Introduction to the Detection Analytics Core

Judith C. Maro, PhD[1]; Joshua C. Smith, PhD[2]; Shirley V. Wang, PhD[3]

1. Harvard Medical School and Harvard Pilgrim Health Care Institute 2. Vanderbilt University 3. Brigham and Women's Hospital and Harvard Medical School

February 18, 2021

Signal Detection: Past and Present

Routine Signal Identification Practices @ FDA

Case Reports in the FDA Adverse Event Reporting System (FAERS) and published medical literature

- Review of individual reports/articles
- Disproportionality analyses (i.e., Multi-Gamma Poisson Shrinker with Empirical Bayes Geometric Mean)
  - These indicate when reports of a particular exposure-outcome pairing are occurring more frequently than expected based on the total volume of reports received

Cumulative analyses

- Cumulative review of FAERS, literature, and Sponsor's periodic safety reports
- Risk-based approach* to frequency and product selection

* https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/UCM567959.pdf

FAERS is a Key Source Leading to Regulatory Action

Post-market drug safety evidence sources: an analysis of FDA drug safety communications

Chieko Ishiguro Research Expert[1], Marni Hall[2], George A. Neyarapally[2,*] and Gerald Dal Pan[2]

Version of Record online: 3 OCT 2012

DOI: 10.1002/pds.3317

Issue

Pharmacoepidemiology and Drug Safety

Volume 21, Issue 10, pages 1134–1136, October 2012

- 57% of FDA Drug Safety Communications were informed by FAERS data

Evaluation of FDA safety-related drug label changes in 2010

Jean Lester[1,2,*], George A. Neyarapally[2], Earlene Lipowski[1], Cheryl Fossum Graham[2], Marni Hall[2] and Gerald Dal Pan[2]

Version of Record online: 2 JAN 2013

DOI: 10.1002/pds.3395

Issue

Pharmacoepidemiology and Drug Safety

Volume 22, Issue 3, pages 302–305, March 2013

- Most common evidence sources:
  - Spontaneous reports (52%)
  - Clinical trials (16%)
  - Pharmacokinetic studies (11%)

FAERS: Advantages and Disadvantages

Advantages

- Good for detecting rare and acute events
- Captures all products and settings of use
- Can provide a patient perspective

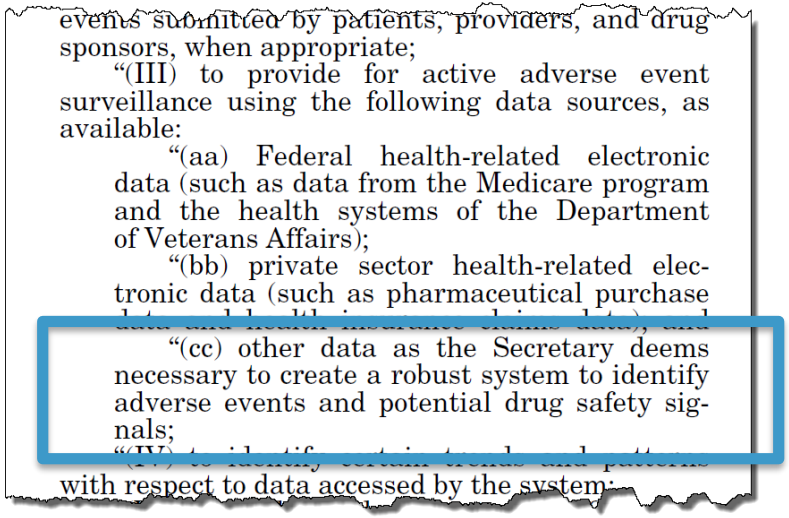
Disadvantages

- **Unknown denominator**, underreporting, stimulated reporting, variable information quality, etc.
- Performs poorly for long latency, high background rates, or idiopathic causes
- Cannot quantify/contextualize risk

https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard

Opportunity: Sentinel as Active Surveillance

- Longitudinal data provides denominator (i.e., exposure and event capture are not dependent on voluntary process)
- Ability to control for confounding variables
- Support from Institute of Medicine's 2007 *Future of Drug Safety*
- Inclusion in 2007 FDA Amendments Act



events submitted by patients, providers, and drug sponsors, when appropriate;

"(III) to provide for active adverse event surveillance using the following data sources, as available:

"(aa) Federal health-related electronic data (such as data from the Medicare program and the health systems of the Department of Veterans Affairs);

"(bb) private sector health-related electronic data (such as pharmaceutical purchase

"(cc) other data as the Secretary deems necessary to create a robust system to identify adverse events and potential drug safety signals;

with respect to data accessed by the system;

https://www.gpo.gov/fdsys/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf

Signal Identification in Sentinel as Compared to FAERS

Similarities

General Safety Net: No need to specify exposure-outcome pair of interest

Hypothesis Generation: Both produce hypotheses that necessitate further investigation

Tree Structure: Both can use data structured in hierarchical trees

Differences

Different Data Sources: Longitudinal data sources are compatible with familiar epidemiologic designs that analyze singular exposure-outcome pairings

Different Analytic Datasets: Longitudinal data can be analyzed as summary-level datasets rather than patient-level datasets

Multiplicity Control: Some methods have formal control for multiple hypothesis testing

Different Comparison Groups: Epidemiologic design dictates choice of comparison / referent group

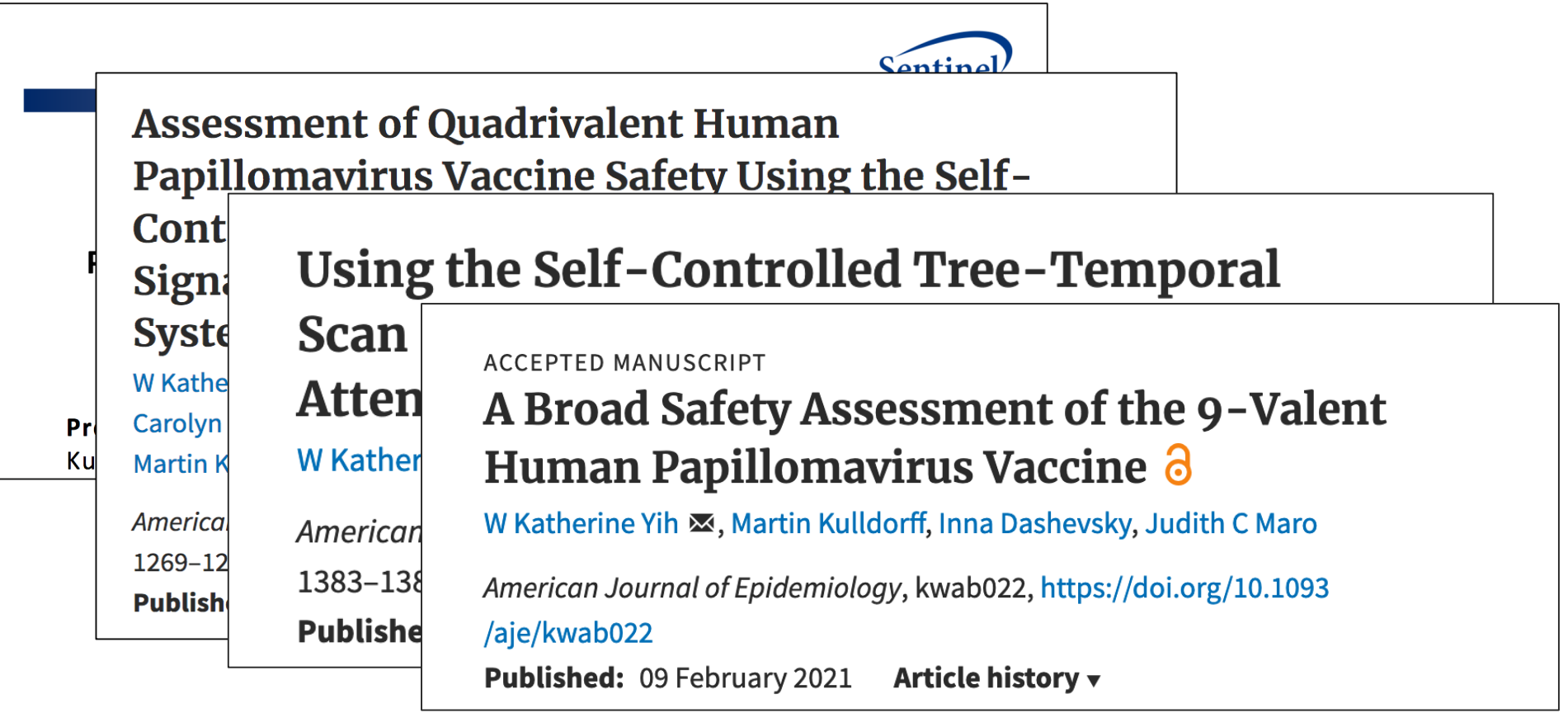
Signal Identification within the Sentinel System

Signal Identification Methods

Study Designs	TreeScan Analytics	Information Component Temporal Pattern Discovery (ICTPD)	Sequence Symmetry Analysis
Self-Controlled Design	X	X	X
Propensity Score or other Fixed Ratio Match Design	X		
Stratified Cohort Design	X		

https://www.sentinelinitiative.org/methods-data-tools/signal-identification-sentinel-system

Self-Controlled Designs (Tree-Temporal)



https://www.sentinelinitiative.org/assessments/vaccines-blood-biologics/postlicensure-medical-product-safety-data-mining-power, ; https://academic.oup.com/aje/article/187/6/1269/4904164; https://doi.org/10.1093/aje/kwz104, https://doi.org/10.1093/aje/kwab022

Propensity Score Matched Designs

Epidemiology. 29(6):895–903, NOV 2018
DOI: 10.1097/EDE.0000000000000907, PMID: 30074538
Issn Print: 1044-3983
Publication Date: 2018/11/01

Print

Data M
Proper
Statist

Shirley V. Wan
Gagne; Elisabe
Sebastian Sch

+ Author Infor

A General Propensity Score for Signal Detection using Tree-Based Scan Statistics

Shirley V Wang[1], Joshua J Gagne[1], Judith C Maro[2], Sushama Kattinakere[1], Danijela Stojanovic[3], Efe Eworuke[3], Elande Baro[4], Rita Ouellet-Hellstrom[3], Michael Nguyen[3], Elisabetta Patorno[1], Sandra DeLuccia[2], Ella Pestine[2], Yong Ma[4], Martin Kulldorff[1]

1. Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 2. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA; 3. Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD; 4. Office of Biostatistics, Center for Drug Evaluation and Research, FDA, Silver Spring, MD

https://pubmed.ncbi.nlm.nih.gov/30074538/; https://www.sentinelinitiative.org/sites/default/files/communications/publications-presentations/General_Propensity_Score_for_Signal_Detection_Using_TreeScan.pdf

Stratified Cohort Designs with Referent Cohort

https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.3423; https://doi.org/10.3390/pharmaceutics5010179; https://egems.academyhealth.org/articles/10.5334/egems.225/, https://doi.org/10.1093/aje/kwaa288

Signal Detection – Looking to Incorporate Structured and Unstructured EHR Data

Innovation Center Master Plan Published

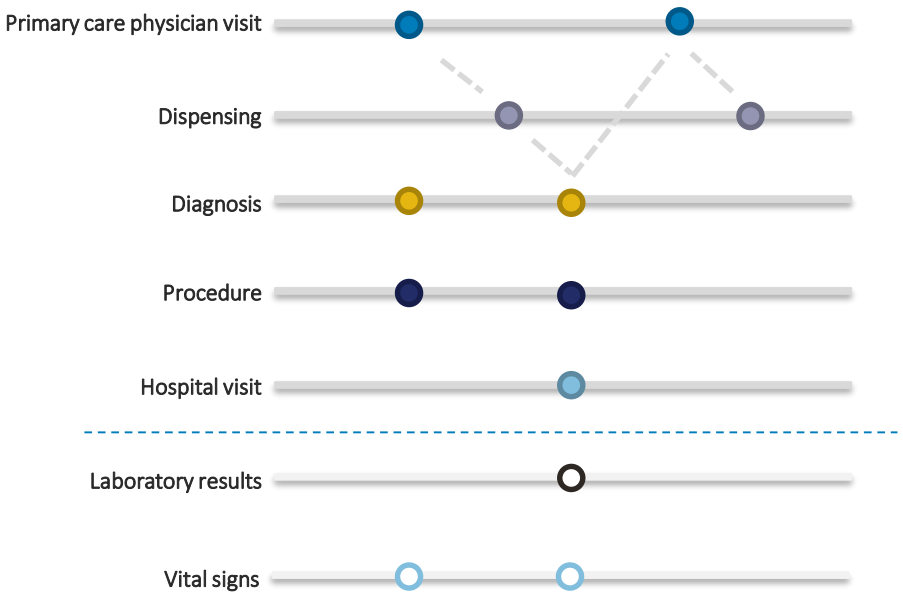
- "Electronic health record data offer a potentially promising complementary source of information for medical project safety signal detection but may require **different signal detection approaches** to account for and leverage differences in data content and structure…The Innovation Center will develop a methodological framework for **electronic health record-based signal detection** to address general safety use cases as well as the specific pregnancy, birth outcomes, and cancer use cases."

https://www.sentinelinitiative.org/sites/default/files/communications/publications-presentations/IC-Master-Plan.pdf

So, What's so Different about EHR Data?

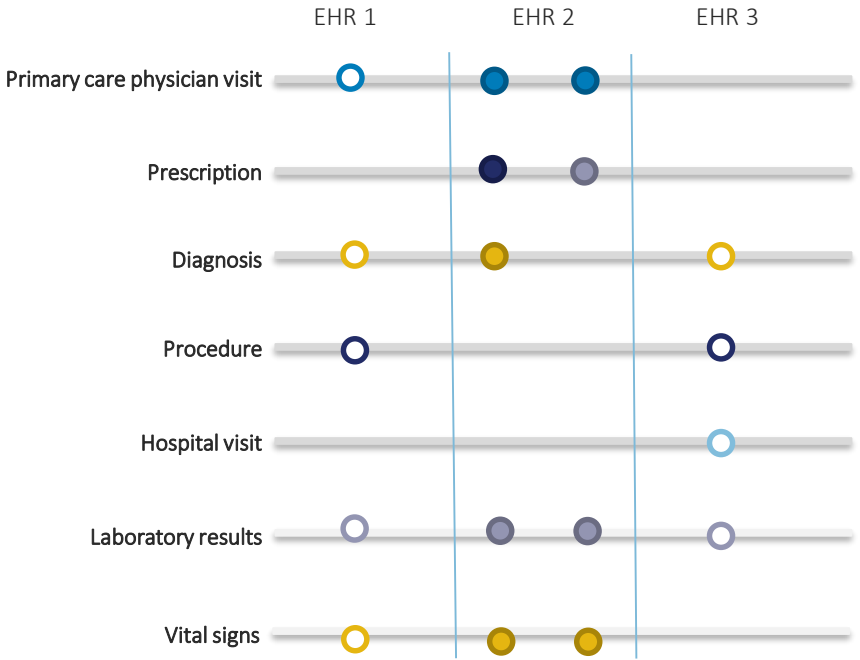
Claims Data

Comprehensive data across all encounters and settings
Miss some clinical detail



U.S. EHR Data

Detailed data within a single encounter that miss other encounters



Solid circles = captured data; Open circles = missing data

EHR Challenges Ahead…

- Standalone EHR data reaches back to techniques without a denominator (e.g., disproportionality analyses) because there is not a concept of complete capture of well-defined person time.

- Unstructured EHR data has promise but many challenges
  - How to filter, prioritize
  - How to annotate timing properly

- There have been efforts to leverage unstructured text in spontaneous reports.
  - Journal of Biomedical Informatics December 2015 Supplement

https://www.sciencedirect.com/journal/journal-of-biomedical-informatics/special-issue/10THTCB2D7L

Introduction to the Detection Analytics Core

Signal Detection Using Unstructured EHR Data

ADE Discovery from EHR Notes

Vanderbilt Study (2012)

"Exploring Adverse Drug Effect Discovery from Data Mining of Clinical Notes"

- Used NLP to identify findings/symptoms/diseases from Admission History & Physical Exam (H&P) notes current drugs from patients' medication lists
- "Snapshot in time"
- Extracted concepts represented as Drugs or "clinical manifestations" based on UMLS and RxNorm Semantic Types

Drugs & Clinical Manifestations

Drug Semantic Types

- Clinical Drug, Antibiotic, Pharmacologic Substance

Clinical Manifestation (CM) Semantic Types

- Anatomical Abnormality, Injury or Poisoning, Congenital Abnormality, Finding, Sign or Symptom, Acquired Abnormality, Clinical Attribute, Disease or Syndrome, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function

NLP Tools

KnowledgeMap Concept Indexer (KMCI)

- *Concept Recognition & Negation*

Sectag

- *Note section headers*

Medex

- *Medication extraction*

RxNorm

- *Normalize clinical drugs to medication ingredient*

NLP on Clinical Notes

Name: Doe, Jane **Date:** 10/10/2019
MRN: 12345678

History of Present Illness: Ms. Doe is an 80 yof with a PMHX of hypertension, congestive heart failure, hypothyroidism, who presents with a complaint of generalized weakness without fever, chills, or night sweats.

She reports having a dry cough for months. Denies abdominal pain, nausea, vomiting or diarrhea. In the ED, she was noted to have bradycardia with heart rate in the 35-40 range…

Family History: Father – MI at age 64; Sister – Alzheimer's disease.

Medications:
furosemide 80 mg tablet; 1 tablet by mouth daily
levothyroxine 112 mcg tablet; 1 tablet by mouth daily
omeprazole 20 mg capsule; 1 capsule by mouth daily
hydromorphone 2 mg tablet; 1 tablet by mouth every 8 hours
…

NLP on Clinical Notes

Name: Doe, Jane
MRN: 12345678

Date: 10/10/2019

History of Present Illness: Ms. Doe is an 80 yof with a PMHX of hypertension, congestive heart failure, hypothyroidism, who presents with a complaint of generalized weakness ~~without~~ fever, chills, or night sweats.

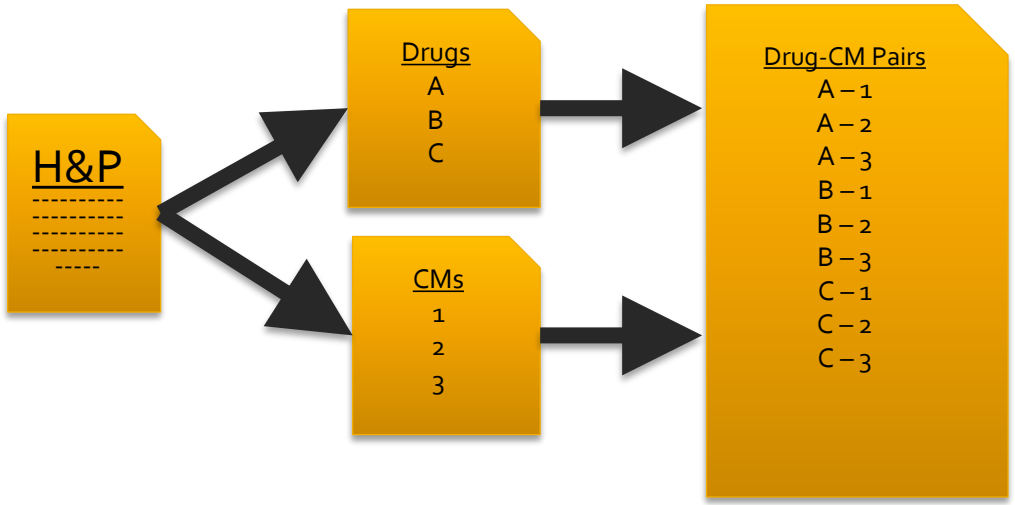
She reports having a dry cough for months. ~~Denies~~ abdominal pain, nausea, vomiting or diarrhea. In the ED, she was noted to have bradycardia with heart rate in the 35-40 range…

~~**Family History:** Father – MI at age 64; Sister – Alzheimer's disease.~~

Medications:
furosemide 80 mg tablet 1 tablet by mouth daily
levothyroxine 112 mcg tablet 1 tablet by mouth daily
omeprazole 20 mg capsule 1 capsule by mouth daily
hydromorphone 2 mg tablet 1 tablet by mouth every 8 hours
…

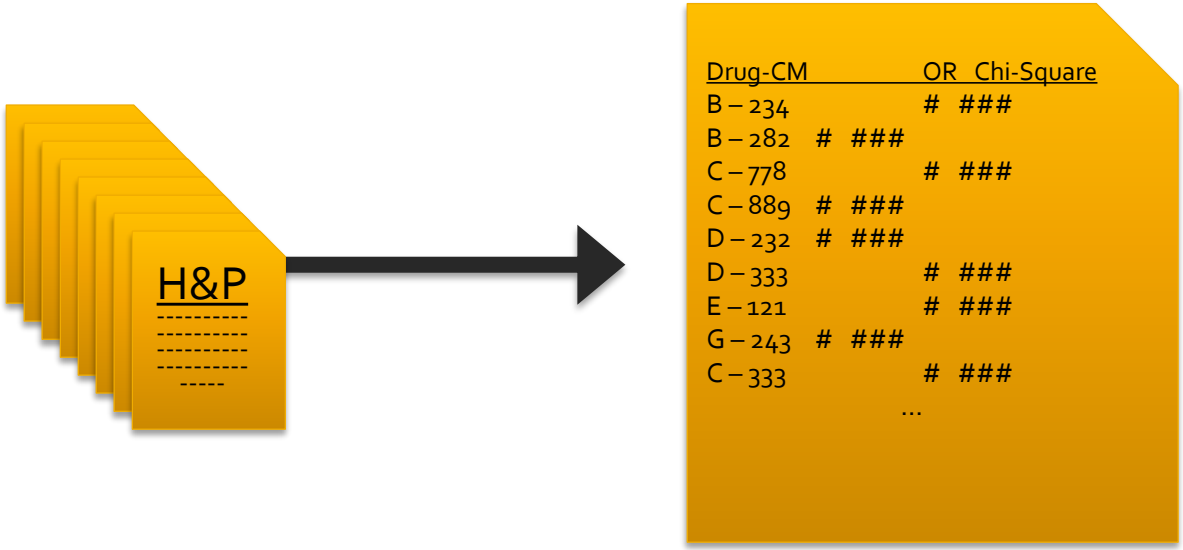
Procedure

Given an H&P note, we first extract all of the patients current drugs and clinical manifestations (CMs)



Procedure

Using 366,545 Admission H&Ps, we analyzed Drug-Clinical Manifestation correlations



Procedure

We calculated the odds ratio and Pearson's Chi-square for each drug-CM pair.

- Bonferroni correction to correct for multiple testing
- Required pairs to co-occur in at least 100 notes.

We also utilized a reference standard, based on drug product labels and other sources, to highlight indications and known ADEs.

Drug	Clinical Manifestation	Odds Ratio	Chi Square
Drug X	Known Indication	#	#
Drug X	Known Adverse Effect	#	#
Drug X	Other association	#	#

Correlations…

Drug	Clinical Manifestation	Odds Ratio	Chi Square
Drug X	Known Indication	#	#
Drug X	Known Indication	#	#
Drug X	Confounder	#	#
Drug X	Known Indication	#	#
Drug X	Known Adverse Effect	#	#
Drug X	Confounder	#	#
Drug X	Known Adverse Effect	#	#
Drug X	Confounder	#	#
Drug X	Confounder	#	#
Drug X	Confounder	#	#

What you actually find is **significant confounding** making it difficult to separate the signal from the noise…

Results

We processed 366,545 Admission H&Ps:

- **809,478** drug-CM pairs
- 1755 distinct drugs
- 10,723 distinct clinical manifestations.

After requiring a min 100 co-occurrences:

- 75,749 drug-CM pairs
- 666 distinct drugs
- 2182 distinct clinical manifestations.

Analysis of Drug-CM Pairs

After the Bonferroni correction, there were **39,304** pairs with a significant chi-square.

Based on our reference standard:

- 10,500 were known ADEs
- 3417 were Indications (INDs).

Selected Results

- Our top-ranked correlations, Rofecoxib, Rosiglitazone, Statins, and Insulin

Results – Top Correlations

Drug	Clinical Manifestation	count	odds	chisq	?	Expert Reviewer
Thyroxine	Hypothyroidism	13422	59.93	**122517.76**	IND	
Dornase Alfa	Pancreatic Insufficiency	773	637.71	**105067.22**		Confounder, due to CF
Dornase Alfa	Cystic Fibrosis	1418	1658.53	**90518.37**	IND	
Tobramycin	Pancreatic Insufficiency	647	368.44	**72462.81**		Confounder, due to CF
Tobramycin	Cystic Fibrosis	1212	346.65	**64923.85**	IND	
Allopurinol	Gout	2778	79.57	**61419.85**	IND	
Insulin	Diabetes Mellitus, Insulin-Dependent	6179	32.76	**55082.08**	IND	
Furosemide	Congestive heart failure	11955	12.04	**44120.11**	IND	
Nitroglycerin	Coronary Arteriosclerosis	10379	17	**42400.06**	IND	
Colchicine	Gout	1650	90.31	**40544.75**	IND	
Insulin	Diabetes Mellitus	11478	10.59	**36228.16**	IND	
Lactulose	Hepatic Encephalopathy	747	116.26	**35601.03**	IND	
Aspirin	Coronary Arteriosclerosis	19026	6.83	**35539.91**		Prophylaxis and early RX; IND
Statins	Hyperlipidemia	15536	7.73	**35356.23**	IND	
valacyclovir	Graft-vs-Host Disease	765	96.44	**33656.09**		Confounder, herpes prophylaxis or transplant patient treatment
Albuterol	Asthma	9549	10.01	**32429.05**	IND	
donepezil	Dementia	901	96.29	**31183.81**	IND	

Results – Top Correlations (cont.)

Drug	Clinical Manifestation	Count	odds	chisq	?	Expert Review
Nitroglycerin	Chest Pain	9501	11.42	**29787.4**	IND	
clopidogrel	Coronary Arteriosclerosis	7289	14.41	**28112.3**		IND
Illicit Drugs	abnormal bruising	728	87.24	**28061.35**		Too broad
Digoxin	Congestive heart failure	4728	15.16	**26264.48**	IND	
Sinemet	Parkinson Disease	756	115.75	**25794.43**		IND
latanoprost	Glaucoma	663	97.64	**24977.36**	IND	
Statins	Coronary Arteriosclerosis	15692	5.34	**24296.85**		IND
mesalamine	Crohn's disease	610	101.51	**23912.73**	IND	
Cocaine	Cocaine Abuse	552	98.09	**23906.61**		Trivial
Albuterol	Exacerbation of asthma	2553	30.84	**23675.4**		IND
Aspirin	Hypertensive disease	33022	4.51	**23593.69**	IND	Confounding
Hydroxychloroquine	Lupus Erythematosus, Systemic	572	86.91	**23029.24**	IND	
mesalamine	Ulcerative Colitis	423	105.89	**22653.9**	IND	
Levetiracetam	Seizures	2804	37.2	**22565.16**	IND	
Insulin	Diabetes Mellitus, Non-Insulin-Dependent	7624	7.81	**22347.89**	IND	
Statins	Hypertensive disease	30117	4.65	**22289.33**		Confounding
Tamsulosin	Benign prostatic hypertrophy	1430	31.73	**22267.01**	IND	
Insulin	Diabetic Ketoacidosis	2014	47.45	**22213.8**	IND	

Results – Rofecoxib

Drug	Clinical Manifestation	Count	odds	chisq	?
rofecoxib	Degenerative polyarthritis	250	3.35	**318.07**	IND
rofecoxib	Obesity	253	2.58	**188.01**	
rofecoxib	Hypertensive disease	598	2	**138.74**	AE
rofecoxib	Arthritis	157	2.63	**135.18**	IND
rofecoxib	Prothrombin time increased	101	3.1	**129.33**	
rofecoxib	Rheumatoid Arthritis	212	2.21	**113.16**	IND
rofecoxib	Congestive heart failure	170	2.32	**107.98**	AE
rofecoxib	Metabolic Diseases	216	2.1	**100.06**	
rofecoxib	Myocardial Infarction	189	2.17	**98.77**	AE
rofecoxib	Chest Pain	267	1.95	**94.2**	AE
rofecoxib	Coronary Arteriosclerosis	248	1.98	**92.85**	
rofecoxib	White blood cell count increased	233	1.96	**86.54**	
rofecoxib	Mental Depression	238	1.9	**80.1**	
rofecoxib	Shortness of Breath	260	1.77	**66.08**	
rofecoxib	Lupus Erythematosus, Discoid	145	1.99	**61.54**	

Results – Rofecoxib (cont.)

Drug	Clinical Manifestation	Count	odds	chisq	?
rofecoxib	Gastroesophageal reflux disease	212	1.8	**60.93**	AE
rofecoxib	Adverse Event Associated with the Gastrointestinal System	107	2.1	**55.35**	
rofecoxib	Back Pain	119	2.02	**54.62**	IND
rofecoxib	Swelling	113	1.93	**44.95**	
rofecoxib	Pain	521	1.49	**44.48**	IND
rofecoxib	Hypothyroidism	129	1.83	**42.89**	
rofecoxib	Osteoporosis	114	1.87	**41.58**	
rofecoxib	Asthenia	137	1.76	**39.41**	AE
rofecoxib	Gastrointestinal tract finding	112	1.85	**39.09**	
rofecoxib	Diabetes Mellitus	198	1.6	**36.66**	
rofecoxib	Chronic Obstructive Airway Disease	126	1.67	29.89	
rofecoxib	Urinary tract infection	121	1.67	28.81	AE
rofecoxib	Anemia	135	1.61	27.52	
rofecoxib	Lesion	273	1.44	27.43	
rofecoxib	Cerebrovascular accident	136	1.57	24.86	AE

Results – Rosiglitazone

Drug	Clinical Manifestation	Count	odds	chisq	?
rosiglitazone	Diabetes Mellitus, Non-Insulin-Dependent	608	9.11	**2416.6**	IND
rosiglitazone	Diabetes Mellitus	745	8.77	**2334.6**	IND
rosiglitazone	Hypertensive disease	1028	5.02	**849.05**	
rosiglitazone	Obesity	420	3.85	**611.06**	
rosiglitazone	Hyperlipidemia	384	3.44	**475.98**	
rosiglitazone	Coronary Arteriosclerosis	396	2.78	**320.68**	
rosiglitazone	Gastroesophageal reflux disease	300	2.13	**139.92**	
rosiglitazone	Lupus Erythematosus, Discoid	209	2.37	**139.78**	
rosiglitazone	hypercholesterolemia	164	2.54	**133.66**	
rosiglitazone	Anicteric	808	1.82	**123.49**	
rosiglitazone	Arthritis	177	2.36	**119.52**	
rosiglitazone	Angina Pectoris	116	2.71	**113.74**	
rosiglitazone	Chronic Obstructive Airway Disease	197	2.18	**107.52**	
rosiglitazone	Dyspnea on exertion	153	2.21	**89.42**	
rosiglitazone	Congestive heart failure	190	2.06	**88.6**	AE
rosiglitazone	Shortness of Breath	326	1.79	**87.12**	
rosiglitazone	Orthopnea	136	2.27	**86.62**	
rosiglitazone	Myocardial Infarction	214	1.95	**83.16**	AE
rosiglitazone	Paroxysmal atrial tachycardia	296	1.79	**81.46**	
rosiglitazone	Anemia	193	1.9	**70.68**	AE
rosiglitazone	Visual impairment	111	2.23	**68.43**	

Results – Statins

Drug	Finding	Count	odds	chisq	?
Statins	Hyperlipidemia	15536	7.73	**35356.23**	IND
Statins	Coronary Arteriosclerosis	15692	5.34	**24296.85**	
Statins	Hypertensive disease	30117	4.65	**22289.33**	IND
Statins	hypercholesterolemia	7825	6.36	**17068.75**	IND
Statins	Myocardial Infarction	8511	3.15	**7456.66**	IND
Statins	Stenosis	4370	4.24	**6425.6**	
Statins	Diabetes Mellitus	10110	2.6	**5996.25**	IND
Statins	Diabetes Mellitus, Non-Insulin-Dependent	7419	2.78	**5339.97**	IND
Statins	Peripheral Vascular Diseases	3751	4.1	**5321.13**	IND
Statins	Congestive heart failure	6713	2.8	**4941.87**	
Statins	Angina Pectoris	3702	3.49	**4204.19**	IND
Statins	Epilepsy	7834	2.34	**3880.28**	
Statins	Cerebrovascular accident	6866	2.44	**3838.77**	IND

Results – Statins (cont.)

Drug	Finding	cocount	odds	chisq	det
Statins	Ischemic cardiomyopathy	1815	5.28	**3582.31**	
Statins	Retina-normal	1449	5.82	**3173.03**	
Statins	Gastroesophageal reflux disease	8464	2.05	**2977.79**	
Statins	Ischemia	3301	3	**2957.79**	IND
Statins	Obesity	7689	2.1	**2917.42**	IND
Statins	Arthritis	5041	2.42	**2856.74**	
Statins	Chronic Obstructive Airway Disease	5779	2.28	**2828.4**	IND
Statins	Dyslipidemias	1840	4.25	**2826.03**	IND
Statins	Mental Depression	8899	1.98	**2820.19**	AE
Statins	Memory Impairment	312	1.77	**84.57**	
...					
Statins	Memory Loss	376	1.22	**13.19**	
...					
Statins	Memory observations	112	1.41	**11.3**	

Result – Insulin

Drug	Finding	Odds Ratio	Chi Square
Insulin	Diabetes Mellitus, Insulin-Dependent	32.76	55082
Insulin	Diabetes Mellitus	10.59	36228
Insulin	Diabetes Mellitus, Non-Insulin-Dependent	7.81	22347
Insulin	Diabetic Ketoacidosis	47.45	22213
Insulin	Retinal Diseases	17.9	10865
Insulin	Hyperglycemia	7.74	8990
Insulin	Hypertensive disease	3.52	8286
Insulin	Diabetic Neuropathies	15.75	7283
Insulin	Coronary Arteriosclerosis	3.29	6426
Insulin	Congestive heart failure	3.8	6190
Insulin	Neuropathy	6.74	6163
Insulin	Obesity	3.07	4888
Insulin	Hyperlipidemia	3.03	4690
Insulin	Diabetic Nephropathy	16.42	4524
Insulin	hypoglycemia	6.94	4457
Insulin	Diabetic Retinopathy	5.99	4167
Insulin	Kidney Diseases	5.63	3954
Insulin	Peripheral Vascular Diseases	3.82	3140
Insulin	Proliferative diabetic retinopathy	20.35	3109
Insulin	Foot Ulcer	14.33	2683
Insulin	Ketoacidosis	27.75	2340
Insulin	Foot Ulcer, Diabetic	16.94	2308

Discussion

Significantly correlated drug-CM pairs seemed reasonable, representing both **known ADEs** or **indications**.

Correlations representing unrecognized ADEs were potentially discoverable **before** they were known.

NLP is sometimes coarse and the ambiguous nature of some CM concepts can be a problem.

Confounding due to co-morbid conditions and symptoms of a disease was very prevalent.

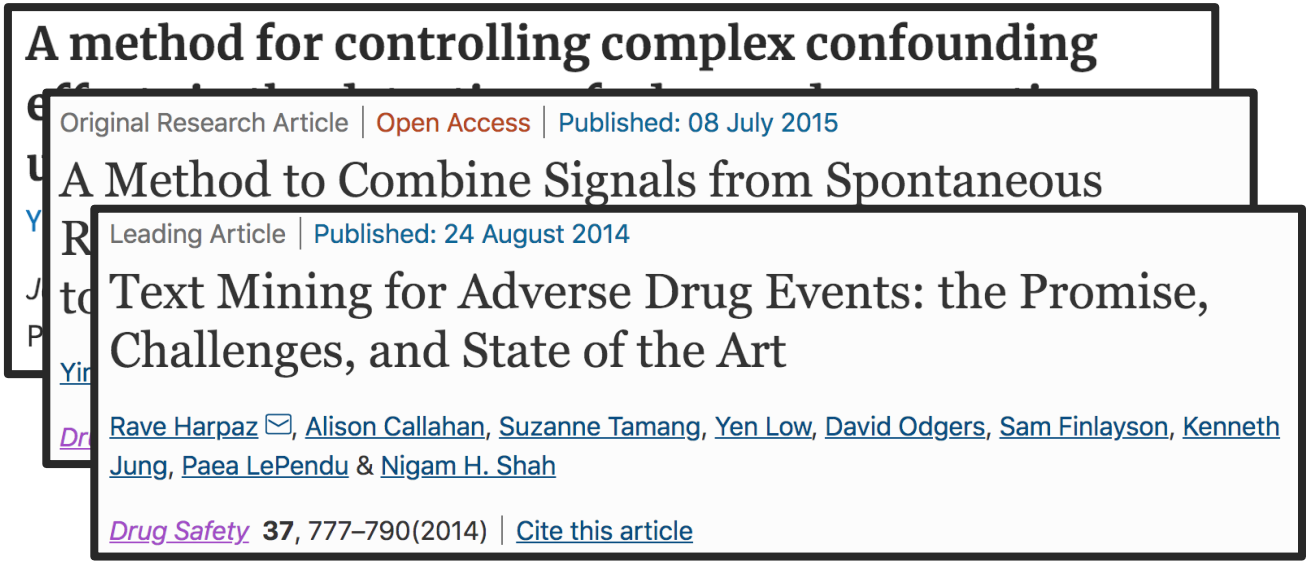
Confounding

The vast amount of unstructured EHR data exacerbates the problem of confounding by introducing many conditions.

Adverse Effect signals are likely to be:

- confounded by co-medication
- confounded by indication
- confounded by comorbidity
- or any combination of the three.

Confounding



Questions & Opportunities

What methods should we use to adjust for confounding?

Do we focus on disproportionality analysis, or other approaches using regression or epidemiologic study design?

How do we deal with timing, missingness in data?

How will be combine NLP data with claims/labs?

How can we best normalize concepts extracted using NLP?

Do we need a reference standard (indications, known ADEs in a computable format)?