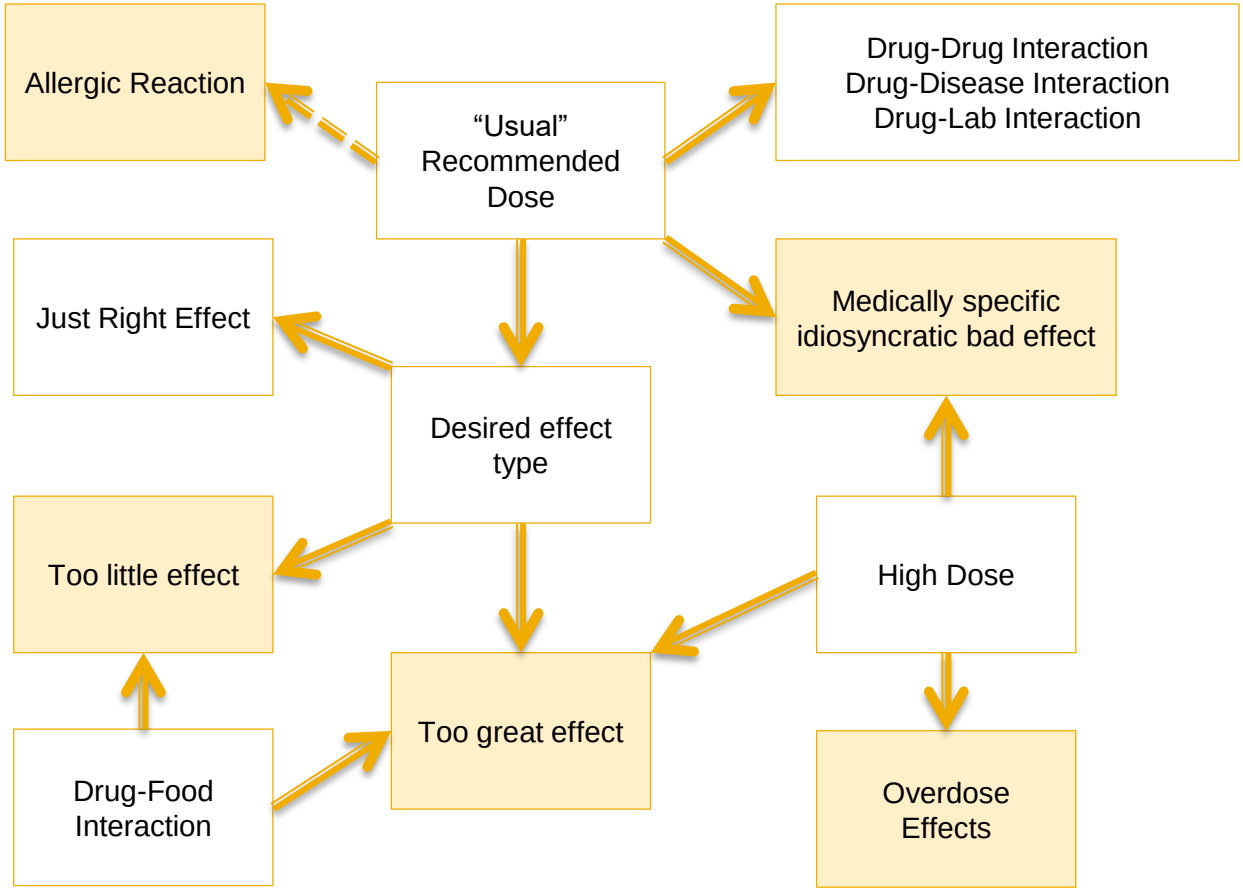
Exploring Adverse Drug Effect Discovery from Data Mining of Clinical Notes

Slides from Josh Smith, PhD

Overview

- Hypothesis: Drugs and findings that co-occur together more often in clinical documents do so for a cause-and-effect reason (e.g., indication/treatment or AE).
- Goal: Identify the drug and finding concepts that co-occur in a large number of H&P notes.
- Implementation: If we can rule out those which co-occur for known reasons (indications or adverse effects), those that remain could be unknown side effects.

Adverse Effects



Premarket Evaluation of Drugs

- FDA requires testing for both efficacy and safety before a drug is sold in the marketplace

Phase 1. 20-80 people; initial testing to determine metabolism, pharmacologic action, safe dosage range, and side effects

Phase 2. 100-300 people; controlled trial; tests drug effectiveness in treating a particular condition; further evaluates drug safety

Phase 3. 1000-3000 people; controlled or uncontrolled; risk-benefit relationship

Understanding Clinical Trials - ClinicalTrials.gov. Available from: http://clinicaltrials.gov/ct2/info/understand

Premarket Evaluation of Drugs

- AEs do not always arise in pre-market trials
  - Trials are relatively brief and small; cannot detect rare effects occurring in fewer than 1 in 10,000
  - Inclusion criteria are often restrictive in terms of age, race, gender, and health status
  - Trials rarely emulate the varying conditions of medication use by the public
  - Trials cannot fully explore effects of comorbid conditions, ranges of dosing, duration, and interactions with other medications

Examples of significant AEs

- **Rofecoxib** (Vioxx) – a COX-2 selective NSAID
- 1999: Approved by the FDA to treat arthritis, acute pain in adults, and dysmenorrhea.
- Sold over 80 million prescriptions worldwide over 5 years.
- 2004: Withdrawn from the market after evidence linked the drug with myocardial infarction (MI) and stroke

Recent Examples - Rosiglitazone

- **Rosiglitazone** (Avandia) – drug used for treatment of Type 2 diabetes
  - 1999: FDA approval
  - 2006: Sales of ~$2.2 billion
  - 2007: Linked with MI and other CV effects
  - 2010: Removed from the market in Europe and severely restricted by the FDA
- Other recent examples include **Statins** and **Risperidone** (Risperdal)

Prevalence of AEs

- 2004 – A study[1] of ~20,000 inpatients showed 6.5% of admissions were associated with ADRs
  - ADR directly led to admission in 80% of cases
- 2008 – A systematic review[2] of 25 studies involving 100,000+ patients showed ~5.3% of admission were associated with ADRs
- 2010 – An inpatient study[3] in a Dutch hospital found 19-29% of admissions were due to ADRs

1. Pirmohamed, et al. Adverse drug reactions as cause of admission to hospital… BMJ 2004 Jul;329(7456):15–19.
2. Kongkaew, et al. Hospital Admissions Associated with Adverse Drug Reactions… Ann Pharmacother 2008 Jul;42(7/8):1017–1025.
3. Atiqi, et al. Prevalence of iatrogenic admissions to the Departments of Medicine/Cardiology/ Pulmonology in a 1,250 bed general hospital. Int J Clin Pharmacol Ther 2010 Aug;48(8):517–524.

Pharmacovigilance

- Comprises the detection, assessment, understanding, and prevention of AEs at both normal and excessive doses
- Began in the 1960's after the tragic widespread effects of Thalidomide use during pregnancy
- Includes pre-marketing risk assessment, ongoing risk minimization, and **post-marketing surveillance**

Post-Marketing Surveillance

- Spontaneous Reports
  - Reports of suspected AEs by hospitals, healthcare professionals, patients, and drug manufacturers
- Spontaneous Reporting Systems (SRS)
  - FDA MedWatch and AERS
  - EMA EudraVigilance
  - WHO International Drug Monitoring Programme

Spontaneous Reporting Systems

- Modern methods use analysis of aggregate data to detect and evaluate AE "signals"
  - Popular methods include the Proportional Reporting Ratio (PRR), Reporting Odds Ratio (ROR), Pearson's Chi-Square Test, and the Poisson Probability Test.
  - Other methods that have been used include correlation analysis, regression analysis, neural networks, empirical Bayes screening, and many others.

Shortcoming of SRS

- Problems with spontaneous reports
  - Under-, Over-, and Duplicate Reporting
  - Voluntary
  - Serious vs. non-serious AEs
  - Limited temporal information
- Many methods are "numerator-based"
- These methods are for identifying hypotheses; definitive proof still requires clinical and experimental research.

AE Detection From Other Sources

- SRS were once the only sources of large scale AE data, but EMRs now contain similar data.
- It has been shown that AE information is present in EMRs.[1]
- As late as 2005, few, if any, major stakeholders had the goal of hypothesis-free examination of large databases in efforts to find new AEs.[2]
- While some methods compensate for SRS shortcoming, the field must explore EMRs as a data source.

1. Wang, et al. Selecting information in electronic health records for knowledge acquisition. JBMI 2010 Aug;43(4):595–601.
2. Roden DM. An underrecognized challenge in evaluating postmarketing drug safety. Circulation 2005 Jan;111(3):246–248.

AE Discovery From EMRs

- Carol Friedman and colleagues at Columbia
  - Used NLP to extract diseases and related symptoms from discharge summaries, as well as drug concepts.[1]
  - Used co-occurrence statistics to find correlations between seven drugs/drug classes and their known AEs from ~25,000 discharge summaries.[2]
  - Based on dates of discharge summaries, showed that some AEs appeared in notes in significant number before they were "discovered." [2]

1. Wang, et al. Automated knowledge acquisition from clinical narrative reports. AMIA Annu Symp Proc 2008;783–787.
2. Wang, et al. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. JAMIA 2009 Mar;16(3):328–337.

AE Discovery From EMRs

- FDA Sentinel Initiative[1]
  - "Create a linked, sustainable system that will draw on existing automated healthcare data from multiple sources to actively monitor the safety of medical products continuously and in real time."
  - Will make use of data mining on healthcare information from sources such as hospitals and insurance companies.

1. FDA's Sentinel Initiative. Available from: http://www.fda.gov/Safety/FDAsSentinelInitiative/default.htm

Overview: Feasibility of Detecting AEs

- Using NLP, we extracted drug and finding concepts from a corpus of H&Ps.
- We calculated the statistical correlation between drugs and findings appearing in the notes.
- We used DEB to categorize known drug-finding pairs.
- We did an informal exploratory analysis to determine if drug-finding correlations seemed valid and if recently discovered AEs could be detected.

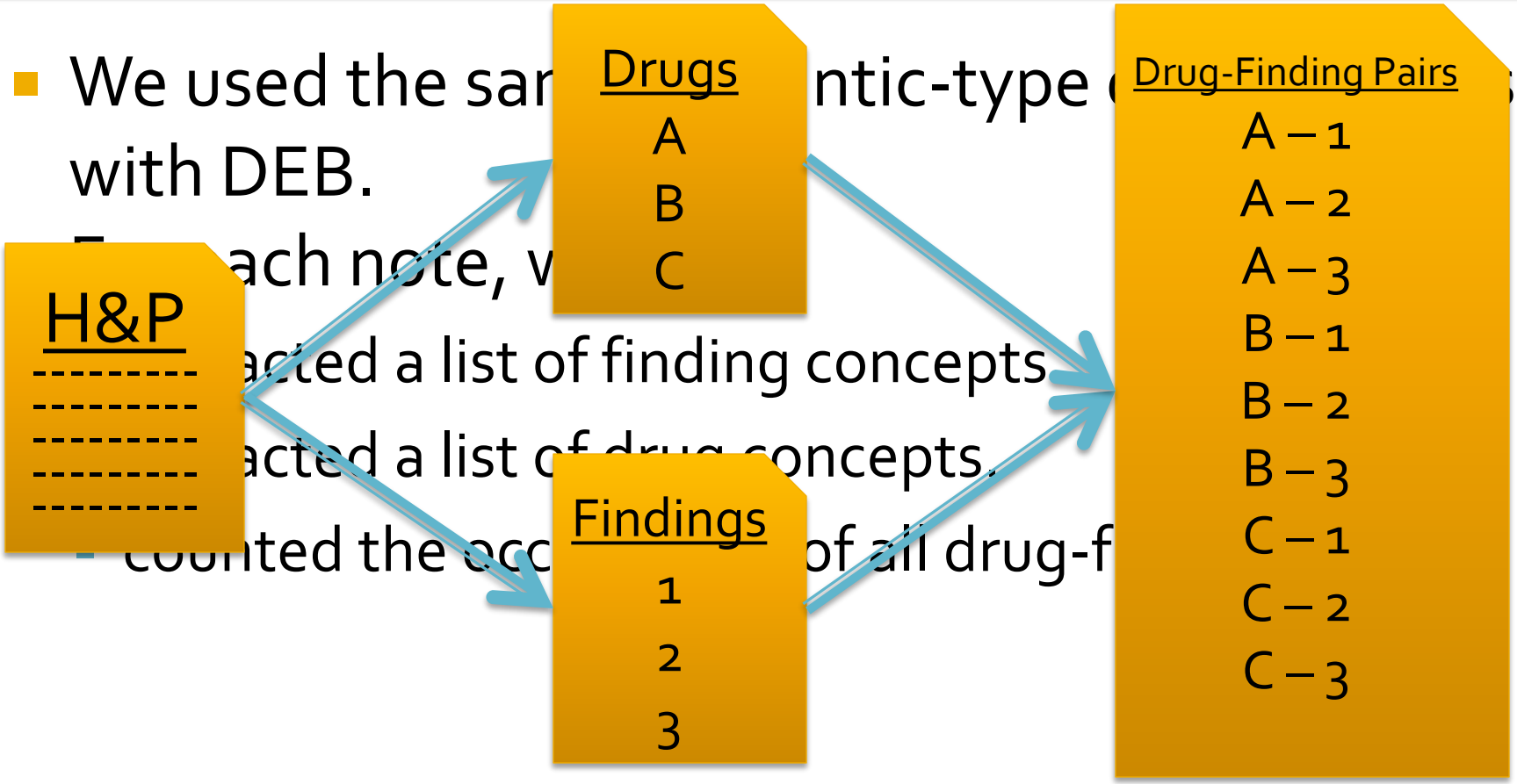
H&P Notes

- We obtained 500,000 "likely H&Ps" from the SD, but needed to further refinement.
- We used SecTag to identify sections in the note, then defined criteria for a "high quality" H&P
- We required that at least 2/3 criteria be met.
  - 366,545 H&Ps remained out of ~500,000

Criterion	Sections Present in the note
1	(History of Present Illness OR Past Medical History) AND Physical Exam
2	At least 4 of (Review of Systems | Chief Complaint | Assessment | Family Medical History | Medications | (History of Present Illness | Past Medical History))
3	At least 5 of (Vital Signs | Pulmonary Exam | Cardiovascular Exam | Neurological Exam | HEENT Exam | Abdominal Exam | Lymphatic Exam | Extremity Exam | General Exam)

Extracting Drug-Finding Pairs

- We used the sar… ntic-type …
  with DEB.
  - …ach note, w…
    - …acted a list of finding concepts.
    - …acted a list of drug concepts.
    - counted the occ… of all drug-f…



Extracting Findings from H&Ps

- Used KMCI and SecTag to identify findings in H&Ps
- Added CUIs to "**finding list**" if…
  - it was of the relevant semantic type.
  - it had not been seen before in current note.
  - it was not in an excluded section of an H&P note.
    - We did not include concepts from *Family Medical History* sections, as they were less likely to be related directly to the patient

Extracting Drugs from H&Ps

- Using SecTag, extracted H&P sections to use as input to MedEx
  - Included H&P sections likely to contain past or current medications, such as History of Present Illness, Past Medical History, Medications, Medication History, etc.
  - Did not include "Plan" sections, which contain future medications.
- Added CUIs to "**drug list**" if…
  - it had appropriate semantic type.
  - at least one MedEx signature field was present.
  - it had not been seen before in current note.

Filtering Drug-Finding Pairs

- Drug mapping to generic ingredients
  - Mapping to drug ingredients using RxNorm, per DEB methods
  - Improved over DEB by creating manual mapping for frequently appearing drugs that lacked mappings
- Removing irrelevant concepts
  - Manually removed most frequently occurring non-descriptive and irrelevant extracted concepts
  - Examples: “blood for culture,” “date of admission,” “annual exam,” etc.

Analysis of Drug-Finding Pairs

- For each drug-finding pair, we calculated the odds ratio and Pearson's Chi-square in set of H&Ps.
- For a drug-finding pair to remain in the analysis, we required it to appear in at least 100 notes.
- Since we performed the Chi-square test ~100,000 times, we used a Bonferroni correction to correct for multiple testing $p <= 10^{-8}$.

Identifying Known D-F Pairs

- KMCI, MedEx, and DEB all use UMLS CUIs to designate concepts, so checking for a drug-finding pair in DEB required only simple matching.

Informal Analysis Questions

1. Do drug-finding pair classifications derived by applying the DEB to clinical notes appear reasonable most of the time?
2. When DEB classifications appear to be incorrect, is there at least a plausible reason that explains the mistake?
3. Does there appear to be, at face value, a known clinical reason that can be cited to explain those clinical-note-derived drug-finding pairs that are highly statistically correlated?
4. Can the method rank recently discovered drug-AE pairs high enough to suggest future potential for more careful side effect discovery?

Results

- We started with 500,000 H&P notes.
- After using SecTag to filter out lower-quality H&Ps, we had ~366,600 H&Ps.
- From them, we extracted 809,478 drug-finding pairs composed of 1755 distinct drugs and 10,723 distinct findings.
- After requiring a min co-occurrence count of 100, we retained 75,749 pairs composed of 666 distinct drugs and 2182 distinct findings.

Results

- After the Bonferroni correction, there were 39,304 pairs with a significant chi-square value.
- 10,500 were identified as AEs, 3417 as INDs.
- We show the highest ranked correlations overall and for each of the following drugs:
  - Rofecoxib
  - Rosiglitazone
  - Risperidone
  - Statins

Results – Top-Ranked Correlations

Drug	Finding	cocount	odds	chisq	det	Review by Expert
Thyroxine	Hypothyroidism	13422	59.93	**122517.76**	IND	OK
Dornase Alfa	Pancreatic Insufficiency	773	637.71	**105067.22**		Confounder, due to CF
Dornase Alfa	Cystic Fibrosis	1418	1658.53	**90518.37**	IND	OK
Tobramycin	Pancreatic Insufficiency	647	368.44	**72462.81**		Confounder, due to CF
Tobramycin	Cystic Fibrosis	1212	346.65	**64923.85**	IND	OK
Allopurinol	Gout	2778	79.57	**61419.85**	IND	OK
Insulin	Diabetes Mellitus, Insulin-Dependent	6179	32.76	**55082.08**	IND	OK
Furosemide	Congestive heart failure	11955	12.04	**44120.11**	IND	OK
Nitroglycerin	Coronary Arteriosclerosis	10379	17	**42400.06**	IND	OK
Colchicine	Gout	1650	90.31	**40544.75**	IND	OK
Insulin	Diabetes Mellitus	11478	10.59	**36228.16**	IND	OK
Lactulose	Hepatic Encephalopathy	747	116.26	**35601.03**	IND	OK
Aspirin	Coronary Arteriosclerosis	19026	6.83	**35539.91**		Prophylaxis and early RX; IND
Statins	Hyperlipidemia	15536	7.73	**35356.23**	IND	OK
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	OK
donepezil	Dementia	901	96.29	**31183.81**	IND	OK

Results – Top-Ranked Corr. (cont.)

Drug	Finding	cocount	odds	chisq	det	Review by Expert
Nitroglycerin	Chest Pain	9501	11.42	**29787.4**	IND	OK
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	OK
Sinemet	Parkinson Disease	756	115.75	**25794.43**		Multi-component drug; IND
latanoprost	Glaucoma	663	97.64	**24977.36**	IND*	OK
Statins	Coronary Arteriosclerosis	15692	5.34	**24296.85**		IND
mesalamine	Crohn's disease	610	101.51	**23912.73**	IND	OK
Cocaine	Cocaine Abuse	552	98.09	**23906.61**		Trivial
Albuterol	Exacerbation of asthma	2553	30.84	**23675.4**	AE	**Incorrect – IND**
Aspirin	Hypertensive disease	33022	4.51	**23593.69**	IND	**Confounder**, stroke/MI prophylaxis
Hydroxychloroquine	Lupus Erythematosus, Systemic	572	86.91	**23029.24**	IND	OK
mesalamine	Ulcerative Colitis	423	105.89	**22653.9**	IND	OK
Levetiracetam	Seizures	2804	37.2	**22565.16**	IND	OK
Insulin	Diabetes Mellitus, Non-Insulin-Dependent	7624	7.81	**22347.89**	IND	OK
Statins	Hypertensive disease	30117	4.65	**22289.33**	IND	**Confounder**, stroke/MI prophylaxis
tamsulosin	Benign prostatic hypertrophy	1430	31.73	**22267.01**	IND	OK
Insulin	Diabetic Ketoacidosis	2014	47.45	**22213.8**	IND	OK

Results – Rofecoxib

Drug	Finding	cocount	odds	chisq	det
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	Finding	cocount	odds	chisq	det
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**	AE*
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	Finding	cocount	odds	chisq	det
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**	AE
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 – Risperidone

Drug	Finding	cocount	odds	chisq	det
Risperidone	Schizophrenia	308	42.2	**8710.9**	IND
Risperidone	Iron deficiency anemia	[illegible]	[illegible]	[illegible]	
Risperidone	Mental disorders	121	25.9	**2422.33**	IND
Risperidone	Dementia	269	13.01	**2372.91**	IND
Risperidone	HYPOKINESIS GLOBAL	164	16.7	**2039.94**	
Risperidone	Poor historian	136	12.5	**1253.73**	
Risperidone	Hypothyroidism	411	4.84	**919.79**	AE*
Risperidone	Epilepsy	497	4.2	**829.72**	IND
Risperidone	Bipolar Disorder	178	6.12	**661.89**	IND
Risperidone	Abnormal mental state	191	5.42	**594.93**	
Risperidone	Diabetes Mellitus, Insulin-Dependent	202	5.06	**564.57**	
Risperidone	Agitation	210	4.83	**544.8**	IND
Risperidone	Congestive heart failure	327	3.3	**413.96**	
Risperidone	Diabetes Mellitus	444	2.83	**378.33**	AE
Risperidone	Hypovolemia	156	4.48	**374.39**	
Risperidone	Psychiatric problem	169	4.27	**372.97**	
Risperidone	Coronary Arteriosclerosis	455	2.73	**354.89**	
Risperidone	Obesity	393	2.84	**351.31**	AE
Risperidone	Diabetes Mellitus, Non-Insulin-Dependent	329	2.95	**333.44**	IND
Risperidone	Hypertensive disease	937	2.41	**321.51**	AE*
Risperidone	Rhonchi	115	4.62	**297.31**	

Results – Statins

Drug	Finding	cocount	odds	chisq	det
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**	

Informal Analysis Questions

1. Do drug-finding pair classifications derived by applying the DEB to clinical notes appear reasonable most of the time?
2. When DEB classifications appear to be incorrect, is there at least a plausible reason that explains the mistake?
3. Does there appear to be, at face value, a known clinical reason that can be cited to explain those clinical-note-derived drug-finding pairs that are highly statistically correlated?
4. Can the method rank recently discovered drug-AE pairs high enough to suggest future potential for more careful side effect discovery?

Discussion

- There were many significantly correlated drug-finding pairs that seemed reasonable, representing both known AEs or Indications.
- Confounding due to co-morbid conditions and symptoms of a disease was common and must be addressed.
- The DEB pair classifications seemed to be correct in most cases, but there are some obvious mistakes that need to be addressed (e.g. asthma vs. "exacerbation of asthma).

Limitations

- Exploratory study
  - Lack of formal analysis of drug-finding correlations
  - Only reviewed correlations for a few drugs
- Ambiguous nature of finding concepts a problem
- Co-morbidities and confounders problematic
- NLP is sometimes coarse
- Statistical analysis approach must be refined

Conclusion

- The goal of this project was to explore the feasibility of adverse drug effect discovery from data mining on H&P notes.
- We have shown that there are potentially interesting and useful AE signals within a large collection of H&Ps, but more must be done to clearly separate this signal from the noise.