This Challenge asks Solvers to work on improving methods to identify subgroups of patients that will derive a more pronounced benefit from a medical treatment. Effective methods in this domain can have important implications for medical care by identifying ideal treatments for individual patients, as underscored by the recently announced Precision Medicine initiative (http://www.nih.gov/precisionmedicine/). As an example, Vectibix (panitumumab) has been shown to be effective only in patients whose tumors do not have certain KRAS mutations (http://www.accessdata.fda.gov/drugsatfda_docs/label/2009/125147s080lbl.pdf).

Here are a number of key concepts that define this Challenge:

1. **Treatment effect**: To determine its effectiveness, a novel treatment needs to be compared against a standard treatment, which can be placebo (i.e. sugar pills or a sham treatment) or the best established treatment. As such, the response of interest for a patient is defined to be the added benefit he or she receives from the novel treatment as compared to the standard treatment. This response is commonly referred to as the ***treatment effect***. In this Challenge we consider treatments that ***reduce*** certain symptom scores. Mathematically, the treatment effect for a patient would be zero if the patient receives no additional benefit from the novel treatment; but if the patient indeed benefits from the novel treatment, the treatment effect would manifest as a negative number on the scale of the response variable (since the aforementioned goal of treatment is to reduce a symptom score). Once we have defined treatment effect for each patient as above, we can refer to treatment effect of any group of patients as the average of the individual treatment effects of the group members. Note that the treatment effect, both at the individual and group level, is rarely known but can be estimated from appropriate data.
2. **Patient subgroups**: The gold standard of demonstrating effectiveness of a novel treatment is to compare it with a standard treatment in a randomized controlled double-blind clinical trial. In this Challenge, we are interested in analyzing data from such trials to identify subgroups of patients who enjoy an enhanced treatment effect. That is, the treatment effect associated with such a subgroup is meaningfully greater than the treatment effect associated with the corresponding complementary subgroup consisting of all patients who do NOT belong to the subgroup. Often, analyses to identify these subgroups are performed for the purpose of tailoring the novel treatment, so that it can be given to those patients who derive the most benefit from it. The subgroups under consideration in the analysis are defined by one or more of a number of patient characteristics (which we will refer to as the ***X***'s) that can include demographic (e.g. age, sex), disease description (e.g. duration, severity), and biomarkers measured by molecular technologies (e.g. genetics, protein levels).
3. **Multiplicity**: If the number of ***X***'s in the analysis is moderate or even large, as is often the case, the number of possible subgroups they can define can be exceedingly large, given the possibility of defining a subgroup based on a number of ***X***'s, such as "patients who are ≥50 years old AND having more severe disease". It is crucial to account for this level of multiplicity in the analysis, in order to guard against false positive findings, which often occur when an analysis identifies a subgroup that seems to have an enhanced treatment effect in the data at hand, but subsequent clinical trials fail to show this enhanced treatment effect (i.e. fail to replicate the finding). Obviously, a subgroup finding that cannot be replicated has no value in the treatment of future patients. Given the considerations of multiplicity and practicality of clinical application, all else being similar, simple subgroups (i.e. those defined by a small number of ***X***'s) are preferred over complex ones.

Solutions to this Challenge will be computer algorithms that successfully identify patient subgroups within data provided for this Challenge.

TESTING DATA

The Prodigy testing data consists of 1,200 datasets combined into one file, *InnoCentive_9933623_Data.csv* - see Challenge Attachments, with the following columns:

- ***dataset***: Dataset label with values 1-1200. Each dataset consists of all the rows that share the same value in this column.
- ***id***: Patient identification number within each dataset with values 1-240.
  - *Note: Patients with the same "id" in different datasets (e.g. the 3rd patient in dataset #1 and the 3rd patient in dataset #2) are unrelated.*
- ***trt***: Treatment label with values 0 and 1 indicating standard and novel treatments, respectively.
- ***y***: Response (symptom level after treatment) of each patient, with lower values being more desirable for patients.
- ***x1, x2, …, x20***: Genetic characteristics of each patient with values 0, 1, and 2.
- ***x21, x22, …, x40***: Patient characteristics that take on continuous values.

To perform the analysis for this Challenge, the Solver needs to take these steps upon downloading the combined data:

1. Partition the entire file into 1,200 separate datasets according to the ***dataset*** column.
2. Analyze each of the datasets to:
   a. Attempt to identify a patient subgroup as defined by one or more of the ***x***'s, such that the average treatment effect in this subgroup is meaningfully better than that in the complementary subgroup. In this Challenge, "meaningfully better" is defined as an enhanced treatment effect of -0.6 or greater (i.e. greater in the negative direction).
   b. Create a binary vector of subgroup membership for all the patients in this dataset, with the value "1" indicating that the corresponding patient belongs to the identified subgroup, and the value "0" indicating the opposite.
   c. IMPORTANT: If no meaningful subgroup is identified in Step (a), the binary vector in Step (b) should either have value "1" for all patients or have value "0" for all patients.
3. Combine the results, once all datasets have been analyzed as above, by "binding" the subgroup membership vectors into a 240-by-1201 matrix as follows:
   - Column #1, to be named "***id***": the vector of id with values from 1 to 240
   - Column #2, to be named "***dataset_1***": the subgroup membership vector for the 1st dataset
   - Column #3, to be named "***dataset_2***": the subgroup membership vector for the 2nd dataset
   - …
   - Column #1201, to be named "***dataset_1200***": the subgroup membership vector for the 1200th dataset
4. Format the matrix as a CSV file and submit as a **compressed zip** file to Prodigy using the "Test Solution" tab above.

TRAINING DATA

Training data to assist in method development is provided in *InnoCentive_9933623_Training_Data.csv* (four datasets in one file) and *InnoCentive_9933623_Training_Data_truth_subjects.csv* (perfect solution.) The description of the data generating model for these training datasets is as follows:

	Training data	**Predictive marker(s)**	**"Perfect" Subgroup**
dataset="1"	#1	None	None
dataset="2"	#2	1 continuous (x29)	x29 < 51.3
dataset="3"	#3	1 ordinal (x5)	x5 = 1 or 2
dataset="4"	#4	1 continuous (x22) & 1 ordinal (x4)	{x4 = 1 or 2} & {x22 > 57.7}

Notes:

- Similar to the Prodigy data, 4 training datasets have been combined into one single file. The individual datasets can be obtained by utilizing the "dataset" column of the meta data file.
- Dataset #1 was generated with the same treatment effect for all patients. That is, the only reason for the differences among observed responses to the same treatment across patients is random noise. Hence, there is NO meaningful subgroup for this dataset, and the perfect solution should have "0" for all subjects. (We use "all 0's" since the overall treatment effect here is lower than the required threshold. However, a solution with "all 1's" would receive the same score.)
- Dataset #2: there is a single subgroup associated with the data generating model. To obtain the perfect solution, one needs to determine two aspects: (1) ***x29***, and only ***x29***, is associated with treatment effect; (2) the cutoff value for this continuous variable that defines the patient subgroup is ***x29 < 51.3***. The perfect solution assigns "1" to any subject whose ***x29*** value is less than 51.3, and "0" to all other subjects. However, a solution that comes close to the perfect solution—for example "*x29 < 60*"—will also be assigned a high score.
- Dataset #3: there is a single subgroup associated with the data generating model. To obtain the perfect solution, one needs to determine two aspects: (1) ***x5***, and only ***x5***, is associated with treatment effect; (2) for the 3 levels of this ordinal variable, "1" and "2" should be combined to form the patient subgroup. The perfect solution assigns "1" to any subject whose ***x5*** value is either 1 or 2, and "0" to any subject whose ***x5*** value is 0.
- Dataset #4: there are two variables associated with the data generating model. To obtain the perfect solution, one needs to determine: (1) both ***x4*** and ***x22*** (but no other variables) are associated with treatment effect; (2) for the 3 levels of the ordinal variable ***x4***, "1" and "2" should be combined to form the patient subgroup; (3) the cutoff value for the continuous variable ***x22*** that defines the patient subgroup is ***x22 > 57.7***; (4) the two variables should be combined to form the ideal subgroup. The perfect solution assigns "1" to any subject who meet both criteria (that is, ***x4*** is 1 or 2, AND ***x22 > 57.7***), and "0" to all other subjects. There are various ways in which an imperfect solution can receive high or moderately high scores; for example, if one identifies only one of the two variables, or if one uses a cut-off value for ***x22*** that is different from 57.7.

SCORING

When a solution is submitted to Prodigy, it will be automatically evaluated and receive a score between 0 and 100, with higher scores representing better performance. Here is some pertinent information as related to the scoring:

- The 1,200 datasets have been generated under a variety of models to enable a comprehensive assessment of each submission. Both false positive (i.e. identifying a subgroup when there is none) and false negative (i.e. failing to identify a subgroup when there is one) results will reduce the performance score.
- The result of each dataset will first be scored individually, and then all dataset scores will be combined as a weighted average to produce the overall score. The weights reflect the relative importance of the various data generating models.
- The result of an individual dataset will be scored as follows:
  - The true treatment effect for the identified subgroup is computed based on the data generating model for the dataset.
  - Any result is given a score between 0 and 100, with two anchor points:
    - The score 0 is given to the trivial "coin toss" solution, which is a binary subgroup membership vector obtained by flipping a fair coin 240 times (240 is the number of patients in each dataset).
    - The score 100 is given to the perfect solution, which is:
      - ***If there is no meaningful subgroup for this dataset***: The entire group of patients (i.e. the subgroup membership vector consists of all 1's or of all 0's);
      - ***If there is at least one meaningful subgroup for this dataset***: The subgroup consisting of all of—and only—those patients whose treatment effects exceed the meaningful threshold (i.e. -0.6, or more negative).
    - All other solutions will receive a score between 0 and 100 on a linear scale, depending on how the treatment effect in the subgroup compare with those of the trivial and perfect solutions.

PRODIGY SUBMISSION FOR SCORING

The Prodigy online scoring tool and leaderboard will be used to track Solver performance on the testing data for this Challenge. The matrix of subgroup membership vectors described under the **TESTING_DATA** heading above should be in CSV format as demonstrated in the example submission file *InnoCentive_9933623_example_submit.csv* and submitted for online scoring in **compressed zip** file form using the "Test Solution" tab above or "Test your Solution" link below. Solvers are limited to **5 submissions for scoring per day**.

Uploads to the online scoring tool DO NOT constitute a full submission to this Challenge – Solvers must submit a full submission through their Project Room to be considered for an award. Please use the "My Solution" tab or "Submit Solution" button above for full submissions to this Challenge. While feedback provided by the "Test Solution" function is designed to indicate real-world performance on a common benchmark, any awards are contingent upon theoretical evaluation and experimental validation of the submitted solution by the Seeker and neither InnoCentive nor the Seeker makes any warranties or claims regarding the accuracy of online scores.

SOLUTION REQUIREMENTS

To be eligible for an award, the Solver's method must achieve a score that exceeds 60 on the testing data using the Prodigy online scoring tool, and 55 on the validation data held by the Seeker and used during the evaluation phase of the Challenge. If multiple submissions meet both requirements, performance on the validation data will be the primary deciding factor along with the supporting rationale for the algorithm employed.