Towards the Creation of a Large Corpus of Synthetically-Identified Clinical Notes

Willie Boag, Tristan Naumann, Peter Szolovits
Clinical Decision Making Group, MIT

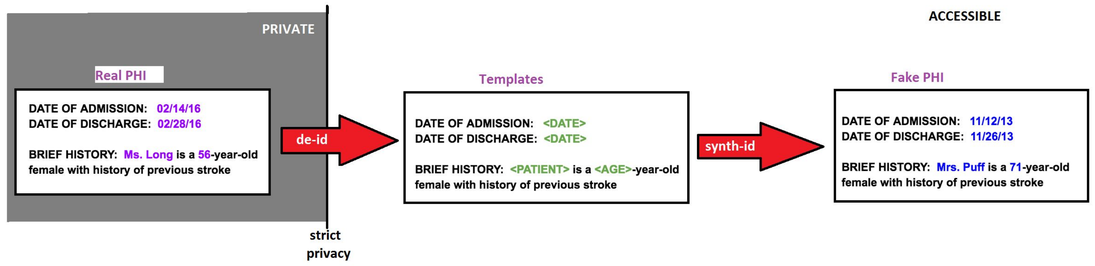
Clinical Notes

Clinical notes often describe the most important aspects of a patient's physiology and are therefore critical to medical research.

These notes are typically inaccessible to researchers without prior removal of sensitive protected health information (PHI), a natural language processing (NLP) task referred to as de-identification.

Tools to automatically de-identify clinical notes are needed, but are difficult to create without access to those very same notes containing PHI.

Synthetic Identification



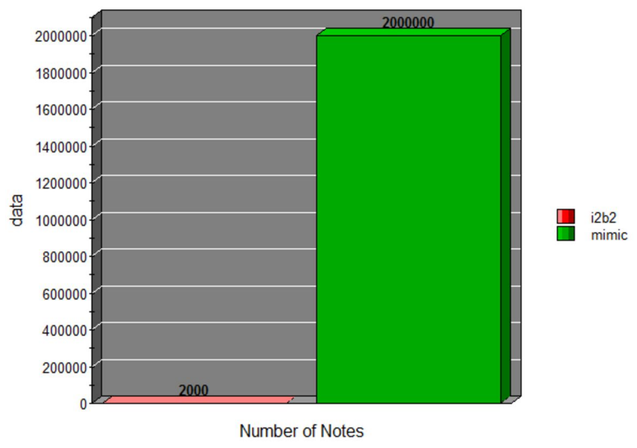
We are building a publically accessible De-Identification dataset from the MIMIC notes, which are easilly accessible.

Table 1: Distribution of de-identified protected health information (PHI) in MIMIC-III v1.4 notes by category. Particularly noteworthy are the ~2.5 million instances of PHI among ~500 million tokens.

Category	Notes	Contain PHI		Tokens	PHI Instances	
Case Management	967	954	(98.65%)	131806	9860	(7.48%)
Consult	98	98	(100%)	71453	1843	(2.58%)
Discharge summary	59652	59651	(99.99%)	80986971	2632527	(3.25%)
ECG	209051	133146	(63.69%)	5856486	135048	(2.31%)
Echo	45794	45794	(100%)	14817189	127233	(0.86%)
General	8301	5200	(62.64%)	1688905	36923	(2.19%)
Nursing	223556	188691	(84.40%)	56107626	1048996	(1.87%)
Nursing/other	822497	561187	(68.23%)	104063367	1718441	(1.65%)
Nutrition	9418	9196	(97.64%)	3068351	204730	(6.67%)
Pharmacy	103	96	(93.20%)	34466	1100	(3.19%)
Physician	141624	141047	(99.59%)	115484159	3475738	(3.01%)
Radiology	522279	522278	(99.99%)	102460089	3097379	(3.02%)
Rehab Services	5431	5010	(92.24%)	2125724	52955	(2.49%)
Respiratory	31739	10395	(32.75%)	4717416	14662	(0.31%)
Social Work	2670	2609	(97.72%)	779550	38691	(4.96%)
Total	2083180	1685352	(80.90%)	492393558	12596126	(2.56%)

The i2b2 2014 de-id challenge only has 2,000 notes of de-identified text. Such a small dataset is insufficient for building modern, deep models.

There are 2,000,000 notes in the MIMIC database.



De-Identification

Sensitive PHI must be removed from electronic healthcare records before they can be released for research.

DATE OF ADMISSION: 11/12/13
DATE OF DISCHARGE: 11/26/13

BRIEF HISTORY: Mrs. Puff is a 71-year-old female with history of previous stroke; renal carcinoma; presenting after a fall and possible syncope. While walking, she accidentally fell to her knees and did hit her head on the ground.

DATE OF ADMISSION: <DATE>
DATE OF DISCHARGE: <DATE>

BRIEF HISTORY: <PATIENT> is a <AGE>-year-old female with history of previous stroke; renal carcinoma; presenting after a fall and possible syncope. While walking, she accidentally fell to her knees and did hit her head on the ground.

Until recently, the state-of-the-art from the i2b2 shared task for de-id was a Conditional Random Field.

We trained a CRF baseline using: unigram features, context features, and comparisons to lists of common PHI words.

The system was trained to identify a variety of PHI tags, including: names, hospitals, companies, locations, dates, and identifying numbers (e.g. SSN).

Results

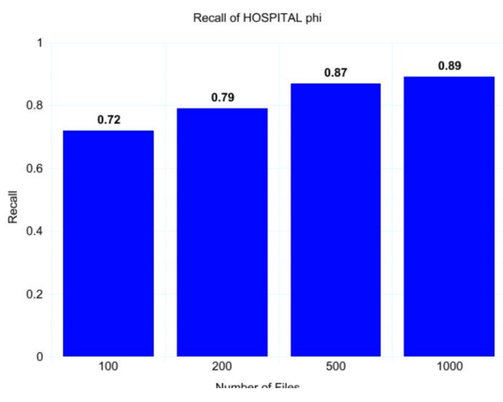
PATIENT_NAME			
# training files	precision	recall	f1
100	0.89	0.78	0.83
200	0.90	0.87	0.88
500	0.91	0.93	0.92
1000	0.92	0.96	0.94

HOSPITAL			
# training files	precision	recall	f1
100	0.92	0.72	0.81
200	0.93	0.79	0.86
500	0.94	0.87	0.91
1000	0.95	0.89	0.92

Unsurprisingly, the CRF continued to improve as data size grew.

Future Work:

- Currently training deep LSTM models on the silver MIMIC data
- Getting the up-to-date rule-based script used to originally de-id the MIMIC data.
- More sophisticated synth-id (e.g. preserving name form)



Acknowledgements

This research was funded in part by the Intel Science and Technology Center for Big Data, the National Library of Medicine Biomedical Informatics Research Training grant (NIH/NLM 2T15 LM007092-22), the MIT-MGH Challenge grant and NIH grant 1R01MH106577. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1122374. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.