Automation Bias in the Clinical Deployment of AI

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 an AI startup, licensee of technology for Machine Learning-based Automated Quality Assurance in Radiation Oncology
## Acknowledgements

<table>
<thead>
<tr>
<th>Contributions</th>
<th>Acknowledgements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ale Berlin</td>
<td>Hitesh Dama</td>
</tr>
<tr>
<td>Peter Chung</td>
<td>David Hodgson</td>
</tr>
<tr>
<td>Leigh Conroy</td>
<td>Tony Fyles</td>
</tr>
<tr>
<td>Tim Craig</td>
<td>Kathy Han</td>
</tr>
<tr>
<td>Chris McIntosh</td>
<td>Kesavi Kanagasabai</td>
</tr>
<tr>
<td></td>
<td>Aly Khalifa</td>
</tr>
<tr>
<td></td>
<td>Tony Tadic</td>
</tr>
<tr>
<td></td>
<td>Jeff Winter</td>
</tr>
<tr>
<td></td>
<td>Aaron Babier</td>
</tr>
<tr>
<td></td>
<td>Tim Chan</td>
</tr>
<tr>
<td></td>
<td>Andrea McNiven</td>
</tr>
<tr>
<td>Peter Chung</td>
<td>David Hodgson</td>
</tr>
<tr>
<td>Tony Fyles</td>
<td>Kesavi Kanagasabai</td>
</tr>
<tr>
<td>Tony Tadic</td>
<td>Jeff Winter</td>
</tr>
<tr>
<td>Chris McIntosh</td>
<td>Aly Khalifa</td>
</tr>
<tr>
<td>Tim Craig</td>
<td>Peter Chung</td>
</tr>
<tr>
<td>Chris McIntosh</td>
<td>Hitesh Dama</td>
</tr>
<tr>
<td>Peter Chung</td>
<td>David Hodgson</td>
</tr>
<tr>
<td>Tony Fyles</td>
<td>Kesavi Kanagasabai</td>
</tr>
<tr>
<td>Tony Tadic</td>
<td>Jeff Winter</td>
</tr>
<tr>
<td>Chris McIntosh</td>
<td>Aly Khalifa</td>
</tr>
<tr>
<td>Peter Chung</td>
<td>Tim Craig</td>
</tr>
<tr>
<td>Tony Fyles</td>
<td>Andrea McNiven</td>
</tr>
<tr>
<td>Tony Tadic</td>
<td>Chris McIntosh</td>
</tr>
<tr>
<td>Chris McIntosh</td>
<td>Peter Chung</td>
</tr>
<tr>
<td>Tony Fyles</td>
<td>Hitesh Dama</td>
</tr>
<tr>
<td>Tony Tadic</td>
<td>David Hodgson</td>
</tr>
<tr>
<td>Chris McIntosh</td>
<td>Kesavi Kanagasabai</td>
</tr>
<tr>
<td>Peter Chung</td>
<td>Aly Khalifa</td>
</tr>
<tr>
<td>Tony Fyles</td>
<td>Tim Craig</td>
</tr>
<tr>
<td>Tony Tadic</td>
<td>Andrea McNiven</td>
</tr>
</tbody>
</table>

*The Princess Margaret Cancer Foundation*
Rapid-Learning System for Cancer Care

Amy P. Abernethy, Lynn M. Etheredge, Patricia A. Ganz, Paul Wallace, Robert R. German, Chalapathy Neti, Peter B. Bach, and Sharon B. Murphy

Abstract

Compelling public interest is propelling national efforts to advance the evidence base for cancer treatment and control measures and to transform the way in which evidence is aggregated and applied. Substantial investments in health information technology, comparative effectiveness research, health care quality and value, and personalized medicine support these efforts and have resulted in considerable progress to date. An emerging initiative, and one that integrates these converging approaches to improving health care, is “rapid-learning health care.” In this framework, routinely collected real-time clinical data drive the process of scientific discovery, which becomes a natural outgrowth of patient care. To better understand the state of the rapid-learning health care model and its potential implications for oncology, the National Cancer Policy Forum of the Institute of Medicine held a workshop entitled “A Foundation for Evidence-Driven Practice: A Rapid-Learning System for Cancer Care” in October 2009. Participants examined the elements of a rapid-learning system for cancer, including registries and databases, emerging information technology, patient-centered and data-driven clinical decision support, patient engagement, culture change, clinical practice guidelines, point-of-care needs in clinical oncology, and federal policy issues and implications. This Special Article reviews the activities of the workshop and sets the stage to move from vision to action.

J Clin Oncol 28:4268-4274. © 2010 by American Society of Clinical Oncology
Computer-Aided Diagnosis and Decision Support Systems

Increase in Data Required for Medical Decision Making Relative To Human Cognitive Capacity

A plethora of data:

→ NOT synthesized medical knowledge

→ confounds generation of needed information

→ obscures decision-making
AI for the Clinic ➔ AI in the Clinic

- Tension between Accuracy and Interpretability
- Machine Prediction meets Human Judgement
- Workflow Integration
Do no harm: a roadmap for responsible machine learning for health care

Jenna Wiens1,2,20*, Suchi Saria3,4,20, Mark Sendak5, Marzyeh Ghassemi2,3,4, Vincent X. Liu6, Finale Doshi-Velez10, Kenneth Jung11, Katherine Heller12,13, David Kale14, Mohammed Saeed15, Pilar N. Ossorio16, Sonoo Thadaney-Israeli17 and Anna Goldenberg6,13,18,19,20*

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).

The potential impact of machine learning (ML) in health care warrants genuine enthusiasm, but its limited adoption in clinical care to date indicates that many of the current strategies are far from optimal. Although successful translation requires bringing together expertise and stakeholders from many disciplines, the development of ML solutions is currently occurring in silos. Past work has tackled particular aspects of data-analysis challenges separating causation from correlation, identifying lurking biases in data and regulating predictive analytics. Here, we take a step back, providing an overarching view of the barriers to deployment and translational impact. With a view toward accelerating safe, ethically responsible and meaningful progress in ML for health care, we lay out critical steps to consider when designing, testing and deploying new solutions. ML deployment in any field should be carried out by interdisciplinary teams including knowledge experts, decision-makers and users (Table 1). Accordingly, this roadmap is intended for a broad audience, while making specific recommendations for critical contributors to such initiatives.

Choosing the right problems

Progress in ML for health care to date has been limited by the lack of well-defined questions and a dearth of annotated datasets. Many ML researchers remain focused on questions for which annotations are readily available, without necessarily questioning the clinical relevance of the problems and their solutions. For example, a popular benchmark challenge in the community focuses on predicting in-hospital mortality on the basis of data collected during the first 48 hours after admission to the intensive care unit. Clearly annotated data are publicly available, and in recent years, performance on this task has approached an area under the curve (Box 1) of 0.9 (ref. 7). However, assessing clinical utility requires careful evaluation of the model against the scenario in which the model will be used. For example, a model may learn to associate patterns of end-of-life care with a high risk of mortality, as a result, despite the high area under the curve, 5

Choosing the right problems

- clinical relevance?
- appropriate data?
- collaborators?
- definition of success?

Rigorous evaluation and thoughtful reporting

- model use?
- sensical predictions?
- shared model/code?
- failure modes?

Making it to market

- medical device?
- model updates?

Developing a useful solution

- data provenance?
- ground truth?

Considering the ethical implications

- ethicist engagement?
- bias correction?

Deploying responsibly

- prospective performance?
- clinical trial?
- safety monitoring?

Box 1 | Glossary of ML terms

| Area under the curve | A measure of discriminative performance that summarizes the trade-off between sensitivity and specificity. |
| Label leakage | When the labels (that is, outcomes) of interest are erroneously (perhaps implicitly) included in the input. |
| Model | A learned mapping from some input (for example, covariates representing a patient) to some output (for example, risk of mortality). |
| Operating regime | Most models output continuous estimates that can then be thresholded; the choice of threshold corresponds to a specific sensitivity, specificity, positive predictive value and so forth, and reflects the operating regime. |
| Stepped-wedge trial | Participants receive treatments in ‘waves’ rather than complete randomization (for example, intervention is applied gradually one unit at a time). |

Fig. 1 | A roadmap for deploying effective ML systems in health care. By following these steps and engaging relevant stakeholders early in the process, many issues stemming from the complexity of adopting ML in practice can be successfully avoided.
DEEP TROUBLE FOR DEEP LEARNING

ARTIFICIAL-INTELLIGENCE RESEARCHERS ARE TRYING TO FIX THE FLAWS OF NEURAL NETWORKS.

A self-driving car approaches a stop sign, but instead of slowing down, it accelerates into the busy intersection. An accident report later reveals that four small rectangles had been stuck to the face of the sign. These fooled the car’s onboard artificial intelligence (AI) into misreading the word ‘stop’ as ‘speed limit 60’.

Such an event hasn’t actually happened, but the potential for sabotaging AI is very real. Researchers have already demonstrated how to fool an AI system into misreading a stop sign, by carefully positioning stickers on it. They have deceived facial-recognition systems by sticking a printed pattern on glasses or hats. And they have tricked speech-recognition systems into hearing phantom phrases by inserting patterns of white noise in the audio.

These are just some examples of how easy it is to break the leading pattern-recognition technology in AI, known as deep neural networks (DNNs). These have proved incredibly successful at correctly classifying all kinds of inputs, including images, speech and data on consumer preferences. They are part of daily life, running...

© 2019 Springer Nature Limited. All rights reserved.

10 OCTOBER 2018 | VOL 574 | NATURE | 362
Fooling AI

Deep Neural Network (DNNs) are Brilliant at Image Recognition ... But are Easily Hacked

- Stop Sign + Stickers
- 45 km/h Speed Limit Sign
- Abstract Pattern
- Starfish
- Seemingly Identical
- Panda + Noise
- Sloth + Racecar
- Panda
- Gibbon
- Starfish
- Racecar

Override
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.
Radiomic Signature for Outcome Prediction

Hypothesis → Computers are better equipped to extract hidden information in the images

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,s

a Department of Medical Biophysics, University of Toronto; b Department of Radiation Oncology, University of Toronto; c Ontario Institute of Cancer Research, Toronto; d IBBME, University of Toronto; e Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto; f The Techna Institute for the Advancement of Technology for Health, Toronto, Canada; g Department of Radiation Oncology (MAASTRO), GROW Research Institute, Maastricht University, the Netherlands; h Dana-Farber Cancer Institute, Brigham and Women’s Hospital, Harvard Medical School, Boston, USA; i Princess Margaret Cancer Centre, University Health Network; and s Vector Institute, Toronto, Canada
Image Signal Dependence of Radiomic Signature

Does randomizing image voxel locations affect prognostic accuracy of a common radiomic signature?

Original Patient Images

Voxel Randomized Images

Bottom line
Imaging signal intensities not pertinent to model performance

AUC = 0.64 (CI=0.60-0.68)

AUC = 0.64 ± 0.0003
Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer

Chris McIntosh, Leigh Conroy, Michael C. Tjong, Tim Craig, Andrew Bayley, Charles Catton, Mary Gospodarowicz, Joelle Helou, Naghmeh Isfahanian, Vickie Kong, Tony Lam, Srinivas Raman, Padraig Warde, Peter Chung, Alejandro Berlin and Thomas G. Purdie
Clinical Validation Framework for Prospective Deployment

New Standard of Care for Curative Intent Prostate Radiation Therapy

- Proof of Concept
- Technical Validation
- Retrospective Validation
- Feasibility Study
- Retrospective Clinical Validation
- Prospective Clinical Deployment

Don’t Have to Repeat These Steps

The Chasm
Blinded Head to Head Evaluation with Standardized Form

Prostate AutoPlanning Evaluation

Reviewer

Patient ID (First 3 digits of MRN only)

Qualitative Evaluation

- Target Coverage
- OAR Sparing
- High Dose Conformity
- Rectal Dose Gradient
- Lateral Dose Symmetry

Treatment Plan Acceptability

- Prostate_A1
- Prostate_A2

Treatment Plan Selected

- Prostate_A1
- Prostate_A2
- Equivalent ➔ go with Prostate_A1
- Equivalent ➔ go with Prostate_A2
- Do not like either plan

Pepsi Challenge (~Turing Test)

- Prostate_A1
- Prostate_A2

Comments (optional)

Submit

Plan Comparison. Which plan has better:

- Plan Acceptability
- Preferred Plan

- Which plan do you think is the Automated Plan?

Prostate_A1
Prostate_A2

They are Equivalent

Target Coverage
OAR Sparing
High Dose Conformity
Rectal Dose Gradient
Lateral Dose Symmetry

- Acceptable
- Unacceptable

Preferred Plan

Prostate_A1
Prostate_A2

Equivalent ➔ go with Prostate_A1
Equivalent ➔ go with Prostate_A2
Do not like either plan

Do not like either plan

Automated Plan

Prostate_A1
Prostate_A2

*VTTLU[ZVW[PVUHS

Submit
### Prostate AutoPlanning Evaluation

**Reviewer**

**Patient ID (First 3 digits of MRN only)**

<table>
<thead>
<tr>
<th>Qualitative Evaluation</th>
<th>Prostate_A1</th>
<th>Prostate_A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Coverage</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>OAR Spacing</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>High Dose Conformity</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Rectal Dose Gradient</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Lateral Dose Symmetry</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment Plan Acceptability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate_A1</td>
</tr>
<tr>
<td>Prostate_A2</td>
</tr>
</tbody>
</table>

**Pepsi Challenge (~Turing Test)**

<table>
<thead>
<tr>
<th>Treatment Plan Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate_A1</td>
</tr>
<tr>
<td>Prostate_A2</td>
</tr>
<tr>
<td>Equivalent go with Prostate_A1</td>
</tr>
<tr>
<td>Equivalent go with Prostate_A2</td>
</tr>
<tr>
<td>Do not like either plan</td>
</tr>
</tbody>
</table>

**Comments (optional)**

---

### Consensus

**Actual Selected**

**Perceived Selected**

**Quantitatively Superior**
Radiation Oncologist Preference Toward ML or Human?

- ML Model is Applicable:
  - Simulation (n = 50): 94%
  - Deployment (n = 50): 86%
  - Overall (n = 100): 90%

- ML Selected for Treatment:
  - Simulation (n = 47): 83%
  - Deployment (n = 43): 61%
  - Overall (n = 90): 72%
Radiation Oncologist Preference Toward ML or Human?

- **ML Superior on Consensus Review**
  - Simulation (n = 47) - 81%
  - Deployment (n = 43) - 72%

- **ML Selected for Treatment**
  - Simulation (n = 47) - 83%
  - Deployment (n = 43) - 61%
Radiation Oncologist Preference Toward ML or Human?

- ML Superior on Consensus Review:
  - Simulation (n = 47): 81%
  - Deployment (n = 43): 72%

- ML Selected when Superior:
  - Simulation (n = 47): 92%
  - Deployment (n = 43): 71%
Radiation Oncologist Preference Toward ML or Human?

Toward ML = 6

Toward Human = 13

Number of Plans with Radiation Oncologist Showing Preference
Is ML+Humans The Answer?

Prospective Study → Human in the Loop

Phase 0.1 Feasibility

Phase 1 Prospective

Phase 2 Collaboration

Prospective Study vs Human in the Loop

Phase 0.1 Feasibility

Phase 1 Prospective

Phase 2 Collaboration

Feasibility

Prospective

Collaboration

Vs

+
Performance stability evaluation of atlas-based machine learning radiation therapy treatment planning in prostate cancer

Leigh Conroy\textsuperscript{1,2,3}, Aly Khalifa\textsuperscript{3}, Alejandro Berlin\textsuperscript{1,2,4}, Chris McIntosh\textsuperscript{3,4,5,6,7} and Thomas G Purdie\textsuperscript{1,2,3,4}

\textsuperscript{1} Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto, Canada
\textsuperscript{2} Department of Radiation Oncology, University of Toronto, Toronto, Canada
\textsuperscript{3} Department of Medical Biophysics, Faculty of Medicine, University of Toronto, Toronto, Canada
\textsuperscript{4} Techna Institute, University Health Network, Toronto, Canada
\textsuperscript{5} Peter Munk Cardiac Centre, University Health Network, Toronto, Canada
\textsuperscript{6} Joint Department of Medical Imaging, University Health Network, Toronto, Canada
\textsuperscript{7} Vector Institute, Toronto, Canada

Keywords: machine learning, automated treatment planning, atlas-selection, dose prediction, external beam radiotherapy, quality assurance, benchmarking
How Stable is Our AI Model to Input Data Variations?

→ Test Performance Impact of Input Data Variations
→ Benchmarked Against Ground Truth “Base” Plan (Testing ↔ Training)

- Presence/absence of organs
- Organ contouring practice
- Prostate margins

Courtesy of Leigh Conroy
Atlas-Based Treatment Planning

**Training**
- Images, ROI & Dose
- Feature Extraction
- Feature-to-Dose Mapping
- Atlas Selection Learning

**Novel Patient**
- Images & ROIs
- Feature Extraction
- Atlas Selection
- Dose Prediction
- Final Deliverable Plan

Images & ROIs → Feature Extraction → Atlas Selection → Dose Prediction → Final Deliverable Plan

Atlas Selection Learning

Courtesy of Leigh Conroy and Chris McIntosh
Impact of Image Input Change on Atlas Selection

→ Atlases selected used to indicate magnitude the AI process is disturbed
→ The number of Flips is the change in the atlases selected between the ground truth “base” plan and the plan with input changes (0-5)

![Graph showing impact of image input change on atlas selection](image)

- **P8: All mean = 3.6 [2.1, 5.0]**
- **Number of Patients**
- **Number of Flips**

- **Atlases Selected for Ground Truth “Base” Plan**
  - A, B
  - C, D, E
  - F, G

- **Atlases Selected for Plan with Input Changes**
  - C, D, E
  - F, G

- **Flips from “Base” Plan**
  - 1, 2, 3

Courtesy of Aly Khalifa, Chris McIntosh, Jeff Winter
How Stable is Our ML Model to Input Data Variations?

Treatment Planning Scenarios Investigated

<table>
<thead>
<tr>
<th>Scenario</th>
<th>P0(Base)</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5-P7</th>
<th>P8 (All)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTV</td>
<td>10/7 mm</td>
<td>5 mm</td>
<td>0 mm</td>
<td>——</td>
<td>——</td>
<td>——</td>
<td>0 mm</td>
</tr>
<tr>
<td>Bladder/Rectum</td>
<td>Wall 3mm</td>
<td>——</td>
<td>——</td>
<td>Wall 5mm</td>
<td>Whole Organ</td>
<td>Present (Distant)</td>
<td>Present (Distant)</td>
</tr>
<tr>
<td>Small/Large Bowel</td>
<td>None</td>
<td>——</td>
<td>——</td>
<td>——</td>
<td>——</td>
<td>——</td>
<td>——</td>
</tr>
</tbody>
</table>

Present (Distant)
5 ➔ Small Bowel
6 ➔ Large Bowel
7 ➔ S+L Bowel

Courtesy of Leigh Conroy
Using Atlas Distance To Understand Model Behaviour

Number of Atlas Changes from Base Plan Atlases for each Scenario (25x5 Atlases)

➔ PTV Margins and Inclusion of Small and Large Bowel Most Impactful

Spatial Distance

Dose-Volume Histogram Based Distance

Courtesy of Leigh Conroy
Potential Clinical Practice Change

- New contour script
- Decreased PTV margins
- New dose fractionations
- Use of OAR spacer

Courtesy of Leigh Conroy
Challenges and Opportunities

**AI → Automation** will be an essential component for many processes in radiation oncology

Radiation treatment planning using AI is promising for clinical use
→ in the clinic now
→ need to understand potential automation bias

**Human in the Loop**
→ a pre-cursor to complete automation
→ promote human|machine collaboration

**AI performance** evaluation should be done initially and routinely
→ Evaluate in the real clinical environment against standard practice
→ Benchmark against a gold standard
→ Assess need for re-training (or not?) even with observed practice changes