



HealthFacts™ Datawarehouse & Active Pharmacovigilance

**Doug McNair MD PhD
SVP Cerner
LifeSciences**

Advances in Therapeutics Safety Analytics

■ Comprehensive Product and Procedure Vigilance

- ◇ Pharmacovigilance, medical device vigilance, services vigilance, preventive policies vigilance, etc.
- ◇ Business model mainly involves pharmacovigilance, though...

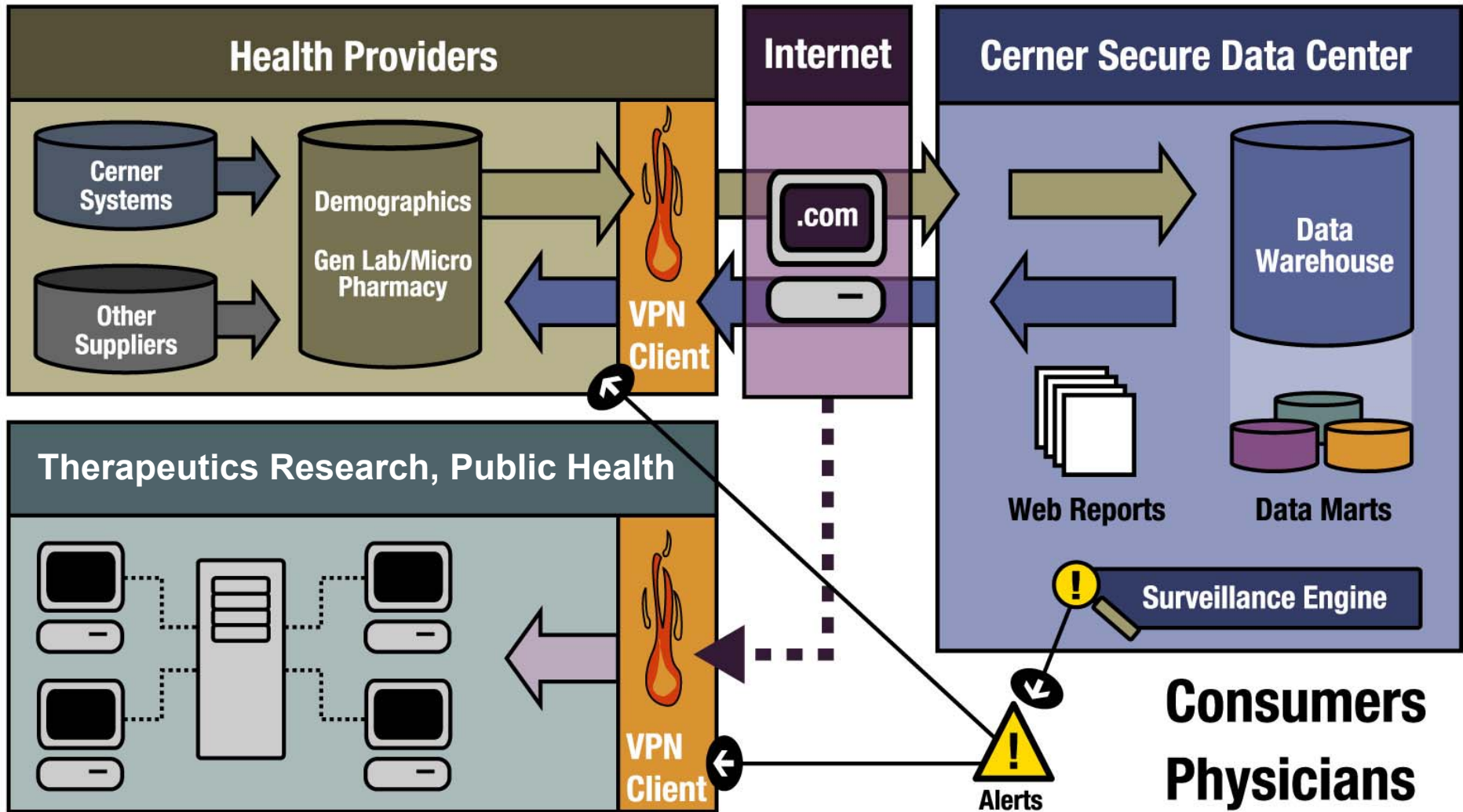
■ Vigilance processes

- ◇ 'Passive' or reactive pharmacovigilance, signal established by mining and analyzing spontaneous ADR reports submitted by humans, coded by humans
- ◇ 'Active' or pro-active pharmacovigilance, signal established by mining EMR-derived warehouses, AE ascription by machine, AE coding mostly by machine

Pro-Active vs. Reactive Vigilance

- Capable of detecting, assessing, trending, and dispositioning risks that have not yet been noticed or reported by humans
- Capable of detecting multi-variate relationships and time-oriented sequential patterns that are too complex for humans to notice
- Higher sensitivity signal-detection than ADRs, but usually less specific
- Not affected by sociological, medicolegal, and other factors that inhibit submission of spontaneous ADRs or confound analysis of reports
- For known AEs and risk signals, passive and active vigilance analytics can be check-and-balance on each other
- Machine learning: pattern recognition requires human supervision and human curation of the machine-identified categories

Cerner's Datamining Discovery/Vigilance Data Assets



**Consumers
Physicians**

HealthFacts contains...

- **Demographics of the patients – age, gender, race, payor, etc.**
- **Diagnosis at time of visit/discharge – principal diagnosis and up to 8 secondary diagnoses**
- **Locations of services/patients (e.g., clinic, ED, ICU, etc.) and the locations where medication and laboratory orders are placed. These can be used to ascertain sequences and patterns of care delivery processes.**
- **Time stamped service (procedures/treatment) records; Health Facts has time minute-wise timestamps for medication and laboratory orders and results.**
- **ICD-9 procedure codes (up to 6 per visit) have date-wise time resolution.**
- **Duration/route/quantity/time of meds administration and other medication uses. Each medication order has the brand/generic name of the medication, a start and stop date to compute duration, the route of administration, and dose quantity.**
- **Specialties of admission/surgical/attending physicians; Health Facts contains the medical specialty of the admitting physician when available as well as the ordering physicians for medication and laboratory orders. Currently the data warehouse does not contain the surgical or attending physician information.**
- **Charge information from the UB-92/CMS1500 billing.**

Overview - HealthFacts

■ Time scope

- ◇ January 2000 – current

■ Data types

- ◇ Pharmacy

Captures the pharmacy orders dispensed by pharmacy

Does not include medications managed by other departments (e.g. Surgery, Radiology)

- ◇ Chemistry, Hematology, UA, Microbiology, Immunology-Serology

Does not include lab sections Blood Bank, Anatomic Pathology, HLA, Genomics at this time

- ◇ UB-92, CMS1500 charges:

Does not contain revenue codes or detail charge master data

- ◇ Diagnoses

- ◇ Procedures

■ Pharmacy

- ◇ Medical specialty for the ordering physician per order if available
- ◇ Type of clinical care provided at the location where the medication order was requested from as well as dispensed from if available
- ◇ Route of administration
- ◇ Frequency of the order
- ◇ SVP, LVP – diluent, volume to infuse, infusion rate
- ◇ Order start, stop, discontinued dates-times
- ◇ Total quantity of doses dispensed per order

Data Elements

■ Laboratory

- ◇ Detail lab procedure name, character or numeric result, unit of measurement.
- ◇ Dates and times for the lab order, drawn, received at the lab, performed, verified, canceled, and completed for each result
- ◇ Medical specialty for the ordering physician per order if available.
- ◇ Type of clinical care provided at the location where the order was placed if available

■ Microbiology

- ◇ Dates and times for order, drawn, received by the lab, canceled, completed for each microbiology order
- ◇ Organism name if there was a positive finding. NULL organism name if nothing found.
- ◇ Medical specialty for the ordering physician per order if available.
- ◇ Type of clinical care provided at the location where the order was placed if available.
- ◇ Type of result, for example, preliminary, final, amended final.
- ◇ Specimen source, for example blood, urine, CSF, sputum, nose.
- ◇ Microbiology susceptibility:
 - The ordered procedure code*
 - The organism species code*
 - The antimicrobials that were used for susceptibility testing.*
 - Numeric MIC or KB interpretation of the susceptibility results.*
 - The specimen source (anatomical site)*
 - The performed and verified dates and times for the numeric and interpretation fields.*

Encounters with Medication, Laboratory and Billing

96% of the Inpatient encounters have medication, general lab or billing.

70% of Emergency and 60% of Outpatient visits have activity in the 3 main subject areas.

Patient type	Encounters w/ Med, Lab, and Diag	Hospitals	% Encounters w/ Med, Lab, and Diag	Encounters w/ med or lab or diag	Hospitals	% / Med or Lab or Diag	% w/ no activity (have "encounter" with nothing else)
Billing			0.00%	70,282	4	84.70%	15.30%
Client	1	1	0.00%	802,372	22	69.96%	30.04%
Clinic	819	9	0.02%	1,624,498	34	45.68%	54.32%
Community	2	1	0.00%	88,280	12	81.28%	18.72%
Day Surgery	1,106	6	1.10%	68,300	15	67.75%	32.25%
Dental			0.00%	25	7	0.21%	99.79%
Dialysis	40	2	0.39%	8,808	15	86.74%	13.26%
Emergency	417,153	49	7.36%	3,917,108	84	69.15%	30.85%
Home Health			0.00%	2,234	5	78.97%	21.03%
Hospice	229	3	14.10%	1,564	5	96.31%	3.69%
Inpatient	1,334,784	56	47.63%	2,683,904	88	95.78%	4.22%
Institution			0.00%	71,380	9	68.88%	31.12%
Laboratory	8	5	0.00%	469,930	25	59.54%	40.46%
Long Term Care (LTC)	159	1	0.46%	29,679	9	85.89%	14.11%
Newborn	4,489	8	11.10%	35,829	16	88.56%	11.44%
Non- diagnostic Outpatient	24	1	75.00%	31	1	96.88%	3.13%
Non-Patient	1	1	0.00%	209,020	11	58.03%	41.97%
Not Mapped			0.00%	60,980	17	83.13%	16.87%
NULL	1	1	0.05%	610	18	32.14%	67.86%
Nursing Home			0.00%	8,186	8	81.35%	18.65%
Observation	3,055	11	3.31%	78,649	18	85.24%	14.76%
Observation / Short stay / 24 hr stay	24,859	22	12.90%	131,939	37	68.45%	31.55%
Obstetrics	3,818	4	17.26%	13,005	8	58.79%	41.21%
Occupational Health			0.00%	2,778	6	14.12%	85.88%
Other Specialty	2,001	3	0.73%	96,416	27	35.30%	64.70%
Outpatient	88,842	43	0.46%	11,610,929	89	60.46%	39.54%
Outpatient Surgery	48,405	19	9.38%	348,263	34	67.50%	32.50%
Physician Business			0.00%	41,616	3	63.64%	36.36%
Preadmit	31,336	21	6.35%	318,803	64	64.60%	35.40%
Quality Control			0.00%	769	4	50.56%	49.44%
Radiology	357	2	0.38%	4,397	5	4.72%	95.28%
Recurring	13,604	20	1.73%	335,605	55	42.57%	57.43%
Research			0.00%	30	2	20.83%	79.17%
Schedules	9	2	0.01%	3,449	6	2.29%	97.71%
Series	334	3	0.23%	73,890	9	50.00%	50.00%
Skilled Nursing Facility	1,197	6	8.24%	10,866	15	74.85%	25.15%
Test			0.00%	1	1	33.33%	66.67%
Unknown / Invalid	5,313	11	0.40%	756,917	40	57.43%	42.57%
Urgent			0.00%	6,105	2	29.24%	70.76%
Totals	1,981,946	56	5.17%	23,987,447	97	62.61%	37.39%

ICD-9 Diagnoses and Procedures by Patient Type

Green 4+	Blue 2.25-3.9	Red 1-2.25	Min Diagnosis	Max Diagnosis	Avg Diagnoses / Encounter	Encounters w/ Diagnoses	Patient Type	Min Procedure	Max Procedure	Avg Procedures / Encounter	Encounters w/ procedures	Green 2.01+	Blue 1.5-2	Red 0-1.5
Outpatient			1	9	1.61	7,697,961	Outpatient	1	6	1.97	613,812			
Emergency			1	9	2.31	2,198,476	Emergency	1	6	1.62	504,871			
Inpatient			1	9	5.18	1,675,974	Inpatient	1	6	2.34	1,095,667			
Unknown / Invalid			1	9	2.00	246,659	Unknown / Invalid	1	6	1.42	53,135			
Recurring			1	9	1.66	232,373	Recurring	1	6	2.25	11,405			
Preadmit			1	9	2.29	200,433	Preadmit	1	6	1.67	165,868			
Outpatient Surgery			1	9	2.88	164,751	Outpatient Surgery	1	6	1.71	162,201			
Non-Patient			1	9	1.82	102,401	Non-Patient	1	6	1.94	106			
Clinic			1	9	2.12	85,401	Clinic	1	6	1.61	19,252			
Other Specialty			1	9	2.28	74,494	Other Specialty	1	6	1.39	7,986			
Observation / Short stay / 24 hr stay			1	9	3.22	69,744	Observation / Short stay / 24 hr stay	1	6	2.05	47,974			
Client			1	9	1.55	68,210	Client	1	6	1.62	191			
Laboratory			1	9	2.22	39,454	Laboratory	1	6	2.08	274			
Series			1	9	1.37	36,741	Series	1	6	2.20	126			
Newborn			1	9	2.22	18,569	Newborn	1	6	1.34	8,746			
Obstetrics			1	9	3.36	9,087	Obstetrics	1	6	2.31	7,994			
Not Mapped			1	9	3.81	9,083	Not Mapped	1	6	1.79	4,166			
Observation			1	9	2.76	7,992	Observation	1	6	1.73	5,607			
Day Surgery			1	9	3.12	6,668	Day Surgery	1	6	1.96	6,359			
Schedules			1	9	1.52	3,358	Schedules	1	3	1.32	720			
Radiology			1	9	3.12	2,538	Radiology	1	6	2.11	2,524			
Long Term Care (LTC)			1	9	3.25	2,524	Long Term Care (LTC)	1	4	1.38	48			
Skilled Nursing Facility			1	9	7.55	1,478	Skilled Nursing Facility	1	6	2.83	767			
Hospice			1	9	1.13	1,299	Hospice	1	4	2.07	14			
Community			1	9	3.32	394	Community	1	6	3.09	347			
Billing			1	9	3.07	339	Billing	1	6	2.13	8			
Urgent			1	9	2.98	114	Urgent	1	4	1.20	98			
Occupational Health			1	9	2.65	97	Occupational Health	1	4	1.62	42			
Dialysis			1	9	4.64	75	Dialysis	1	6	2.58	40			
NULL			1	9	6.51	43	NULL	1	6	2.82	22			
Nursing Home			1	9	7.44	36	Nursing Home	1	5	2.40	35			
Non- diagnostic Outpatient			1	9	3.23	31	Non- diagnostic Outpatient	1	6	3.45	31			
Institution			1	9	5.78	9	Institution	1	6	3.17	6			
Dental			1	9	6.00	2	Dental	1	1	1.00	1			
Home Health			1	5	3.50	2	Home Health	1	1	1.00	1			
Quality Control			1	7	5.50	2	Quality Control	0	0	-	-			
Total Encounters w/ Diagnoses						12,956,812	Total Encounters w/ procedures				2,720,444			

Total Encounters: 35,836,796 Encounters with Billing: 13,178,959

2,720,444



Encounter Summary

(Visit Types)

2000-July 2006	Encounters	Facilities **	Contributors	Distinct Patient/Patient Group
Emergency	5,664,403	86	42	3,244,183
Inpatient (2)	2,842,758	88	42	1,861,170
Clinic (2)	3,576,770	37	16	698,690
Outpatient (8)	22,148,468	95	42	6,586,442
Other (24)	4,082,380	91	42	1,930,276
Total	38,314,926	97	42	10,516,678

Emergency

Inpatient – Inpatient, Newborn

Clinic – Clinic, Urgent

Outpatient -Client, Day Surgery, Dialysis, Laboratory, Non-Patient, Obstetrics, Outpatient, Outpatient Surgery

Other - Billing, Community, Dental, Home Health, Hospice, Institution, Long Term Care (LTC), Non- diagnostic Outpatient, Not Mapped, NULL, Nursing Home, Observation, Observation / Short stay / 24 hr stay, Occupational Health, Other Specialty, Physician Business, Preadmit, Quality Control, Radiology, Recurring, Schedules, Series, Skilled Nursing Facility, Unknown / Invalid

Encounters Over Time

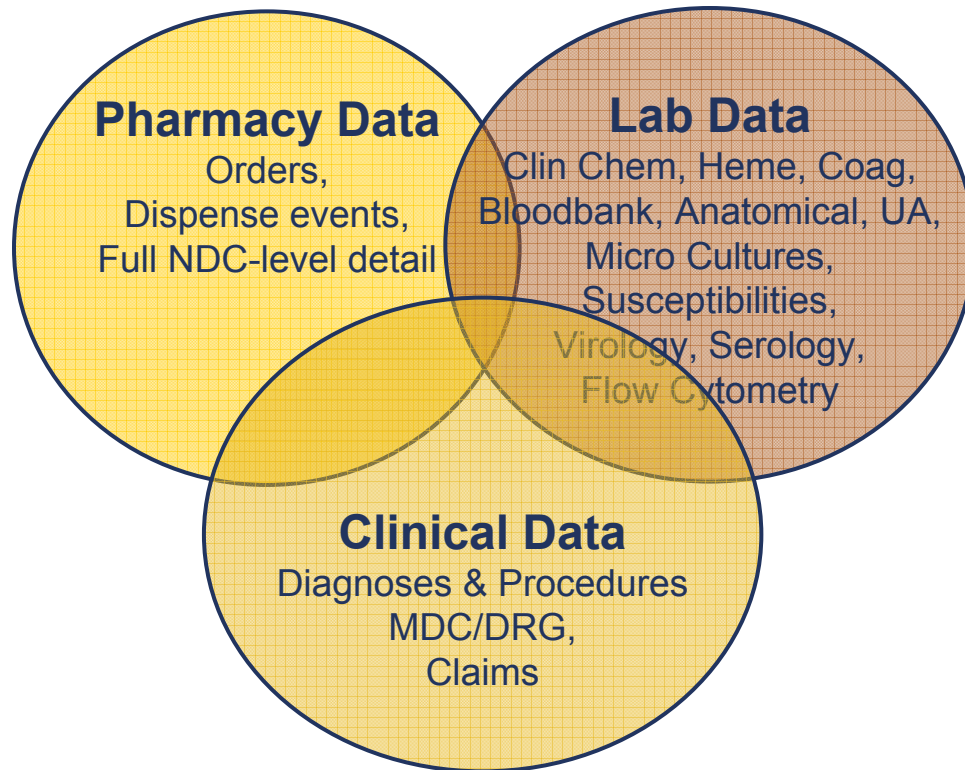
Demonstrates how the Health Systems change over time

Year	All Encounters	Patients	Hospitals	Health Systems
2000	4,173,618	2,015,331	38	23
2001	6,918,645	2,903,877	40	24
2002	6,316,161	2,791,647	47	26
2003	5,065,827	2,245,758	46	25
2004	5,155,186	2,053,912	56	26
2005	6,237,834	2,421,473	70	32
2006	4,447,655	2,139,118	74	31
Total thru July 2006	38,314,926			

Demonstrates how facilities with Inpatients change over time and how many have the three main data sources

Year	Inpatients w/ Med or Lab or Diag	Facilities	% of Inpt	Inpatients w/ Med & Lab & Diagnosis	Facilities	% of Inpt
2000	277,337	26	93.21%	96,498	15	32.43%
2001	460,720	37	95.92%	240,173	23	50.00%
2002	469,555	40	95.99%	239,211	30	48.90%
2003	377,943	42	96.36%	231,368	35	58.99%
2004	345,275	47	95.27%	207,358	36	57.21%
2005	428,242	57	96.80%	245,576	38	55.51%
2006	324,518	65	95.22%	74,598	35	21.89%
	2,683,590	89	95.68%	1,334,782	56	47.59%

HealthFacts™ -- 70+ U.S. Source Institutions



- Daily datafeeds
- Anonymized EMR-level detail result info, not just orders or claims
- Longitudinal tracking of patients (eMPI)
- Date-time stamped to minute-wise resolution
- Time series information to help analyze sequences of events
- Interactions between concomitant meds
- Interactions between comorbid clinical conditions

Rofecoxib, Celecoxib, Naproxen

Vioxx

	AMI/CVA +	AMI/CVA -	
Rx +	2,894	35,458	38,352
Rx -	203,947	18,717,005	18,920,952
	206,841	18,752,463	18,959,304

Chi-square= 14800, p < 0.0001
SRR = 6.9 (6.7 - 7.2)

Celebrex

	AMI/CVA +	AMI/CVA -	
Rx +	3,002	22,093	25,095
Rx -	203,839	18,730,370	18,934,209
	206,841	18,752,463	18,959,304

Chi-square= 27500, p < 0.0001
SRR = 11.0 (10.6 - 11.4)

naproxen

	AMI/CVA +	AMI/CVA -	
Rx +	1,215	18,741	19,956
Rx -	205,626	18,733,722	18,939,348
	206,841	18,752,463	18,959,304

Chi-square= 4620, p < 0.0001
SRR = 5.6 (5.3 - 5.9)

Rofecoxib, Celecoxib, Naproxen

Vioxx	Exp +	Exp -	
Rx +	11	38,341	38,352
Rx -	1,401	18,919,551	18,920,952
	1,412	18,957,892	18,959,304

Chi-square= 20.5, p < 0.001
SMR = 3.8 (2.1 - 6.9)

Celebrex	Exp +	Exp -	
Rx +	20	25,075	25,095
Rx -	1,392	18,932,817	18,934,209
	1,412	18,957,892	18,959,304

Chi-square= 167, p < 0.0001
SMR = 10.7 (6.9 - 16.5)

naproxen	Exp +	Exp -	
Rx +	3	19,953	19,956
Rx -	1,409	18,937,939	18,939,348
	1,412	18,957,892	18,959,304

Chi-square= 0.7, p > 0.4
SMR = 2.0 (0.4 - 6.2)

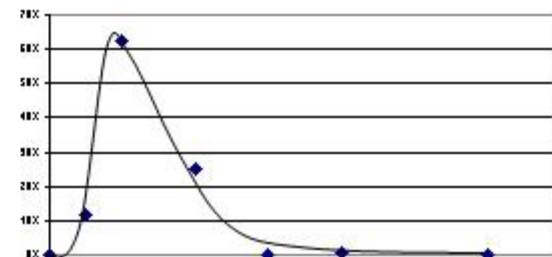
Dosing – 24h Exposure

■ Exposure = Strength X Dose_Qty X Frequency

e.g., “Vioxx 12.5 mg ii P.O. bid” = 12.5 X 2 X 2 = 50 mg/24h

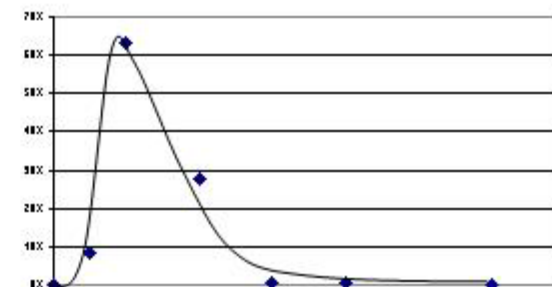
VIOXX DOSE_QTY=1	FREQ STRENGTH	QD 1	BID 2	TID 3	QID 4
12.5 mg		3216	407	5	2
25 mg		16677	1040	36	7
50 mg		5801	77	2	1

mg/24h	N	Prevalence
12.5 mg	3216	11.8%
25 mg	17084	62.7%
50 mg	6843	25.1%
75 mg	36	0.1%
100 mg	84	0.3%
150+ mg	3	0.0%



CELEBREX DOSE_QTY=1	FREQ STRENGTH	QD 1	BID 2	TID 3	QID 4
100 mg		1594	2960	40	4
200 mg		8944	5143	55	7
400 mg		99	59	1	0

mg/24h	N	Prevalence
100 mg	1594	8.4%
200 mg	11904	63.1%
400 mg	5246	27.8%
600 mg	55	0.3%
800 mg	66	0.3%
1200+ mg	1	0.0%



... Do the same totals for Dose_Qty = 2, 3, ...

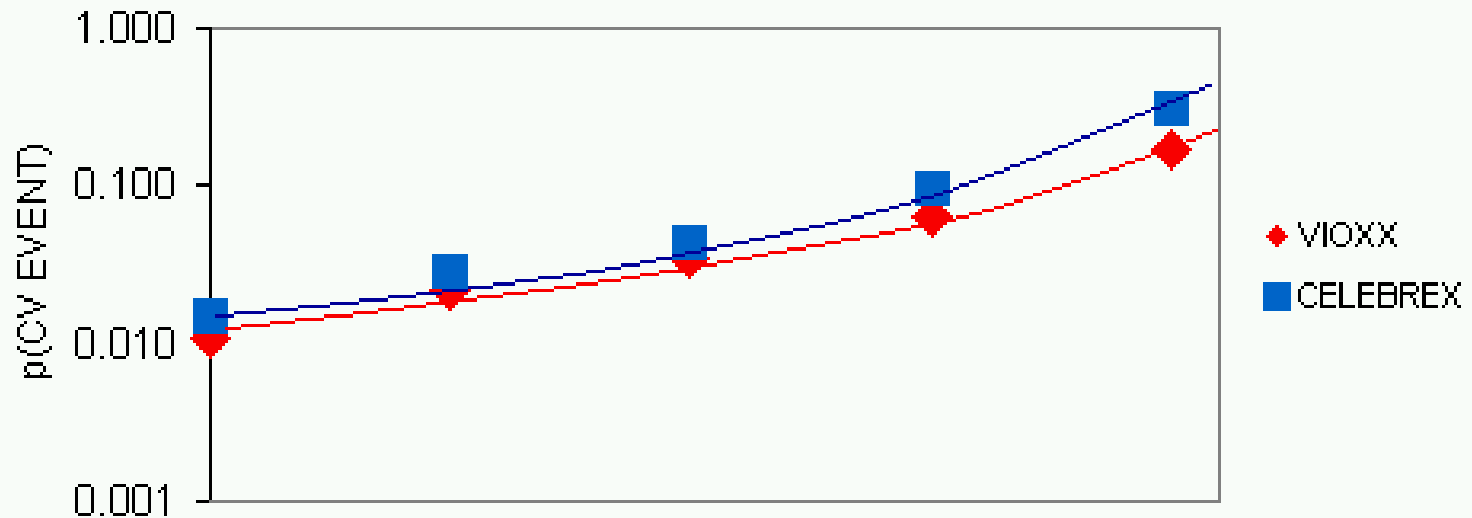
Ischemic Cardiac and CNS Events: They probably “shot the wrong horse”

Probit Regression

coeffs
intercept
b(dose)

VIOXX	CELEBRX
-2.1044	-2.1022
0.1444	0.1989

log DOSE (mg/kg)



HealthFacts™ 2000-2006 Oncology Case Counts

		Encounters	Patients
■ Breast Ca	-	85,237	38,427
■ Prostate Ca	-	65,164	32,109
■ Lung, NOS	-	57,292	25,815
■ CRC	-	43,306	20,833
■ NHL	-	43,054	17,898
■ MM	-	14,535	5,511
■ MDS	-	12,670	5,955
■ CLL	-	11,234	5,508
■ RCC	-	10,254	6,251
■ Glioma	-	8,161	4,063
■ Melanoma, NOS	-	7,284	4,372
■ Hodgkins	-	7,216	3,049
■ AML	-	6,355	2,248
■ ALL	-	5,824	1,862
■ H&N Ca	-	5,696	3,124
■ CML	-	2,978	1,298
■ Melanoma, mets	-	971	442
■ MF cutaneous T-cell	-	742	434
■ GIST	-	88	62
■ Sezary	-	37	24

Example: Classical Endpoints in Oncology

- **Survival**
- **Response Rate (RR)**
- **Time to progression (TTP)**
- **Disease-free Survival (DFS)**

balance of benefits, in the context of toxicities ...

Endpoints for Regulatory Approval in Cancer

- US & Canada** Overall survival in randomized trials is the accepted standard for all recent approvals
- Europe** An intermediate position: randomized trials usually required, but other endpoints (i.e. RR, TTP and QoL) and/or methodologies (i.e. meta-analyses, observational databases and registries) utilized
- Japan** Objective response rate in non-randomized trials (with independent committee assessment) for all recent approvals

Therapeutics Amenable to Study with HealthFacts™

■ Targeted therapies and translational research

- ◇ Concurrent development of biomarker diagnostics (predictive of efficacy, predictive of tox) with therapeutics
- ◇ Carve-out populations and personalized marker-based Rx identified, quantified, and validated by datamining in controlled observational datawarehouse

■ Cost-effectiveness endpoints

■ “Minimal clinical effectiveness” for regulatory approval

■ Dose-dense regimens

■ Disease-progression endpoints

- ◇ http://www.fda.gov/cder/drug/cancer_endpoints/default.htm

Cerner Observational EMR-Derived Datawarehouses

■ Our datawarehouses derived from EMRs have:

- ◇ HIPAA-compliant deidentification/anonymization
- ◇ Longitudinal record-linkage (eMPI) across visits / encounters / admissions
- ◇ Cross-walked version-controlled nomenclature codesets
- ◇ Detail comprehensive data to support in silico discovery and development, as well as post-market brand management, medical affairs, active pharmacovigilance, and other uses
- ◇ 60+ U.S. domestic institutions, daily feeds (OLAP ETL), > 3TB and growing
- ◇ Ambulatory clinic as well as ED and inpatient cases
- ◇ Data-cleaning and QA
- ◇ Can create sponsor-specific extracts and datamarts

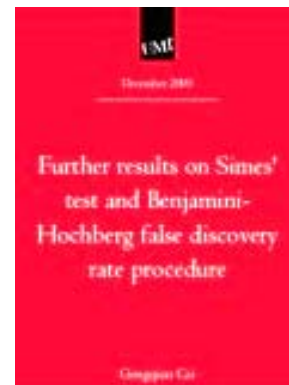
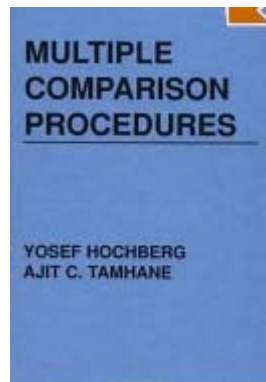
■ Characterization of comorbid dx's and concomitant meds

■ All items are date-time stamped

- ◇ Most are minute-wise granularity
- ◇ We can evaluate sequences of events, exposures (“process”)
- ◇ ‘Administrative/claims’ datawarehouses CANNOT do this!

False Discovery Rate (FDR) Control

- Machine-learning based signal-detection and pattern-recognition algorithms must be managed
- The more variables you monitor, the more likely you are to detect a 'signal'
- Analogy to genomics microarray chips (e.g., Affymetrix)
- False-Discovery Rate (FDR): 'false-positive' Type I error
- Benjamini-Hochberg, Tibshirani, other methods for FDR control



False Discovery Rate Control

- Introduced by Benjamini and Hochberg (1995)
- Want to control the number of incorrect rejections, V , as a proportion of the total number of rejections, R

$$\text{FDR} = E\left(\frac{V}{R} \mid R > 0\right) \Pr(R > 0)$$

- Stepwise p-value adjustment guarantees: $\text{FDR} \leq \alpha$
- Finner and Roters (2001) showed that: $\text{FDR} = \alpha\pi_0$
where π_0 is the proportion of true null hypotheses.
- FDR is expectation, so control is “on average”

Adaptive Control of FDR

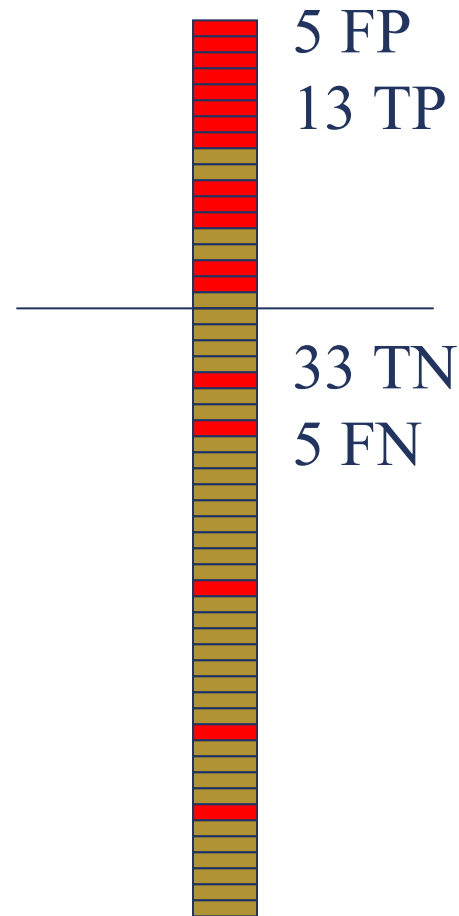
- **Control of the FDR at level α requires estimation of the proportion of true null hypotheses, π_0**
- **In the RiskExplorer™ pharmacovigilance setting, this is the proportion of medications which do *not* exhibit non-proportional AE**
- **In the genomics microarray setting, this is the proportion of genes on the array which do *not* undergo differential expression**
- **Estimation of this quantity is not straightforward. Although Storey (2002) and Storey and Tibshirani (2003) have proposed methods for this, they can produce biased estimates (Black, 2004).**
- **Newton et al. (2003) demonstrated that their Bayesian approach provides unbiased adaptive FDR control**

Multiple-Testing Corrections

- On an array of 10,000 spots, a p-value of 0.0001 may not be significant
- On an analysis of 100,000 combinations of clinical and medications regimen variables, a p-value of 0.00001 may not be significant
- Bonferroni correction: divide your p-value cutoff by the number of measurements
- For significance of 0.05 with 10,000 spots, you need a p-value of 5×10^{-6} , according to Bonferroni method
- Bonferroni is excessively conservative because it assumes that all variables are perfectly statistically and causally independent, which is not true

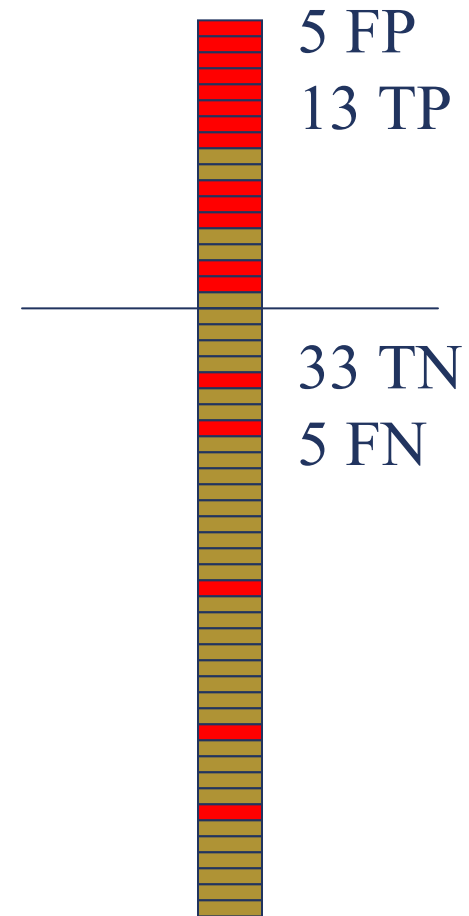
Types of Statistical Errors

- **False positive (Type I error, FDR):** the evidence suggests that a signal is present, but it actually is not
- **False negative (Type II error):** a signal is present, but the evidence and analytics failed to indicate it
- Typically, researchers are more concerned about false positives
- Without doing many (expensive) replicates, there will always be some false negatives
- Regulators and public health officials tend to be more concerned about false negatives



False Discovery Rate

- The false discovery rate (FDR) is the percentage of AEs above a given position in the ranked list that are expected to be false positives
- **False positive rate:** percentage of non-proportionately incident AEs that are flagged
- **False discovery rate:** percentage of events that are 'proportional' (not different from expected PRR rate) but are flagged as signal



$$\text{FDR} = \text{FP} / (\text{FP} + \text{TP}) = 5/18 = 27.8\%$$

$$\text{FPR} = \text{FP} / (\text{FP} + \text{TN}) = 5/38 = 13.2\%$$

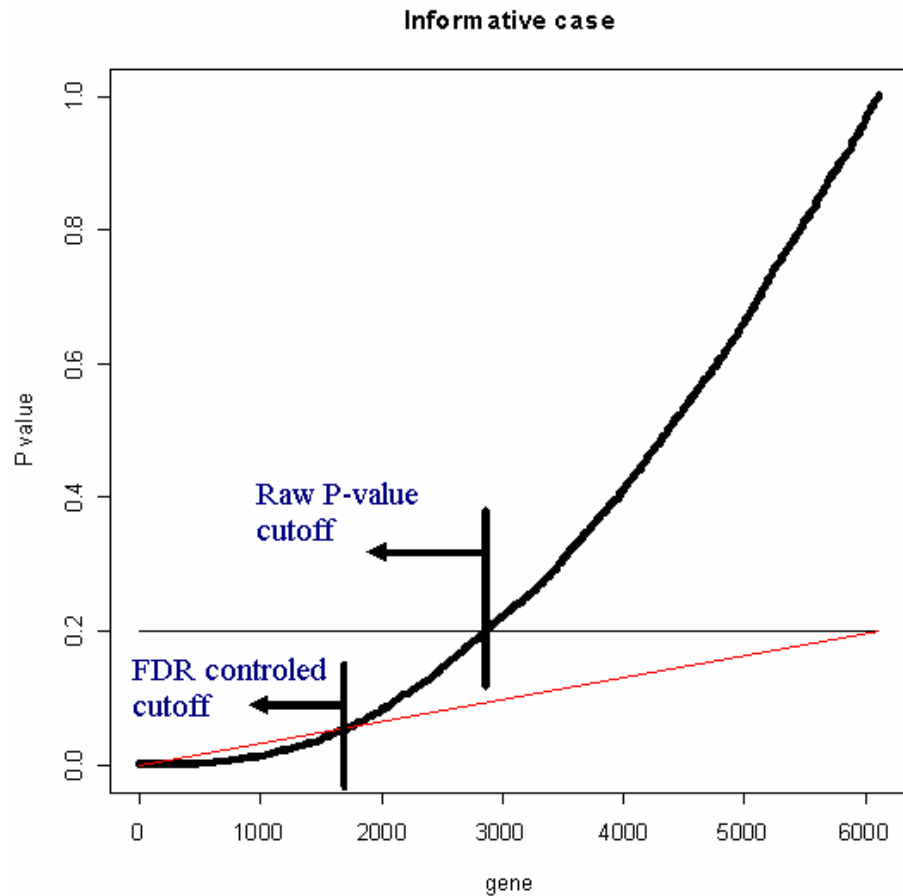
Controlling the FDR

- Order the unadjusted p-values $p_1 \leq p_2 \leq \dots \leq p_m$.
- To control FDR at level α ,

$$j^* = \max \left\{ j : p_j \leq \frac{j}{m} \alpha \right\}$$

- Reject the null hypothesis for $j = 1, \dots, j^*$.
- This approach is conservative if many genes are differentially expressed.

FDR-adjusted P-value



- Sort the variables' raw P-values
- Reject the test with the i -th smallest P-value if $P_i < f/i$, where f is a pre-defined FDR level

Consider the Sets of Variables, not Individual Ones

- **Many variables are cross-correlated with each other**
- **Many diagnoses-conditions-diseases are co-associated**
- **Many treatments are cross-correlated, too**
- **Associations with age, gender, venue, procedures, other variables**
- **We have prior knowledge to group genes together**
 - ◇ GO terms, CBO terms and 'facets'
 - ◇ Metabolic pathways
 - ◇ Functional group
 - ◇ Clustering
- **Design statistics on sets of variables, not individual genes or individual drugs, between conditions**
- **Combine weak information of variables in the datasets, and reduce the number of hypothesis tests**

Active PV and Genomics FDR Analogies

- In linkage studies, a LOD score (log-odds) of 3 or more is usually taken as significant. Why?
- Morton (1955) used the following calculation. Define the posterior Type I error rate (PER) as the probability that a locus is NOT linked, given that it is called 'significant'

nominal significance level

$$\text{PER} = \Pr(L = 0 | D = 1) = \frac{\alpha \Pr(L = 0)}{\alpha \Pr(L = 0) + \pi \Pr(L = 1)}$$

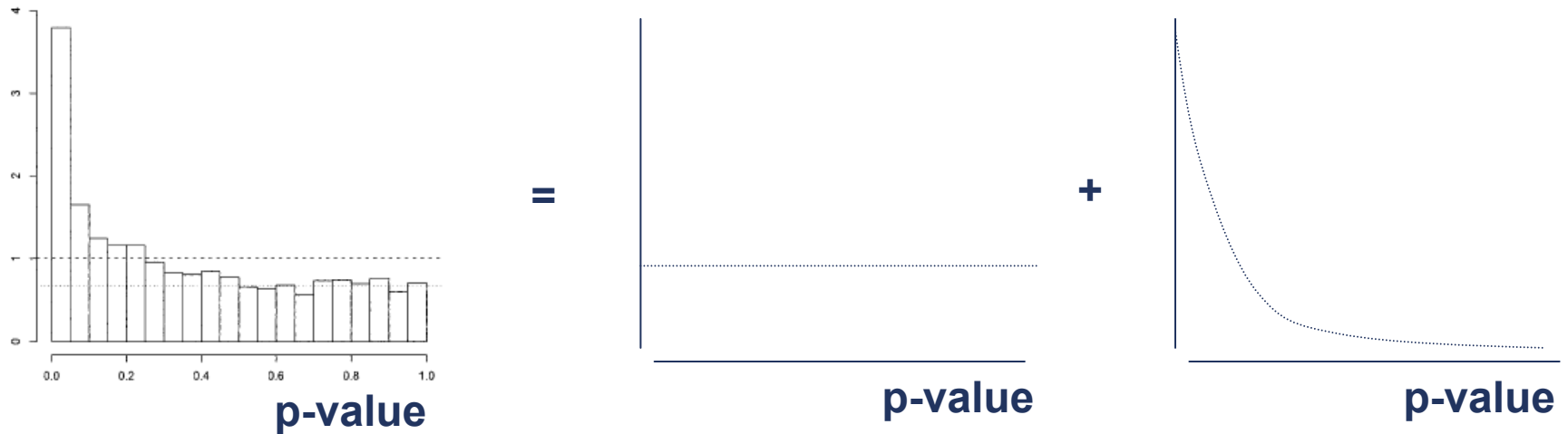
Fernando et al (2004)

average power of test

- In humans the prior probability that a random biomarker is within detectable linkage of the trait locus is 0.02; i.e. $\Pr(L = 1) = 0.02$
- Suppose we have power (π) of 90% (perhaps rather optimistic). To give a PER of 5%, we need to set $\alpha = 0.001$ (i.e. a LOD score of 3)

FDR Control Methods

- An estimate of $E[S(t)] = S(t)$; i.e. the number of p-values less than or equal to a threshold, t
- We estimate $E[F(t)]$ by exploiting our knowledge that null p-values are uniformly distributed

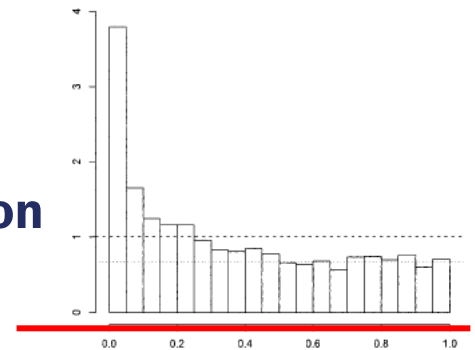


FDR Control 'Rubber' Meets the Road

- We want to estimate π_0 ; the proportion of truly null features among all features (of which there are m)
- We make the assumption that p-values greater than some tuning parameter λ , only represent truly null features
- An estimate of π_0 is

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \dots, m\}}{m(1 - \lambda)}$$

- In effect, this is like top-slicing the p-value distribution



Limitations/Criticisms of FDR Control

- **Accurate estimation of FDR relies on a reasonable proportion of all features being ‘alternative’**
 - ◇ Otherwise estimation of π_0 becomes tricky
- **FDR effectively models the distribution of p-values under the null and the alternative**
 - ◇ Why not directly model the observations themselves under the null and alternative?
 - ◇ Would allow more detailed statements about the alternative
 - ◇ Improvements in FDR estimation (e.g. Jain et al 2005 for microarrays) can come from more detailed modelling of data (e.g. variance as a function of expression level)
- **Systematic biases in p-value distributions can arise from poor modelling, rather than particular ‘alternative’**
 - ◇ e.g. rejection of coalescent ‘null’ distribution does not necessarily imply selection
 - ◇ e.g. permutation assumes exchangeability under the null, which may not be true (stratification)

A HealthFacts™ Example

94 p-values computed from the 274 variables in data on 30,471 exposures to a drug were less than or equal 0.002

The average number of p-values less than or equal to 0.002 in 2199 permuted datasets was 10.3

An initial estimate of FDR is $10.3/94 \approx 10.9\%$

This is 'too high', from pharma-sponsor's perspective

This is 'acceptable', from regulator's and tort lawyers' view

Estimating the False Discovery Rate (FDR)

The initial estimate of 10.9% is too high because the estimate is based on assuming that all null hypotheses H_0 are true

Estimating the False Discovery Rate (FDR)

488 p-values computed from the observed data were greater than or equal to 0.9

The average number of p-values greater than or equal to 0.9 in the 2199 permuted data sets was 827.7

Thus we estimate that $488/827.7 \approx 59\%$ of the null hypotheses are true in our dataset

Estimating the False Discovery Rate (FDR)

Our revised estimate of the FDR for $p \leq 0.002$ is

$$\frac{(10.273) \frac{488}{827.699}}{94} \approx 6.4\%$$

Active PV Concluding Remarks

- **Any discovery and machine-learning app in Datamining and Bioinformatics must deal with high-dimensionality and multiple-comparisons corrections**
- **Thresholding for selecting the findings worthy of attention and monitoring is context-dependent, not one-size-fits-all**
- **FDR is well-established approach for multiplicity correction**
- **Demonstrated useful properties of procedures**
 - ◇ Theory for FDR control is well-established in the literature
 - ◇ Acceptance in epidemiology and bioinformatics > 10 years
 - ◇ Still opportunities for authoring new intellectual property for Cerner in this space
- **Anonymized observational EMR-derived datawarehouses with longitudinal linkage capability are vital assets**
 - ◇ Able to do 24hr and cumulative exposures for multivariate dose-response/dose-tox modeling