1. Problem Statement and Motivation

**Context:** Without sites there would be no clinical trials and no new medications.

Takeda has **multiple clinical site options** to choose from and thus, site performance is critical.

**Sub-objectives**

1. Will a site be non-enrolling?
2. Will a site be a low-moderate or high enroller?
3. What's the time point when a site will never enroll?
4. What is the optimal site selection for a given study?

**Why does it matter?**

- Getting drugs out faster to patients in need
- Better allocate resources on new drugs development
- Decrease costs and study delays:
  - A typical trial can cost ~$86M
  - Delayed trials take +1-6 months

2. Data

We created 3 distinct datasets for our analysis leveraging internal trials data mostly after 2010.

~140 modelling factors including both site and study characteristics

3. Exploratory Analysis

**# of sites per performance category**

<table>
<thead>
<tr>
<th>% of sites</th>
<th>non-enrolling</th>
<th>low- moderate</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>15%</td>
<td>35%</td>
<td>50%</td>
<td>10%</td>
</tr>
<tr>
<td>10%</td>
<td>30%</td>
<td>10%</td>
<td>60%</td>
</tr>
</tbody>
</table>

~30% of all site-studies are non enrolling, with that % decreasing in recent years

4. Methodology

**3 Predictive Machine Learning Models**

1. Classification model to predict probability of non-enrolling sites
2. Multi classification model to predict low-med-high enrolling sites
3. Survival model (log-logistic AFT) to predict time inflection point of never enrolling

**1 Prescriptive Optimization model**

**Dynamic optimization model for site selection:**
- Using closed form expression of classification models as constraints
- Accounting for complex interactions with dynamic optimization
- Maximize expected enrollment while minimizing costs
- Control over minimal proportion of high enroller and minimal proportion of low enrollers
- Piecewise linear approximation of the sigmoid

These 4 analytics models built will allow Takeda to act on three different parts of the site selection & management process

5. Results

**Built high performing models:**

- AUC for best performing model: ~0.93
- 2nd model: ~0.8 AUC
- 3rd model: ~0.7 C-index

That provided actionable recommendations: Identified subset of most impactful site and study characteristics affecting enrollment

**Impact and Correlation Matrix**

**Impact of our work**

$200M Avg. 5-year cost savings just by considering non-enrolling sites. Could even be more (i.e.: entry to market saving)

**Accelerate drug development:** Getting drugs out faster to patients – advancing society

**Implementation**

Currently testing our solution on a two-study pilot

Our project will be implemented in the clinical analytics hub in a 2–3-year horizon

**Future Areas of work**

Data collecting: Incorporate external data and plan to collect further information from CROs

Expand scope: Include in analysis other KPIs (i.e.: retention) and to include ongoing effect of other actions

**Higher performing sites contribute to most of total patient enrollment (~50%) while being the minority in number of sites (~20%)**

**Impact and Correlation Matrix**

<table>
<thead>
<tr>
<th>Impact</th>
<th>Correlation</th>
<th>Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient experience</td>
<td>Site density per State</td>
<td>Free experience working with site</td>
</tr>
<tr>
<td>Site density per City</td>
<td>Site density per PCU</td>
<td>Pediatric study</td>
</tr>
<tr>
<td>Site density per PCU</td>
<td>Site density per PCU</td>
<td>Site density per PCU</td>
</tr>
</tbody>
</table>

Both study and site characteristics significantly affect enrollment outcome

**Faculty Advisor**

Retsef Levi

**Takeda Team**

Saurabh Awasthi
Stephen Cue
Melissa Chiasson

**Capstone Team**

Maria Camila Marenco
Aziz Ayed

**Shujaullah Mohammed**

Scion Li