Creating Innovations that Matter
Deep Learning for Medical Imaging

Christine Swisher, PhD

Philips Research North America
IMAGENET Accuracy Rate

ML watches YouTube for three straight days! (and learns to recognize cats?!!)

a plate of food, food on a plate, a blue cup on a table, a plate of food, a blue bowl with red sauce, a bowl of soup, a cup of coffee, ...
Adapting to Artificial Intelligence
Radiologists and Pathologists as Information Specialists

Artificial intelligence—the mimicking of human cognition by computers—was once a topic in science fiction but is becoming reality in medicine. The combination of big data and artificial intelligence, referred to by some as the fourth industrial revolution, will change radiology and pathology along with other medical specialties. Although reports of radiologists and pathologists being replaced by computers seem exaggerated, these specialties must plan strategically for a future in which artificial intelligence is part of the health care workforce. Radiologists have already converted to digital technology. In 1960, Lord predicted that “a scanner and computer to examine chest films—so as to separate the clearly abnormal chest film from the abnormal chest films.” Lord further suggested that “the abnormal chest films would be marked for later study by the radiologist.” This progress in imaging has changed the work of radiologists. Radiology once confined to projectional images, such as chest radiographs, has become more complex and data rich. Cross-sectional imaging such as CT and magnetic resonance, by showing anatomy with greater clarity, has made diagnosis simpler in many instances; for example, a ruptured aneurysm is inferred on a chest radiograph but actually seen on CT. However, this has come at a price—the amount of data has increased markedly. For example, a radiologist typically views 4000 images in a CT scan of multiple body parts (“pan scan”) in patients with multiple trauma. The abundance of data has changed how radiologists interpret images, from pattern recognition, with clinical context, to searching for needlestick injuries, from inference to detection. Their radiologist, once a maestro with a chest ra-

“Deep learning technology applied to medical imaging may become the most disruptive technology radiology has seen since the advent of digital imaging.” —Nadim Daher

“Radiologists and pathologists need not fear artificial intelligence but rather must adapt incrementally to artificial intelligence, retaining their own services for cognitively challenging tasks.” —Eric Topol
Chest Defense: Deep Learning Spots Disease Early Using Chest X-Rays
Deep Learning is Everywhere!
A Street Vendor in China

Deep Learning Service - System Development & Testing
Caffe installation: 10 Yuan = $1.5
CNN: 5 Yuan = $0.75 per layer
RNN: 8 Yuan = $1.2 per layer
The three rules of meaningful ML innovation still apply

1. Eyes on the Prize
2. Involvement of the World Outside of ML
3. Meaningful Evaluation Methods
“With this positive trial result (NLST), we have the opportunity to realize the greatest single reduction of cancer mortality in the history of the war on cancer.”

– James Mulshine, MD
Three rules of meaningful ML innovation

1. Eyes on the Prize
   • How significant is the impact of a solution to the problem?
   • How many lives would it change? What is a severe unmet need we can overcome?
   • What would constitute a meaningful improvement over the status quo?

2. Involvement of the World Outside
   • Co-creation with clinicians
   • Feedback from hospital infrastructure and hospital administrator
   • Involve experts in business models, marketing & sales
   • *Know your data!!!*

3. Meaningful Evaluation Methods
   • Performance in multisite clinical trials
   • Machine vs Human vs Machine + Human
   • Improvement of clinical outcome
Lung Screening at a Glance

It Causes a Lot of Deaths
Lung cancer is the number-one cancer killer, taking more lives than colon, breast and prostate cancer combined.

Urgent need: Lung cancer kills 450 people every day in the US alone.

Source: Onco Iss 2014

Early Diagnosis is Critical
Reduced Mortality:
Generally, early detection can increase five-year survival by nearly 90%.

Source: NEJM 2006

Expected Widespread Adoption
In 2015, the CMS added annual screening for lung cancer with LDCT ensuring that 3-4 million high-risk patients could get lifesaving intervention regardless of income level.

Source: NYTimes 2014.
Recommendation by NCCN and USPSTF.
Failure to screen lawsuits favor patients
Ex: DC jury awards $5M for failure to screen for cancer

Christine Swisher - Guest Seminar, MIT: Machine Learning for Healthcare, Spring 2017
Lung Screening at a Glance

**IT CAUSES A LOT OF DEATHS**
Lung cancer is the number-one cancer killer, taking more lives than colon, breast and prostate cancer combined.

**Urgent need:** Lung cancer kills 450 people every day in the US alone.

Source: Onco Iss 2014

**EARLY DIAGNOSIS IS CRITICAL**
Reduced Mortality:
Generally, early detection can increase five-year survival by nearly 90%.

Source: NEJM 2006

**EXPECTED WIDESPREAD ADOPTION**
In 2015, the CMS added annual screening for lung cancer with LDCT ensuring that 3-4 million high-risk patients could get lifesaving intervention regardless of income level.
Source: NYTimes 2014.
Recommendation by NCCN and USPSTF.

Failure to screen lawsuits favor patients
Ex: DC jury awards $5M for failure to screen for cancer

**Asymptomatic**
- Stage I: 90% 5yr OS
- Stage IV: 1% 5yr OS

**Symptomatic**
- 58% 5yr OS

Christine Swisher - Guest Seminar, MIT: Machine Learning for Healthcare, Spring 2017
Challenges for Adoption of LDCT

Cognitive Challenges:
• Vast majority are negative ~89.4%
• Satisfaction of search
• Volume and complexity of information

False Positives
• 96.4% FP of positive readings by LDCT
• Most have noninvasive imaging follow-up
• Invasive diagnosis procedure: 2.6%
• Complication rate: 1.4% (0.06% Major)

Overdiagnosis: More than 18% seem to be indolent.
• Bronchioloalveolar carcinoma 79%; NSCLC 22% are overdiagnosed
• Risk: 11% by LDCT vs no screening and 9% vs CXR (lifetime follow-up)
Challenges for Adoption of LDCT

**Cognitive Challenges:**
- Vast majority are negative ~89.4%
- Satisfaction of search
- Volume and complexity of information

**False Positives**
- 96.4% FP of positive readings by LDCT
- Most have noninvasive imaging follow-up
- Invasive diagnosis procedure: 2.6%
- Complication rate: 1.4% (0.06% Major)

**Overdiagnosis:** More than 18% seem to be indolent.
- Bronchioloalveolar carcinoma 79%; NSCLC 22% are overdiagnosed
- Risk: 11% by LDCT vs no screening and 9% vs CXR (lifetime follow-up)
Class Imbalance

True positives and rare incidental findings, by virtue of being rare, are underrepresented. If not accounted for properly, the class imbalance will occur biasing the model to predict the healthy-label.

- 1000 samples (963 Negative; 37 positives)
- Network learns that all are negative
- Accuracy of 96.3% and PPV = 0
Class Imbalance

True positives and rare incidental findings, by virtue of being rare, are underrepresented. If not accounted for properly, the class imbalance will occur biasing the a model to predict the healthy-label.

- Augmentation of underrepresented class*
- Train on an easier problem
- Weight the loss function
- Pre-training for lower level features

*Underrepresented class should have examples of various ways rare class can present.

18% are indolent (BAC 79%; broadly NSCLC 22%)
Goals

1. *Reduce time* and *cognitive load* for radiologists reading LDCT images
2. Reduce unnecessary escalation and resultant complications due to *false positives* reads
Three rules of meaningful ML innovation

1. Eyes on the Prize
   • How significant is the impact of a solution to the problem?
   • How many lives would it change? What is a severe unmet need we can overcome?
   • What would constitute a meaningful improvement over the status quo?

2. Involvement of the World Outside
   • Co-creation with clinicians
   • Feedback from hospital infrastructure and hospital administrator
   • Involve experts in business models, marketing & sales
   • *Know your data!!!*

3. Meaningful Evaluation Methods
   • Performance in multisite clinical trials
   • Machine vs Human vs Machine + Human
   • Improvement of clinical outcome

Christine Swisher - Guest Seminar, MIT: Machine Learning for Healthcare, Spring 2017
Value

Hospital:
- Reduce costs associated with unnecessary care escalation (10BE/yr on US health system)
- Reduced mis-diagnoses and resultant resource utilization
- Identify high risk patients for follow-up

Patient:
- Improved outcomes (quality of life, mortality, cost)

Staff:
- Increase staff efficiency (improve throughput/reduce radiologist man hours)

Health System:
- Estimates of total health expenditures for a national screening program range from $1B to $3B annually, constituting a 20% increase in expenditure for lung cancer overall.
What is the FDA approval process?

“Soft” Use-Case
• Current regulatory situation is reminiscent of the early days of computer-aided detection (CADe) devices.
• Cleared under the 510[k] process

“Hard” Use-case:
• Likely regulated as Class 2, even Class 3
• Requires a large randomized clinical trial
• Similar to computer-aided diagnosis (CADx) applications, which required premarket approval (PMA) process.
Data

New AI method (Deep Learning)

Most learning algorithms

Unique challenges for medical images
Image characteristics are 3+ dimensional
Commonly used transfer learning input that leverages the 3D structures

This is just one simple example. There are many approaches to take 3D structures into account. There are obvious limitations to this approach.
Image characteristics are 3+ dimensional
Multimodal, Multiple Reconstructions, Registration Challenges

Sarah Nelson. UCSF’s Neuroradiology Research Laboratory.
Scale Variance

Cat

Also a Cat

Negative Finding

Positive Finding

Follow-up Diagnostic Tests

Christine Swisher - Guest Seminar, MIT: Machine Learning for Healthcare, Spring 2017
The scope of this paper is more about the value of HDR. Here, we are highlighting the insight that going from a HDR to LDR (e.g. 16-bit to 8-bit image) will destroy important image characteristics and reduce performance in computer vision tasks. This is particularly important in radiology and pathology, where images tend to have a higher dynamic range than natural images.

Swisher* & Vinegoni*. Nature Communications (2016); *Contributed equally.

Christine Swisher - Guest Seminar, MIT: Machine Learning for Healthcare, Spring 2017
High Dynamic Range

Images with high dynamic range do better in computer vision tasks

The scope of this paper is more about the value of HDR. Here, we are highlighting the insight that going from a HDR to LDR (e.g. 16-bit to 8-bit image) will destroy important image characteristics and reduce performance in computer vision tasks. This is particularly important in radiology and pathology, where images tend to have a higher dynamic range than natural images.

Swisher* & Vinegoni*. Nature Communications (2016); *Contributed equally.
The scope of this paper is more about the value of HDR. Here, we are highlighting the insight that going from a HDR to LDR (e.g. 16-bit to 8-bit image) will destroy important image characteristics and reduce performance in computer vision tasks. This is particularly important in radiology and pathology, where images tend to have a higher dynamic range than natural images.

Swisher* & Vinegoni*. Nature Communications (2016); *Contributed equally.

Christine Swisher - Guest Seminar, MIT: Machine Learning for Healthcare, Spring 2017
Commonly used transfer learning input that leverages the full dynamic range

LDR = Low Dynamic Range; rHDR = reconstructed High Dynamic Range image at a Low dynamic range
This is just one simple example. There are many approaches to utilize HDR characteristics. There are obvious limitations to this approach.
Downsampling - We must be creative in how we tackle dimensionality

Examples of healthy tissue and typical interstitial lung disease patterns (link to paper). Left to right: Healthy, ground glass opacity, micronodules, consolidation, reticulation, honeycombing, combination of ground glass and reticulation.

Clinical significant features look like noise.

Still looks like a woman.
## Transfer Learning

<table>
<thead>
<tr>
<th></th>
<th>Very Similar Dataset</th>
<th>Very Different Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Small dataset</strong></td>
<td>Use Linear classifier on top layer</td>
<td>This is going to be challenging!</td>
</tr>
<tr>
<td><strong>Large dataset</strong></td>
<td>Fine-tune a few layers</td>
<td>Fine-tune a large number of layers</td>
</tr>
</tbody>
</table>
Value of pre-training for DL tasks:

Aids in ambitious DL tasks:
Learning the ‘easier’ localization (regression) task served as ‘stepping stone’ for learning the detection task: the weights learned for localization were close enough to what was needed for detection to allow convergence.

Multitask Capability:
Network detects and localizes

Transparency:
Easier to understand and justify the output of DNNs

Re-use:
Re-use successful DNNs for new tasks
“Deep Learning is a black box” – most physicians

Feature understanding

Uncertainty


Christine Swisher - Guest Seminar, MIT: Machine Learning for Healthcare, Spring 2017
“Deep Learning is a black box” – most physicians

Feature understanding

Uncertainty

Paper from Philips Research [link]

Must read blog [Link]
Three rules of meaningful ML innovation

1. Eyes on the Prize
   • How significant is the impact of a solution to the problem?
   • How many lives would it change? What is the severe unmet need we can overcome?
   • What would constitute a meaningful improvement over the status quo?

2. Involvement of the World Outside
   • Co-creation with clinicians
   • Feedback from hospital infrastructure and hospital administrator
   • Involve experts in business models, marketing & sales
   • *Know your data!!!*

3. Meaningful Evaluation Methods
   • Generalization - multisite clinical trails, sustainability to changes in technology
   • Machine vs Human vs Machine + Human
   • Improvement of clinical outcome
# Probability of Cancer in Pulmonary Nodules Detected on First Screening CT

## Table 1. Prediction Model for the Probability of Lung Cancer in Pulmonary Nodules

<table>
<thead>
<tr>
<th>Predictor Variable</th>
<th>Model A</th>
<th>Model B</th>
<th>Model C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex, female</td>
<td>1.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family history of lung cancer, yes, no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupational history, yes, no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model score</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Model A**
- Predictive accuracy: 0.75 (Mean of 6 Observers)
- Radiologist Alone

**Model B**
- Predictive accuracy: 0.75 (Mean of 6 Observers)
- Radiologist + Model

**Model C**
- Predictive accuracy: 0.85 (Mean of 6 Observers)
- Model Alone

---

Christine Swisher - Guest Seminar, MIT: Machine Learning for Healthcare, Spring 2017

---

**Link to article**
Look at the star in the center
There is an X in this image, can you find it?
How many people noticed the T?
Think about where and when the algorithm will be used so that it will actually deliver improved clinical outcomes.
Example of meaningful evaluation metric

\[(\text{AI + Pathologist}) > \text{Pathologist}\]

* Error rate defined as 1 – Area under the Receiver Operator Curve
** A study pathologist, blinded to the ground truth diagnoses, independently scored all evaluation slides.

Speaker for next week
Three rules of meaningful ML innovation

1. Eyes on the Prize
   • How significant is the impact of a solution to the problem?
   • How many lives would it change? What is a severe unmet need we can overcome?
   • What would constitute a meaningful improvement over the status quo?

2. Involvement of the World Outside
   • Co-creation with clinicians
   • Feedback from hospital infrastructure and hospital administrator
   • Involve experts in business models, marketing & sales
   • Know your data!!!

3. Meaningful Evaluation Methods
   • Generalization - multisite clinical trials, sustainability to changes in technology
   • Machine vs Human vs Machine + Human
   • Improvement of clinical outcome
Acknowledgements

Philips Research - Eindhoven
Dimitrios Mavroeidis
Stojan Trajanovski
Jack He
Ulf Grossekathofer
Erik Bresch
Binyam Gebre
Teun van Den Heuvel
Bas Veeling
Devinder Kumar
Vlado Menkovski

Philips HealthCare
Homer Pien

Philips Research - Hamburg
Tobias Klinder
Rafael Wiemker

Philips Research – North America
Sadid Hasan
Jonathan Rubin
Cristhian Potes
Yuan Ling
Joey Liu

Lahey Medical Center
Sebastian Flacke
Christoph Wald
Brady Mckee
Ali Ardestani

MGH
Anthony Samir
John Gilbertson

Christine Swisher - Guest Seminar, MIT: Machine Learning for Healthcare, Spring 2017