Cohort Diagnostics

Gowtham Rao
06/16/2020
Overview

• Motivation

• Uses cases → software functionality
Motivation: Natural history of an observation study.

- Research question
- **Protocol**: Full (human-readable)
- **Study package**: Full (machine-readable)
- Register protocol & study package
- Execute study
- Results

**Wrong**: Changing after results
Motivation: Reduce bias/Improve confidence

Effect size estimates in published observational studies

To make sure our study will provide a reliable answer

(before looking at the answer)

Changing the question/study design/dissemination based on the results leads to unreliable evidence
Diagnostics as a best practice

Research question

Write protocol
- **Protocol**: Full (human-readable) description
  - **Full diagnostics**

Implement study
- **Study package**: Full (machine-readable) implementation
  - **Execute diagnostics**

Without looking at the outcome of interest!

Register protocol & study package

Execute study

Results

CORRECT
Diagnostics in OHDSI

Phenotype Diagnostics
• Clinical practice experience
• Data domain expertise
• Coding experience

PheValuator

Something was missing

- Earlier in the process (phenotyping)
  - Before phevaluator, negative control, PS
- Focussed specifically on building cohorts
  - Feedback on impact of specification choices
  - Clinical description → literature review → building cohorts
- Enable conversation between investigators
  - Compare the impact of specification choices across data sources.
Overview

• Motivation

• Uses cases → software functionality
New R-package - Overview
Functionality – Cohort Counts

Cohort Counts

Description
A table showing the number of cohort entries and unique persons per cohort per database. Because one person can have more than one cohort entry, the number of entries can be higher than the number of persons.

Options
You can select multiple databases in the side bar to see counts from different databases side-by-side.

What to look for
- Are there cohorts that are empty in some databases?
- Are the relative counts (relative to the other cohorts in the same database) comparable across databases? Note that the color bars show the relative counts.
- Are the cohorts of expected and sufficient size? For example, if we want to study the effect of an exposure, a rule-of-thumb is that we require at least 2,000 in the exposure cohort.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>CCAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>New users of ACE inhibitors as first-line monotherapy for hypertension (2)</td>
<td>846,195</td>
</tr>
<tr>
<td>New users of Thiazide-like diuretics as first-line monotherapy for hypertension (2)</td>
<td>327,805</td>
</tr>
<tr>
<td>Angioedema outcome</td>
<td>136,844</td>
</tr>
<tr>
<td>Acute myocardial infarction outcome</td>
<td>516,340</td>
</tr>
<tr>
<td>ACES with bad inclusion criterion</td>
<td>571,991</td>
</tr>
</tbody>
</table>

Showing 1 to 6 of 6 entries
## Functionality – Cohort Counts

<table>
<thead>
<tr>
<th>Cohort Description</th>
<th>CPRD Entries</th>
<th>CPRD Subjects</th>
<th>IBM_CCAE Entries</th>
<th>IBM_CCAE Subjects</th>
<th>IBM_MDCD Entries</th>
<th>IBM_MDCD Subjects</th>
<th>IBM_MDCR Entries</th>
<th>IBM_MDCR Subjects</th>
<th>JMDC Entries</th>
<th>JMDC Subjects</th>
<th>OPTUM_EXTENDED_DOD Entries</th>
<th>OPTUM_EXTENDED_DOD Subjects</th>
<th>OPTUM_PANTHER Entries</th>
<th>OPTUM_PANTHER Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Cohort Diagnostics] Bacterial Pneumonia</td>
<td>26379</td>
<td>23,366</td>
<td>1,311,797</td>
<td>854,006</td>
<td>749,001</td>
<td>435,821</td>
<td>660,162</td>
<td>392,526</td>
<td>216,228</td>
<td>89,760</td>
<td>1,093,643</td>
<td>910,276</td>
<td>1,487,700</td>
<td>668511</td>
</tr>
<tr>
<td>[Cohort Diagnostics] Hospitalizations with Pneumonia</td>
<td>746,591</td>
<td>625,020</td>
<td>747,009</td>
<td>514,456</td>
<td>506,500</td>
<td>706,669</td>
<td>54,465</td>
<td>48,550</td>
<td>1,832,905</td>
<td>910,717</td>
<td>730,028</td>
<td>319412</td>
<td>1,211,964</td>
<td>910,717</td>
</tr>
<tr>
<td>[Cohort Diagnostics] Pneumonia</td>
<td>220,447</td>
<td>183,475</td>
<td>4,126,188</td>
<td>3,278,670</td>
<td>2,157,561</td>
<td>1,285,845</td>
<td>2,069,702</td>
<td>1,206,277</td>
<td>441,396</td>
<td>217,765</td>
<td>3,003,301</td>
<td>4,654,768</td>
<td>319412</td>
<td>319412</td>
</tr>
<tr>
<td>[Cohort Diagnostics] Pneumonia with viral code events</td>
<td>304,682</td>
<td>29,689</td>
<td>157,532</td>
<td>70,760</td>
<td>68,501</td>
<td>42,206</td>
<td>40,470</td>
<td>6,156</td>
<td>140,995</td>
<td>136,030</td>
<td>96,890</td>
<td>319412</td>
<td>94089</td>
<td>94089</td>
</tr>
<tr>
<td>[Cohort Diagnostics] Pneumonia without non-viral source events</td>
<td>202,079</td>
<td>186,365</td>
<td>3,559,238</td>
<td>3,144,556</td>
<td>1,710,475</td>
<td>1,326,548</td>
<td>1,570,292</td>
<td>1,141,432</td>
<td>230,695</td>
<td>277,160</td>
<td>3,490,792</td>
<td>2,871,082</td>
<td>3,789,057</td>
<td>307,833</td>
</tr>
</tbody>
</table>

### Concept Name

- Pneumonia
- Ornithosis with pneumonia
- Idiopathic pneumonia
- Fungal pneumonia
- Atypical pneumonia
- Acute ulcerative gastroenteritis complicating pneumonia
- Acute pneumonia due to coccidioidomycosis
Functionality – Incidence Rate

Incidence Rate

Description
A graph showing the incidence rate, optionally stratified by age (in 10-year bins), gender, and calendar year.

The incidence rate is computed as $\frac{\text{number of people first entering the cohort}}{\text{number of years people were eligible to enter the cohort}}$ for the first time. The eligible person time is defined as the time when

- A person was observed in the database (based on the observation_period table).
- Had the required amount of prior observation time as specified in the cohort entry event criteria. For example, if the cohort definition requires 365 days of observation prior to cohort entry, patients are not eligible to enter the cohort in the first 365 days of their observation period, and this time is not counted in the eligible time.
- If the person enters the cohort, then only the time up to cohort entry. Because we only consider the first cohort entry, persons are no longer eligible to enter to cohort after their first entry.

Options
You can select multiple databases in the side bar to see graphs from different databases in the same plot.

Select the cohort to explore in the side bar.
At the top of the plot, you can choose whether to stratify the data by age, gender, or calendar year.
If you move the mouse over the plot, you can see the precise value.

What to look for
- Are the observed incidence rates in line with expectations? For example, if we have an estimate of the population incidence based on an external source, is the incidence rate comparable to that estimate?
- Are the age and gender distributions in line with expectations? For example, are contraceptives only prescribed in women?
- Is the incidence rate stable over time? If there are sudden peaks or drops, this may indicate coding issues.
Functionality – Incidence Rate
Dramatic drop around 2008
Functionality – Incidence Rate Increase around 2015
Functionality – Incidence Rate
Progressive increase over time
Functionality – Incidence Rate
Progressive decrease over time
Functionality – Incidence Rate
Abortion among Male
Functionality – Time distributions

Time Distributions

Description
Boxplot and a table showing the distribution of time (in days) before and after the cohort index date (cohort start date), and the time between cohort start and end date. This information is shown for all cohort entries, so not limited to the first per person.

The boxplot shows:
- Whiskers: The minimum and maximum observed number of days.
- Box: The 25th to 75th percentile.
- Line: The median.

The table shows the same information and more:
- Average: The mean of the distribution
- SD: Standard Deviation
- Min: The minimum
- P10: The 10th percentile
- P25: The 25th percentile
- Median: The median (50th percentile)
- P75: The 75th percentile
- P90: The 90th percentile
- Max: The maximum

Options
You can select multiple databases in the side bar to see time distributions from different databases in the same plot and table.

Select the cohort to explore in the side bar.

What to look for
- For exposure cohorts: is there sufficient time after index (either within the cohort for on-treatment analyses, or until the end of observation for intent-to-treat type analyses) to observe the outcome of interest?
- Are there many cohorts with length = 0 when this is not expected?
- Are the distributions comparable across databases?
Functionality – Time distributions

Time Distributions

observation time (days) after index

observation time (days) prior to index

time (days) between cohort start and end
### Functionality – Included concepts

A table showing the (source) concepts observed in the database that are included in a concept set of a cohort. The Subjects column contains the number of subjects in the entire database that have the specific concept. This count is not restricted to only those people in the cohort. Source concepts are identified in the `source_concept_id` fields of the Common Data Model, e.g., `drug_source_concept_id` and are used to identify the specific source codes used in a database. Standard concepts are found using the `concept_id` fields, e.g., `drug_concept_id`, and use the same coding system across all databases.

#### Options
You can select a database in the side bar to see the concepts and counts observed in that database. Select the cohort and the specific concept set within that cohort to explore in the side bar. You can switch between Source Concepts and Standard Concepts at the top of the table.

#### What to look for
- Are there source codes included that shouldn't be? For example, in a concept set for hypertensive disorder, are hypotension codes included by accident?
- Are all expected codes present? For example, if we have a list of ICD-10 codes that have been used in literature to identify a cohort, are all those codes present?
Functionality – Orphan concepts

Orphan (Source) Concepts

Description
A table showing the (source) concepts observed in the database that are not included in a concept set of a cohort, but maybe should be. The following logic is used to identify concepts that might be relevant:

1. Given a concept set expression, find all included concepts.
2. Find all names of those concepts, including synonyms, and the names of source concepts that map to them.
3. Search for concepts (standard and source) that contain any of those names as substring.
4. Filter those concepts to those that are not in the original set of concepts (i.e. orphans).
5. Restrict the set of orphan concepts to those that appear in the CDM database as either source concept or standard concept.

The Subjects column contains the number of subjects in the entire database that have the specific concept. This count is not restricted to only those people in the cohort. Source concepts are identified in the _source_concept_id fields of the Common Data Model, (e.g. drug_source_concept_id) and are used to identify the specific source codes used in a database. Standard concepts are found using the _concept_id fields (e.g. drug_concept_id), and use the same coding system across all databases.

Options
You can select a database in the side bar to see the concepts and counts observed in that database.

Select the cohort and the specific concept set within that cohort to explore in the side bar.

What to look for
- Are there concepts that are not included in the concept but should be? Note that the provided list likely contains many false positives.
## Functionality – Orphan concepts

<table>
<thead>
<tr>
<th>Count</th>
<th>Concept ID</th>
<th>Standard</th>
<th>Vocabulary</th>
<th>Code</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>9,882,695</td>
<td>80180 S</td>
<td>SNOMED</td>
<td>SNOMED</td>
<td>398273005</td>
<td>Osteoarthritis</td>
</tr>
<tr>
<td>5,317,647</td>
<td>44631489</td>
<td>ICD9CM</td>
<td>ICD9CM</td>
<td>715.90</td>
<td>Osteoarthritis, unspecified whether generalized or localized, site unspecified</td>
</tr>
<tr>
<td>3,707,579</td>
<td>40797570 S</td>
<td>SNOMED</td>
<td>SNOMED</td>
<td>213873007</td>
<td>Osteoarthritis of knee</td>
</tr>
<tr>
<td>3,354,361</td>
<td>75617 S</td>
<td>SNOMED</td>
<td>SNOMED</td>
<td>213873000</td>
<td>Degenerative joint disease involving multiple joints</td>
</tr>
<tr>
<td>3,295,489</td>
<td>45586894</td>
<td>ICD10CM</td>
<td>ICD10CM</td>
<td>M19.90</td>
<td>Unspecified osteoarthritis, unspecified site</td>
</tr>
<tr>
<td>2,670,722</td>
<td>4035439 S</td>
<td>SNOMED</td>
<td>SNOMED</td>
<td>213962000</td>
<td>Idiopathic osteoarthritis</td>
</tr>
<tr>
<td>2,317,279</td>
<td>44636171</td>
<td>ICD9CM</td>
<td>ICD9CM</td>
<td>715.00</td>
<td>Osteoarthritis, generalized, site unspecified</td>
</tr>
<tr>
<td>1,837,048</td>
<td>44630325</td>
<td>ICD9CM</td>
<td>ICD9CM</td>
<td>715.10</td>
<td>Osteoarthritis, localized, primary, lower leg</td>
</tr>
<tr>
<td>1,287,693</td>
<td>44636173</td>
<td>ICD9CM</td>
<td>ICD9CM</td>
<td>715.96</td>
<td>Osteoarthritis, unspecified whether generalized or localized, lower leg</td>
</tr>
<tr>
<td>1,104,623</td>
<td>35206765</td>
<td>ICD10CM</td>
<td>ICD10CM</td>
<td>M15.0</td>
<td>Polyosteoarthritis, unspecified</td>
</tr>
<tr>
<td>1,037,592</td>
<td>44625195</td>
<td>ICD9CM</td>
<td>ICD9CM</td>
<td>715.09</td>
<td>Osteoarthritis, generalized, multiple sites</td>
</tr>
<tr>
<td>954,396</td>
<td>72993 S</td>
<td>SNOMED</td>
<td>SNOMED</td>
<td>213829007</td>
<td>Localized, primary osteoarthritis</td>
</tr>
<tr>
<td>810,869</td>
<td>35206772</td>
<td>ICD10CM</td>
<td>ICD10CM</td>
<td>M17.0</td>
<td>Bilateral primary osteoarthritides of knee</td>
</tr>
<tr>
<td>703,515</td>
<td>4079749 S</td>
<td>SNOMED</td>
<td>SNOMED</td>
<td>2139872002</td>
<td>Osteoarthritis of hip</td>
</tr>
<tr>
<td>696,168</td>
<td>44636174</td>
<td>ICD9CM</td>
<td>ICD9CM</td>
<td>715.35</td>
<td>Osteoarthritis, localized, not specified whether primary or secondary, lower leg</td>
</tr>
<tr>
<td>612,020</td>
<td>72990 S</td>
<td>SNOMED</td>
<td>SNOMED</td>
<td>98600001</td>
<td>Localized osteoarthritis uncertain if primary or secondary</td>
</tr>
<tr>
<td>610,355</td>
<td>45577163</td>
<td>ICD9CM</td>
<td>ICD9CM</td>
<td>M17.11</td>
<td>Unilateral primary osteoarthritis, right knee</td>
</tr>
<tr>
<td>574,398</td>
<td>44536881</td>
<td>ICD9CM</td>
<td>ICD9CM</td>
<td>326202002</td>
<td>Osteoarthritis involving multiple sites but not designated as generalized</td>
</tr>
<tr>
<td>573,064</td>
<td>73840 S</td>
<td>SNOMED</td>
<td>SNOMED</td>
<td>213813003</td>
<td>Localized, primary osteoarthritides of the shoulder region</td>
</tr>
<tr>
<td>546,943</td>
<td>35206767</td>
<td>ICD10CM</td>
<td>ICD10CM</td>
<td>M17.0</td>
<td>Osteoarthritis of knee, unspecified</td>
</tr>
<tr>
<td>520,222</td>
<td>44625197</td>
<td>ICD9CM</td>
<td>ICD9CM</td>
<td>715.89</td>
<td>Seropositive rheumatoid arthritis</td>
</tr>
<tr>
<td>384,095</td>
<td>4035611 S</td>
<td>SNOMED</td>
<td>SNOMED</td>
<td>213979005</td>
<td>Osteoarthritis, unspecified whether generalized or localized, pelvic region and thigh</td>
</tr>
<tr>
<td>378,690</td>
<td>4462480 S</td>
<td>ICD9CM</td>
<td>ICD9CM</td>
<td>715.95</td>
<td>Osteoarthritis involving multiple sites but not designated as generalized, pelvic region and thigh</td>
</tr>
<tr>
<td>378,690</td>
<td>40483794 S</td>
<td>SNOMED</td>
<td>SNOMED</td>
<td>445478004</td>
<td>Degenerative joint disease of pelvis</td>
</tr>
</tbody>
</table>
Orphan codes

<table>
<thead>
<tr>
<th>Concept Name</th>
<th>Descendants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide attempt</td>
<td>Y</td>
</tr>
<tr>
<td>Intentional harming self</td>
<td>Y</td>
</tr>
</tbody>
</table>

Orphans: matches not included in concept set
- Electrocution, intent to self-farm

Data in Common Data Model

- Descendants
- Synonyms of descendants
- Source concepts mapping to included concepts

Search matches
- Suicide attempt
- Electrocution, intent to self-farm

- All standard concepts
- All source concepts

Search terms
- Suicide attempt
- Suicidal poisoning
- Self-harm
- Self-inflicted injury

Substring search

Concepts to search
- Cough
- Suicide attempt
- Acetaminophen
- Electrocution, intent to self-farm

Orphan codes

<table>
<thead>
<tr>
<th>Concept Name</th>
<th>Descendants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide attempt</td>
<td>Y</td>
</tr>
<tr>
<td>Intentional harming self</td>
<td>Y</td>
</tr>
</tbody>
</table>

Orphans: matches not included in concept set
- Electrocution, intent to self-farm
Functionality – Inclusion Rule Statistics

Inclusion Rule Statistics

Description
A table showing the number of subject that match specific inclusion rules in the cohort definition. Note that this table will be empty if no inclusion rules have been specified.

The table contains the following columns:
- Sequence: The order in which the inclusion rules are applied to the cohort.
- Name: The name of the inclusion rule.
- Meet: The number of cohort entries that meet the entry event definition and the specific inclusion rule indicated in the row.
- Gain: The number of cohort entries that would be gained if this inclusion rule was dropped.
- Total: The number of cohort entries meeting the entry event definition. In other words, the number of cohort entries before applying any of the inclusion rules.
- Remain: The number of cohort entries remaining after applying the specific inclusion rule, and all preceding rules.

Options
You can select a database in the side bar to see the inclusion rules statistics observed in that database.
Select the cohort to explore in the side bar.

What to look for
- Are there inclusion rules that nobody meets in a database? For example, requiring a specialist visit that isn’t recorded in a specific database.
- Are there inclusion rules that have no effect in a database? For example, requiring no occurrence of a prior disease code that is not recorded in a database.
- Are there inclusion rules that drastically reduce the population? In this case we might worry about generalizability. For example, if we require a diagnostic procedure, and only a small fraction meets this criteria, we may wonder if this identifies a special population that differs from the overall population in significant ways.
Functionality – Index Event Breakdown

Description
A table showing the concepts belonging to the concept sets in the entry event definition that are observed on the index date. In other words, the table lists the concepts that likely triggered the cohort entry. The counts indicate number of cohort entries where the concepts was observed on the index date. Note that multiple concepts can be present on the index date, so the sum of counts might be greater than the cohort entry count.

Options
You can select multiple databases in the side bar to see counts from different databases side-by-side.
Select the cohort to explore in the side bar.

What to look for
- Is one concept unexpectedly dominating? For example, if our cohort identifies exposure to drugs in a class, but we notice almost everyone enters the cohort based on a single drug, we may wonder whether our results will generalize to the class.
- Are the highest ranking concepts different across databases? For example, is everyone in one database initiating high-dose prescriptions, and everyone in another database low-dose prescriptions?
<table>
<thead>
<tr>
<th>Concept ID</th>
<th>Name</th>
<th>CPRD Count</th>
<th>IBM_CCAE Count</th>
<th>IBM_MDCD Count</th>
<th>IBM_MDCR Count</th>
<th>JMDC Count</th>
<th>OPTUM_EXTENDED_DOD Count</th>
<th>OPTUM_PANTHER Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>80009</td>
<td>Rheumatoid arthritis</td>
<td>76,744</td>
<td>446,615</td>
<td>102,530</td>
<td>125,411</td>
<td>116,722</td>
<td>650,465</td>
<td>997,312</td>
</tr>
<tr>
<td>80009</td>
<td>Rheumatoid arthritis</td>
<td>76,744</td>
<td>446,615</td>
<td>108,530</td>
<td>135,411</td>
<td>116,722</td>
<td>650,465</td>
<td>997,312</td>
</tr>
<tr>
<td>4083556</td>
<td>Seronegative rheumatoid arthritis</td>
<td>6,866</td>
<td>18,587</td>
<td>3,700</td>
<td>3,011</td>
<td>374</td>
<td>28,775</td>
<td>72,288</td>
</tr>
<tr>
<td>4083556</td>
<td>Seronegative rheumatoid arthritis</td>
<td>6,866</td>
<td>18,587</td>
<td>3,700</td>
<td>3,011</td>
<td>374</td>
<td>28,775</td>
<td>72,288</td>
</tr>
<tr>
<td>4035611</td>
<td>Seropositive rheumatoid arthritis</td>
<td>2,280</td>
<td>26,096</td>
<td>6,444</td>
<td>3,736</td>
<td>108</td>
<td>45,150</td>
<td>127,277</td>
</tr>
<tr>
<td>4035611</td>
<td>Seropositive rheumatoid arthritis</td>
<td>2,280</td>
<td>26,096</td>
<td>6,444</td>
<td>3,736</td>
<td>108</td>
<td>45,150</td>
<td>127,277</td>
</tr>
<tr>
<td>41144444</td>
<td>flare of rheumatoid arthritis</td>
<td>710</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Functionality – Cohort Characterization

Cohort Characterization

Description

A table showing cohort characteristics (covariates). These characteristics are captured on or before the cohort start date. There is a Pretty and a Raw version of this table.

The Pretty table shows a few selected covariates. These are all binary covariates, and the table shows the proportion (%) of the cohort entries having the covariate.

The Raw table shows all captured covariates. These include binary and continuous covariates (e.g. the Charlson comorbidity index). For each covariate the table lists the mean, which for binary covariates is equal to the proportion, and the standard deviation (SD).

Options

- You can select multiple databases in the sidebar to see cohort characteristics from different databases side-by-side in the same table.
- Select the cohort to explore in the sidebar.
- Select either the Pretty or the Raw table at the top of the table.

What to look for

- Are the characteristics of the cohort as expected? For example, do people have the expected comorbidities?
- Do the characteristics of the cohort differ much per database?
Functionality - Cohort Overlap (Subjects)

A Venn diagram showing the overlap between two cohorts, and a table listing several overlap statistics. The diagram shows the overlap in terms of subjects. It shows the number of subjects that belong to each cohort and to both. The diagram does not consider whether the subjects were in the different cohorts at the same time.

The table shows the same information and more:

- Subject in either cohort: The number of subjects that enter one or both cohorts. (The union)
- Subject in both cohorts: The number of subjects that enter both cohorts, although not necessarily at the same time. (The intersection)
- Subject in target not in comparator: The number of subjects that enter the target cohort, but not the comparator cohort. (Subtracting the comparator from the target)
- Subject in comparator not in target: The number of subjects that enter the comparator cohort, but not the target cohort. (Subtracting the target from the comparator)
- Subject in target before comparator: The number of subjects that enter both cohorts, but enter the target cohort before entering the comparator cohort. This number considers only the first entry per cohort per person.
- Subject in comparator before target: The number of subjects that enter both cohorts, but enter the comparator cohort before entering the target cohort. This number considers only the first entry per cohort per person.
- Subject in target and comparator on same day: The number of subjects that enter both cohorts on the same date. This number considers only the first entry per cohort per person.
- Subject having target start during comparator: The number of subjects that enter the target cohort during the comparator cohort, meaning comparator cohort start date -> target cohort start date <= comparator cohort end date. This number considers only the first entry per cohort per person.
- Subject having comparator start during target: The number of subjects that enter the comparator cohort during the target cohort, meaning target cohort start date <= comparator cohort start date = comparator cohort end date. This number considers only the first entry per cohort per person.
Functionality - Cohort Overlap (Subjects)

Depression

3,432,163

Fluoxetine

369,251

Cohort Overlap Statistics

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject in either cohort</td>
<td>3,933,103</td>
</tr>
<tr>
<td>Subject in both cohort</td>
<td>369,251</td>
</tr>
<tr>
<td>Subject in target not in comparator</td>
<td>3,432,163</td>
</tr>
<tr>
<td>Subject in comparator not in target</td>
<td>151,689</td>
</tr>
<tr>
<td>Subject in target before comparator</td>
<td>281,314</td>
</tr>
<tr>
<td>Subject in comparator before target</td>
<td>59,572</td>
</tr>
<tr>
<td>Subject in target and comparator on same day</td>
<td>28,365</td>
</tr>
<tr>
<td>Subject having target start during comparator</td>
<td>50,555</td>
</tr>
<tr>
<td>Subject having comparator start during target</td>
<td>280,654</td>
</tr>
</tbody>
</table>
Functionality – Compare Cohort Characteristics

Compare Cohort Characteristics

Description
A table or plot showing cohort characteristics (covariates) for two cohorts side-by-side. These characteristics are captured on or before the cohort start date. There is a Pretty and a Raw version of the table.

The Pretty table shows a few selected covariates. These are all binary covariates, and the table shows the proportion (% of the cohort entries having the covariate, as well as the standardized difference of the mean (zStdDiff). The Raw table shows all captured covariates. These include binary and continuous covariates (e.g., the Charlson comorbidity index). For each covariate the table lists the mean, which for binary covariates is equal to the proportion, the standard deviation (SD), and the standardized difference of the mean (zStdDiff).

The plot shows all covariates, include binary and continuous covariates. The x-axis represents the mean value in the target cohort, the y-axis the mean value in the comparator cohort. Each dot represents a covariate, and the color indicates the absolute value of the standardized difference of the mean.

Options
You can select a database in the side bar.
Select the cohort to explore in the side bar.
Select either the Pretty, the Raw table, or the plot at the top of the screen.
In the plot, you can move the mouse pointer over a dot to see information on that covariate.

What to look for
- Are there major differences between the two cohorts? For example, if we wish to compute a propensity score between two cohorts, concepts that have very high proportion in one cohort and a very low proportion in the other may lead to a perfectly predictive model.
- In general, how comparable are the cohorts? If we wish to compare two exposures, but the cohorts differ over many characteristics, we may be able to fit a propensity model and compute an estimate, but we may have concerns over the generalizability of the results.
Functionality – Compare Cohort Characteristics
Functionality – Compare Cohort Characteristics

Crohns

Ulcerative colitis
Benefits of Cohort Diagnostics

• CohortDiagnostics has already proven instrumental in improving cohort definitions (how did we do without?)
• Provides insight into behavior of cohort definitions across research networks
• How to record and communicate the higher standard we’re now using?
CohortDiagnostics in a distributed setting

- Zipped set of CSV files
- Enforcing a minimum cell count
  (Fully human-reviewable, privacy preserving)
Overview

• Motivation

• Uses cases ➔ software functionality
Walk thru Use case

Backup
### Clinical review

**Complex syndrome – mostly clinical diagnosis (no testing)**

- Skin manifestations
- Joint manifestations
- Fever

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Mechanism</th>
<th>Clinical features</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Immediate reaction (within one hour)</td>
<td>IgE-mediated, immediate-type hypersensitivity</td>
<td>Antigen exposure causes IgE-mediated activation of mast cells and basophils, with release of vasoactive substances, such as histamine, prostaglandins, and leukotrienes.</td>
</tr>
<tr>
<td>II</td>
<td>Cell-mediated or delayed hypersensitivity</td>
<td>Antigen exposure activates T cells, which then mediate tissue injury. Depending upon the type of T cell activation and the other effector cells recruited, different subtypes can be differentiated (ie, types IVA to IVD).</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IgE: immunoglobulin E; Fc: IgG; Fc portion of Immunoglobulin G; SJS/TEN: Stevens-Johnson syndrome/toxic epidermal necrolysis; AGEP: acute generalized exanthematous pustulosis; DRESS/DHS: drug rash with eosinophilia and systemic symptoms/drug-induced hypersensitivity syndrome.

Data review

Hierarchy

Not coded in data!!
Literature review

In 1990!

 Estimate from spontaneous report: 0.0018% (100x lower) (Platt et al.)

Very rare phenotype!

reported more frequently in children than in adults with an overall occurrence ranging from 0.5% (1 in 200) in one trial, to 0.024% (2 in 8,346) in overall clinical trials (with an incidence in children in clinical trials of 0.055%). The worldwide reporting rate for serum-sickness-like reactions in adults is very rare (<0.01%).
Lets start building phenotypes

• Broad definition vs Narrow definition
  – What is broad? What is narrow?
  – How about ‘Very broad’? Or even ‘Very very broad’
  – How about ‘very narrow’? Or ‘very very narrow’!

Since this is a syndrome – lets break it down into ‘Components’

Syndrome – ‘combination of components temporally related’

<table>
<thead>
<tr>
<th>Symptoms</th>
<th>Signs</th>
<th>Clinical diagnosis</th>
<th>Common Treatment/Investigations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itch/Pruritus</td>
<td>Rash (2)</td>
<td>Type 1 HS reaction</td>
<td>Anti-histamine (diphenhydramine, cetirizine)</td>
</tr>
<tr>
<td>Polyarthralgia (1)</td>
<td>Fever (3) +/-</td>
<td>Type 2 HS reaction</td>
<td>Steroids (Prednisolone, methylprednisolone)</td>
</tr>
<tr>
<td>Joint pain</td>
<td>Hypertension</td>
<td>Type 3 HS reaction</td>
<td>Skin biopsy</td>
</tr>
<tr>
<td>Joint Swelling</td>
<td>Tachycardia</td>
<td>Type 4 HS reaction</td>
<td>Complement testing</td>
</tr>
<tr>
<td>Fever</td>
<td>Anosmia</td>
<td>Polyarthritis (1)</td>
<td></td>
</tr>
<tr>
<td>Wheezing</td>
<td>Exanthema</td>
<td>Small joint arthritis</td>
<td></td>
</tr>
<tr>
<td>Arthus reaction</td>
<td>Purpura</td>
<td>Large joint arthritis</td>
<td></td>
</tr>
<tr>
<td>Vasculitis</td>
<td>Necrotizing capillaritis/ neutrophilic vasculitis</td>
<td>Stevens-Johnson Syndrome</td>
<td></td>
</tr>
<tr>
<td>Ulcerative (2)</td>
<td>Tumor epithelial NaCl crystals</td>
<td>Vasculitis</td>
<td>Phototherapy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sweet syndrome</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Contact dermatitis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Eczema</td>
</tr>
</tbody>
</table>
# Component - Itch

<table>
<thead>
<tr>
<th>Subject</th>
<th>Concept ID</th>
<th>Vocabulary</th>
<th>Code</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,357,083</td>
<td>44826848</td>
<td>ICD9CM</td>
<td>692.6</td>
<td>Contact dermatitis and other eczema due to plants [except food]</td>
</tr>
<tr>
<td>990,335</td>
<td>44833794</td>
<td>ICD9CM</td>
<td>698.9</td>
<td>Unspecified pruritic disorder</td>
</tr>
<tr>
<td>562,985</td>
<td>44822164</td>
<td>ICD9CM</td>
<td>691.0</td>
<td>Diaper or napkin rash</td>
</tr>
<tr>
<td>475,888</td>
<td>3520881</td>
<td>ICD10CM</td>
<td>L25.9</td>
<td>Unspecified contact dermatitis, unspecified cause</td>
</tr>
<tr>
<td>412,390</td>
<td>44825651</td>
<td>ICD9CM</td>
<td>692.89</td>
<td>Contact dermatitis and other eczema due to other specified agents</td>
</tr>
<tr>
<td>405,289</td>
<td>35208496</td>
<td>ICD10CM</td>
<td>L29.9</td>
<td>Pruritus, unspecified</td>
</tr>
<tr>
<td>365,481</td>
<td>44826852</td>
<td>ICD9CM</td>
<td>698.3</td>
<td>Lichenification and lichen simplex chronicus</td>
</tr>
</tbody>
</table>
Itch or pruritis and-or urticaria (exclude diaper rash, contact dermatitis)

Flat – uniform
Passes my qualitative review!!
Rinse and repeat for each component

<table>
<thead>
<tr>
<th>Symptoms</th>
<th>Signs</th>
<th>Clinical diagnosis</th>
<th>Common treatment/investigations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itch/Pruritis</td>
<td>Rash (2)</td>
<td>Type 1 HS reaction</td>
<td>Anti-histamines (diphenhydramine, cetirizine)</td>
</tr>
<tr>
<td>Polyarthritis (1)</td>
<td>Fever (3) +/-</td>
<td>Type 2 HS reaction</td>
<td>Steroids (Hydrocortisone, prednisone)</td>
</tr>
<tr>
<td>Joint pain</td>
<td>Hypotension</td>
<td>Type 3 HS reaction</td>
<td>Skin biopsy</td>
</tr>
<tr>
<td>Joint Swelling</td>
<td>Tachycardia</td>
<td>Type 4 HS reaction</td>
<td>Complement testing</td>
</tr>
<tr>
<td>Hives</td>
<td>Angioedema</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wheezing</td>
<td>Exanthema</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arthus reaction</td>
<td>Purpura</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neutrophilia/Neutropenia</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eosinophilia</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thrombocytopenia</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bronchospasm</td>
<td></td>
<td>Anaphylaxis</td>
</tr>
<tr>
<td></td>
<td>Urticaria (2)</td>
<td>Toxic Epidermal Necrolysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Eczema</td>
</tr>
</tbody>
</table>

---

(1) Indicates the most common manifestation.
(2) Indicates an additional symptom.

43
Composite outcome definition

Cohort Entry Events

Events having any of the following criteria:

- a condition occurrence of [EPI_767-Q] Skin manifestations...

- an observation of [EPI_767-Q] Skin manifestations...

with continuous observation of at least 0 days before and 0 days after event index date

Limit initial events: all events per person.

Restrict initial events

Inclusion Criteria

1. [EPI_767-Q] Joint manifestations (broad) - any musculoskeletal pain not including arthropathy, muscle pain

With at least 1 using all occurrences of:

- a condition occurrence of [EPI_767-Q] Joint manifestation...

where event starts between 0 days before and 0 days after index start date

The index date refers to the event from the Cohort Entry criteria.
Composite definition