Domo Arigato Mr. Roboto

Pouring through spreadsheets of data to pick out individuals that might be at risk of breast and ovarian cancer even without a known genetic risk factor sounds like something you might hire an intern to do. In fact, it requires some pretty specialized skills to determine what family history indicators might make someone more at risk than another. Genetic counselors are excellent at finding these, sometimes nuanced, indicators but are limited on time. This challenge requires the development of an AI that can not only find those individuals that may be at risk, but learn as it goes to build a more robust risk model based on both genetic and family history data.

Long Description

In large population-based genetic studies, it is common to have thousands of individuals provide not only a DNA sample, but also detailed information about their personal history and family history of disease. In most cases, that information is obtained via paper or an online survey. Regardless, the data ends up on a giant spreadsheet, typically sorted by participant ID or name so that it can eventually be linked back up with any results that are found in the genetic analysis.

Interestingly, most studies only link back to information provided by the participant if there is a genetic finding to help clarify and validate the finding. In a recent study at HudsonAlpha the research team took a different approach and systematically reviewed every participant's personal and family history of breast and ovarian cancer. In order to accomplish that review, a team of genetic counselors would review, row by row, each participant's data and categorize it into one of a few categories (descriptions included in the data for this challenge) and then analyzed how many of the participants needed follow-up conversations about their risk for breast and ovarian cancer despite any genetic result finding. With diseases like breast and ovarian cancer where environmental and genetic factors along with strong family history of the disease all contribute to the personal risk of an individual, it is imperative that those in the moderate to high risk categories are provided the opportunity for care they need.

This challenge is designed around a simple idea, create an AI system that replaces the genetic counselor's analysis of every participant's personal and family history and categorize those individuals into the appropriate group. The system will need to be able to learn from human verification and recategorization. A large, deidentified dataset is provided for both building the initial model and testing. A document describing the genetic counselor decision-making process is also included.

To complete the entire challenge, participants should consider creating a functional product that contains both parts; Model and End-to-End. However, participants may also choose to focus on only one of the sub-challenges.

1. Model
   To complete this sub-challenge of Domo Arigato Mr. Roboto teams will need to create a functional model for sorting patient data into the three risk categories based on personal and family history data. The model should include not only the

ability for the data to be sorted correctly, but it should also have the ability to correctly assign a category to new patient data. Finally, the completion of this sub-section should include the ability for the model to learn from human modification of the categorization. For example, if a genetic counselor sees a risk category assignment made by the model that he/she does not agree with and changes the assigned category, the program should be 'intelligent' enough to learn from that moving forward. Of course, parameters ought to be placed around that functionality so that the system can truly learn from the process and not become increasingly poor at categorization with the introduction of human intervention on unique cases.

2. End-to-End
   To accomplish true integration of AI into the process, the entire process from data collection to output of categorization must be considered. To successfully complete this sub-challenge, teams will need to create an end-to-end solution to reliably collect data from patients (you do NOT need to consider HIPAA standards or FHIR integration), handle both discrete data input as well as free-text, run the AI to categorize each patient, provide an interface for human recategorization that feeds back into the AI model, and create a user interface for healthcare providers to easily see patients in each of the risk categories, the details for each patient that was used for classification, and any genetic findings for that patient. This is both a UI and data collection/storage sub-challenge. The end-to-end solution should be robust and be able to store and retrieve thousands of records. Historical look up will be necessary.