

**OBSERVATIONAL  
MEDICAL  
OUTCOMES  
PARTNERSHIP**

**Managing Data Quality for an Active  
Surveillance System**

Christian Reich, MD, PhD

**OMOP Symposium**

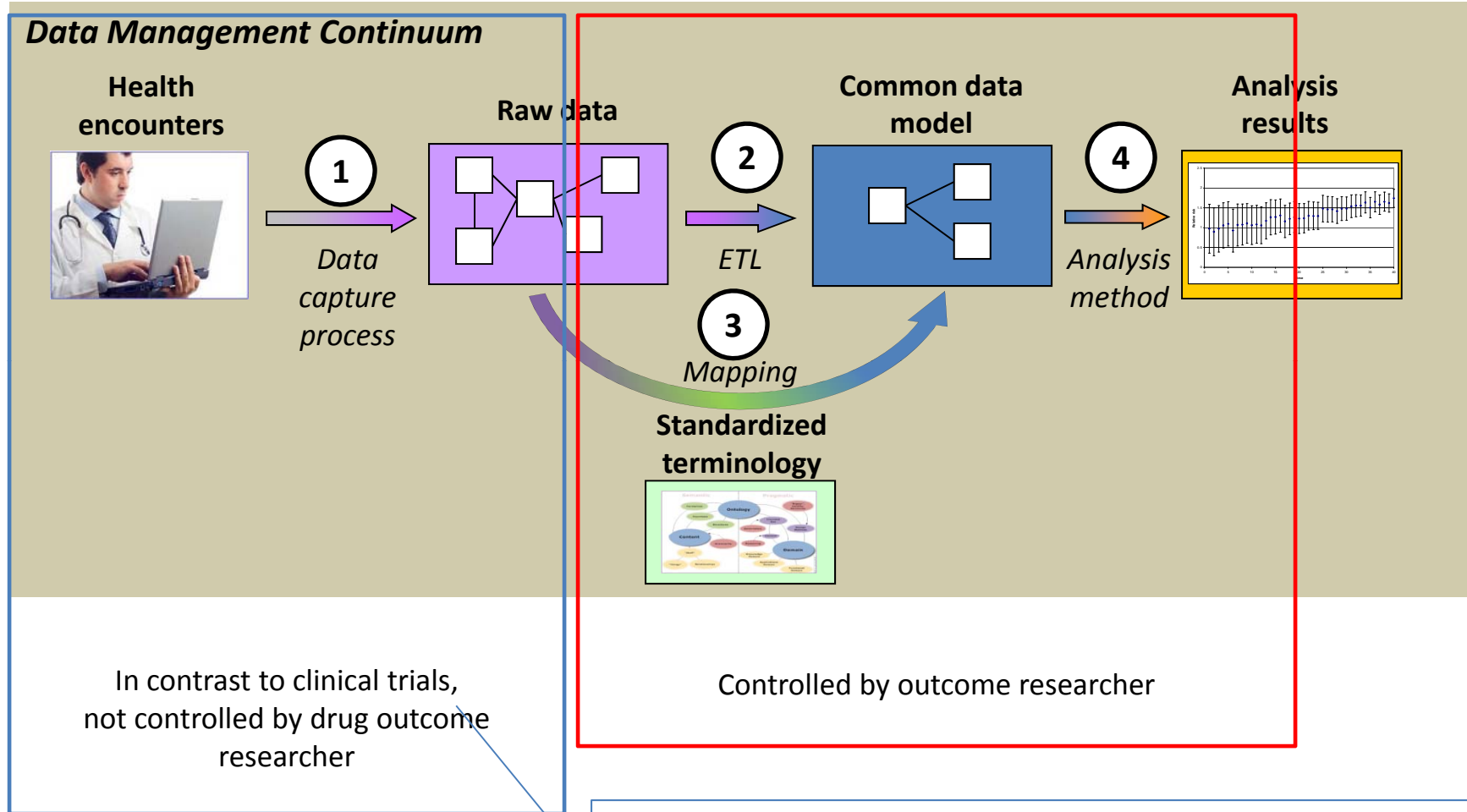
January 11, 2011

**FOUNDATION**  
FOR THE  
National Institutes of Health

# Challenge

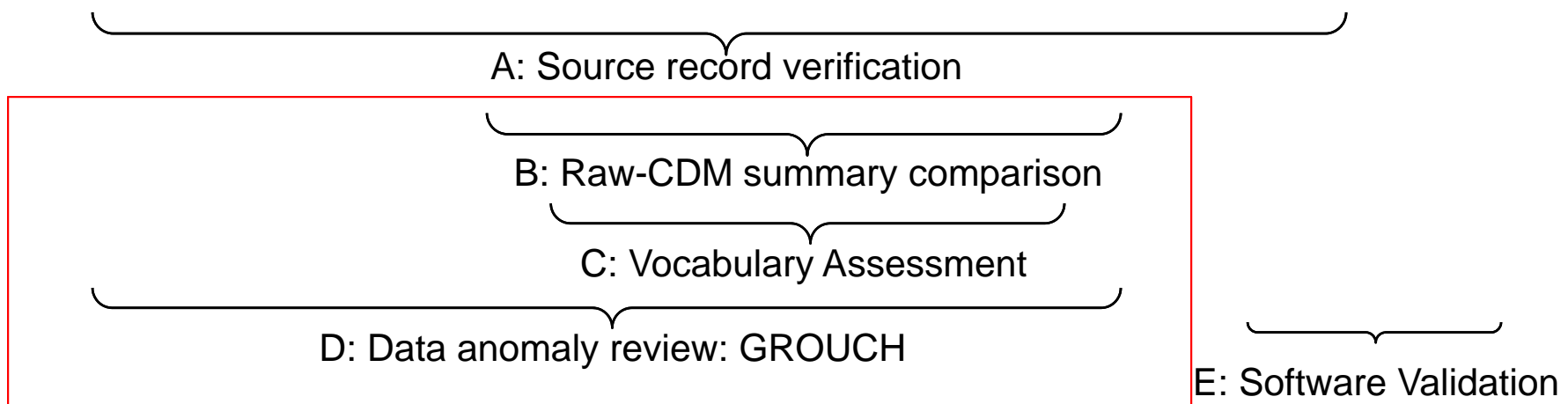
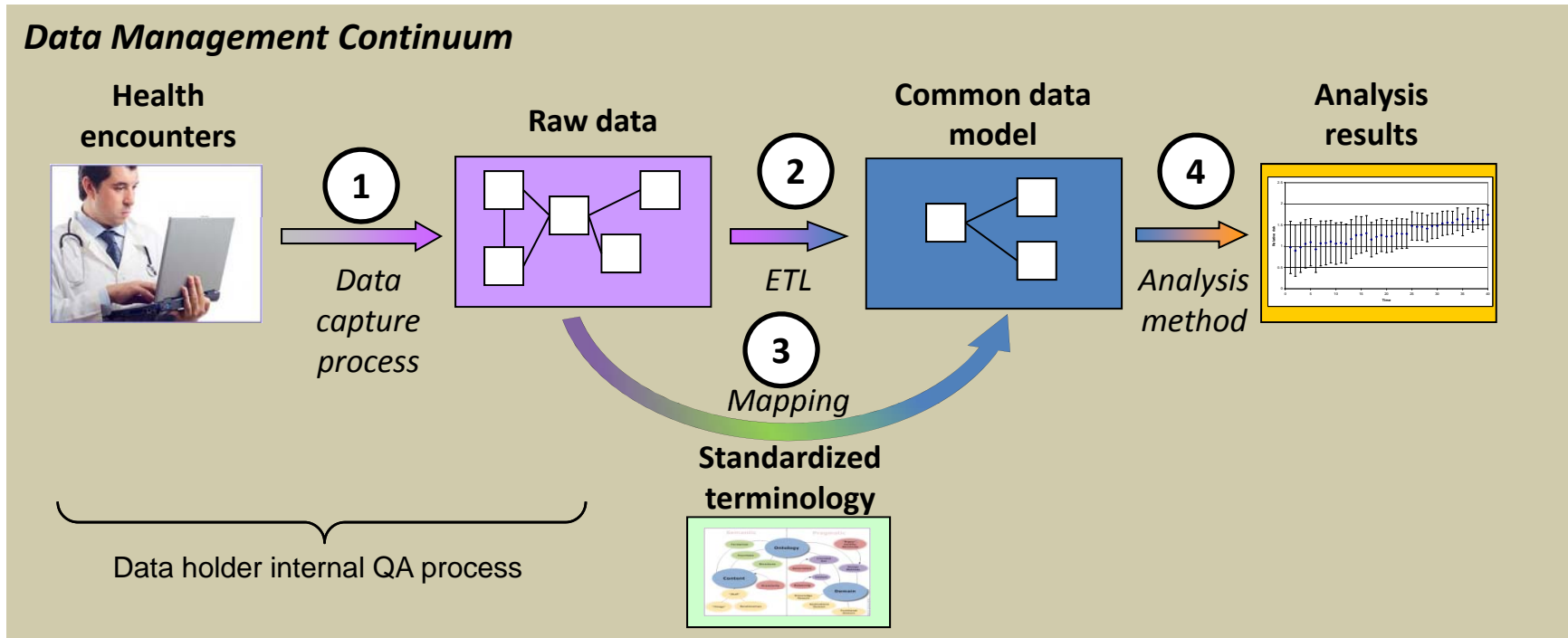
- Active surveillance relies on high quality of data
  - OMOP has manipulated data in two ways:
    - Converting format to Common Data Model
    - Standardizing terminologies
1. Have we impaired the ability to detect drug-outcome associations?
  2. How do we find out in the absence of a Gold Standard?
- Development of a Quality System with Standardized Tools

# Data Management Continuum in Active Surveillance

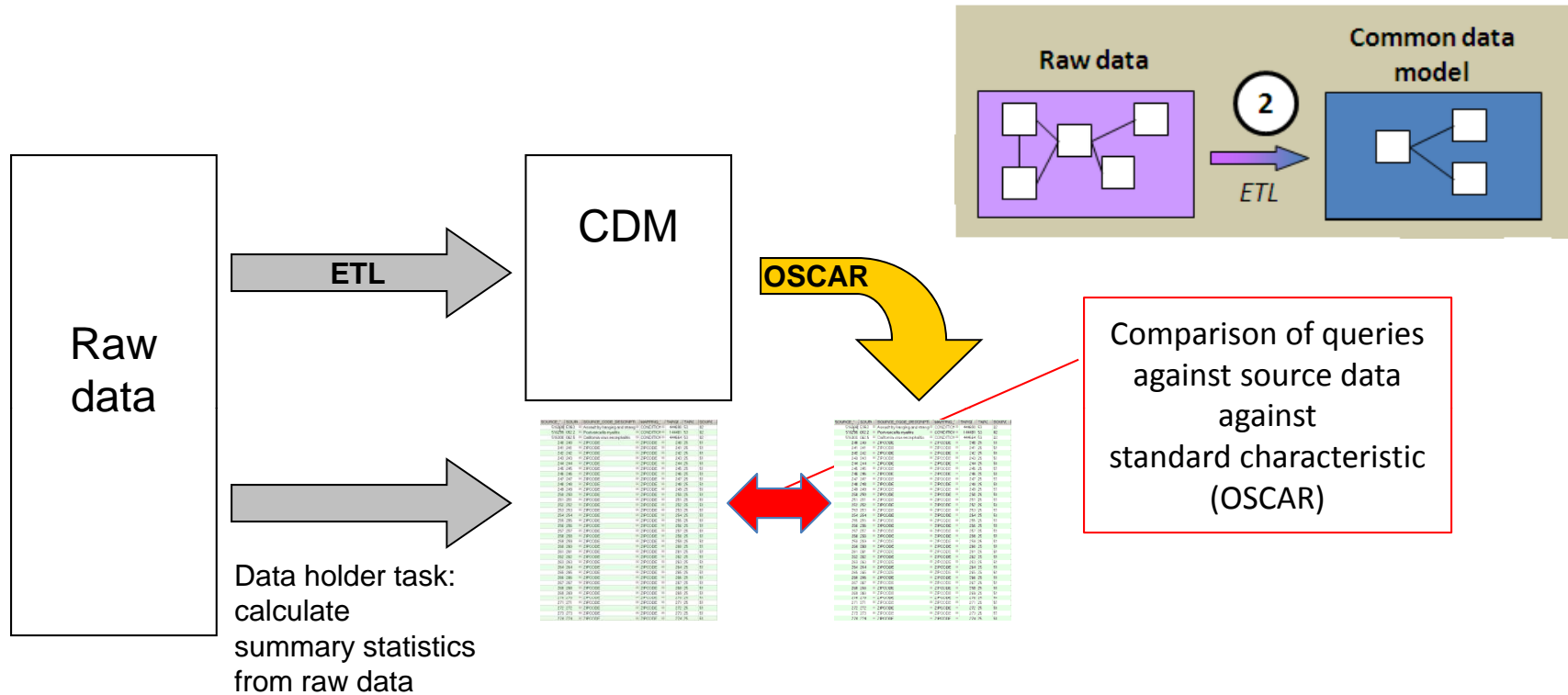


- Sources of error and bias:
- Insurance policies: Variations in coverage, frequent changes
  - Incomplete documentation
  - Miscoding
  - Transaction errors with insurance

# Proposed Quality System



# B: 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

## B: Raw-CDM Summary Comparison - Results

### Thomson Reuters databases:

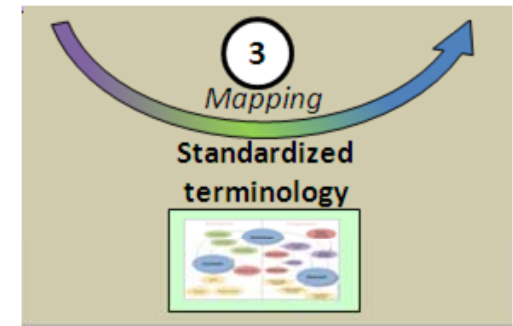
Issue	Impact on HOI or DOI
Zip codes 001-009 incorrectly loaded	No effect on HOI or DOI, no method taking geographical region into account
Procedure drug mapping incorrect, small (%) number of extra procedure drugs	No effect on DOI
Drug quantity rounded, errors in quantity for fractions (like ½ for ointments,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

### GE database:

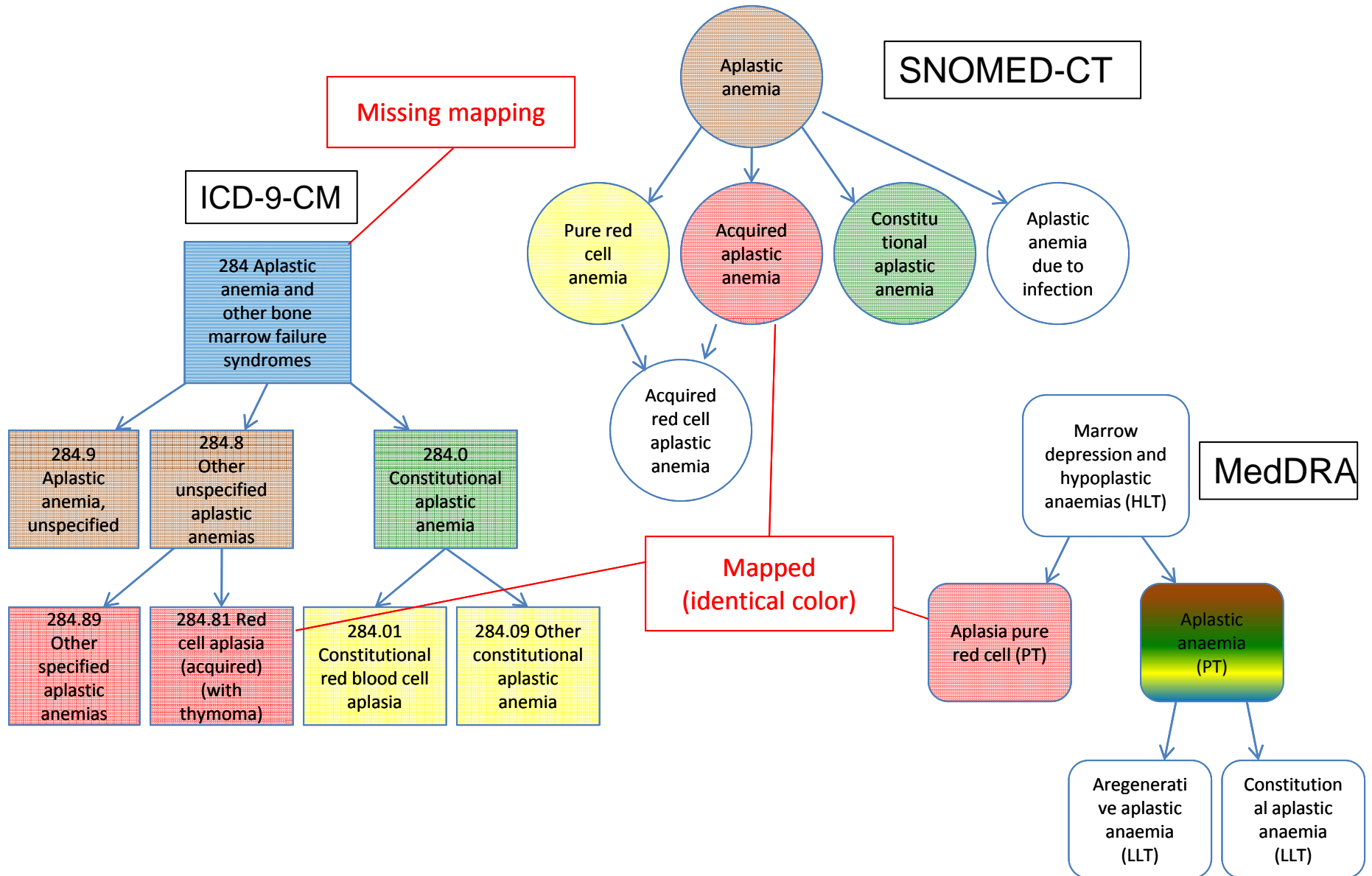
Issue	Impact on HOI or DOI
Gender by age calculated based on 2008, not 2009	No effect on methods
Drug exposure length incorrectly programmed, resulting in values deviating in 3.72% of cases	Small effect on DOI era length
Condition length incorrectly programmed, resulting in values deviating in a small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

## C: Vocabulary Assessment - Conditions

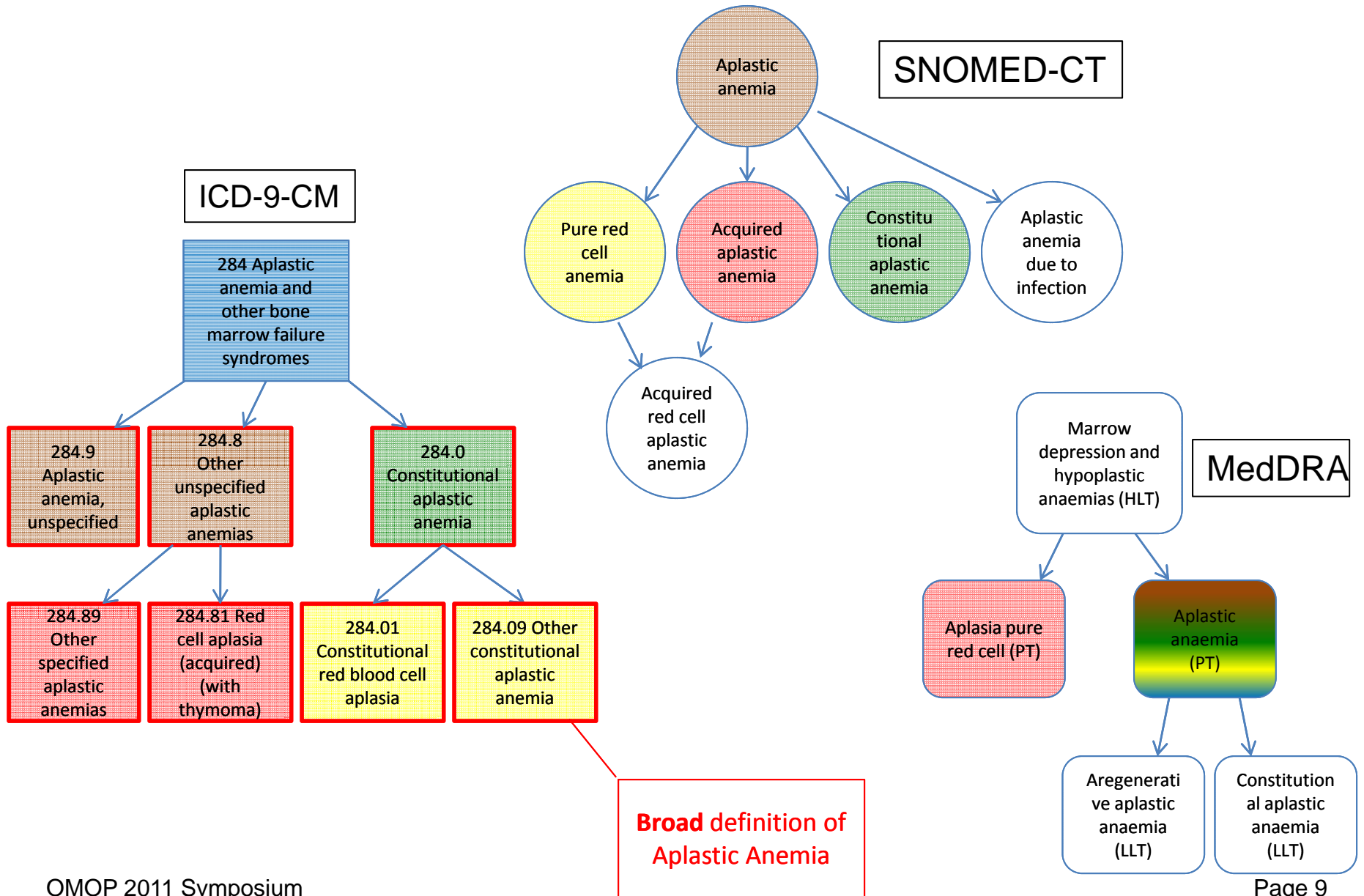
- Potential for quality issues:
  - Incorrect mapping
  - Incomplete mapping
  - Semantic mismatch
  - Hierarchy mismatch
- Quality check SNOMED vs. ICD-9 vs. MedDRA
  1. Spot checking
  2. Comparing record numbers
  3. Comparing whether drug-outcome associations can be reproduced in selected methods
- 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



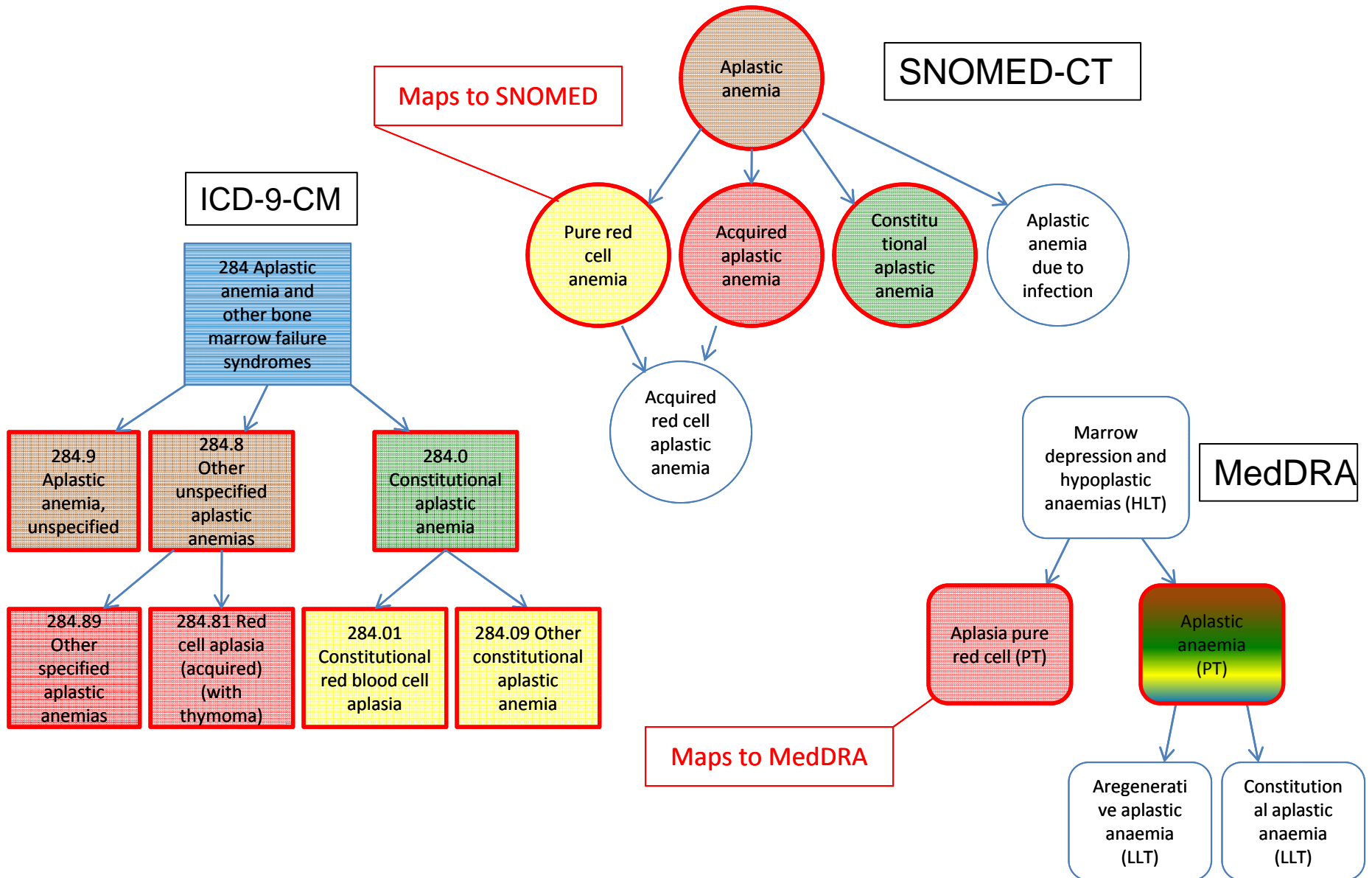
# Terminology Mapping Artifacts



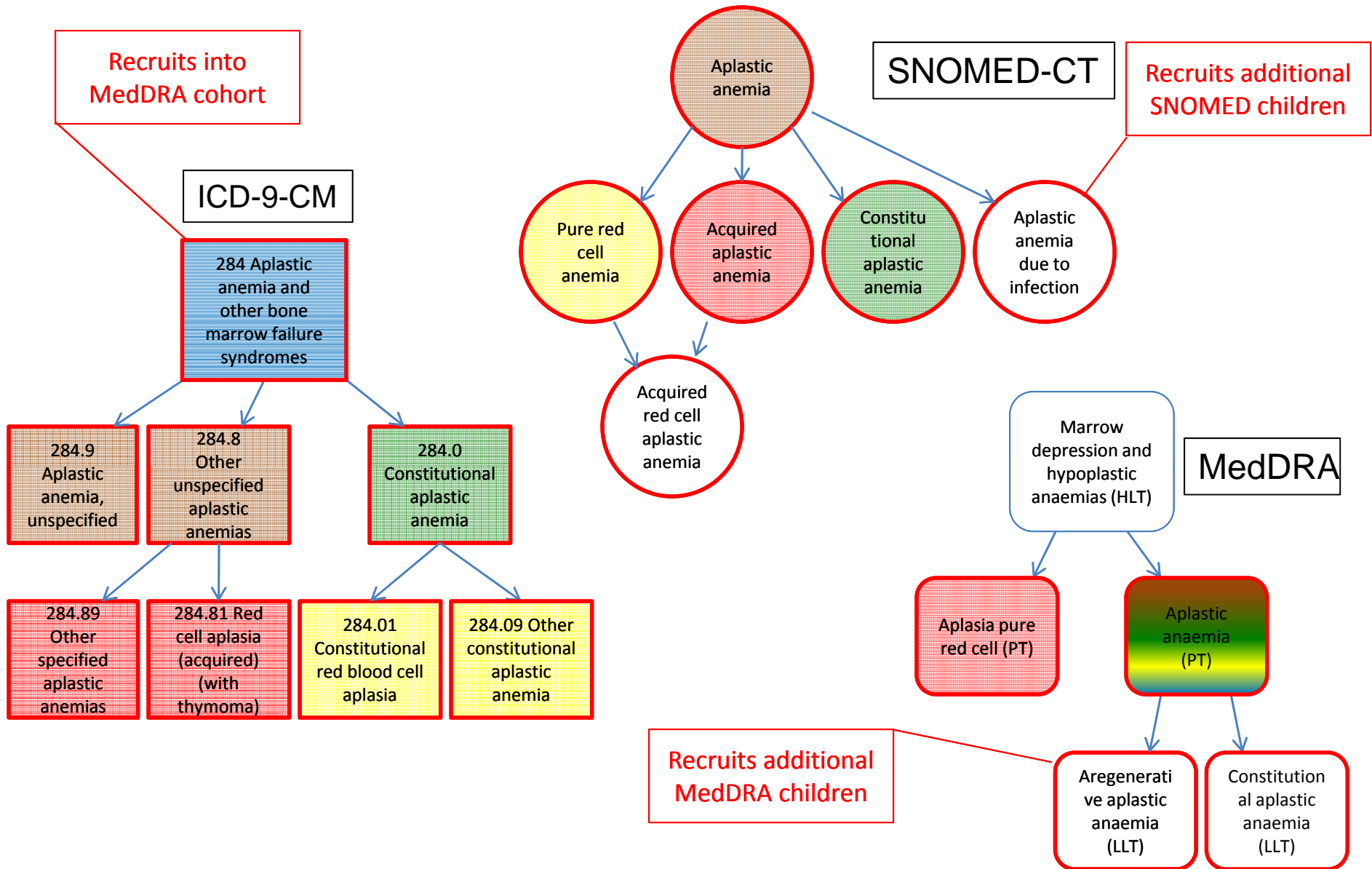
# Terminology Mapping Artifacts



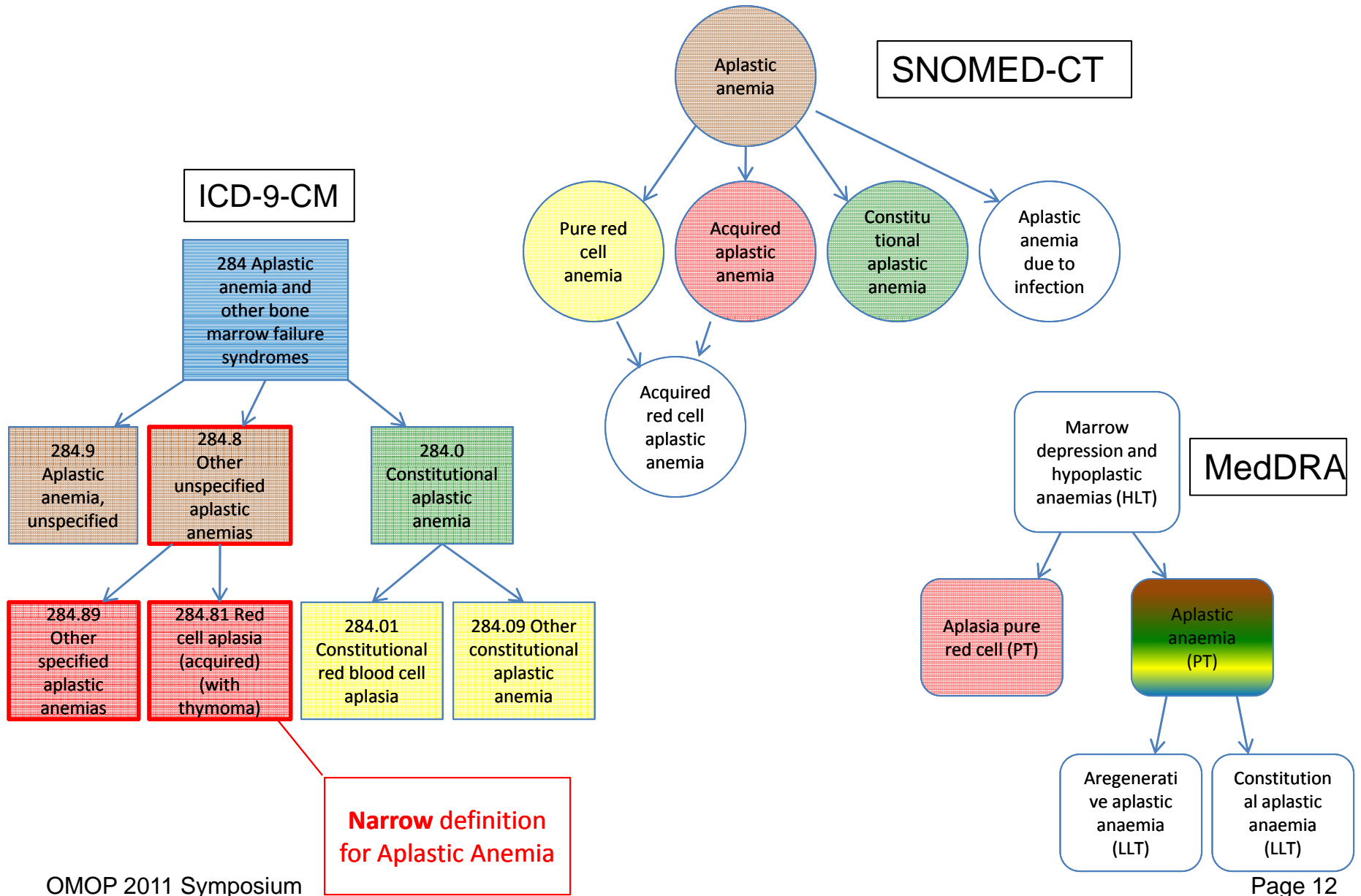
# Terminology Mapping Artifacts



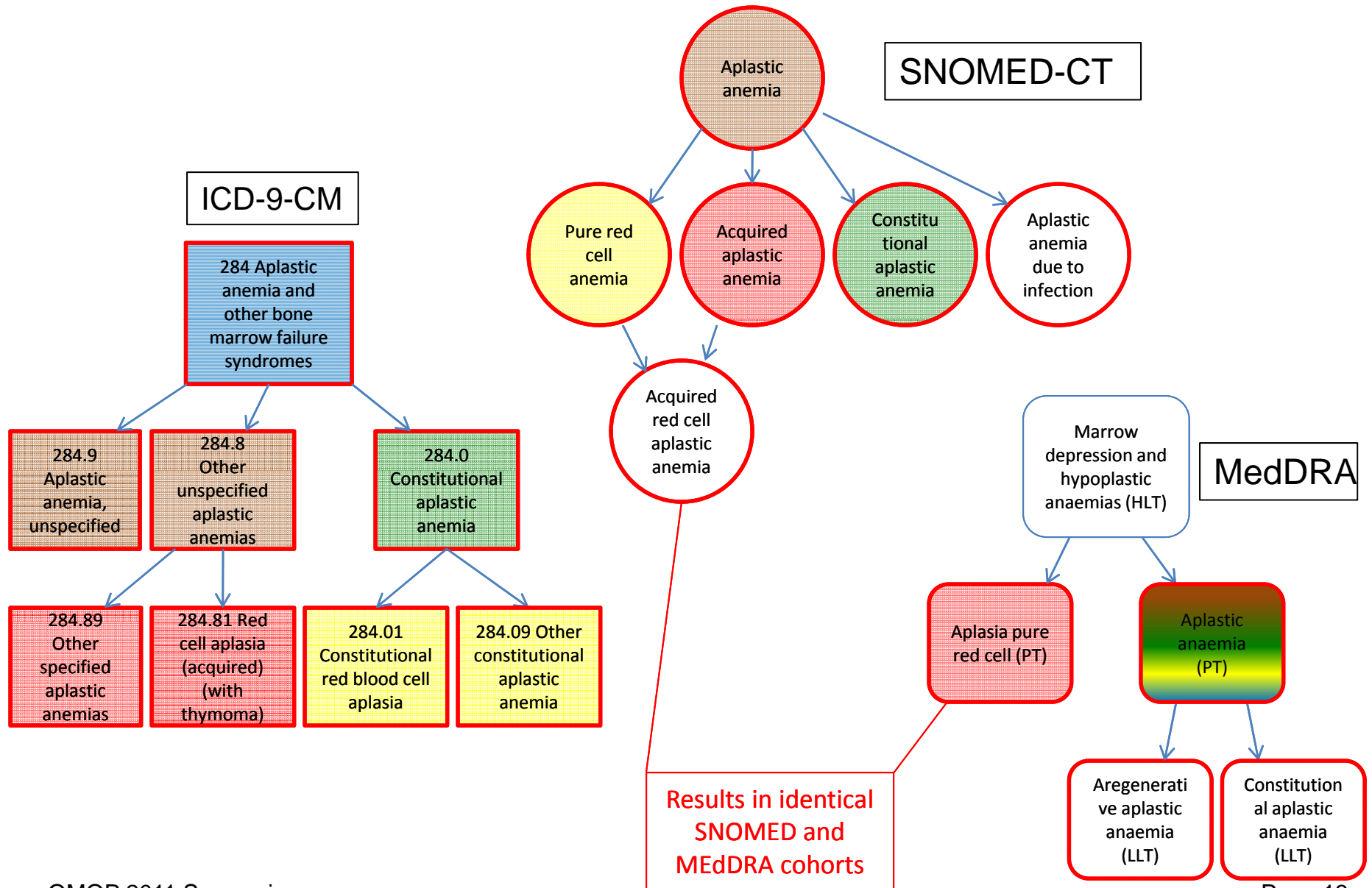
# Terminology Mapping Artifacts



# Terminology Mapping Artifacts



# Terminology Mapping Artifacts

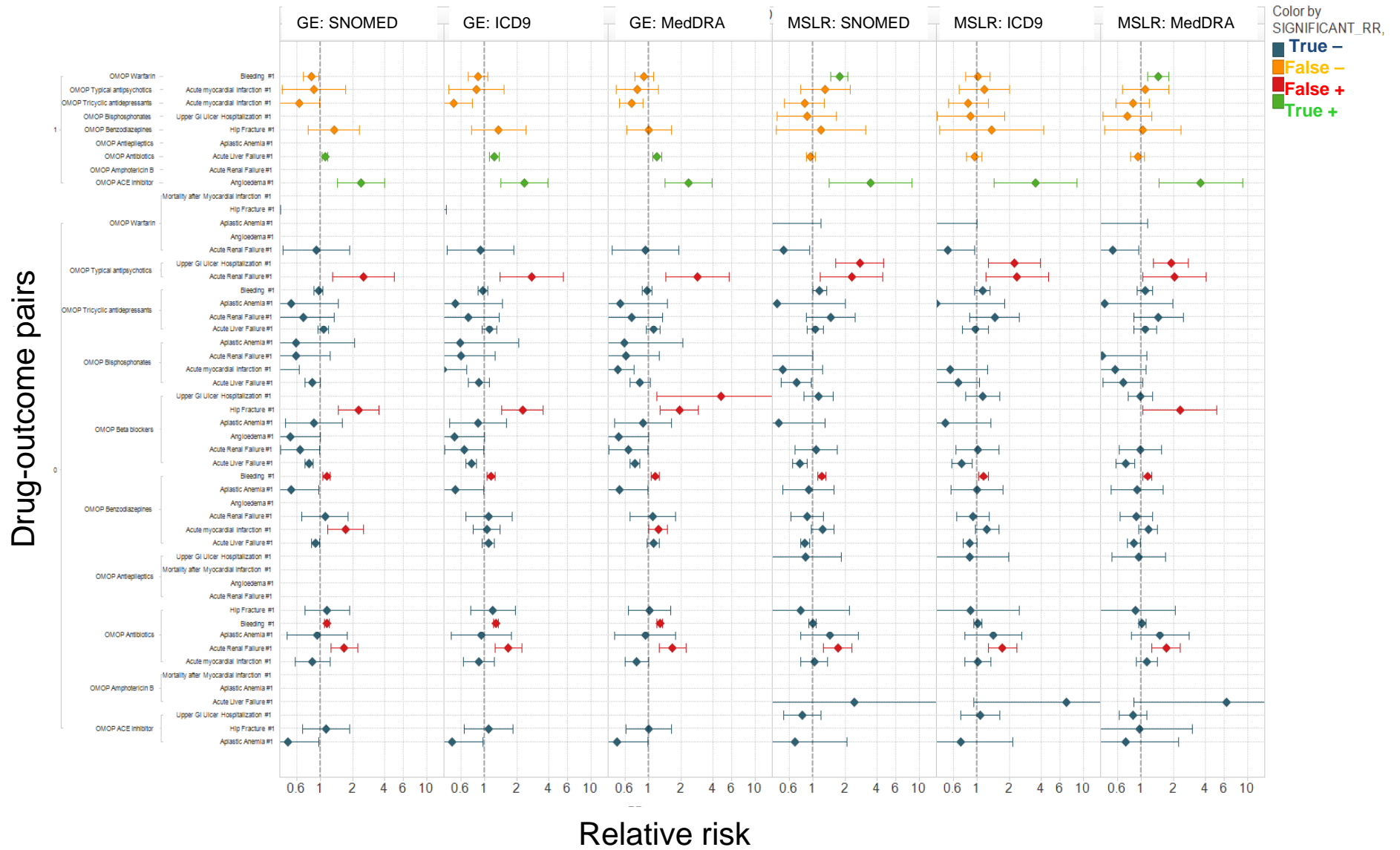


# Summary of Terminology Mapping Artifacts

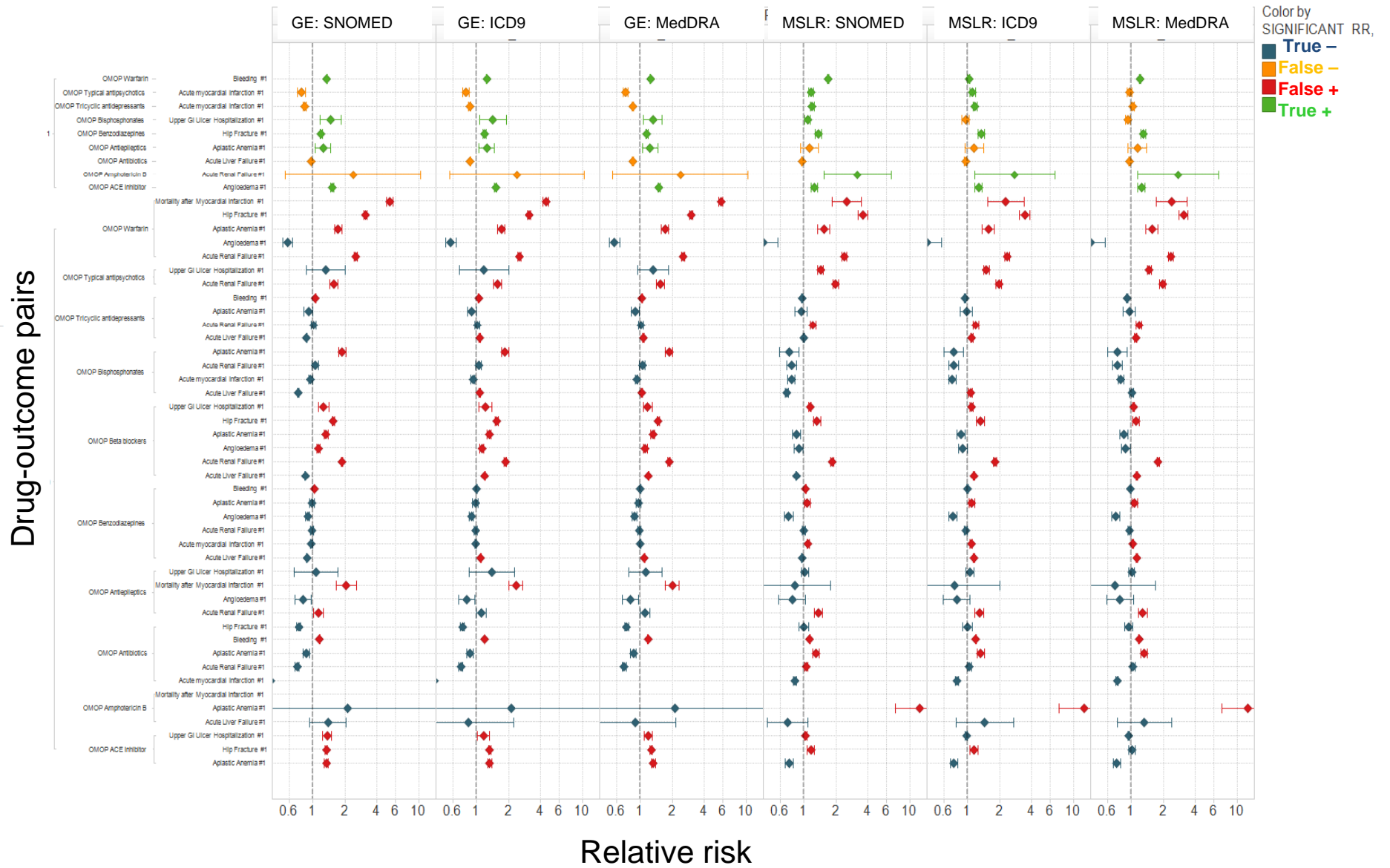
Artifact	Resulting in
1. Codes are wrongly mapped	Wrong data
2. Codes are not mapped	Missing data
3. Many to one mapping	Recruiting data for related codes
4. Child concepts of mapped codes	Recruiting data for related codes

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

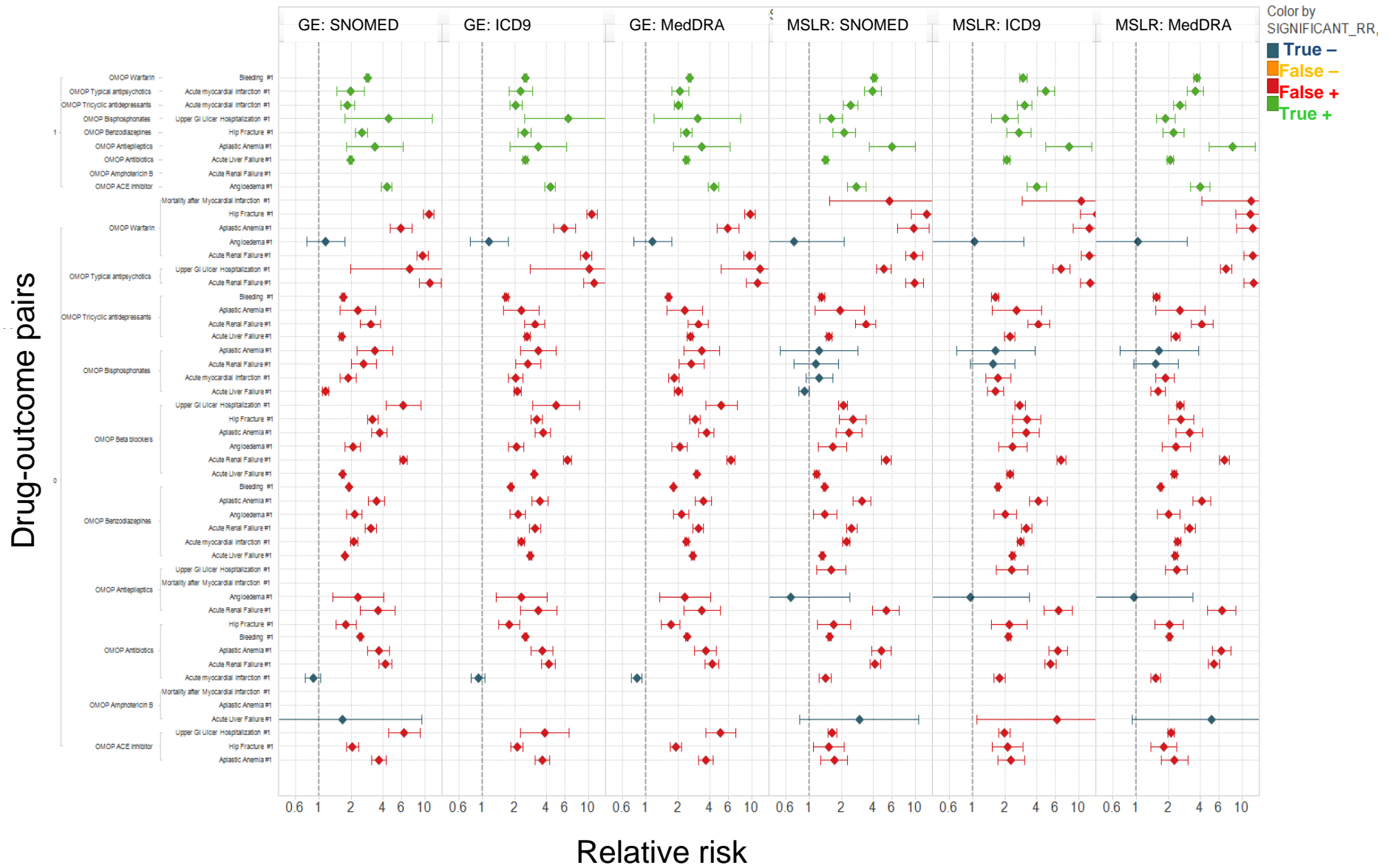
# Sensitivity to Vocabulary: Method HDPS



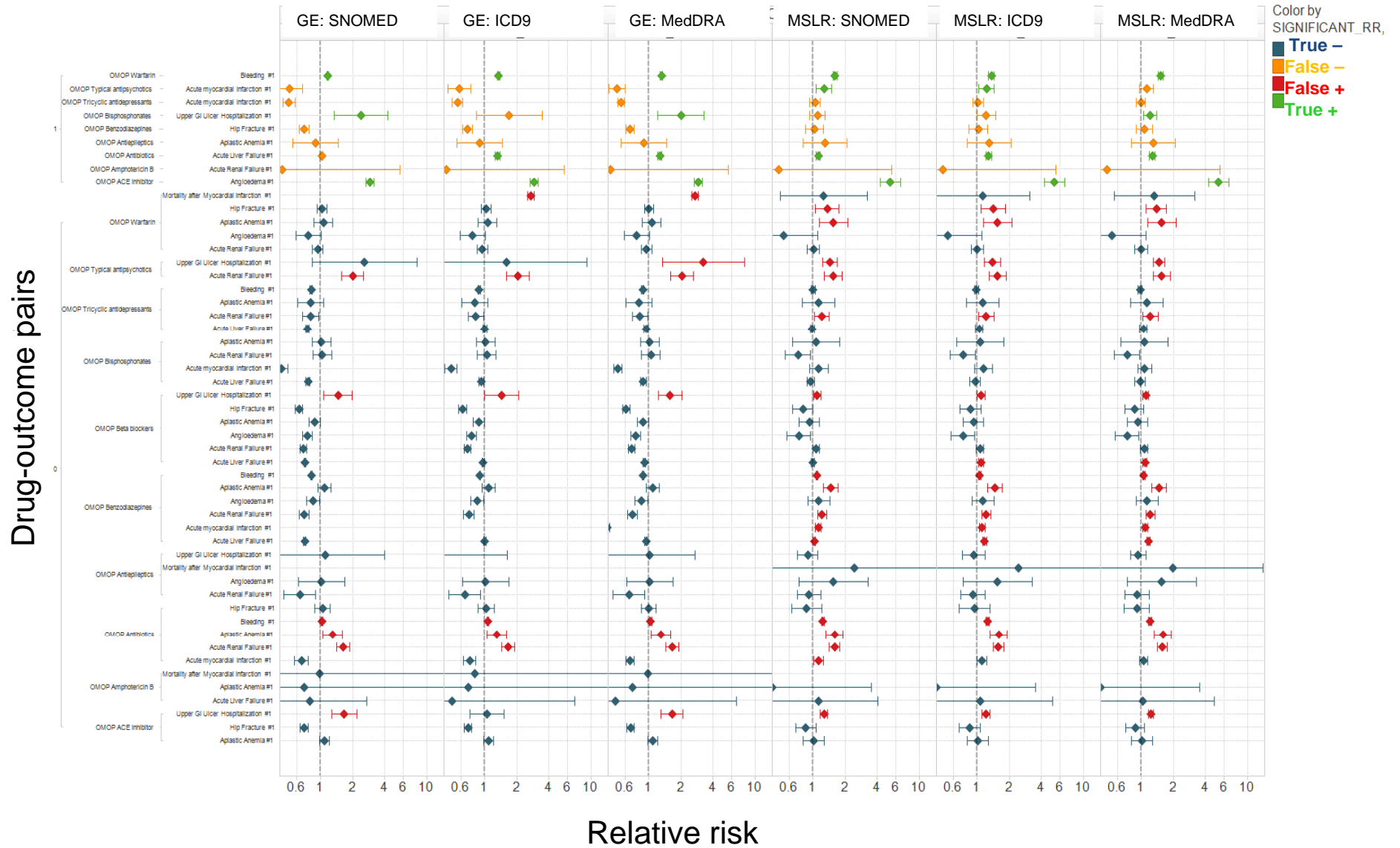
# Sensitivity to Vocabulary: Method DP



# Sensitivity to Vocabulary: Method OS

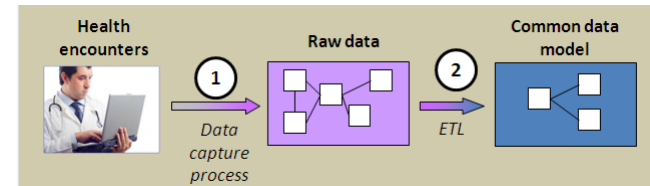


# Sensitivity to Vocabulary: Method USCCS

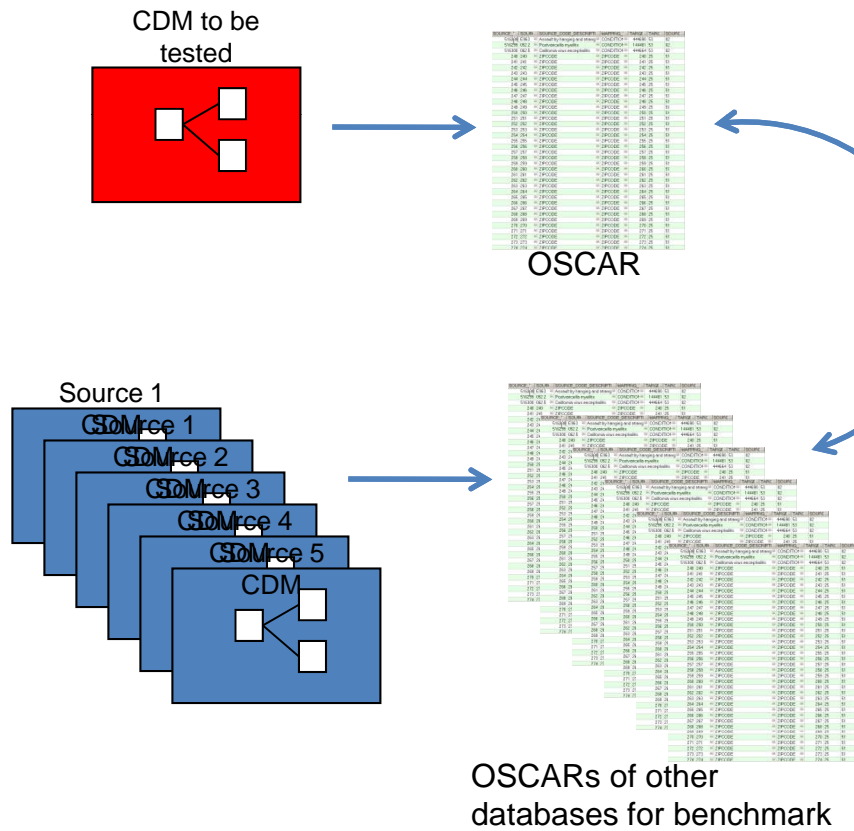


# D: Data Anomaly Review – GROUCH

GROUCH produces a summary report from OSCAR for each concept:



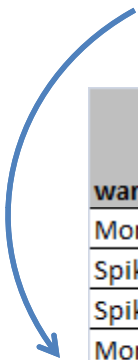
## GROUCH detects data anomalies:



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

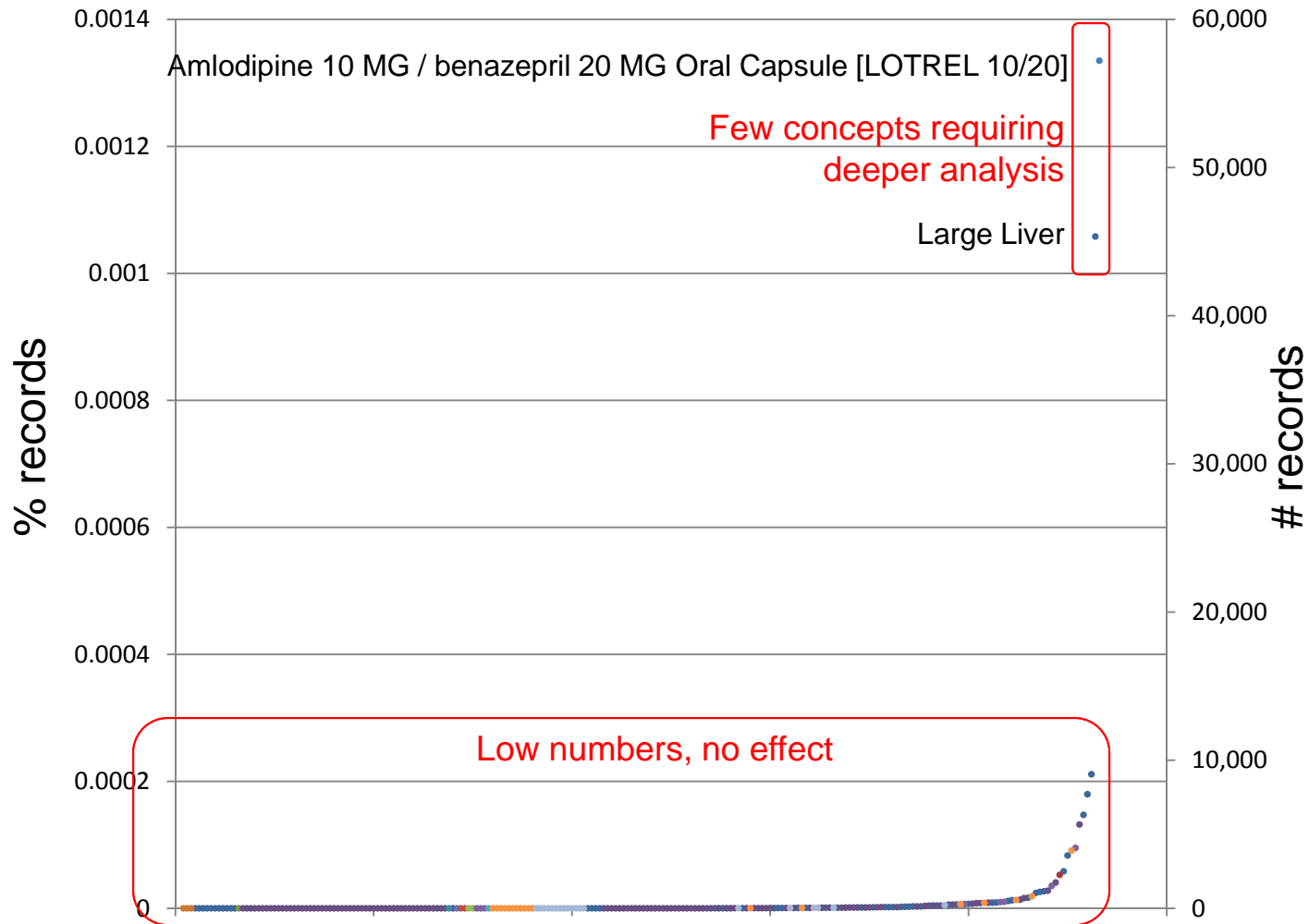
Conclusions: MSLR has large spikes in enrollment at start of each year

# 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 noticeably 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

# Summary

## General

- Data quality assurance procedures are necessary across the data management continuum to ensure a reliable active surveillance system
- OMOP tools and processes promote transparency and facilitate shared understanding across the data network and central coordinating center
- ETL validation through raw-summary comparison enables complete specification and evaluation of decision rules
- OSCAR provides a comprehensive summary of each database to facilitate evaluation and comparison of source population and data characteristics
- GROUCH produces a comprehensive set of data anomaly checks. That allows to identify, diagnose and treat issues of potential concern prior to drug safety analyses. This is a job shared between all stakeholders

# Summary

## Standard vocabularies

- Enables integration of disparate data sources, but requires evaluation to ensure clinical concepts are adequately represented
- Vocabulary mapping had variable effect on the number of observed cases across Health Outcomes of Interest
- Choice of ICD9, SNOMED or MedDRA as standard had negligible impact on the estimates of drug-outcome associations

## Conclusion

OMOP's data quality evaluation provides confidence in the integrity of the OMOP experimental results, but further work is needed to establish robust data quality tools and standards for a national surveillance system.