Using Artificial Intelligence Techniques to Match Patients with Their Best Clinical Trial Options

Gina Tourassi
Ioana Danciu
Health Data Sciences Institute
Oak Ridge National Laboratory

Gil Alterovitz
US Department of Veterans Affairs, Presidential Innovation Fellows program

Introduction

The Presidential Innovation Fellows, US Department of Veterans Affairs, and the Oak Ridge National Laboratory Health Data Sciences Institute are coordinating this Data Challenge, which draws on resources across a dozen federal agencies and departments. The related project, Health Tech Sprint, emphasizes the need for open federal data for artificial intelligence (AI) applications as defined by the newly signed OPEN Government Data Act under the Foundations for Evidence-based Policymaking Act (signed Jan 15, 2019).

Novel therapeutics, such as those under development in clinical trials, are often a treatment option for patients with serious and life-threatening diseases such as cancer. Increasing patient awareness of clinical trials is believed to be a factor in reducing time for participant recruitment, a very large cost category in clinical trials. Thus, applying AI to help patients and their health care providers find clinical trials of novel therapeutics may improve patient care and, by aiding in recruitment, reduce drug development costs.

For AI to be useful in trial matching, both representative patient data and clinical trial eligibility information, ideally in a structured format, are needed. In addition, expert-based guidance on matching patients to trials, including which criteria are matched, is useful for building and testing models.

The AI-able data ecosystem seeks to enable AI by bringing together an ensemble of interlinked datasets with data suitable for AI in a given use case. Having this information in the public domain enables standardization by facilitating testing across different approaches. This challenge features the first such standardized dataset ensemble related to clinical trial matching, with the various interlinked datasets provided.

Datasets

Three datasets are available to Data Challenge participants:

1. A subset of eligibility criteria translated into machine-readable code from a selected group of cancer clinical trials.
2. Records based on callers to the NCI's Cancer Information Service that have been enhanced with synthetic data and translated into machine-readable code.
3. Participant records matched against clinical trials for which the eligibility criteria and participant data were previously translated into machine-readable code.

A second version of the third dataset, produced by oncology professionals, serves as a comparison dataset for the matches identified through the application of AI.
For more information on the above datasets and potential approaches on usage, please see https://digital.gov/2019/02/27/how-a-health-tech-sprint-inspired-an-ai-ecosystem/.

In addition to the datasets provided, participants are encouraged to use other publicly available datasets. For example, National Cancer Institute (NCI)-funded cancer clinical trials, including API with annotations on disease eligibility criteria for all trials, is available at https://clinicaltrialsapi.cancer.gov

Challenge Questions

Challenge questions are listed below. However, participants are encouraged to suggest and tackle challenge questions different from those listed below. Innovative use of the provided data is strongly encouraged.

1. Data representation
    - Develop novel big data structures to represent the clinical trials and the patient data that accommodate the interaction of the three datasets. The ultimate goal is to support thousands of clinical trials being matched with millions of people.
2. Algorithm development
    - Develop novel algorithms for finding the most suitable matches between patients and clinical trials.
3. Visualization/human computer interaction
    - Develop visualization and/or human-computer interaction solutions to enable medical providers to effectively leverage the data for clinical decision support.

Notes on the Challenge Questions:

- A participant may choose to do any question(s) they prefer. Completing all three questions is optional.
- Regarding approaches to question 2, our preference is to receive solutions involving machine learning techniques.