Use of Social Media for Data Mining in Pharmacovigilance

Nabarun Dasgupta, MPH, PhD
Chief Science Officer, Epidemico, Inc.
Disclosures

Epidemico is commercializing aspects of this research.

Epidemico is wholly-owned subsidiary of Booz Allen Hamilton.

US FDA

01

FDA has been funding research technology for 3 years.
Office of the Chief Scientist (OCS)
Center for Tobacco Products (CTP)
Office of International Programs (OIP)

EUROPE - IMI

02

A public-private partnership entering Year 2. The WEB-RADR project is supported by the Innovative Medicines Initiative Joint Undertaking (IMI JU) under grant agreement n° 115632, resources of which are composed of financial contributions from the European Union’s Seventh Framework Programme (FP7/2007-2013) and EFPIA companies’ in kind contribution.

INDUSTRY

03

GSK has been a development partner to extend technology and obtain GxP software validation, over the last 2 years.
01
Background
Why bother looking at social media?
The Drug I’m Taking - Responsible for Monitoring
02
Methods
How to tame social media
Data Processing Overview

01. Acquire
Collect social media data from APIs, third-party authorized resellers, and automated scraping

02. Filter
Automated processes to identify relevant posts and clean data

03. Curate
OPTIONAL: remove false positives, disentangle attributions, expand dictionary, correct geography

04. Statistics… and Determining Causation
Automated statistics and visual display serve hypothesis generation. HOWEVER, determining causation requires information from other sources.
Automated Processes
For making sense of social media data

1. Automated NLP
   Bayesian classifier
   Machine learning
   Identify Events

2. Translate Vernacular
   Internet vernacular to MedDRA preferred terms

3. Reduce multiple copies
   Remove identifiers
   Consolidate

4. Geotag
   Use post geodata or profile location
   (~50% of Proto-AEs)
Proto-AE
Posts with Resemblance to Adverse Events

Albuterol has me feeling all sorts of dizzy and weak this morning 😞

Asthenia
MedDRA: 10003549
ICD-10: R42

Dizziness
MedDRA: 10013573
ICD-10: R53

For more examples see demo.medwatcher.org
Visual snow, don't want to eat, cross vision, lost my appetite

apetite surpressed googley eyed no appetitey
doublevision blind appetite is nonexistent blurry vision miss feeling hungry

blindness can't eat

I couldn't see double vision, blurry vision change blurry vision making me eat like a mouse

killed my apetite googly eyed lost teh appetite

lost apetite crosseved killed my appetite seeing weird colour

cross eyed seeing weird colour stomach small changes in vision

lack of apetite never want to eat seeing weird color

never hungry

Typos
visual snow  dont want to eat  apetite surpresred  googley eyes  blind  appetite is nonexistent  anorexic  double vision  #notevenhungry  lost appetite  lost their eyesight  cant eat  killed my apetite  googly eyed  lost apetite  crosseyed  killed cross eyed  seeing weird colour lack of apetite  never want to 视觉受损
MedDRA 10047571
Visual impairment
SNOMED 397540003

Decreasdeed appetite
MedDRA 10061428
Loss of appetite
SNOMED 79890006
System Statistics
As of October 23, 2015

63,000,000
Posts collected
Public posts from Facebook, Twitter & patient forums, back to 2012

2,500
Regulated products
Drugs, medical devices, vaccines, biologics, tobacco products, supply chain

380,000
Manually classified
Certified coders train machine learning classifiers & build dictionary

10,000
Vernacular English phrases
Updated daily by human curators. Mandarin, French, Spanish & Dutch next.
System Performance
How well does automated processing work?

Across all products, all time, all data sources

88% Sensitivity or Recall
Automated tools identify 9 out of 10 adverse events.

68% Positive Predictive Value or Precision
7 out of 10 posts contain AE information. Can increase to 100% with manual curation. May vary by product subset.
How to build a machine learning classifier

Congratulations! You just trained the classifier.

You are all MedDRA Certified Coders, right?

Instant Feedback: % Confidence that this is a Proto-AE

The Old Spice High Endurance deodorant is irritating the hell out of my skin. Too much alcohol in it I think.
Indicator score calculation

How well does automated processing work?

Albuterol has me feeling all sorts of dizzy and weak this morning 😩

1. Tokenization

- albuterol
- albuterol has
- albuterol has me
- has me
- feeling
- feeling all
- feeling all sorts
- sorts
- sorts of dizzy
- dizzy
- dizzy and weak
- weak
- weak this
- weak this morning
- morning
- morning 😩

2. Score calculation

For each token “w”

\[ b(w) = \frac{\text{(number of positives containing w)}}{\text{(total number of positives)}} \]

\[ g(w) = \frac{\text{(number of negatives containing w)}}{\text{(total number of negatives)}} \]

\[ p(w) = \frac{b(w)}{b(w) + g(w)} \]
03

One Applied Example

albuterol or salbutamol

Established oral and inhaled drug for asthma & COPD
Highly genericized market
A priori selected demonstration drug at demo.medwatcher.org
Methods

Albuterol/salbutamol has been a demo drug from project inception due to client interest

01. Public Facebook and Twitter
   1 year ending October 23, 2015
   albuterol, salbutamol, brand names

02. Proto-AEs
   English language only
   Use 0.7 indicator score threshold
   Multiplicate copies consolidated

03. No curation
   Need to check for false positives

04. Statistics
   Tree plots, top events, PRR
   Comparison to deduplicated UK spontaneous reports
   Drug Analysis Prints form MHRA
Spam Happens
Exclude n=297 Twitter posts mentioning *Ventolin*, having a link, 19-Mar-2015 to 02-Apr-2015
3% of Mentions were Proto-AEs

Mentions represent the patient voice.
They exclude most spam, news, coupons, and corporate announcements.
Albuterol/Salbutamol by System Organ Class
24-Oct 2014 to 23-Oct 2015, Facebook & Twitter
n=1,946 Proto-AEs

Rectangle size = Proto-AE count  |  Color intensity = PRR

- Cardiac disorders
- Ear and labyrint disor
ders
- Eye disorders
- Gastrointestinal disorders
- General disorders and administration site conditions
- Immune system disorders
- Infections and infestations
- Injury, poisoning and procedural complications
- Investigations
- Metabolism and nutrition disorders
- Musculoskeletal and connective tissue disorders
- Nervous system disorders
- Oncology, malignant and neoplastic (including cysts and polyps)
- Respiratory, thoracic and mediastinal disorders
- Skin and subcutaneous tissue disorders
- Social circumstances
- Surgical and medical procedures
- Vascular disorders
Albuterol/Salbutamol – Top 5 by Count
24-Oct 2014 to 23-Oct 2015, Facebook & Twitter
n=1,946 Proto-AEs
Public Facebook and Twitter combined.

- Tremor: 433
- Drug ineffective: 204
- Altered state of consciousness: 175
- Malaise: 136
- Cough: 123

Represents the things that most concern patients
Albuterol/Salbutamol – Top 5 by PRR
24-Oct 2014 to 23-Oct 2015, Facebook & Twitter
n=1,946 Proto-AEs

Unadjusted PRRs calculated against all other products in social listening database (n=2,551 medical products), 5 or more events

- Tremor: 35.7%
- Palpitations: 32.5%
- Feeling jittery: 31.1%
- Tachycardia: 26.6%
- Wheezing: 23.6%

Corresponds well with product label
Correlation: Spontaneous Report vs. Social Media

Non-parametric density (isometric plot) at HLGT level

**Spontaneous**

- 07-Apr-1969 to 26-Aug-2015
- 46.5 years
- 2,199 unique non-fatal reports
- n=3,851 reactions at HLGT
- 156 unique HLGT

**Social**

- 1 year
- 1,946 unique non-fatal reports
- N=2,419 reactions at HLGT
- 57 unique HLGT

More common in spontaneous data

Respiratory disorders NEC
- (cough, lung disorder, oropharyngeal pain, rhinorrhea, chest pain, dyspnoea)

Epidermal & dermal conditions
- (pruritus, skin discomfort, rash, angioedema, urticaria)

Movement disorders
- (tremor)

Therapeutic/non-therapeutic effects
- (drug ineffective)

Sleep disorders
- (insomnia, somnolence)

General system disorders
- (malaise, feeling jittery, pain, condition aggravated, fatigue, feeling abnormal, asthenia)
04
Social Media in Practice
Projects already underway for safety surveillance
Social Data in Concert with Spontaneous
Rectifying a false negative in an sponsor’s safety database

1. Routine Portfolio Surveillance
   Social media monitoring using Epidemico platform

2. Twitter Post Identified
   “Did this product cause x, y, z? I’ve had 3 administrations of this product and ever since then have felt a, b, c.”

3. Review Company’s Safety Database
   A couple of cases had been previously reported to the sponsor but had not crossed the signal threshold.

4. Submit to Regulators
   Assessment of causality and importance led the sponsor to initiate a submission with possible label change
Monitoring Discussions about Clinical Trials
Protecting validity and blinding

Discontinuation
I'm in a [Sponsor] double blind #### trial and could receive the placebo up until week 16 at which point I'm guaranteed to receive the drug. What I don't understand is that I can't be told if I was on the placebo for the first 16 weeks. I'm not seeing any positive benefit from whatever I've been taking for 8 weeks so far, and I would really like to know at week 16 if it's the drug because then I can bail from the study at that time and move on to something else…

Unblinding
[My brother] started on Thursday. He told me yesterday that he tried to chew a capsule (I'm not sure why, I think he wanted to see if it had a flavor) & it was hard as a rock. I'm assuming maybe it was enteric [sic] coated or something. He took the no flavor to mean he's on the placebo, but I actually wonder if they would bother enteric coating a placebo. Probably doesn't matter & I know we're not supposed to know anyway, just wondered if anyone else's was enteric coated.
Beyond Pharmacovigilance
Hospital sentiment, counterfeits
Collaboration with WHO-UMC & MedDRA MSSO

**Identify Products, Translate Vernacular to MedDRA**
Collaboration with WHO-UMC & MedDRA MSSO

**01**

**Remove Spam, Identify AE , Benefits, Sentiment**
Reduce noise, provide indicator scores

**02**

**Entity Extraction for Causal Determination**
Demographics, time to onset, dose, route, medical conditions, lab results, challenge

**03**

**Anonymize by Removing PII**
Names, hospital name, medical record #, phone, email

**04**

Coming 1Q2016: Free Public API
In appreciation for the public investment in this tool (FDA, IMI & Web-RADR)

For social media, clinical trial patient diaries, case safety reports, scientific literature
Limitations
What does “bias” mean in this setting?
Challenges
A few among many

01 Causality & Quantitative Emphasis
How well can patients assess causality? How much should we care?
Is a solely quantitative approach justified?

02 Patient Privacy vs. Muting the Patient Voice
Are we willing to disenfranchise the patient voice on privacy grounds?

03 Stability of Data Availability & Emerging Arenas
How do you build tools for an ever-shifting data environment?

04 Regulation Unclear
Can we all agree that nobody wants to see each Tweet as a separate case report in a safety database? So, why is everyone so afraid?
Are social media data biased?  
YES! All data are biased. But, can we understand the bias and make use of the information?
What we know, how we know it
Each source of product safety information provides different perspectives on patient experience

Hypothetical data for illustrative purposes only.

1. Patient-Reported Outcomes
Social media can elucidate what patients experience most frequently.

2. Seriousness
Traditional PV focus on most serious of events may be less central in social media.

3. Completeness
Social media posts may by less complete. Conversely, contain less sensitive identifying information than electronic health records.
Thank you
nabarun@epidemico.com
@epidemico
@WEBRADR
demo.medwatcher.org