Editorial

Explainable *vs.* interpretable artificial intelligence frameworks in oncology

Dimitris Bertsimas[1], **Georgios Antonios Margonis**[2,3]

[1]Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA; [2]Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA; [3]Department of General and Visceral Surgery, Charité Campus Benjamin Franklin, Berlin, Germany

Correspondence to: Georgios Antonios Margonis, MD, PhD. Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA. Email: margonig@mskcc.org.

Comment on: Ladbury C, Zarinshenas R, Semwal H, *et al.* Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. Transl Cancer Res 2022;11:3853-68.

Keywords: Explainable artificial intelligence (XAI); interpretable artificial intelligence; oncology

Submitted Oct 19, 2022. Accepted for publication Dec 31, 2022. Published online Jan 19, 2023.
doi: 10.21037/tcr-22-2427
View this article at: https://dx.doi.org/10.21037/tcr-22-2427

The review by Ladbury and colleagues is non-systematic in nature and serves to provide examples of how explainable artificial intelligence (XAI), specifically SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), can be applied to almost all stages of oncologic patient care including *cancer diagnosis*, *patient prognostication*, and *optimal selection* of treatment (1-3). However, it is also important to discuss its limitations and caveats, especially given its increasing use in oncology.

To demonstrate the applicability of XAI in *cancer diagnosis*, the authors highlighted a study by a group from Seoul National University Hospital that developed a model to guide patient selection for prostate biopsy based on the model-predicted likelihood of a patient having prostate cancer (4). This would help minimize unnecessary invasive procedures. Although the authors were able to train an extreme gradient-boosting algorithm with excellent discriminatory ability [area under the curve (AUC): 0.945], the algorithm was essentially a black box method. This is a barrier to clinical implementation since clinicians are unlikely to base decisions on a model's output without knowing how the predictions were made. To mitigate this issue, the researchers used SHAP to provide a mathematically based reasoning of how the model yielded its output. Specifically, they calculated the SHAP value of each predictor, which reflects the association of the predictor with the probability of having prostate cancer and therefore the need for biopsy. As noted above, the discriminatory ability of the algorithm was excellent. However, the calibration of the model was problematic, with the model underestimating the risk of prostate cancer by as much as 25% on visual estimation. This was not discussed or acknowledged by the authors in the manuscript, but is of high clinical importance as several patients with high risk of prostate cancer may not have undergone biopsy if the models' recommendations were followed. Thus, we call for increased attention to calibration, as this study shows that excellent discriminatory ability does not necessarily equate to good calibration. Unfortunately, poor calibration is a known weakness of several machine learning (ML) based models, and most studies fail to report this statistical property (5).

Aside from diagnosis, *patient prognostication* is another important aspect of oncologic care for both the physician and the patient. XAI certainly has a place in prognostication, given the well-known accuracy-interpretability trade-off; accurate models suffer from limited interpretability ("black box" models) while interpretable models are limited by lack of accuracy. Although XAI has been utilized to "explain" these highly accurate black box models, it is only as good as the model that it attempts to explain, and as demonstrated by the example above, the model with the highest AUC is not always the best model. "Black box" models tend to be more flexible and thus prone to overfitting, which can hinder generalizability.

This brings us to the use of inherently interpretable *vs.* explainable (e.g., SHAP or LIME) methods. Some inherently interpretable models such as decision trees may have comparable discriminatory performance to complex (and opaque) models. The authors of the review referenced a study by our group that used a novel type of decision tree, or the optimal classification trees (OCTs), to predict prognosis in patients with resected pancreatic ductal adenocarcinoma (PDAC) (6,7). In that study, the AUCs of the 1-year OCT *vs.* 1-year XGBoost were comparable at 0.638 *vs.* 0.654, respectively. Similarly, the AUCs of the 3-year OCT *vs.* 3-year XGBoost were remarkably close at 0.675 *vs.* 0.690, respectively. To assess whether the OCT-determined cut-offs for the prognostic factors were internally valid, we performed an analysis of interactions with the SHAP method in independently devised XGBoost prognostic models. The same cut-offs were identified, which attests to the usefulness of both OCT and SHAP in capturing interactions between prognostic factors. It also suggests that SHAP can be used as a complementary internal validation method. Another study that exemplifies the complementary use of XAI and other methodologies comes from a group at the City of Hope Medical Center. They studied a cohort of patients with prostate cancer and used SHAP to model and visualize previously unknown but clinically important nonlinear relationships among predictors of mortality (8). Importantly, they replicated their results using conventional regression methods. This is significant because regression methods can only capture interactions if previously specified, which is problematic if we want to derive new knowledge. Furthermore, testing widely for interactions in linear models can lead to erroneous results. Thus, XAI can screen for interactions which can then be verified by linear regressions models.

Aside from informing prognosis, predictive models have also been used in oncology to establish risk thresholds above which treatment is indicated. However, risk of an outcome (e.g., recurrence) cannot be the sole determinant of allocating a treatment, as high baseline risk does not necessarily equate to good treatment response. Conversely, patients with lower risk who respond well to a treatment may benefit from intervention. Thus, we need a composite measure of baseline risk and response to treatment to predict who will benefit most from a treatment and guide our treatment algorithm. Predicting who will benefit from a particular treatment is at the core of *precision medicine*, and we believe it is the most exciting and clinically important aspect of XAI use. The City of Hope group conducted a study that used XAI to identify subgroups of patients with completely resected stage III-N2 non-small cell lung cancer who may benefit from adjuvant radiotherapy (9). Specifically, they used SHAP to screen for non-linear interactions between number of positive lymph nodes and adjuvant radiotherapy, and assessed their impact on overall survival. They found that radiotherapy was associated with improved OS only in those with more than three positive lymph nodes. Importantly, they internally validated this finding by performing multivariable Cox regression analyses both in the entire cohort as well as in those with three or more positive lymph nodes. Of note, radiation was not independently prognostic in the overall analysis but was prognostic in the sub analysis of patients with a high number of positive lymph nodes. This highlights a fundamental limitation of Cox regression analysis, which is *largely ignored* in almost all clinical studies. Unfortunately, it appears that this limitation is shared by the SHAP feature importance plots, which also underestimated the importance of adjuvant radiotherapy. In fact, the mean SHAP value for radiotherapy was the lowest among all examined prognostic factors, which means that radiotherapy had on average the smallest contribution to the model's prediction. This limitation may not be unique to this study, as a recent study by Dunn and colleagues showed that SHAP and XGBoost consistently underestimated the importance of key features while they assigned significant importance to irrelevant features (10). Importantly, the City of Hope group avoided this limitation by ignoring the results of the SHAP feature importance plots; instead, they employed SHAP dependence and interaction plots to screen for non-linear interactions between number of positive lymph nodes and adjuvant radiotherapy. This idea was based on existing literature suggesting that the extent of nodal involvement is both prognostic of overall survival and predictive of response to adjuvant radiotherapy.

The findings by the City of Hope group are impressive but still have two main weaknesses. The first is the lack of external validation to demonstrate generalizability, which precludes clinical use. The second is that SHAP visualization can only test two factors at a time for an interaction (in this case, number of lymph nodes and receipt of radiation therapy). Thus, it is possible that there exist higher order interactions that cannot be visualized by SHAP, which means that this analysis may have the same limitations as the one-variable-at-a-time sub-group analyses of randomized controlled trials (RCTs) that compare groups of patients who differ systematically on a single variable; in reality, individual patients may differ from one another across many variables simultaneously.

Decision trees use numerous factors to split a cohort into several groups and thus could remedy the aforementioned deficiency of XAI. Our group has recently introduced a novel methodology called optimal policy trees (OPTs) (11). OPTs are globally optimal decision trees that are trained using a matrix of probabilities of what the outcome of interest (e.g., recurrence) *would have been* if a given patient was treated with all possible treatment options; one is the option that was used in reality, and all others are the counterfactuals. We recently tested this methodology by training OPTs to determine the optimal margin width in colorectal liver metastases and to assess whether optimal margin width should be individualized based on patient characteristics (12). Of note, we used SHAP analysis in that study to internally validate the OPT recommendations, and also externally validated our findings in an independent cohort.

Collectively, XAI and interpretable AI (IAI) should be viewed as complementary and not competing analytical frameworks. In our opinion, IAI is more appropriate to estimate heterogeneity treatment effects (HTE) due to its granularity, while XAI has proved its value in internally validating IAI. In any case, it is important that external validations are also performed given the high stakes of making treatment recommendations in oncology.

Acknowledgments

Funding: This work was funded by National Cancer Institute award (No. P30-CA008748). The funding organizations had no role in the design and conduct of the study; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Translational Cancer Research*. The article did not undergo external peer review.

Conflicts of Interest: Both authors have completed the ICMJE uniform disclosure form (available at https://tcr.amegroups.com/article/view/10.21037/tcr-22-2427/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Ribeiro MT, Singh S, Guestrin C, editors. "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016.
2. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems 2017;30:4768-77.
3. Ladbury C, Zarinshenas R, Semwal H, et al. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. Transl Cancer Res 2022;11:3853-68.
4. Suh J, Yoo S, Park J, et al. Development and validation of an explainable artificial intelligence-based decision-supporting tool for prostate biopsy. BJU Int 2020;126:694-703.
5. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17:230.
6. Bertsimas D, Margonis GA, Huang Y, et al. Toward an Optimized Staging System for Pancreatic Ductal Adenocarcinoma: A Clinically Interpretable, Artificial Intelligence-Based Model. JCO Clin Cancer Inform 2021;5:1220-31.
7. Bertsimas D, Dunn JJML. Optimal classification trees. 2017;106:1039-82.
8. Li R, Shinde A, Liu A, et al. Machine Learning-Based Interpretation and Visualization of Nonlinear Interactions in Prostate Cancer Survival. JCO Clin Cancer Inform 2020;4:637-46.
9. Zarinshenas R, Ladbury C, McGee H, et al. Machine learning to refine prognostic and predictive nodal burden thresholds for post-operative radiotherapy in completely

resected stage III-N2 non-small cell lung cancer. Radiother Oncol 2022;173:10-8.

10. Dunn J, Mingardi L, Zhuo YD. Comparing interpretability and explainability for feature selection. 2021. doi: 10.48550/arXiv.2105.05328.
11. Amram M, Dunn J, Zhuo YD. Optimal Policy Trees. Machine Learning 2022;111:2741-68.
12. Bertsimas D, Margonis GA, Sujichantararat S, et al. Using Artificial Intelligence to Find the Optimal Margin Width in Hepatectomy for Colorectal Cancer Liver Metastases. JAMA Surg 2022;157:e221819.

Cite this article as: Bertsimas D, Margonis GA. Explainable *vs.* interpretable artificial intelligence frameworks in oncology. Transl Cancer Res 2023;12(2):217-220. doi: 10.21037/tcr-22-2427