Automated Planning - Technical Considerations for Ensuing Safety and Quality

Automated Planning and Data-driven Plan Quality Control

Tom Purdie, PhD, MCCPM
Princess Margaret Cancer Centre
Disclosures

Method and System for Automated Planning of Radiation Therapy technology patented in PCT/CA2011/001130

Automated Quality Assurance (QA) and Planning technology patented in WO2014197994 A1

Receive royalties from RaySearch Laboratories for license of technology for Automated Breast Treatment Planning and Machine Learning-based Automated Treatment Planning

Have an equity interest in bridge7, licensee of technology for Machine Learning-based for Automated Quality Assurance in Radiation Oncology
Outline

Responsible Machine Learning (ML)
Regulatory Process with AI/ML
Risks of Automated Planning
Open Knowledge Based Planning (KBP) Challenge
Are We Ready to Use Deep Neural Networks Clinically?

Deep neural network tested for diagnosing skin lesions representing the most common and deadliest skin cancers with 21 board-certified dermatologists used as control

**Bottom line:**
- Algorithm matched the performance of dermatologists

---

**LETTER**

doi:10.1038/nature21056

**Dermatologist-level classification of skin cancer with deep neural networks**

Andre Esteva1*, Brett Kuprel1*, Roberto A. Novoa2,3, Justin Ko2, Susan M. Swetter2,4, Helen M. Blau5 & Sebastian Thrun6

1Department of Electrical Engineering, Stanford University, Stanford, California, USA. 2Department of Dermatology, Stanford University, Stanford, California, USA. 3Department of Pathology, Stanford University, Stanford, California, USA. 4Dermatology Service, Veterans Affairs Palo Alto Health Care System, Palo Alto, California, USA. 5Baxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, California, USA. 6Department of Computer Science, Stanford University, Stanford, California, USA.

*These authors contributed equally to this work.
Promise | Potential for AI to Infer New Knowledge

- AI recognizes complex data in images and can give quantitative metrics.

- Radiomics ➔ High throughput extraction of quantitative image features (non-human detectable) for biomarker discovery.

Aerts et al. Nat Commun 2014
Are we Inferring New Knowledge?

Radiotherapy and Oncology

Volume 130, January 2019, Pages 2-9

Original article

Vulnerabilities of radiomic signature development: The need for safeguards

Mattea L. Welch a, f, i, Chris McIntosh e, f, i, Benjamin Haibe-Kains a, c, i, j, Michael F. Milosevic b, e, i, Leonard Wee g, Andre Dekker g, Shao Hui Huang b, i, Thomas G. Purdie b, e, f, i, Brian O'Sullivan b, i, Hugo J.W.L. Aerts h, David A. Jaffray a, b, d, e, f, i, j

Purdie | Automated Planning - Technical Considerations for Ensuring Safety and Quality
AAPM | COMP Virtual Meeting 2020 | July 14, 2020
Responsible Use of Machine Learning

Do no harm: a roadmap for responsible machine learning for health care

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni and Anna Goldenberg

Interest in machine-learning applications within medicine has been growing, but few studies have progressed to deployment in patient care. We present a framework, context and ultimately guidelines for accelerating the translation of machine-learning-based interventions in health care. To be successful, translation will require a team of engaged stakeholders and a systematic process from beginning (problem formulation) to end (widespread deployment).
New Regulatory Approach for AI/ML Needed

AI/ML-Based SaMD are highly iterative, autonomous, and adaptive

Need to facilitate rapid cycle of constant improvement while providing effective safeguards
New Regulatory Approach for AI/ML Needed

![Diagram showing Good Machine Learning Practices and processes involving data selection, model training and tuning, model validation, data for re-training, culture of quality and organizational excellence, premarket assurance of safety and effectiveness, review of SaMD pre-specifications and algorithm change protocol, model monitoring, and real-world performance monitoring.]

Legend:
- AI Model Development
- Proposed TPLC Approach
- AI Production Model
- AI Device Modifications
- Real-World Performance Monitoring
Where do Medical Physicists fit in?

Physicists are technical experts with clinical domain expertise. We are uniquely positioned to shape the future of AI in Medical Physics.

- Data: governance, collection, curation
- Problem definition
- Model development, testing, and tuning
- Workflow design, validation, and implementation
- Supervision, maintenance
- Development and execution of routine QA

What does a QA program for AI look like?

We are no longer testing for constancy.

Courtesy of Leigh Conroy
Metrics to Evaluate Machine Learning Models

- **Classification Accuracy** ➔ Number of correct predictions
- **Logarithmic Loss** ➔ Penalize false classifications
- **Confusion Matrix** ➔ Complete performance of model
- **Area Under the Curve** ➔ Probability of ranking +data higher than –data
- **F1 Score** ➔ Precision of model based on precision and recall
- **Mean Absolute Error** ➔ Prediction minus truth | No direction
- **Mean Squared Error** ➔ Prediction minus truth | Penalize larger errors
Sources of Error in AI-based Error Detection

**Class Imbalance**
There are many more examples of non-errors than errors

**Bias & Underfitting**
Lots of historical data... not collected for this purpose
Inaccurate assumptions/simplifications of the model

**Variance & Overfitting**
Detect the signal, not the noise!
Inter-institutional applicability
Risks of Automated Planning

Poor Model ➔ Include Sub-Optimal Plans in Training

- Complex plans
- Simple plans
- New plans
- Poor quality plan
Risks of Automated Planning

Model Error ➔ Over-fitting resulting in lack of generalization for new patients

- Plans of type A with standardized planning
- Plans of type B with standardized planning
- Plans of type C with standardized planning
Risks of Automated Planning

Model Error ➔ Over-fitting resulting in lack of generalization for new patients

- Plans of type A with standardized planning
- Plans of type B with standardized planning
- Plans of type C with standardized planning
- Plans of type D with inconsistent planning
Risks of Automated Planning

Model Error ➔ Over-fitting resulting in lack of generalization for new patients

- Plans of type A with standardized planning
- Plans of type B with standardized planning
- Plans of type C with standardized planning
- Plans of type D with inconsistent planning
Risks of Automated Planning

Unrealistic Dose Prediction ➔ Optimization Loop Cannot Achieve Predicted Dose
Risks of Automated Planning

Unrealistic Dose Prediction ➔ Optimization Loop Cannot Achieve Predicted Dose
Risks of Automated Planning

Outlier Data

Model Training ➔ Model Testing ➔ Data Collection ➔ Data Review ➔ Remove

- Acceptable
- Unacceptable

Sub-optimal Plans

Bias Stability ➔ Data Half-Life
How many samples do I need in my training dataset?

Accuracy | Number of Training Atlases
Accuracy is the % of voxels with similar radiation dose to ground truth
The Open Knowledge-Based Planning Challenge (OpenKBP)

An AAPM Grand Challenge

The aim of the OpenKBP Challenge is to advance fair and consistent comparisons of dose prediction methods for knowledge-based planning (KBP). Participants of the challenge will use a large dataset to train, test, and compare their prediction methods, using a set of standardized metrics, with those of other participants.
AAPM Grand Challenge 2020

Open Knowledge-Based Planning (KBP)

**Objective**
Implement the most accurate KBP dose prediction method on a large open-access dataset

**Evaluation**
All models are trained on the same data, and evaluated with standard metrics

Courtesy of Aaron Babier and Tim Chan
AAPM Grand Challenge 2020
Open Knowledge-Based Planning (KBP)

March 1 — April 30

Prize (Top 2):
Free registration to the 2020 AAPM/COMP Annual Meeting in Vancouver
The Competition

**Objective**
Predict Dose using CT images

**Dose Prediction Method**

Planning CT + ROIs

Predicted Dose

Planning CT + ROIs

**Dose Prediction Method**

Planning CT + ROIs

Predicted Dose

**3D**
Predict Full 3D Dose Distributions

**DVH**
Predict DVH Curves

Courtesy of Aaron Babier and Tim Chan
DVH Error Metric

Average difference between ground truth and predictions were evaluated at two and three DVH metrics for OARs and targets, respectively:

OAR DVH metrics
- $D_{0.1cc}$: Highest dose received by any 0.1cc of OAR volume
- $D_{\text{mean}}$: Mean dose to OAR

Targets DVH metrics
- $D_{1\%}$: Highest dose received by any 1% of target volume
- $D_{95\%}$: Highest dose received by any 95% of target volume
- $D_{99\%}$: Highest dose received by any 99% of target volume

Courtesy of Aaron Babier and Tim Chan
## Leaderboard Streams

<table>
<thead>
<tr>
<th></th>
<th>3D Leaderboard</th>
<th>DVH Leaderboard</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ranking</strong></td>
<td></td>
<td>Lowest average error</td>
</tr>
<tr>
<td><strong>Ground truth</strong></td>
<td>Plan dose distribution</td>
<td>Plan DVH</td>
</tr>
<tr>
<td><strong>Error measure</strong></td>
<td>Voxel-by-voxel mean absolute error</td>
<td>Mean absolute error at dose volume metrics</td>
</tr>
<tr>
<td><strong>3D stream</strong></td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td><strong>DVH stream</strong></td>
<td>✗</td>
<td>✔</td>
</tr>
</tbody>
</table>

Courtesy of Aaron Babier and Tim Chan
Leaderboard Phases

**Validation phase**
Each team’s best validation submission was displayed on a public leaderboard

**Testing Phase**
Each team made one submission to the testing phase, which was used to determine winners
Scores were blinded until competition closed
Summary of Participation

<table>
<thead>
<tr>
<th></th>
<th>Registration</th>
<th>Active in Validation</th>
<th>Active in Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total participants</td>
<td>195</td>
<td>73</td>
<td>54</td>
</tr>
<tr>
<td>Total teams</td>
<td>129</td>
<td>44</td>
<td>28</td>
</tr>
<tr>
<td>Number of submissions</td>
<td>---</td>
<td>1750</td>
<td>28</td>
</tr>
</tbody>
</table>

New interest to research community

- 57% say primary research is Not “Medical Physics”
- 62% say NEVER done KBP research before
Thanks for your attention