Developing Machine Learning Tools for Cancer Treatment Strategies

Jiaming Zeng
Advisor: Ross Shachter
AAAI-20 Doctoral Consortium

Background

In medicine, clinical randomized trials are the only accepted standard for deciding which treatment is better than the other. Retrospective studies from observational data are often plagued by confounding and selection bias[1]. The readily available covariates cannot be used to adjust for confounding. We explore whether we can perform robust causal inference from observational data by including word embeddings from clinical text.

Aims and Contributions

1. **Build language models from clinical data for decision-making and inference.**
   - Established protocol for building natural lang. models from clinical data
   - Publications: AAAI-20 SA (Monday night)
2. **Develop ML tools with unstructured and high-dimensional data for medical decision support.**
   - Status: [in-progress]
   - Adapted causal inference methods to work with clinical notes
3. **Validate developed decision support tools against existing clinical trials.**
   - Status: [in-progress]
   - ***Showed that unstructured notes make a different in correcting for selection bias***

Dataset

- **Source**: Stanford Cancer Institute Research Database (SCIRDB)
- **Total**: 4,420 patients
  - Localized prostate, oropharynx, and esophagus
- **Timeframe**: 2008 – 2019
- **Notes**: 483,782 clinical notes
- **Additional Data**: California Cancer Registry (CCR)
  - Initial treatment information: all treatments performed within 6 months of initial diagnosis
  - Date of death, date of diagnosis, etc.
- **Testing**: reserved 10% of patients for testing

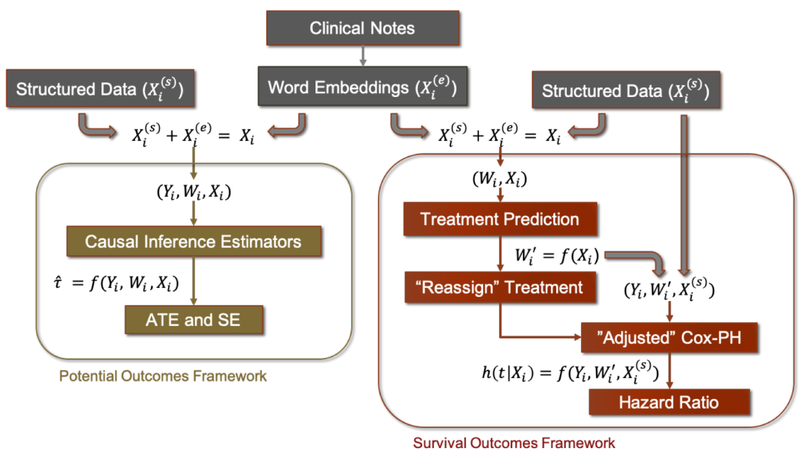
Stanford Cancer Institute
Stanford MEDICINE

Methodology

Natural Language Processing (NLP) Models

- **Notes**: 483,782 clinical notes (excluded 10% for testing)
- **Model**: Doc2vec[3]
  - Trained 324 doc2vec models for generating word embeddings
    - vector size, vs = [100, 300, 500]
    - the learning rate, α = [0.0025, 0.025, 0.25]
    - epochs, e = [5, 10, 30]
    - window size, w = [3, 5]: The maximum distance between the current and predicted word within a sentence
    - sample, s = [1e-4, 1e-2, 0]: threshold for configuring which higher-frequency words are randomly down sampled
    - distributed memory, dm = [0, 1]
- **Task**: Initial Line of Treatment Prediction

Causal Inference



Potential Outcomes Framework

- For a set of *i.i.d.* subjects $i = 1, \ldots, n$ we observe:
  - $X_i \in \mathbb{R}^m$: covariates
  - $Y_i \in \mathbb{R}$: outcome
  - $W_i \in \{0, 1\}$: treatment assignment
- Estimate average treatment effect (ATE) for the treated

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0) \,|W_i = 1]$$

- Assumptions:
  - Unconfoundedness

$$W_i \perp (Y_i(0), Y_i(1))|X_i$$

  - Overlap

$$0 < pr(W_i = 1|X_i = x) < 1$$

Survival Outcomes Framework

- Kaplan-Meier survival plots are used to compare two treatments.
- **Hazard rate** $[h(t|X)]$: probability patient will fail in time interval t given covariates X.
- **Hazard ratio (HR)**: ratio of hazard rate between the two treatments
- **Cox-Proportional Hazard Regression Model:**
  - $b_0(t)$: baseline hazard function
  - b_j: hazard ratio of specific variable

$$\log h(t|X_i) = b_0(t)\exp\left(\sum_{j=1}^{m+1} b_j(X_{ij} - \bar{X}_i)\right)$$

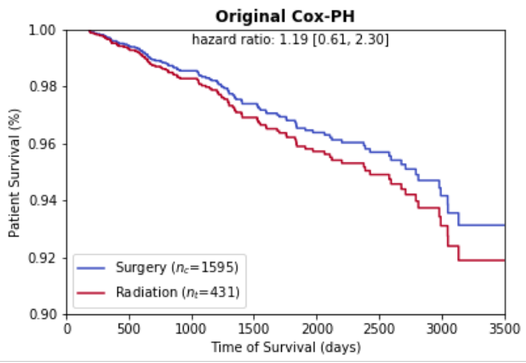
Case Study: Prostate Cancer[2]

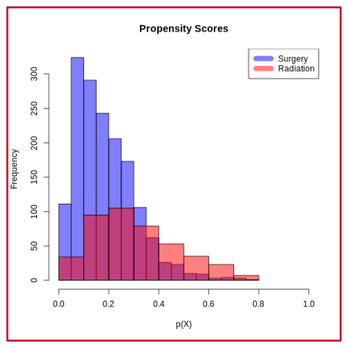
Treatments (W):
- Surgery ($W_i = 0$)
- Radiation ($W_i = 1$)

Outcome (Y): Days of Survival

Covariates (X):
- *Structured*: age, race, ethnicity, cancer stage
- *Unstructured*: embeddings of clinical notes

Total Patients: $n = 2026$
- $n_c = 1595, n_t = 431$





References:
[1] Payal D Soni, Holly E Hartman, Robert T Dess, Ahmed Abugharib, Steven G Allen, Felix Y Feng, Anthony L Zietman, Reshma Jagsi, Matthew J Schipper, and Daniel E Spratt. 2019. Comparison of Population-Based Observational Studies With Randomized Trials in Oncology. Journal of Clinical Oncology (2019), JCO– 780 18.
[2] Freddie C Hamdy, Jenny L Donovan, J Athene Lane, Malcolm Mason, Chris Metcalfe, Peter Holding, Michael Davis, Tim J Peters, Emma L Turner, Richard M Martin, et al. 2016. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. New England Journal of Medicine 375, 15 (2016), 1415–1424.
[3] Le Q, Mikolov T. Distributed representations of sentences and documents. InInternational conference on machine learning 2014 Jan 27 (pp. 1188-1196).
[4] Donald B Rubin.1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology 66, 5 (1974), 688.

Results

Natural Language Processing (NLP) Models

[See AAAI-20 Student Poster (Monday night)]

Causal Inference

Potential Outcomes Framework

Gold standard: ATE to be within the range of [-60, 60]. (Days of survival difference to be less than 2 months).
- Achieve ATE of 30 days or below from estimators.
- Standard errors are huge on many estimates.
- **Findings**: there is signal in using clinical text to inform ATE estimators.

Table 3: Best Performing Average Treatment Effect Estimators

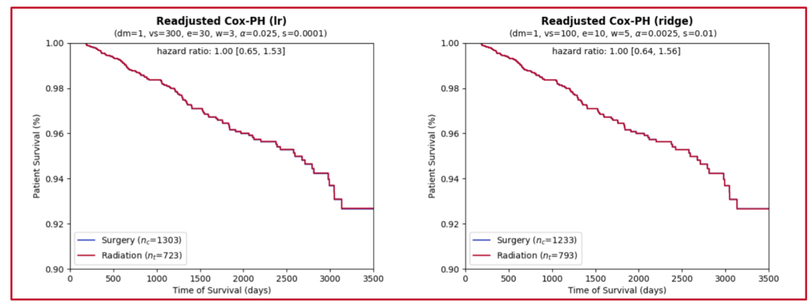
Method	Word Embedding Hyperparameters: *dm*	*vector size*	*α*	*window*	*epochs*	*sample*	Estimates: *est*	*(s.e.)*
Difference in Means	1	500	0.025	3	30	0	-203.29	50.27
Bias Corrected Matching	0	100	0.025	5	10	0	-18.40	24.36
Outcome Models								
Linear	0	100	0.025	3	5	0	34.51	44.85
Random Forest	0	100	0.025	3	10	0	14.55	44.44
Neural Net	1	100	0.0025	3	10	0.0001	4.12	52.15
Propensity Score Weighting								
Linear	1	100	0.0025	3	10	0.0001	-2.31	128.77
Random Forest	0	100	0.025	3	5	0.01	-0.14	98.88
Neural Net	1	500	0.0025	5	5	0.01	0.60	180.92
Doubly Robust Methods								
Linear	1	100	0.0025	3	10	0.0001	-5.77	68.46
Random Forest	0	300	0.025	3	10	0.0001	-8.22	49.32
Neural Net	0	300	0.025	3	10	0.01	-4.34	303.04
Generalized Random Forest	0	100	0.025	3	5	0	-1.83	55.59

Survival Outcomes Framework

Gold standard: Hazard ratio to be between [0.95, 1.05]. (Comparable hazard between surgery and radiation).
- "Readjusted" Cox-PH plots with Linear Regression and Ridge Regression achieved HR = 1.
- ***However***, prediction accuracy extremely low. → Shows that a random assignment of patients do just as well.

Table 5: Best Performing Survival Outcome Analysis

Method	Word Embedding Hyperparameters: *dm*	*vector size*	*α*	*window*	*epochs*	*sample*	Accuracy: *est*	F1-Score: *surg.*	*rad.*	Hazard Ratio: *mean*	*95% CI*
Linear Regression	1	300	0.025	3	30	0.0001	0.63	0.74	0.38	1.0	[0.65, 1.53]
Ridge Regression	1	100	0.0025	5	10	0.01	0.65	0.75	0.43	1.0	[0.64, 1.56]
Random Forest	1	300	0.0025	3	5	0.01	0.80	0.88	0.37	0.59	[0.35, 1.02]
Boosting	0	500	0.25	3	30	0.0001	0.79	0.87	0.42	0.61	[0.36, 1.03]



Summary

Significant Findings:
- Clinical notes can be employed to adjust for selection bias.

Challenge:
- Model selection for word embedding hyperparameters (No reliable without gold standard of RCT)

Next step:
- Explore if GANs can be used to augment the data for better adjustment of confounders.