Translating minimally-supervised NLP to interpretable clinical AI

John Kang, MD, PhD

@DrJohnKang
Potential conflicts of interest

• Family member works for Change Healthcare
Outline: a tour of how NLP has leveraged unstructured data, unsupervised learning, and transfer learning to extract information from patient records

1. Supervised models for clear outcomes using structured data
2. Supervised models for clear outcomes using unstructured data
3. Supervised models for unclear outcomes using unstructured data
4. Supervised models for unclear outcomes using pre-trained unstructured data
5. Unsupervised models for unclear outcomes using pre-trained unstructured data
Real world trials using supervised learning on structured data
But can we go deeper?

Can we also use unstructured data (>80% of data*)

Survival prognosis with deep learning of structured variables AND clinical notes

- 1,390,032 provider notes
- 12,876,137 lab values (200 most common labs)
- 1,451,740 vital signs
- 357,981 diagnoses (500 most common codes)
- 1,162,164 procedures (500 most common codes)
- 1,834,477 medication orders (500 most common meds)

**Table 2. Survival model coefficients for selected note text terms**

<table>
<thead>
<tr>
<th>Term</th>
<th>Coefficient*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptoms/appearance</td>
<td></td>
</tr>
<tr>
<td>Gachectic</td>
<td>0.020</td>
</tr>
<tr>
<td>Fatigued</td>
<td>0.0059</td>
</tr>
<tr>
<td>Ascites</td>
<td>0.0085</td>
</tr>
<tr>
<td>Completely asymptomatic</td>
<td>-0.0054</td>
</tr>
<tr>
<td>Anxious</td>
<td>-0.0031</td>
</tr>
<tr>
<td>Feel well</td>
<td>-0.0073</td>
</tr>
<tr>
<td>Cancer location/response</td>
<td></td>
</tr>
<tr>
<td>Disease progression</td>
<td>0.012</td>
</tr>
<tr>
<td>Leptomeningeal</td>
<td>0.0067</td>
</tr>
<tr>
<td>Mixed response</td>
<td>0.014</td>
</tr>
<tr>
<td>Innumerable pulmonary</td>
<td>0.0046</td>
</tr>
<tr>
<td>Minimal progression</td>
<td>-0.0012</td>
</tr>
<tr>
<td>Oligometastatic</td>
<td>-0.0066</td>
</tr>
<tr>
<td>Systemic therapy agents</td>
<td></td>
</tr>
<tr>
<td>Nivolumab</td>
<td>-0.00065</td>
</tr>
<tr>
<td>Liposomal doxorubicin†</td>
<td>0.011</td>
</tr>
<tr>
<td>Anastrozole†</td>
<td>-0.00051</td>
</tr>
<tr>
<td>Leuprolide†</td>
<td>-0.0037</td>
</tr>
<tr>
<td>Tamoxifen</td>
<td>-0.0034</td>
</tr>
</tbody>
</table>

*A positive coefficient indicates shorter survival.
†Brand name converted to generic name for display.
But what if the outcome is not self-labeled?

Can we predict *ill-defined* events like cancer progression or treatment response?
ConvNet-based architecture on clinical text can detect cancer outcomes
What if we wanted to understand the words in clinical notes?

Can semantic understanding of increase **performance** and/or improve **interpretability**?
A primer on word embeddings

“You shall know a word by the company it keeps”

Mikolov et al. 2011 (Word2Vec)

Guillaume Desagulier tutorial
https://corpling.hypotheses.org/495
Combining semantic map with word embeddings increases interpretability.
What if we do not even know what the labels are?

Can labels be extracted from unlabeled text?
Significant resources used to track radiation oncology research

2013

**The Profession**

**National Institutes of Health Funding in Radiation Oncology: A Snapshot**

Michael Steinberg, MD, William H. McBride, PhD, DSc, Erina Vlashi, PhD, and Frank Pajonk, MD, PhD

Department of Radiation Oncology, David Geffen School of Medicine at University of California, Los Angeles (UCLA), and Jonsson Comprehensive Cancer Center at UCLA, Los Angeles, California

Received Dec 30, 2012, and is revised Jan 22, 2013. Accepted for publication Jan 27, 2013

"At the start of fiscal year 2013 we extracted records for 952 individual grants, which were active at the time of analysis from the NIH database... Our analysis identified 197 grants in radiation oncology."

2014

**Education Original Article**


Paul E. Wallner, DO, Mitchell S. Anscher, MD, Christopher A. Barker, MD, Michael Bassetti, MD, PhD, Robert G. Bristow, MD, PhD, Yong I. Cha, MD, PhD, Adam P. Dicker, MD, PhD, Silvia C. Formenti, MD, **Edward E. Graves, PhD, Stephen M. Hahn, MD, Tom K. Hei, PhD, Alec C. Kimmelman, MD, PhD, David G. Kirsch, MD, PhD,** Kevin R. Kozlik, MD, PhD, **Theodore S. Lawrence, MD, PhD, Brian Marples, PhD, William H. McBride, DSc, Ross B. Mikkelsen, PhD, Catherine C. Park, MD, Joanne B. Weidhaas, MD, PhD, **Anthony L. Zietman, MD, PhD, and Michael Steinberg, PhD**

"The first was... to congress about actual radiation oncology funding levels; the second was a review of the publicly available grant system database... To differentiate biological research from clinical trials and physics research, all radiation oncology grants... were hand-curated, separating the biology grants from the clinical and physics grants. Further, the biology grants were then subdivided by research topic."

2017

**The Profession**

**Analysis of the 2017 American Society for Radiation Oncology (ASTRO) Research Portfolio**

James B. Yu, MD, Tyler F. Beck, PhD, Mitchell S. Anscher, MD, Andrew M. Baschnagel, MD, Kristy K. Brock, PhD, David J. Carlson, PhD, **Michael M. Dominiello, DO**, Randall J. Kimple, MD, PhD, Jonathan P. Kuliszy, MD, Marc S. Mendonca, PhD, **Omar Y. Hian, MD, PhD**, Anurag K. Singh, MD, **Eduardo G. Moros, PhD**, and Judith C. Keen, PhD

*Department of Radiation Oncology, Yale School of Medicine, New Haven, Connecticut; American Society for Radiation Oncology, Arlington, Virginia; Department of Radiation Oncology, MD Anderson Cancer Center, Houston, Texas; Department of Human Oncology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin; Departments of Imaging Physics and Radiation Physics, MD Anderson Cancer Center, Houston, Texas; Department of Radiation Oncology, Karmanos Cancer Institute, Detroit, Michigan; Department of Radiation Oncology, Wexner Comprehensive Cancer Center, Columbus, Ohio; Department of Radiation Oncology, Roswell Park Comprehensive Cancer Center, Buffalo, New York; and Department of Radiation Oncology, Moffitt Cancer Center, Tampa, Florida

Received Mar 22, 2016. Accepted for publication Jul 22, 2016.

Of the grants submitted... a significant number of grants were categorized as "unknown."
What ideas/topics/themes are being funded by NCI in Dept. of Radiology or Radiation Oncology?

Beidler/Nguyen et al., in prep
Methods/results

- 7k grant abstracts converted to BioWordVec embeddings (trained on biomedical+clinical data)

- clustered using combined hierarchical/K-means clustering

- used k=15 centroids (per elbow plot of clustering performance) and k=60 (more realistic)

- manual validated ~5% of grants over 4 raters (different training level) with good concordance
5-7k abstracts → 15 domains

Abstract distribution on TSNE (k=15)

Change in funding per year (k=15)

- 3 fastest growing (15 topics):
  - (0) MRI treatment response
  - (1) Mechanisms of therapy resistance
  - (2) Targeted therapies

- 3 slowest growing (15 topics):
  - (14) Breast screening
  - (13) Pharmacological pathways
  - (12) Tumorigenesis and resistance mechanisms

Beidler/Nguyen et al., in prep
5-7k abstracts->60 domains

• 3 fastest growing (60 topics):
  • (0) Imaging Biomarkers
  • (1) AI Decision Support & Imaging software
  • (2) Radio-pharmaceuticals

• 3 slowest growing (60 topics):
  • (59) Breast cancer CAD
  • (58) Multimodal Imaging R&D
  • (57) Stress Response
Manual validation shows reasonable concordance between human and machines

Not shown: experience level correlates with concordance
Grants wrongly labelled by the algorithm tended to be further from centroid
Funding topics have “emerged” and “disappeared” in last 20 years

First Fiscal Year of the newly emerging research topics

- Cancer Therapy Mechanism
- Image-guided therapy
- Radioimmunotherapy
- Image-guided Drug Delivery
- AI Decision Support
- Metabolic Imaging Markers
- Radio-pharmaceuticals

2000 2001 2003 2004 2005 2006

Last Fiscal Year of the disappeared research topics

- DNA Repair (base mismatch)
- DNA Mutagenesis & Repair
- Epigenetics
- Breast Cancer CAD

2010 2015 2016 2018 2020
Limitations

- Grants further away from the centroid may not seem like they belong
- If new data is entered, the new optimal clustering may appear different
- 1 grant : 1 topic
If a clinical model performs great but never affects patients, is it a useful model?
Collaborators for this work

• Eric Ford

• Kevin Lybarger
  (UW→George Mason)
  [https://www.kevinlybarger.me](https://www.kevinlybarger.me)

Mentees/trainees for this work

• August Anderson

• Peter Beidler

• Mark Nguyen

• Joseph Tsai

• Qian Zhang
  (UW→Northwestern)