The Problem

- Clinical trial descriptions are written at 17th-grade reading level.
- Fewer eligible and interested people are able to understand these descriptions.
- This reduces the pool of potential candidates.

The data

- 268,000 trial descriptions from ClinicalTrials.gov
- XML format
- Most information of interest contained in unstructured text fields

Supervised learning: multiclass regression

- Development of annotation codebook
- Regression models to estimate: number of visits, number of interventions, active period, evaluation period, total study period, and whether or not the trial is of indefinite length
- Tf-idf feature generation

Model performance

Figure 4: Normalized error rates for regression models: KNN, Extra Tree, and Decision Tree

Future directions

- Transformer language model-based feature generation for regression
- Simplification and/or abstract summarization using neural models

Keyword indicator matching

<table>
<thead>
<tr>
<th>Type of information</th>
<th>XML field</th>
<th>Indicator</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>size of study</td>
<td>brief_summary</td>
<td>( \text{#} + ) patients</td>
<td>3.02</td>
</tr>
<tr>
<td>size of study</td>
<td>brief_summary</td>
<td>projected accrual</td>
<td>0.0</td>
</tr>
<tr>
<td>size of study</td>
<td>brief_summary</td>
<td>subjects</td>
<td>10.31</td>
</tr>
<tr>
<td>size of study</td>
<td>brief_summary</td>
<td>total of ( \text{#} + ) patients</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Figure 2: Example per-document keyword frequencies

Dependency parsing

Sentence: Functional biomechanical outcomes will be measured at 6 months and 12 months using DXA at the Biodynamics Lab.
Extract: measured at 6 months and 12 months
Sentence: Patients who were discharged after an uneventful ERCP were contacted by telephone within 5 days to capture delayed occurrence of the primary end point.
Extract: contacted within 5 days
Sentence: All subjects will receive surgical treatment of their SCCis.
Extract: subjects will receive treatment SCCis
Sentence: Patients receive tacrolimus IV over 24 hours or orally daily on days -3 to 35 and oral mycophenolate mofetil twice daily on days -3 to 28 as graft-vs-host disease (GVHD) prophylaxis.
Extract: Patients receive tacrolimus IV

Figure 3: Extracting information using dependency parsing

Wikification

Key inclusion criteria: - males and females, 18 to 65 years of age, with HIV infection and a body mass index of 18.0 to 35.0 kg/m\(^2\). - HIV-infected participants receiving a treatment regimen containing only atazanavir/ritonavir, 300/100 mg once daily (OD) orodovir, 300 mg OD at least 1 other nucleotide reverse transcriptase inhibitor continuously for at least 3 months prior to study day 1. - Plasma HIV RNA levels of CD4 count > 200 cells/mm\(^3\). - No history of virologic failure on a protease inhibitor (PI), documented phenotypic PI resistance, or primary PI mutations, according to International AIDS Society recommendations. - No documented phenotypic resistance to atazanavir or primary genotypic mutations causing resistance to atazanavir. - History of Gilbert's syndrome, hemophilia, chronic pancreatitis, hypochlorhydria, achlorhydria, clinically relevant gastrosophageal reflux disease, hiatal hernia, or peptic/gastric ulcer disease. - Intractable diarrhea (≥ 6 loose stools/day for at least 3 consecutive days) within 30 days prior to study day 1. - Recent (within 6 months prior to study day 1) drug or alcohol abuse. - Evidence of organ dysfunction or any clinically significant deviation from normal in physical examination, vital signs, electrocardiogram (ECG).

Figure 4: Automatic linking of Wikipedia articles