Assessing the Population Representativeness of Related Clinical Trials Using Public Data Resources

Zhe He\textsuperscript{1}, Patrick Ryan\textsuperscript{1,2}, Shuang Wang\textsuperscript{3}, Chunhua Weng\textsuperscript{1}

\textsuperscript{1}Department of Biomedical Informatics, Columbia University
\textsuperscript{2}Janssen Research and Development
\textsuperscript{3}Department of Biostatistics, Columbia University
Disclosure

• All the authors have no relationships with commercial interests to disclose.
Generalizability of Clinical Trials

- Research studies that investigate how well a treatment for a medical condition work in people
- Lack of population representativeness \(\rightarrow\) Compromised generalizability
- Overly stringent eligibility criteria may
  - Overestimate the efficacy of the treatments (Blanco et al., 2008)
  - Exclude patients who may benefit from the trial (Okuda et al., 2010)

Dementia Trials


(Schoenmaker & Van Gool. 2004)
"Roughly 53 percent of new cancer diagnoses are in people 65 or older, but this age group accounts for just 33 percent of participants in cancer drug trials."

Opportunity

• Many clinical trials on the same condition use similar or identical eligibility criteria (Hao et al. 2013)

• Opportunity: The official trial registry ClinicalTrials.gov
  • 180,000 + summaries of studies in 180+ countries
  • Characteristics of trials: e.g., study type, study design, phase, interventions, etc.
  • Free-text eligibility criteria

Target Population vs. Patient Population in EHRs

Aims for This Work

1. To visualize differences between real-world patient population and collective target population of related trials

2. To quantify the population representativeness of clinical trials using national survey data
Transforming CT.gov to Computable Format

- Trial summaries from CT.gov
- Extracting metadata of trials
- Indexing trials by conditions
- Extracting categorical features
- Extracting numeric features

Identifying and formalizing numeric expressions

Matching n-grams against UMLS?

Yes

Normalizing measurement units

Converting exclusion criteria to inclusion criteria

Inclusion criteria:
HbA1c value between 7.5% and 11%

Numeric Features:
["HbA1c", ">=", 7.5, "%"]
["HbA1c", "=", 11, "%"]

Visual aggregate analysis of eligibility features of clinical trials

Zhe He\textsuperscript{a}, Simona Carini\textsuperscript{b}, Ida Sim\textsuperscript{b}, Chunhua Weng\textsuperscript{a,*}

\textsuperscript{a} Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA
\textsuperscript{b} Division of General Internal Medicine, University of California, San Francisco, San Francisco, CA 94143, USA

\textbf{A B S T R A C T}

\textbf{Objective:} To develop a method for profiling the collective populations targeted for recruitment by multiple clinical studies addressing the same medical condition using one eligibility feature each time.

\textbf{Methods:} Using a previously published database COMPACT as the backend, we designed a scalable method for visual aggregate analysis of clinical trial eligibility features. This method consists of four modules for eligibility feature frequency analysis, query builder, distribution analysis, and visualization.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{distribution.png}
\caption{Distribution of number of studies for HBA1C values}
\end{figure}
Patient Population from NHANES

- National Health and Nutrition Examination Survey (NHANES), a continuous cross-sectional health survey conducted by the NCHS of CDC
- Combined five two-year survey cycle: 2003-2012
- Diabetes?
  - told by health professional to have diabetes
  - had an HbA1c measurement
- Excluding Type 1 diabetes patients (Dodd et al. 2009)
  1. first diagnosed with diabetes before age 30
  2. taking insulin
- Out of 3,082 T2DM samples, 2,695 have values for age, HbA1c, and BMI

Patient Population from NHANES

- Sample weight: WTMEC10YR = 1/5 * WTMEC2YR
- 2,695 samples ➔ 15,575,484 T2DM patients in the U.S. national population (P1)

**Baseline characteristics of patient cohort.**

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample, n</td>
<td>2,695</td>
</tr>
<tr>
<td>Population, n</td>
<td>15,575,484</td>
</tr>
<tr>
<td>Age (mean ± stddev)</td>
<td>60.6 ± 13.3</td>
</tr>
<tr>
<td>Gender, male%</td>
<td>48.2</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
</tr>
<tr>
<td>Mexican American, %</td>
<td>8.5</td>
</tr>
<tr>
<td>Other Hispanic, %</td>
<td>5.5</td>
</tr>
<tr>
<td>Non-Hispanic White, %</td>
<td>61.8</td>
</tr>
<tr>
<td>Non-Hispanic Black, %</td>
<td>16.0</td>
</tr>
<tr>
<td>Other races, %</td>
<td>8.2</td>
</tr>
<tr>
<td>HbA1c, % (mean ± stddev)</td>
<td>7.2 ± 1.7</td>
</tr>
<tr>
<td>BMI, kg/m² (meaw ± stddev)</td>
<td>32.9 ± 7.6</td>
</tr>
</tbody>
</table>
Visualization for Age

- **T2DM patients**
- **Target populations of T2DM trials**

Age vs. Percentage of patients vs. Percentage of trials
Visualization for HbA1c

- T2DM patients
- Target populations of T2DM trials

Percentage of patients vs. HbA1c Value (%)
Visualization for BMI

- **T2DM patients**
- **Target populations of T2DM trials**

**BMI Value (kg/m²)**

**Percentage of patients**

**Percentage of trials**
Limitation in Visualization

- Not informative for certain features
- Difficult to compare population representativeness w.r.t. different features
Generalizability Index for Study Traits (GIST)

To quantify the representativeness of a given patient population in a set of trials

\[
GIST = \sum_{i=1}^{N} \frac{T}{\sum_{j=1}^{T} I([i_{low}, i_{high}] \subset w_j)} \times \frac{P}{\sum_{k=1}^{P} I(i_{low} \leq y_k < i_{high})}
\]

\(N\): number of distinct value intervals of the quantitative feature

\(T\): number of trials

\(P\): number of patients

\(w_j\): inclusion value interval of the quantitative feature for the \(j^{th}\) study

\(y_k\): observed value of the quantitative feature for the \(k^{th}\) patient

# GIST Scores for Different Phases of Trials

<table>
<thead>
<tr>
<th>Phase</th>
<th>Variable</th>
<th>Age</th>
<th>HbA1c</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Age</td>
<td>0.60 (N=368)</td>
<td>0.55 (N=141)</td>
<td>0.64 (N=204)</td>
</tr>
<tr>
<td>II</td>
<td>HbA1c</td>
<td>0.77 (N=517)</td>
<td>0.45 (N=244)</td>
<td>0.71 (N=194)</td>
</tr>
<tr>
<td>III</td>
<td>BMI</td>
<td>0.87 (N=766)</td>
<td>0.38 (N=438)</td>
<td>0.77 (N=356)</td>
</tr>
<tr>
<td>VI</td>
<td>HbA1c</td>
<td>0.80 (N=484)</td>
<td>0.42 (N=306)</td>
<td>0.69 (N=194)</td>
</tr>
<tr>
<td>All</td>
<td>BMI</td>
<td>0.77 (N=2702)</td>
<td>0.44 (N=1463)</td>
<td>0.69 (N=1274)</td>
</tr>
</tbody>
</table>
Single Feature vs. Multiple Features

• Single-feature GIST score is not efficient for comparing the population representativeness of multiple sets of trials

• Eligibility features may have inherent correlations
  • E.g., Impaired fasting glucose is correlated with age (Cowie et al. 2006)

Visualization of the Eligibility Space

HbA1c, BMI, and age in T2D studies recruiting female
Multivariate GIST

\[
GIST = \sum_{n=1}^{N=\prod_{c=1}^{C} N_c} \left( \sum_{j=1}^{T} \prod_{c=1}^{C} \prod_{i_c=1}^{N_c} I(i_{c,low} \leq w_{j,c} < i_{c,high}) \right) \sum_{k=1}^{P} \prod_{c=1}^{C} \prod_{i_c=1}^{N_c} I(i_{c,low} \leq y_{k,c} < i_{c,high})
\]

C: number of features

\(N_c\) is the number of distinct intervals within each study trait \(c\)

N: total number of combination of distinct intervals across all study traits

T: number of trials

P: number of patients

\(w_{j,c}\): the inclusion interval of trait \(c\) for the \(j^{th}\) study

\(y_{k,c}\): observed value of the trait \(c\) for the \(k^{th}\) patient
# Multivariate GIST Scores of T2DM Trials

<table>
<thead>
<tr>
<th>Trial Characteristic</th>
<th>Number of Trials</th>
<th>GIST Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study Type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interventional</td>
<td>2731</td>
<td>0.43</td>
</tr>
<tr>
<td>Observational</td>
<td>423</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>Start Date</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01/2003 – 12/2004</td>
<td>333</td>
<td>0.52</td>
</tr>
<tr>
<td>01/2005 – 12/2006</td>
<td>477</td>
<td>0.51</td>
</tr>
<tr>
<td>01/2007 – 12/2008</td>
<td>743</td>
<td>0.47</td>
</tr>
<tr>
<td>01/2009 – 12/2010</td>
<td>853</td>
<td>0.45</td>
</tr>
<tr>
<td>01/2011 – 12/2012</td>
<td>752</td>
<td>0.44</td>
</tr>
</tbody>
</table>
## NHANES or EHR?

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>NHANES</th>
<th>EHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Nationally representative population</td>
<td>Potentially-biased patients (treatment-seeking)</td>
</tr>
<tr>
<td>Data Quality</td>
<td>Structured and readily analyzable. No NLP needed</td>
<td>Partially structured. NLP needed</td>
</tr>
<tr>
<td>Richness</td>
<td>Limited data</td>
<td>Rich data</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>Yes</td>
<td>Almost no</td>
</tr>
<tr>
<td>Suitability</td>
<td>Population-based studies</td>
<td>Longitudinal/observational studies</td>
</tr>
</tbody>
</table>
Summary & Future Work

• Assessing population representativeness of multiple trials using national survey data

• Will apply this method on other medical conditions

• Long-term goal
  • Improving the transparency of population representativeness of related trials
  • Facilitating cost-effective data-driven precision design of clinical trials
Acknowledgments

Funding support:

• **U.S. National Library Medicine**
  
  R01 LM009886 (PI: Weng)

• **National Center for Advancing Translational Science**
  
  UL1 TR000040 (PI: Ginsberg)
Thank you!

Questions?

Contact:
Zhe He
zh2132@cumc.columbia.edu
Additional Slides
Selecting Frequent Eligibility Features

- Age, HbA1c and BMI were identified as frequently used quantitative features used by 99.0%, 53.6%, and 46.6% of T2DM trials, respectively.

- Gender is chosen as a categorical feature.
**Patient Population from NHANES**

Numbers of NHANES samples with diabetes, Type 2 diabetes, and Type 1 diabetes in each two-year group.

<table>
<thead>
<tr>
<th>Survey Year</th>
<th>Total number of samples</th>
<th>Number of samples with diabetes</th>
<th>Number of samples with Type 2 diabetes</th>
<th>Number of samples with Type 1 diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003-2004</td>
<td>10,122</td>
<td>559</td>
<td>525</td>
<td>34</td>
</tr>
<tr>
<td>2005-2006</td>
<td>10,348</td>
<td>521</td>
<td>487</td>
<td>34</td>
</tr>
<tr>
<td>2007-2008</td>
<td>10,149</td>
<td>777</td>
<td>721</td>
<td>56</td>
</tr>
<tr>
<td>2009-2010</td>
<td>10,537</td>
<td>739</td>
<td>688</td>
<td>51</td>
</tr>
<tr>
<td>2011-2012</td>
<td>9,756</td>
<td>708</td>
<td>661</td>
<td>47</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>50,912</strong></td>
<td><strong>3,304</strong></td>
<td><strong>3,082</strong></td>
<td><strong>222</strong></td>
</tr>
</tbody>
</table>

Out of 3,082 T2DM samples, 2,695 have values for age, HbA1c, and BMI.
Generalizability Index for Study Traits (GIST)

Set of trials

- Trial 1: Age between 20-50
- Trial 2: Age >= 20

Set of Patients

- 10 patients < 20
- 60 patients 20-50
- 30 patients > 50

<table>
<thead>
<tr>
<th>Age interval</th>
<th>Percentage of trials</th>
<th>Percentage of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 20</td>
<td>0%</td>
<td>10%</td>
</tr>
<tr>
<td>20 - 50</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>50 - inf</td>
<td>50%</td>
<td>30%</td>
</tr>
</tbody>
</table>

\[
GIST = 0\% \times 10\% + 100\% \times 60\% + 50\% \times 30\% = 0.75
\]