Cancer Treatment Classification with Electronic Medical Health Records

Jiaming Zeng, Imon Banerjee, Michael Gensheimer, Daniel Rubin

Stanford University

Stanford Human-Centered Artificial Intelligence

Abstract

We built a natural language processing (NLP) language model that can be used to extract cancer treatment information using structured and unstructured electronic medical records (EMR). Our work appears to be the first that com- bines EMR and NLP for treatment identification.

Knowing the sequence of treatments administered to a cancer patient is important for personalized medicine and se- quential treatment planning. Our final goal is to leverage the full EMR, including the information available in the clinical notes, to build causal models for treatment effectiveness. For that purpose, we need a sufficiently large dataset with la- beled treatment information. However, cancer registries only record the initial line of treatment, even that requires hours of expensive manual labour.

We aim to build a NLP language model that can extract longitudinal treatment information using a combination of structured and unstructured EMR data. The extracted treatments can then be used for future analysis and treatment planning.

Some related works include [3] and [4].

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**: ICD9 procedure codes, medication names, count of different note types
- **Ground Truth**: 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

Table of Demographics

Characteristics		Prostate	Oropharynx	Esophagus
Gender	*male*	2,145	274	167
	female	0	46	58
Race	*white*	1,532	229	162
	black	92	12	3
	asian	190	17	18
	other	248	44	33
	unknown	83	18	9
Ethnicity	*hispanic*	148	18	18
	non-hispanic	1,183	284	196
	unknown	114	18	11
Age (years)	*≤ 25*	0	1	0
	25-50	63	41	10
	50-60	578	113	42
	60-70	1,030	119	81
	70-80	399	39	66
	80-90	74	3	23
	>90	1	4	3
Cancer Stage	*stage 1*	347	13	35
	stage 2	1,425	22	69
	stage 3	69	45	87
	stage 4	47	216	2
	unknown	247	24	32

Stanford Cancer Institute

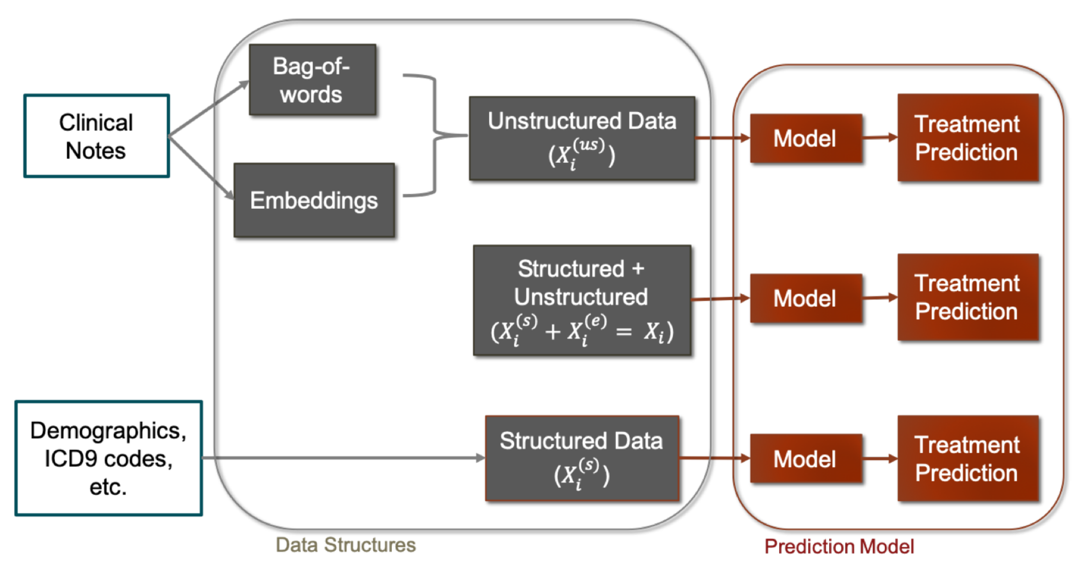
Stanford MEDICINE

Methodology

Natural Language Processing (NLP) Models

- **Notes**: 483,782 clinical notes (excluded 10% for testing)
- **Baseline:** Bag-of-words
- **Model**: Doc2vec[2]
  - Trained 324 doc2vec models for generating embeddings[1]
    - 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]

Treatment Prediction Models



Models

Model	Parameters
Logistic Regression (LR)	C = logspace(-4, 4, 20) solver = [newton-cg, lbfgs, saga, sag]
Ridge Regression (RR)	$\alpha = [1^{-15}, 1^{-10}, 1^{-8}, 1^{-5}, 1^{-4}, 1^{-3}, 1^{-2}, 1, 5, 10]$ (reduce some alphas)
Random Forest (RF)	n_estimators ∈[100, 500] max_features = [auto, sqrt] min_sample_split = [2, 5, 10] bootstrap = [True, False]
Stochastic Gradient Boosting (SGB)	max_depth = [3, 4, 5, 6, 7] learning_rate = [0.001, 0.05, 0.1] n_estimators ∈[100, 500] booster = [gbtree, gblinear, dart] gamma = [0, 1, 5, 10] subsample = [0.8, 1] colsample_bytree = [0.3, 0.8] reg_alpha = [0.0001, 0.001, 0.01, 0.1, 1, 10, 100] reg_lambda ∈[0.01, 1.0, 0.1]

Conclusions

Significant Findings/Contributions:

- Clinical notes can be very effective in performing treatment prediction.
- Concatenating structured and unstructured data allow us to benefit from both data formats.
- Building a set of institution specific doc2vec NLP language models.

Challenge:

- Missing data in EMR data. Missing icd9 codes in structured data really throws off analysis.

Next step:

- Extending the treatment prediction model to be part of a larger treatment decision analysis.
- Explore other ways of combining the structured and unstructured data.

Results

Predicted Treatment Classes

- For each cancer type, we combined treatments that were normally administered together with the guide of a clinician.
- We selected for patients with at least one note.

Cancer	Initial Line of Treatment	Entries	Total Notes	Mean notes
Prostate	Surgery	1642	30312	18.46
	Radiation (+hormone)	503	9371	18.63
	Hormone	61	942	15.44
Oropharynx	Chemo, Radiation	179	8827	49.31
	Surgery (+other)	141	5753	40.80
Esophagus	Surgery (+other)	150	16294	108.63
	Chemo, radiation	75	3156	42.08

Prostate Cancer

- Inclusion of notes information improves structured data performance
- For prostate, had to run two separate experiments. Will fix in later run.

Data Format	Methods	Overall	Hormone	Radiation(+hormone)	Surgery
Structured	Boosting	0.973	0.750	0.965	0.987
Bag-of-words	Linear Regression	0.968	0.750	0.965	0.981
Doc2vec	Random Forest	0.991	-	0.982	0.994
Structured+BOW	Boosting	0.982	0.875	0.982	0.987
Structured+doc2vec	**Random Forest**	**0.995**	**-**	**1.0**	**0.994**

Oropharynx Cancer

- Inclusion of notes definitely helped. However, just using notes seem to perform the best.
- Hypothesis: structured data has lots of missing information.

Model	Method	Overall	Surgery (+other)	Chemo, Radiation
Structured	Random Forest	0.750	0.667	0.909
Bag-of-words	Boosting	0.750	0.714	0.818
Doc2vec	**Random Forest**	**0.844**	**0.762**	**1.0**
Structured+BOW	Boosting	0.750	0.714	0.818
Structured+Doc2vec	Linear Regression	0.813	0.762	0.910

Esophagus Cancer

- Unstructured data outperforms structured data.
- Hypothesis: the treatment types are too similar. Hence, the structured data does not have enough information to distinguish them.

Model	Method	Overall	Surgery (+other)	Chemo, Radiation
Structured	Linear Regression	0.913	0.882	1.000
Bag-of-words	Boosting	0.957	0.941	1.000
Doc2vec	**Boosting**	**1.0**	**1.0**	**1.0**
Structured+BOW	Linear Regression	0.913	0.941	0.833
Structured+Doc2vec	**Ridge/LR/RF**	**1.0**	**1.0**	**1.0**

References

[1] Caselles-Dupré H, Lesaint F, Royo-Letelier J. Word2vec applied to recommendation: Hyperparameters matter. In Proceedings of the 12th ACM Conference on Recommender Systems 2018 Sep 27 (pp. 352-356). ACM.

[2] Le Q, Mikolov T. Distributed representations of sentences and documents. InInternational conference on machine learning 2014 Jan 27 (pp. 1188-1196).

[3] Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, Amin S, Liu H. A clinical text classification paradigm using weak supervision and deep representation. BMC medical informatics and decision making. 2019 Dec;19(1):1.

[4] Zhu H, Ni Y, Cai P, Qiu Z, Cao F. Automatic extracting of patient-related attributes: disease, age, gender and race. Studies in health technology and informatics. 2012;180:589-93.