Journal for ImmunoTherapy of Cancer

Letter to the editor: radiomics analysis for predicting pembrolizumab response in patients with advanced rare cancers

Mateus Trinconi Cunha ,[1] Vinicius Jardim Carvalho,[2] Rafael Maffei Loureiro ,[3] Carlos Eduardo Brantis-de-Carvalho,[4] Murilo Bicudo Cintra,[5,6] Gilberto de Castro Junior[1]

To cite: Cunha MT, Carvalho VJ, Loureiro RM, *et al*. Letter to the editor: radiomics analysis for predicting pembrolizumab response in patients with advanced rare cancers. *Journal for ImmunoTherapy of Cancer* 2021;**9**:e003044. doi:10.1136/jitc-2021-003044

Accepted 14 June 2021

Linked

► http://dx.doi.org/10.1136/jitc-2021-003299

[1]Medical Oncology, Instituto do Cancer do Estado de Sao Paulo, Faculdade de Medicina da Universidade de Sao Paulo, Sao Paulo, Brazil
[2]Computer Science, Instituto de Matematica e Estatistica, Universidade de Sao Paulo, Sao Paulo, Brazil
[3]Department of Radiology, Hospital Israelita Albert Einstein, Sao Paulo, Brazil
[4]Faculdade de Medicina, Centro Universitario Saude ABC, Santo Andre, SP, Brazil
[5]Radiology, Diagnosticos da America, Barueri, Brazil
[6]Radiology, Instituto do Cancer do Estado de Sao Paulo, Faculdade de Medicina da Universidade de Sao Paulo, Sao Paulo, Brazil

Correspondence to
Dr Mateus Trinconi Cunha; mateustcunha@gmail.com

ABSTRACT

A commentary on the original research article: 'Radiomics analysis for predicting pembrolizumab response in patients with advanced rare cancers'. Of note, the predictor selection process, the cross-validation method, along with the lack of final testing of the developed model with a separated data set may mask overfitting, overestimating performance metrics.

In the original research article, Colen *et al*[1] use classic statistics and machine learning methods in order to identify significant radiomics features and predict pembrolizumab response in advanced rare cancers. This novel approach raises relevant hypotheses and may eventually prove useful in the expansion of the therapeutic arsenal for some patients.

However, the encouraging results obtained are, until further clarification, to be interpreted with caution. Machine learning is the software-mediated attempt to produce accurate output from previously unseen data through mostly automatic adjustment of parameters based on previous experience.[2] Effectively, the 'learning' step in this study occurs in a supervised fashion, that is, feeding the algorithm examples of labeled data (ie, the characteristics of each patient along with the label of 'responder' or 'non-responder'). The learning algorithm then builds models to predict each patient's label as accurately as possible.[3]

After initial training, model validation is carried out. This is usually done by splitting the data set into training and validation sets: two groups with no overlapping patients, each used exclusively in their respective phase. To increase the model's generalization capability and decrease any sample selection bias, resampling methods are used. Bootstrapping is the process of resampling data with replacement, usually producing several new groups of different training and test data sets, sometimes containing multiple instances of the same original cases, while omitting others. Cross-validation comprises resampling without replacement, systematically producing k surrogate data sets, with n original cases being part of the validation data set exactly once. This is called k-fold cross-validation. A special case is leave-one-out cross-validation (LOOCV), in which the training set consists in all cases but one, and the remaining case is used as a one-case validation set. The process repeats until all cases are separately used as validation. LOOCV is usually reserved for small data sets, in which the omission of a significant part of the training data (ie, 10%–20%) might hinder algorithm learning and thus performance.[4]

Following the validation phase, the investigators may adjust the algorithm's hyperparameters and try again until satisfactory performance is achieved. Since many changes are made to make the model more accurate for the validation data, overfitting may occur. This usually causes high performance metrics in the validation set, with poor prediction capability in a distinct dataset. To detect such phenomena, testing on sequestered, previously unseen data is performed, differences in model metrics are analyzed, methodology problems are addressed, and the process is repeated.[4]

In the study, Colen *et al*[1] address the objective with an admittedly small, but multidimensional patient data set, using LOOCV to assess model accuracy and C-statistic. However, caveats to their study design should be noted. Regarding feature selection, both in tables 3 and 4, multiple instances of the same feature in different levels of grayscale can be seen. While their relevance was reportedly identified by a sound method (L1 penalty), one cannot but wonder their collinearity (assessed by variance inflation factor[5]), and

whether data preprocessing or the usage of other selection methods (wrapper or embedded methods) would change the outcomes. This must be carefully considered when small-n-large-p-problems, known to lead to feature selection instability, are involved.[6 7] In relation to cross-validation, while LOOCV maximizes training data, testing a single point at a time implies a large variance in error and a similarly high variance of CIs. The method underestimates error rates, especially in small samples with high dimensionality (ie, few patients with several features), which can explain the reported results.[8 9] The lack of cross-validation on blocks of correlated data may introduce another bias in the study: the algorithm might have been able to distinguish between different primary sites, and correlating tumor origin to outcome, always guessing the correct label, leading to accuracy, and C-statistic inflation. For example, penile carcinomas, small cell malignancies of non-pulmonary origin, and retroperitoneal spindle cell sarcoma had no responders in the sample, yielding always perfect predictions of no response in the test, while this might not hold true in external validation.[9]

Additionally, the lack of testing in a separated set after cross-validation hinders the credibility of the outstanding metrics achieved—at least until independent verification.[8]

In order to address the outlined issues, the following procedures might be applied: assessment of feature collinearity and usage of different methods of feature selection might help with the small-n-large-p-problem; and the separation of lesions (in case of metastatic sites) into distinct data points, as well as data amplification methods (such as synthetic minority over-sampling technique[10]) may help increase the data set. After a larger amount of data is achieved, other resampling strategies (k-fold cross-validation or bootstrapping) may be employed, and more data (with no synthetic points) can be spared for final testing and overfitting assessment. The analysis of one lesion may not be a surrogate marker for cancer response to immunotherapy, but it may be an interesting hypothesis generator. Also, other predictive models for treatment response based on voting on the probability of response for each tumor in a patient may be developed from the original algorithms.

Twitter Mateus Trinconi Cunha @MateusTrinconi, Rafael Maffei Loureiro @RMaffeiLoureiro and Gilberto de Castro Junior @GilbertodeCas13

Contributors All author contributed equally to letter conception, design, data analysis and interpretation, manuscript writing, and final approval.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests GdC has received personal fees from AstraZeneca, Bayer, Bristol-Myers Squibb, Boehringer Ingelheim, Janssen, Lilly, Merck Serono, Merck Sharp and Dohme, Novartis, Pfizer, Roche, Teva, and Yuhan, none related to this publication.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

ORCID iDs
Mateus Trinconi Cunha http://orcid.org/0000-0001-9101-8553
Rafael Maffei Loureiro http://orcid.org/0000-0002-1635-2225

REFERENCES

1 Colen RR, Rolfo C, Ak M, *et al*. Radiomics analysis for predicting pembrolizumab response in patients with advanced rare cancers. *J Immunother Cancer* 2021;9:e001752.
2 Koza JR, Bennett FH, Andre D, *et al*. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: Gero JS, Sudweeks F, eds. *Artificial Intelligence in Design '96*. Dordrecht: Springer Netherlands, 1996: 151–70.
3 Chen T, Guestrin C, Association for Computing Machinery. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016:785–94.
4 Burkov A. Basic Practice. In: *The Hundred-Page machine learning book*, 2019. http://themlbook.com/wiki/doku.php
5 Grønning B, Nilsson JC. Multiple regression: a primer. *Stat Med* 2001;20:1888–9.
6 Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. *IEEE Trans Pattern Anal Mach Intell* 1997;19:153–8.
7 He Z, Yu W. Stable feature selection for biomarker discovery. *Comput Biol Chem* 2010;34:215–25.
8 Rao RB, Fung G, Rosales R, Society for Industrial and Applied Mathematics. On the dangers of cross-validation. An experimental evaluation. *Proceedings of the 2008 SIAM International Conference on Data Mining*, Philadelphia, PA, 2008:588–96.
9 Varoquaux G, Raamana PR, Engemann DA, *et al*. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *Neuroimage* 2017;145:166–79.
10 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: synthetic minority Over-sampling technique. *Jair* 2002;16:321–57.