Letter to the Editor: "A radiomics-based decision support tool improves lung cancer diagnosis in combination with the Herder score in large lung nodules"

Antonio Eleuteri[a,b,c,*]

[a]NHS Digital, Liverpool University Hospitals, NHS Foundation Trust, United Kingdom
[b]Department of Physics, School of Physical Sciences, University of Liverpool, United Kingdom
[c]School of Medical Sciences, Faculty of Biology, Medicine, and Health, University of Manchester, United Kingdom

eBioMedicine
2023;94: 104688
Published Online 28 June 2023
https://doi.org/10.1016/j.ebiom.2023.104688

Dear Editor,

I read with great interest the publication "A radiomics-based decision support tool improves lung cancer diagnosis in combination with the Herder score in large lung nodules".[1] In this multi-centre study, the authors describe the implementation of a decision support tool based on classification models of malignant large lung nodules, using clinical and radiomics predictors.

While I appreciate that this decision support tool may represent an improvement over existing clinical models, I would like to outline some issues in the modelling and validation of the tool reported in the above publication. The goal of this letter is to promote theoretically sound approaches to data analysis based on rigorous statistical principles, and as such, it is in no way intended to single-out the authors of the manuscript or to undermine their intellectual integrity.

The concerns I will raise are unfortunately all too common, and I have seen similar issues in many papers published in top rank journals.

Statistical issues in radiomics model development

Data splitting is a widely used technique,[2,3] since it guarantees nearly unbiased results. However, it greatly reduces the sample size for both model development and model testing (reducing the efficiency of all statistics). The split may be fortuitous; if the process were repeated with a different split, different assessments of predictive accuracy may be obtained.[2,3] In the present case, considering the limited sample size, the clustered structure, its high dimensionality, and the presence of an external data set (which allows one of the most stringent forms of validation), it's difficult to justify the 70%/30% split in favour of an efficient resampling approach.[2,3] Furthermore, a procedure like nested cross-validation may be necessary if feature selection is deemed important (see further below).

The authors split the data according to "Study ID", but it is not clear from the text whether this ID also includes information about the institution to which the patients belong. It is important to recognize that there are at least two levels of clustering in the data: Institution, Patient (with multiple observations per patient, with potential serial correlations due to the longitudinal acquisition of CT scans), and possibly scan vendor and reconstruction kernel. Some of these clusters are nested (patient within institution, assuming patients haven't moved institutions between a scan and another), and others may be crossed (e.g. if an institution has more than one scanner). Clustering structures can detrimentally affect the calculation of the test statistics (and resulting p-values) of the fitted models unless robust covariance clustering techniques (e.g. sandwich estimators), or more complex models (e.g. linear mixed effects models) are adopted[2,4]; in the most extreme cases, statistics which looked significant may not prove to be as such after clustering, thus increasing the probability of type-I errors.[2,4] There is no suggestion in the paper that clustering structures have been handled in the fitting of the radiomics and clinical models. In general, all calculations that involve estimates of variability (confidence intervals, standard errors etc.) must account for clustering, when the estimators assume independent observations.[2,4]

The authors used univariate logistic regression to screen for useful predictors. To quote Prof. Harrell[2]

"[Univariable screening] is just a form of forward stepwise variable selection in which insignificant variables from the first step are not reanalysed in later steps. It is thus even worse than stepwise modelling as it can miss important variables that are only important after adjusting for other variables. Overall, neither univariable screening nor stepwise variable selection in any way solves the problem of "too many variables, too few subjects," and they cause severe biases in the resulting

DOIs of original articles: https://doi.org/10.1016/j.ebiom.2023.104687, https://doi.org/10.1016/j.ebiom.2022.104344
*NHS Digital, Liverpool University Hospitals, NHS Foundation Trust, United Kingdom.
E-mail address: Antonio.Eleuteri@rlbuht.nhs.uk.

multivariable model fits while losing valuable predictive information from deleting marginally significant variables."

Although the authors applied the B–H procedure to mitigate the issue of multiple testing, it does nothing to address the other problems of univariable screening, as outlined by Prof. Harrell[2]

- All R^2 (and similar statistics) are biased high.
- The F and χ^2 statistics don't have the claimed distribution.
- The standard errors of regression coefficient estimates are biased low, and confidence intervals are falsely narrow.
- The procedure is made arbitrary by collinearity.
- The regression coefficients are biased high and need shrinkage. For a positive association, it can be shown for example that $E[\widehat{\beta} \mid p<0.05, \widehat{\beta}>0]>\beta$.[5]

The authors' concerns about "potential issue of collinearity between the Brock, Herder and PET status," and results "suggesting a high level of collinearity between Herder and PET status" reinforce the arbitrariness of the univariable screening procedure, and the potential unreliability of the findings.

If one really insists on using p-values to screen predictors, a reasonable threshold that does allow for deletion of *some* variables is $\alpha = 0.5$ (quite far from the authors' use of $\alpha = 0.05$).[6]

Incidentally, it's not clear why the B–H procedure (which controls the false discovery rate) was used, and not the Holm procedure, which controls the family-wise error rate, and is thus more stringent in controlling type-I errors. The consequence of this choice is that there is a decrease in type-II errors, in favour of an increase in type-I errors; but since no justification is provided, it's hard to tell whether this was intentional.

The authors correctly applied cross-validation (CV) after univariable screening for features on the training set (this step is usually a source of errors when applied to a full data set), however it should be noted that such a two-step procedure is statistically inefficient, and a nested CV procedure (where feature selection happens inside a CV loop) is generally to be preferred.[3,7] As previously mentioned, the existence of an external validation test set (which provides the most stringent form of validation) makes the use of data splitting redundant and wasteful of data.[2,3]

It should be noted that the radiomic features extracted by the TexLab 2.0 software based on segmented images (manually or automatically, e.g. by nnUNet), are intrinsically affected by uncertainty. Said uncertainty can negatively affect a model when the features are used as predictors[8,9] (see below for further comments.) Inference on model parameters and estimation of confidence limits on model predictions require quantification of input uncertainty to be reliable.[2,8]

Assessment of the model's performance is incomplete, as the authors only report the *discrimination*[2] (i.e. the model's ability to separate subjects' outcomes) in the form of AUC (and sensitivity, specificity etc.), but no mention at all is made of *calibration/reliability*[2] (i.e. the model's ability to make *unbiased* estimates of the outcomes.) It is quite simple to see that a naïve model that outputted 51% for all positive outcomes, and 49% for all negative outcomes, would have a perfect discrimination ability (AUC = 1, sensitivity 100%, specificity 100% etc.), but it would be completely unreliable and pretty much useless as a decision support tool. Calibration graphs (i.e. predicted vs. smoothed observed probabilities), Brier score (in its many shapes) etc. are all useful and they must always be reported.[2] For example, in the above case the Brier score would be approximately 0.25 (assuming for the sake of example, equal classes' prevalence), which practically corresponds to a "coin toss" prediction (despite the model showing a perfect AUC). Assessment of calibration may reveal that, despite a reasonable AUC, *recalibration*[2] may be necessary to effectively use the model as a decision support tool.

Statistical issues in clinical model development

The same considerations as above apply *mutatis mutandis*, with the following additional comments.

It's not clear why patients with no recorded PET data were considered PET negative; is it safe to assume that this data is MAR[2] (Missing at Random) or MCAR[2] (Missing Completely at Random)? In the absence of such guarantees, a safer approach is to use imputation procedures.[2,10]

As previously hinted, a subtle issue in the fitting and validation of the combined clinical-radiomics model, is that using the output of the LN-RPV model as a predictor for the clinical-radiomics model is equivalent to observing LN-RPV with measurement errors represented by the uncertainty associated with the coefficients of the LN-RPV model.[8] The magnitude of this measurement error is directly related to how reliable is the estimate of the covariance matrix of the model's parameters (which would at least in principle allow to estimate the variance of the output of the LN-RPV model). Ignoring clustering and using univariable screening casts doubts on the reliability of the estimates, thus indirectly affecting the quality of the clinical model.

The measurement error should not be ignored as it may result in biased estimates of the parameters of the clinical-radiomics model, and unreliable estimates of all standard errors, with attendant negative impact on

inference. Special regression analysis techniques must be employed to account for measurement error.[8,9]

The same considerations as described for the radiomics and clinical model development apply to the fusion models.

Auto-segmentation

The estimates of the performance of the Dice score should account for the clustered nature of data in the calculation of the standard errors. Neglecting serial correlations (which seem likely in this case, considering that nodules may grow and/or shrink in time) may produce falsely narrow standard errors. Similar arguments can be applied to the calculation of the ICC and attendant p-values.

Furthermore, due to nnUNet being an inherently non-identifiable black-box model, it may be practically impossible to estimate the uncertainty of the model's parameters, which as previously noted may indirectly affect the performance of the radiomics model if used as a pre-processing step to the extraction of the radiomics features (which are themselves affected by uncertainty). Reproducibility of the radiomics model in presence of auto-segmentation as reported by the authors is reassuring, but care is necessary when applying automated black-box models.

Conclusions

I hope this letter will help the research community in improving the quality of their data analysis and encourage them in seeking the help of expert statisticians if there are any doubts about the formal validity of the statistical procedures. In this context, it is always useful to remember the words of Prof. G. P. Box:

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

Declaration of interests

The author discloses no financial or personal relationships with other people or organizations that could inappropriately influence his work.

References

1 Hunter B, Chen M, Ratnakumar P, et al. A radiomics-based decision support tool improves lung cancer diagnosis in combination with the Herder score in large lung nodules. *eBioMedicine*. 2022;86:104344. https://doi.org/10.1016/j.ebiom.2022.104344.

2 Harrell FE Jr. *Regression modeling strategies*. 2nd ed. New York: Springer; 2015.

3 Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*. 2nd ed. New York: Springer; 2009.

4 Zeileis A, Köll S, Graham N. Various versatile variances: an object-oriented implementation of clustered covariances in R. *J Stat Software*. 2000;95(1):1–36.

5 Chatfield C. Model uncertainty, data mining and statistical inference (with discussion). *J Roy Stat Soc A*. 1995;158:419–466.

6 Steyerberg EW, Eijkemans MJC, Harrell FE Jr, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med*. 2000;19:1059–1079.

7 Cawley GC, Talbot NLC. on over-fitting in model selection and subsequent selection bias in evaluation, 2010 performance evaluation. *J Mach Learn Res*. 2010;11:2079–2107.

8 Ogburn EL, Rudolph KE, Morello-Frosch R, Khan A, Casey JA. A warning about using predicted values from regression models for epidemiologic inquiry. *Am J Epidemiol*. 2021;190(6):1142–1147.

9 Brakenhoff TB, Mitroiu M, Keogh RH, Moons K, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. *J Clin Epidemiol*. 2018;98:89–97.

10 Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models. *J Am Stat Assoc*. 2005;100(469):332–346.