UNITED STATES

DEPARTMENT OF VETERANS AFFAIRS

Contact: Kelly Peterson
University of Utah, Salt Lake City UT
kelly.peterson@hsc.utah.edu

Can Electronic Clinical Notes Identify Travelers with Zika?

Kelly Peterson, MS[1,2], Daniel Denhalter, MSPH[1,2], Olga V. Patterson, PhD[1,2], Makoto Jones, MD, MS[1,2]
[1] VA Salt Lake City Healthcare System, [2]Division of Epidemiology, University of Utah

powered by: python, learn, jupyter Notebook

VINCI VA Informatics and Computing Infrastructure

IDEAS VA Salt Lake Informatics, Decision-Enhancement, and Analytic Sciences Center

BACKGROUND / OBJECTIVES

Travel history can help differentiate a public health emergency from a travel-related infection by providing information on exposure. However such information is often available only in unstructured clinical documents.[1] We are not aware of existing work has reported on feasibility of automated extraction of travel history, likely due to a need for annotated data and a process for selecting data.

We aimed to assess feasibility of extracting past travel history mentions from the electronic health record in an automated fashion by first annotating a dataset and then developing a machine learning model to extract travel history locations from clinical documents.

METHODS

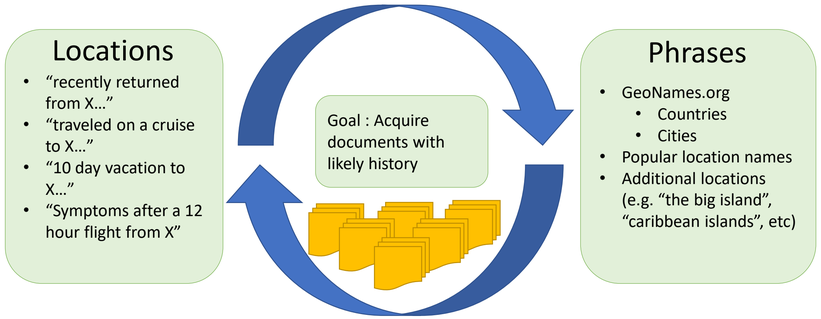
In collaboration with the National Biosurveillance Integration Center (NBIC), clinical notes were extracted from patient records for encounters with Zika, dengue and chikungunya virus testing in the Department of Veterans Affairs.

- Extracted 4,584 snippets from a set of >250k using a semi-automated bootstrapping process to identify documents containing potentially relevant information using locations and phrases (see right).[2]
- Manually annotated snippets for travel affirmation and locations visited including negation states (i.e. "pt denies travel to Puerto Rico")
- However, time period of the travel (if stated) was not annotated.
- Trained a Conditional Random Field (CRF) model to extract affirmed travel locations outside of the continental US.

Pt recently returned from a trip to Germany and Italy.	1 Evidence 2 Locations Status: Affirmed
Travel to Miami? [] Y [x] N	1 Evidence 1 Locations Status: Negated
Pt plans to travel to Japan in the Fall.	No annotation

ANNOTATION RESULTS

Bootstrapping selection of annotation corpus



Annotation agreement results

Type	Agree	Disagree	Missing	% Agree	Kappa
Text Span	769	24	75	89%	0.651

Annotation results

Unique Snippets	Travel Evidence (Any)	Evidence (Positive)	Evidence (Negated)
4584	3006 (65.6%)	2659 (58%)	347 (7.6%)

Most frequent annotated locations

Positive Locations	Negated Locations
Iraq	Liberia
Mexico	Guinea
Dominican Republic	Sierra Leone
Costa Rica	Democratic Republic of Congo
Vietnam	West Africa

MODELING RESULTS

Model contextual features

Feature Types	Examples (i – 2, i – 1, i)
Part of Speech	Verb, Preposition, Noun,…
Terms	"returned", "from", "Uganda", …
Character n-grams	"ret", "ned", "fro", "rom", "uga", "nda", …
Lexical form (capitalization, digit, punctuation)	"Capital=False", "Capital=False", "Capital=True", …
Term stems	"return", "from", "uganda", …
Word Embedding Clusters[3]	"213", "10", "320", …
Word Embedding "Compound Clusters"[3]	"213_10", "10_320", …
Gazetteer Match (geonames.org lexicon)	False, False, True, …

Model performance on 356 held out test documents

Positive Predictive Value	Sensitivity	F1
85.6	76.7	80.9

Most frequent model extractions from over 250k notes

Extracted Location	Count
Vietnam	571
Iraq	569
Mexico	392
Costa Rica	301
Dominican Republic	259
Afghanistan	254
Jamaica	160
Honduras	134

CONCLUSION / DISCUSSION

- Targeted travel history extraction is feasible in a large medical system with acceptable accuracy
- Approach capable of extracting novel locations that would not necessarily be found in a curated list (e.g. Mexican Riviera, Baja Cruise)
- Further research could incorporate automated extractions into models improving the early detection of autochthonous transmission
- Approach may also be beneficial to other biosurveillance (e.g. screening for CRE)
- Automated extraction now deployed in operations for continued validation

REFERENCES

1. Chapman WW, Gundlapalli AV, South BR, Dowling JN. Natural language processing for biosurveillance. InInfectious Disease Informatics and Biosurveillance 2011 (pp. 279-310). Springer, Boston, MA.
2. Alba PR, Patterson OV, Viernes B, Denhalter DW, Bailey N, Wilson A, Kamauu AW, DuVall SL. The Super Annotator: A Method of Semi-Automated Rare Event Identification for Large Clinical Data Sets. In AMIA 2016.
3. Guo J, Che W, Wang H, Liu T. Revisiting embedding features for simple semi-supervised learning. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014 (pp. 110-120).

ACKNOWLEDGEMENTS

- Project funded by Department of Homeland Security, National Biosurveillance Integration Center
- This project was supported with facilities at the VA SLC IDEAS Center Salt Lake City, Utah and data resources from VA Informatics and Computing
- The views expressed are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Presented at the 2018 ID Week Annual Conference