Natural Language Processing for Health (HLP)
September 2018
Tweet @UPennHLP #HLPMeeting

Graciela Gonzalez-Hernandez
Contact: gragon@upenn.edu
Twitter: @gracielagon

https://healthlanguageprocessing.org
Program & Speaker List

♦ Welcome/Introduction – Graciela Gonzalez-Hernandez, Ph.D.

♦ KEYNOTES (15 minutes + 15 minutes for questions each)

Deep neural networks and distant supervision for geographic location mention extraction
Arjun Magee
Dept. of Biomedical Informatics
Arizona State University

Social Media Mining for Pharmacovigilance: challenges and opportunities: Case Control Studies from Twitter?
Graciela Gonzalez-Hernandez, Ph.D
Health Language Processing Lab – Penn IBI
University of Pennsylvania
Deep neural networks and distant supervision for geographic location mention extraction

Arjun Magge$^{1,2}$, Davy Weissenbacher $^3$, Abeed Sarkar $^3$, Matthew Scotch $^{1,2}$, and Graciela Gonzalez $^3$

$^1$ Department of Biomedical Informatics, Arizona State University
$^2$ Biodesign Center for Environmental Health Engineering, Biodesign Institute, Arizona State University
$^3$ Department of Biostatistics, Epidemiology and Informatics, The Perelman School of Medicine, University of Pennsylvania

Sept 6, 2018
Phylogenetic tree and spread reconstruction

- Virus phylogeography and epidemiology research relies on nucleotide sequence repositories like **GenBank**
GenBank Overview

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research. 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample GenBank record for a Saccharomyces cerevisiae gene demonstrates many of the features of the GenBank flat file format.
Zika virus isolate Brazil-ZKV2015, complete genome

GenBank: KU497555.1

FASTA  Graphics

LOCUS     KU497555   10793 bp  RNA  linear  VRL 18-FEB-2016
DEFINITION Zika virus isolate Brazil-ZKV2015, complete genome.
ACCESSION KU497555
VERSION   KU497555.1
KEYWORDS  .
SOURCE     Zika virus
ORGANISM  Virus; ssRNA viruses; ssRNA positive-strand viruses, no DNA
  stage; Flaviviridae; Flavivirus.
REFERENCE  1 (bases 1 to 10793)
  AUTHORS  Calvet, G., Aguiar, R.S., Melo, A.S., Sampaio, S.A., de Filippis, I.,
           Fabri, A., Araujo, E.S., de Sequeira, P.C., de Mendonca, M.C., de
           Oliveira, L., Tschoeke, D.A., Schrago, C.G., Thompson, F.L., Brasil, P.,
           Dos Santos, F.B., Nogueira, R.M., Tanuri, A. and de Filippis, A.M.
  TITLE    Detection and sequencing of Zika virus from amniotic fluid of
           fetuses with microcephaly in Brazil: a case study
  JOURNAL  Lancet Infect Dis 16 (6), 653-660 (2016)
  PUBMED   26897108
REFERENCE  2 (bases 1 to 10793)
  AUTHORS  Tanuri, A., Bispo, A., Thompson, F., Santana, R., Tschoeke, D., de
           Oliveira, L. and Guerra, C.
  TITLE    Direct Submission
  JOURNAL  Submitted (06-JAN-2016) UFRJ, UFRJ, Avenida Carlos Chagas Filho,
           373, Rio de Janeiro, Rio de Janeiro 21040-900, Brazil
  COMMENT  ##Assembly-Data-START##

Location/Qualifiers
  1..10793
  /organism="Zika virus"
  /mol_type="genomic RNA"
  /isolate="Brazil-ZKV2015"
  /isolation_source="amniotic liquid"
  /host="Homo sapiens"
  /db_xref="taxon:64320"
  /country="Brazil"
  /collection_date="30-Nov-2015"
  /101..10372
  /codon_start=1
  /product="polyprotein"
  /protein_id="AMD16557.1"
  /translation="MKNPKKKKGGRIVNMLKRGVARVS"
<table>
<thead>
<tr>
<th>LOCUS</th>
<th>KU497555</th>
<th>10793 bp</th>
<th>RNA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Zika virus isolate Brazil-ZKV2015, complete genome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACCESSION</td>
<td>KU497555</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VERSION</td>
<td>KU497555.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>Zika virus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Zika virus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFERENCES</td>
<td>1 (bases 1 to 10793)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td>Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOURNAL</td>
<td>Lancet Infect Dis 16 (6), 653-660 (2016)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PUBMED</td>
<td>26897108</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

/organism="Zika virus"
/host="Homo sapiens"
/collection_date="30-Nov-2015"
/country="Brazil"
Detection and state of Paraíba in Brazil: a case study.

Calvet G1, Aguilar RS2, Schrager CG2, Thompson IC2

Abstract

BACKGROUND: The incidence of microcephaly in Brazil in 2015 was associated with genetic factors and several causative agents. Epidemiological factors were associated with the introduction of Zika virus. We aimed to detect the state of microcephaly in pregnant women in Brazil whose fetuses were diagnosed with microcephaly and associated with infection with Zika virus.

METHODS: In this case study, amniotic fluid samples from two pregnant women diagnosed with microcephaly were obtained, on the recommendation of amniocentesis at 28 weeks' gestation. The women had presented clinical manifestations that could have been symptoms of Zika virus infection. The samples were centrifuged, DNA and RNA were extracted from the purified tissue, reverse transcription PCR and viral metagenomic next-generation sequencing were performed. Recombination events were done by comparing the Brazilian Zika virus strain and isolates from similar regions in Brazil.
Problem: Locations in GenBank metadata are not sufficient for Phylogeography research

- Especially for countries like USA, Canada, Russia, China, Brazil

Solution: Enrich location information in GenBank by extracting locations from the associated PubMed article using Natural Language Processing (NLP)
1. Named Entity Recognition
   - identifying words of interest in text (usually nouns)
   - e.g. names, genes, proteins, locations, organizations, time, etc.

1. Concept Resolution
   - perform disambiguation by assigning a unique gazetteer ID
   - e.g. Paris can refer to Paris, Texas, USA or Paris, France

1. Determine Location of Infected Host (LOIH)
   - assign probabilities to all identified locations using heuristics
1. **Named Entity Recognition**
   - identifying words of interest in text (usually nouns)
   - e.g. names, genes, proteins, locations, organizations, time, etc.

1. **Concept Resolution**
   - perform disambiguation by assigning a unique gazetteer ID
   - e.g. Paris can refer to Paris, Texas, USA or Paris, France

1. **Determine Location of Infected Host (LOIH)**
   - assign probabilities to all identified locations using heuristics
Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study.

Calvet G1, Aguiar RS2, Melo ASO3, Sampaio SA4, de Filippis I5, Fabri A4, Araujo ESM4, de Sequeira PC4, de Mendonça MCL4, de Oliveira L2, Tschoeke DA6, Schrago CG7, Thompson FL7, Brasil P1, Dos Santos FB4, Nogueira RMR4, Tanuri A2, de Filippis AMB8.

Abstract

BACKGROUND: The incidence of microcephaly in Brazil in 2015 was 20 times higher than in previous years. Congenital microcephaly is associated with several congenital causal agents. Epidemiological data suggest that microcephaly cases in Brazil might be associated with the Zika virus. We aimed to detect and sequence the Zika virus genome in amniotic fluid samples of two pregnant women who were diagnosed with microcephaly.

METHODS: Amniotic fluid samples from two pregnant women from the state of Paraíba in Brazil, whose fetuses had been diagnosed, on the recommendation of the Brazilian health authorities, by ultrasound-guided transabdominal amniocentesis at 18 weeks' and 10 weeks' gestation, respectively, with clinical manifestations of Zika virus infection, including fever, myalgia, and rash. After the amniotic fluid samples were centrifuged, the purified virus particles before the viral genome was identified by quantitative reverse transcription PCR and viral metagenomic next-generation sequencing. Phylogenetic reconstruction and investigation of recombination events were done by comparing the Brazilian Zika virus genome with sequences from other Zika strains and from flaviviruses that occur in similar regions in Brazil.
1. Named Entity Recognition
   - identifying words of interest in text (usually nouns)
   - e.g. names, genes, proteins, locations, organizations, time, etc.

1. Concept Resolution
   - perform disambiguation by assigning a unique gazetteer ID
   - e.g. Paris can refer to Paris, Texas, USA or Paris, France

1. Determine Location of Infected Host (LOIH)
   - assign probabilities to all identified locations using heuristics
Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study.

Calvet G1, Aguiar RS2, Melo ASO3, Sampaio SA4, de Filippis I5, Fabri A4, Araujo ESM4, de Sequeira PC4, de Mendonça MCL4, de Oliveira L2, Tschoeke DA6, Schraro CG2, Thompson FL7, Brasil P1, Dos Santos FB4, Nogueira RMR4, Tanuri A2, de Filippis AMB8.

Abstract

BACKGROUND: The incidence of microcephaly in Brazil in 2015 was 20 times higher than in previous years. Congenital microcephaly is associated with the Zika virus. These data raise the hypothesis that microcephaly cases in Brazil might be associated with the Zika virus genome.

METHODS: The infection was diagnosed by ultrasound-guided transabdominal amniocentesis. Classic manifestations were central nervous system involvement and absence of calcification. Phylogenetic reconstruction was performed using DNA sequences recovered from the virus genome identified by quantitative reverse transcription PCR and viral metagenomic next-generation sequencing. Phylogenetic reconstruction and investigation of recombination events were done by comparing the Brazilian Zika virus genome with sequences from other Zika strains and from flaviviruses that occur in similar regions in Brazil.
1. Named Entity Recognition
   – identifying words of interest in text (usually nouns)
   – e.g. names, genes, proteins, locations, organizations, time, etc.

1. Concept Resolution
   – perform disambiguation by assigning a unique gazetteer ID
   – e.g. Paris can refer to Paris, Texas, USA or Paris, France

1. Determine Location of Infected Host (LOIH)
   – assign probabilities to all identified locations using heuristics
Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil, a case study.

Calvet G1, Aguilar RS2, Melo ASO3, Sampaio SA4, de Filippis J5, Fabri A4, Araujo ESM4, de Sequeira PC4, de Mendonça MCL4, de Oliveira L2, Tschoeke DA6, Schrago CG7, Thompson FL7, Brasil P1, Dos Santos FB4, Nogueira RMR4, Tanuri A2, de Filippis AMB8.

+ Author information

Abstract

BACKGROUND: The incidence of microcephaly in Brazil in 2015 was 20 times higher than in previous years. Congenital microcephaly is associated with microcephaly, associated with microcephaly, and pregnant women...

METHODS: The diagnosis of microcephaly was confirmed by reverse transcription PCR and viral metagenomic next-generation sequencing. Phylogenetic reconstruction and investigation of recombination events were done by comparing the Brazilian Zika virus genome with sequences from other Zika strains and from flaviviruses that occur in similar regions in Brazil.

Locations: Brazil, Paraiba

GeonamesID: 3469034, 3393098

LOIH Probability: 0.00, 1.00
Natural Language Processing

1. Named Entity Recognition
   - identifying words of interest in text (usually nouns)
   - e.g. names, genes, proteins, locations, organizations, time, etc.

1. Concept Resolution
   - perform disambiguation by assigning a unique gazetteer ID
   - e.g. Paris can refer to Paris, Texas, USA or Paris, France

1. Determine Location of Infected Host (LOIH)
   - assign probabilities to all identified locations using heuristics
Ambiguity in Natural Language

<table>
<thead>
<tr>
<th>Description</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>in <strong>May, Russia</strong> in 2010.</td>
<td>✔️</td>
</tr>
<tr>
<td>found in <strong>May</strong> 2013.</td>
<td>✗</td>
</tr>
<tr>
<td>pigs, <strong>turkey</strong> and quail</td>
<td>✗</td>
</tr>
<tr>
<td>University of <strong>Las Vegas</strong>.</td>
<td>✗</td>
</tr>
</tbody>
</table>
Why deep neural nets?

- Rule based systems
- Machine Learning and Deep learning
  - Better performance with more annotated data
- Most times, you can only annotate a few articles.
  - So, distant supervision?
Dataset

- A set of 60 full-text articles (~300,000 words) from Pubmed containing 1881 location annotations
  - 48 for training and 12 for testing

- Distant supervision
  - Use GenBank articles where locations are known
  - Generate positive and negative examples based on rules
  - They are noisy! But, that’s okay.
  - We use them to generate ~8 million training instances (words)
Collecting Distant Supervision Samples

Process records which have fine-grained locations

/country="USA: San Felasco State Park, Alachua, Florida"
/PUBMED 25536075
Isolation of Tacaribe Virus, a Caribbean Arenavirus, from Host-Seeking Amblyomma americanum Ticks in Florida. We report the re-isolation of the virus from a pool of 100 host-seeking Amblyomma americanum (lone star ticks) collected in a Florida state park in 2012.

At least ten of these viruses are associated with human disease in many parts of the world including western Africa, Argentina, Bolivia, Venezuela and Brazil [2].

All tick trapping was performed in accordance with the Florida Department of Environmental Protection Research and Collection Permit #05231210.

Only host-seeking tick species common in Florida were collected because these species are most likely to attach to a person and take a blood meal.

In 2013, ticks were collected using the same methods as 2012 from two additional Florida state parks: Manatee Springs State Park in Chiefland, Florida (29° 29’47.401” N, 82°58’4.429” W) and O’Leno State Park in High Springs, Florida (29°55’11.863” N and 82° 35’15.427” W), to determine if the virus could be detected in other locations in North Central Florida (Fig. 1).

A total of 500 host-seeking ticks were collected from three state parks located in North Central Florida, including the original field site where ticks were trapped for virus isolation attempts (Fig. 1).
Collecting Negative Samples

PubMed articles with sufficient location information

- Mentions of locations available in GenBank
  - Positive Examples
- Sentences without potential locations
- Sentences with potential locations
  - Contains collocated words?
    - Yes
      - Compile words collocated with locations
      - Negative Examples
    - No
      - Trash

Annotated PubMed Articles
Filter them based on some guidelines

**Positive Examples**
- Ticks in Florida in a Florida state park with the Florida Department of
- common in Florida were collected two additional Florida state parks
- Chiefland, Florida (29
- Springs, Florida (29
- North Central Florida (Fig
- North Central Florida, including University of Florida Interdisciplinary Center in Central Florida, USA

**Negative Examples**
- <PAD> <PAD> Gene UL111A encodes
- <PAD> Gene UL111A encodes viral
- Gene UL111A encodes viral interleukin
- UL111A encodes viral interleukin - encodes viral interleukin - 10
- viral interleukin - 10 (interleukin - 10 (Lockridge et al
- - 10 (Lockridge et al
- - 10 (Lockridge et al
- (Lockridge et al.)
- Lockridge et al.
Layered training

Distant Supervision

- GenBank Metadata
- PubMed Central Article
- Distant Supervision Examples
- Random Model Weights
- Initialization
- DistSup-NER Training Phase-1
  - Pre-trained Word Embeddings
  - Feature Embeddings
- DistSup-Trained Model Weights

Supervision

- Annotated PubMed Articles
- NER Training Phase-2
  - Pre-trained Word Embeddings
  - Feature Embeddings
- NER-Trained Model Weights

Production

- PubMed Article
- Waco, TX, Dallas, ...
- List of Toponyms
## Improved performance

<table>
<thead>
<tr>
<th>Implementation</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge-based</td>
<td>0.58</td>
<td>0.88</td>
<td>0.70</td>
</tr>
<tr>
<td>CRF-All</td>
<td>0.85</td>
<td>0.76</td>
<td>0.80</td>
</tr>
<tr>
<td>Stanford-NER</td>
<td>0.89</td>
<td>0.85</td>
<td>0.872</td>
</tr>
</tbody>
</table>

(Weissenbacher et. al. 2015)

(Weissenbacher et. al. 2017)

- **92.7%** on tokenwise evaluation
- **91.5%** on strict evaluation
Limitations and Future Work

• Potential for improving performance
  – Deal with table data
  – Second layer of supervision for trying advanced recurrent models like Bi-LSTM-CRFs
• What is the improvement to resolution/normalization?
• Any improvements to the phylogeographic models?
• Distant supervision and Supervision - a systematic analysis of how much data for both is sufficient
• Validate with other entities like hosts, virus, genes etc.
Tell your peers:
Even though the field says country
please add in addition to country information:
• state
• county (if available)
• city (if available)
Softwares and Applications

NER (source code) : https://github.com/amagge/ner-topo-ff

GeoBoost v1 : https://tinyurl.com/geoboost (Tahsin et. al. 2017)

ZoDo (under development): https://zodo.asu.edu/zodo

ZooPhy (under development): https://zodo.asu.edu/zoophy
Acknowledgments

For the travel grant to present at ISMB

Bio-Ontologies
Funding and Acknowledgments

Principal Investigators:
Dr. Matthew Scotch
Dr. Graciela Gonzalez

Collaborators:
Dr. Davy Weissenbacher
Dr. Abeed Sarker

Annotators:
Karen O’Connor
Megan Rorison
Briana Trevino
Thank you!

Questions?
Social Media Mining for Pharmacovigilance: challenges and opportunities

Case-control studies from Twitter???

Health Language Processing Lab – Penn IBI
Graciela Gonzalez-Hernandez, PhD
email: gragon@pennmedicine.upenn.edu

Funded by NLM/NIH Grant R01LM011176
SM data for pharmacovigilance studies

- There are about 38,220 tweets / minute about the user’s current medical conditions\(^1,2,3\)
- Patient reporting brings different perspective, more detail, info on severity and impact of ADRs in daily life. (34 studies - PMID 27558545).
- Abundant adverse event reports in SM, with a higher frequency of adverse events, particularly for ‘mild’ adverse events. (51 studies = PMID 26271492).

\(^1\)http://www.pewinternet.org/fact-sheets/health-fact-sheet/
\(^3\)http://www.internetlivestats.com/twitter-statistics/
Work during first funding cycle

- Our prior work addressed the challenges of automatically collecting and processing SM reports on medication side effects.
- It resulted in over 16 publications, numerous annotated datasets, and novel automatic language processing (NLP) methods for side effect mention extraction and normalization to a standardized vocabulary (the UMLS/MedDRA).
Overview

- Develop novel NLP methods to leverage SM data for specific pharmacovigilance efforts that are hindered by known drawbacks of SRSs.
- We focus on methods to facilitate the use of SM data for exploring
  - (a) factors affecting medication adherence and persistence among the general population (Aim 1), and
  - (b) possible associations between medications taken during pregnancy and pregnancy outcomes (Aim 2).
- These are areas of significant impact for which SM data could meaningfully complement current PV efforts.
Social Media Mining pipeline

Data collection → Annotation → Classification

Concept Mapping → Concept extraction → Analysis

Legend:
- Amazon
- Daily Strength

Perelman School of Medicine
University of Pennsylvania
The Aims

- Develop and evaluate NLP methods to identify non-adherence and non-persistence and related information from Twitter data.

- Develop and evaluate NLP methods to identify medication use during pregnancy and pregnancy outcomes from Twitter data.

- Develop and evaluate methods for automatic selection of control groups to address the challenge faced when information from SM is to be used for epidemiological studies.
Aim 1

- Develop and evaluate NLP methods to identify non-adherence and non-persistence and related information from Twitter data. The methods will
  - dynamically collect a cohort of SM users that stopped taking or switched medications, did not fill a prescription, or altered their treatment,
  - extract information from the user’s timeline (publicly available postings over time) and conversation threads (postings by the user and others in reply to a posting of interest) relevant to
    - (a) an expressed reason for these actions,
    - (b) dosage/duration of treatment,
    - (c) concomitant treatments, and
    - (d) diagnosed health conditions.
Adherence/persistence studies from SM

- Social media may be particularly useful for identifying sources of intolerability that lead to non-adherence/non-persistence
- These are often not reported by physicians or patients through standard means because are considered “mild”, “not serious” or are unexpected
- Significant problem, given that, on average:
  - 30% of treated patients have a beneficial response
  - 30% do not respond
  - 10% have only side effects
  - 35%-70% are non-adherent / non-persistent, often due to side-effects or perceived/real non-response
6-month persistence rate

- prostaglandin analogs 47%
- statins 56%
- bisphosphonates 56%
- oral antidiabetics 66%
- angiotensin II receptor blocker 63%
- overactive bladder medications 28%
“I stopped taking” & “made me”

If anyone’s wondering which I doubt, the reason I stopped taking my antidepressants was because it messed with my appetite and made me feel extra drowsy and just emotionally numb. I constantly felt like a zombie, so I figured I see how I felt without them.

And well without them while in a better state of mind. I’ve gone without them before but those were my darker days aka like 2 weeks ago lol

I used to be on it. It made me nauseous, gave me headaches, made me bloated & made my period pains worse. I stopped taking it.

It’s a cheaper version of Triphasil. After that, I promised not to use cheap contraceptives again. The expensive ones have less side effects
Aim 2

- Develop and evaluate NLP methods to identify medication use during pregnancy and pregnancy outcomes from Twitter data.

  - Development and evaluation of NLP methods to dynamically collect a cohort of SM users who report a pregnancy, and
  - Methods to extract information from the user’s timeline to
    - (a) distinguish when mention of a medication indicates possible intake of it,
    - (b) infer the estimated pregnancy timeframe (beginning and end of pregnancy), and
    - (c) extract or infer pregnancy outcomes from those postings (including at least live birth, fetal death, hemorrhage, miscarriage, low-birth weight, pre-term birth, and reported congenital malformations).
Case-control study with SM data?

- Select cohort of pregnant women from SM\(^1\)
  - About 120 thousand, 700 million tweets
- Within that, find cases of interest
  - “Women who gave birth to a child with a birth defect and whose public tweets include tweets during pregnancy”
- Annotate (100\% of the data found)
- Find matching (control) subjects
  - “Women pregnant around the same time, for whom there is no evidence that their child was born with a birth defect”

From Twitter, “I am 12 weeks pregnant”
From Twitter, noise

my son is 15 months and my wife is 12 weeks pregnant. when I am home it's funny dealing with his high energy and tantrums

'I am 12 Weeks Pregnant!:,' Janet Mbugua Reveals She Is Expecting Baby Number Two  classic105.com/i-am-12-weeks-...
Finding cases – birth defects cohort

Klein et al, 2018 (in preparation)
# Birth defects data from Social Media

<table>
<thead>
<tr>
<th>Age</th>
<th>Cases (n=197)</th>
<th>Controls (n=196)</th>
<th>OR or t-test [95% CI]</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median Age (IQR)</td>
<td>23 (20 to 28)</td>
<td>21 (19 to 23)</td>
<td>2 (1 to 3)</td>
<td>0.0001</td>
</tr>
<tr>
<td>Mean Age (range)</td>
<td>25 (17 to 42)</td>
<td>22 (16 to 37)</td>
<td>2.52 (1.38 to 3.66)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Women &lt;30 years</td>
<td>80% (134/168)</td>
<td>91% (129/141)</td>
<td>0.37 (0.17 to 0.77)</td>
<td>0.004</td>
</tr>
<tr>
<td>Women &lt;35 years</td>
<td>93% (156/168)</td>
<td>98% (138/141)</td>
<td>0.28 (0.05 to 1.08)</td>
<td>0.04</td>
</tr>
<tr>
<td>Missing data on age</td>
<td>14% (28/196)</td>
<td>28% (55/196)</td>
<td>0.43 (0.25 to 0.73)</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th></th>
<th></th>
<th>chi² = 23.69, d.f. = 5, P &lt; 0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasian</td>
<td>73% (120/164)</td>
<td>55% (102/184)</td>
<td>2.19 (1.36 to 3.54)</td>
</tr>
<tr>
<td>Black</td>
<td>13% (22/164)</td>
<td>27% (51/184)</td>
<td>0.40 (0.22 to 0.72)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>9% (14/164)</td>
<td>12% (21/184)</td>
<td>0.72 (0.33 to 1.56)</td>
</tr>
<tr>
<td>Asian</td>
<td>2% (4/164)</td>
<td>3% (5/184)</td>
<td>0.90 (0.17 to 4.24)</td>
</tr>
<tr>
<td>Other (Islander, Native American/Indian, Multiracial/Mixed)</td>
<td>2% (4/164)</td>
<td>2% (5/184)</td>
<td>0.90 (0.17 to 4.24)</td>
</tr>
<tr>
<td>Missing data on race</td>
<td>16% (32/196)</td>
<td>6% (12/196)</td>
<td>0.99 (1.44 to 6.58)</td>
</tr>
</tbody>
</table>

Klein et al, 2018 (in preparation)
Thank you!

gragon@pennmedicine.upenn.edu
Twitter: @gracielagon
HLP lab (datasets and software available):
https://healthlanguageprocessing.org