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

@gracielagon

Funded by NLM/NIH Grant 5R01LM011176

ASCPT 2018 #Science Session – Orlando, Florida

Perelman School of Medicine
UNIVERSITY of PENNSYLVANIA

Social media as an “online health report”?

- **26% of internet users actively discuss health information. Of that group …[1]**
  - 30% *changed behavior* as a result
  - 42% discussed *current* medical conditions
- **“Extrapolating” this to Twitter…[2,3]**
  - Given 317 million active monthly users (Q3 2015): about 24 million would change their health behavior
  - Given 350,000 tweets/minute: about 38,220 tweets / minute about their current medical conditions

[1]http://www.pewinternet.org/fact-sheets/health-fact-sheet/
[2]http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
[3]http://www.internetlive**stats**.com/twitter-**statistics**/

SM data for pharmacovigilance studies

- **Patient reporting brings different perspective** (34 studies): more detail, info on severity and impact of ADRs in daily life. (PMID 27558545).
- **Abundant adverse event reports in SM** (29 studies): a higher frequency of adverse events are found in SM, particularly for ‘mild’ adverse events. (PMID 26271492).

Perelman School of Medicine
University of Pennsylvania

Social Media Mining pipeline

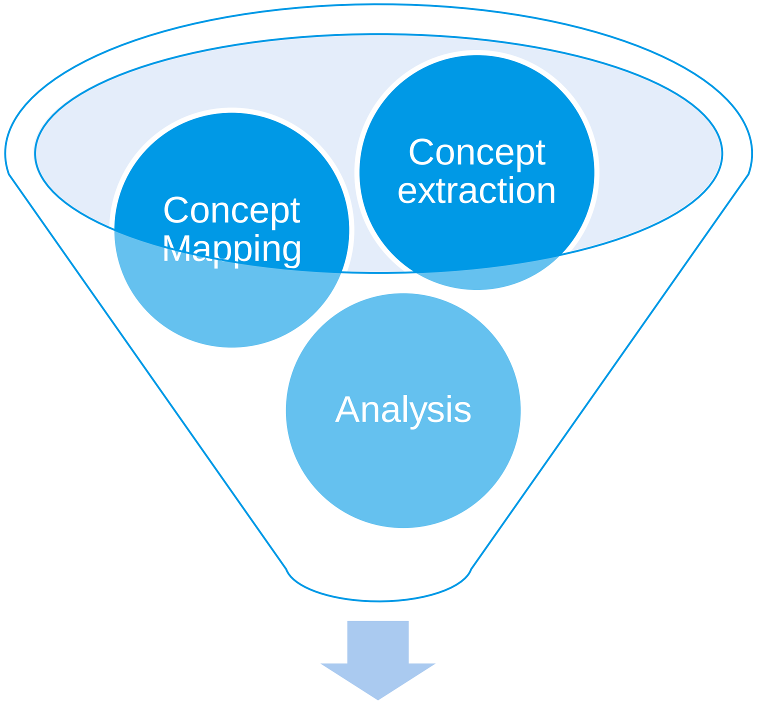
amazon

DailyStrength

Data collection

Annotation

Classification



Adverse event annotation example

Indication: mania (C0338831)

Indication: depression (C001157)

ADR: drowsiness (C0013144)

Works to calm **mania** or **depression** but **zonks** me and scares me about **diabetes** issues reported.

Other: diabetes

Indication: crying (C0010399)

ADR: emotional indifference (C0001726)

stops me from **crying** most of the time, **blocks most of my feelings**

Perelman School of Medicine
University *of* Pennsylvania

Concept extraction: finding adverse events

- Given a set of postings, need to find a specific concepts automatically
- An algorithm is trained to find (extract) mentions of AEs and other concepts
- Our approach: conditional random fields and deep learning clusters to capture descriptive expressions beyond what lexicon-based approaches could[1]

b) *I felt awful, it made my* ***stomach hurt***$_{ADR}$ *with bad* ***heartburn***$_{ADR}$ *too,* ***horrid taste in my mouth***$_{ADR}$ *tho it does tend to clear up the* ***infection***$_{Indication}$.

1 Nikfarjam *et al.* Pharmacovigilance from social media.. cluster features. **JAMIA**. 2015.

Unsupervised learned clusters

Cluster	Topic	Examples of clustered words
c_1	Drug	abilify, adderall, ambien, ativan, aspirin, citalopram, effexor, paxil, …
c_2	Signs/Symptoms	hangover, headache, rash, hive, …
c_3	Signs/Symptoms	anxiety, depression, disorder, ocd, mania, stabilizer, …
c_4	Drug dosage	1000mg, 100mg, .10, 10mg, 600mg, 0.25, .05, ...
c_5	Treatment	anti-depressant, antidepressant, drug, med, medication, medicine, treat, …
c_6	Family member	brother, dad, daughter, father, husband, mom, mother, son, wife, …
c_7	Date	1992, 2011, 2012, 23rd, 8th, april, aug, august, december, …

Perelman School of Medicine
UNIVERSITY of PENNSYLVANIA

Classification task examples

- Adverse event mention classification[1]
- Drug abuse classification[2]
- Medication intake mention classification[3]
- Pregnancy announcement classification[4]

1 Sarker and Gonzalez. Portable automatic text classification. **J Biomed Inform.** 2015.
2 Sarker *et al.* Social media mining for toxicovigilance. **Drug Saf**. 2016.
3 Klein *et al.* Detecting Personal Medication Intake in Twitter. ACL **BioNLP Workshop 2017**
4 Sarker *et al* Discovering cohorts of pregnant women **..** **J Med Internet Res**. 2018

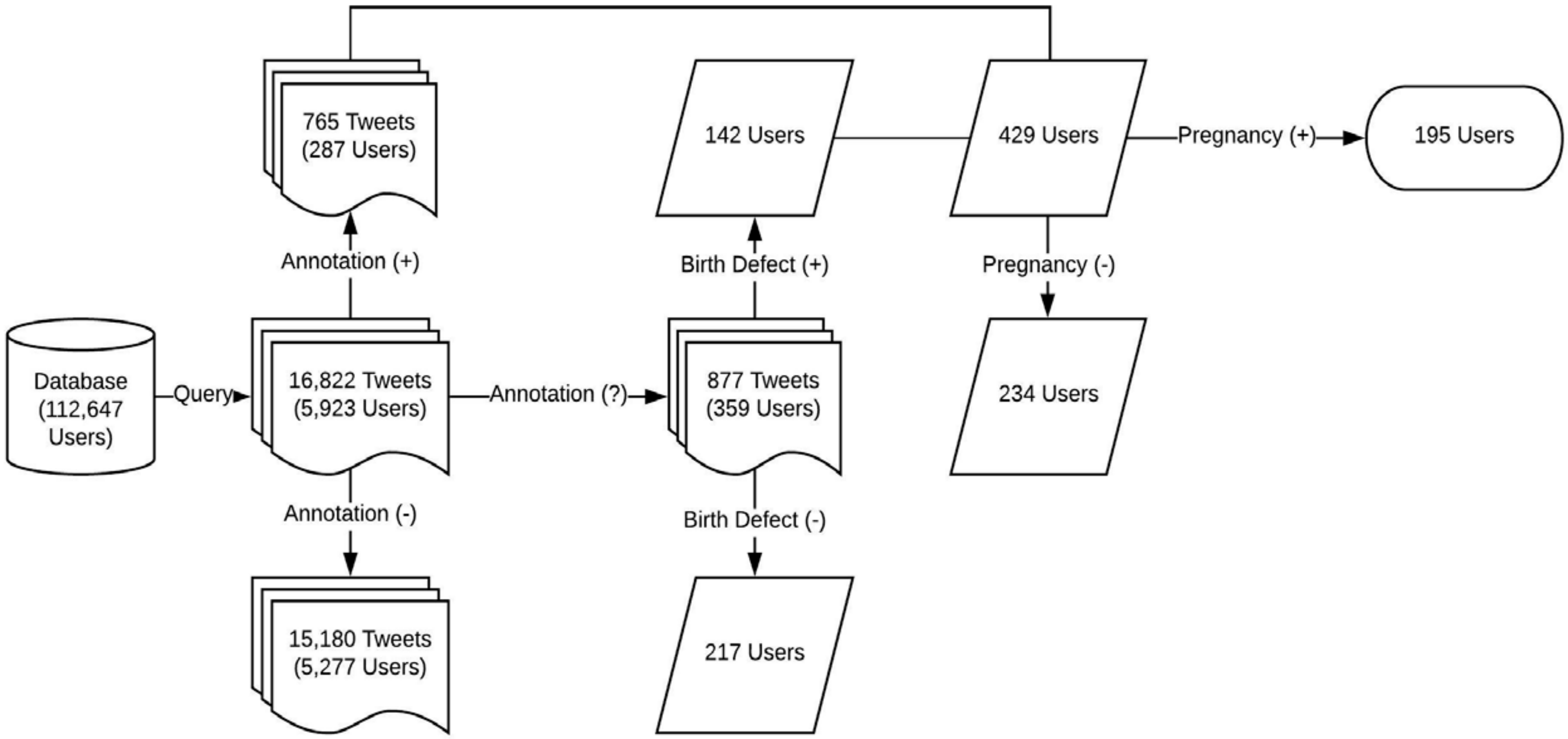
Case-control study with SM data?

- **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”*

1. Sarker *et al* Discovering cohorts of pregnant women **..** **J Med Internet Res**. 2018

From Twitter, “I am 12 weeks pregnant”

Finding cases – birth defects cohort



Birth defects data from Social Media

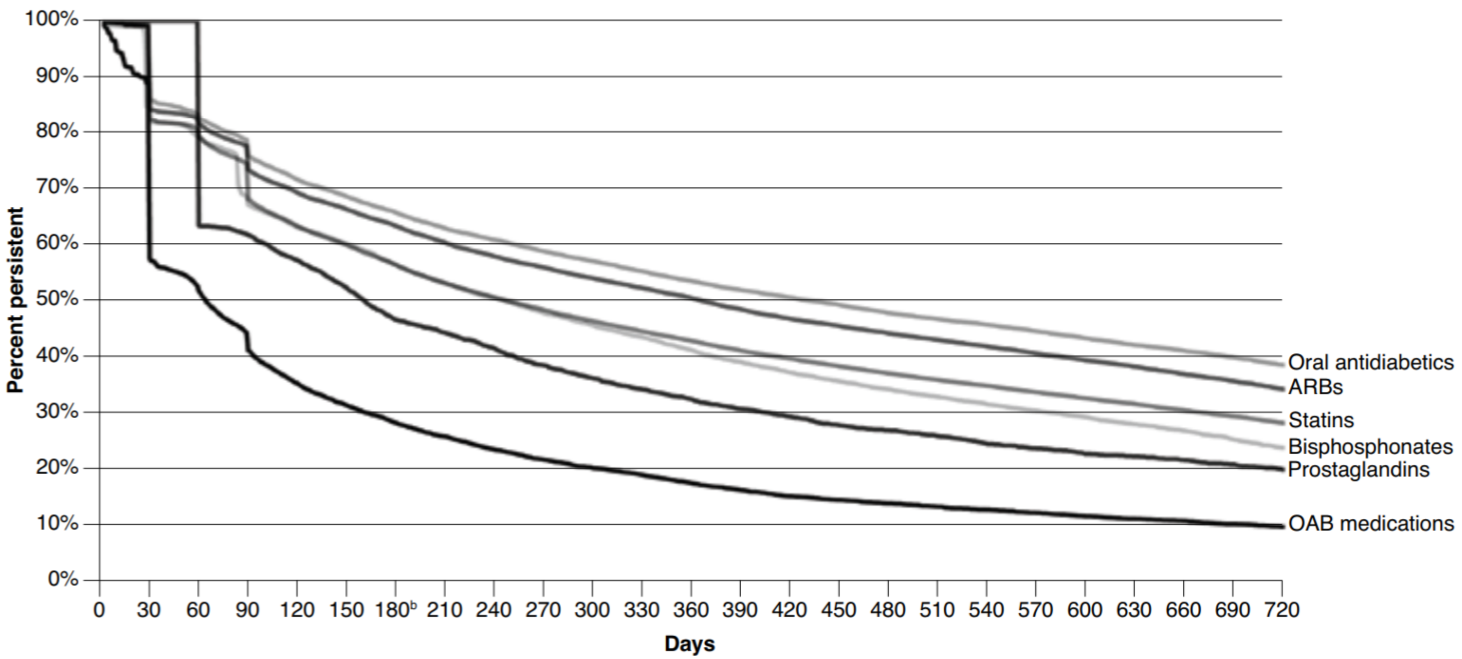
	Cases (n=197)	Controls (n=196)	OR or t-test [95% CI]	P-value
Age				
Median Age (IQR)	23 (20 to 28)	21 (19 to 23)	2 (1 to 3)	0.0001
Mean Age (range)	25 (17 to 42)	22 (16 to 37)	2.52 (1.38 to 3.66)	<0.0001
Women <30 years	80% (134/168)	91% (129/141)	0.37 (0.17 to 0.77)	0.004
Women <35 years	93% (156/168)	98% (138/141)	0.28 (0.05 to 1.08)	0.04
Missing data on age	14% (28/196)	28% (55/196)	0.43 (0.25 to 0.73)	0.0008
Race/Ethnicity				
Caucasian	73% (120/164)	55% (102/184)	2.19 (1.36 to 3.54)	chi^2 = 23.69, d.f. = 5 P < 0.001
Black	13% (22/164)	27% (51/184)	0.40 (0.22 to 0.72)	
Hispanic	9% (14/164)	12% (21/184)	0.72 (0.33 to 1.56)	
Asian	2% (4/164)	3% (5/184)	0.90 (0.17 to 4.24)	
Other (Islander, Native American/Indian, Multiracial/Mixed)	2% (4/164)	2% (5/184)	0.90 (0.17 to 4.24)	
Missing data on race	16% (32/196)	6% (12/196)	0.99 (1.44 to 6.58)	

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

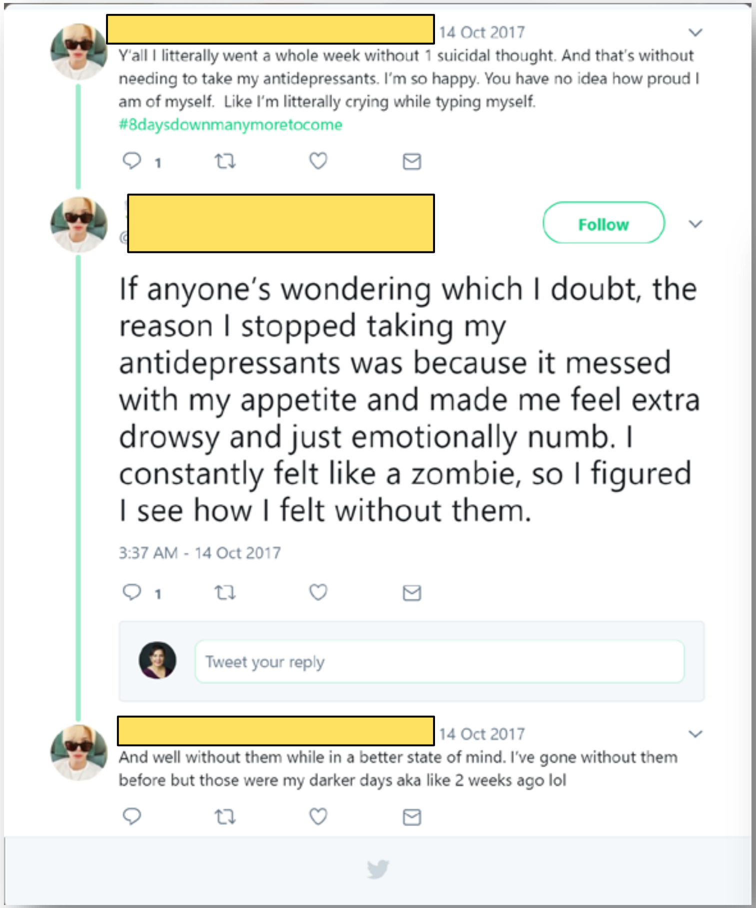
FIGURE 2 Time to Discontinuation[a] of 6 Chronic Therapy Classes, Allowing for 60-Day Treatment Gap

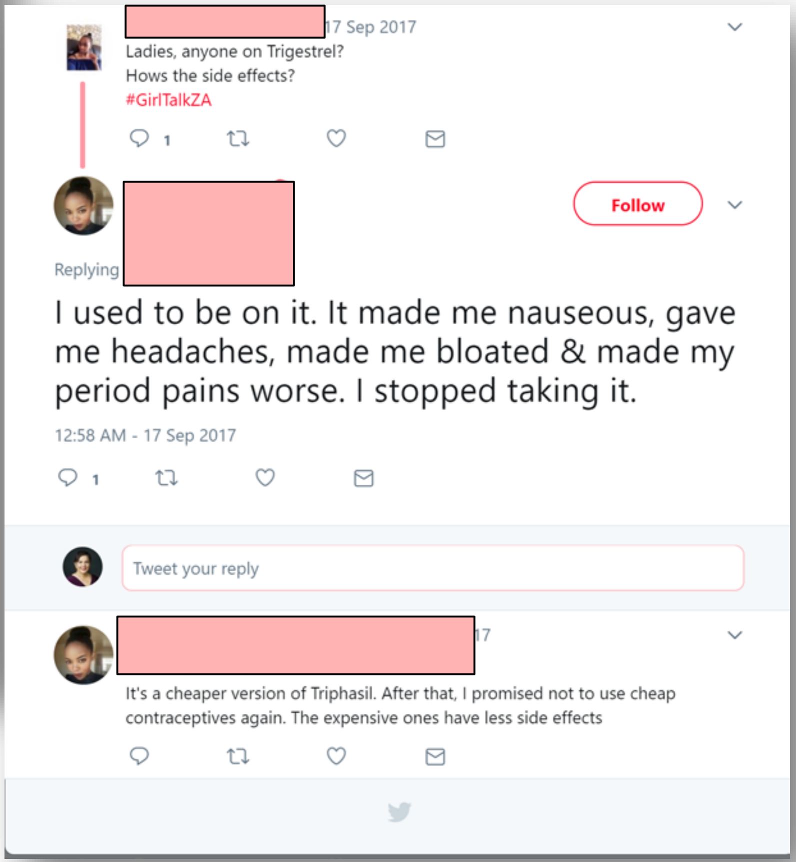


- prostaglandin analogs 47%
- statins 56%
- bisphosphonates 56%
- oral antidiabetics 66%
- angiotensin II receptor blocker 63%
- overactive bladder medications 28%

Comparing Adherence and Persistence Across 6 Chronic Medication Classes
Jason Yeaw, J Manag Care Spec Pharm, 2009 Nov;15(9):728-740.

“I stopped taking” & “made me”





Social media data: challenges

- **Incompleteness:**
  - Not all health conditions may be revealed
  - Complete data about individual cases may be difficult to obtain: taking drug X, but dosage, frequency, length of treatment, may be missing
  - Participants from the cohort may dropout at higher rates
- **Accessibility:**
  - Data from social media is not easily collected: API limitations
  - Not easily processed once collected: challenging to process using automatic methods
  - Data collection methods may have to be changed frequently over time
- **Authenticity:**
  - Bots – a large portion of social media is generated by bots, making it harder to mine reliable data
  - *Automatic* processing of postings is often misleading: for example postings mentioning a drug might not necessarily mean intake.

Thank you!

gr

Tw

HI

ht