International Journal of Radiation Oncology biology • physics

www.redjournal.org

EDITORIAL

Flogging a Dead Salmon? Reduced Dose Posterior to Prostate Correlates With Increased PSA Progression in Voxel-Based Analysis of 3 Randomized Phase 3 Trials

Jane Shortall, PhD,* Giuseppe Palma, PhD,† Hitesh Mistry, PhD,* Eliana Vasquez Osorio, PhD,* Alan McWilliam, PhD,*,‡ Ananya Choudhury, PhD,*,‡ Marianne Aznar, PhD,* Marcel van Herk, PhD,* and Andrew Green, PhD*

**Division of Cancer Science, University of Manchester, Manchester, United Kingdom; †National Research Council, Institute of Biostructures and Bioimaging, Napoli, Italy; and ‡Christie Medical Physics & Engineering, The Christie NHS Foundation Trust, Manchester, United Kingdom*

Received Nov 5, 2020, and in revised form Jan 7, 2021. Accepted for publication Jan 10, 2021.

Image-based data mining (IBDM) and voxel-based analysis (VBA) have shown great promise in the retrospective analysis of routine clinical data, offering a way to analyze a patient population without selection. In radiation therapy, IBDM analyzes whole dose distributions for their effect on a given outcome with no prior assumptions. Recent extensions in IBDM methodologies include per voxel survival analyses, allowing the generation of more robust and testable hypotheses. An essential aspect to produce solid conclusions includes applying the correct statistical techniques to account for multiple testing. There is a growing interest in applying IBDM/VBA techniques in radiation therapy, which is a relatively new area with limited examples in the literature. We present this editorial to discuss guidelines for best practice. In doing so, we highlight 3 recently published papers applying IBDM techniques to radiation therapy data[1-3] and suggest alternative, more robust analysis methods.

We focus on Marcello et al's recent paper "Reduced dose posterior to prostate correlates with increased PSA progression in voxel-based analysis of 3 randomized phase 3 trials."[1] The work aims to investigate "the association between 3-dimensional planned dose and prostate specific antigen (PSA) progression" in men with intermediate- and high-risk prostate cancer using 3 VBA statistical methods. The authors report regions posterior to the prostate where reduced dose was significantly correlated with PSA progression.

Multiple Comparisons and Survival Analysis

IBDM presents a serious multiple comparisons problem. This is due to the analysis of several million voxels against a single outcome, meaning that, were no action taken, many voxels would appear significant purely by chance. This challenge was identified in neuroscience research by Bennet et al,[4] who found significant activation of brain regions in a dead salmon shown pictures of human interactions.

Corresponding author: Andrew Green, PhD; E-mail: andrew.green-2@manchester.ac.uk

Disclosures: M.vH. and A.C. were supported by NIHR Manchester Biomedical Research Centre.

Int J Radiation Oncol Biol Phys, Vol. 110, No. 3, pp. 696–699, 2021
0360-3016/$ - see front matter
https://doi.org/10.1016/j.ijrobp.2021.01.017

The apparent significance of these activations was an artefact of having not adequately controlled for multiple comparisons in the analysis. Multiple comparisons can be corrected for by constructing the distribution of a summary statistic under the null hypothesis of no significant interaction. This is done by permuting the event labels when analyzing the data and then taking a threshold at, for example, the 95th percentile ($P = .05$) of the distribution. Since being proposed by Chen et al,[5] this approach has been applied to several IBDM analyses in radiation therapy.

This leads us to our first concern, which is exemplified in Marcello et al's analyses, for the implementation of VBA in radiation therapy dose distributions. To correctly address the multiple comparisons problem, a distribution of summary statistics should be constructed, for example the maximum T statistic over the field-of-view in each permutation, from which a threshold on T for significance can be derived. However, in all 3 of their papers, the authors have applied permutation testing on a per-voxel basis, without calculating T, using instead the difference in mean dose per voxel (which is a nonpivotal quantity) and without inspecting the distribution of a maximum statistic. We further note that, despite making reference to Chen et al,[5] their methodology is unclear.[1-3]

This is also relevant in the authors' application of per voxel Cox regression models. In their section on "univoxel" regressions, there is no mention of multiple comparison corrections at all; the authors instead rely on analytical P values per voxel. This is statistically incorrect and leads to spurious "significant" regions being identified, like the excited brain of a dead salmon.

We contend that the proper way of applying IBDM/VBA includes a full permutation test, permuting event labels for each patient in the analysis and recalculating the entire hazard ratio map. From these recalculated maps, a summary statistic can be derived, for example, the most extreme hazard ratio (which can be positive or negative), and used to construct the null distribution. Thresholds are then taken to determine a threshold hazard ratio for significance. This approach is illustrated by Green et al[6] using a lung cancer data set. The regions identified as significantly associated with overall survival in this work agree with those identified using a per-voxel t test on the same data.[7]

For in-depth discussion of permutation testing, we direct the reader to a critical tutorial review by Groppe et al.[8] We highlight permutation testing as an effective way of correcting for the multiple comparisons problem as it relies on minimal assumptions. There are other parametric approaches for dealing with the multiple comparisons problem; however, these can reply on assumptions that are not always met.[9]

The Bonferroni correction, where the desired alpha level is divided by the number of hypotheses to calculate a corrected P value, is a common approach to correcting for multiple comparisons and controlling the family wise error rate. However, this method does not consider the high correlation between neighboring voxels, which, as discussed later, should be considered in radiation therapy VBA and is therefore extremely conservative.[9]

Alternatively, random field theory, where images are split into small volumes representing a volume of special correlation (resolution elements or resels), does consider the correlation between voxels. Random field theory involves thresholding a smoothed image using Euler characteristics to control the family wise error rate. The derivation of the number and size of resels, which are spatially inhomogeneous in dose distributions, is an assumption that is difficult to justify in radiation therapy, however.[9] For further discussion of correcting for multiple comparisons we direct the reader to the voxel-based analysis methodological cookbook[9] and Groppe et al's[8] tutorial.

In their multivoxel Cox analysis, Marcello et al apply least absolute shrinkage and selection operator (LASSO) variable selection, with each voxel being considered as a variable in one large model. Although this approach is interesting, we disagree with the authors' claim that it accounts for multiple comparisons. LASSO tests many regularization penalty values using cross-validation to find the optimum turning parameter, λ, based on various criteria such as number of variables to be included in a model.[10] Although, as Cohen et al[10] explain, "these penalized methods are, implicitly, doing multiple testing of the potential factors that can enter the model," postselection inference requires care. For instance, the final model depends on the order of variables entering the model.[11,12] The LASSO variable selection process does not test any hypothesis and, as correcting for the multiple comparisons problem is not the "ultimate goal" of finding the optimum λ, "properties of this multiple testing aspect of the methodology are typically ignored."[10]

Hence, LASSO alone does not correct for multiple comparisons, and proper accounting for the problem would require further steps (eg, bootstrapping). We therefore believe that the analysis presented by the authors fails to properly consider multiple comparisons. A more suitable application of LASSO and a way to obtain realistic P values for selected variables are outlined by Taylor and Tibshirani.[13]

Furthermore, adjusted P values and postselection inference for LASSO and other classical variable selection models, such as stepwise regression, should be handled with care. For instance, the number of predictors we select from, and the order in which they are selected, affects the inference of a model. For example, the strongest predictor selected from a group of 100 variables should have a larger adjusted P value than if it were selected from a group of 10. Moreover, the appropriate P values should also be obtained for each step of the selection process, rather than for the first or last selection event only. Taylor et al[14] recommend writing the selection events in "polyhedral" form, using nested models to adjust for the effects of variable selection. We direct the reader to Taylor et al[14] for a more detailed

discussion of postselection inference using LASSO and stepwise regressions.

Marcello et al take an interesting approach to the intervoxel correlation in dose distributions (an unavoidable consequence of radiation therapy dose distributions), which they claim is accounted for by the application of LASSO. LASSO focuses on a few scattered voxels (those identified as predictive for PSA recurrence are scattered around the anatomy and are distant to organs at risk, lymph node, or other organ). Due to the shallow dose gradients used in radiation therapy, neighboring voxels are likely to have similar or correlated dose features. As such, it is questionable whether detecting a few scattered voxels really shows us anything meaningful. We propose that it is important to consider intervoxel correlations in the dose distributions between neighboring voxels when applying IDBM/VBA to radiation therapy to improve inference of any significant regions.

For this reason, it may be more appropriate to apply a regularization without variable selection (eg, Tikhonov regularization). Tikhonov regularization, where the signal is smoothed according to some cost function, has been used in functional magnetic resonance imaging (fMRI) studies to improve predictive models for brain functionality in neuroimaging.[15,16] The importance of spatial correlation is primary in fMRI neuroimaging to explore functional connectivity of the brain (ie, studying networks in the brain that are structurally and/or functionally connected).[17] Similarly, assessing intervoxel correlations during IBDM/VBA in radiation therapy could aid our understanding of dose-volume relationships and how organs behave as in a serial, parallel, or some hybrid, such as the rectum.

We suggest similar methods to those used in fMRI studies could be employed in the future to address issues of intervoxel correlations in radiation therapy dose distributions.

Additionally, any survival analysis performed as part of a predictive model should also be carefully implemented. Alongside the previously mentioned assertions, there are several curious choices made in Marcello et al's[1] application of survival analyses that we feel would be useful to highlight. In their uni-voxel analysis, dose (a continuous variable) is dichotomized at the median dose in each voxel independently. Aside from the fact that dichotomizing a continuous variable is poor statistical practice, dichotomizing each voxel independently means it becomes impossible to compare dose differences. It is also not stated which variables aside from dose were significant and included in the final model.[18] To obtain a meaningful predictive model, we recommend that continuous (eg, dose) and categorical (eg, Gleason grade) variables be handled appropriately and authors present their final models with clarity.

Spatial Registration

Aside from the issues regarding the statistical analysis used in this paper, we would query the fundamental data handling and processing. The first step in IBDM is to spatially normalize the radiation therapy data to a common frame of reference. This is achieved through the use of rigid and nonrigid registration, both of which are susceptible to uncertainty, partly due to the selection of a template anatomy.

Marcello et al test the effect of template selection by qualitatively assessing the patterns of dose-progression relationship in 3 templates. In this respect, we note that several quantitative metrics for evaluating the mismatch of the dose-response patterns are available to provide a more precise insight into the map's similarity (eg, Dice index over the volume or, more appropriately, the Dice index over P value). In addition, although this shows that the registration is largely insensitive to the selection of template anatomy, there is no attempt to assess the uncertainty introduced by the registration process itself. This is possible using structures defined in the routine radiation therapy workflow, as was done in McWilliam et al's[7] analysis of survival in lung cancer. A robust quantification of random registration uncertainty should be incorporated in the analysis by blurring the dose distributions. Information regarding the registration algorithm used by the authors to allow independent validation of the results is welcome.

Handling Patient Data

Although the study accesses a large cohort of high-quality data, its full potential has not been utilized. Intermediate- and high-risk patients in trial A (randomised androgen deprivation and radiotherapy [RADAR]) are analyzed, yet the authors "validate" results merely by repeating analysis for intermediate-risk patients only. Furthermore, there are inconsistent definitions of intermediate- and high-risk disease between cohorts; whereas patients with stage $\leq$T2a disease are classed as intermediate-risk in trial A, patients with disease stage T1b−T3a are defined as intermediate-risk in trials B and C (a randomised trial of high dose therapy in localized cancer of the prostate using conformal radiotherapy techniques [RT01], conventional or hypofractionated high dose intensity modulated radiotherapy for prostate cancer [CHHiP]). More consistent stratification of patients would have allowed validation of both intermediate- and high-risk groups.

Finally, the end-point used in this work is inconsistent between the 3 cohorts (nadir + 2 ng/mL [trials A and C] vs 150% nadir + 2 ng/mL at 6 months from beginning of radiation therapy [trial B]), with no clear mutual follow-up time. The authors "divided (patients) according to whether they experienced an end-point event at any time during follow-up"; binarizing the time-to-event variable like this means valuable information is lost. Furthermore, because the follow-up is inconsistent between patients, special care is required when censoring data. For example, a fixed maximum time-to-event, say T0, could be defined and patients divided based on whether they did or did not

experience a mutual end-point by T0. Patients who had follow-up <T0 and did not develop the endpoint during their follow-up time would be excluded from the study. As there is no mention of this type of data processing, it is possible that censoring was not properly considered.

Conclusions

IBDM has potential to generate robust and testable clinical hypotheses, but only when applied correctly. We have presented a discussion of best practice for the implementation of IBDM/VBA in radiation therapy dose distributions here, highlighting several methodological shortcomings in 3 analyses presented by Marcello et al as examples. Although the data used in these analyses are of the highest quality, the improper application of IBDM methodology may lead to conclusions that are not fully supported by the data. In future, to avoid detecting excited regions in the brain of a dead salmon, authors would be advised to follow established methodologies more closely and refer to the voxel-based analysis methodological cookbook[9] and other publications in which the necessary methodologies are discussed. We would also like to emphasize the need for the development of robust methodologies to account for intervoxel correlations in radiation therapy dose distributions to improve the application of IBDM/VBA in this field.

References

1. Marcello M, Denham JW, Kennedy A, et al. Reduced dose posterior to prostate correlates with increased PSA progression in voxel-based analysis of 3 randomised phase 3 trials. *Int J Radiat Oncol* 2020;108:1304-1318.
2. Marcello M, Denham JW, Kennedy A, et al. Relationships between rectal and perirectal doses and rectal bleeding or tenes-mus in pooled voxel-based analysis of 3 randomised phase III trials. *Radiother Oncol* 2020;150:281-292.
3. Marcello M, Denham JW, Kennedy A, et al. Increased dose to organs in urinary tract associates with measures of genitourinary toxicity in pooled voxel-based analysis of 3 randomized phase III trials. *Front Oncol* 2020;10:1174.
4. Bennett C, Miller M, Wolford G. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. *Neuroimage* 2009;47:S125.
5. Chen C, Witte M, Heemsbergen W, Herk MV. Multiple comparisons permutation test for image based data mining in radiotherapy. *Radiat Oncol* 2013;8:293.
6. Green A, Vasquez Osorio E, Aznar MC, McWilliam A, van Herk M. Image based data mining using per-voxel cox regression. *Front Oncol* 2020;10:1178.
7. McWilliam A, Kennedy K, Hodgson C, Osorio EV, Faivre-Finn C, Van Herk M. Radiation dose to heart base linked with poorer survival in lung cancer patients. *Eur J Cancer* 2017;85:106-113.
8. Groppe DM, Urbach TP, Kutas M. Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. 1993.
9. Palma G, Monti S, Cella L. Voxel-based analysis in radiation oncology: A methodological cookbook. *Phys Med* 2020;69:192-204.
10. Cohen A, Kolassa J, Sackrowitz HB. Penalized likelihood and multiple testing. *Biometrical J* 2019;61:62-72.
11. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. *Ann Stat* 2014;42:413-468.
12. Efron B. Estimation and accuracy after model selection. *J Am Stat Assoc* 2014;109:991-1007.
13. Taylor J, Tibshirani R. Post-selection inference for ℓ1-penalized likelihood models. *Can J Stat* 2018;46:41-61.
14. Taylor J, Tibshirani RJ, Brant R, Storey RD. Statistical learning and selective inference. *Proc Natl Acad Sci U S A* 2015;112:7629-7634.
15. Nunez-Elizalde AO, Huth AG, Gallant JL. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage* 2019;197:482-492.
16. Vakorin VA, Borowsky R, Sarty GE. Characterizing the functional MRI response using Tikhonov regularization. *Stat Med* 2007;26:3830-3844.
17. van den Heuvel MP, Hulshoff Pol HE. Exploring the brain network: A review on resting-state fMRI functional connectivity. *Eur Neuropsychopharmacol* 2010;20:519-534.
18. Altman DG, Royston P. The cost of dichotomising continuous variables. *Br Med J* 2006;332:1080.