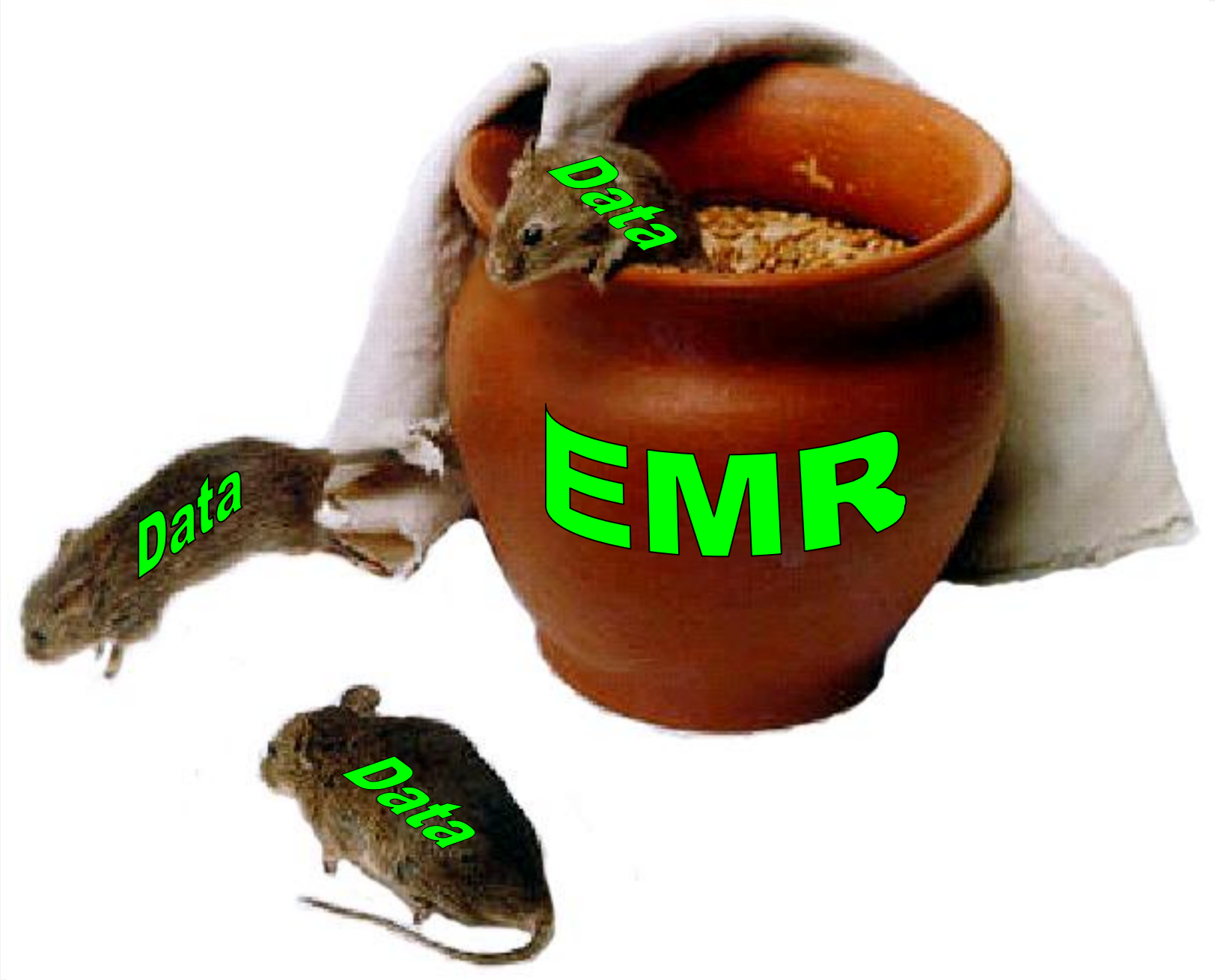
Characteristics, challenges, and determinants of data quality

J. Marc Overhage, MD, PhD
Chief Medical Informatics Officer
Siemens Health Services


Data
EMR
Data
Data

Using Data from Care Process

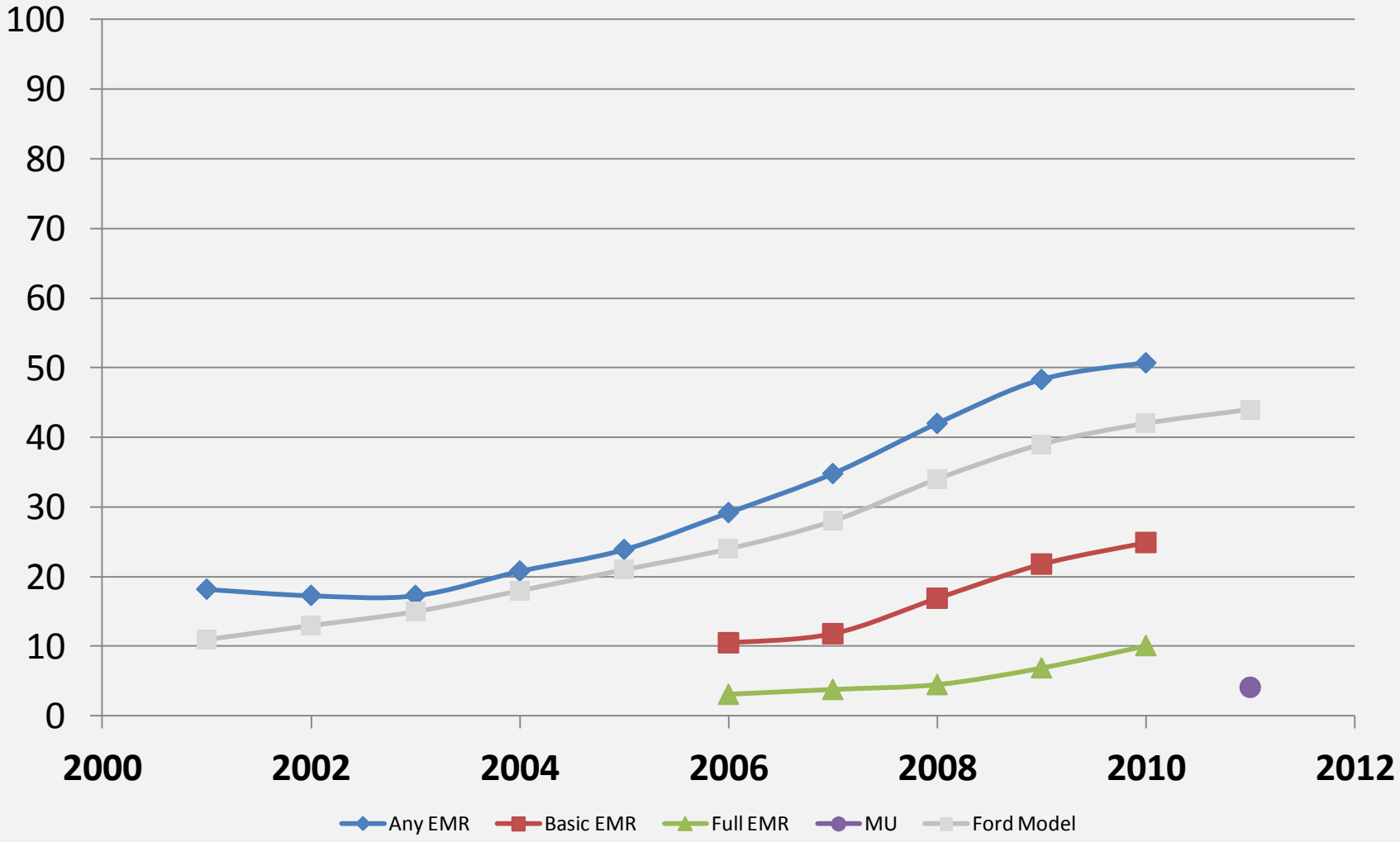
- Benefits from readily available data
- But…
  - Data may be incomplete
  - Data may lack detail
  - Data may be biased
  - Data may be incomparable

Seeking a balance

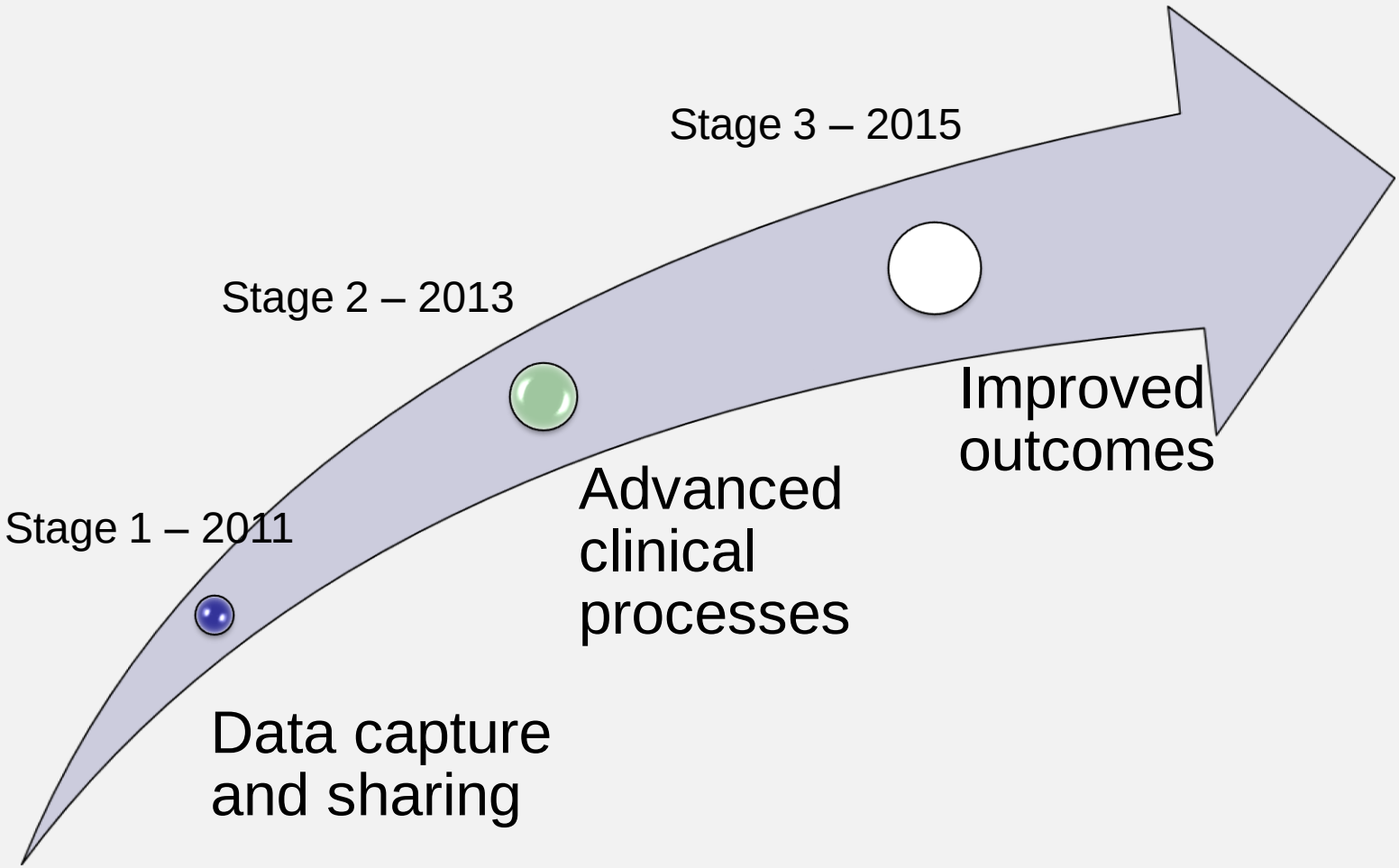
- Data from the clinical care process often not fit for reuse
- Dedicated data collection costly or impossible
- Recording “everything” about “everyone” is impossible
- How to collect data in the primary care process that can be reused with minimal drawbacks (e.g., bias, detail)?


Provider
Payor

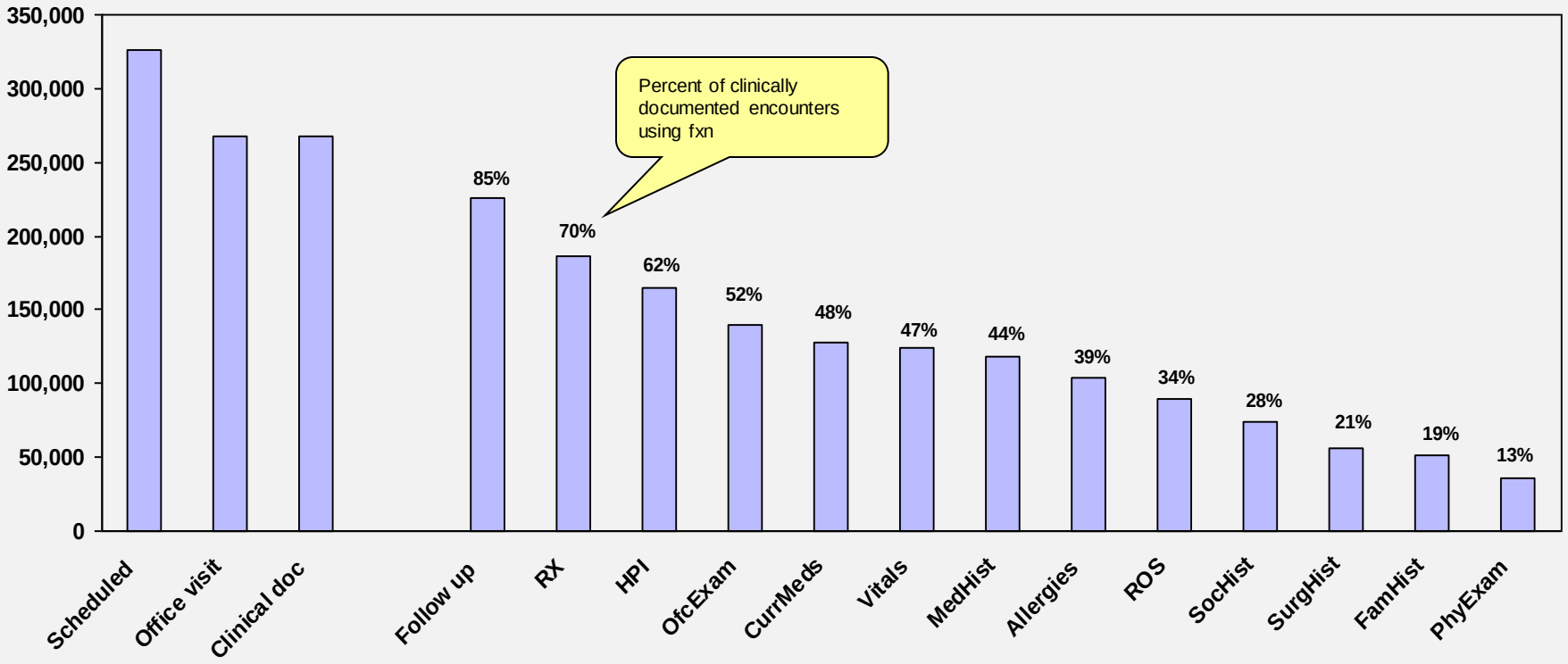
Ambulatory EMR Adoption



Meaningful Use (original)



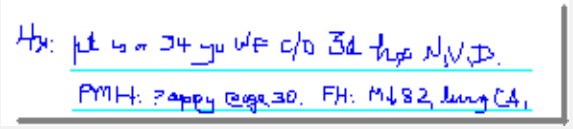
Utilization of Available Functionality



67 Practices Representing 189 Clinicians

MAeHC 2007

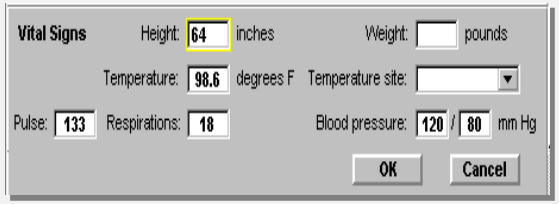
Challenges in Data Capture



- Images

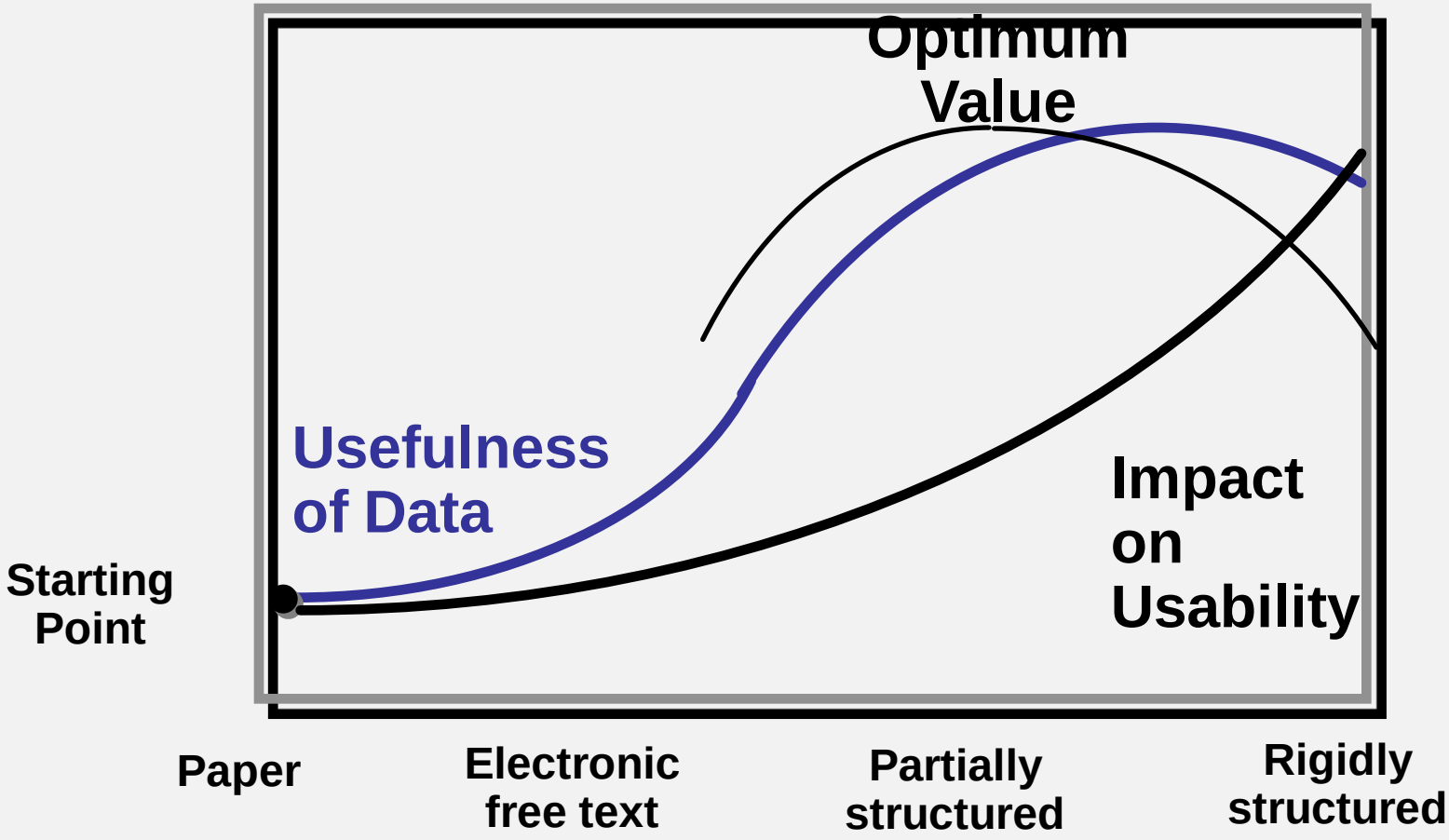
HPI: Patient is a 38 year old white female complaining of a 3 day history of nausea, vomiting and diarrhea.
PMH: questionable appendectomy
FH: mother died at age 82 of lung

- Narrative text (labeled)



- Structured data

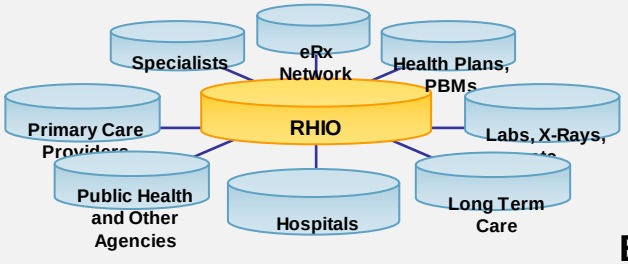
Cost-Value Tradeoff



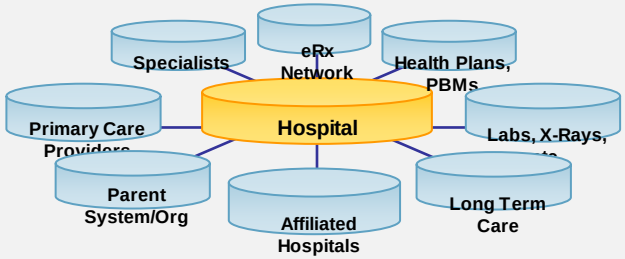

S-5

HIE Diversity

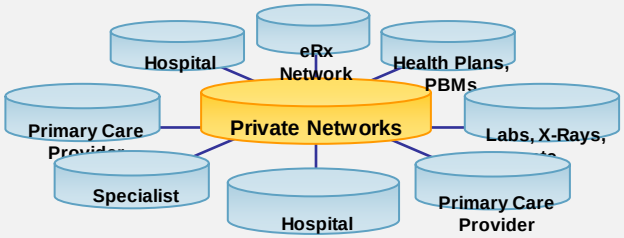
Community/Population Centric



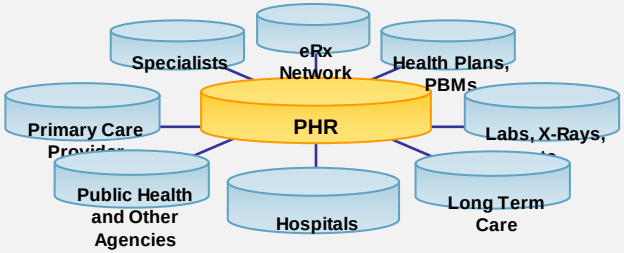
Provider Centric



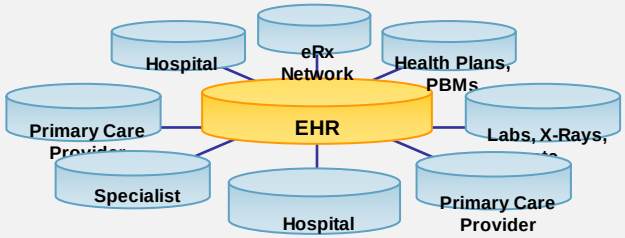
Emerging Private Service Providers and Networks



Person Centric

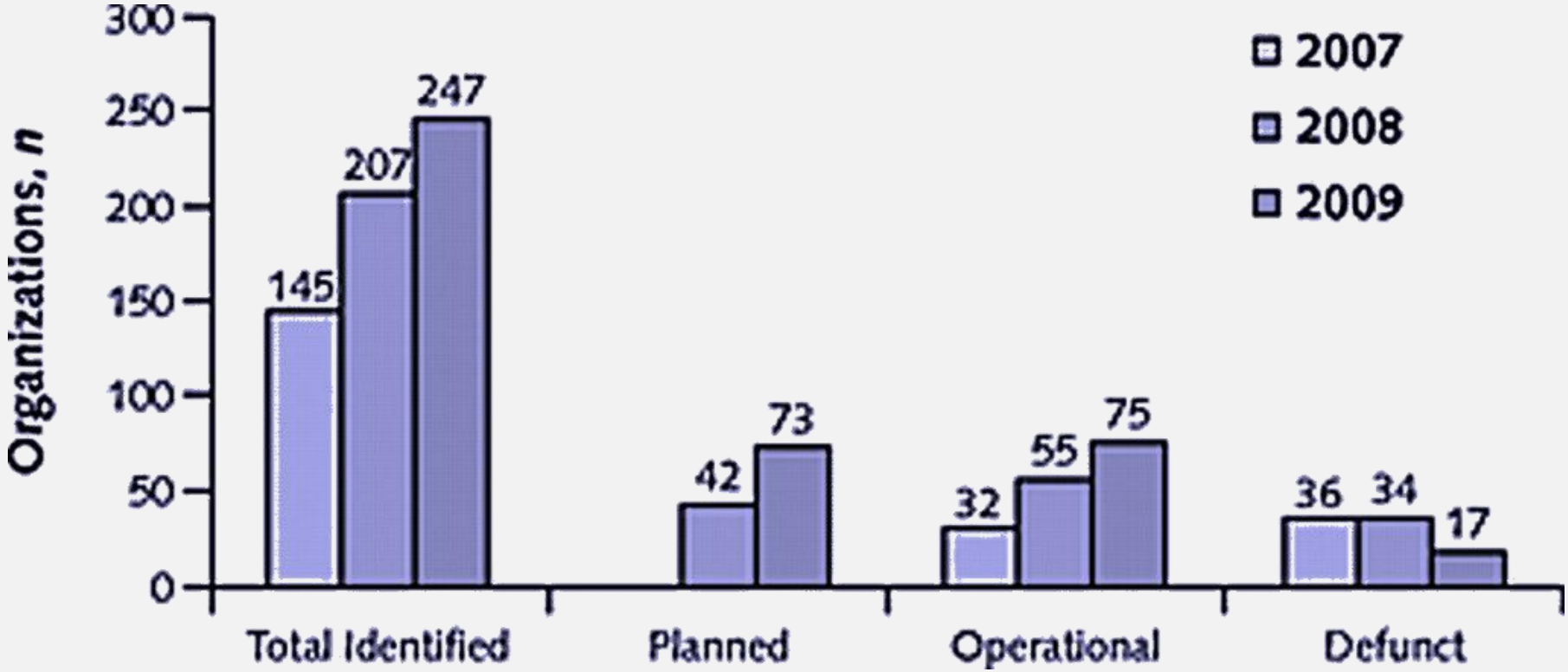


EHR Vendor Centric



* *Source:* The National Alliance for Health Information Technology Report to the Office of the National Coordinator for Health Information Technology on Defining Key Health Information Technology Terms, April 28, 2008

Community HIE Growth

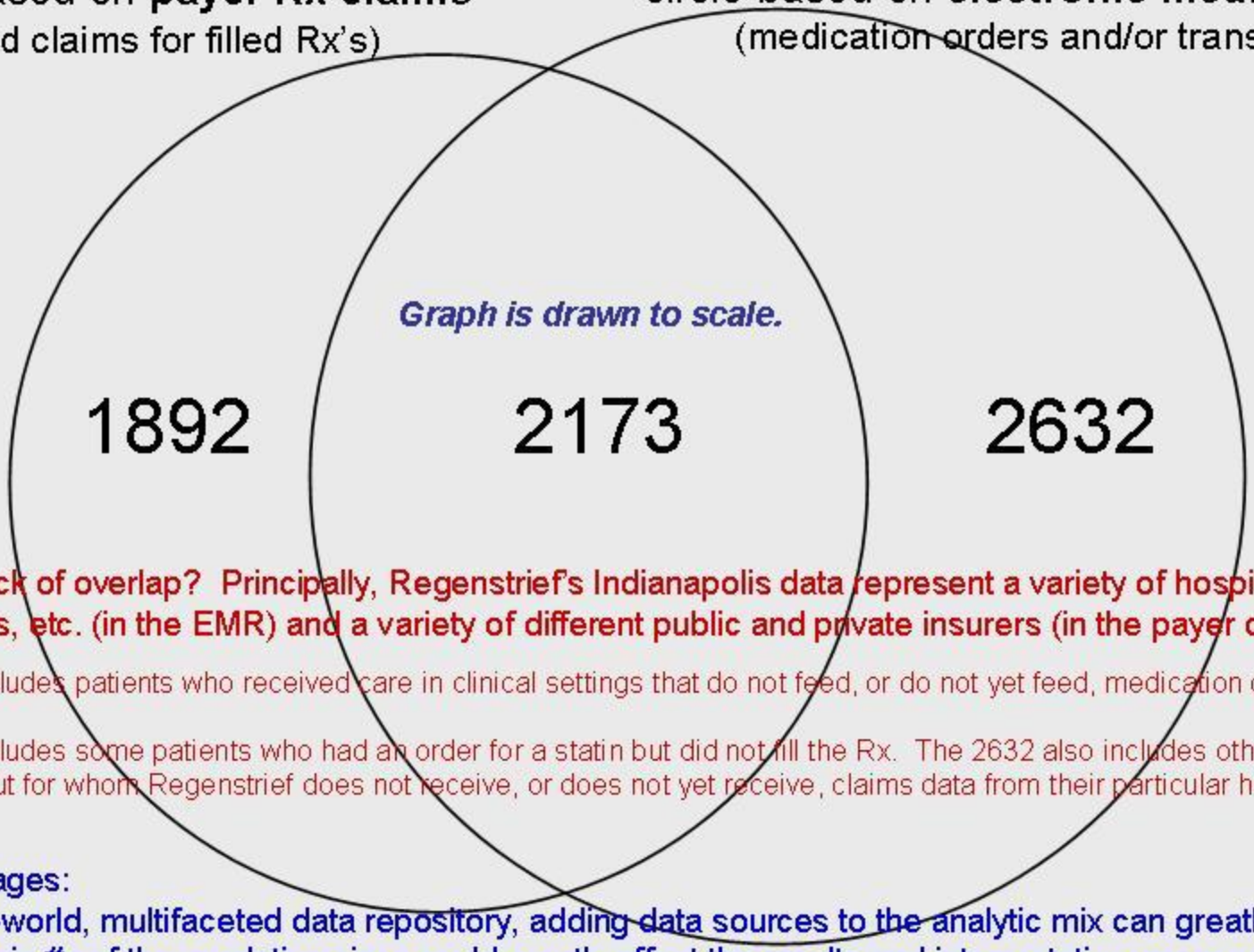


Adler-Milstein J et al. Ann Intern Med 2011;154:666-671

N of Patients with Statin

RI

This Venn diagram shows the numbers of patients identified by Regenstrief Institute as exposed to statins.
The 1892 (found by claims alone), 2632 (found by EMR alone), and 2173 (found in both data types) are 3 non-overlapping sets, totaling 6697 people.



Why the lack of overlap? Principally, Regenstrief's Indianapolis data represent a variety of hospitals, practices, laboratories, etc. (in the EMR) and a variety of different public and private insurers (in the payer claims).

The 1892 includes patients who received care in clinical settings that do not feed, or do not yet feed, medication data to Regenstrief.

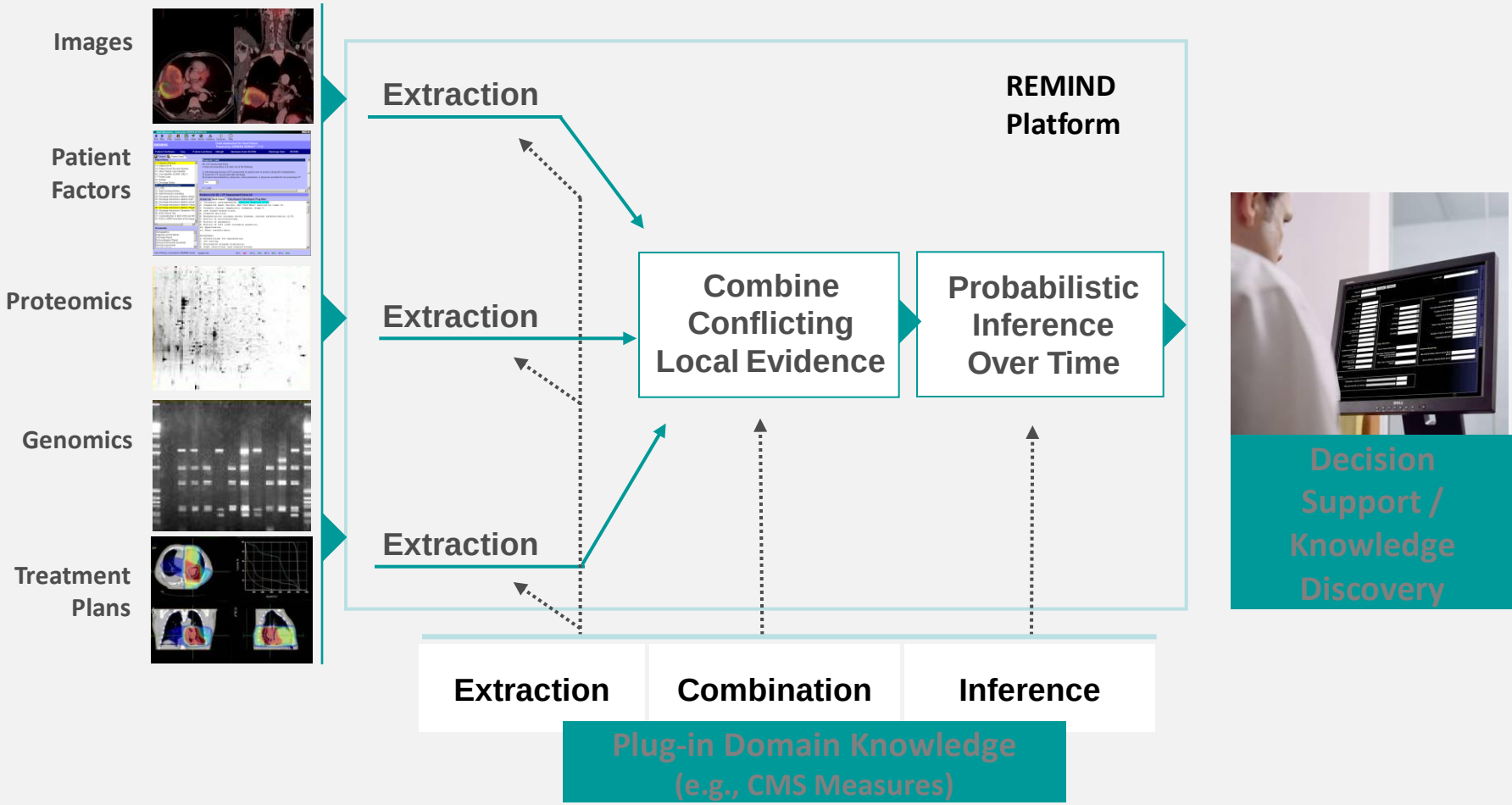
The 2632 includes some patients who had an order for a statin but did not fill the Rx. The 2632 also includes other patients who did fill their Rx but for whom Regenstrief does not receive, or does not yet receive, claims data from their particular health care payer.

Two messages:

1) in a real-world, multifaceted data repository, adding data sources to the analytic mix can greatly boost the #s
2) changes in #s of these relative sizes could greatly affect the results and interpretation; more work is needed for communities to understand the implications such multi-faceted data have for future pharmacovigilance

REMIND Knowledge Platform*: Architecture

Reliable Extraction & Meaningful Inference from Nonstructured Data



REMIND Example

File | Abstraction | Reports | Export | Import | Tools | Help | Save | Verify | Siemens EDM

Visit List | Visit Details

Encounter Id	Patient Id	First Name	Last Name	Admission Date	Discharge Date	Status	Sampled	Stratum
1110008251	1110008251	Bruce	Mack	10/10/2009	10/15/2009	Un-Verified	N/A	N/A

Questions | Documents

- 10. Discharge Status
- 11. Comfort Measures Only
- 12. Clinical Trial
- 13. Transfer From Another ED
- 14. Arrival Date
- 15. Arrival Time
- 16. LVSD
- 17. Initial ECG Interpretation
- 18. Fibrinolytic Administration
- 19. Fibrinolytic Administration Date

Acute Myocardial Infarction 4.7 | Question help

16. Is the left ventricular systolic function (LVSF) documented as an ejection fraction (EF) less than 40% or a narrative description consistent with moderate or severe systolic dysfunction? (LVSD)

- ○ Yes
- ◉ No

Evidences (1)

▼ Cardiac Cath Note (1)	
EF 54%	04/04/2009

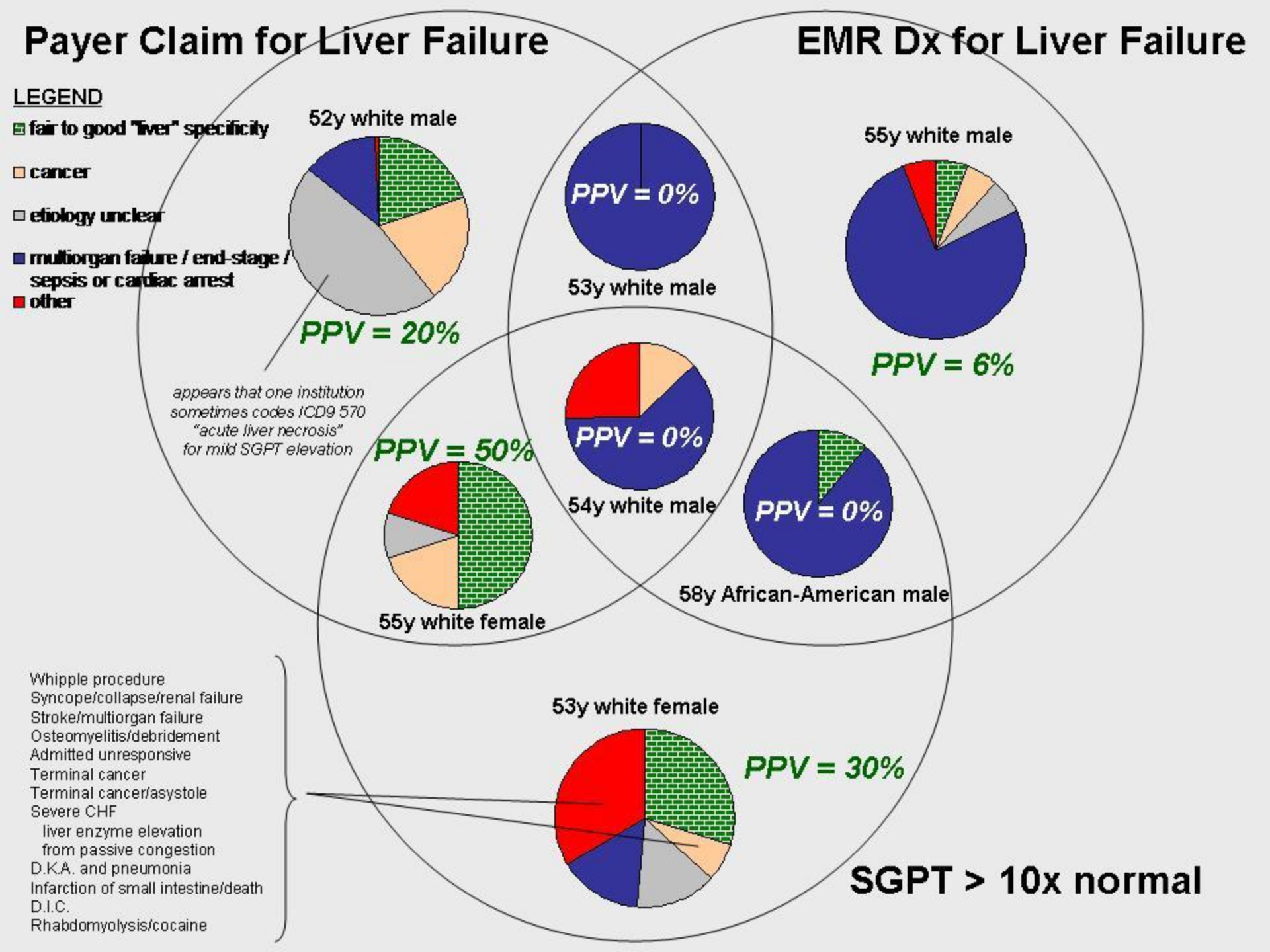
Errors | Comments | Evidences | Change Log

VENTRICULOGRAPHY: Ventriculography revealed an overall preserved left ventricular ejection fraction, EF 54%. There was some inferior hypokinesis. The left ventricular end diastolic pressure was normal at 4 mmHg. Central aortic pressure was 129/59 mmHg.

ANGIOGRAPHY: Adequate cine angiograms were obtained. Circulation is right dominant. The left main gives rise to the LAD and circumflex systems. The circumflex has luminal irregularities present throughout and is compromised by a small first marginal, a large second, with some luminal irregularities. The LAD has a large septal system. Around the first septal and diagonal there is a 50-70% lesion. Also prior to the takeoff of the second diagonal there appears to be a 50% lesion. The LAD continues to

Logged in as: 'sqmuser' Facility: 'Health Enterprise' Displayed: 21 Selected: 1

Soarian® Quality Measures powered by REMIND™ Platform


Payer Claim for Liver Failure
EMR Dx for Liver Failure
LEGEND
fair to good "liver" specificity
cancer
etiology unclear
multiorgan failure / end-stage / sepsis or cardiac arrest
other
52y white male
PPV = 20%
appears that one institution sometimes codes ICD9 570 "acute liver necrosis" for mild SGPT elevation
PPV = 0%
53y white male
55y white male
PPV = 6%
PPV = 50%
55y white female
PPV = 0%
54y white male
PPV = 0%
58y African-American male
53y white female
PPV = 30%
Whipple procedure
Syncope/collapse/renal failure
Stroke/multiorgan failure
Osteomyelitis/debridement
Admitted unresponsive
Terminal cancer
Terminal cancer/asystole
Severe CHF
liver enzyme elevation
from passive congestion
D.K.A. and pneumonia
Infarction of small intestine/death
D.I.C.
Rhabdomyolysis/cocaine
SGPT > 10x normal

Quality for purpose

- Clinical care
- Accountable care
- Public health reporting
- CER
- Drug/Device safety
- Health services research

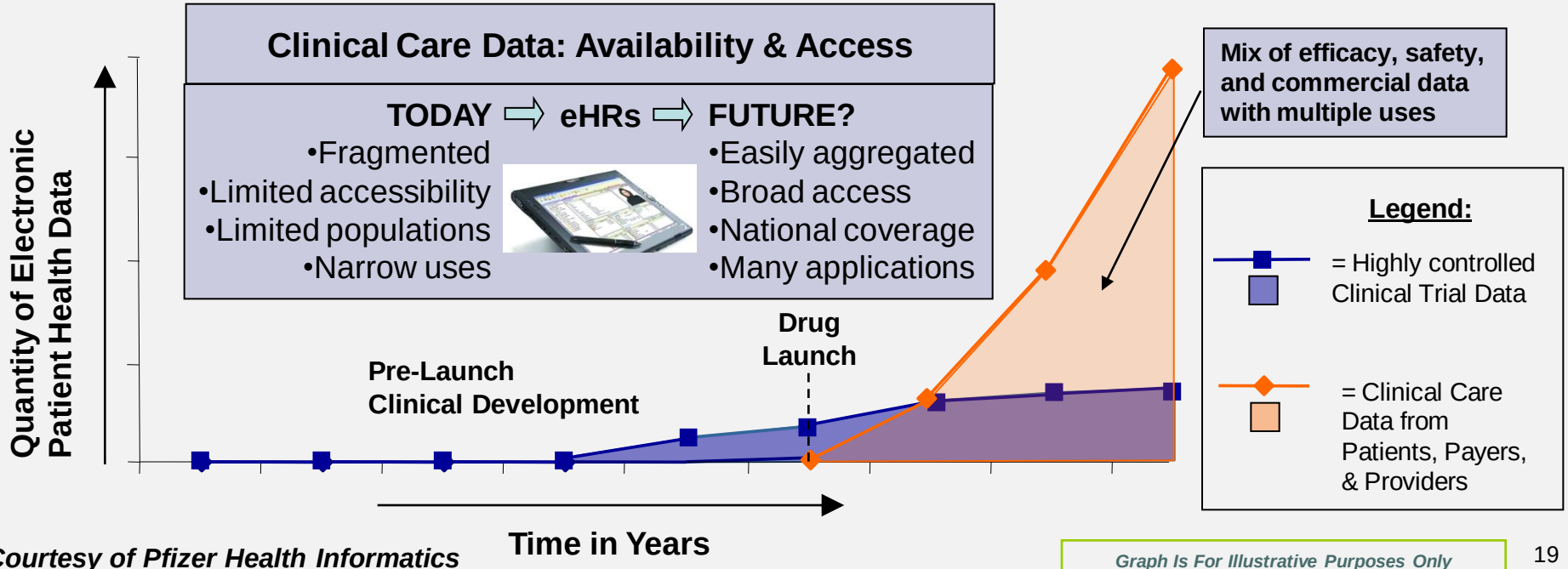
Clinical trials vs. clinical practice

Clinical Trials:

Data are high integrity due to validation, but are sourced from limited patient populations

Post-launch Clinical Care:

Today, data from payers & providers are lower quality, fragmented, and challenging to access



Courtesy of Pfizer Health Informatics

Graph Is For Illustrative Purposes Only

Draft – For discussion only

Numerous data sources to support VBHC analyses, but not all data sources are equivalent

Data

Numerous potential data sources

- Clinical trials
- Claims
- Disease registry
- Modalities + service providers
- EMR
- Patient sources

Three major dimensions determine utility of a data source for a business question

Variables Depth	**Type and number of variables within a dataset** • Determines what can be analyzed • Determines whether analysis can be adjusted for case-mix • Needs to be optimized to contain cost and complexity
Observations Breadth	**Size of the dataset, both in time and number of patient** • Impacts the "power" of the analyses (how small an effect can be detected) • Impacts strength of the conclusions • Should be maximized as long as quality can be maintained
Quality control	**Degree of syntactic and semantic consistency within a dataset and between datasets; validation – 'correctness' of each field, reliability of clinical reporting** • Impacts whether analysis can be "trusted" • Determines whether data can be integrated between datasets or organizations • Impacts whether analyses can be compared across datasets (uniformity) and populations (generalizability)

Critical capability in value-based health care: leveraging the right data to meet business requirements

Source: Stakeholder interviews, BCG analysis

Data

Major dimensions composed of numerous factors

Variables
Depth

	Factors	Rationale
Measures	Outcomes measures	• Improvement in outcomes (cost and quality) is ultimate goal of VBHC
	Relevant process measures	• Understanding drivers of outcomes enables quality improvement
	Financial measures	• Understanding cost and utilization • Enables cost-effectiveness research
	Patient-centered measures	• Patient-generated data, e.g. assessment of health and well-being via satisfaction or survey results • Supplements clinical findings
Complete-ness	Number of variables	• Enables greater diversity of analyses • Enables risk-adjustment/case-control
	Granularity of variables	• More granular variables enable more detailed analyses
Context	Risk-adjustment data	• Demographics, time, setting, source (obj./subjective) enable analyses to be placed in context
	Patient ID	• Enables segmentation of data (e.g. based on demographics) • Enables follow-up with providers/pts

Observations
Breadth

	Factors	Rationale
Population / Sample	Number of patients	• Improves detection of small differences • Decreases need for risk-adjustment
	Penetration	• Improves applicability of findings to population • Decreases need for risk adjustment
	Number of records	• Enables ID of subsegments of patients / outcomes • Improves precision, validity of data
	Skew / Generalizability	• 'Balanced' population enables analyses which are more generalizable across populations
	Internal distribution/ Comparability	• Dataset requires large enough sample for each provider to enable comparison across providers
Time and Setting	Longevity / Temporal extent	• Enables general trending over time
	Longitudinality / Temporal consistency	• Enables trending of specific patients over time
	Longitudinality / Across care settings	• Enables linkage of patient data across care settings within an episode

Quality control

	Factors	Rationale
Process	Intent	• Data collected for a specific purpose more likely to be relevant to the question and higher quality
	Validation	• Increases confidence that findings are accurate (e.g. collected in controlled environment; double entry in clinical trials)
	Fidelity	• Increases confidence that findings represent the 'real world' (e.g. that an outcome in one setting means the same as in another; 'apples to apples')
	Timeliness	• Increases relevance of data
Technical	Structure	• 'Syntactic' consistency • Enables automated analysis and integration of datasets
	Coding	• 'Semantic' consistency • Enables confidence in internal data validity and comparison across datasets
	Linkability	• Linkage of patient data across datasets (may be done without identification) enables construction of integrated datasets with greater depth and breadth

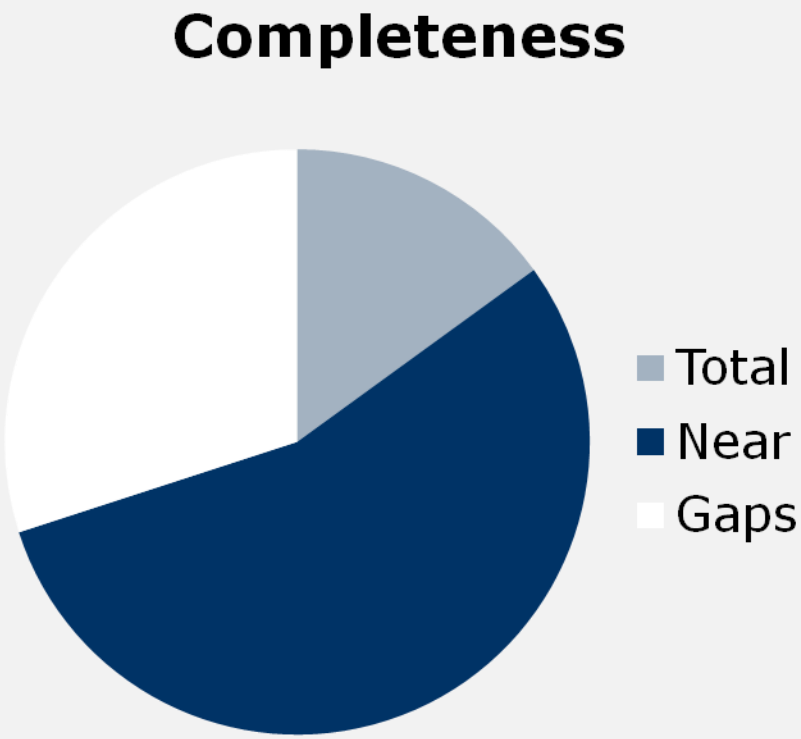
Source: Stakeholder interviews, BCG analysis

Pharmaceutical Questions

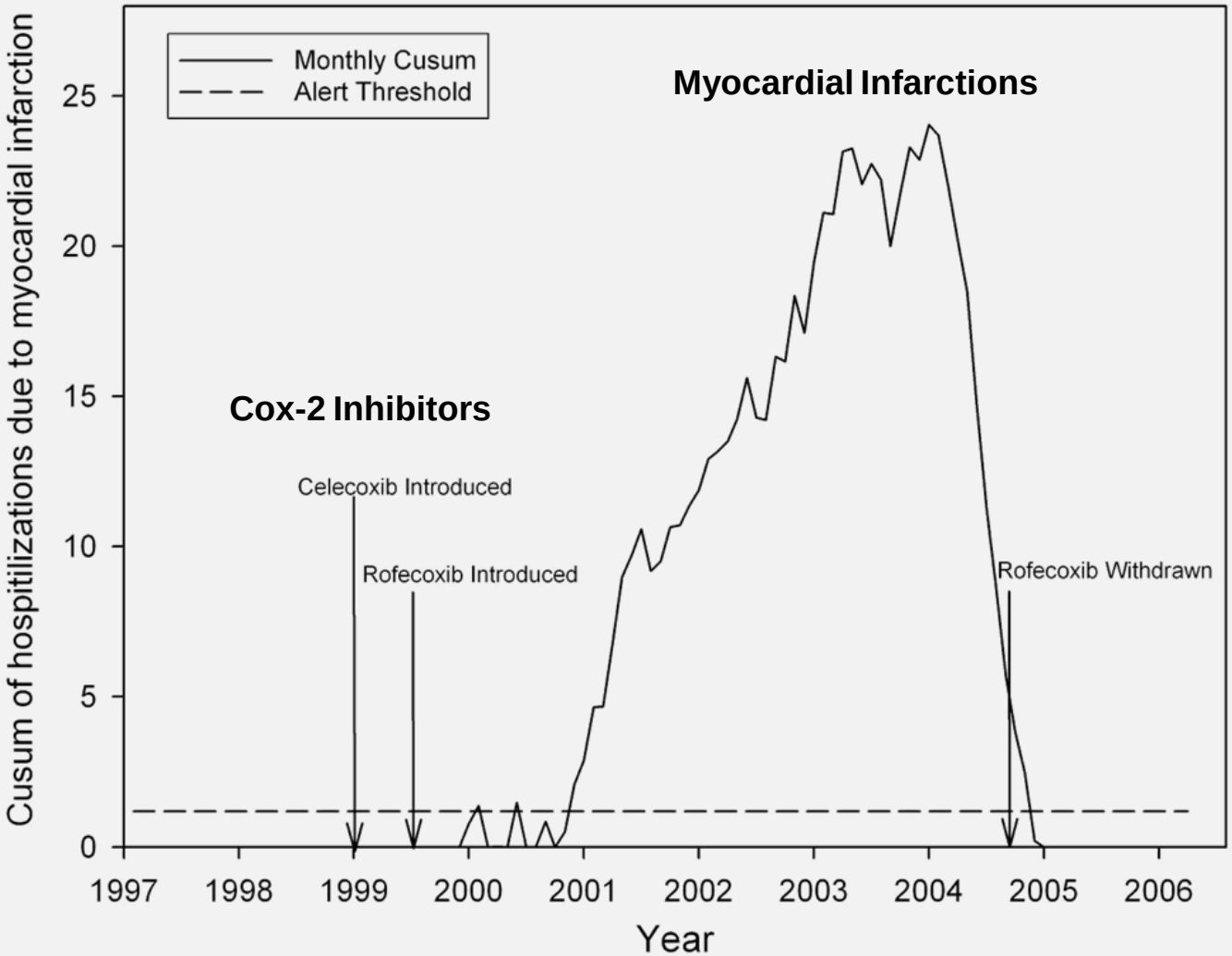
Questions

- 10 companies
- 10 questions per company

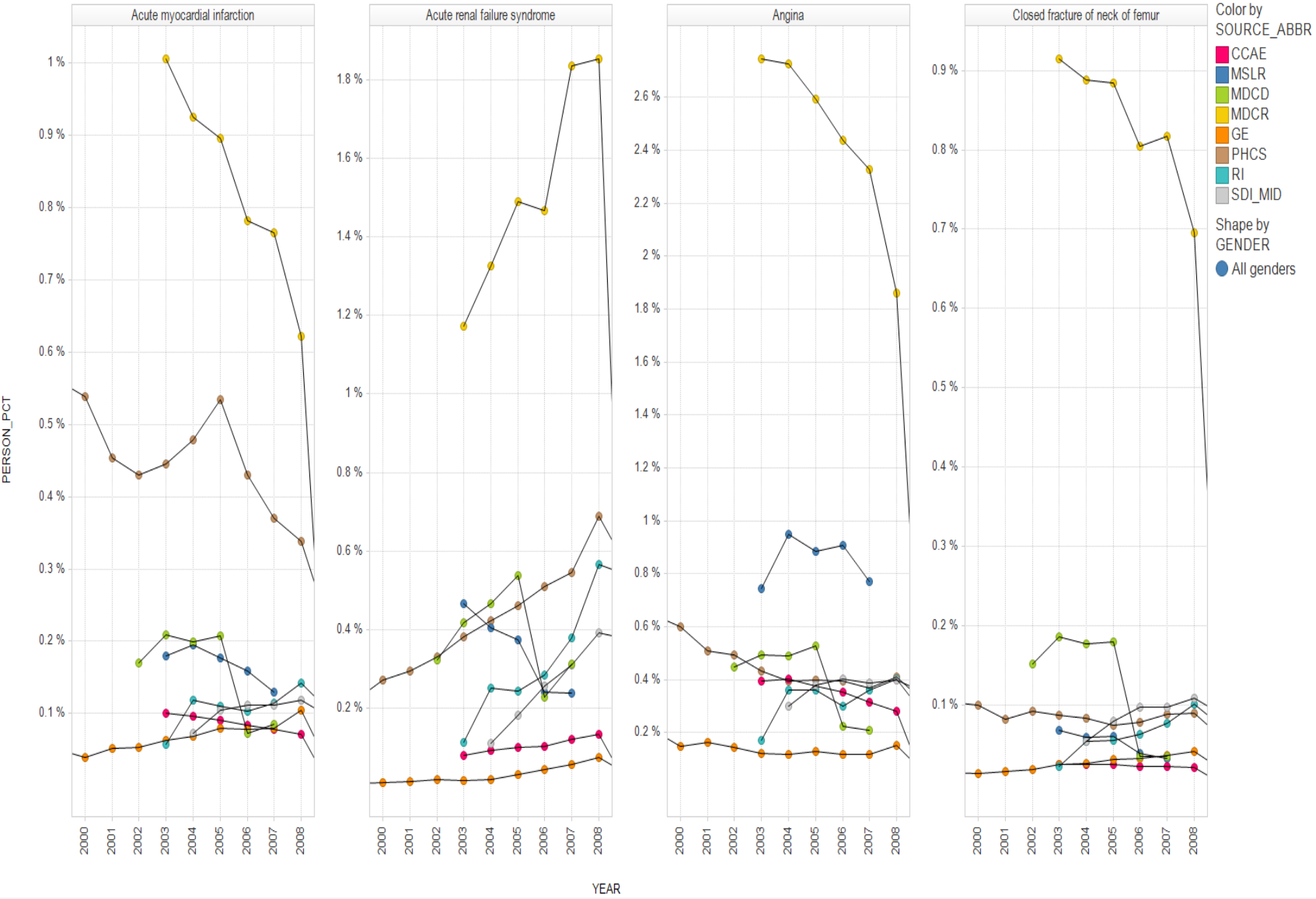
Answers

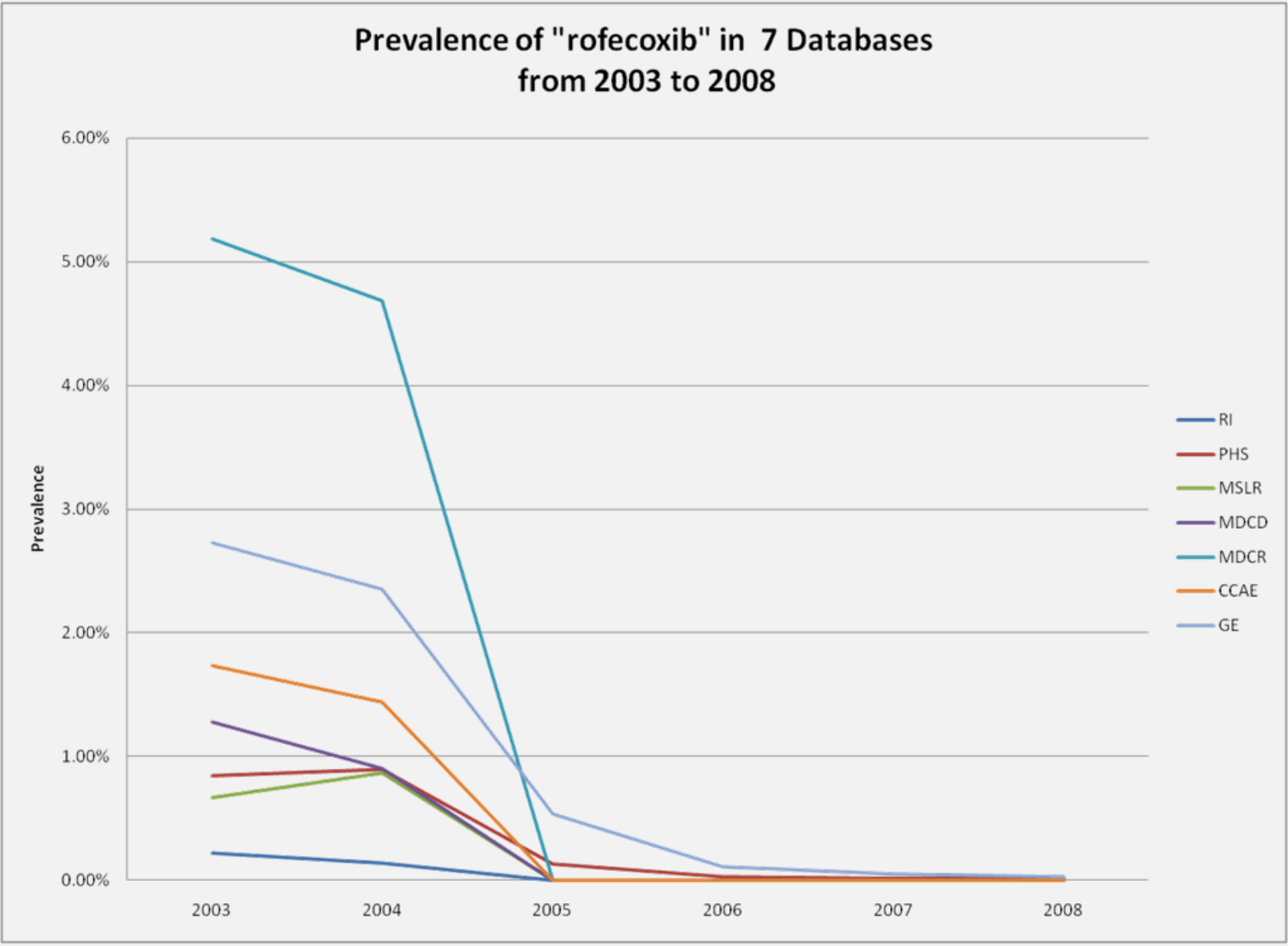


Monitoring Adverse Drug Events

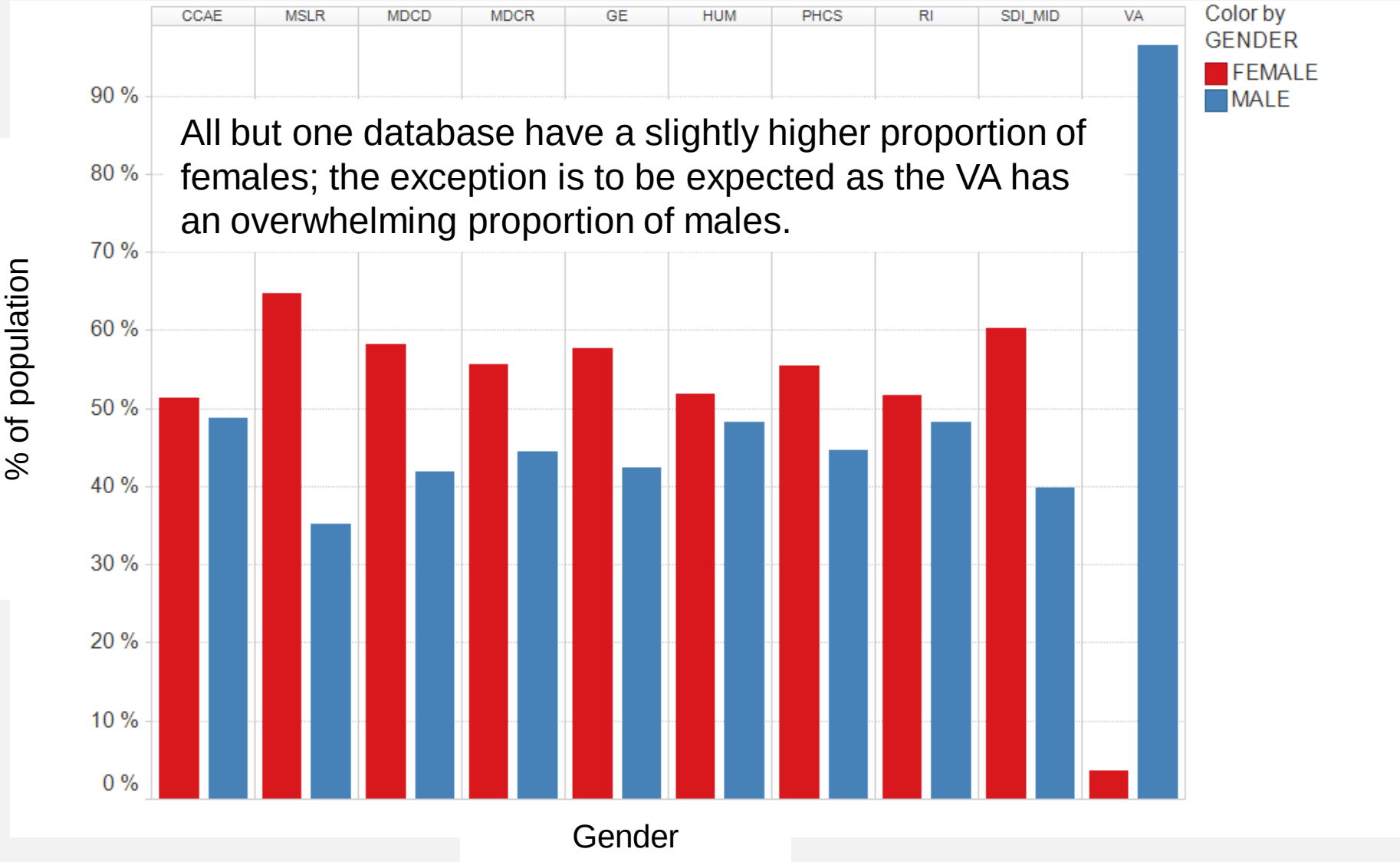


Brownstein, et al. 2007. Plos One.


Acute myocardial infarction
Acute renal failure syndrome
Angina
Closed fracture of neck of femur
PERSON_PCT
YEAR
Color by SOURCE_ABBR
CCAE
MSLR
MDCD
MDCR
GE
PHCS
RI
SDI_MID
Shape by GENDER
All genders

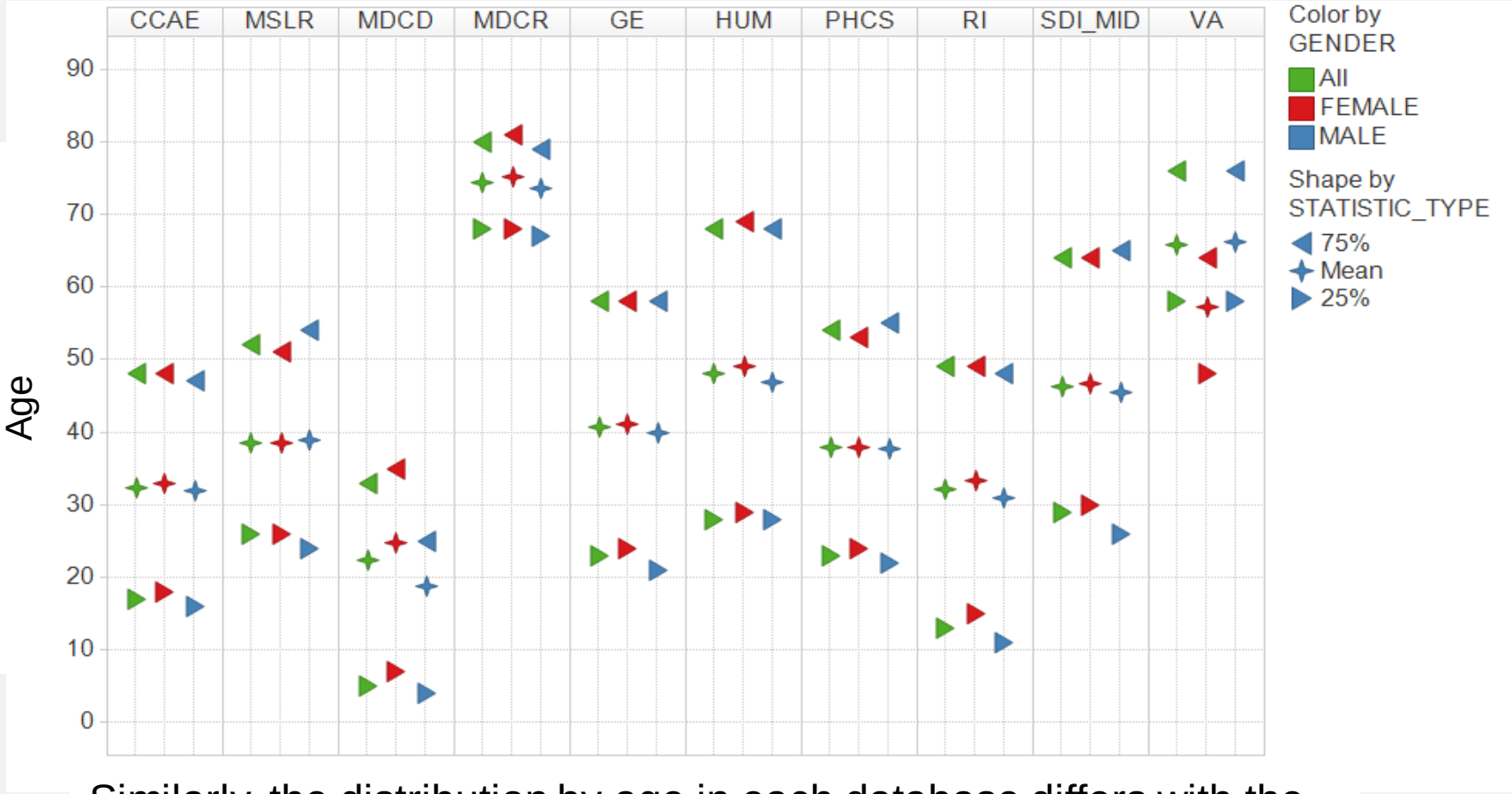

Prevalence of "rofecoxib" in 7 Databases
from 2003 to 2008
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
Prevalence
2003
2004
2005
2006
2007
2008
RI
PHS
MSLR
MDCD
MDCR
CCAE
GE

Sources by Gender



All but one database have a slightly higher proportion of females; the exception is to be expected as the VA has an overwhelming proportion of males.

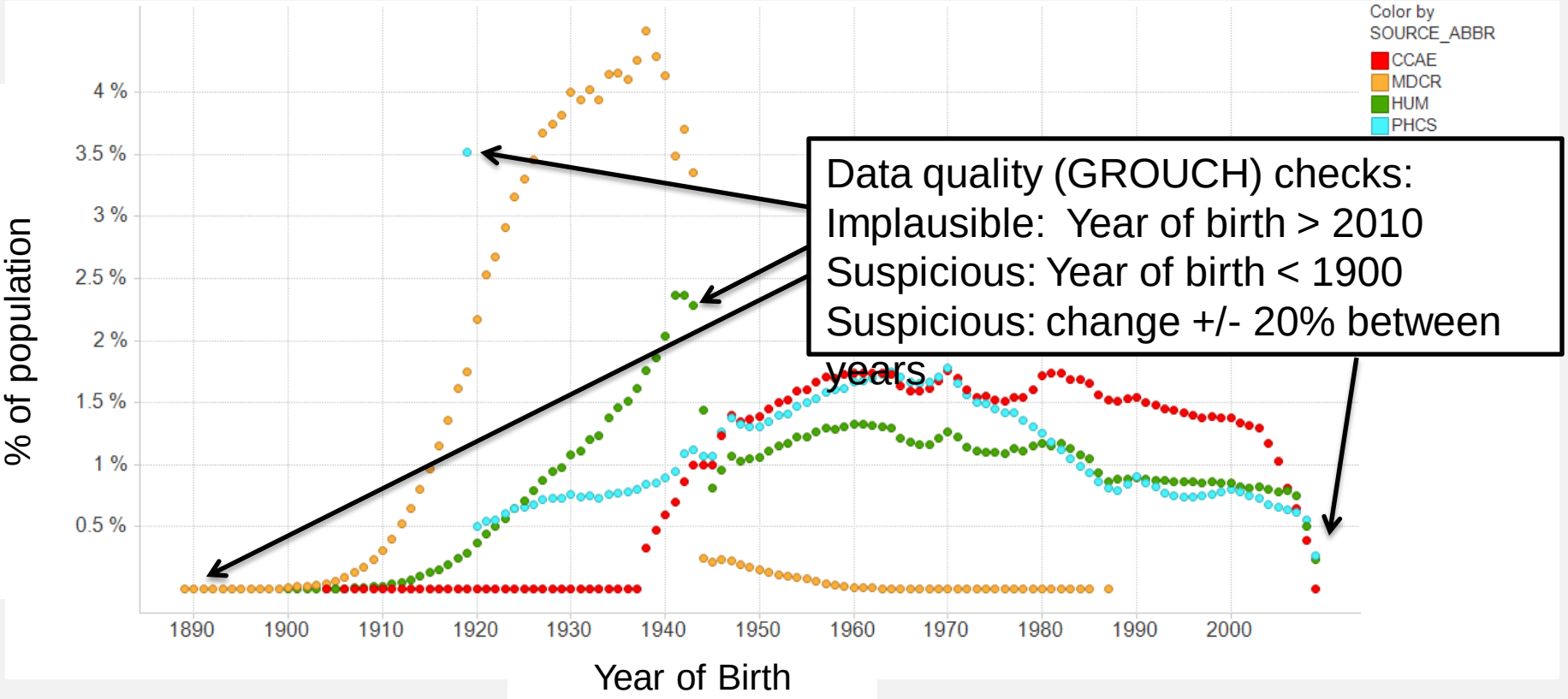
Sources by age distribution



Similarly, the distribution by age in each database differs with the most striking difference as expected in the older ages in Medicare. Medicaid data shows a gender imbalance in age, as females are older than males.

Perfect example of the potential diversity that a data network can bring and the promise of generalizability.

Age distribution



Data quality (GROUCH) checks:
Implausible: Year of birth > 2010
Suspicious: Year of birth < 1900
Suspicious: change +/- 20% between years

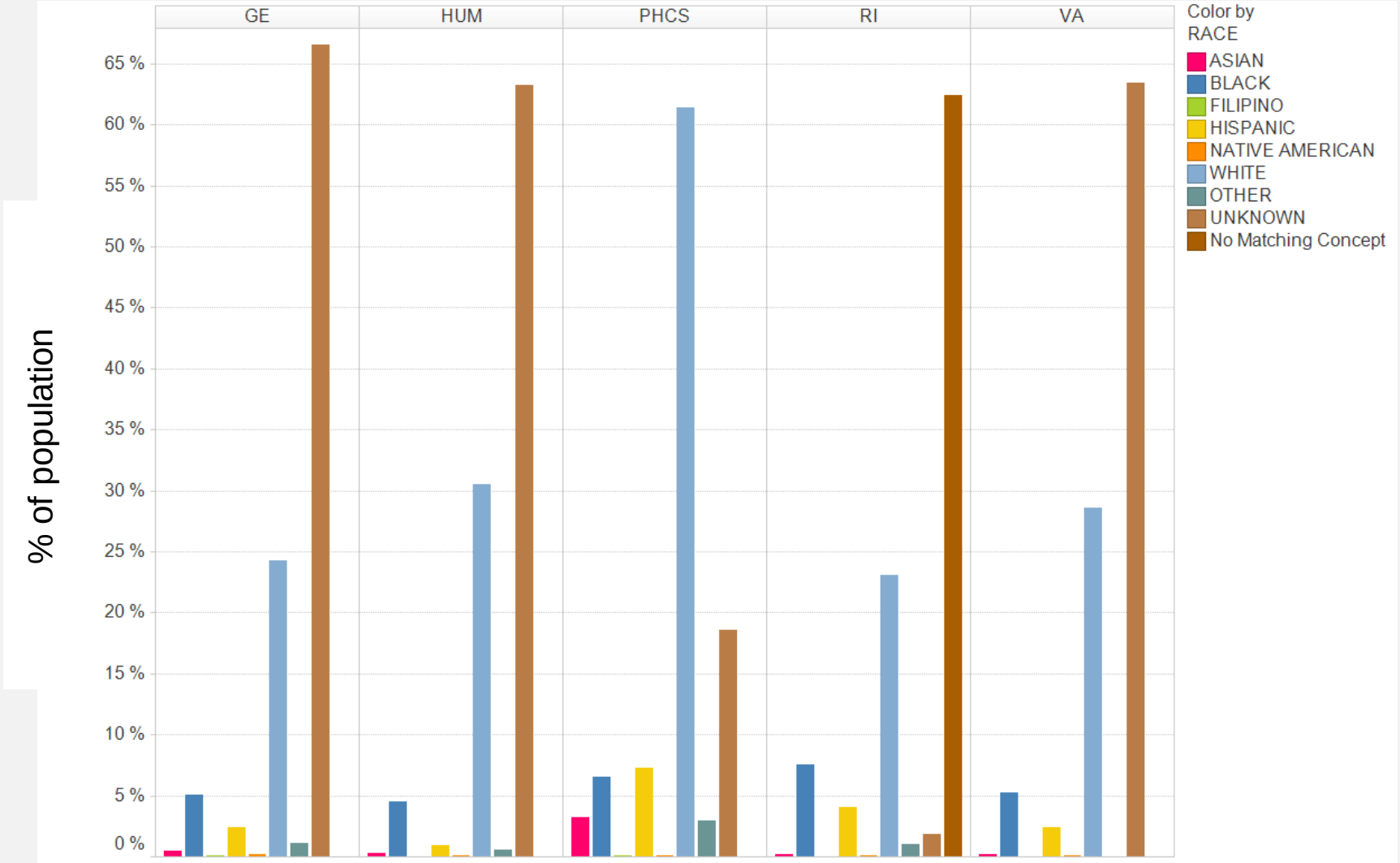
CCAE, being a privately insured population, primarily reflects employed and their dependen

In contrast, MDCR represents patients with supplemental Medicare benefits, so primarily refl
und
bot
retir
po
dis

Humana, as a large insurer providing coverage to

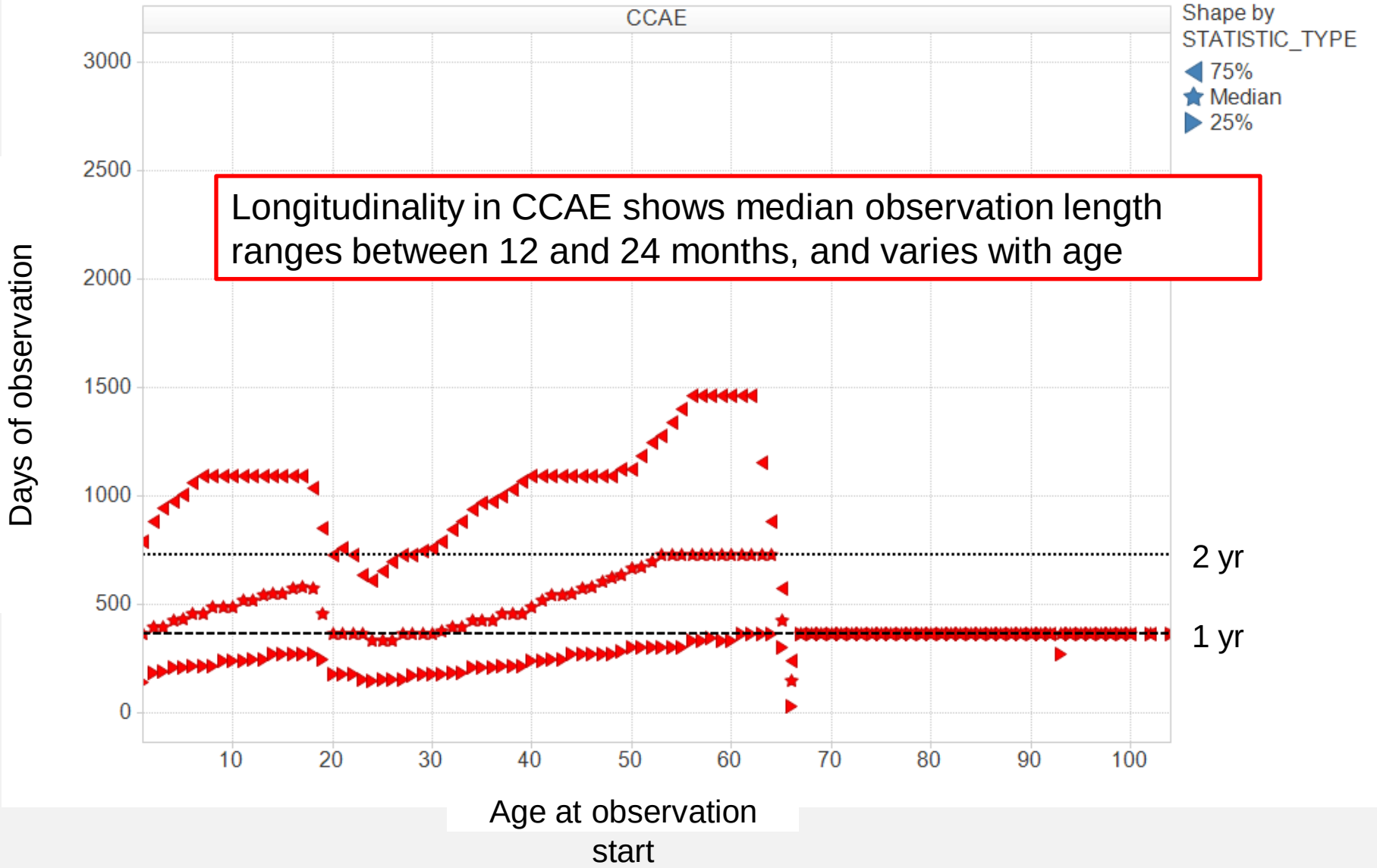
Partners HealthCare System, as a clinical system providing care to patients of varied insurance coverage, shows a more uniform age distribution.

Race distribution



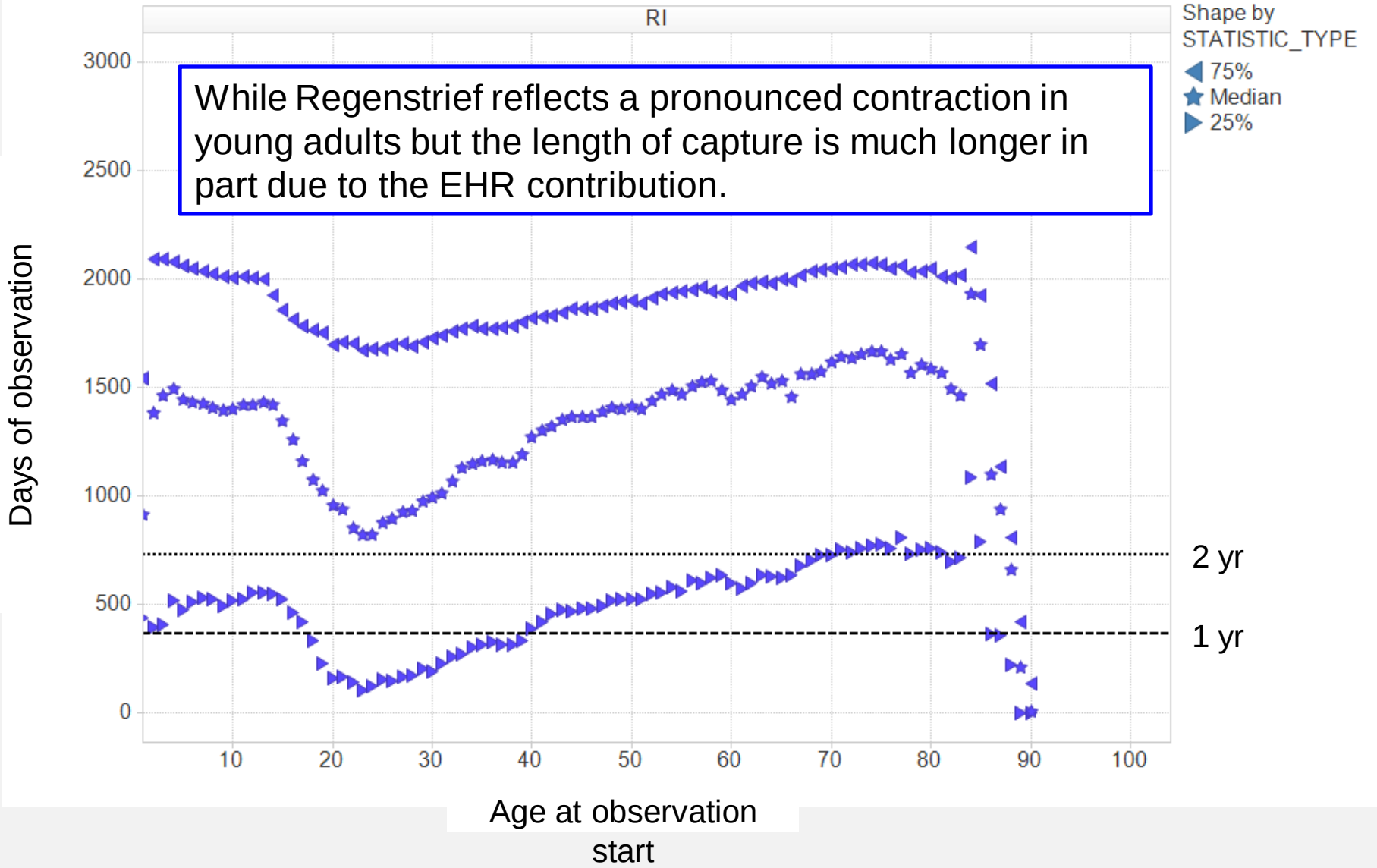
Ethnic diversity is a concept that we would like to see more cogently and consistently represented.

Observation period length



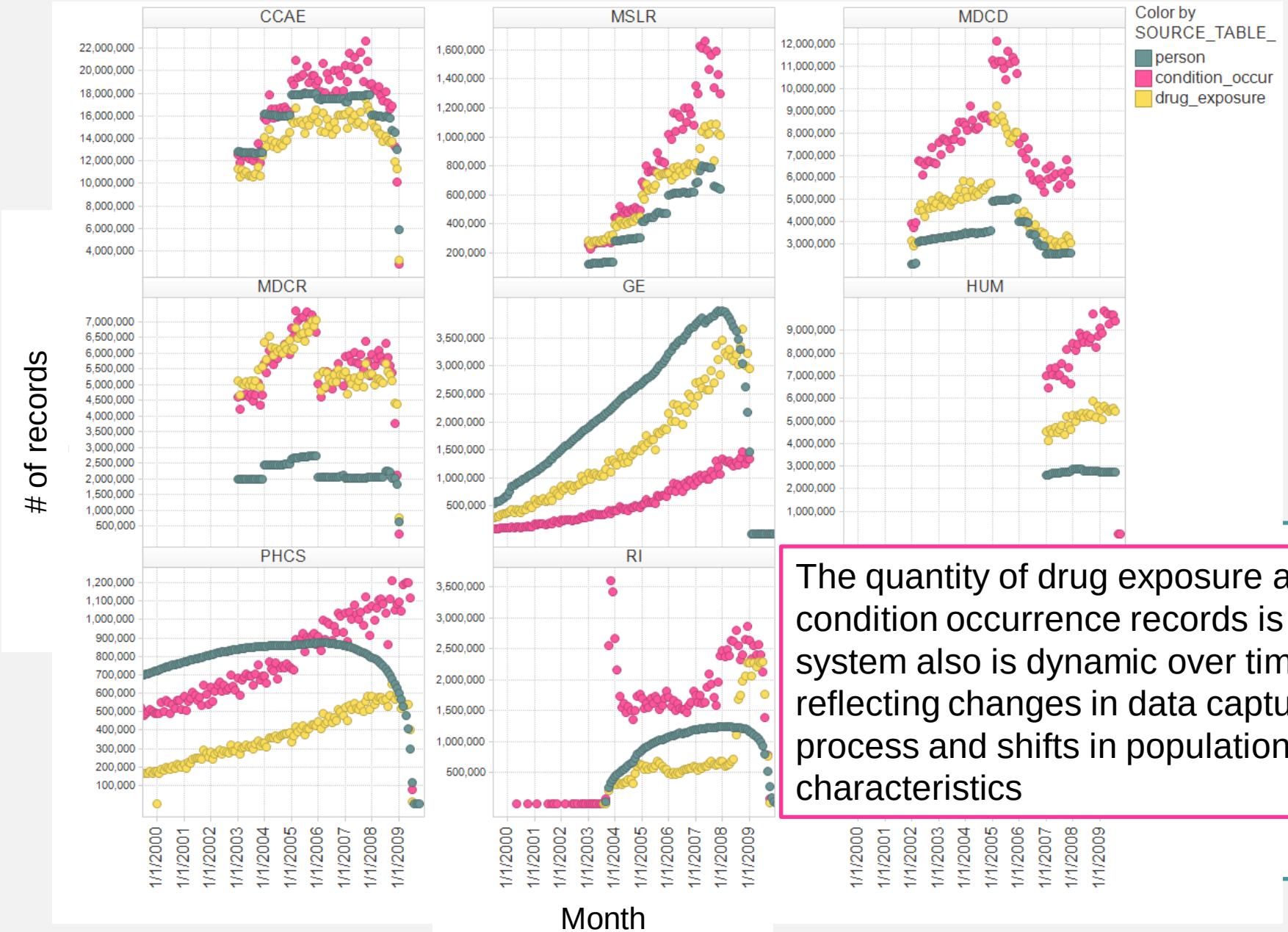
Longitudinality in CCAE shows median observation length ranges between 12 and 24 months, and varies with age

Observation period length



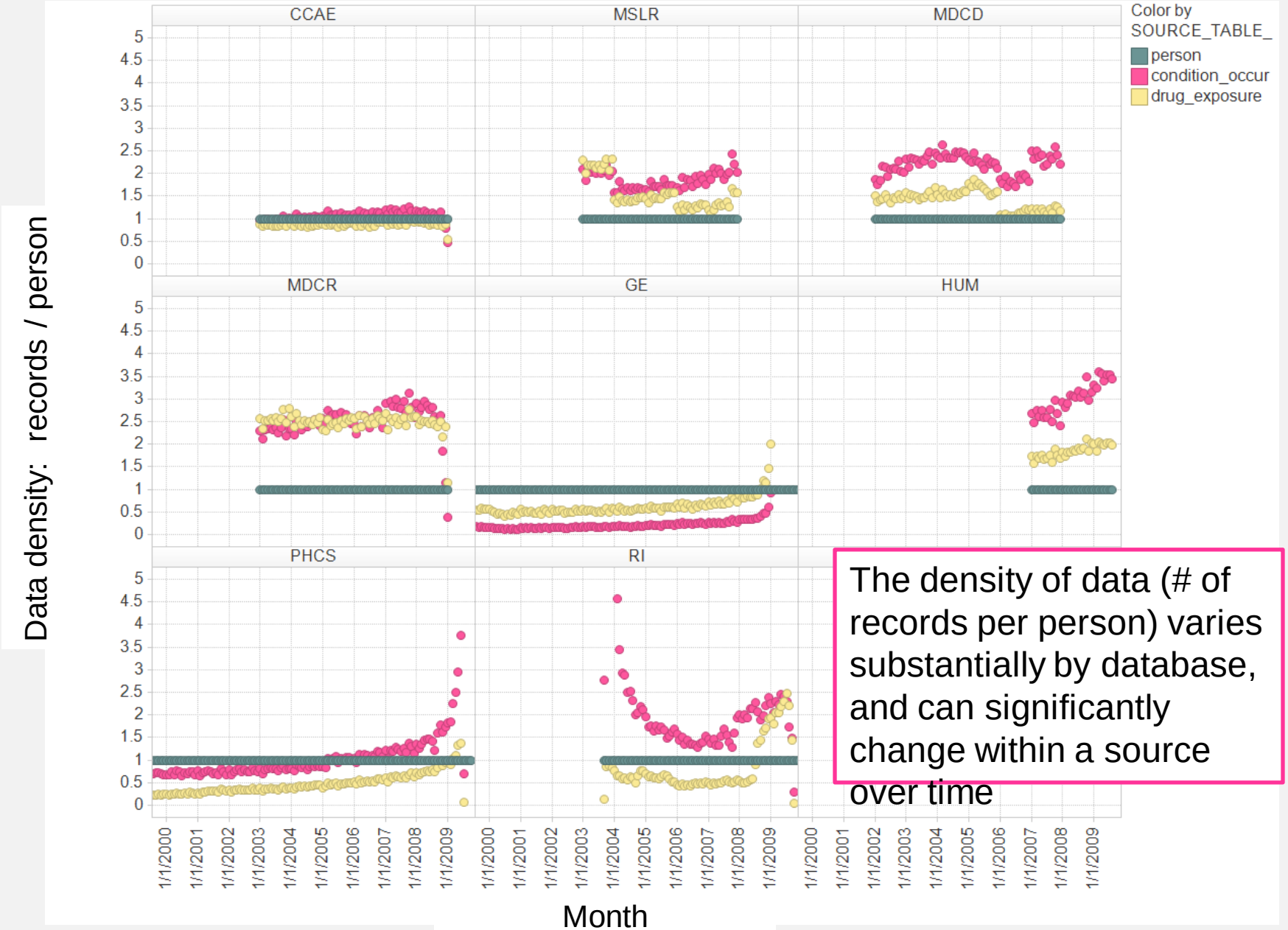
While Regenstrief reflects a pronounced contraction in young adults but the length of capture is much longer in part due to the EHR contribution.

Records over time



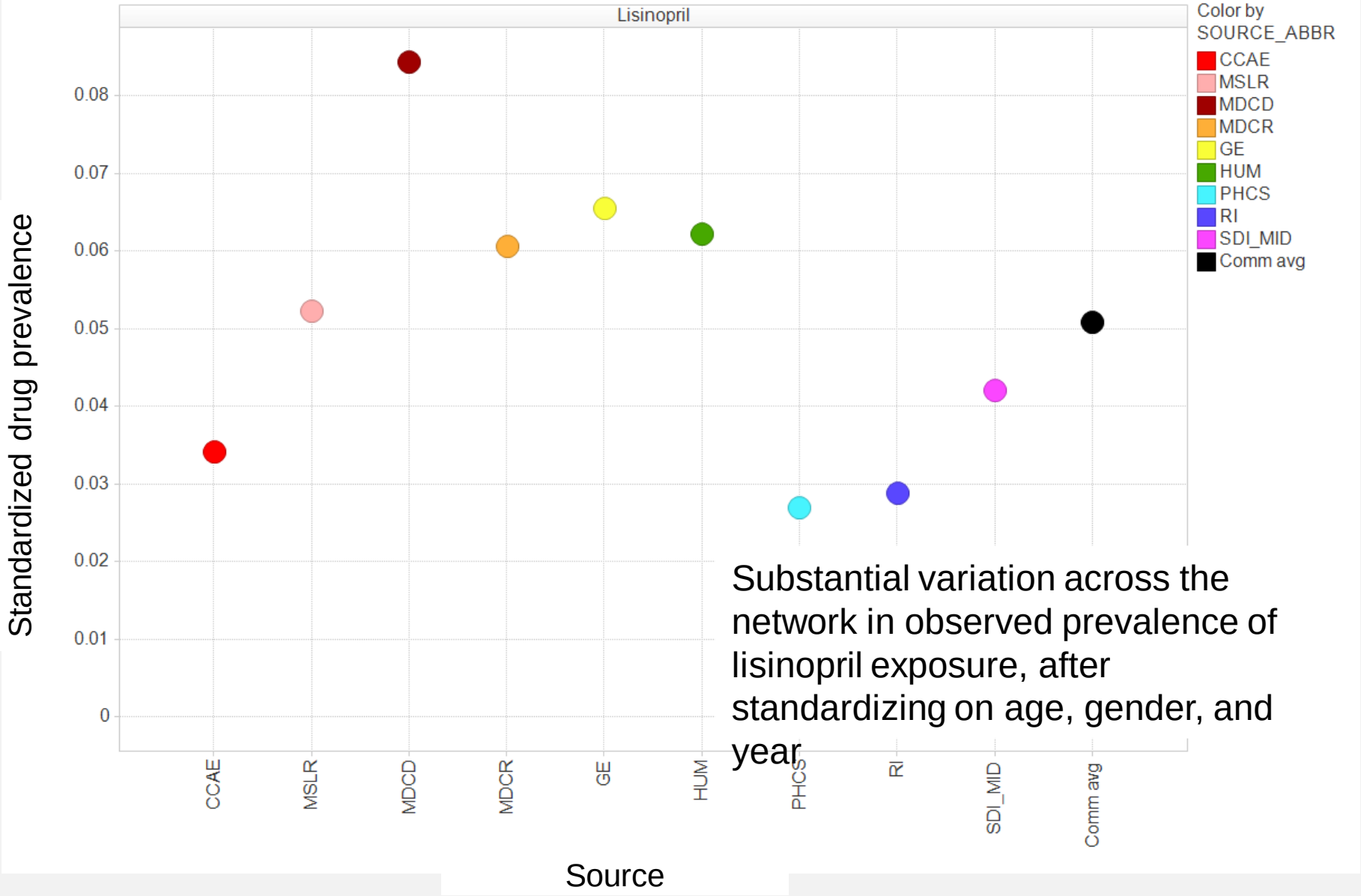
The quantity of drug exposure and condition occurrence records is the system also is dynamic over time, reflecting changes in data capture process and shifts in population characteristics

Records per person over time



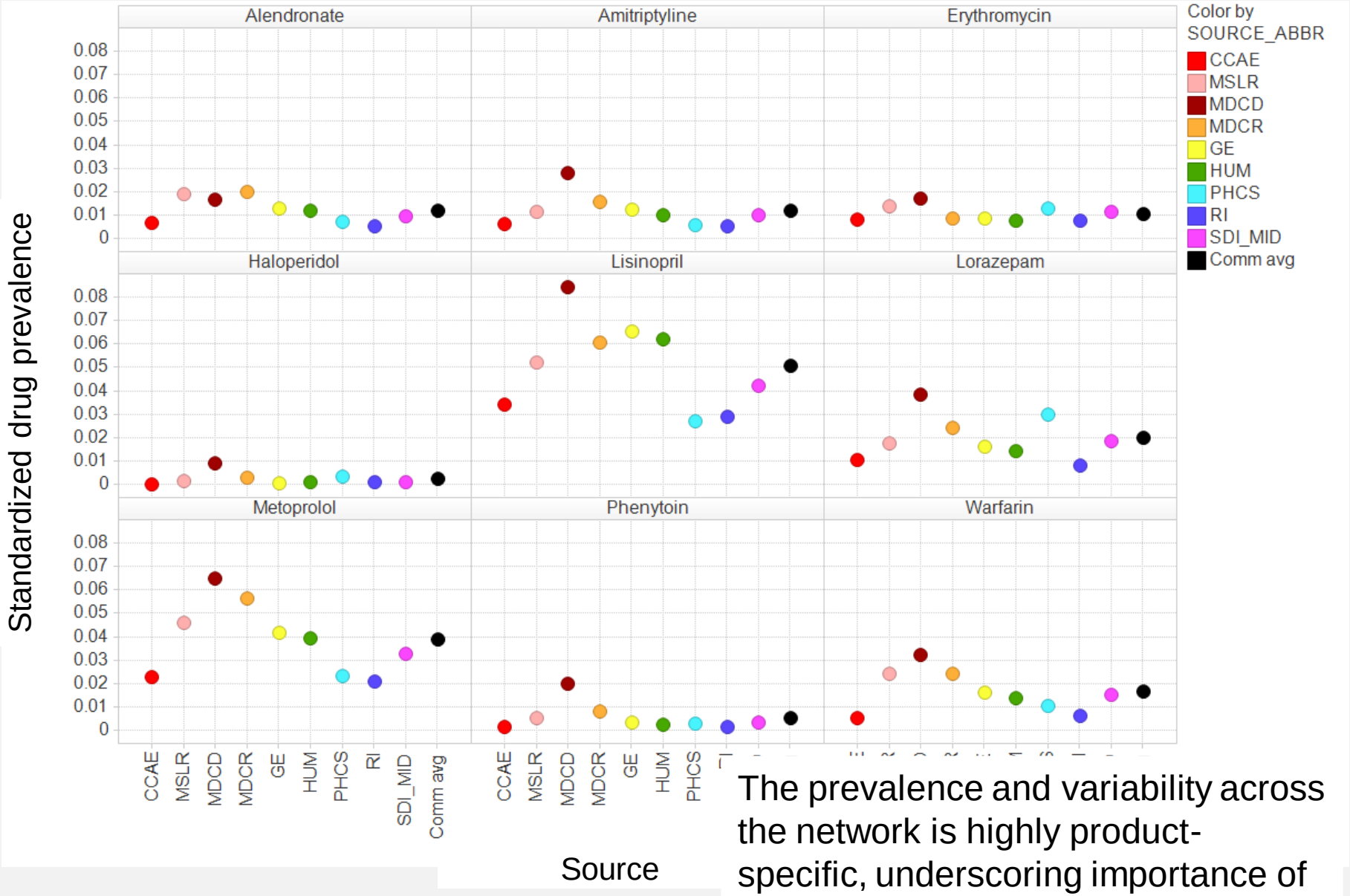
The density of data (# of records per person) varies substantially by database, and can significantly change within a source over time

Standardized drug prevalence



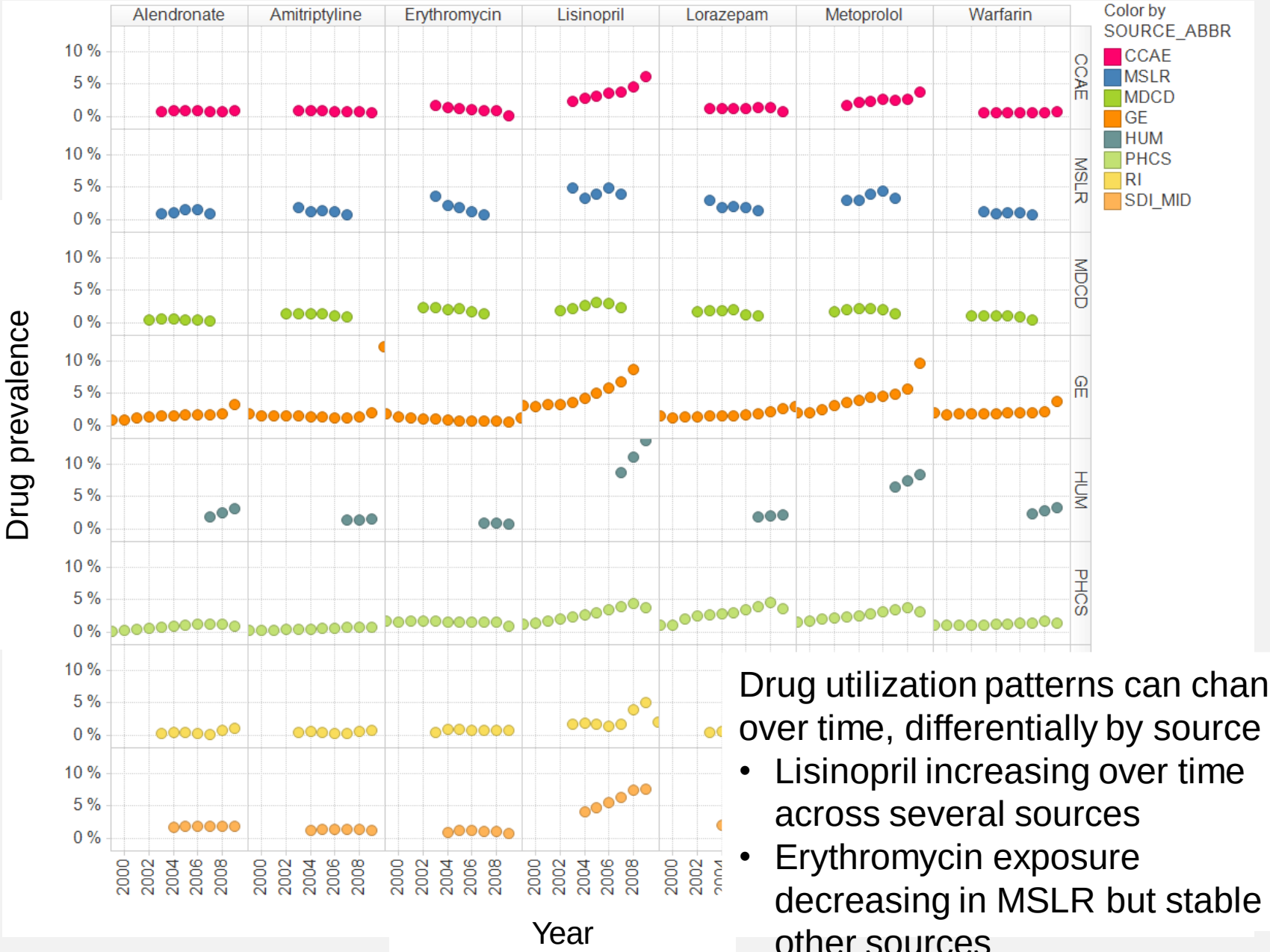
Substantial variation across the network in observed prevalence of lisinopril exposure, after standardizing on age, gender, and year

Standardized drug prevalence



The prevalence and variability across the network is highly product-specific, underscoring importance of efficient exploration of summary statistics

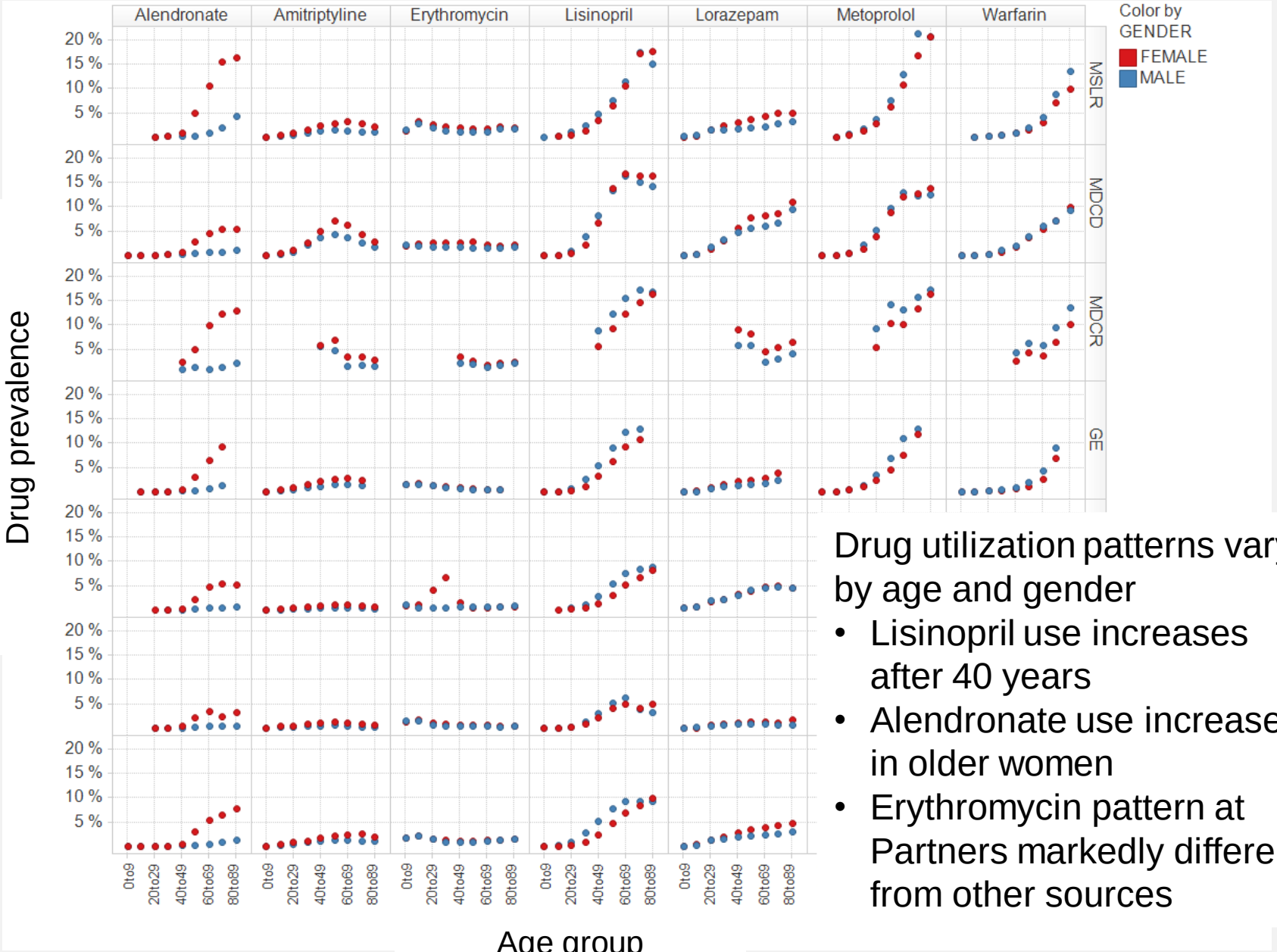
Drug prevalence by year



Drug utilization patterns can change over time, differentially by source

- Lisinopril increasing over time across several sources
- Erythromycin exposure decreasing in MSLR but stable in other sources

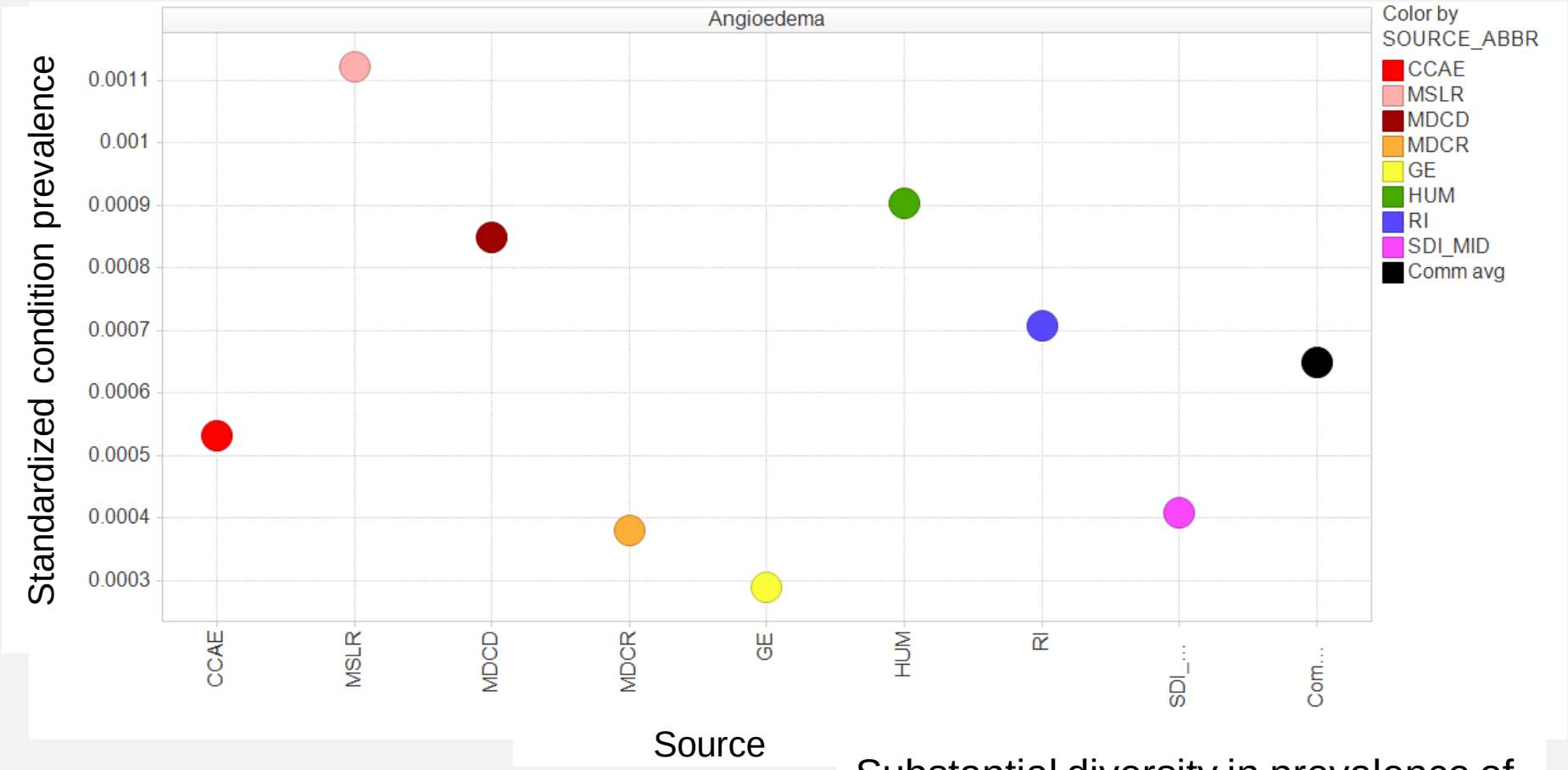
Stratified drug prevalence by age group



Drug utilization patterns vary by age and gender

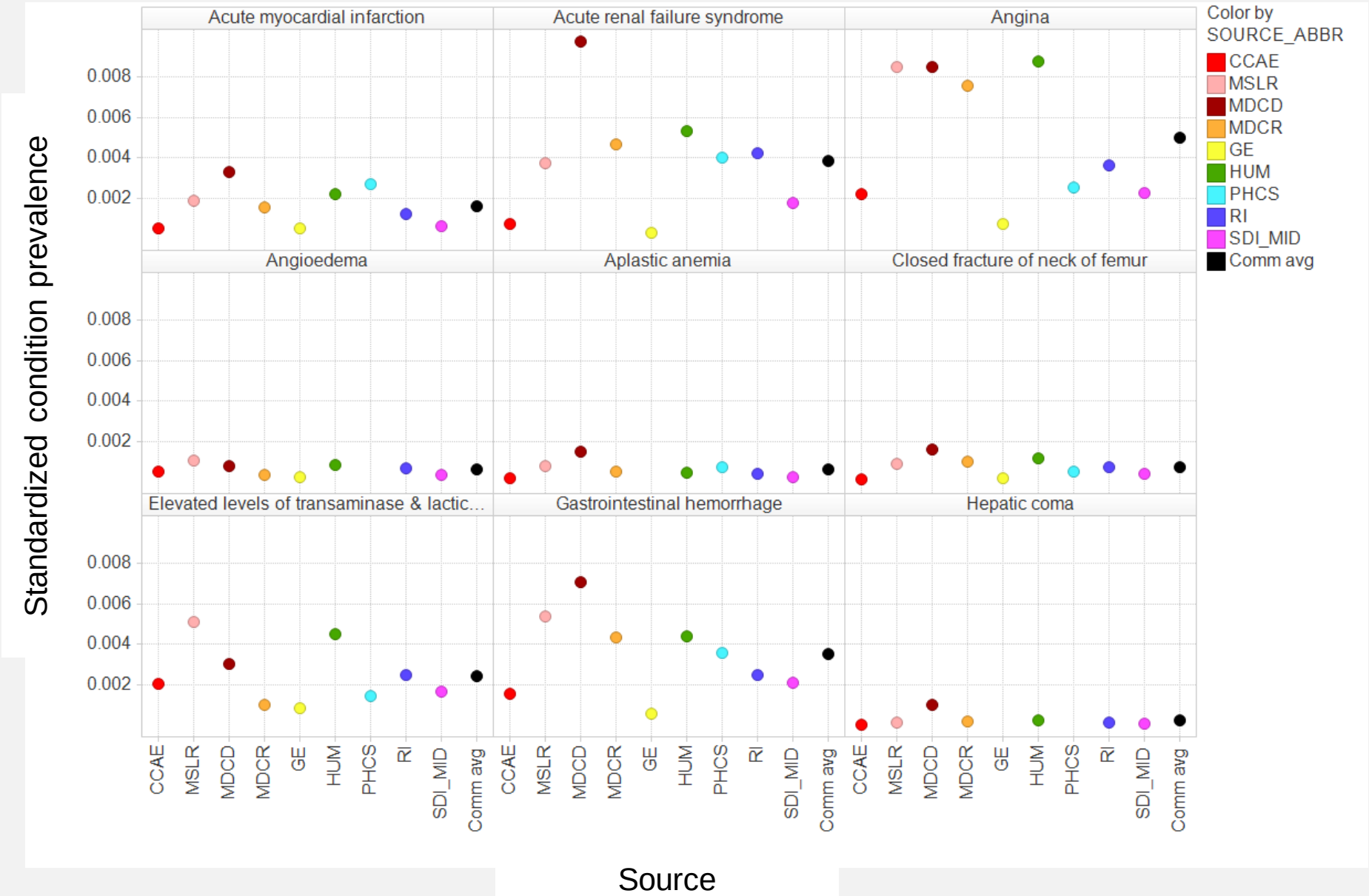
- Lisinopril use increases after 40 years
- Alendronate use increases in older women
- Erythromycin pattern at Partners markedly different from other sources

Standardized condition prevalence

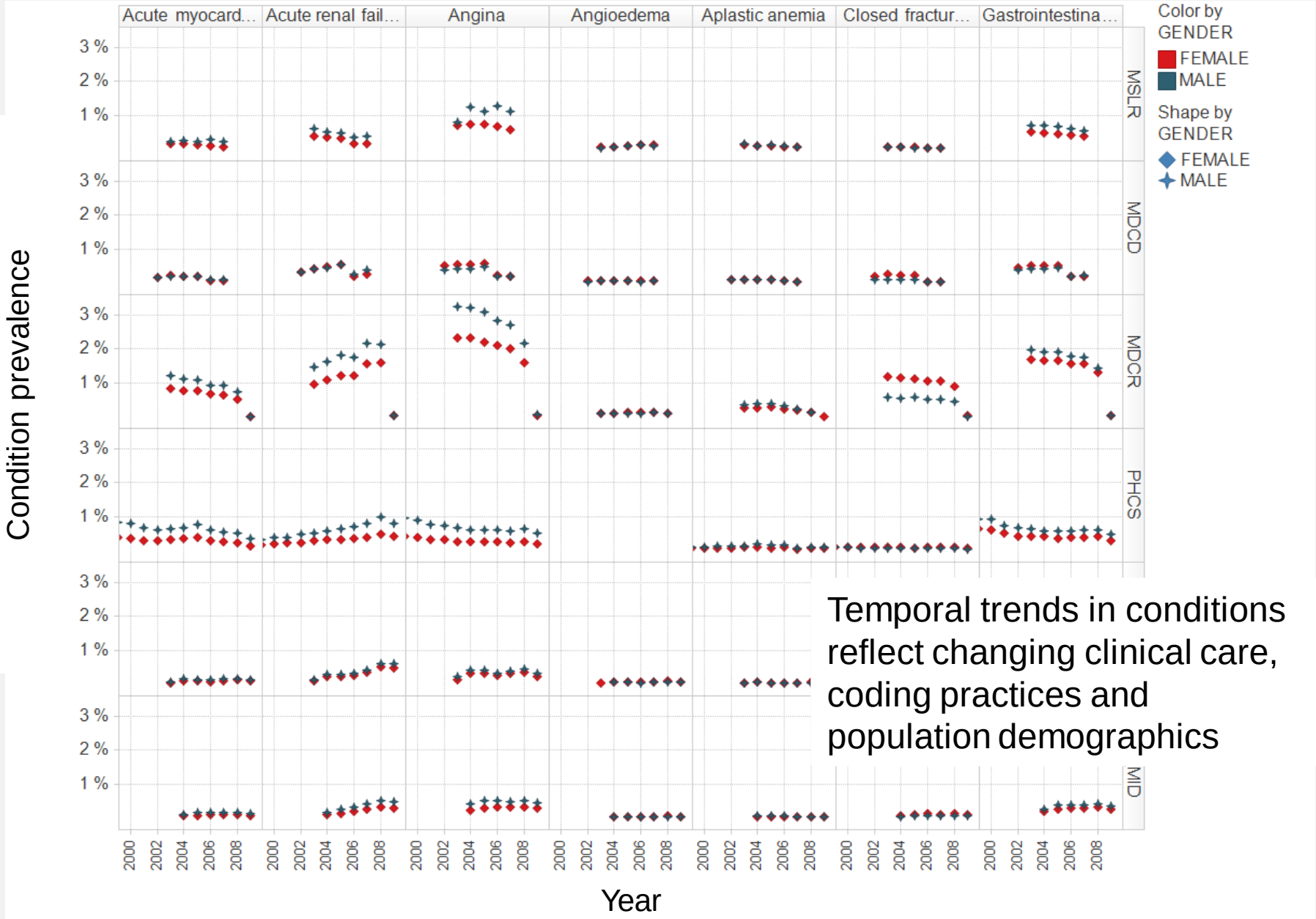


Substantial diversity in prevalence of condition occurrence across sources

Standardized condition prevalence

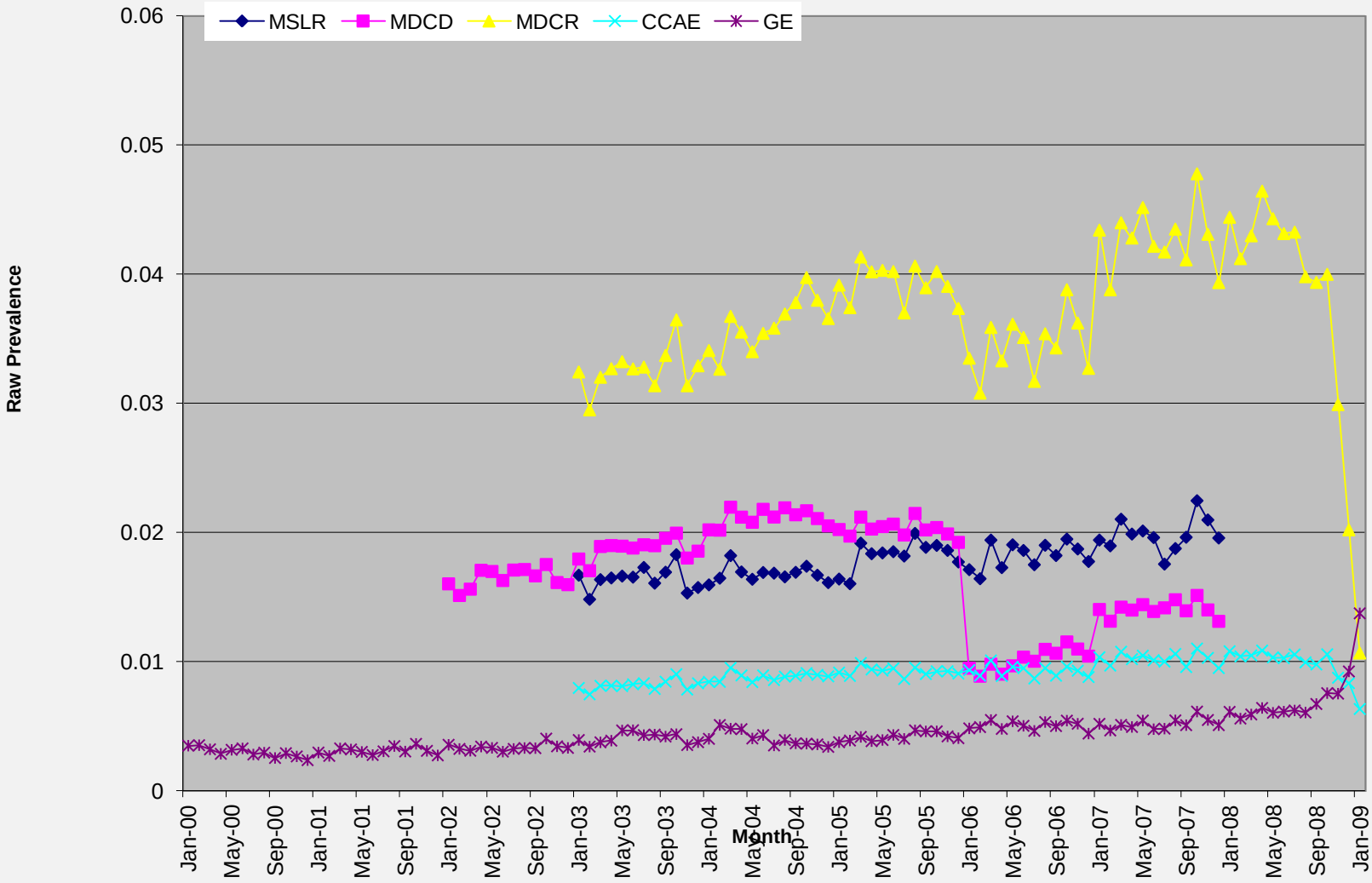


Stratified condition prevalence by year

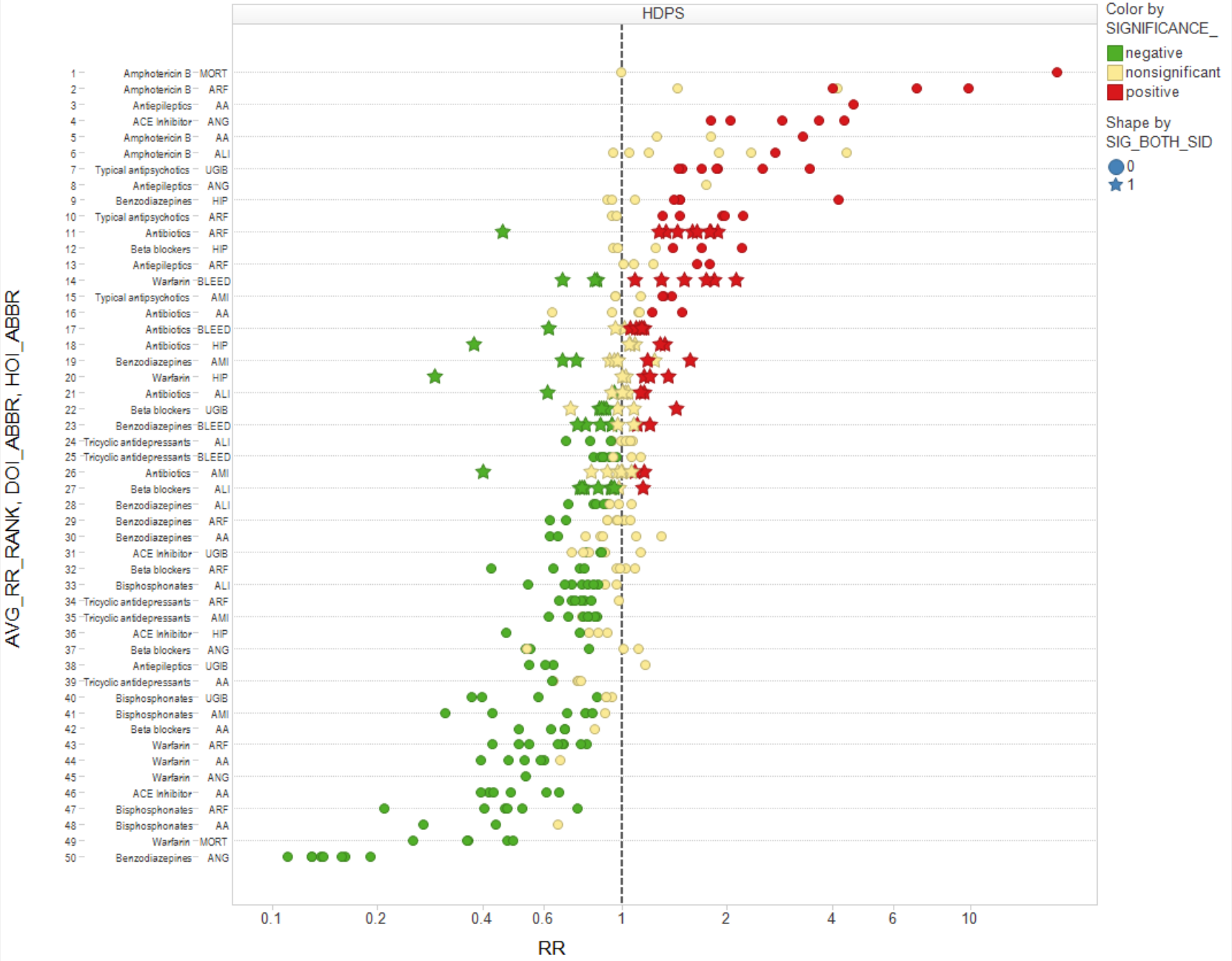


Temporal trends in conditions reflect changing clinical care, coding practices and population demographics

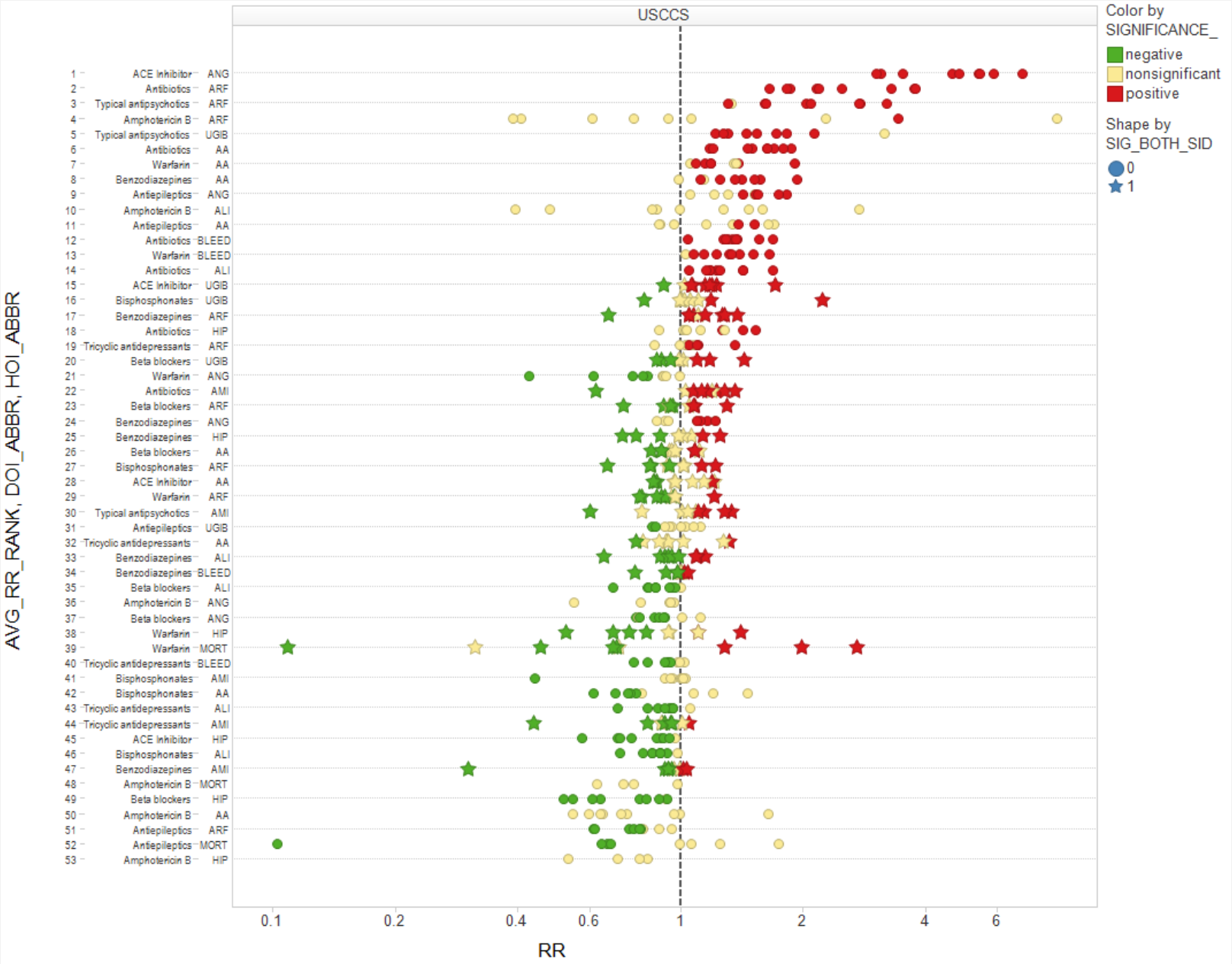
Essential Hypertenson



Heterogeneity Across Databases



Heterogeneity Across Databases

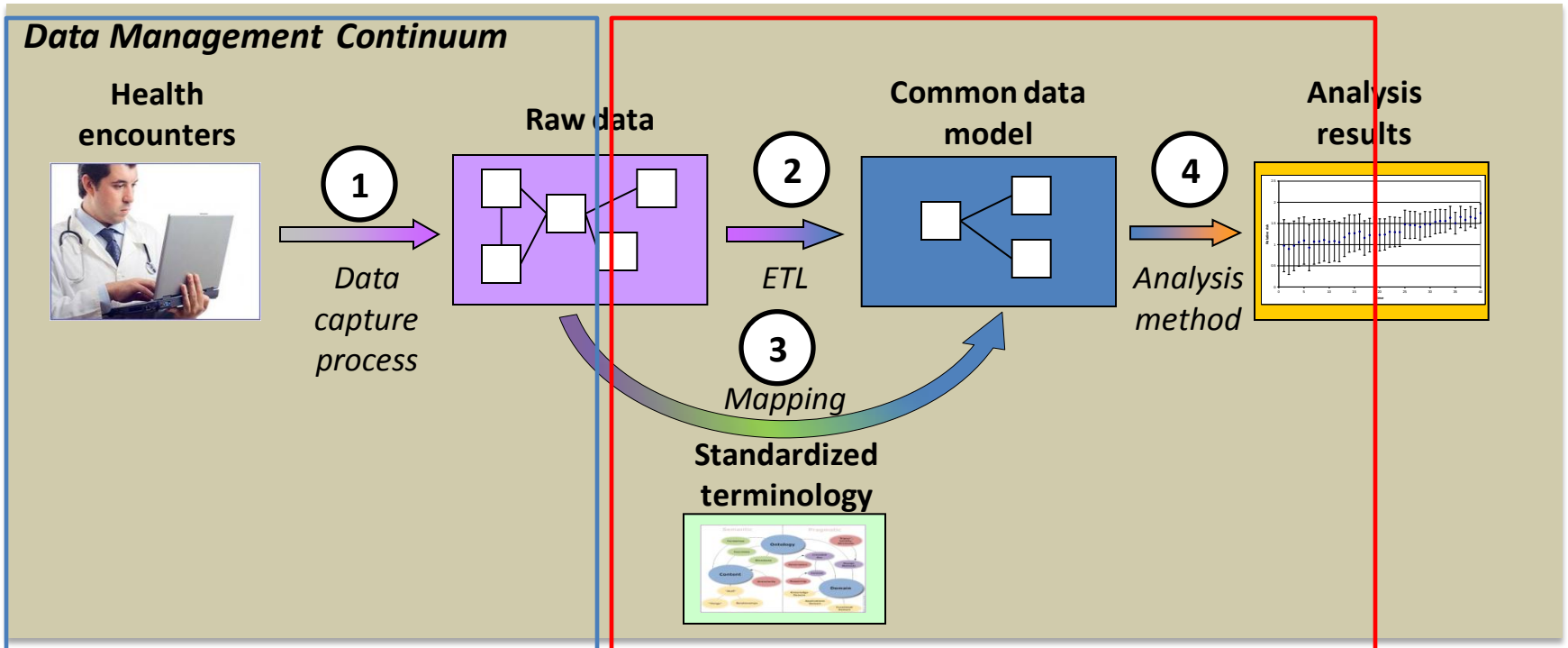


Distributed queries unambiguously define a population from a larger set

Questions about disease outbreaks, prevention activities, health research, quality measures, etc.

OBSERVATIONAL
MEDICAL
OUTCOMES
PARTNERSHIP

Data Management Continuum



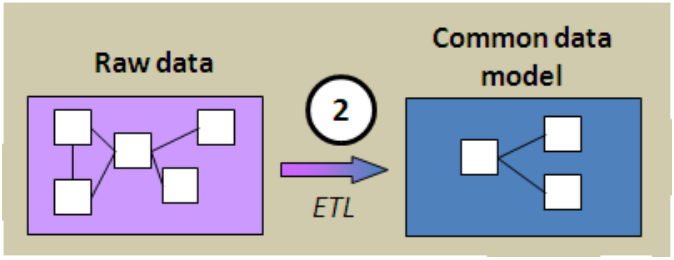
In contrast to clinical trials, not controlled by drug outcome researcher

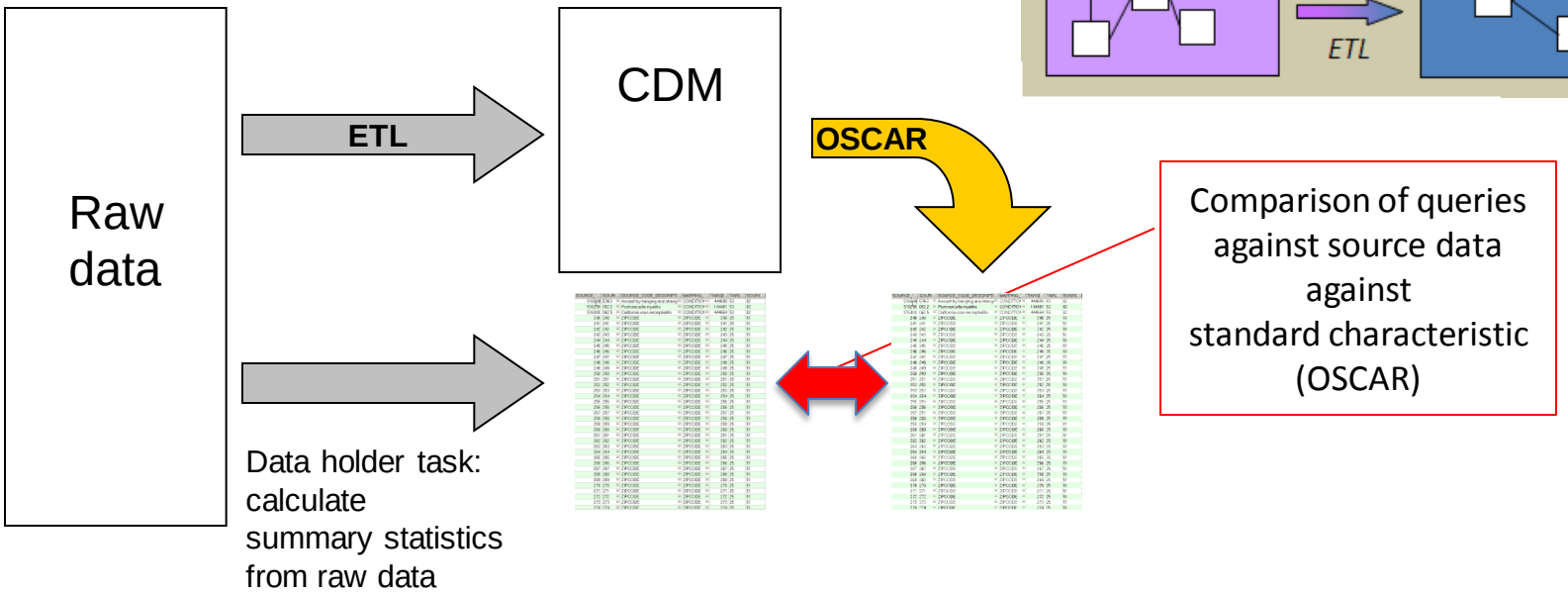
Controlled by outcome researcher

Sources of error and bias:

- Insurance policies: Variations in coverage, frequent changes
- Incomplete documentation
- Miscoding
- Transaction errors with insurance

Raw-CDM Summary Comparison





Tested in GE

- Person
  - Gender
  - Race
  - Year of Birth
  - Gender by Age
- Drug
  - Counts of codes
  - Refills
  - Quantity
  - Stop Reason
- Condition
  - Counts of codes
  - Discharge Status

Tested in Thomson Reuters

- Person
  - Gender
  - Year of Birth
  - Geographical region
- Drug
  - Quantity
  - Refill
  - Days Supply,
- Condition
  - Counts of codes
  - Discharge Status
- Procedure
  - Counts of codes
- Visit
  - Counts of codes
  - Start dates, end dates

Raw-CDM Summary Comparison - Results

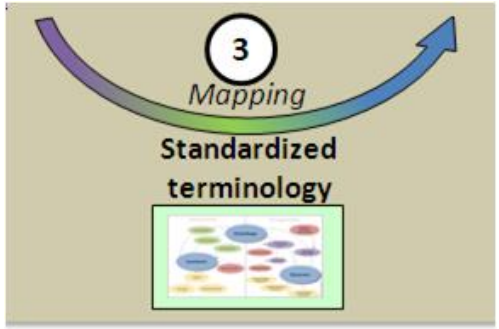
Thomson Reuters databases:

Issue	Impact on HOI or DOI
Zip codes 001-009 incorrectly loaded	No effect on HOI or DOI, no method taking geographical region into account
Procedure drug mapping incorrect, small (%) number of extra procedure drugs	No effect on DOI
Drug quantity rounded, errors in quantity for fractions (like ½ for ointments, etc.)	No effect on DOI, no method taking drug quantity into account

GE database:

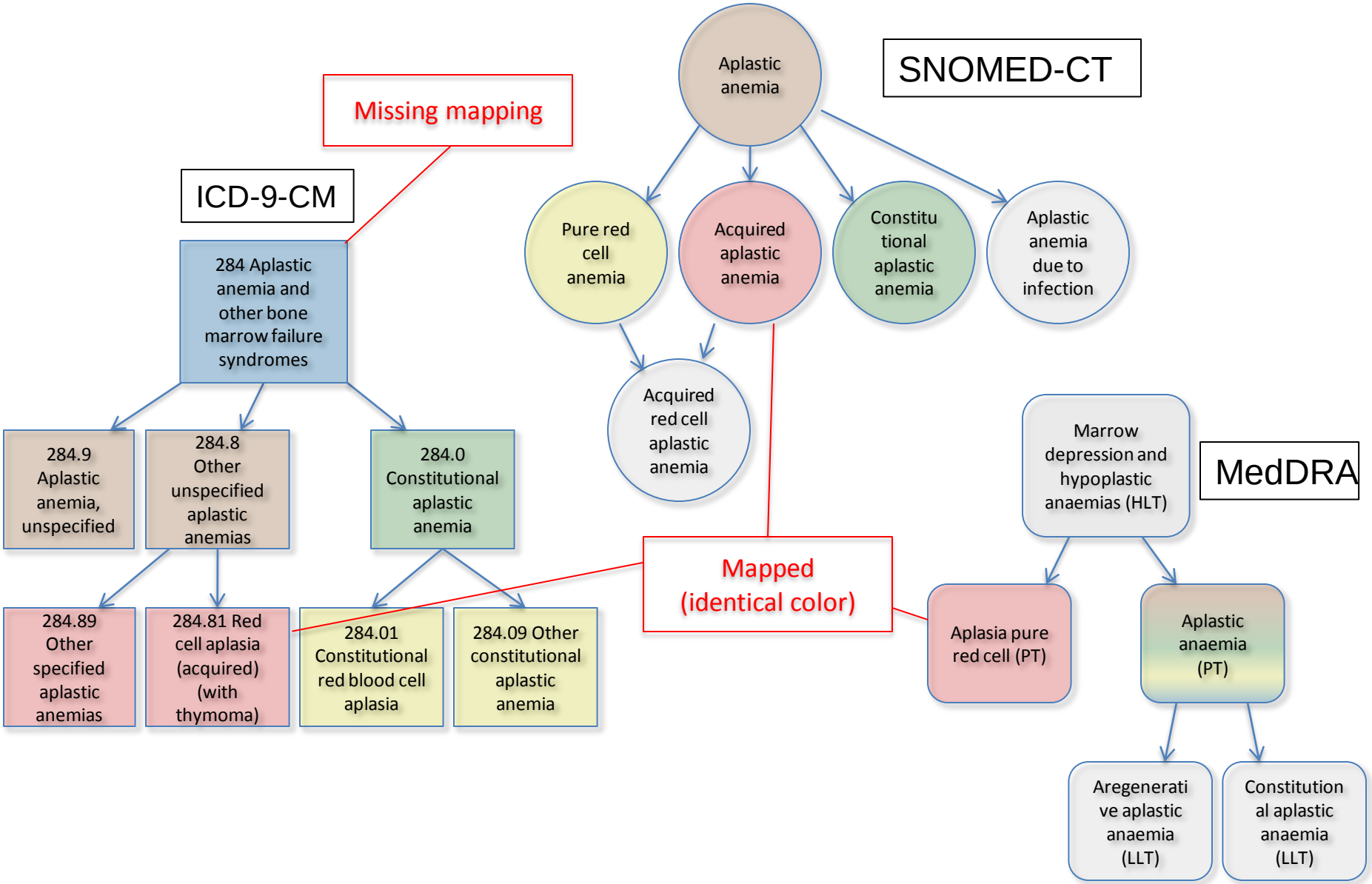
Issue	Impact on HOI or DOI
Gender by age calculated based on 2008, not 2009	No effect on methods
Drug exposure length incorrectly programmed, resulting in values deviating in 3.72% of cases	Small effect on DOI era length
Condition length incorrectly programmed, resulting in values deviating in a small number of cases	Possibly small effect on HOI eral length

Vocabulary Assessement - Conditions



- Potential for quality issues:
  - Incorrect mapping
  - Incomplete mapping
  - Semantic mismatch
  - Hierarchy mismatch
- Quality check SNOMED vs. ICD-9 vs. MedDRA
  1. Spot checking
  2. Comparing record numbers
  3. Comparing whether drug-outcome associations can be reproduced in selected methods
- Test: OMOP HOI
  - Original definition: ICD-9 codes
    - Only HOI used that have no additional diagnostic/therapeutic procedure, lab test, radiology test or EKG definition

Terminology Mapping Artifacts



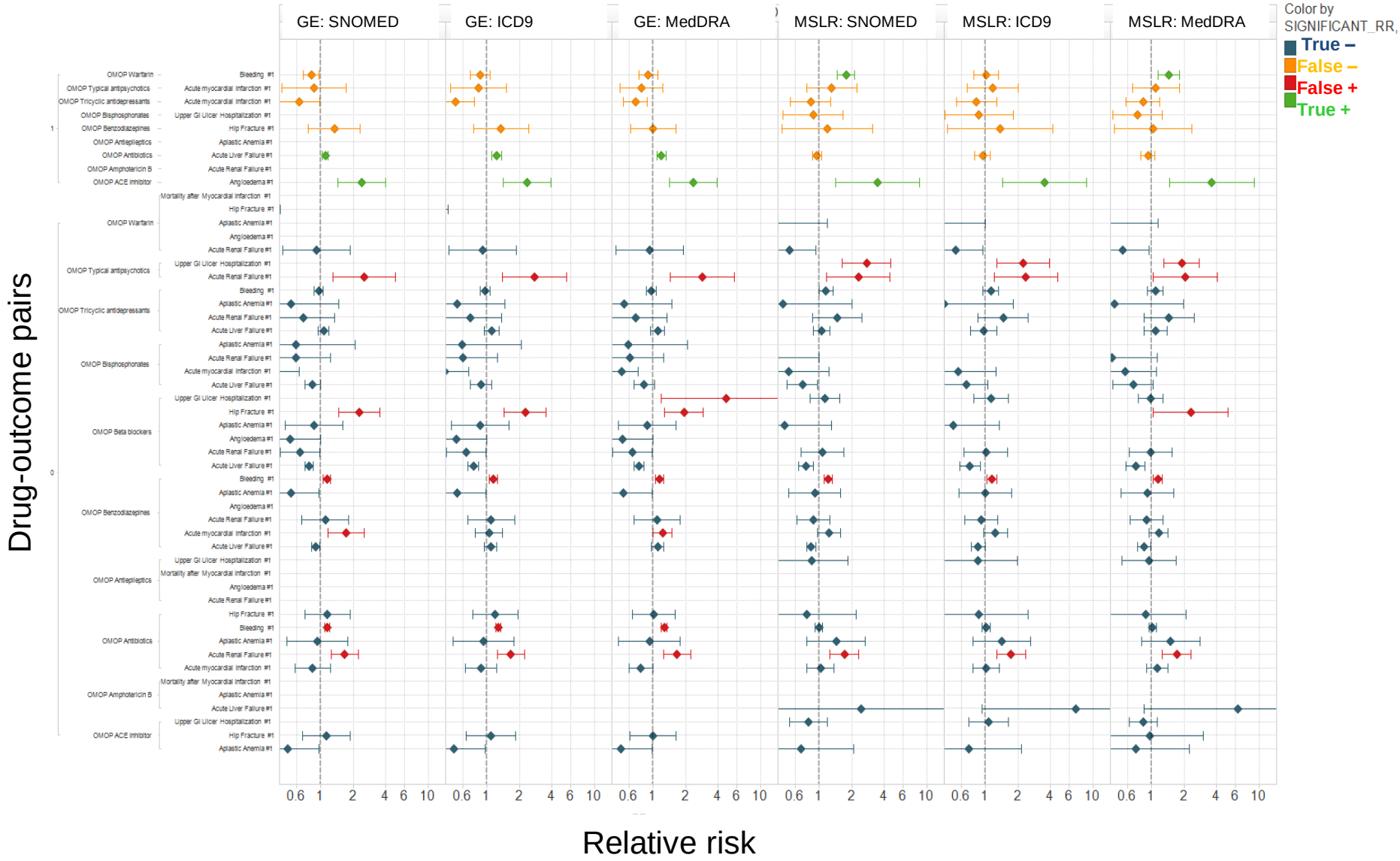
Summary of Terminology Mapping Artifacts

Artifact	Resulting in
1. Codes are wrongly mapped	Wrong data
2. Codes are not mapped	Missing data
3. Many to one mapping	Recruiting data for related codes
4. Child concepts of mapped codes	Recruiting data for related codes

What are the effects of these artifacts on a method's ability to detect drug-outcome relationships?

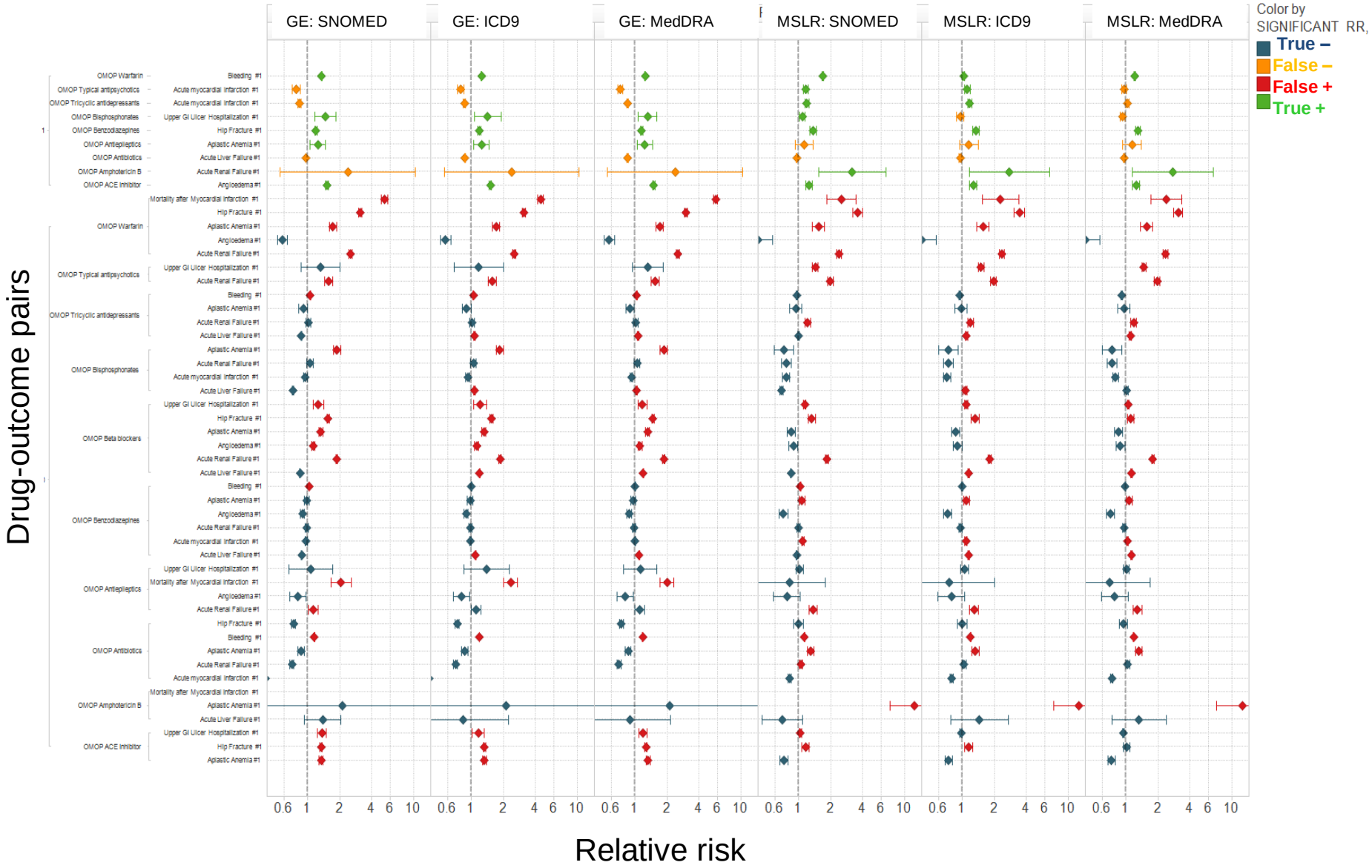
OBSERVATIONAL
MEDICAL
OUTCOMES
PARTNERSHIP

Sensitivity to Vocabulary: Method HDPS



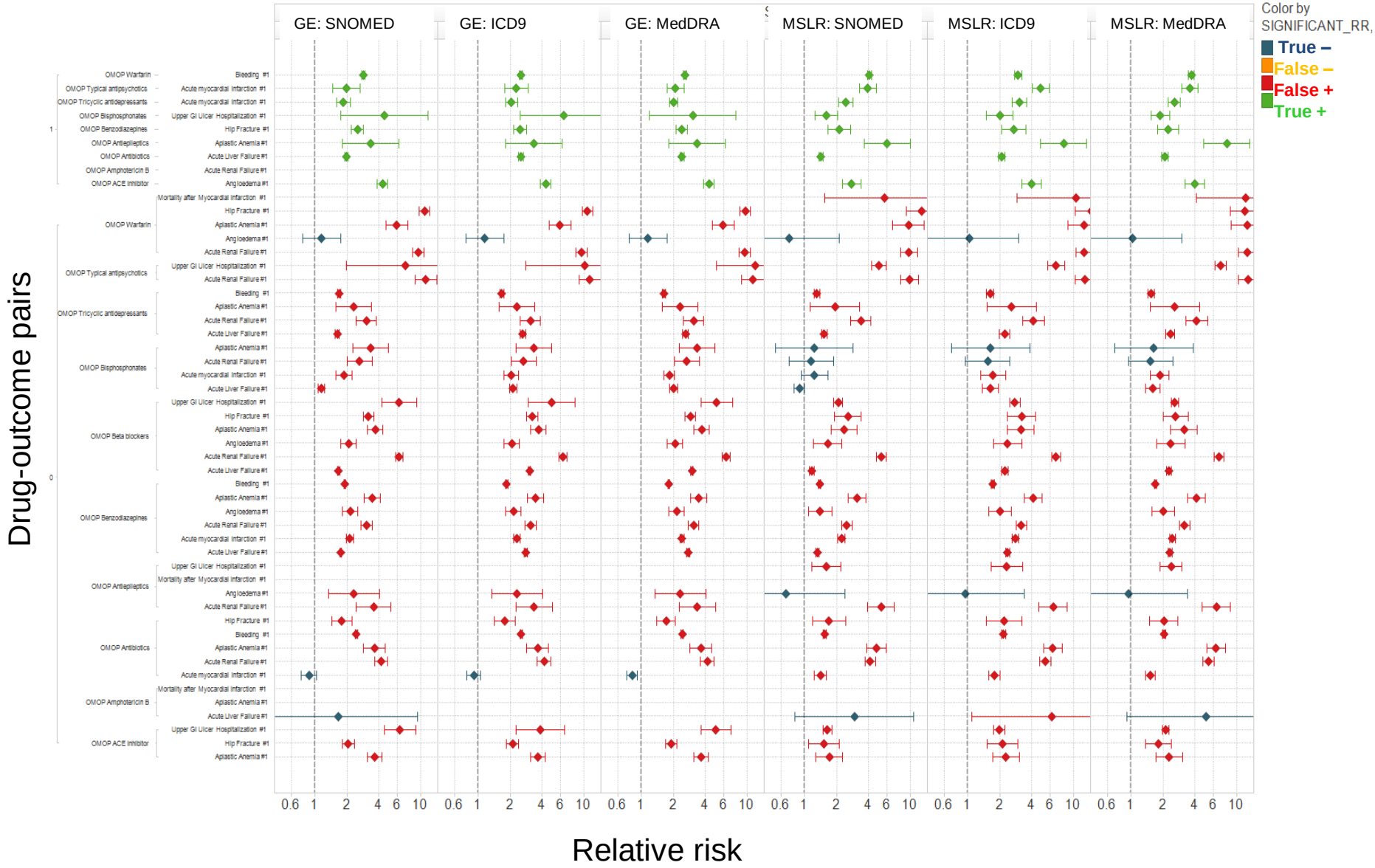
OBSERVATIONAL
MEDICAL
OUTCOMES
PARTNERSHIP

Sensitivity to Vocabulary: Method DP



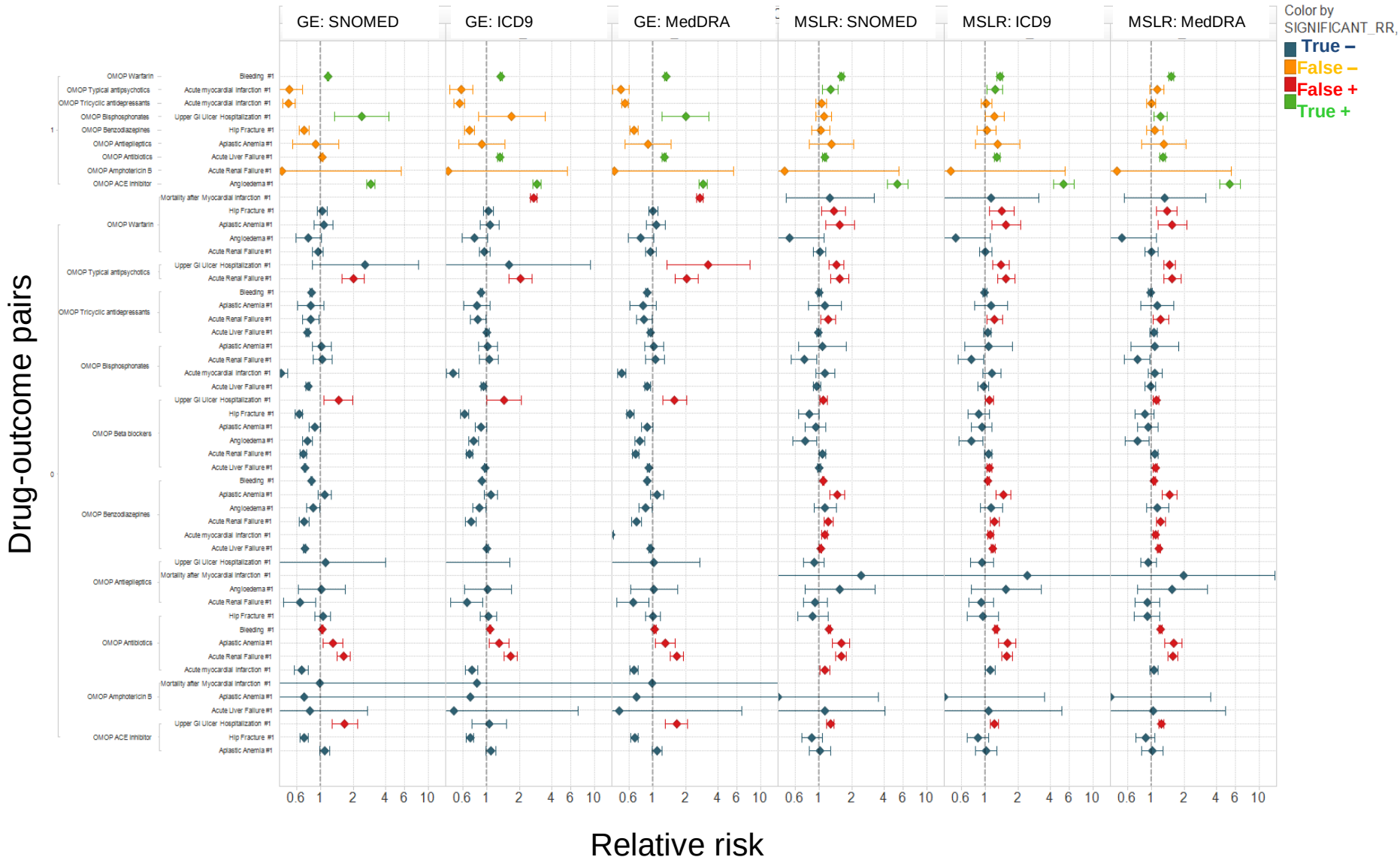
OBSERVATIONAL
MEDICAL
OUTCOMES
PARTNERSHIP

Sensitivity to Vocabulary: Method OS



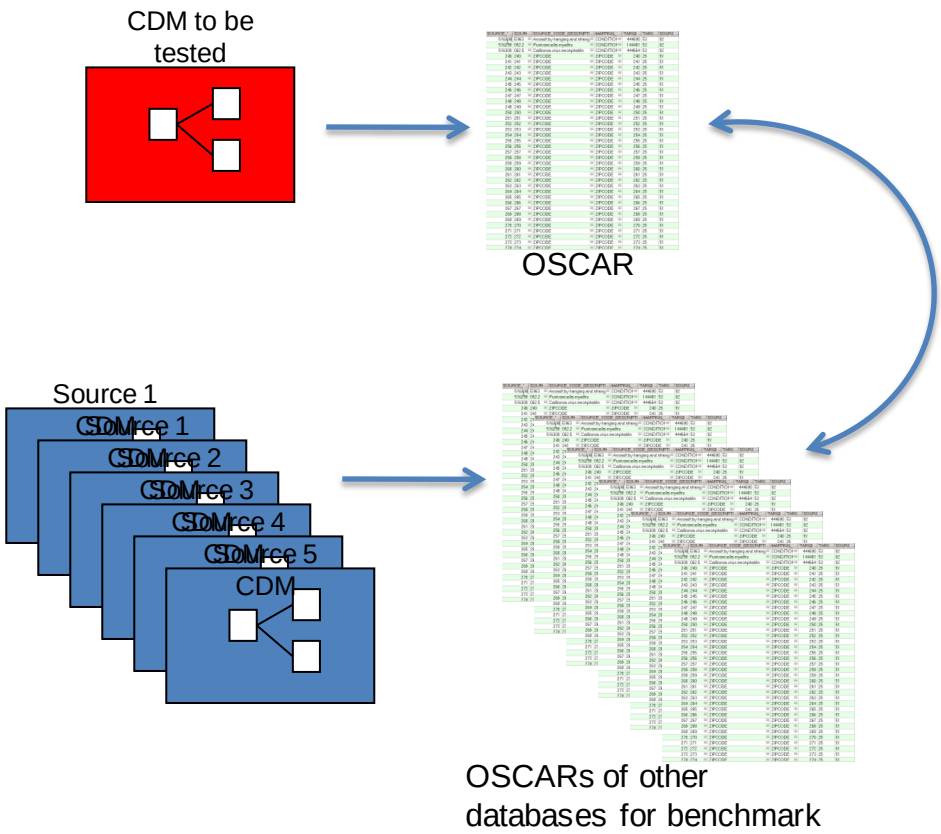
OBSERVATIONAL
MEDICAL
OUTCOMES
PARTNERSHIP

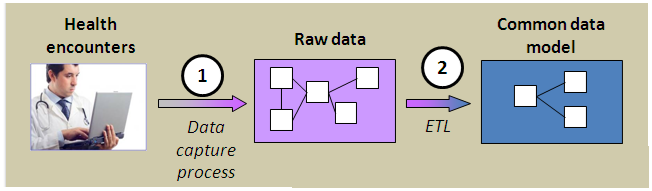
Sensitivity to Vocabulary: Method USCCS



GROUCH

GROUCH produces a summary report from OSCAR for each concept:





GROUCH detects data anomalies:

1. Concept –
   existence and relative frequency of codes compared to benchmark
   - Invalid concepts
   - Concepts appear in one source, not in others
   - Prevalence in one source is statistically different from others
2. Boundary –
   suspicious or implausible values
   - Dates outside range (e.g. drug end date < drug start date)
   - Implausible values (e.g. year of birth > 2010)
   - Suspicious data (e.g. days supply > 180)
3. Temporal –
   patterns over time
   - Unstable rates over time

Summary MSLR GROUCH – Temporal Checks

Warning text	Number of affected Variables	Total amount of warnings
Spike (Gain/loss of 20% or more followed by a 20% loss/gain)	2	8
More than a 100% growth from previous timepoint	2	6

warning_text	VARIABLE_NAME	Observation month or Year of Birth	statistic_value
More than a 100% growth from previous timepoint	observation_month	01/01/2006	612768
Spike (Gain/loss of 20% or more followed by a 20% loss/gain)	observation_month	01/01/2006	612768
Spike (Gain/loss of 20% or more followed by a 20% loss/gain)	observation_month	09/01/2007	835548
More than a 100% growth from previous timepoint	observation_month	01/01/2004	668573
Spike (Gain/loss of 20% or more followed by a 20% loss/gain)	observation_month	02/01/2003	182644
Spike (Gain/loss of 20% or more followed by a 20% loss/gain)	observation_month	09/01/2007	424651
Spike (Gain/loss of 20% or more followed by a 20% loss/gain)	observation_month	12/01/2005	531596
More than a 100% growth from previous timepoint	observation_month	01/01/2004	281564
Spike (Gain/loss of 20% or more followed by a 20% loss/gain)	year_of_birth	1900	5
Spike (Gain/loss of 20% or more followed by a 20% loss/gain)	year_of_birth	1901	0
Spike (Gain/loss of 20% or more followed by a 20% loss/gain)	year_of_birth	1904	0
More than a 100% growth from previous timepoint	year_of_birth	1908	17
More than a 100% growth from previous timepoint	year_of_birth	1909	44
More than a 100% growth from previous timepoint	observation_month	01/01/2004	364802

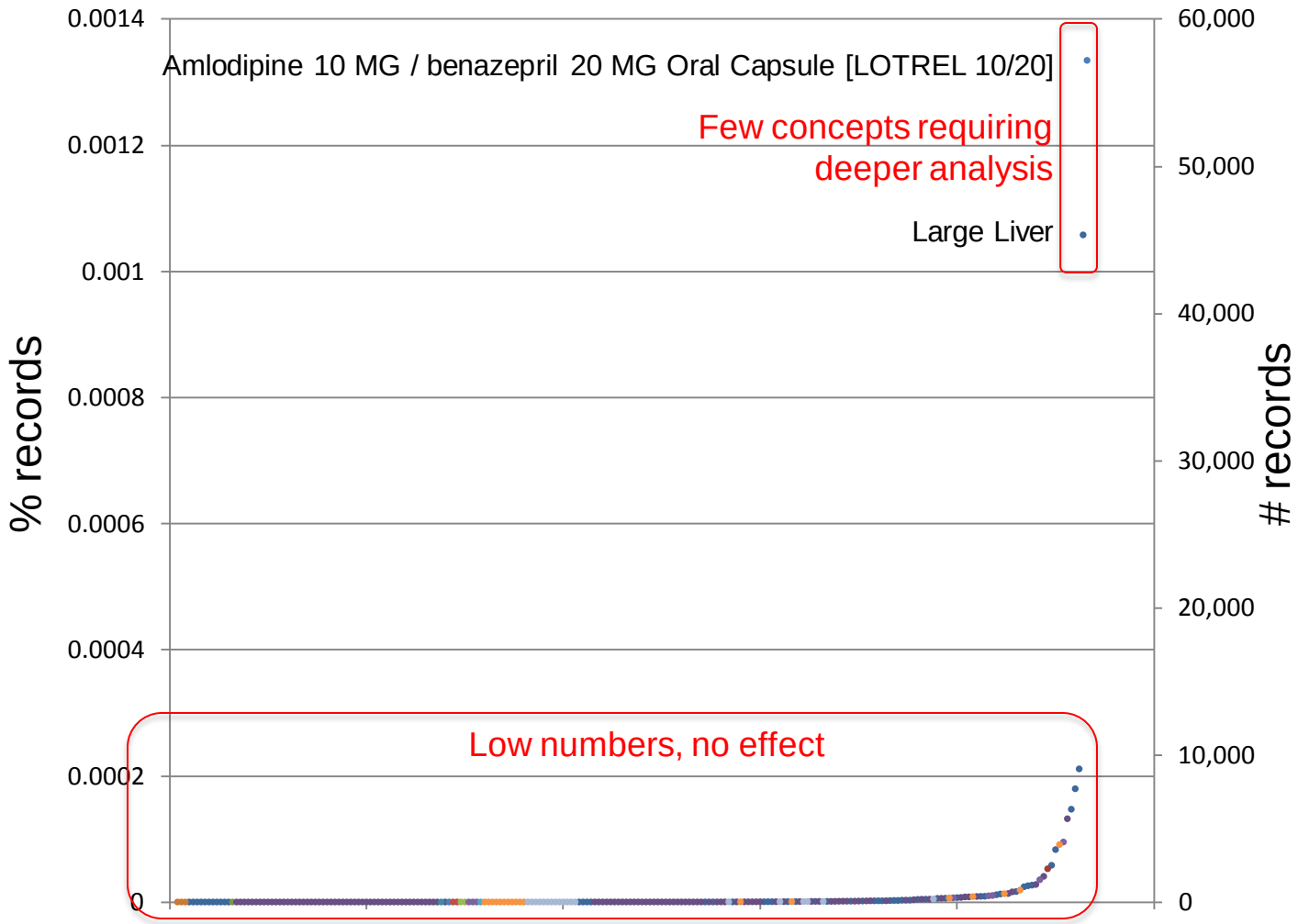
Conclusions: MSLR has large spikes in enrollment at start of each year

Summary MSLR GROUCH – Concept Checks

Warning text	Number of affected Variables	Total amount of warnings	Affecting a HOI or DOI	..and >.1% of records
Concept not in vocabulary	5	5	0	0
Concept only found in this source	7	3445	14	0
Concept only in all other sources EXCEPT this source	6	4984	167	0
Concept exists at a rate more than 3 standard deviations from the mean of the other sources	11	5217	126	2
Average number of records per person more than 3 standard deviations from the mean of the other sources	0	0	0	0
Maximum number of records per person more than 3 standard deviations from the average maximum of the other sources	0	0	0	0
Concept only found in this source (Male)	3	1016	22	0
Concept only found in this source (Female)	3	835	12	0
Concept only in all other sources EXCEPT this source (Male);	3	4790	121	0
Concept only in all other sources EXCEPT this source (Female);	3	3773	95	0
Concept exists at a rate more than 3 standard deviations from the mean of the other sources (Male)	3	3465	67	0
Concept exists at a rate more than 3 standard deviations from the mean of the other sources (Female)	3	4129	83	0

126 concepts are observed at a notacibly different frequency in MSLR compared to other databases
2 of them are not very rare in the cohort

GROUCH Warning affecting HOI and DOI

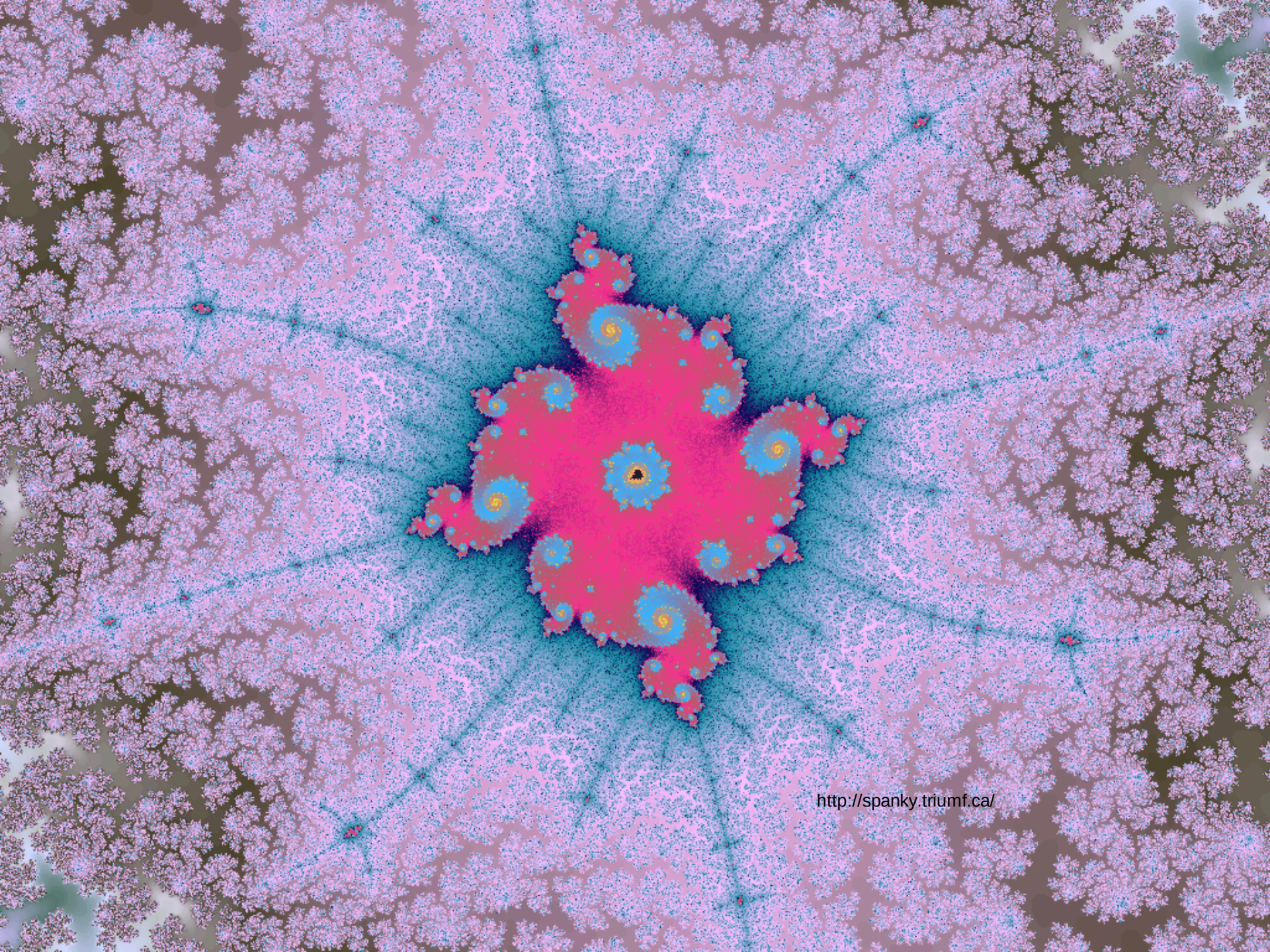


HOI and DOI concepts: Frequency > 3 standard deviation from average

Summary MSLR GROUCH – Boundary Checks

Warning text	Number of affected Variables	Total amount of warnings	Affecting a HOI or DOI
Year of Birth before 1900	1	2	0
Year of Birth after 2010	0	0	0
Date before Earliest Observation Start Date for the Datasource	0	0	0
Date after Last Observation End Date for the Datasource	1	1	0
Days_supply is a missing value	1	1	0
Days_supply is a negative value	1	1	0
Days_supply is a more than 180 days	1	1	0
Refill count is a missing value	1	1	0
Refill count is a negative value	0	0	0
Refill count is more than 10	1	1	0
Drug Quantity is a missing value	1	1	0
Drug Quantity is a negative value	1	1	0
Drug Quantity is more than 600	1	1	0
Drug Exposure Count is a negative value	0	0	0
Drug Exposure Count is more than 100	1	1	0
Condition occurrence count is a negative value	0	0	0
Condition occurrence count is more than 1,000	1	18	0
Age at earliest observation date < 0	0	0	0
Age at earliest observation date > 110	7	21	0
Invalid period length of Period (end date is before start date)	0	0	0
Length is longer than the longest possible length of observation	1	6	0

Conclusion: Small numbers, many of the warning legitimate healthcare situations


http://spanky.triumf.ca/

Key Points

- Data are not patients
- Data are Swiss cheese
- Data hide their meaning
- Data are dynamic over time
- Data may be truncated temporally
- Data are not data
- Data are biased
- Data are never as abundant as they appear
- Not all data comes from patients


The patient is waiting!