

Manual registration interface showing source images (top panel) and target images (bottom panel) from a 4D CT set. Green crosshairs indicate manually registered source landmarks. Expert readers navigated through the three-dimensional target image space to locate the landmark displayed in the source image; the location in the target image was recorded at the point where the reader positioned the computer cursor and clicked on the computer mouse.

Figure 2.

Illustration of the mis-registration effect caused by uncertainty in the landmark location in the source image. In the source image, the vector $\zeta_p = (\zeta_{p1}, \zeta_{p2}, \zeta_{p3})'$ is the true location of a chosen landmark; $\mathbf{w}_p = (w_{p1}, w_{p2}, w_{p3})'$ is the expert's (latent) continuous estimate of ζ_p ; (t_{p1}, t_{p2}, t_{p3}) denote the discretized, voxel coordinates of $\mathbf{w}_p$. A DIR algorithm maps these points in the source image to points in the target image, $(f(\zeta_{p1}), f(\zeta_{p2}), f(\zeta_{p3}))$ and $(f(t_{p1}), f(t_{p2}), f(t_{p3}))$, through a vector-valued function $f(\cdot)$. The vector $(y_{p1}, y_{p2}, y_{p3})'$ is the discretized, voxel coordinates of $(f(t_{p1}), f(t_{p2}), f(t_{p3}))$.

Figure 3.

The empirical bias for the estimate of the DIR algorithm error variance σ_{1j2}^2 different degrees of bias for expert readers.

Figure 4.

Boxplots for posterior draws of the reading error τ_{pjk}^2 for three expert readers in three dimensions and five images. Under each dimension and image, three boxplots, from left to right, are for expert readers 1, 2 and 3, respectively.

Table 1

Simulation results for estimating the reader error variance τ_{p1k}^2 and DIR algorithm variance σ_{p1l}^2 based on the joint modeling and two-stage modeling.

Image	Joint modeling						Two-stage modeling					
	Reader (τ_{p1k}^2)			DIR (σ_{p1l}^2)			Reader (τ_{p1k}^2)			DIR (σ_{p1l}^2)		
	1	2	3	1	2	3	1	2	3	1	2	3
Scenario 1												
Bias	1	-0.130	0.090	0.118	2.971	2.311	0.048	0.020	0.194	4.892	4.143	
	2	-0.087	0.025	0.034	3.007	2.308	0.086	-0.020	0.165	4.721	4.129	
	3	0.139	0.021	0.072	3.155	2.308	0.081	0.015	0.053	4.886	4.057	
	4	-0.121	0.028	0.112	3.129	2.280	0.029	0.052	0.132	4.924	4.087	
	5	0.138	0.028	0.107	3.159	2.258	0.038	0.036	0.102	4.858	4.145	
MSE	1	0.566	0.669	1.455	19.66	11.10	0.716	0.759	1.473	38.28	22.55	
	2	0.565	0.676	1.399	21.39	11.03	0.693	0.790	1.503	33.54	22.61	
	3	0.616	0.725	1.434	20.75	11.29	0.730	0.729	1.575	37.76	21.43	
	4	0.540	0.695	1.437	24.82	11.54	0.711	0.855	1.612	38.49	22.13	
	5	0.491	0.685	1.410	22.66	10.50	0.677	0.798	1.501	36.70	23.01	
Coverage	1	0.939	0.929	0.946	0.954	0.975	0.927	0.930	0.946	0.647	0.708	
	2	0.921	0.938	0.934	0.919	0.965	0.942	0.943	0.947	0.621	0.685	
	3	0.938	0.940	0.951	0.927	0.983	0.920	0.931	0.940	0.636	0.719	
	4	0.940	0.937	0.949	0.931	0.927	0.932	0.928	0.944	0.632	0.698	
	5	0.921	0.936	0.944	0.930	0.963	0.934	0.925	0.949	0.619	0.712	
Scenario 2												
Bias	1	0.002	0.018	0.009	1.528	0.531	0.005	0.021	0.005	1.674	0.850	
	2	0.005	0.019	0.018	1.457	0.542	0.004	0.023	0.004	1.483	0.877	
	3	0.007	0.020	0.012	1.463	0.567	0.006	0.021	0.013	1.547	0.868	
	4	0.005	0.017	0.009	1.526	0.525	0.008	0.016	0.018	1.508	0.786	
	5	0.001	0.016	0.011	1.390	0.496	0.008	0.025	0.012	1.590	0.747	
MSE	1	0.014	0.018	0.041	12.32	3.220	0.019	0.021	0.038	13.83	4.269	
	2	0.012	0.017	0.038	11.48	3.095	0.015	0.021	0.043	13.59	3.817	

Joint modeling										Two-stage modeling									
Reader (τ_{p1k}^2)					DIR (σ_{p1l}^2)					Reader (τ_{p1k}^2)					DIR (σ_{p1l}^2)				
Image	1	2	3		1	2	3	1	2	1	2	3		1	2	3	1	2	
Coverage	3	0.014	0.023	0.038	11.22	2.854	0.015	0.023	0.043	11.11	3.487								
	4	0.014	0.021	0.049	13.32	2.858	0.018	0.020	0.048	11.21	3.446								
	5	0.013	0.021	0.040	10.61	2.413	0.014	0.020	0.042	11.13	3.050								
	1	0.939	0.958	0.930	0.934	0.926	0.967	0.940	0.925	0.912	0.927								
	2	0.943	0.943	0.937	0.930	0.936	0.947	0.941	0.932	0.937	0.930								
	3	0.937	0.940	0.943	0.941	0.951	0.946	0.924	0.938	0.932	0.917								
	4	0.948	0.920	0.945	0.941	0.936	0.944	0.963	0.935	0.938	0.915								
	5	0.936	0.955	0.943	0.937	0.937	0.959	0.953	0.949	0.935	0.932								

Table 2

Posterior mean and standard error of σ_{pjl}^2 and γ_{pjl} . The posterior standard error is shown in parentheses.

Image	Dimension 1		Dimension 2		Dimension 3	
	MLS	OFM	MLS	OFM	MLS	OFM
$\hat{\sigma}_{pjl}^2$						
1	1.7(0.4)	20.4 (6.9)	2.3 (1.5)	107.8 (72.3)	0.7 (0.3)	2.0 (0.6)
2	1.7 (0.4)	10.6 (4.1)	2.6 (2.2)	96.6 (90.5)	0.7 (0.4)	4.8 (11.0)
3	1.7 (0.3)	32.3 (13.8)	3.0 (1.5)	195.2 (218.2)	0.7 (0.2)	7.0 (2.5)
4	1.7 (0.3)	13.9 (3.1)	2.5 (1.9)	41.0 (29.2)	0.7 (0.2)	5.4 (7.3)
5	1.7 (0.3)	10.4 (9.8)	7.7 (8.2)	36.6 (41.3)	0.8 (0.3)	2.2 (0.8)
$\hat{\gamma}_{pjl}^2$						
1	.009 (.003)	.029 (.008)	.015 (.009)	.011 (.005)	.007 (.005)	.040 (.013)
2	.007 (.002)	.029 (.011)	.004 (.004)	.007 (.004)	.001 (.001)	.026 (.012)
3	.026 (.007)	.027 (.009)	.019 (.008)	.017 (.007)	.029 (.009)	.032 (.009)
4	.019 (.005)	.040 (.009)	.024 (.012)	.021 (.009)	.016 (.007)	.028 (.010)
5	.022 (.006)	.023 (.012)	.014 (.007)	.018 (.011)	.025 (.009)	.036 (.012)

Table 3

Posterior mean and standard error of registration errors for three expert reader and the MLS and OFM algorithms based on the joint modeling and two-stage modeling. The posterior standard error or 95% credible interval is shown in parentheses.

Image	Reader			DIR algorithm		
	1	2	3	OFM	MLS	Difference (95% CI)
Joint modeling						
1	0.51 (0.22)	0.46 (0.20)	0.62 (0.27)	9.64 (6.10)	1.93 (0.96)	7.71 (−0.45, 23.25)
2	0.44 (0.19)	0.37 (0.16)	0.60 (0.26)	8.70 (6.02)	1.99 (1.03)	6.71 (−0.79, 22.49)
3	0.57 (0.26)	0.47 (0.21)	0.83 (0.38)	12.62 (8.74)	2.07 (1.01)	10.55 (0.24, 30.98)
4	0.43 (0.19)	0.38 (0.17)	0.52 (0.22)	6.76 (3.88)	1.97 (0.96)	4.79 (−0.84, 14.43)
5	0.58 (0.26)	0.48 (0.23)	0.87 (0.44)	5.71 (3.93)	2.72 (1.69)	2.99 (−3.30, 14.04)
Two-stage modeling						
1	0.51 (0.24)	0.44 (0.23)	0.75 (0.33)	10.80 (8.47)	2.05 (1.09)	8.75 (−0.58, 30.56)
2	0.44 (0.19)	0.28 (0.14)	0.70 (0.32)	8.77 (6.64)	2.03 (1.07)	6.74 (−0.90, 22.60)
3	0.57 (0.27)	0.40 (0.20)	1.04 (0.48)	11.93(7.40)	2.14 (1.09)	9.79 (−0.00, 28.40)
4	0.43 (0.20)	0.33 (0.16)	0.61 (0.27)	6.64 (3.67)	2.08 (1.06)	4.56 (−1.15, 13.45)
5	0.56 (0.27)	0.45 (0.23)	1.07 (0.55)	5.68 (3.61)	2.66 (1.72)	3.02 (−3.41, 12.96)