Unboxing Artificial Intelligence "Black-Box" Models – A Novel Heuristic

Sarah Weppler[1,2], Harvey Quon[3,4], Nabhya Harjai[5], Craig A. Beers[5], Lukas Van Dyke[2], Charles Kirkby[1,4,6], Colleen Schinkel[2,4], Wendy Smith[1,2,4]

[1]Department of Physics and Astronomy, University of Calgary, [2]Department of Medical Physics, Tom Baker Cancer Centre,
[3]Department of Radiation Oncology, Tom Baker Cancer Centre, [4]Department of Oncology, University of Calgary,
[5]Cumming School of Medicine, University of Calgary, [6]Department of Medical Physics, Jack Ady Cancer Centre

Purpose

Artificial intelligence (AI) models capture complex associations in a dataset; however, it can be difficult to translate these associations into useful clinical practice guidelines. We introduce a novel heuristic to simplify AI models into intuitively understandable sets of inequalities and apply the approach to: (1) deformable image registration (DIR) quality assurance (QA), and (2) patient selection criteria for adaptive radiation therapy (ART).

Heuristic Overview

- Conceptually, the heuristic increments the values of variables to identify when AI model predictions "transition" from one classification to another.
- Depending on the nature of the data, clinical application, and underlying AI model, *OR* or *AND* inequality simplifications may be more appropriate.

Definitions

i: index of the input datapoint, e.g., patient (n total)
j: index of the input variable, e.g., parotid gland volume (m total)
$\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m})$: input datapoints
δ_i: model output, for example, providing a "normal" or "violation" classification
f: "black-box" AI model, e.g., $f(\boldsymbol{x}_i) = \delta_i$

Summary of the Simplification Heuristic

1) For each j:
   i. For each i:
      Fix the value of all variables $\neq j$, and simulate 1000 new datapoints $\tilde{x}_{i,j}$ by incrementing from $\min_{i=1,\dots,n}(x_{i,j})$ to $\max_{i=1,\dots,n}(x_{i,j})$. Each produces $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \dots, \tilde{x}_{i,j}, \dots, x_{i,m-1}, x_{i,m})$.
   ii. Input each simulated datapoint into the AI model and record when model output transitioned from "normal" to "violation".
   iii. Assess the frequency of transition values across all i. Extract local maxima, c_j, as candidate cutoff criteria for inequalities of the form:
      If $x_{i,1} \leq c_1$ *OR* $x_{i,2} \leq c_2$ *OR* ... then "violation."
2) Assess the sensitivity and specificity of combinations of candidate cutoff criteria. Select cutoffs to maximize the performance of the simple criteria.
3) Verify simple criteria performance on an external validation dataset.

Modification: To derive AND criteria, modify 1)i. and 2) to the following:

1)i. For each i:
   Fix the value of all low-importance predictors. Substitute in sets of high-importance predictor values from other datapoints $i = 1, \dots, n$. For example, if the values of $x_{i,1}$ and $x_{i,2}$ are of high importance, simulate: $\boldsymbol{x}_i = (x_{\tilde{\imath},1}, x_{\tilde{\imath},2}, x_{i,3}, \dots, x_{i,m})$ for $\tilde{\imath} = 1, \dots, n$.
2) Starting with the highest-importance variable, determine the most-effective cutoff value. Fix this value and determine the cutoff value for the next most important variable. Continue this process until all cutoff values are selected.

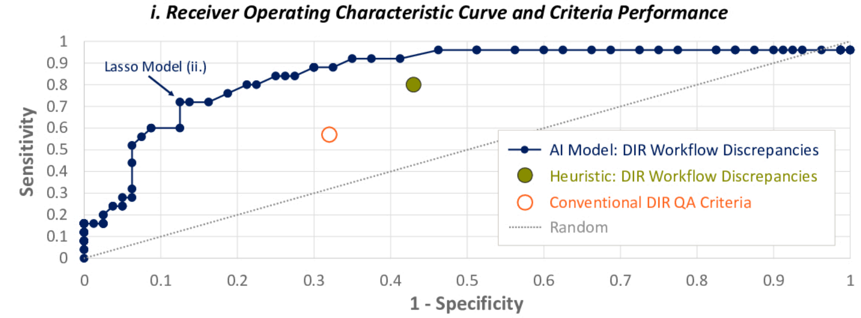
Application #1 – Workflow-Specific DIR QA Criteria

Clinical Problem: DIR algorithms may not be interchangeable in a given workflow. AI models can predict which algorithm differences will lead to workflow discrepancies, (e.g., workflows selecting head and neck patients for adaptive replanning).

AI Modelled Solution: Lasso logistic regression modelled differences in contours produced by two DIR algorithms that would give discrepant replan indications.

Implications of the Heuristic: The heuristic converted the lasso model into simple OR-type clinical practice guidelines (Figure 1). Performance was poorer than the AI model, but achieved greater sensitivity (prioritized in this setting) than conventional DIR QA criteria (i.e., naïve use of TG-132 criteria for algorithm comparisons).

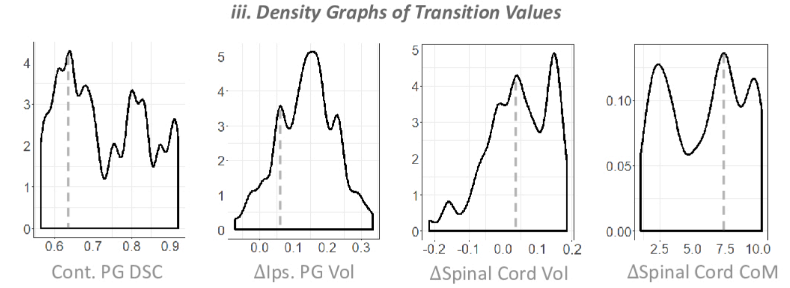
i. Receiver Operating Characteristic Curve and Criteria Performance



ii. Reference Lasso Logistic Regression Model

If $-0.35 + (0.87 \times \text{Brainstem DSC}) + (-0.54 \times \text{Cont. PG DSC}) + (0.23 \times \text{Spinal Cord DSC})$
$+ (-0.06 \times \text{Spinal Cord MDA}) + (1.05 \times \Delta\text{Ips. PG Vol}) + (0.73 \times \Delta\text{Spinal Cord Vol})$
$+ (0.01 \times \Delta\text{Spinal Cord CoM}) \geq \text{logit}(0.58)$
then DIR algorithms are likely to lead to discrepant workflow output.

iii. Density Graphs of Transition Values



iv. Heuristically Simplified DIR QA Criteria

If Cont. PG DSC $\leq$ 0.65
OR ΔIps. PG Vol $\geq$ 10%
OR ΔSpinal Cord Vol $\geq$ 5%
OR ΔSpinal Cord CoM $\geq$ 8.0mm
then DIR algorithms are likely to lead to discrepant workflow output.

Figure 1: Reference AI model and heuristically simplified DIR QA criteria. i. Receiver operating characteristic curve for the lasso model obtained by varying the minimum probability required to predict discrepant workflow output. ii. Reference lasso model maximizing the sum of sensitivity and specificity. iii. Density graphs of transition values for high importance variables (dashed lines indicate local maxima used in the final criteria). iv. Final format of the simplified QA criteria. Note: abbreviations are included below.

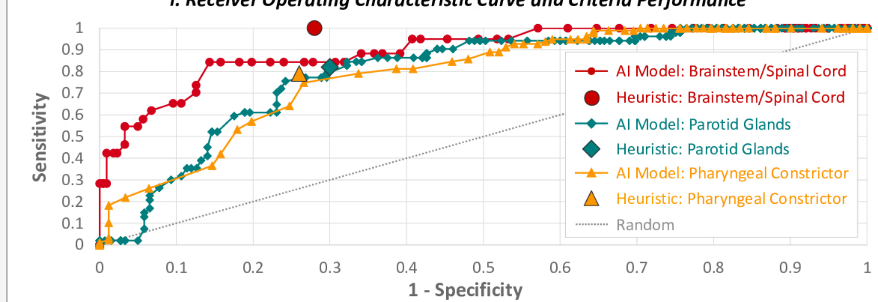
Application #2 – ART Patient Selection Criteria

Clinical Problem: Inter-fractional changes in head and neck patient and tumor anatomy can affect the accuracy of delivered dose. ART can correct dose deviations but is a resource intensive process. AI models may predict which patients are at risk of structure overdoses and improve resource allocation.

AI Modelled Solution: Random forests predicted overdoses of priority organs-at-risk

Implications of the Heuristic: The heuristic converted random forest models into simple AND criteria (Figure 2). Results appear promising as sensitivity is prioritized.

i. Receiver Operating Characteristic Curve and Criteria Performance



ii. ART Priority Organs-At-Risk and Heuristically Simplified Patient Selection Criteria

Brainstem/Spinal Cord	If Planned spinal cord D0.03cc $\geq$ 43 AND Planned cont. parotid gland Dmean $\geq$ 20 Gy AND Planned cont. submand. gland Dmean $\geq$ 34 Gy AND Planned pharyngeal constrictor Dmean $\geq$ 45 Gy AND Planned brainstem D0.03cc $\geq$ 16 Gy AND Planned ips. parotid gland Dmean $\geq$ 25 Gy AND ΔNeck diameter $\geq$ 5 mm then dose increases to brainstem/spinal cord are likely.
Parotid Glands	If Planned cont. parotid gland Dmean $\geq$ 23 Gy AND Planned cont. submand. gland Dmean $\geq$ 33 Gy AND Planned pharyngeal constrictor Dmean $\geq$ 45 Gy AND Planned ips. parotid gland Dmean $\geq$ 24 Gy AND Planned spinal cord D0.03cc $\geq$ 41 Gy AND Planned brainstem D0.03cc $\geq$ 16 Gy AND Planned low-dose CTV D20% $\geq$ 64 Gy AND Planned ips. submand. gland Dmean $\geq$ 61 Gy then dose increases to parotid glands are likely.
Pharyngeal Constrictor	If Planned pharyngeal constrictor Dmean $\geq$ 49 Gy AND Planned cont. parotid gland Dmean $\geq$ 19 Gy AND Planned cont. submand. gland Dmean $\geq$ 34 Gy AND Initial low-dose CTV volume $\geq$ 197cc AND Planned ips. parotid gland Dmean $\geq$ 21 Gy AND Planned spinal cord D0.03cc $\geq$ 40 Gy AND Planned brainstem D0.03cc $\geq$ 16 Gy AND ΔFace diameter $\geq$ 6 mm then dose increase to pharyngeal constrictor is likely.

Figure 2: Performance of AI models and simplified ART patient selection criteria. i. Receiver operating characteristic curves for AI models obtained by varying the minimum probability required to predict dose increases. Models predicted dose increases to specific priority organs at risk. ii. Simplified ART patient selection criteria. Note: abbreviations are included below.

Conclusion

This heuristic technique is capable of simplifying classification criteria for the applications investigated and may be valuable in scenarios where full AI models cannot be integrated with the clinic.

Abbreviations: DSC: Dice similarity coefficient; MDA: mean distance to agreement; ΔVol: change in volume; ΔCoM: change in centre of mass position; PG: parotid gland; Ips.: ipsilateral; Cont.: contralateral; submand.: submandibular

Email: Sarah.Weppler@ahs.ca