A Method for Analyzing Commonalities in Clinical Trial Target Populations

Zhe (Henry) He¹, Simona Carini², Tianyong Hao¹, Ida Sim², and Chunhua Weng¹

¹Department of Biomedical Informatics, Columbia University
²Department of Medicine, University of California, San Francisco

AMIA 2014 Annual Symposium
November 18 2014
Disclosure

• All the authors disclose that they have no relationships with commercial interests.
Learning Objective

After this session, the learners should be aware of a database called COMPACT and its use for analyzing common eligibility features for related clinical trials.
Motivation

• **Observation 1**: Clinical studies are often criticized for the lack of generalizability

![Graph showing age distribution of patients enrolled in studies on dementia vs. demented patients in the general population](image)


• **Observation 2**: Many clinical trials on the same condition use similar or identical eligibility criteria (Hao et al. 2013)
Motivation (cont.)

- Make transparent the design pattern of research eligibility criteria
- Tradeoffs between internal validity and external validity (generalizability).
- Opportunity: The official trial registry ClinicalTrials.gov
  - 170,000 + summaries of studies in 180+ countries
Our long-term research agenda to analyze commonalities in target populations

Eligibility criteria text

Discrete data elements

Contextual and temporal information of features

Semantic equivalence between eligibility features

e.g., “Diabetic with HbA1c > 7.0% after injection of insulin”

e.g., Data elements: “diabetic”, “insulin”, “HbA1c” with its property “>7.0%”

e.g., HbA1c > 7.0% after insulin

e.g., “HbA1c > 7.0%” ≡ “diabetic”
Innovation

HbA1c $\leq 10.0$

Hemoglobin A1c $< 11.0$

HbA1c $\geq 7.0$

HbA1c $> 11.0$

A1c $> 7.0$

A1c less than 6.5

Unifying names and units

Database

Distribution of number of studies over boundary values for HBA1C

Number of studies

Boundary Values of HBA1C

Number of studies

Number of studies with Lower bound

Number of studies with Upper bound
COMPACT (Commonalities in Target Populations in Clinical Trials)

- Metadata of clinical trials, e.g., condition, study design, phase
- Structured eligibility criteria
- Common eligibility features
  - Numeric, e.g., BMI, blood pressure
  - Categorical, e.g., gravidity, lung cancer
- Properties of numeric features by disease domain
Pipeline of Constructing the COMPACT Database

1. Downloaded trial summaries from ClinicalTrials.gov
2. Extracting metadata of trials → Metadata table
3. Indexing trials by conditions → Trials_disease table
4. Extracting categorical eligibility features from eligibility criteria text → Categorical_features table
5. Extracting and analyzing numeric expressions in eligibility criteria text → Numeric_features table
Extracting Metadata of Trials

• Downloaded 159,891 trial summaries from ClinicalTrials.gov
• Excluded 704 trials with no or non-informative eligibility criteria text
• Extracted structured metadata of trials
• Retrieved NCT IDs for each condition in a list on ClinicalTrials.gov
  • Synonyms of conditions were consolidated, e.g., “heart attack” and “myocardial infarction”
Extracting Categorical Features

- Unsupervised tag mining of eligibility criteria text (Miotto et al. 2013)

- N-grams in eligibility criteria text that
  - Completely or partially match a UMLS concept assigned 27 semantic types that are relevant to clinical study domain
  - Normalized to Concept Unique Identifier of the UMLS
  - Appeared in at least 5% of trials on the same condition
  - Excluded concepts of the semantic type “Body Part, Organ, or Organ Component”, e.g., eye, ear.

Inclusion criterion: “HbA1c value between 7.5% and 11%”

Numeric expressions extracted by Valx (Hao et al. 2013):

["HbA1c", ">=", 7.5, "]%""] ["HbA1c", "<<=", 11.0, "]%""]

Properties of the numeric feature “HbA1c”:

• Value range: [7.5, 11.0]
• Boundary values: 7.5 and 11

Extracted in total 1,045,893 numeric expressions

Data Storage

Metadata
NCT02097342: Type 2 diabetes, phase 4, ...
NCT01784848: hypertension, phase 3, ...
.....

Categorical features
NCT02097342: metformin, diabetes
NCT01784848: hepatic, smoker ...
.....

Numeric features
NCT02097342: HbA1c < 7.5%, BMI >= 20 kg/m2
NCT00859300: Age > 18, ...
.....
Database Schema of COMPACT

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Numeric_features</th>
<th>Categorical_features</th>
<th>Trials_disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCTID</td>
<td>NCTID</td>
<td>NCTID</td>
<td>NCTID</td>
</tr>
<tr>
<td>brief_title</td>
<td>section</td>
<td>variable</td>
<td>disease</td>
</tr>
<tr>
<td>official_title</td>
<td>variable</td>
<td>semantic_group</td>
<td></td>
</tr>
<tr>
<td>location</td>
<td>value_range</td>
<td>value_width</td>
<td></td>
</tr>
<tr>
<td>agency</td>
<td>boundary_value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>start_date</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>minimum_age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maximum_age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>enrollment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>condition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>study_pop</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intervention_type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intervention_name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>authority</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>source</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>study_type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intervention_model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allocation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time_perspective</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What are the commonalities in Type 2 diabetes trials?
Frequent Categorical Features

Type 2 diabetes trials that recruit patients with HbA1c >= 7.0%

<table>
<thead>
<tr>
<th>Categorical features used in the inclusion criteria</th>
<th>Categorical features used in the exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>diabetes mellitus non-insulin-dependent</td>
<td>diabetes mellitus insulin-dependent</td>
</tr>
<tr>
<td>Disorders</td>
<td>Disorders</td>
</tr>
<tr>
<td>520</td>
<td>236</td>
</tr>
<tr>
<td>74.3%</td>
<td>33.7%</td>
</tr>
<tr>
<td>sulfonlurea compounds</td>
<td>pharmacologic substance</td>
</tr>
<tr>
<td>Chemicals &amp; Drugs</td>
<td>Chemicals &amp; Drugs</td>
</tr>
<tr>
<td>118</td>
<td>229</td>
</tr>
<tr>
<td>16.9%</td>
<td>32.7%</td>
</tr>
<tr>
<td>antidiabetics</td>
<td>allergy severity - severe</td>
</tr>
<tr>
<td>Chemicals &amp; Drugs</td>
<td>Disorders</td>
</tr>
<tr>
<td>94</td>
<td>224</td>
</tr>
<tr>
<td>13.4%</td>
<td>32.0%</td>
</tr>
<tr>
<td>pharmacologic substance</td>
<td>gravidity</td>
</tr>
<tr>
<td>Chemicals &amp; Drugs</td>
<td>Disorders</td>
</tr>
<tr>
<td>91</td>
<td>223</td>
</tr>
<tr>
<td>13.0%</td>
<td>31.9%</td>
</tr>
<tr>
<td>contraceptive methods</td>
<td>malignant neoplasm</td>
</tr>
<tr>
<td>Procedures</td>
<td>Disorders</td>
</tr>
<tr>
<td>83</td>
<td>190</td>
</tr>
<tr>
<td>11.9%</td>
<td>27.1%</td>
</tr>
</tbody>
</table>
### Frequent Numeric Features

<table>
<thead>
<tr>
<th>Numeric features used in the inclusion criteria</th>
<th></th>
<th></th>
<th></th>
<th>Numeric features used in the exclusion criteria</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Numeric features</strong></td>
<td><strong>Semantic group</strong></td>
<td><strong># Trials</strong></td>
<td><strong>Perc.</strong></td>
<td><strong>Numeric features</strong></td>
<td><strong>Semantic group</strong></td>
<td><strong># Trials</strong></td>
<td><strong>Perc.</strong></td>
</tr>
<tr>
<td>HbA1c</td>
<td>Physiology</td>
<td>663</td>
<td>94.7%</td>
<td>Creatinine</td>
<td>Chemicals &amp; Drugs</td>
<td>114</td>
<td>16.3%</td>
</tr>
<tr>
<td>BMI</td>
<td>Physiology</td>
<td>370</td>
<td>52.8%</td>
<td>Systolic blood pressure</td>
<td>Physiology</td>
<td>85</td>
<td>12.1%</td>
</tr>
<tr>
<td>Age</td>
<td>Physiology</td>
<td>327</td>
<td>46.7%</td>
<td>Diastolic blood pressure</td>
<td>Physiology</td>
<td>84</td>
<td>12%</td>
</tr>
<tr>
<td>Glucose</td>
<td>Chemicals &amp; Drugs</td>
<td>114</td>
<td>15.9%</td>
<td>ALT</td>
<td>Chemicals &amp; Drugs</td>
<td>73</td>
<td>10.4%</td>
</tr>
</tbody>
</table>

- **HbA1c (Glycohemoglobin):** average blood sugar level
- **BMI (body mass index):** weight(kg)/(height(m))^2
## Top Five Collective Value Ranges

<table>
<thead>
<tr>
<th>HbA1c value ranges</th>
<th>Number of trials</th>
<th>BMI value ranges</th>
<th>Number of trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>[7.0, 10.0]</td>
<td>228</td>
<td>(-∞, 45.0]</td>
<td>113</td>
</tr>
<tr>
<td>(7.0, +∞)</td>
<td>97</td>
<td>(-∞, 40.0]</td>
<td>104</td>
</tr>
<tr>
<td>(-∞, 7.0]</td>
<td>88</td>
<td>[25.0, 40.0]</td>
<td>72</td>
</tr>
<tr>
<td>[7.0, 11.0]</td>
<td>75</td>
<td>(-∞, 40.0)</td>
<td>61</td>
</tr>
<tr>
<td>[7.0, +∞)</td>
<td>57</td>
<td>(-∞, 35.0]</td>
<td>58</td>
</tr>
</tbody>
</table>
Collective Boundary Values of HbA1c

HbA1c boundary values in Type 2 Diabetes Trials

Number of Trials

HbA1c Value (%)
Collective Value Range Widths of HbA1c

HbA1c Value Range Width in Type 2 Diabetes Trials
VITTA: http://is.gd/VITTA
Discussion & Limitations

• Utility of the COMPACT database
• Limitations of NLP techniques
  • Partial semantics of the eligibility features
  • Accuracy of parsing numeric expressions
• Limitations of ClinicalTrials.gov
  • Partial or condensed eligibility criteria
  • Condition indexing errors of trials
Future Work

• Improve free-text parsers
• Sophisticated feature analysis
• Large-scale user evaluation
• Comparative analysis between patient population and target population
• Analysis of multiple features simultaneously
• ……

Summary

• COMPACT - a new resource for analyzing common eligibility features in clinical trials

• VITTA (http://is.gd/VITTA) – a Web-based visual analysis tool of clinical trial target populations

• To improve the transparency of design patterns of eligibility criteria for trials in a certain disease domain
Acknowledgments

Columbia University Medical Center

• Gregory Hruby
• Alexander Rusanov

(current) Mount Sinai

• Riccardo Miotto

Funding support:

• NLM R01 LM009886 (PI: Weng)
• NCATS UL1 TR000040 (PI: Ginsberg)
Thank you!

Zhe He, Ph.D.
zh2132@cumc.columbia.edu
http://www.columbia.edu/~zh2132/

Feedback:
http://is.gd/yhMnDX