Characteristics, challenges, and determinants of data quality

J. Marc Overhage, MD, PhD
Chief Medical Informatics Officer
Siemens Health Services
Spontaneous Generation?
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)?
Ambulatory EMR Adoption

The graph shows the adoption rates of different types of EMR systems over the years 2000 to 2012. The categories include Any EMR, Basic EMR, Full EMR, MU, and Ford Model. The x-axis represents the years from 2000 to 2012, and the y-axis represents the percentage adoption rates.

- Any EMR shows a steady increase from 10% in 2000 to around 50% in 2012.
- Basic EMR starts at a lower percentage and also shows a steady increase, reaching around 30% in 2012.
- Full EMR starts at a very low percentage and shows a slight increase over the years.
- MU and Ford Model have very low adoption rates and are not significantly represented in the graph.

The overall trend indicates a growing adoption of EMR systems in the ambulatory sector.
Data capture and sharing

Advanced clinical processes

Improved outcomes

Stage 1 – 2011

Stage 2 – 2013

Stage 3 – 2015

Meaningful Use (original)
Utilization of Available Functionality

67 Practices Representing 189 Clinicians
Challenges in Data Capture

- Images

- Narrative text (labeled)

- Structured data

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.
Cost-Value Tradeoff

- Paper
- Partially structured
- Rigidly structured

Usefulness of Data
Starting Point
Optimum Value
Impact on Usability
Electronic free text
Partially structured
Rigidly structured
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
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

Extraction

Combine Conflicting Local Evidence

Probabilistic Inference Over Time

Extraction

Extraction

Extraction

Extraction

Plug-in Domain Knowledge (e.g., CMS Measures)
REMIND Example

Acute Myocardial Infarction 4.7

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

VENTRICULOGRAPHY: Ventriculography revealed an overall preserved left ventricular ejection fraction, EF 51%. There was some infero basal hypokinesis. The left ventricular end diastolic pressure was normal at 4 mmHg. Central aortic pressure was 125/59 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
Payer Claim for Liver Failure

LEGEND
- fair to good "liver" specificity
- cancer
- etiology unclear
- multiorgan failure / end-stage / sepsis or cardiac arrest
- other

appears that one institution sometimes codes ICD9 570 "acute liver necrosis" for mild SGPT elevation

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

EMR Dx for Liver Failure

52y white male
PPV = 20%

53y white male
PPV = 0%

54y white male
PPV = 0%

55y white male
55y white female
55y white female
PPV = 50%
53y white female
58y African-American male

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

Clinical Care Data: Availability & Access

<table>
<thead>
<tr>
<th>TODAY</th>
<th>eHRs</th>
<th>FUTURE?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Fragmented</td>
<td>• Easily aggregated</td>
<td></td>
</tr>
<tr>
<td>• Limited accessibility</td>
<td>• Broad access</td>
<td></td>
</tr>
<tr>
<td>• Limited populations</td>
<td>• National coverage</td>
<td></td>
</tr>
<tr>
<td>• Narrow uses</td>
<td>• Many applications</td>
<td></td>
</tr>
</tbody>
</table>

Mix of efficacy, safety, and commercial data with multiple uses

Legend:
- = Highly controlled Clinical Trial Data
- = Clinical Care Data from Patients, Payers, & Providers

Graph is For Illustrative Purposes Only

Courtesy of Pfizer Health Informatics
Numerous data sources to support VBHC analyses, but not all data sources are equivalent.

Numerous potential data sources:
- Clinical trials
- Disease registry
- EMR
- Claims
- Modalities + service providers
- Patient sources

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

- **Variables**
  - 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**
  - Size of the dataset, both in time and number of patients
    - 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
Major dimensions composed of numerous factors

### Variables
**Depth**

<table>
<thead>
<tr>
<th>Factors</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcomes measures</td>
<td>Improvement in outcomes (cost and quality) is ultimate goal of VBHC</td>
</tr>
<tr>
<td>Relevant process measures</td>
<td>Understanding drivers of outcomes enables quality improvement</td>
</tr>
<tr>
<td>Financial measures</td>
<td>Understanding cost and utilization</td>
</tr>
<tr>
<td>Patient-centered measures</td>
<td>Patient-generated data, e.g., assessment of health and well-being via satisfaction or survey results; Supplements clinical findings</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of variables</td>
</tr>
<tr>
<td>Granularity of variables</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk-adjustment data</td>
</tr>
<tr>
<td>Patient ID</td>
</tr>
</tbody>
</table>

### Observations
**Breadth**

<table>
<thead>
<tr>
<th>Factors</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>Improves detection of small differences</td>
</tr>
<tr>
<td>Penetration</td>
<td>Decreases need for risk-adjustment</td>
</tr>
<tr>
<td>Number of records</td>
<td>Improves applicability of findings to population</td>
</tr>
<tr>
<td>Skew / Generalizability</td>
<td>Decreases need for risk adjustment</td>
</tr>
<tr>
<td>Internal distribution/Comparability</td>
<td>Enables ID of subsegments of patients / outcomes</td>
</tr>
<tr>
<td></td>
<td>Improves precision, validity of data</td>
</tr>
<tr>
<td></td>
<td>‘Balanced’ population enables analyses which are more generalizable across populations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time and Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longevity / Temporal extent</td>
</tr>
<tr>
<td>Longitudinality / Temporal consistency</td>
</tr>
<tr>
<td>Longitudinality / Across care settings</td>
</tr>
</tbody>
</table>

### Quality control
**Process**

<table>
<thead>
<tr>
<th>Factors</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intent</td>
<td>Data collected for a specific purpose more likely to be relevant to the question and higher quality</td>
</tr>
<tr>
<td>Validation</td>
<td>Increases confidence that findings are accurate (e.g., collected in controlled environment; double entry in clinical trials)</td>
</tr>
<tr>
<td>Fidelity</td>
<td>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’)</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Increases relevance of data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
</tr>
<tr>
<td>Coding</td>
</tr>
<tr>
<td>Linkability</td>
</tr>
</tbody>
</table>

Source: Stakeholder interviews, BCG analysis

BCG VBHC Frameworks for IOM-v2.pptx

THE BOSTON CONSULTING GROUP
Pharmaceutical Questions

Questions

• 10 companies
• 10 questions per company

Answers

Completeness

- Total
- Near
- Gaps
Monitoring Adverse Drug Events

Myocardial Infarctions

Cox-2 Inhibitors

Celecoxib Introduced
Rofecoxib Introduced
Rofecoxib Withdrawn

Prevalence of "rofecoxib" in 7 Databases from 2003 to 2008
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.
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.
CCAE, being a privately insured population, primarily reflects employed and their dependents, so underrepresents those older than 65 years.

In contrast, MDCR represents patients with supplemental Medicare benefits, so primarily reflects persons older than 65 years, but underrepresents those younger than retirement age.

Humana, as a large insurer providing coverage to both privately insured and the Medicare populations, is observed to combine the two age distribution patterns.

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

Data quality (GROUCH) checks:
- Implausible: Year of birth > 2010
- Suspicious: Year of birth < 1900
- Suspicious: change +/- 20% between years.
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.
While Regenstrief reflects a pronounced contraction in young adults but the length of capture is much longer in part due to the EHR contribution.
Each database captures data for a different span of time. The number of persons observed in any given month varies substantially, e.g.,

- Humana maintains a consistent population size
- GE is consistently growing as more practices adopt the Centricity EHR system
- Thomson databases are observed to have annual changes as new data sources are aggregated

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.
Substantial variation across the network in observed prevalence of lisinopril exposure, after standardizing on age, gender, and year.
The prevalence and variability across the network is highly product-specific, underscoring importance of efficient exploration of summary statistics.
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
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
Stratified condition prevalence by year

Temporal trends in conditions reflect changing clinical care, coding practices and population demographics.
Essential Hypertension
Heterogeneity Across Databases
Distributed queries unambiguously define a population from a larger set.

Questions about disease outbreaks, prevention activities, health research, quality measures, etc.
In contrast to clinical trials, not controlled by drug outcome researcher

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

Controlled by outcome researcher
Raw-CDM Summary Comparison

**Tested in GE**
- Person
  - Gender
  - Race
  - Year of Birth
  - Gender by Age
- Drug
  - Counts of codes
  - Refills
  - Quantity
  - Stop Reason

**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

Data holder task: calculate summary statistics from raw data
### Thomson Reuters databases:

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

### GE database:

<table>
<thead>
<tr>
<th>Issue</th>
<th>Impact on HOI or DOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender by age calculated based on 2008, not 2009</td>
<td>No effect on methods</td>
</tr>
<tr>
<td>Drug exposure length incorrectly programmed, resulting in values deviating in 3.72% of cases</td>
<td>Small effect on DOI era length</td>
</tr>
<tr>
<td>Condition length incorrectly programmed, resulting in values deviating in a small number of cases</td>
<td>Possibly small effect on HOI era length</td>
</tr>
</tbody>
</table>
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

ICD-9-CM

284 Aplastic anemia and other bone marrow failure syndromes

284.9 Aplastic anemia, unspecified
284.8 Other unspecified aplastic anemias
284.0 Constitutional aplastic anemia
284.89 Other specified aplastic anemias
284.81 Red cell aplasia (acquired) (with thymoma)
284.01 Constitutional red blood cell aplasia
284.09 Other constitutional aplastic anemia

SNOMED-CT

Aplastic anemia

Pure red cell anemia

Acquired aplastic anemia

Constitutional aplastic anemia

Aplastic anemia due to infection

Marrow depression and hypoplastic anaemias (HLT)

MedDRA

Aplasia pure red cell (PT)

Aplastic anaemia (PT)

Aregenerative aplastic anaemia (LLT)

Constitutional aplastic anaemia (LLT)

Mapped (identical color)

Missing mapping
## Summary of Terminology Mapping Artifacts

<table>
<thead>
<tr>
<th>Artifact</th>
<th>Resulting in</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Codes are wrongly mapped</td>
<td>Wrong data</td>
</tr>
<tr>
<td>2. Codes are not mapped</td>
<td>Missing data</td>
</tr>
<tr>
<td>3. Many to one mapping</td>
<td>Recruiting data for related codes</td>
</tr>
<tr>
<td>4. Child concepts of mapped codes</td>
<td>Recruiting data for related codes</td>
</tr>
</tbody>
</table>

What are the effects of these artifacts on a method’s ability to detect drug-outcome relationships?
Sensitivity to Vocabulary: Method HDPS

Drug-outcome pairs

Relative risk
Sensitivity to Vocabulary: Method DP

Drug-outcome pairs

Relative risk
Sensitivity to Vocabulary: Method OS

Drug-outcome pairs

Relative risk
Sensitivity to Vocabulary: Method USCCS

Drug-outcome pairs

Relative risk
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

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

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

**Conclusions:** MSLR has large spikes in enrollment at start of each year
## Summary MSLR GROUCH – Concept Checks

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

126 concepts are observed at a notably 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

Amlodipine 10 MG / benazepril 20 MG Oral Capsule [LOTREL 10/20]

Few concepts requiring deeper analysis

Large Liver

Low numbers, no effect
## Summary MSLR GROUCH – Boundary Checks

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

**Conclusion:** Small numbers, many of the warning legitimate healthcare situations
Current Level of Interoperability

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