Review paper

Voxel-based analysis in radiation oncology: A methodological cookbook

G. Palma^{*}, S. Monti, L. Cella

National Research Council, Institute of Biostructures and Bioimaging, Napoli, Italy

ARTICLE INFO

Keywords:

Outcome analysis
Voxel-based analysis
Elastic image registration
Multiple comparison problem
Normal tissue complication probability

ABSTRACT

Recently, 2D or 3D methods for dose distribution analysis have been proposed as evolutions of the Dose Volume Histogram (DVH) approaches. Those methods, collectively referred to as pixel- or voxel-based (VB) methods, evaluate local dose response patterns and go beyond the organ-based philosophy of Normal Tissue Complication Probability (NTCP) modelling. VB methods have been introduced in the context of radiation oncology in the very last years following the virtuous example of neuroimaging experience. In radiation oncology setting, dose mapping is a suitable scheme to compare spatial patterns of local dose distributions between patients who develop toxicity and who do not.

In this critical review, we present the methods that include spatial dose distribution information for evaluating different toxicity endpoints after radiation therapy. The review addresses two main topics. First, the critical aspects in dose map building, namely the spatial normalization of the dose distributions from different patients. Then, the issues related to the actual dose map comparison, *i.e.* the viable options for a robust VB statistical analysis and the potential pitfalls related to the adopted solutions. To elucidate the different theoretical and technical issues, the covered topics are illustrated in relation to practical applications found in the existing literature.

We conclude the overview on the VB philosophy in radiation oncology by introducing new phenomenological approaches to NTCP modelling that accounts for inhomogeneous organ radiosensitivity.

1. Background

The improvement of our understanding of the mechanisms underlying radiation-induced damage to normal tissue in cancer patients treated with radiation therapy (RT) has long been recognized as one of the key challenges in Radiation Oncology [1]. In a RT treatment, the goal of high and highly tailored tumor dose distribution with maximum normal tissue sparing generally produces uniformly distributed doses in the target, while an unevenly distribution in the surrounding healthy structures is expected. The dose-volume histogram (DVH), since its introduction in the late 1970s [2], has proved to be a valuable tool for summarizing the target or organ-at-risk 3D dose distribution in a 2D graph [3].

Normal tissue damage has historically been assessed by organ-at-risk DVHs [4] and most available biological models developed in RT to predict normal tissue rely on them rather than on the complete radiation dose distribution information [5]. This, however, has manifold implications.

First, the use of DVH implies treating the organ-at-risk as a homogeneous entity in its response to radiation, losing the fine-structure biology within the organ-at-risk [6]. This goes against the increasing

evidences showing heterogeneous regional organ radiation sensitivity for many organs and many radiation-induced effects [7,8]. Organ regional differences in radiation sensitivity have been demonstrated for lung [9], urinary bladder [10], parotid gland [11] or white matter structures [12].

Furthermore, the use of DVH implies the application of an organ-based approach that relies on a *a priori* definition of the anatomical region involved in a given toxicity outcome. This hampers investigations on the interplay among different organs and their contribution to a specific radiation induced toxicity. Interactions of radiation effects between functionally dependent organs, such as heart and lung [13–15], are thus not taken into consideration.

A non-standardized anatomy-based definition and delineation for the considered organ-at-risk may ultimately influence the consistency of toxicity prediction [16,17].

It is against this background that various methods have been proposed as alternative to the DVH in order to take into account the spatial information of dose distributions. Since the pioneering paper by Thames et al. in 2004 [7] and its first attempts to introduce *in silico* normal tissue complication probability (NTCP) models based not on DVHs but rather on the 3D co-registered dose distributions, we have

^{*} Corresponding author.

E-mail address: giuseppe.palma@ibb.cnr.it (G. Palma).

<https://doi.org/10.1016/j.ejmp.2019.12.013>

Received 24 July 2019; Received in revised form 16 December 2019; Accepted 17 December 2019

Available online 08 January 2020

1120-1797/ © 2019 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. All rights reserved.

Fig. 1. Results of a voxel-based t -test analysis assessing patterns of gray matter loss in a group of patients with Alzheimer's disease (AD) compared with a group of healthy controls. The voxels that showed significantly reduced gray matter volume in the AD subjects compared with controls are shown using three different display techniques. Reprinted from [23] – copyright 2009 Society for Neuroscience.

seen an increasing spread of studies exploring local dose response patterns for the analysis of different toxicity endpoints [18–20].

In particular, new 2D or 3D approaches collectively referred to as pixel- or voxel-based (VB) methods have been introduced in the context of radiation oncology in the very last years following the virtuous example of neuroimaging community.

VB analysis is indeed a widely used technique for assessing structural or functional changes in neuroanatomical images. It involves a voxel-wise comparison of the intensity (in Magnetic Resonance (MR) imaging) or activity (in nuclear medicine) of brain tissue between groups of individuals [21,22]: following the spatial normalization of all the images to the same anatomical space, a statistical analysis is performed to make inferences on group differences. Since its introduction, VB morphometry has proven to be a very useful tool for investigating the pathophysiological patterns of several neuroinflammatory or neurodegenerative diseases, such as multiple sclerosis, parkinsonisms, Alzheimer's disease (Fig. 1), etc. [23].

In RT, VB analysis refers to the analysis of dose distributions in which local dose response patterns are evaluated beyond the organ-based philosophy, by means of computational tools able to fully exploit the information included in the 3D dose distribution. Its most common application can be found in radiation induced morbidity studies to identify heterogeneous patterns of radiosensitivity [5,24–39]. Alongside, few studies explored possible VB correlations between treatment failure and dose distribution around the target [28,40].

2. The backbone of a VB analysis

A VB analysis consists of two main processes (Fig. 2). First, a spatial normalization of the different anatomies in the analyzed cohort of patients to a common coordinate system (CCS). Once the RT dose maps have been anchored in a way that they can be directly compared point-by-point, regions in which significant correlations exist between a clinical outcome and the local dose release are identified by means of statistical inference on the spatial signature of dose response.

2.1. Spatial normalization

The anatomical differences among patients, which in principle

hinder the direct comparison of local dose delivered to each subject, can be overcome following two alternative strategies: the definition of a coordinate patch on a hollow organ, or the Elastic Image Registration (EIR) of the 3D patient space.

2.1.1. Charts on organ walls

For hollow organs, such as the rectum or the bladder, it is reasonably expected that the dose to the inner volume (occupied by urine or rectal filling, respectively) does not contribute to the analyzed outcome. Therefore, the analysis is focused on their walls, which are then considered as surfaces (2D differentiable manifolds $\{\Omega_i\}$) embedded in the 3D body space [41]. Hence, one coordinate chart f_i is chosen for each patient i to map an open subset O_i of the organ wall into an open subset of $\mathbb{R}^2$

$$f_i: O_i \subset \Omega_i \rightarrow \mathbb{R}^2$$

From a mathematical viewpoint, a single chart, in general, cannot be sufficient to map the entire organ; therefore, a radiobiological comprehensive choice should try to select a homeomorphism defined on a large organ domain. For instance, in the case of the bladder, whose Lusternik-Schnirelmann category is 2 since it is homeomorphic to a 2-sphere, a single chart can map at most all the points but one.

The common assumption, then, is that those charts provide a way to identify the local doses – delivered to the different patients – that can be safely compared at a small-scale anatomical level. This is obtained by computing the doses D_i that were delivered in the preimages of a given point $\vec{x}$ of the 2D Euclidean space. The scheme results in the computation of the organ Dose-Surface Maps (DSMs) for each patient, which are thus defined as

$$\text{DSM}_i: f_i(O_i) \rightarrow \mathbb{R}$$

$$\vec{x} \rightarrow D_i(f_i^{-1}[\vec{x}])$$

The anatomical examples found in the literature fall into two geometrical classes: organs that can be assimilated to a cylinder (rectum or esophagus – Fig. 3) and the bladder, which is, as said above, topologically equivalent to a sphere.

In the first case, a very robust approach has been proposed and applied for the modelling of esophagus [28,42] and rectum

Fig. 2. Flowchart of a typical voxel-based analysis. The structural images of the analyzed dataset are first spatially normalized to a common coordinate system (here, via an elastic image registration with a subject-specific censoring mask). The obtained transformations are then used to warp the dose maps. Finally, the statistical analysis is performed according to a statistic designed to potentially include non-dosimetric explanatory variables.

[24,25,43,44], which can be described as follows. For each point along the central axis of the organ, the closed curve given by the intersection of the organ with the plane orthogonal to the axis is parametrized by the azimuth from a fixed point (usually the most posterior point of the curve). A normalization of the axial coordinate is then recommended in order to rescale the organ length to a dimensionless constant.

Concerning the bladder, the issue has been addressed in very few studies, in which the strategy for the Cartesian unfolding of the organ wall closely follows the techniques developed for cylinder-like structures. In particular, both Palorini et al. [38] and Yahya et al. [39] defined somehow a reference point of the bladder that should be mapped in the origin of the DSM reference domain. Then, they parametrized the intersection of the organ wall with each axial plane according to the azimuth from the sagittal plane through the reference point. To account for the axial position, the plain craniocaudal distance from the reference point was provided, with an arbitrary, non-normalized length (ranging from 25 mm to 45 mm, depending on the paper) defining the axial extent of the considered organ portion O_i . However, as acknowledged by the authors themselves, this choice provides unreliable results in the upper portions of the bladder [38]. We suggest that, in order to satisfy the assumption on anatomical accuracy of the chosen charts, a consistently better option would imply to rescale the organ axial extent to a dimensionless constant.

It is worth mentioning that, in general, the domain subregions of the DSMs cannot be required to be representative of the actual number of endothelial cells present in their organ preimages (this would be desirable in order to assign comparable weights to equivalent portions of

tissue, as mentioned in [38,43,44]). Indeed, if this requirement can be roughly met for properly designed charts of the rectum or of the esophagus (a cylinder is locally isometric to the Euclidean plane), it is necessarily violated for the bladder, whose scalar curvature is non-null.

A completely different solution, as suggested in [38], would be to adopt a spatial normalization that employ the EIR approach. While this approach will be described in the following section as a standalone procedure, we argue that EIR could also be used as preprocessing step for the extraction of a more anatomically accurate DSM.

2.1.2. Elastic image registration

For a generic organ, the mapping of the doses to a CCS can be obtained through tailored EIR, and results are usually more accurate than those achieved with geometric mapping even for 2D organs [45]. This strategy assumes the availability of a set of dose maps, each of which is paired with a registered (*i.e.*, spatially aligned) structural image I_i with a sufficient tissue contrast that can guide the normalization process:

$$I_i: \Omega_i \subset \mathbb{R}^3 \rightarrow \mathbb{R}$$

In accordance to the type of structural image, an appropriate reference image has to be chosen

$$\rho: \text{CCS} \subset \mathbb{R}^3 \rightarrow \mathbb{R}$$

The spatial normalization by means of EIR consists of two main steps. First, an inter-patient registration of each I_i to ρ is performed by finding the transform

$$\tau_i: \text{CCS} \rightarrow \Omega_i$$

Fig. 3. A geometrical model of the rectum. A straight hollow tube may be bent and the slices orthogonal to the central axis can be deformed to match a real rectum shape. Reprinted from [41] with permission from Elsevier.

that minimizes a real cost function $C_{el}(I_i \circ \tau_i, \rho)$. A methodological challenge arises when the tumor lesion is inside the organ-at-risk. In this case, a lesion-masking approach should be used to censor the contribution of the Gross Tumor Volume (GTV) to the EIR cost function. Under these conditions, defining the set $G_i \subset \Omega_i$ as the support of the GTV mask for the i -th patient, the cost function to be minimized is given by

$$C_{el}(\tau_i, I_i, \rho) = C_M(\tau_i, I_i \circ \tau_i, \rho)$$

where C_M must be independent on the definition of both $I_i \circ \tau_i$ and ρ in the preimage of G_i under τ_i .

Finally, once for each patient i the transformation τ_i has been obtained, the propagation of the associated dose map to the CCS is performed as

$$D_{i,\tau} = D_i \circ \tau_i$$

The most used schemes for inter-patient EIR and consequent dose warping are the Demons and B-spline algorithms, each of which is known to have pros and cons [46]. The specific impact of the EIR scheme on the VB analysis of regional dose effects on radiation-induced morbidity has been evaluated for a specific RT setup in [32].

In the VB literature, I_i can be either a usual medical image (i.e., a planning CT), a structural model derived from the organ contours, or a combination of both.

Among the first group, an EIR method based on B-spline parametrization [47,48] has been implemented in [31] and [27] and applied respectively to the planning CTs of lungs and head and neck cancer RT patients, after a Hounsfield Units thresholding to exclude from the registration process bone in the lung dataset and air cavities in the head and neck dataset.

Another CT-based approach has been proposed in [34] and [5], where it has been applied to thoracic and head and neck regions respectively. Briefly, for each patient, a binary mask was computed as the union and dilation of the organ-at-risk segmentations of the treatment plan (heart and lungs in [34], oral cavity, larynx and esophagus in [5]). The masks were then used to crop the field of view and were multiplied with the planning CT of each patient to hide inessential interindividual anatomic differences. Finally, the masked CTs were registered to the CCS via a log-diffeomorphic demons approach [49].

In the above works, the reference image was chosen among the planning CT of the cohort of patients under investigation, defining a study-specific CCS.

A different choice concerning the CCS has been made in [36] and in [35], where a synthetic CT phantom from the extended cardiac torso (XCAT) [50] was chosen as reference anatomy. Those studies, in addition, dealt for the first time with the lesion masking issue due to the presence of a target volume inside the organ-at-risk. The actual EIR was based on B-spline parametrization.

Several studies highlighted the impact of image quality (signal-to-noise ratio, tissue contrast, etc.) on the accuracy of the spatial normalization [51–53]. The variability of tissue contrasts between patients' and reference images (due, for instance, to different scanners or to variable acquisition parameters) can be handled in most cases by a proper choice of the EIR cost function (e.g., preferring an analysis of the joint probability distribution rather than a simpler radiometric distance). Conversely, an insufficient contrast-to-noise ratio can pose a more serious problem.

To deal with this issue, the EIR based on 3D structural model extracted from the organ-at-risk contouring has been applied in the pelvis region, where poor soft-tissue contrast, large inter-individual variability and differences in bladder and rectum filling make CT intensity-based registration not sufficiently accurate. In [29] and [30] a rectal structural model has been conceived. Briefly, the rectum centerlines were computed with a skeletonization method. Then, for each patient, a Laplace's equation with constant Dirichlet conditions on the centerline and on the rectal boundary was defined, and the structural image was

then obtained as the solution of the equation inside the organ and as the Euclidean distances to the rectal boundary outside of it. The rectal structural models were registered onto a study-specific anatomical template's model by means of an accelerated version of the Demons algorithm. A similar approach was described in [33] for the whole pelvis.

The work by [26] have described an EIR method based on the combination of the grayscale of CT with structural model from the contoured organs, in a group of prostate cancer patients. Euclidean distance maps were computed within the prostate, the bladder and the rectum from each patient and then added to the individual CT scans. These images were then spatially aligned to a study specific CCS by means of Demons algorithm.

A spatial normalization based on multimodal imaging has been devised in [54] to extend the scope of VB analysis to the field of brain irradiation, where potentially large inaccuracies may be introduced in the parenchyma by EIRs based on the planning CT alone. To overcome this issue, the proposed method was based on the combined use of planning CT with MR images, which well characterize the cerebral anatomy and soft tissue contrast. It consisted of three main steps: first, an intra-patient registration of the planning CT onto the MR image was performed to anchor the TP information to an image with high soft-tissue contrast; then, each MR image was normalized onto an MR-based CCS; finally, doses were propagated to the CCS by combining the previously obtained deformation fields.

2.1.3. Evaluation of elastic image registration

The assessment of the performances of an EIR method is fundamental to estimate the uncertainties of the spatial normalization or even to determine an exclusion criterion to remove from the successive statistical analysis the subjects in which EIR failed to produce accurate results.

A simple and powerful, yet qualitative, method to evaluate the success of the anatomical matching is the visual inspection of the voxel-wise averages and standard deviations of the patients' anatomical images used to guide the registration, as computed before and after spatial normalization (Fig. 4). A sharp appearance of the voxel-wise average is associated with a consistently good anatomical alignment of the patients' image to the reference, while a blurred look indicates a diffuse mismatch of anatomical structures among the subjects. On the other hand, the standard deviation highlights misalignments in the image edges, and is particularly sensitive at pointing out the presence of potential outliers in an overall accurate spatial normalization of the cohort.

As for the quantitative evaluation, the most commonly used geometrical scores are the Dice Index (DI) and the Hausdorff Distance (HD), which reflect the overall geometric overlap between structures belonging to a patient and the corresponding ones in the reference.

The DI [55] is a similarity measure between sets:

$$DI(\Omega_i, \Omega_\rho) = \frac{2 |\Omega_i \cap \Omega_\rho|}{|\Omega_i| + |\Omega_\rho|}$$

where Ω_i and Ω_ρ are the structures under analysis for the i -th patient and the reference, respectively.

The HD [56] measures the maximum distance between the structures and is defined as

$$HD(\Omega_i, \Omega_\rho) = \max(d_1(\Omega_i, \Omega_\rho), d_1(\Omega_\rho, \Omega_i))$$

with

$$d_1(Q, W) = \min_{q \in Q} \max_{w \in W} \|q - w\|$$

A modified version of the HD (MHD) [57] keeps into account the Euclidean distance between the points of the structures but retains a higher robustness against the presence of registration outliers:

Fig. 4. Visual assessment of the accuracy of the match of planning CTs from a cohort of lung cancer patients before (first column) and after (second column) the spatial normalization via elastic image registration. Voxel-wise averages (second row) and standard deviations (third row) of the individual anatomical images are shown in the RGB color-coded fusion images (first row; red channel for standard deviation and cyan channel for average). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\text{MHD}(\Omega_i, \Omega_p) = \max(d_2(\Omega_i, \Omega_p), d_2(\Omega_p, \Omega_i))$$

where

$$d_2(Q, W) = \frac{1}{|Q|} \int \inf_{w \in W} \|q - w\| dq$$

However, these measures only reflect overall geometric overlap between structures and do not show local mapping errors within them, consequently they have to be interpreted with caution. In addition, the values of DI and MHD strictly depend on the structure size: for a given mismatch between the structure boundary, the smaller the structure, the lower the DI and the higher the MHD. Consequently, since a cut-off value cannot be defined *a priori*, these metrics are not adequate for comparing EIR methods if they are applied on different anatomical regions and structures. They are, instead, suitable to compare the organ overlap before and after the EIR process, thus obtaining a measure of the anatomical match improvement.

In order to get a first insight on the EIR behavior inside the structures, Mean Centerline Dispersion (MCD) or Center of Mass Standard Deviation (CMSD) can be measured.

MCD [29] can be used on tubular organs and is a measure of dispersion between the structure centerlines:

$$\text{MCD}(\Lambda_i, \Lambda_p) = \int_0^1 \|\Lambda_i(t) - \Lambda_p(t)\| dt$$

where Λ_i and Λ_p are the centerlines of the structure under analysis for the i -th patient and the CCS, respectively. The centerlines are described using B-splines and are parametrized in the unit interval.

CMSD [31] is a measure of the registration uncertainties and is given by

$$\text{CMSD} = \sqrt{\frac{\sum_{i=1}^P (\vec{R}_i - \langle \vec{R} \rangle)^2}{P - 1}}$$

where P is the number of patients in the cohort and $\vec{R}_i$ is the center of mass of the structure under analysis for the i -th patient.

To shed more light on the accuracy of the deformation within the organ (*i.e.*, at the voxel level) similarity metrics, such as the Root Mean Squared Error (RMSE) and the Mutual Information (MI), have to be used. Their choice depends on the relationship between contrasts in the reference and the individual images of the patients. If we assume that equivalent points in the patients' images have the same intensity, RMSE can be used (*e.g.*, in the case of study-specific CCS), while MI provides meaningful results even if the reference shows radiometric properties

Fig. 5. Sagittal and axial images of significantly activated brain voxels in the dead salmon during the mentalizing task, according to a threshold of 0.001 on uncorrected p values and 3 voxel extent of connected clusters. Reprinted from (Bennet) with permission from the authors.

different from the patients' image (e.g., if a synthetic phantom is chosen as reference).

RMSE [58] is defined as

$$\text{RMSE}(I_i, \rho) = \sqrt{\frac{1}{|\Omega|} \int_{\Omega} |I_i(\vec{x}) - \rho(\vec{x})|^2 d\vec{x}}$$

MI [59] is given by

$$\text{MI}(I_i, \rho) = \int_{\rho[\text{CCS}]} \int_{I_i[\Omega_i]} p(x, y) \log \frac{p(x, y)}{p_i(x)p_p(y)} dx dy$$

where $p(x, y)$ is the joint probability density function of I_i and ρ , and $p_i(x)$ and $p_p(y)$ are the marginal probability density functions of I_i and ρ , respectively.

In the context of RT, however, different mismatches quantified by the same anatomical matching metrics may have highly different impact of the VB analysis, according to their locations. Moderate misalignments in homogeneous dose regions have almost no consequence in the statistical analysis, but even small misalignments occurring in regions with a steep dose gradient could result in sensibly different VB results. To take into account this issue, specific dosimetric measures have been conceived: the Relative Difference of Area (RDA) of DVHs and the Dose Organ Overlap (DOO).

The RDA [26] evaluates DVH differences after dose mapping, taking into account the information in the CCS. It is computed as:

$$\text{RDA}(\text{DVH}_i, \text{DVH}_\rho) = \frac{\int_0^{D_{\max}} |\text{DVH}_i - \text{DVH}_\rho| du}{\max\{\int_0^{D_{\max}} \text{DVH}_i du, \int_0^{D_{\max}} \text{DVH}_\rho du\}}$$

where DVH_i and DVH_ρ are computed for the warped dose $D_{i,\tau}$ on the warped Ω_i and on Ω_{Ref} , respectively.

The DOO [29] is a score that measures the coincidence of both the organs and dose distribution. It is defined as

$$\text{DOO}(D_i, \Omega_i, \Omega_\rho) = \frac{\int_{\Omega_i \cap \Omega_\rho} D_i(\vec{x}) d\vec{x}}{\int_{\Omega_i \cup \Omega_\rho} D_i(\vec{x}) d\vec{x}}$$

2.2. Statistical analysis

Once the dose maps are spatially normalized, we can move on to the voxel-based statistical analysis of dose distributions. Conceptually, the analysis of dose maps can be designed following an arbitrary hypothesis test conceived for standard global explanatory variables (EVs). Classical

examples of analysis include:

- Two-sample tests, in which patients' dose maps may be grouped, for instance, according to a side effect or to the tumor control;
- Regression analysis, in which patients' dose maps may be correlated to an ordinal or continuous variable (e.g. toxicity score, organ function measures, survival, etc.);
- Analysis of variance between multiple groups comparing, for instance, dose maps of patients treated by different RT modalities.

2.2.1. Multiple comparison problem

With the possible exception of the regression analysis, which can also estimate voxel-wise a dose-response curve, all the previous examples are performed to derive a map of the statistical significance of a given effect. A most intuitive approach consists in testing voxel-wise the null hypothesis, for instance on dose differences between the groups of patients. Therefore, one could voxel-wise compare the local dose values by a classical parametric (e.g. t) or non-parametric (e.g. U) test, thus filling voxel-by-voxel a significance map of dose differences. However, in this approach there is a relevant issue, which is related to the massive multiple comparison problem that arises when testing a null hypothesis on a huge number of variables, namely the dose delivered to each voxel of the reference domain. The issue has been long studied in the genomics [60] and neuroimaging [22] setting, two very data-intensive disciplines with a high risk for false positives. In particular, Bennett et al. [61] performed a functional MR imaging study on a dead salmon to illustrate the impact of this issue via a *reductio ad absurdum*. The mentalizing task for the dead salmon formally consisted in determining the emotions expressed in a series of pictures of humans. During the administration of the task, the salmon was scanned to detect the Blood Oxygen Level Dependent (BOLD) contrast. The BOLD contrast was then processed according to the voxel-wise t -test, and two clusters of seemingly significant brain activation were found inside the dead salmon CNS (Fig. 5), which were clearly false positives (or type I errors). In the same study, it was specifically showed that lowering by some amount the standard p -value threshold α for statistical significance or setting a minimum size for the activation clusters are ineffective at controlling type I errors. The reason is easily found by considering the definition of α , which implies that if N independent tests are performed, the probability of claiming at least one false positive is given by $1-(1-\alpha)^N$. Alternatively, one could think that the expected fraction of false positives relative to the total number of tested hypotheses can be as high as α . From a still different perspective, for large N value, the non-negligible fraction of false positives could represent the entire set of seemingly

activated voxels. This is exactly what was found in the educational experiment on the dead salmon.

In order to counteract the multiple comparison issue, two main concepts are usually introduced: the Family-Wise Error Rate (FWER) and the False Discovery Rate (FDR).

2.2.2. Family-wise error rate

The FWER is defined as the probability that at least one false positive is found in a family of tests. Several strategies have been developed to control the FWER.

The most common approach in studies involving multiple mutually independent tests is to use the Bonferroni correction or one of its more sophisticated variants, like the Holm-Bonferroni or the Šidák corrections. However, in particular for VB analyses, in which the signal in neighboring voxels are likely to be highly correlated, such corrections are highly conservative, thus leading to a very low power of the test and increasing the incidence of false negatives.

Conversely, in the image analysis, two schemes are highly effective at controlling the FWER with a justified impact on the test power.

The Random Field Theory (RFT) [62] keeps into account the correlation among voxel signals by introducing the concept of resolution element (resel), which intuitively corresponds to the volume inside which the spatial coherence extends within the image. Assuming that the considered statistic t under the null hypothesis should be drawn from a specific known distribution (e.g., the normal distribution), the RFT identifies the statistic threshold for controlling the FWER by modelling the map of the null statistic as a smooth homogeneous random field (RF) evaluated over the specified lattice of voxels. For such mathematical entities, it is possible to calculate – in terms of the number of resels within the field of view – the expected value $\bar{\chi}(t)$ of the Euler Characteristic χ for a given superlevel set $L_T^+(\text{RF}) = \{\vec{x} \in \Omega | \text{RF}(\vec{x}) > T\}$ of the RF. That topological invariant of a space X essentially expresses the number of connected components minus the alternating sum of the number of κ -dimensional holes:

$$\chi(X) = \sum_{\kappa=0}^{\infty} (-1)^{\kappa} \text{rank} H_{\kappa}(X)$$

where $H_{\kappa}(X)$ represents the κ -th homology group of X . At high values of the threshold T , it can be shown that holes tend to disappear in $L_T^+(\text{RF})$, so that, for T close to the global maximum Θ of RF, $\chi(L_T^+(\text{RF}))$ is 0 if $T > \Theta$, and 1 otherwise. Therefore, for a given significance α level corrected for FWER, it is possible to derive the corresponding threshold on the statistic as $\bar{\chi}^{-1}(\alpha)$, which can be considerably lower than Bonferroni threshold, at least if the spatial coherence in the dataset extends for several voxels [22].

However, the RFT relies on several assumptions that may not be justified in a number of VB analyses, especially in the context of RT. Among those, it is often said in functional MR imaging studies that the resel size is determined by the full width at half maximum of the smoothing kernel used to filter the inaccuracies of the spatial normalization of the data [22]. However, the intrinsic spatial correlation of the original data can be far greater than the one introduced by the smoothing process, in particular in the cases of nuclear medicine images or dose distributions. Some strategies exist to blindly derive the resel size of a given dataset from the analysis of the least-squares residuals of the statistical model fitted voxel-wise, without the knowledge of the smoothing details [22,63]. Unfortunately, the spatial dependence of the derived resel size explicitly highlights a potential flaw in the homogeneity assumption of the RFT, which requires extra efforts to be overcome [22].

Most importantly, however, the assumptions of a parametric approach, such as the RFT, are not always met. The non-parametric alternative provided by the permutation test theory relies on minimal assumptions and overcomes, in a convenient and simple way, a large number of RFT limitations [22,64,65]. Essentially, if a statistic t is

designed for a voxel-wise analysis according to a given test, the $t_0(x_j)$ map is first computed for the original dataset. Then, assuming that the classification variable of each subject (e.g., the RT outcome) – hereafter label – can be exchanged between the subjects under the null hypothesis, M new datasets are derived by reshuffling the labels according to M random permutations (in the following we assume that $P! \gg M \gg 1$). Without making any assumption on the expected null distribution of the statistic, it would be then possible to find voxel-wise a statistic uncorrected threshold $t_{\text{th}}(x_j)$ corresponding to a given α level by finding the $(1-\alpha)$ quantile of the t distribution as it could be estimated by the $\{t_k(x_j) \mid k = 1, \dots, M\}$ histogram on the M permutations for the j -th voxel. Conversely, in order to find a threshold which controls the FWER to an α level, it is sufficient to use a single statistic threshold $t_{\text{th}0}$ for all the voxels, defined as the $(1-\alpha)$ quantile of the distribution of the maximum values of the statistic map $\{\max_j t_k(x_j) \mid k = 1, \dots, M\}$. Hence, the null hypothesis is rejected in voxels in which $t_0(x_j) > t_{\text{th}0}$, instead of the weaker condition $t_0(x_j) > t_{\text{th}}(x_j)$.

A higher power of the permutation test can be achieved by a step-down version of the above method, which iteratively identifies the most significant voxel and cuts it out before assessing the remaining voxels. Strictly speaking, the t values from the unpermuted dataset are sorted descendingly as $\{t_0(x_{s_j}) \mid j = 1, \dots, N\}$. Then, if we define for arbitrary j and k

$$\tilde{t}_k(x_{s_j}) = \max\{t_k(x_{s_j}), \dots, t_k(x_{s_N})\}$$

it is possible to count how many times the t statistic on a given voxel s_j can be more extreme than the observed value just by chance:

$$\xi_j = |\{t_0(x_{s_j}) < \tilde{t}_k(x_{s_j})\}|$$

Hence, the adjusted p values can be finally computed as

$$p(x_{s_j}) = \begin{cases} \frac{\xi_j}{M}, & j = 1 \\ \max\left(p(x_{s_{j-1}}), \frac{\xi_j}{M}\right), & j > 1 \end{cases}$$

The above procedure highlights a relevant issue tied to the randomness introduced in the p estimation by the casual choice of M permutations. The events counted by ξ_j are roughly drawn from a binomial distribution with parameters M and $p(x_{s_j})$, hence the fraction ξ_j/M is approximately normal with mean $p(x_{s_j})$ and standard deviation $\sqrt{p(x_{s_j}) \cdot [1 - p(x_{s_j})]/M}$. Therefore, in order to keep the relative uncertainty of the p estimation below some value η at a given confidence level γ , the number of permutations should be

$$M > 2 \cdot \frac{1-p}{\eta^2 \cdot p} \cdot (\text{erf}^{-1} \gamma)^2$$

Typically, 10^4 permutations are computed in order to obtain a 95% confidence level uncertainty smaller than 10% at p values as low as 0.05.

Permutation tests based on the maximum statistic has been designed in a number of studies on RT outcome analysis. In a great part of studies on independent two-samples test, the voxel-wise t statistic is based on the local doses and essentially expresses the effect (i.e., the dose difference between groups) normalized by the estimated standard deviation of the dose [5,25,28,34,39]. In presence of noisy data, it has been suggested that a more powerful test can be obtained with the adoption of a pseudo- t statistic, in which the voxel-wise standard deviation, whose estimation is intrinsically noisier than the effect evaluation, is pooled with those of neighbor voxels to produce a smooth standard deviation map [22].

To include non-dosimetric covariates in the analysis, a well-established strategy consists in defining voxel-wise a General Linear Model (GLM) in which the dose represents the dependent variable and the RT outcome, as well as the non dosimetric covariates, are included as explanatory variables [32]. To deal with the special challenge of GTV

inside the organ-at-risk in radiation-induced toxicity analyses [35,36], two alternative strategies can be adopted. One consists in computing, for each voxel, the t statistic based on the only patients in which the voxel falls out of the GTV censoring mask. The other requires to include in the GLM the GTV mask as a voxel-wise explanatory variable.

Beside the valuable robustness and simplicity of the permutation testing, the permutation approach perfectly fits the Threshold-Free Cluster-Enhancement (TFCE) strategy. The TFCE technique was conceived to continuously boost belief in extended areas of signal without relying on arbitrary settings of cluster-based thresholding [66]. Briefly, each $t_k(x_j)$ map is filtered according to:

$$\text{TFCE}[t_k](\vec{x}) = \int_0^{t_k(\vec{x})} |\hat{L}_h^-(t_i)|^E h^H dh$$

where $\hat{L}_h^-(t_k)$ represents the connected sublevel set of t_k at level h that includes $\vec{x}$. The parameters E and H rule the weight assigned to different types of signal clusters in the statistic map. In particular, a higher value of H assigns larger weights to higher clusters, whereas increasing E gives more weight to spatially larger clusters.

Notably, in the particular case of DSM analysis, the notion of connectedness of the sublevel sets in the DSM domain may need to be fixed in order to restore the topology of the organ manifold (e.g., in the case of the rectum an equivalence relation can be naturally defined for the points with the same axial coordinate and azimuth equal to 0 or 2π , so that the quotient space of the DSM domain under such relation can be naturally identified with a cylinder).

The permutation test is then run on $\{\text{TFCE}[t_k]\}$ rather than on plain $\{t_k\}$. With a proper choice of TFCE parameters it is possible to sensibly enhance the power of the test, still guaranteeing a strong control over FWER. For instance, the RFT shows that, by preprocessing the statistic map with a TFCE filter with $H = 2$ and $E = 2/\dim(\Omega)$, it is possible to approximate the behavior that a Fisher's combined probability test would have under the assumption of independent tests, even though for 3D applications an empirical rule sets $E = 0.5$ [66].

A comparison between the classical permutation test on the unfiltered statistic map and the test on TFCE-processed data was provided in [34]. The authors showed that, adopting the TFCE, it was possible to highlight a far larger cluster S_{TFCE} of significant dose differences than the one found without statistic enhancement (S_{NoFilt}). In particular, given the cluster S_{Uncorr} of significant dose differences derived with no correction for multiple comparison, it was shown that $S_{\text{NoFilt}} \subset S_{\text{TFCE}} \subset S_{\text{Uncorr}}$ (Fig. 6).

2.2.3. False discovery rate

The FDR represents the expected proportion of false positives among the rejected null hypotheses. Controlling the FDR at level α is therefore less conservative than controlling the FWER at the same level.

The most common strategy to control FDR is the Benjamini-Hochberg [67] procedure, which was devised under the hypothesis of independent tests. Briefly, the uncorrected p -values are sorted ascendingly as $\{p_{s_j} | j = 1, \dots, N\}$, so that $p_{s_{j_1}} \leq p_{s_{j_2}}$ if $j_1 < j_2$, and the largest J is

found such that $p_{s_J} < \alpha J/N$, where N represents the number of tests (i.e., voxels). Then, the null hypothesis corresponding to $\{p_{s_j} | j = 1, \dots, J\}$ are rejected. It can be shown that the Benjamini-Hochberg procedure controls conservatively the FDR in case of independent or positive dependent test statistics, whereas, in general case of arbitrary correlation structure between voxels, the Benjamini-Yekutieli procedure [68] should be applied by substituting α with

$$\frac{\alpha}{\sum_{j=1}^N \frac{1}{j}}$$

Compared to the FWER, the correction for FDR has been applied in a much lower number of RT studies. An example can be found in [30], where the Benjamini-Hochberg procedure was applied to derive a proper cutoff for the voxel-wise significant p -values of dose differences associated to rectal bleeding in prostate cancer patients.

2.2.4. How to compare p -maps

Significance maps derived according to different statistical tests can be compared by suitable concordance measures.

A class of distances parametrized by the real parameter $a \geq 1$ can be defined by considering the p -maps as elements of a Banach space given by the set of measurable real functions on the analyzed organ Ω , equipped with the natural generalization of the a -norm:

$$A_a = \left(\int_{\Omega} |A|^a d\lambda \right)^{1/a} = \left(\int_{\Omega} |A(x)|^a dx \right)^{1/a}$$

for a generic significance map A . In this context, given the finite extension of the organ, it is possible to define a normalized Minkowski distance between two significance maps A and B as:

$$d_a(A, B) = \left(\frac{1}{\lambda(\Omega)} \int_{\Omega} |A - B|^a d\lambda \right)^{1/a}$$

Being the probability maps $0 \leq A, B \leq 1$, it follows that $0 \leq d_a \leq 1$. A well-known result of the Jensen's inequality guarantees that, for given A and B maps, d_a is an increasing function of a , ranging from the arithmetic mean of the voxel-wise differences between A and B (for $a = 1$) to their maximum (for $a \rightarrow \infty$). To summarize the a -distances, it may be useful to define a global metric as

$$\bar{d}(A, B) = \int_0^1 d_{1/n}(A, B) dn$$

which ranges from 0 (for perfect match) to 1 (for the worst match).

A different approach consists in measuring the match of the positive clusters at a given significance level p . To this aim, a similarity measure, parametrized by $p_0 < p \leq 1$ ($p_0 = \min[\min(A), \min(B)]$), can be defined as

$$\text{DI}_p(A, B) = \text{DI}[L_p^-(A), L_p^-(B)]$$

where $L_p^-(Y)$ is the sublevel set of Y at level p . An overall measure of

Fig. 6. Voxel-based analyses of radiation induced lung damage in a cohort of Hodgkin's lymphoma survivors. Coronal view of lung subregions showing a statistically significant dose difference between groups according to permutation test on max t (left), permutation test on max TFCE(t) (center) and uncorrected voxel-wise t -test (right). Reprinted from [34] with permission from Elsevier.

the DI_p similarity between maps is provided by the mean DI_p

$$\overline{DI_p}(A, B) = \frac{1}{1 - p_0} \int_{p_0}^1 DI_p(A, B) dp$$

which ranges from 0 (for the worst match) to 1 (for the perfect match).

In the context of functional MR studies, a simplified version of the DI_p measure was used to face the problem of statistic map comparison [69], along with a visual representation of map differences based on the Bland-Altman plots [70]. In particular, for a given pair of statistic maps A and B , a density image shows in the form of a 2D histogram the relationship between the average (abscissa = $[A + B]/2$) and the difference of statistic values (ordinate = $A - B$) at corresponding voxels. The degree of disagreement between A and B is therefore assessed by the distance of points from the abscissa axis.

The above measures are useful if we are interested in an absolute comparison of the significance maps. For instance, we may be evaluating the impact of two different VB methods (e.g., different spatial normalization approaches, alternative correction procedures for multiple comparisons, etc.) on spatial patterns that can be identified for the same radiobiological issue (i.e., null hypothesis) from the same dataset: both d_a and DI_p could successfully quantify the different power of two statistical schemes or the localization of an effect according to alternative EIR algorithms.

However, as long as we change the dataset (e.g., with a different cardinality or with a different outcome distribution) or the null hypothesis (e.g., by the introduction of covariates or by changing the outcome), the measures provided by d_a and DI_p could be misleading, since we just cannot expect that the effect is being evaluated with the same power or is even the same. In this case, a different similarity measure DI_V , first described in [32], can be adopted to compare equivalent clusters (of relative volume $0 < V \leq 1$) of most significant voxels in each map. If $g_Y(p)$ is the relative volume of $L_p^-(Y)$

$$g_Y(p) = \frac{\lambda[L_p^-(Y)]}{\lambda[\Omega]}$$

then

$$DI_V(A, B) = DI[L_{g_A^{-1}(V)}^-(A), L_{g_B^{-1}(V)}^-(B)]$$

From an alternative viewpoint, the DI_V corresponds to the DI_p computed on the cumulative distribution functions of the two significance maps:

$$DI_V(A, B) = DI_p(F_A \circ A, F_B \circ B)$$

In [35] it was shown that, if the level sets of the significance maps have measure zero (i.e., $\lambda[L_p(Y)] = 0$ for any $0 < p \leq 1$), then the mean DI_V

$$\overline{DI_V}(A, B) = \int_0^1 DI_V(A, B) dV$$

ranges between $1 - \log 2$ (for the worst match) and 1 (for the perfect match).

In several contexts, the DI_V could provide a suitable way to quantify the spatial trends of significance, by looking at the overlap of the equally ranked voxels. For instance, in [32] DI_V provided a robust way to compare the spatial signatures of radiosensitivity inferred according to two different radiobiological hypotheses (namely, by including or not the effect of age on the observed radiation-induced toxicity). The inclusion of an additional explanatory variable could indeed have reduced the power of one of the VB analyses. Therefore, the p values in each voxel should be more insightfully replaced by their statistical rank, which represents the discrete version of the cumulative distribution function.

Similarly, the DI_V could be exploited to quantify the match of organ radiosensitivity associated to different toxicity outcomes [35,36] or

even the overlap between anatomic regions significantly spared by an RT modality and the clusters in which doses were significantly correlated with radiation-induced morbidity [36].

3. The NTCP perspective

The above described VB approach represents a suitable scheme to identify regions in which significant correlations exist between a clinical outcome, such as a radiation induced morbidity, and the local dose. A further advance of the VB approach to RT outcomes consists in the improvement of the NTCP modelling by the inclusion of the regional inference, in order to account for the spatially driven organ response to dose. Indeed, by VB analysis, we can identify the mean dose to significant regions, at an arbitrary significance level as a toxicity predictor. However, on an RT planning level, relying on such predictors might be a simplification as setting a single constraint on the DVH of an organ at risk.

An attempt to include spatial location of delivered dose in NTCP models was first performed in some *in silico* simulations, in which both the number and the spatial location of radiation-sterilized functional subunits played a role in defining complication probability (the so called “clusters models”) [7]. The cluster model relies on the assumption that a given volume receiving at least a certain dose is more likely to induce a complication if the volume is associated to a connected cluster of functional subunits rather than to scattered hot spots [71].

Another example of the incorporation of spatial information into radiation-induced morbidity prediction is provided by the research from Buettner et al. [44]. For the prediction of rectal bleeding after prostate cancer RT, the authors used the rectum DSM as input for a classification method based on locally connected neural networks able to take into account the topology of the input. A further rectal toxicity predictive model based on parameterized representation of the DSM was proposed in [72]: several dose levels were used to threshold the rectal DSM and a non-linear kernel-based probabilistic model was fed a variety of geometrical measures of the obtained clusters (including spatial extent and eccentricity).

A similar approach to the prediction of gastro-intestinal toxicity was proposed in [73], where the dose parameterization via extent measures of the DSM superlevel set was used for univariate and multivariate logistic regression analyses.

A method to incorporate the spatial information about the 3D dose distribution into a NTCP model has been proposed by Vinogradskiy et al. [74]. In particular, the model was conceived to incorporate the lung dose spatial information into a predictive radiation pneumonitis model. To this end, a Lyman-Kutcher-Burman (LKB) model with n set to 1 is refashioned in order to weight the contribution of the dose to each voxel according to an a priori user-defined spatial weighting map w :

$$D_{eff} = \int_{\Omega} D \cdot (1 + Cw) d\lambda$$

where C is a model fitting parameter that modulates the strength of the spatial weight.

The xerostomia induced by the irradiation of salivary glands in head-and-neck cancer patients was modeled by predictive schemes based on the spatial distribution of delivered dose [75]. The commonly acknowledged high volume effect in the parotids and the sub-mandibular glands allowed for obtaining good performance from a ridge logistic regression model based on the doses to each voxel and other non-dosimetric variables. The approach was also effective at highlighting the regression coefficient spatial patterns, which are suggestive of the gland regions that are most influential on the development of radiation-induced toxicity.

A strategy to extract an NTCP model from the results of a VB sample analysis has been suggested in [33,34]. Once an outcome-related sub-region $L_p^-(A)$ has been identified in the significance map A , it is propagated from the CCS back to each individual native space Ω_i and the

Fig. 7. Schematic representation of the backbone of the PACE normal tissue complication probability (NTCP) scheme. For each voxel, training patients' outcome are linked to the local dose through a logistic regression. Then the PACE prediction on a new patient is estimated computing voxel-wise by the list of logistic regressions a collection of morbidity odds (P map) and associated reliability measures (W map), which are combined to extract the actual NTCP value. Reprinted from [37] with permission from Frontiers Media.

mean doses to each subregion are used to train a predictive model (e.g., via a logistic regression).

A new formalism to predict arbitrary radiation-induced morbidity has been recently presented in [37]. The authors developed a Probabilistic Atlas for normal tissue Complication Estimation in radiation therapy (PACE) able to include, for the first time, regional dose information blindly extracted by a VB regression analysis along with non-simetric variables, allowing for an auto-tuning of the volume effect. The adopted strategy basically consists in keeping the general structure of the classical LKB model

where

$$t = \frac{gEUp - T_{p50}}{\mu \cdot T_{p50}}$$

The usual generalized equivalent uniform dose (gEUD) formula [76] is replaced by a generalized equivalent uniform probability, which is computed by substituting the dose distribution with a collection of outcome odds p , specifically estimated for the spatial position of local dose release (Fig. 7) along with the associated confidence interval (CI):

$$gEUp = \left[\frac{\int_{\Omega} p^{1/\nu} \cdot CI^{-1} d\lambda}{\int_{\Omega} CI^{-1} d\lambda} \right]^{\nu}$$

$$PACE = \frac{1 + \operatorname{erf} \frac{t}{\sqrt{2}}}{2}$$

Similarly to LKB, the PACE model contains three free parameters: ν that accounts for the volume effect; μ controlling the slope of the response sigmoid; and T_{p50} that expresses the tolerance probability.

The PACE model was demonstrated both *in silico* and on a clinical database of thoracic cancer patients, paving the way to a new perspective in NTCP modelling.

4. Closing comments

In the last decade the radiation oncology community become ever more aware of the need for methods capable of evaluating local dose response patterns in order to go beyond the traditional organ-based philosophy of NTCP modelling. Many efforts have been made to include spatial information into prediction of radiation-induced normal tissue toxicity. Among them, new 2D or 3D approaches have been introduced in the context of radiation oncology and in present cookbook we gave a self-contained description of the mathematical methods involved. First, the critical aspects of the spatial normalization of the different anatomies in the analyzed cohort were addressed. In particular, for hollow organs, such as the rectum, esophagus or bladder, wall effects could be considered through the extraction of DSMs. For other organs, instead, such as the lungs or the multiple muscles and structures in the head and neck region, probing the whole volume through a 3D approach is necessary. Then, the key elements related to the dose map comparisons were addressed, along with the possible options for a robust VB statistical analysis and the potential drawbacks related to the different solutions.

One of the major impact of the VB approach consists in the application of the inferred spatial patterns of the organ radiosensitivity on RT outcomes. This refinement of the traditional NTCP strategies, which goes beyond the strict definition of VB analyses, was nonetheless discussed in some detail due to our belief that it could pave the way to new clinical radiobiology studies and open new perspective for the identification of a personalized optimal RT plan [77].

The price for the power of the described VB tools is the complexity of the underlying mathematics. However, throughout the text both theoretical and technical issues inherent to VB analysis were elucidated and illustrated in relation to practical application found in the recent literature. In addition, the free tools for VB analysis available on line were reviewed and reported in the [Appendix](#).

Of course, with great power comes great responsibility: the flexibility of the VB schemes in describing non-trivial radiobiological mechanisms leads to models particularly prone to overfitting issues. Therefore, a proper validation of the findings, which can be still performed according to standard validation schemes, is even more essential than in traditional analyses.

A point that is tied to some extent to the above issue is the robustness of VB schemes against organ motion. Given that VB analyses are performed to look for findings that, in general, are not directly comparable to the results of a traditional DVH analysis, it could however be an inspiring question to ask whether the intra- or inter-fraction motion is more likely to affect the results of a VB or a DVH analysis. To deal with this problem, which has not been yet tackled in adequate detail, a metric should be first identified to bring onto a common battlefield two approaches that usually provide qualitatively different results. In this respect, a possible strategy could consist in reducing both VB and DVH studies into predictive models. This is already done for analyses leading to NTCP models arising from LKB or PACE schemes. Conversely, in case of more conventional sample analyses, the significant DVH parameters or the $L_p^-(A)$ anatomical clusters could be used to train a traditional logistic model or to follow the VB procedure in [33,34]. Therefore, the metric to evaluate the impact of motion on the different analyses can be the difference between the performances of classification models (the area under ROC curve could be useful) when considering the planned or the accumulated dose maps. Then, the robustness of the VB and the DVH statistical findings could be compared

by verifying if the performance difference in VB models is larger or smaller than the performance difference in DVH models.

In conclusion, we believe that this cookbook provides the ingredients and the step-by-step recipes that allow the reader to autonomously build up, starting from the scratch, its own VB analysis package, with the hope that this could help a conscious spread in the RT community of a class of powerful techniques.

Appendix. Available tools

Several free tools are available to implement the EIR and statistical methods described so far. It should be mentioned that so far there are no comprehensive tools to handle the entire VB pipeline, so that one has to pick a single tool for each VB step to build up one's own pipeline.

NiftyReg, Elastix and Plastimatch can be numbered in the group of registration tools.

NiftyReg (<http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg>) is a code based on the C/C++ language, but it is also available an accelerated version for GPU architecture based on CUDA. It implements rigid and affine registration algorithms based on a block-matching approach and a Trimmed Least Square scheme [78]. They use the normalized cross-correlation cost function and a multi-resolution approach. The EIR algorithm is based on the Free-From Deformation presented by Rueckert et al. [79] with B-splines parameterization. Normalized Mutual Information, with Bending-Energy penalty term, is used as cost function. NiftyReg is able to handle lesion censoring masks.

A more flexible solution is provided by Elastix (<http://elastix.isi.uu.nl>) [80,81], an open source software based on the Insight Segmentation and Registration Toolkit (ITK, <https://itk.org/itkindex.html>). It is a collection of rigid and non-rigid registration algorithms with a modular design that allows the user to configure different registration methods by means of a parameter file. It has a command-line interface, which easily enables the call of Elastix within an external batch or function. In addition, nowadays Elastix is accompanied by SimpleElastix [82], which makes it available in languages like C++, Python, Java, R, Ruby, C# and Lua. Elastix provides several choices for the similarity measure (Mean Squared Difference, Normalized Correlation Coefficient, Mutual Information, Normalized Mutual Information), with the possible addition of penalty terms, and for the optimizer of the cost function. It allows to implement multi-resolution registration strategies and the lesion-masking approach. Referring to EIR, Elastix provides B-spline and Thin-plate spline parameterization.

Plastimatch (<https://www.plastimatch.org>) is another open source software based on ITK. It can be used for different purposes, including image registration. As Elastix, it is highly configurable: it provides B-spline parametrization and MI or Mean Squared Error similarity measures. It supports lesion masking, but only in combination with B-spline transform and MI cost function. Compared to Elastix, Plastimatch provides also 4 different versions of the demons approach for EIR: fast symmetric forces demons [83], diffeomorphic demons [84], log-domain demons, symmetric log domain diffeomorphic demons [49]. Plastimatch can be used either via a command-line interface or from a more user-friendly plugin for 3D Slicer (<http://www.slicer.org>).

Other implementations of several versions of the Demons algorithm, different from the ITK ones, are available for Matlab (MATLAB, The MathWorks, Inc., Natick, Massachusetts, United States) on the File Exchange section of the MathWorks web site (<https://it.mathworks.com/matlabcentral/fileexchange/39194-diffeomorphic-log-demons-image-registration> and <https://it.mathworks.com/matlabcentral/fileexchange/39194-diffeomorphic-log-demons-image-registration>).

Of note, all the EIR methods implemented in the above mentioned tools have several options and the accuracy of the normalization results largely depends on the fine tuning of the algorithm parameter for the specific application rather than on the chosen tool itself.

As for the statistical analysis, the main two tools available are Statistical Parametric Mapping software (SPM) and Randomeise.

SPM (www.fil.ion.ucl.ac.uk/spm/) is a Matlab-based software package for statistical processing designed for the analysis of functional neuroimaging data. It contains a module for parametric statistical analyses which makes use of GLMs to describe the variability in the data in terms of experimental and confounding effects, and residual variability. The module uses Gaussian RFT to control the FWER. Nonetheless SPM does not allow to include GTV masks as voxel-wise EVs.

Randomise (<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Randomise>) [85] is an FSL [86] tool conceived for neuroimaging. It implements non-parametric permutation inference using GLM with the possibility to take into account global EVs and a voxel-wise one, e.g. the GTV masks of the patients. It also allows for a TFCE filtering of the statistic maps.

An experimental extension to Randomise is represented by Palm (<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/PALM>), a Matlab tool that implements several additional features, including new regression and permutation strategies.

References

- [1] Rubin P, Michael T, Milano, Louis S, Constine. ALERT• Adverse Late Effects of Cancer Treatment: Volume 2: Normal Tissue Specific Sites and Systems. 2013:189.
- [2] Shipley WU, Tepper JE, Prout Jr. GR, Verhey LJ, Mendiola OA, Goitein M, et al. Proton radiation as boost therapy for localized prostatic carcinoma. *JAMA* 1979;241:1912–5.
- [3] Tommasino F, Nahum A, Cella L. Increasing the power of tumour control and normal tissue complication probability modelling in radiotherapy: recent trends and current issues. *Translational. Cancer Res* 2017;S807–21.
- [4] Yorke ED. Modeling the effects of inhomogeneous dose distributions in normal tissues. *Semin Radiat Oncol* 2001;11:197–209.
- [5] Monti S, Palma G, D'Avino V, Gerardi M, Marvaso G, Ciardo D, et al. Voxel-based analysis unveils regional dose differences associated with radiation-induced morbidity in head and neck cancer patients. *Sci Rep* 2017;7:7220.
- [6] Trott KR, Doerr W, Facchetti A, Hopewell J, Langendijk J, van Luijk P, et al. Biological mechanisms of normal tissue damage: importance for the design of NCTP models. *Radiat Oncol* 2012;105:79–85.
- [7] Thames HD, Zhang M, Tucker SL, Liu HH, Dong L, Mohan R. Cluster models of dose-volume effects. *Int J Radiat Oncol Biol Phys* 2004;59:1491–504.
- [8] Dörr W, Herrmann T, Trott K-R. Normal tissue tolerance. *Trans Cancer Res* 2017;6:S840–51.
- [9] Seppenwoolde Y, De Jaeger K, Boersma LJ, Belderbos JS, Lebesque JV. Regional differences in lung radiosensitivity after radiotherapy for non-small-cell lung cancer. *Int J Radiat Oncol Biol Phys* 2004;60:748–58.
- [10] Improta I, Palorini F, Cozzarini C, Rancati T, Avuzzi B, Franco P, et al. Bladder spatial-dose descriptors correlate with acute urinary toxicity after radiation therapy for prostate cancer. *Phys Med* 2016;32:1681–9.
- [11] van Luijk P, Pringle S, Deasy JO, Moiseenko VV, Faber H, Hovan A, et al. Sparing the region of the salivary gland containing stem cells preserves saliva production after radiotherapy for head and neck cancer. *Sci Transl Med* 2015;7:305ra147.
- [12] Hourri J, Karunamuni R, Connor N, Pettersson N, McDonald C, Farid N, et al. Analyses of regional radiosensitivity of white matter structures along tract axes using novel white matter segmentation and diffusion imaging biomarkers. *Phys Imaging Radiat Oncol* 2018;6:39–46.
- [13] Cella L, D'Avino V, Palma G, Conson M, Liuzzi R, Picardi M, et al. Modeling the risk of radiation-induced lung fibrosis: Irradiated heart tissue is as important as irradiated lung. *Radiat Oncol* 2015;117:36–43.
- [14] Cella L, Palma G, Deasy JO, Oh JH, Liuzzi R, D'Avino V, et al. Complication probability models for radiation-induced heart valvular dysfunction: do heart-lung interactions play a role? *PLoS One* 2014;9:e111753.
- [15] Huang EX, Hope AJ, Lindsay PE, Trovo M, El Naqa I, Deasy JO, et al. Heart irradiation as a risk factor for radiation pneumonitis. *Acta Oncol* 2011;50:51–60.
- [16] Huang BT, Huang RH, Zhang WZ, Lin W, Guo LJ, Xu LY, et al. Different definitions of esophagus influence esophageal toxicity prediction for esophageal cancer patients administered simultaneous integrated boost versus standard-dose radiation therapy. *Sci Rep* 2017;7:120.
- [17] Kong FM, Ritter T, Quint DJ, Senan S, Gaspar LE, Komaki RU, et al. Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs, and brachial plexus. *Int J Radiat Oncol Biol Phys* 2011;81:1442–57.
- [18] Rutkowska E, Baker C, Nahum A. Mechanistic simulation of normal-tissue damage in radiotherapy—implications for dose-volume analyses. *Phys Med Biol* 2010;55:2121–36.
- [19] Rutkowska ES, Syndikus I, Baker CR, Nahum AE. Mechanistic modelling of radiotherapy-induced lung toxicity. *Br J Radiol* 2012;85:e1242–8.
- [20] Lakshminarayanan P, Jiang W, Robertson SP, Cheng Z, Han P, Bowers M, et al. Radio-morphology: Parametric shape-based features in radiotherapy. *Med Phys* 2019;46:704–13.
- [21] Ashburner J, Friston KJ. Voxel-based morphometry—the methods. *Neuroimage* 2000;11:805–21.
- [22] Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE, Penny WD. Statistical parametric mapping: the analysis of functional brain images. Academic press; 2011.
- [23] Whitwell JL. Voxel-based morphometry: an automated technique for assessing structural changes in the brain. *J Neurosci* 2009;29:9661–4.
- [24] Heemsbergen WD, Hoogeman MS, Hart GA, Lebesque JV, Koper PC. Gastrointestinal toxicity and its relation to dose distributions in the anorectal region of prostate cancer patients treated with radiotherapy. *Int J Radiat Oncol Biol Phys* 2005;61:1011–8.
- [25] Wortel RC, Witte MG, van der Heide UA, Pos FJ, Lebesque JV, van Herk M, et al. Dose-surface maps identifying local dose-effects for acute gastrointestinal toxicity after radiotherapy for prostate cancer. *Radiat Oncol* 2015;117:515–20.
- [26] Acosta O, Dreon G, Ospina JD, Simon A, Haigron P, Lafond C, et al. Voxel-based population analysis for correlating local dose and rectal toxicity in prostate cancer radiotherapy. *Phys Med Biol* 2013;58:2581–95.
- [27] Beasley W, Thor M, McWilliam A, Green A, Mackay R, Slevin N, et al. Image-based Data Mining to Probe Dosimetric Correlates of Radiation-induced Trismus. *Int J Radiat Oncol Biol Phys* 2018.
- [28] Chen C, Witte M, Heemsbergen W, van Herk M. Multiple comparisons permutation test for image based data mining in radiotherapy. *Radiat Oncol* 2013;8:293.
- [29] Dreon G, Acosta O, Lafond C, Simon A, de Crevoisier R, Haigron P. Interindividual registration and dose mapping for voxelwise population analysis of rectal toxicity in prostate cancer radiotherapy. *Med Phys* 2016;43:2721–30.
- [30] Dreon G, Acosta O, Ospina JD, Fargeas A, Lafond C, Correge G, et al. Identification of a rectal subregion highly predictive of rectal bleeding in prostate cancer IMRT. *Radiat Oncol* 2016;119:388–97.
- [31] McWilliam A, Faivre-Finn C, Kennedy J, Kershaw L, van Herk M. Data mining identifies the base of the heart as a dose-sensitive region affecting survival in lung cancer patients. *Int J Radiat Oncol Biol Phys* 2016;96:S48–9.
- [32] Monti S, Pacelli R, Cella L, Palma G. Inter-patient image registration algorithms to disentangle regional dose bioeffects. *Sci Rep* 2018;8:4915.
- [33] Mylona E, Acosta O, Lizée T, Lafond C, Crehan G, Magne N, et al. Voxel-based analysis for identification of urethrovaginal subregions predicting urinary toxicity after prostate cancer radiation therapy. *Int J Radiat Oncol Biol Phys* 2019.
- [34] Palma G, Monti S, D'Avino V, Conson M, Liuzzi R, Pressello MC, et al. A voxel-based approach to explore local dose differences associated with radiation-induced lung damage. *Int J Radiat Oncol Biol Phys* 2016;96:127–33.
- [35] Palma G, Monti S, Thor M, Rimmer A, Deasy JO, Cella L. Spatial signature of dose patterns associated with acute radiation-induced lung damage in lung cancer patients treated with Stereotactic Body Radiation Therapy. *Phys Med Biol* 2019.
- [36] Palma G, Monti S, Xu T, Scifoni E, Yang P, Hahn SM, et al. Spatial dose patterns associated with radiation pneumonitis in a randomized trial comparing intensity-modulated photon therapy with passive scattering proton therapy for locally advanced non-small cell lung cancer. *Int J Radiat Oncol Biol Phys* 2019.
- [37] Palma G, Monti S, Buonanno A, Pacelli R, Cella L. PACE: a probabilistic atlas for normal tissue complication estimation in radiation oncology. *Front Oncol* 2019;9:130.
- [38] Palorini F, Cozzarini C, Gianolini S, Botti A, Carillo V, Iotti C, et al. First application of a pixel-wise analysis on bladder dose-surface maps in prostate cancer radiotherapy. *Radiat Oncol* 2016;119:123–8.
- [39] Yahya N, Ebert MA, House MJ, Kennedy A, Matthews J, Joseph DJ, et al. Modeling urinary dysfunction after external beam radiation therapy of the prostate using bladder dose-surface maps: evidence of spatially variable response of the bladder surface. *Int J Radiat Oncol Biol Phys* 2017;97:420–6.
- [40] Witte MG, Heemsbergen WD, Bohoslavsky R, Pos FJ, Al-Mamgani A, Lebesque JV, et al. Relating dose outside the prostate with freedom from failure in the Dutch trial 68 Gy vs. 78 Gy. *Int J Radiat Oncol Biol Phys* 2010;77:131–8.
- [41] Hoogeman MS, van Herk M, de Bois J, Muller-Timmermans P, Koper PC, Lebesque JV. Quantification of local rectal wall displacements by virtual rectum unfolding. *Radiat Oncol* 2004;70:21–30.
- [42] Dankers F, Wijsman R, Troost EG, Monshouwer R, Bussink J, Hoffmann AL. Esophageal wall dose-surface maps do not improve the predictive performance of a multivariable NCTP model for acute esophageal toxicity in advanced stage NSCLC patients treated with intensity-modulated (chemo-)radiotherapy. *Phys Med Biol* 2017;62:3668–81.
- [43] Munbodr R, Jackson A, Bauer J, Schmidlein CR, Zelefsky MJ. Dosimetric and anatomic indicators of late rectal toxicity after high-dose intensity modulated radiation therapy for prostate cancer. *Med Phys* 2008;35:2137–50.
- [44] Buettner F, Gulliford SL, Webb S, Partridge M. Using dose-surface maps to predict radiation-induced rectal bleeding: a neural network approach. *Phys Med Biol* 2009;54:5139–53.
- [45] Rigaud B, Simon A, Castelli J, Lafond C, Acosta O, Haigron P, et al. Deformable image registration for radiation therapy: principle, methods, applications and evaluation. *Acta Oncol* 2019;1–13.
- [46] Hub M, Kessler ML, Karger CP. A stochastic approach to estimate the uncertainty involved in B-spline image registration. *IEEE Trans Med Imaging* 2009;28:1708–16.
- [47] Kanai T, Kadoya N, Ito K, Onozato Y, Cho SY, Kishi K, et al. Evaluation of accuracy of B-spline transformation-based deformable image registration with different parameter settings for thoracic images. *J Radiat Res* 2014;55:1163–70.
- [48] Klein S, Staring M, Pluim JP. Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. *IEEE Trans Image Process* 2007;16:2879–90.
- [49] Vercateren T, Pennec X, Perchant A, Ayache N. Symmetric log-domain diffeomorphic registration: A demons-based approach. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2008. p. 754–61.
- [50] Segars WP, Sturgeon G, Mendonca S, Grimes J, Tsui BM. 4D XCAT phantom for multimodality imaging research. *Med Phys* 2010;37:4902–15.

- [51] Murphy MJ, Wei Z, Fatyga M, Williamson J, Anscher M, Wallace T, et al. How does CT image noise affect 3D deformable image registration for image-guided radiotherapy planning? *Med Phys* 2008;35:1145–53.
- [52] Morrow NV, Lawton CA, Qi XS, Li XA. Impact of computed tomography image quality on image-guided radiation therapy based on soft tissue registration. *Int J Radiat Oncol Biol Phys* 2012;82:e733–8.
- [53] Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. *Med Phys* 2017;44:e43–76.
- [54] Monti S, Paganelli C, Buizza G, Preda L, Valvo F, Magliulo M, et al. A novel framework for spatial normalization of dose distributions in voxel-based analyses of brain irradiation outcomes. *Int J Radiat Oncol Biol Phys* 2019;105:S104.
- [55] Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945;26:297–302.
- [56] Hausdorff F. Dimension und äußeres Maß. *Math Ann* 1918;79:157–79.
- [57] Dubuisson MP, Jain AK. A modified Hausdorff distance for object matching. *Pattern Recognition*, 1994 Vol 1-Conference A: Computer Vision & Image Processing. Proceedings of the 12th IAPR International Conference on: IEEE. 1994. p. 566–8.
- [58] Yin Y, Hoffman EA, Lin CL. Mass preserving nonrigid registration of CT lung images using cubic B-spline. *Med Phys* 2009;36:4213–22.
- [59] Hassan AF, Cailin D, Hussain ZM. An information-theoretic image quality measure: Comparison with statistical similarity; 2014.
- [60] Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Stat Med* 2014;33:1946–78.
- [61] Bennett CM, Miller MB, Wolford GL. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction. *Neuroimage* 2009;47:S125.
- [62] Adler RJ, Taylor JE. Random fields and geometry. Springer Science & Business Media; 2009.
- [63] Worsley KJ, Evans AC, Marrett S, Neelin P. A three-dimensional statistical analysis for CBF activation studies in human brain. *J Cereb Blood Flow Metab* 1992;12:900–18.
- [64] Westfall PH, Young SS. Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons; 1993.
- [65] Holmes AP, Blair R, Watson J, Ford I. Nonparametric analysis of statistic images from functional mapping experiments. *J Cereb Blood Flow Metab* 1996;16:7–22.
- [66] Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* 2009;44:83–98.
- [67] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc: Ser B (Methodol)* 1995;57:289–300.
- [68] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 2001;29:1165–88.
- [69] Bowring A, Maumet C, Nichols TE. Exploring the impact of analysis software on task fMRI results. *Hum Brain Mapp* 2019;40:3362–84.
- [70] Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *J R Stat Soc: Ser D (The Statistician)* 1983;32:307–17.
- [71] Deasy JO, El Naqa I. Image-based modeling of normal tissue complication probability for radiation therapy. *Radiation Oncology Advances*. Springer; 2008. p. 211–52.
- [72] Buettner F, Gulliford SL, Webb S, Partridge M. Modeling late rectal toxicities based on a parameterized representation of the 3D dose distribution. *Phys Med Biol* 2011;56:2103.
- [73] Casares-Magaz O, Muren LP, Moiseenko V, Petersen SE, Pettersson NJ, Høyer M, et al. Spatial rectal dose/volume metrics predict patient-reported gastro-intestinal symptoms after radiotherapy for prostate cancer. *Acta Oncol* 2017;56:1507–13.
- [74] Vinogradskiy Y, Tucker SL, Liao Z, Martel MK. A novel method to incorporate the spatial location of the lung dose distribution into predictive radiation pneumonitis modeling. *Int J Radiat Oncol Biol Phys* 2012;82:1549–55.
- [75] Jiang W, Lakshminarayanan P, Hui X, Han P, Cheng Z, Bowers M, et al. Machine learning methods uncover radiomorphologic dose patterns in salivary glands that predict xerostomia in patients with head and neck cancer. *Adv Radiat Oncol* 2019;4:401–12.
- [76] Luxton G, Keall PJ, King CR. A new formula for normal tissue complication probability (NTCP) as a function of equivalent uniform dose (EUD). *Phys Med Biol* 2008;53:23–36.
- [77] Palma G, Monti S, Conson M, Pacelli R, Cella L. Normal Tissue Complication Probability (NTCP) Models for Modern Radiation Therapy. *Seminars in Oncology*. Elsevier; 2019.
- [78] Modat M, Cash DM, Daga P, Winston GP, Duncan JS, Ourselin S. Global image registration using a symmetric block-matching approach. *J Med Imaging (Bellingham)*. 2014;1:024003.
- [79] Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans Med Imaging* 1999;18:712–21.
- [80] Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging* 2010;29:196–205.
- [81] Shamonin DP, Bron EE, Lelieveldt BP, Smits M, Klein S, Staring M, et al. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Front Neuroinf* 2013;7:50.
- [82] Marstal K, Berendsen F, Staring M, Klein S. SimpleElastix: A user-friendly, multi-lingual library for medical image registration. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016. p. 134–42.
- [83] Rogelj P, Kovačič S. Symmetric image registration. *Med Image Anal* 2006;10:484–93.
- [84] Vercauteren T, Pennec X, Perchant A, Ayache N. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*. 2009;45:S61–72.
- [85] Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. Permutation inference for the general linear model. *Neuroimage*. 2014;92:381–97.
- [86] Woolrich MW, Jbabdi S, Patenaude B, Chappell M, Makni S, Behrens T, et al. Bayesian analysis of neuroimaging data in FSL. *Neuroimage*. 2009;45:S173–86.