Note: There is so much missing data that I am exploring use of the Ambulatory Health Care Data on Emergency Department Length of Stay. After searching through all of this, I found NAMCS and NHAMCS data can also be downloaded from the Inter-University Consortium for Political and Social Research (ICPSR) (This Wiki) This is what HealthData.gov should be! Also analyzing Semantic Medline. ICPSR offers more than 500,000 digital files containing social science research data. Disciplines represented include political science, sociology, demography, economics, history, gerontology, criminal justice, public health, foreign policy, terrorism, health and medical care, early education, education, racial and ethnic minorities, psychology, law, substance abuse and mental health, and more.

HPN Health Prize

Identify patients who will be admitted to a hospital within the next year, using historical claims data.

Description

Evaluation
I heard about this Health Prize at the Strata 2012 Conference. I wrote about it for AOL Government. I registered for the Prize and enhanced their content by building a knowledgebase and dashboard to explore their metadata and data. I emailed the Health Prize organizers (Kaggle) about my work, and used the Health Prize data set for my HDI Forum III: Health Datapalooza Call for Submissions: http://www.health2con.com/devchallenge/hdi/. I am looking for other data sets like the NHAMCS to inventory and analyze for the Health Data Prize to meet the April 4th deadline for proposed use of other data sets.

I emailed HPN/Kaggle the following: I am a Data Scientist/Data Blogger for AOL Government and am doing a story (s) on this prize. I am more interested in the properties and visualization of the data sets than the statistical models. I have received no response.

I submitted this to the HDI 3 to (1) use health data and make a meaningful impact on health, and to (2) show apps and tools that have successfully harnessed data and are aimed at three broad audiences: consumers & patients, providers & payers, public health & communities. My HDI Forum III submission is a business operations apps that is focused on cost, efficiency and quality, and that leverages health data to improve health outcomes, reduce health care costs, and increase value.
I found the following from the Strata 2012 Conference that explained my interests: "I distinctly remember this O'Reilly interview with Jeremy Howard, which was ranked one of the best data interviews of the conference. Those who couldn't be there should also check out the Domain Expertise vs Machine Learning debate (moderator Mike Driscoll's summary [here](http://semanticommunity.info/HPN_Health_Prize) or full video [here](http://semanticommunity.info/HPN_Health_Prize)). Final conclusion from Mike Driscoll:

*Thus who you decide to hire as your first data scientist — a domain expert or a machine learner — might be as simple as this: could you currently prepare your data for a Kaggle competition? If so, then hire a machine learner. If not, hire a data scientist who has the domain expertise and the data hacking skills to get you there.*

This explains my interest: I thought that even with the Kaggle preparation of the Health Prize data set, even more data science work needed to be done to understand what could really be done with the data than just by brute force data mining - statistical modeling.

At the recent Strata 2012 Conference, the top two teams received $80,000 in the second progress prize, split as $50,000 and $30,000. each. Once again, the team of "Market Makers" continued to lead the competition, winning the top spot. Players include David Vogel, CEO and Chief Scientist of Voloridge Investment Management, Dr. Randy Axelrod, President of Randy Axelrod Consulting and Resource Global Health, and Phil Brierley, a data mining expert from Australia. Willem Mestrom, an information analyst from the Netherlands, again won second place in the prize contest, and has added a new team mate, Edward de Grijs, an engineer and software developer also from the Netherlands.

In looking at their work, I learned they are taking a purely statistical data mining approach use multiple models to solve the problem. I am more interested in a visual approach to solving the problem, leading to a statistical model, but that provides insight into the relationships.

I am exploring the use of multivariate techniques, tools (Data Relationships, K-Means Clustering, Line Similarity, and Hierarchical Clustering). I tried Trellis: A split view of a visualization, organized by category in separate panels.

The Data Relationships tool is used for investigating the relationships between different column pairs, using comparison methods such as Linear regression, Spearman R, Anova, Kruskal-Wallis or Chi-square.

K-means Clustering: A tool that helps you group rows into a defined number of clusters based on their similarity. A line chart is needed in order to use the tool.

Line Similarity: A tool where you can compare the similarity of different lines in a line chart against each other.

Hierarchical Clustering: Hierarchical clustering arranges objects in a hierarchy with a treelike structure based on the similarity between them.

Hierarchy: A hierarchy is defined as A) a set of ordered columns where the order defines the hierarchy, B) a timestamp column where it is possible to derive a natural hierarchy or C) an external hierarchical structure where every node maps to one or more rows in the data table through an identifier column.

I am creating 15 Spotfire visualizations to understand what the data will allow one to conclude, if anything, about modeling the length of hospital stay as a function of the variables in the Health Prize data set. So far I do not see how the data set will provide meaningful insight into the relationships beyond just providing a statistical model with considerable variability resulting in limited confidence in the results.
Note

I emailed HPN/Kaggle at https://www.heritagehealthprize.com/contact: I am a Data Scientist/Data Blogger for AOL Government and am doing a story (s) on this prize. I am more interested in the properties and visualization of the data sets than the statistical models. Please see http://semanticommunity.info/HPN_Health_Prize

http://semanticommunity.info/AOL_Gov...Space_Launched

http://semanticommunity.info/AOL_Gov...ion_Investment

Thank you,

Dr. Brand Niemann
Director and Senior Data Scientist
Semantic Community
http://semanticommunity.info
http://gov.aol.com/bloggers/brand-niemann/
703-268-9314

Spotfire Dashboard

For Internet Explorer Users and Those Wanting Full Screen Display Use: Web Player Get Spotfire for iPad App

Media, iframe, embed and object tags are not supported inside of a PDF.

Spotfire Visualizations

What is a Table?

The table in TIBCO Spotfire works much like any other table you might be familiar with. It presents the data as a table of rows and columns, and is used to see details and compare values.

By clicking on a row you mark it, and by dragging the mouse pointer over several rows you can mark more than one row.

You can sort the rows in the table according to different columns by clicking on the column headers, or filter out unwanted rows by using the filters.

All visualizations except the map chart can be set up to show data limited by one or more markings in other visualizations only (details visualizations). They can also be limited by one or more filterings. Another alternative is to set up a visualization without any filtering at all.
What is a Cross Table?

A cross table is a two-way table consisting of columns and rows. It is also known as a pivot table or a multi-dimensional table. Its greatest strength is its ability to structure, summarize and display large amounts of data. Cross tables can also be used to determine whether there is a relation between the row variable and the column variable or not.

What is a Graphical Table?

A graphical table is a summarizing visualization designed to provide a lot of information at one glance. It can be set up to show columns with dynamic items such as sparklines, calculated values or conditional icons. One value is shown for each row as specified on the Rows axis.
What is a Bar Chart?

A bar chart is a way of summarizing a set of categorical data (continuous data can be made categorical by auto-binning). The bar chart displays data using a number of bars, each representing a particular category. The height of each bar is proportional to a specific aggregation (for example the sum of the values in the category it represents). The categories could be something like an age group or a geographical location. It is also possible to color or split each bar into another categorical column in the data, which enables you to see the contribution from different categories to each bar or group of bars in the bar chart.

What is a Line Chart?

Line charts are ideal for showing trends over time. A standard example would be how the stock value for a certain company develops over time on the stock market. However, it does not necessarily need to be time along the X-axis. Any data that behaves like a function with respect to the variable on the X-axis can be plotted. Line charts emphasize time flow and rate of change rather than the amount of change.
What is a Combination Chart?

The combination chart is a visualization that combines the features of the bar chart and the line chart. The combination chart displays the data using a number of bars and/or lines, each of which represent a particular category. A combination of bars and lines in the same visualization can be useful when comparing values in different categories, since the combination gives a clear view of which category is higher or lower. An example of this can be seen when using the combination chart to compare the projected sales with the actual sales for different time periods.

What is a Pie Chart?

Pie charts are circle graphs divided into sectors, each pie sector displaying the size of some related piece of information. Pie charts are used to show the relative sizes of the parts of a whole.
What is a Scatter Plot?

Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another. Each row in the data table is represented by a marker whose position depends on its values in the columns set on the X and Y axes.

A third variable can be set to correspond to the color or size of the markers, thus adding yet another dimension to the plot.

The relationship between two variables is called their correlation. If the markers are close to making a straight line in the scatter plot, the two variables have a high correlation. If the markers are equally distributed in the scatter plot, the correlation is low, or zero. However, even though a correlation may seem to be present, this might not always be the case. Both variables could be related to some third variable, thus explaining their variation, or, pure coincidence might cause an apparent correlation.
What is a 3D Scatter Plot?

3D scatter plots are used to plot data points on three axes in the attempt to show the relationship between three variables. Each row in the data table is represented by a marker whose position depends on its values in the columns set on the X, Y, and Z axes.

A fourth variable can be set to correspond to the color or size of the markers, thus adding yet another dimension to the plot.

The relationship between different variables is called correlation. If the markers are close to making a straight line in any direction in the three-dimensional space of the 3D scatter plot, the correlation between the corresponding variables is high. If the markers are equally distributed in the 3D scatter plot, the correlation is low, or zero. However, even though a correlation may seem to be present, this might not always be the case. The variables could be related to some fourth variable, thus explaining their variation, or pure coincidence might cause an apparent correlation.

You can change how the 3D scatter plot is viewed by zooming in and out as well as rotating it by using the navigation controls located in the top right part of the visualization.

Note: The 3D scatter plot is not supported in TIBCO Spotfire Web Player. It is still possible to open an analysis with a 3D scatter plot in the web player, but the 3D scatter plot will not be shown.

What is a Map Chart?

There are three kinds of map charts in TIBCO Spotfire: maps with interactive shapes, maps with markers or pies, and image backgrounds with markers or pies.

Below is an example of a map chart with interactive shapes, where each shape represents a state in the United States. Each shape in the map is a separate item, and you can interact with those items the same way you do with items in any other visualization.
The interactive shapes can be one of three geometry types: polygons, lines, or points. When polygons are used, as in the example above, the shapes constitute different areas in the map, and these areas will be filled with color. How to color the shapes is defined in the Colors page of the visualization properties, or in the legend. When lines or points are used, the interactive shapes are the actual lines or points. The color you define in the Colors page will be the color of the lines or points. Examples of when maps with lines as interactive shapes could be useful are maps showing highways or a street grid. Below is an example of a map chart with interactive shapes, where each shape represents a highway.

Which geometry type is used in a map is defined in the map data before you load it into your analysis, and this cannot be changed in Spotfire. If the map data comes from an ESRI shape file, you can open the Properties tab of the Data Table Properties dialog to see which geometry type a map has if you are uncertain.

In a map with markers or pies, the map areas are not interactive. Instead, markers or pies are positioned in the different areas. In the example below, the map shows the same geographical area as in the first example, and is also divided into states. But instead of the states being interactive, a marker is placed in each of the states, and you can interact with the markers just as you do with markers in other visualizations.
To set up a map with interactive shapes, or a map with markers or pies, you need a data table containing map data, such as from an ESRI shape file. You can use other types of sources than shape files, but that requires some manual configuration. To learn how to use a non-shape file, or to set up information links with map information, see Configuration of Geographical Data for Map Charts.

If you want the shapes in a map with interactive shapes to be colored by a data table other than the map data table, the two data tables must contain columns that can be matched to each other; State or City, for example. When you have set up a relation between the two data tables using these columns, you can color the shapes by a column in the other data table.

Similarly, if you want the markers in a map with markers or pies to be colored by another data table than the map data table, you must set up a relation between the two data tables. Or, if the data table for markers or pies has columns containing coordinates, you can use these to position the markers or pies in their correct locations on the map.

A third way to set up a map chart is to use a background image and then position markers or pies on top of that image. This works similarly to the map with markers or pies, but with the difference that you do not need to have map data in a data table in order to set it up. However, for the markers to be placed correctly in geographical positions, the data table must contain X and Y coordinates. Below is an example of a map chart where the background is a map image of a part of North America. On top of the background image are markers pointing out cities in the United States.

You can zoom and pan in a map using the navigation controls to the right of the map. Click on the small arrow icon on the map chart title bar (shown on mouse over) to show or hide the navigation controls. The example below shows the same map chart as in the first example, but it has been zoomed in to show only some of the states. To learn more, see Zooming and navigating in the map chart.
Labels can be used in the map chart to identify and describe markers or interactive shapes. In the example above, labels with the state names have been added to the map. Open the Labels page of the Map Chart Properties if you want to modify the labels settings. You can also move the labels to other positions in the map using drag and drop. To learn more about labels in different map types, see How to Use the Map Chart.

A map chart can be used to show other than geographical data. The example below displays different types of failures on a wafer, a semi-conductor material used to manufacture microchips.

The background is an image representing the wafer. The markers in the visualization represent the chips on the wafer, and are placed on the background the same way they are placed on the actual wafer. The colors and labels indicate the six different types of manufacturing failures that have occurred on this wafer. Copying the actual layout of the wafer is a way to enhance the readability of the data. To be able to view the data this way, you need to use tiled markers. This means that all the markers have the same size, and are displayed in a grid-like layout. Go to the Shape page in the Map Chart Properties to change to tiled markers.

Note: If you use tiled markers, and the axes’ scales have a large number of values, then the markers may become too small to be seen. The reason for this is that the grid layout makes it necessary for each value on the scales to have a unique position, even if no marker is located at each of these allocated positions. Therefore, with a large number of values on the scale, the markers must become very small to fit in the grid.

All visualizations except the map chart can be set up to show data limited by one or more markings in other visualizations only (details visualizations). They can also be limited by one or more filterings. Another alternative is to set up a visualization without any filtering at all.
What is a Treemap?

Treemaps are ideal for displaying large amounts of hierarchically structured (tree-structured) data. The space in the visualization is split up into rectangles that are sized and ordered by a quantitative variable.

The levels in the hierarchy of the treemap are visualized as rectangles containing other rectangles. Each set of rectangles on the same level in the hierarchy represents a column or an expression in a data table. Each individual rectangle on a level in the hierarchy represents a category in a column. For example, a rectangle representing a continent may contain several rectangles representing countries in that continent. Each rectangle representing a country may in turn contain rectangles representing cities in these countries. You can create a treemap hierarchy directly in the visualization, or use an already defined hierarchy. To learn more, see the section To Create a Treemap Hierarchy.

A number of different algorithms can be used to determine how the rectangles in a treemap should be sized and ordered. The treemap in Spotfire uses a squarified algorithm.

The rectangles in the treemap range in size from the top left corner of the visualization to the bottom right corner, with the largest rectangle positioned in the top left corner and the smallest rectangle in the bottom right corner. For hierarchies, that is, when the rectangles are nested, the same ordering of the rectangles is repeated for each rectangle in the treemap. This means that the size, and thereby also position, of a rectangle that contains other rectangles is decided by the sum of the areas of the contained rectangles.

What is a Heat Map?

The easiest way to understand a heat map is to think of a table or spreadsheet which contains colors instead of numbers. The default color gradient sets the lowest value in the heat map to dark blue, the highest value to a bright red, and mid-range values to light gray, with a corresponding transition (or gradient) between these extremes. Heat maps are well-suited for visualizing large amounts of multi-dimensional data and can be used to identify clusters of rows with similar values, as these are displayed as areas of similar color.
What is a Parallel Coordinate Plot?

A parallel coordinate plot maps each row in the data table as a line, or profile. Each attribute of a row is represented by a point on the line. This makes parallel coordinate plots similar in appearance to line charts, but the way data is translated into a plot is substantially different.

What is a Summary Table?

The summary table is a visualization that summarizes statistical information about data in table form. The information is based on one data table in TIBCO Spotfire. You can, at any time, choose which measures you want to see (such as mean, median, etc.), as well as the columns on which to base these measures. As you change the set of filtered rows, the Summary Table automatically updates the values displayed to reflect the current selection.
What is a Box Plot?

Box plots are graphical tools to visualize key statistical measures, such as median, mean and quartiles.

A single box plot can be used to represent all the data. It is also possible to visualize separate statistics for subsets by selecting a column for the X-axis.

The individual box plot is a visual aid to examining key statistical properties of a variable. The diagram below shows how the shape of a box plot encodes these properties. The range of the vertical scale is from the minimum to the maximum value of the selected column, or, to the highest or lowest of the displayed reference points.

HPN Health Prize

Source: https://www.heritagehealthprize.com/c/hhp
Identify patients who will be admitted to a hospital within the next year, using historical claims data.

Description

Source: [https://www.heritagehealthprize.com/c/hhp](https://www.heritagehealthprize.com/c/hhp)

More than 71 million individuals in the United States are admitted to hospitals each year, according to the latest survey from the American Hospital Association. Studies have concluded that in 2006 well over $30 billion was spent on unnecessary hospital admissions. Is there a better way? Can we identify earlier those most at risk and ensure they get the treatment they need? The Heritage Provider Network (HPN) believes that the answer is "yes".

To achieve its **goal of developing a breakthrough algorithm that uses available patient data to predict and prevent unnecessary hospitalizations**, HPN is sponsoring the Heritage Health Prize Competition (the "Competition"). HPN believes that incentivized competition is the best way to achieve the radical breakthroughs necessary to begin fixing America's health care system.

The winning team will create an algorithm that **predicts how many days a patient will spend in a hospital in the next year**. Once known, health care providers can develop new care plans and strategies to reach patients before emergencies occur, thereby reducing the number of unnecessary hospitalizations. This will result in increasing the health of patients while decreasing the cost of care. In short, a winning solution will change health care delivery as we know it – from an emphasis on caring for the individual after they get sick to a true health care system.

The Competition runs for two years and offers a US $3 million Grand Prize, as well as six Milestone Prizes totaling $230,000, which are awarded in varying amounts at three designated intervals during the Competition.

This competition has ...

- players
- entries

Started: **5:03 pm, Monday 4 April 2011 UTC**
Ends: **6:59 am, Wednesday 3 April 2013 UTC (729 total days)**

Visit the forum
Download the data
View the leaderboard
Enter competition
Evaluation

Source: https://www.heritagehealthprize.com/c/hhp/details/Evaluation

IMPORTANT NOTE: The information provided below is intended only to provide general guidance to participants in the Heritage Health Prize Competition and is subject to the Competition Official Rules (See Below). Any capitalized term not defined below is defined in the Competition Official Rules. Please consult the Competition Official Rules for complete details.

Entries will be judged based on the degree of accuracy of their predictions of DaysInHospital for Y4 (or if applicable Y5), carried to six (6) decimal places. An entry's predictive accuracy will be judged by comparing (i) the predicted number of days a member will spend in the hospital reflected in the entry and (ii) the actual number of days a member spent in the hospital as reflected in the Scoring Data Set. Prediction accuracy will be evaluated based on the following metric:

\[
\varepsilon = \left( \frac{1}{n} \sum_{i}^{n} \left[ \log(p_i + 1) - \log(a_i + 1) \right]^2 \right)^{1/2}
\]

Where:
1. i is a member;
2. n is the total number of members;
3. p is the predicted number of days spent in hospital for member i in the test period;
4. a is the actual number of days spent in hospital for member i in the test period.

Note: the metric is calculated using the natural log.

Entrants may submit one (1) entry during each calendar day (UTC) beginning on May 4, 2011 and ending at 06:59:59 UTC on April 3, 2013. Eligible Milestone Prize Entries judged to have the two lowest prediction scores as of the applicable Milestone Prize Deadline will win Milestone Prizes. The eligible Grand Prize Entry that the judges determine produces the lowest prediction score that passes the Accuracy Threshold as of the end of the Competition (06:59:59 UTC on April 4, 2013) will win the Grand Prize of US $3 million. The Accuracy Threshold is the accuracy of predictions required to win the Grand Prize. That maximum accuracy is 0.4.

Rules

Source: https://www.heritagehealthprize.com/c/hhp/details/Rules

Please read through these carefully and scroll to the bottom to accept

Dated: May 4, 2011

INTRODUCTION
More than 71 million individuals in the United States visit a hospital each year, according to the latest survey from the American Hospital Association. Studies have concluded that in 2006 well over $30 billion was spent on unnecessary hospital admissions. Is there a better way? Can we identify earlier those most at risk and ensure they get the treatment they need? The Heritage Provider Network (HPN) believes that the answer is "yes".

To achieve its goal of developing a breakthrough algorithm that uses available patient data to predict and prevent unnecessary hospitalizations, HPN is sponsoring the Heritage Health Prize Competition (the "Competition"). HPN believes that incentivized competition is the best way to achieve the radical breakthroughs necessary to begin fixing America's health care system.

The winning team will create an algorithm that predicts how many days a patient will spend in the hospital in the next year. Once known, health care providers can develop new care plans and strategies to reach patients before emergencies occur, thereby reducing the number of unnecessary hospitalizations. This will result in increasing the health of patients while decreasing the cost of care. In short, a winning solution will change health care delivery as we know it from an emphasis on caring for the individual after they get sick to a true health care system.

The Competition runs for two years and offers a US $3 million Grand Prize, as well as six Milestone Prizes totaling $230,000, which are awarded in varying amounts at three designated intervals during the Competition. If you would like to participate, please read the following Official Rules for complete details and entry instructions. By entering the Competition, you acknowledge that you have read, understand and agree to be bound by these Official Rules.

A. HOW TO ENTER

1. OVERVIEW

The Competition begins on April 4, 2011 at 19:00 UTC (or such later times as posted on the Website) and ends at 06:59:59 UTC on April 4, 2013 (the "Competition Period"). If you meet the eligibility requirements and would like to participate, then you must first complete the registration process at http://www.heritagehealthprize.com (the "Website") or http://www.kaggle.com by 06:59:59 UTC on October 4, 2012. After you complete the registration process, you will receive access to the Data Sets (described in Rule 5 below) that will enable you to develop and submit one or more Entries (as defined in Rule 8 below). All Entries must be received during the Competition Period.

2. ELIGIBILITY

The Competition is open to individuals who are the age of majority or older in their places of residence as of time of registration. You are not eligible to participate in the Competition if you are a resident of Cuba, Sudan, Iran, North Korea, Syria or any other country designated by the United States Treasury's Office of Foreign Assets Control (see http://www.treasury.gov/resource-cen...s/default.aspx for additional information). Officers, directors, employees and advisory board members (and their immediate families and members of the same household) of Heritage Provider Network, Inc. ("Sponsor"), Kaggle Pty Ltd, Children's Hospital of Eastern Ontario Research Institute, Inc., University of Ottawa, University of Maryland, Baltimore County and Privacy Analytics Inc. and their respective affiliates, agents, judges and advertising and promotion agencies (collectively, the "Competition Entities") also are not eligible to participate in the Competition.

3. REGISTRATION
To register, visit the Website and follow the onscreen instructions to complete and submit your registration. You may register at any time prior to 06:59:59 UTC on October 4, 2012.

To register, you must provide your full legal name, a display name (which may be the same as or different from your legal name), a username (to log in to the Website), your email address and a password. Only your display name will be visible on the Website. You may (but are not required to) provide additional information about yourself as part of the registration process, all of which information will be visible on the Website. All of the registration information that you provide is collectively referred to as your "Account". (If you have already created an Account at www.kaggle.com, enter your user name and password and follow the on-screen instructions. If you have previously registered at http://www.kaggle.com but have not yet provided your full legal name, you will be required to do so to complete the registration process.)

By registering, you agree that (a) your Account is complete, correct and accurate and (b) your registration may be rejected or terminated and all Entries submitted by you and/or your Team may be disqualified if any of the information in your Account is (or Sponsor has reasonable grounds to believe is) incomplete, incorrect or inaccurate. You are solely responsible for your Account. All registration information is deemed collected in the United States.

After you register individually, you may join a team (a "Team"), as described in Rule 4 below, but you may register only once. If you register for the Competition more than once, you will be, and the remainder of your Team may be, disqualified in Sponsor's sole discretion.

You acknowledge and agree that you are solely responsible for abiding by your employer's policies regarding participation in the Competition. Sponsor disclaims any and all liability or responsibility for disputes arising between an employee and employer related to this Competition.

Once you have completed the registration process, you or your Team will be provided with access to the Data Sets that you use to develop your Entries.

BY REGISTERING, YOU ACCEPT THE CONDITIONS STATED IN THESE OFFICIAL RULES AND THE WEBSITE TERMS OF USE (see http://www.heritagehealthprize.com/pages/Terms), AGREE TO BE BOUND BY THE DECISIONS OF THE JUDGES AND WARRANT THAT YOU ARE ELIGIBLE TO PARTICIPATE IN THE COMPETITION. IF YOU DO NOT ACCEPT ALL OF THESE OFFICIAL RULES AND THE WEBSITE TERMS OF USE, THEN PLEASE DO NOT REGISTER FOR THE COMPETITION. WE RECOMMEND THAT YOU PRINT OUT A COPY OF THESE OFFICIAL RULES FOR YOUR FUTURE REFERENCE.

4. TEAMS

You may participate in the Competition individually or as a member of a Team. A Team may be comprised of up to eight (8) members, each of whom must register individually. Multiple Teams from the same institution/organization are allowed, provided that no individual serves on more than one Team. You may not participate in the Competition individually if you are a member of a Team.

One Team member must serve as the Team leader. Initially, the Team leader will be the Team member who first completes the "Team Wizard" on the Website. You will not be considered a member of a Team unless and until you confirm your Team membership by responding to the Team notification message available through your Account.
To change the Team leader, please use the Website's "Contact Us" form.

In Sponsor's discretion, Teams may be permitted to merge. Team merger requests must be made via the Website's "Contact Us" form no later than 06:59:59 UTC on October 4, 2012. Each registered individual or Team is referred to as an "Entrant" in these Official Rules.

B. DATA

Entrants must use only the data described in Rules 5 - 7 below to develop their Entries. In using such data, Entrants must observe customary and prudent medical data security and privacy practices (including but not limited to data confidentiality, storage and encryption practices with respect to any and all data used in connection with the Competition. Sponsor reserves the right to disqualify Entries and/or prohibit use of other data (as described in Rule 7) by Entrants that it determines, in its discretion, do not comply with such security and privacy practices.

5. DATA SETS

Sponsor will provide Entrants with certain deidentified member data collected during a thirty-six month period. Sponsor reserves the right to modify or add data upon notice to Entrants through their Accounts, via email using the email address associated with their Accounts and/or by such other reasonable means as Sponsor may determine. The member data will be allocated among three data sets (collectively, the "Data Sets"). Entrants must use the Data Sets provided to them solely for purposes of the Competition, including but not limited to preparing their Entries, developing and testing their Prediction Algorithms (as defined below) for accurately predicting the number of days that the members will spend in a hospital (inpatient or emergency room visit) during the 12-month period following the Data Set cut-off date and participating in the forum discussions on the Website. A "Prediction Algorithm" is the algorithm used to produce the data in an Entry taken as a whole (i.e., its particular total configuration) but does not include individual components of the Prediction Algorithm or tools used for analysis or development of the Prediction Algorithm. Except as otherwise provided in Rule 21, promptly after the Competition ends Entrants must cease all use of the Data Sets provided by Sponsor and erase or otherwise destroy all waste paper, fax paper, photocopy paper and other physical media and computer entries or other electronic media that contain the Data Sets.

The Data Sets are:

- the "Training and Validation Data Set", which is to be used by Entrants to develop the algorithms that generate their Entries and evaluate the efficacy of their algorithms;
- the "Feedback Data Set" which will be used to calculate standings on the Leaderboard (described in Rule 11 below); and
- the "Scoring Data Set" which will be used to determine the winners of the Milestone Prizes and Grand Prize.

6. TABLES

Each of the Data Sets will be comprised of tables as follows:

- a. Members Table, which will include:
  - i. MemberID (a unique member ID)
  - ii. AgeAtFirstClaim (member's age when first claim was made in the Data Set period)
  - iii. Sex
b. Claims Table, which will contain certain details of claims related to members.

c. Labs Table, which will contain certain details of lab tests provided to members. The Labs Table will be made available to Entrants on the Website on or about June 4, 2011.

d. RX Table, which will contain certain details of prescriptions filled by members. The RX Table will be made available to Entrants on the Website on or about June 4, 2011.

e. DaysInHospital Tables - Y2 and Y3, which will contain the number of days of hospitalization for each eligible member during Y2 and Y3 and will include:
   - i. MemberID; and
   - ii. DaysInHospital (the number of days in hospital Y2 or Y3, as applicable). These two Tables are intended for use by Entrants to train and validate their Prediction Algorithms.

f. DaysInHospitalTable - Y4, which will contain only MemberID. DaysInHospital data for Y4 are the missing values in the file labeled "Target.csv" which are to be filled in by Entrants to produce Entries.

The Data Sets will be made available to Entrants on the Website as follows:

- April 4, 2011: Claims Table - Y1 and DaysInHospital Table - Y2
- May 4, 2011: All other Data Sets except Labs Table and Rx Table
- June 4, 2011: Labs Table and Rx Table

7. USE OF OTHER DATA

Entrants may use data other than the Data Sets to develop and test their Prediction Algorithms and Entries provided that (i) such data are freely available to all other Entrants and (ii) the data and/or a link to the data are published in the "External Data" topic in the Forums section of the Website within one (1) week of the date on which an Entry that uses such data is submitted to the Website. Entrants may not use new external data in connection with the development of their Entries after 11:59:59 UTC on April 4, 2012 without the prior written permission of Sponsor. Any third-party service provider, consultant or contractor of Sponsor that received or receives data or other information in connection with work performed for or on behalf of Sponsor may not use such data or other information in connection with the Competition.

You may not, however, link the Data Sets to records in other external databases such that new demographic, socioeconomic or clinical information about the members in the Data Sets is gained. Sponsor reserves the right in its sole discretion to disqualify any Entrant who Sponsor discovers has undertaken or attempted to undertake such linking of the Data Sets.

C. ENTRIES

8. ENTRY SUBMISSIONS

Each Entry must be uploaded to the Website in the manner and format specified on the Website. All Entries must be received during the Competition Period, no earlier than 19:00:00 UTC on May 4, 2011. An “Entry” is the data submitted in the manner and format specified on the Website via the Website on Entry form.

Entrants may submit one (1) Entry during each calendar day (UTC) of the Competition Period, beginning on May 4, 2011. Team Entries must be submitted by the Team leader.
Sponsor reserves the right to request that an Entrant submit the Prediction Algorithm associated with an Entry to Sponsor.

9. MILESTONE PRIZE ENTRIES

Each Entrant may (but is not required to) designate one of his/her/its Entries for consideration for each Milestone Prize Round (each, a "Milestone Prize Entry") by following the designation instructions on the Website. All designations for Milestone Prizes Entries must be made by 06:59:59 (UTC) on the following dates ("Milestone Prize Deadlines"):

- **Round 1 Milestone Prizes** 06:59:59 UTC on August 31, 2011
- **Round 2 Milestone Prizes** 06:59:59 UTC on February 13, 2012
- **Round 3 Milestone Prizes** 06:59:59 UTC on September 4, 2012

NOTE: Any Milestone Prize Deadline may be adjusted to a date later than set forth above upon written notice to all Entrants through their Accounts, via email using the email address associated with their Accounts and/or by such other reasonable means as Sponsor may determine. Any revised Milestone Prize Deadline also will be posted on the Website.

A Team's designation of a Milestone Prize Entry must be made by the Team leader. If an Entrant does not designate one Milestone Prize Entry by the applicable Milestone Prize Deadline, his/her/its Entry with the lowest prediction score on the Leaderboard will be automatically designated for judging. Designation of a Milestone Prize Entry may be changed at any time prior to the applicable Milestone Prize Deadline.

10. GRAND PRIZE ENTRIES

Each Entrant may (but is not required to) designate up to five of his/her/its Entries for consideration for the Grand Prize (each, a "Grand Prize Entry") by following the designation instructions on the Website. All Grand Prize Entry designations must be made by 06:59:59 UTC on April 4, 2013 ("Grand Prize Deadline"). A Team's designation of a Grand Prize Entry must be made by the Team leader. Designation of a Grand Prize Entry may be changed at any time prior to the Grand Prize Deadline. If an Entrant does not designate five Grand Prize Entries by the Grand Prize Deadline, his/her/its five Entries with the lowest prediction score on the Leaderboard will be automatically designated for judging.

11. LEADERBOARD

A public leader board ("Leaderboard") will be displayed on the Website throughout the Competition beginning on May 4, 2011. The Leaderboard scores will be determined using the Feedback Data Set and are for informational purposes only and will not be used to determine prize winners, except as described in Rule 10 above. Sponsor reserves the right to disqualify any Entrant who attempts (or who Sponsor suspects of attempting) to reverse engineer, "crack", "hack", manipulate or gain unauthorized access to the Leaderboard or Data Sets.

D. JUDGING

12. JUDGING PROCESS
Judging will be conducted by a panel of independent qualified judges. The decisions of the judges are final in all matters related to the Competition.

Entries will be judged based on the degree of accuracy of their Prediction Algorithm’s predictions of DaysInHospital for Y4 (or if applicable Y5), carried to six (6) decimal places. The predictive accuracy of the Prediction Algorithm used to produce an Entry will be judged by comparing (i) the predicted number of days a member will spend in the hospital reflected in the Entry and (ii) the actual number of days a member spent in the hospital as reflected in the Scoring Data Set. Prediction accuracy will be evaluated by determining the Entry’s error percentage based on the following metric (the "Prediction Error Rate"):

\[
\varepsilon = \sqrt{\frac{1}{n} \sum_{i}^{n} \left[ \log(p_i + 1) - \log(a_i + 1) \right]^2}
\]

Where:
1. \( i \) is a member;
2. \( n \) is the total number of members;
3. \( p \) is the predicted number of days spent in hospital for member \( i \) in the test period;
4. \( a \) is the actual number of days spent in hospital for member \( i \) in the test period.
5. \( \log \) is the natural logarithm function.

Sponsor selected this metric because it favors algorithms that accurately predict members who have fewer days of hospitalization based on the assumption that those members who spend fewer days in a hospital are more likely to have hospitalizations that are preventable.

Once an Entry is selected as eligible for a prize, the conditional winner must deliver the Prediction Algorithm’s code and documentation to Sponsor for verification within 21 days. Documentation must be written in English and must be written so that individuals trained in computer science can replicate the winning results. Source code must contain a description of resources required to build and run the method. Conditional winners must be available to provide assistance to the judges verifying their Entries. Sponsor may require conditional winners to submit computer hardware or a virtual machine instance that runs the Prediction Algorithm’s code. If the judges cannot verify an Entry using the Prediction Algorithm after two attempts, Sponsor reserves the right to disqualify the Entry. Sponsor also reserves the right to test winning Entries on additional data sets. If a Prediction Algorithm fails to produce similar accuracy on any such additional data set, Sponsor may in its sole discretion disqualify the Entry.

13. MILESTONE PRIZE ENTRIES

Milestone Prizes will be conditionally awarded to the eligible Milestone Prize Entries with the two lowest prediction scores as of the applicable Milestone Prize Deadline determined as described in Rule 12.

Selection of conditional Milestones Prize winners by the judges will occur on or about the following dates:

- Round 1 Milestone Prizes on or about September 1, 2011
Round 2 Milestone Prizes on or about February 14, 2012

Round 3 Milestone Prizes on or about September 3, 2012

NOTE: The dates on which judging of Milestone Prizes will occur may be adjusted to dates later than the dates set forth above upon written notice to all Entrants through their Accounts, via email using the email address associated with their Accounts and/or by such other reasonable means as Sponsor may determine. The revised Milestone Prize judging date also will be posted on the Website.

The judges will identify the two conditional winning Milestone Prize Entries for each Round and Sponsor will notify the conditional winners. Once Sponsor has completed notification of conditional winners:

1. conditional winners will have 21 days from receipt of notification to document their methodology as described in Rule 12 above. Sponsor will deliver the Prediction Algorithm and documentation to the judges and also post the information on the Website for review and testing by other Entrants.;
2. other Entrants will have the opportunity to submit comments/complaints relating to conditional winners' methodologies for 30 days after such conditional winners' methodologies are published on the Website.;
3. if the judges determine in their sole discretion that any complaints are valid, conditional winners will have seven (7) days to address them and resubmit their methodologies.; and
4. following the 30-day review period (and any subsequent resubmissions), the judges will award the Milestone Prizes.

Milestone Prize winners will be announced as follows:

Round 1 Milestone Prizes
Conditional winners will be announced at the O'Reilly's Strata Conference currently scheduled to begin on September 19, 2011 in New York, New York.

Round 2 Milestone Prizes
Conditional winners will be announced at the O'Reilly's Strata Conference currently scheduled to begin on February 28, 2012 in Santa Clara, California.

Round 3 Milestone Prizes
Conditional winners will be announced at the O'Reilly's Strata Conference scheduled to be held during Q3 2012 on a date and in a location to be announced by the Conference organizers.

If for any reason the conditional Milestone Prize winners cannot be announced as set forth above due to cancellation or postponement of the applicable Strata Conference or otherwise, then the applicable conditional Milestone Prize winners will be announced within 30 days of the scheduled start date of the applicable Conference described above. The schedule for the announcement of the Round 3 conditional Milestone Prizes winners will be posted on the Website no later than September 4, 2012.

If a conditional Milestone First Prize winner wishes to attend the applicable Strata Conference, Sponsor will provide coach-class air transportation from a major commercial airport near conditional winner's residence within the continental United States to a major commercial airport near the Strata Conference (airports/air carrier(s) determined by Sponsor), two-nights' accommodations (hotel determined by Sponsor) and conference registration fee. If the conditional winner resides within 250 miles of the conference location, Sponsor reserves the right to provide ground transportation in lieu of
air transportation. All other expenses (including, but not limited to taxes, fees or meals, gratuities, insurance, personal and hotel incidentals, transfers and ground transportation) are responsibility solely of conditional winner. If a conference is postponed or cancelled for any reason, Sponsor has no obligation to provide substitute compensation. Conditional winner must travel on the itinerary specified by Sponsor. Travel is subject to availability, blackout dates and other restrictions. Sponsor is not responsible for any cancellation, delay, diversion or substitution, or any act or omission whatsoever by the air carriers or other transportation companies or any other person or entity providing any of these services, including but not limited to any change in services or accommodations necessitated by same. Additional restrictions may apply. If conditional winner is a Team, then Team members must designate in writing one (1) Team member to receive the travel/accommodations described above.

Winners of Milestone Prizes are eligible to win subsequent Milestone Prizes, as well as the Grand Prize.

14. GRAND PRIZE ENTRIES

The Grand Prize will be conditionally awarded to the Entrant whose eligible Grand Prize Entry, as of 06:59:59 UTC on April 4, 2013, produces the lowest Prediction Error Rate determined as described in Rule 12 above. The winning Grand Prize Entry must have a Prediction Error Rate that is at least as low as the threshold Prediction Error Rate to be published on the Website on May 4, 2011 ("Prediction Error Threshold").

The judges will identify the conditional winning Grand Prize Entry and Sponsor will notify the conditional Grand Prize winner. Once Sponsor has completed notification of the conditional Grand Prize winner:

1. conditional Grand Prize winner will have 30 days from receipt of notification to document his/her/its methodology as described in Rule 12 above. Sponsor will deliver the Prediction Algorithm’s code and documentation to the judges;
2. the judges will review the Prediction Algorithm’s code and documentation;
3. if the judges determine, in their sole discretion, that there are any inconsistencies or problems in the Entrant's Entry and the Prediction Algorithm’s code and/or documentation, the conditional Grand Prize winner will have seven (7) days from receipt of notification of such determination to address them and resubmit his/her/its methodology; and
4. the judges will then further review the Entrant's submissions (including any additional resubmissions) and if they determine that all inconsistencies or problems have been properly addressed or resolved the judges will award the Grand Prize to the conditional Grand Prize winner.

If the judges determine, in their sole discretion, that an Entry’s issues are not adequately addressed, then steps 1. through 4. above will be repeated with the Entry that produces the next lowest Prediction Error Rate. This process will continue until the Grand Prize is awarded or there are no further qualifying Entries.

15. CONSOLATION PRIZE

If after judging is completed, no Entrant's Grand Prize Entry produces a Prediction Error Rate that is lower than the Prediction Error Threshold, Sponsor will award a Consolation Prize of US $500,000 to the Entrant that submitted the Entry with the best prediction score determined as described in Rule 12 above. The Consolation Prize will be awarded on or about June 1, 2013.

16. TIES
If any two or more eligible and verified Entries produce the same lowest prediction score as determined by the judges, then the Entry submitted first (as measured by the time on the Website’s servers) will be deemed the winner of the applicable prize.

E. PRIZES

17. DESCRIPTION OF PRIZES

One (1) Grand Prize: US $3,000,000 (awarded as a check made payable to Entrant).

Six (6) Milestone Prizes (each awarded as a check made payable to Entrant):

- **Round 1 Milestone Prizes**
  - First Prize: US $30,000
  - Second Prize: US $20,000

- **Round 2 Milestone Prizes**
  - First Prize: US $50,000
  - Second Prize: US $30,000

- **Round 3 Milestone Prizes**
  - First Prize: US $60,000
  - Second Prize: US $40,000

One (1) Consolation Prize: US $500,000 (awarded as a check made payable to Entrant).

Total ARV of All Prizes: US $3,230,000. (Total ARV if Consolation Prize is awarded in lieu of Grand Prize: $730,000.)

Sponsor is responsible only for prize delivery and is not responsible for prize utility or otherwise. No substitution or transfer of prizes is permitted. All taxes, fees and expenses associated with participation in the Competition or receipt and use of a prize are the sole responsibility of winners.

18. PRIZE AWARD

As a condition of receipt of a prize, winner (and if winner is a Team, each individual Entrant on the winning Team) will be required to complete, sign and return the following documents: (i) an agreement provided by Sponsor evidencing the License described in Rule 21 below, (ii) an Affidavit of Eligibility and Liability Release and where lawful, a Publicity Release and (iii) all applicable U.S. tax withholding-related documentation and, if necessary, provide proof of foreign status, as beneficiaries of U.S.-sourced funds. Any prize may be awarded to an alternate winner if all required documentation is not returned within thirty (30) days after mailing to winner, if prize notification letter/email or prize is returned as undeliverable or if winner does not respond to an email or other communication from Sponsor or judges within ten (10) days of the date sent. Allow 6-8 weeks for prize delivery. By accepting any prize, each winner agrees to use of his/her name, address, likeness and/or prize information by the Competition entities for promotional purposes in any medium without additional compensation to the extent permitted by law.

19. TEAM PRIZE WINNERS

If a Team wins a prize, all Team members must submit a single written statement describing how the prize is to be allocated among the Team members. If the Team fails to submit such statement within 30 days after Sponsor requests
it, then Sponsor will distribute the prize among Team members in equal shares and will have no further obligation to winning Team members.

F. INTELLECTUAL PROPERTY RIGHTS

20. ENTRANT REPRESENTATIONS

By submitting an Entry, each entrant represents that he/she/it has the unrestricted right to submit the Entry and that the Entry and Prediction Algorithm used to produce the Entry (i) were not previously published and have not won any other prize/award, (ii) are the exclusive original work of the entrant (and his/her/its Team, if applicable), true and verifiable, (iii) are not unlawful or plagiarized, as determined by Sponsor in its sole discretion, (iv) do not violate or encourage others to violate any applicable law, statute, ordinance or regulation, (v) do not infringe, misappropriate or violate any third party’s copyright, trademark, patent, literary, trade secret, privacy, publicity, proprietary, contractual or other right; and (vi) do not include any virus, worm, corrupt file, Trojan horse or other forms of corruptive code or content that may harm or compromise the Website and/or the proper conduct of the Competition.

21. LICENSE

By registering for the Competition, each Entrant (a) grants to Sponsor and its designees a worldwide, exclusive (except with respect to Entrant), sub-licensable (through multiple tiers), transferable, fully paid-up, royalty-free, perpetual, irrevocable right to use, not use, reproduce, distribute (through multiple tiers), create derivative works of, publicly perform, publicly display, digitally perform, make, have made, sell, offer for sale and import each Entry and the Prediction Algorithm used to produce the Entry (collectively, the "Licensed Materials"), in any media now known or hereafter developed, for any purpose whatsoever, commercial or otherwise, without further approval by or payment to Entrant (the "License") and (b) represents that he/she/it has the unrestricted (except for ‘Open Source’ restrictions in OSI-approved licenses listed at http://www.opensource.org/licenses/alphabetical or when Sponsor, in its sole discretion, expressly waives such requirement in writing) right to grant the License. Entrant understands and agrees that the License is exclusive except with respect to Entrant: Entrant may use the Licensed Materials solely for his/her/its own patient management and other internal business purposes but may not grant or otherwise transfer to any third party any rights to or interests in the Licensed Materials whatsoever.

G. GENERAL

THE DATA SETS HAVE BEEN DEIDENTIFIED IN ACCORDANCE WITH THE HIPAA PRIVACY RULE, AS WELL AS ACCEPTED HEALTH CARE INDUSTRY STANDARDS AND APPLICABLE LAW. EACH INDIVIDUAL ENTRANT WARRANTS THAT HE/SHE WILL NOT ATTEMPT THROUGH ANY MEANS TO DETERMINE THE IDENTITY OF THE INDIVIDUALS AND/OR HEALTH CARE PROVIDERS REFLECTED IN ANY DATA USED OR OBTAINED IN CONNECTION WITH THE COMPETITION OR TO USE OR SHARE THE DATA SETS FOR ANY PURPOSE OTHER THAN PARTICIPATION IN THIS COMPETITION IN ACCORDANCE WITH THESE OFFICIAL RULES. SPONSOR RESERVES THE RIGHT TO SEEK COMPENSATORY AND PUNITIVE DAMAGES AND OTHER LEGAL AND EQUITABLE RELIEF TO THE FULLEST EXTENT PERMITTED BY LAW FROM ANY INDIVIDUAL WHO USES OR ATTEMPTS TO USE THE DATA SETS IN ANY WAY INCONSISTENT WITH THESE OFFICIAL RULES OR UNDERMINES OR ATTEMPTS TO UNDERMINE THE LEGITIMATE OPERATION OF THIS COMPETITION.

22. RESEARCH
The Data Sets may not be used for any purpose other than participation in the Competition without Sponsor's prior written approval. If you wish to use the Data Sets for research purposes, please contact Sponsor via the Website's "Contact Us" form, including a reasonably detailed description of the proposed research. All such requests will be given careful consideration.

23. EQUIPMENT AND COSTS OF PARTICIPATION

Each Entrant is solely responsible for all equipment, including a computer and modem, necessary to establish a connection to the World Wide Web, access to the World Wide Web and any telephone, data, hosting or other service fees associated with such access, as well as all costs incurred by or behalf of the Entrant in participating in the Competition.

24. DELIVERY AND RECEIPT OF ENTRIES

Sponsor is not responsible for (a) late, lost, stolen, damaged, garbled, incomplete, incorrect or misdirected Entries or other communications, (b) errors, omissions, interruptions, deletions, defects, or delays in operations or transmission of information, in each case whether arising by way of technical or other failures or malfunctions of computer hardware, software, communications devices, or transmission lines, or (c) data corruption, theft, destruction, unauthorized access to or alteration of Entry materials, loss or otherwise. Sponsor is not responsible for electronic communications or emails which are undeliverable as a result of any form of active or passive filtering of any kind, or insufficient space in any email account to receive email messages. Sponsor disclaims any liability for damage to any computer system resulting from participation in, or accessing or downloading information in connection with, the Competition.

25. ADJUSTMENTS TO OFFICIAL RULES AND OTHER MATERIALS

Sponsor reserves the right to correct clerical or typographical errors in Competition materials. Sponsor reserves the right to amend these Official Rules during the Competition Period. Entrants will be notified of any amendment to the Official Rules via a notice posted on the Website and also may be notified via email using the email address provided at registration. By continuing participation in the Competition, an Entrant is deemed to have accepted any such amendment. If an Entrant does not wish to continue to participate in the Competition pursuant to the Official Rules, as amended, such Entrant may terminate his/her/its participation in the Competition by deleting his/her Account on the Website. All Entries remain the property of Sponsor. Entrants are urged to consult the Website regularly during the Competition Period.

26. DISQUALIFICATIONS

Sponsor reserves the right, at its sole discretion, to disqualify any Entrant (and all individual or Team Entries submitted by or on behalf of Entrant) from this Competition or any other competition conducted now or in the future by Sponsor or any of its affiliates if her/his fraud or misconduct affects the integrity of this Competition.

27. RELEASE OF COMPETITION ENTITIES

By participating in the Competition, each Entrant agrees to release, defend, indemnify and hold harmless the Competition Entities from any and against any and all liability, claims, loss or damages (including reasonable attorneys'
fees) arising from or in connection with his/her participation in the Competition or the awarding, receipt and/or use or misuse of a prize or participation in any Competition-related activity.

28. OFFICIAL LANGUAGE

The official language for the Competition is English. All communications with Sponsor and the judges must be in the English language. These Official Rules and all other documents produced by or on behalf of the Sponsor in connection with the Competition are available only in English.

29. GEOGRAPHIC LIMITS

The Competition is void in Cuba, Sudan, Iran, North Korea, Syria, and any other countries designated by Sponsor and/or the United States Treasury’s Office of Foreign Assets Control and where prohibited or restricted by law.

30. GOVERNING LAW

The Competition is subject to all applicable laws, rules and regulations. The Competition will be governed by the internal laws of the State of California (without reference to conflicts of laws principles) and, to the extent applicable, the United States of America. Any and all legal actions or proceedings arising in connection with the Competition must be instituted in a state or federal court in Los Angeles County, California. You agree to submit to the jurisdiction of, and agree that venue is proper in, such courts in any legal action or proceeding relating to the Competition.

31. INTEGRATION

These Official Rules and the Website’s Terms of Use (both as may be amended from time to time), along with official Competition communications from Sponsor, govern and apply to all activity arising out of and relating to the Competition and use of the Website. These Official Rules cannot be modified or superseded except by Sponsor, in its reasonable discretion.

32. WINNERS’ NAMES

Winners’ names will be available on the Website after selection and verification of winners.

QUESTIONS

Please see our FAQs on the Website or contact us through the Contact Us form.

Please remember:

- you can not sign up to Kaggle from multiple accounts and therefore you cannot submit from multiple accounts;
- privately sharing code or data is not permitted (sharing data or code is permissible if made available to all players, such as on the forums); and
- merging teams will be permitted at Kaggle’s discretion - requests will generally be rejected if the number of entries made by the merged team exceeds the number of submissions permissible at the date of the merger request.

Rules Acceptance
By clicking on the "I understand and accept" button below, you are indicating that you agree to be bound to the above rules.

---

**Data Files**

Source: [https://www.heritagehealthprize.com/c/hhp/data](https://www.heritagehealthprize.com/c/hhp/data)

<table>
<thead>
<tr>
<th>File Name</th>
<th>Available Formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHP_release1</td>
<td>.zip (7.28 mb) Superseded by HHP_release2.zip</td>
</tr>
<tr>
<td>HHP_release2</td>
<td>.zip (46.58 mb) Superseded by HHP_release3.zip</td>
</tr>
<tr>
<td>SampleEntry</td>
<td>.csv (1.61 mb)</td>
</tr>
<tr>
<td>Data_Dictionary_release3</td>
<td>.pdf (150.76 kb)</td>
</tr>
<tr>
<td>HHP_release3</td>
<td>.zip (52.69 mb)</td>
</tr>
</tbody>
</table>

**IMPORTANT NOTE:** The information provided below is intended only to provide general guidance to participants in the Heritage Health Prize Competition and is subject to the Competition Official Rules. Any capitalized term not defined below is defined in the Competition Official Rules. Please consult the Competition Official Rules for complete details.

Heritage Provider Network is providing Competition Entrants with deidentified member data collected during a forty-eight month period that is allocated among three data sets (the "Data Sets"). Competition Entrants will use the Data Sets to develop and test their algorithms for accurately predicting the number of days that the members will spend in a hospital (inpatient or emergency room visit) during the 12-month period following the Data Set cut-off date.

**HHP_release3.zip contains the latest files, so you can ignore HHP_release2.zip. SampleEntry.CSV shows you how an entry should look.**

Data Sets will be released to Entrants after registration on the Website according to the following schedule:

- **April 4, 2011**  Claims Table - Y1 and DaysInHospital Table - Y2
- **May 4, 2011**   All other Data Sets except Labs Table and Rx Table
- **June 4, 2011**  Labs Table and Rx Table

Entrants are welcome to use other data to develop and test their algorithms and entries until 11:59:59 UTC on April 4, 2012 if the data are (i) freely available to all other Entrants and (i) published (or a link provided) to the data in the External Data portion of the Forum within one (1) week of an entry submission using the other data. Entrants may not use any data other than the Data Sets after 11:59:59 UTC on April 4, 2012 without prior approval.

**Tables**

Each of the Data Sets will be comprised of tables as follows:
• a. Members Table, which will include:
  ◦ i. MemberID (a unique member ID)
  ◦ ii. AgeAtFirstClaim (member's age when first claim was made in the Data Set period)
  ◦ iii. Sex

• b. Claims Table, which will include:
  ◦ i. MemberID
  ◦ ii. ProviderID (the ID of the doctor or specialist providing the service)
  ◦ iii. Vendor (the company that issues the bill)
  ◦ iv. PCP (member's primary care physician)
  ◦ v. Year (the year of the claim, Y1, Y2, Y3)
  ◦ vi. Specialty
  ◦ vii. PlaceSvc (place where the member was treated)
  ◦ viii. PayDelay (the delay between the claim and the day the claim was paid for)
  ◦ ix. LengthOfStay
  ◦ x. DSFS (days since first service that year)
  ◦ xi. PrimaryConditionGroup (a generalization of the primary diagnosis codes)
  ◦ xii. CharlsonIndex (a generalization of the diagnosis codes in the form of a categorized comorbidity score)
  ◦ xiii. ProcedureGroup (a generalization of the CPT code or treatment code)
  ◦ xiv. SupLOS (a flag that indicates if LengthOfStay is null because it has been suppressed)

• c. Labs Table, which will contain certain details of lab tests provided to members.
• d. RX Table, which will contain certain details of prescriptions filled by members.
• e. DaysInHospital Tables - Y2 and Y3, which will contain the number of days of hospitalization for each eligible member during Y2 and Y3 and will include:
  ◦ i. MemberID;
  ◦ ii. ClaimsTruncated (a flag for members who have had claims suppressed. If the flag is 1 for member xxx in DaysInHospital_Y2, some claims for member xxx will have been suppressed in Y1).
  ◦ iii. DaysInHospital (the number of days in hospital Y2 or Y3, as applicable).

These two Tables are intended for use by Entrants to train and validate their algorithms. DaysInHospital Tables are based on the Claims Table with admissions in Y2 or Y3, as applicable. As a privacy measure, any member who spent more than two weeks in hospital is grouped; they are treated as though they spent 15 days in hospital.
• f. Target - is "DaysInHospital_Y4" but doesn't include DaysInHospital. DaysInHospital data for Y4 are to be filled in by Entrants to produce entries. See SampleEntry.csv as an example.

For more information on Competition Data, please see the Official Rules (particularly Rules 5-7), the FAQs or the Forum or send us a message through the "Contact Us" function.
Dos and Don'ts

Source: https://www.heritagehealthprize.com/c/hhp/details/DosandDonts

IMPORTANT NOTE: The information provided below is intended only to provide general guidance to participants in the Heritage Health Prize Competition and is subject to the Competition Official Rules. Any capitalized term not defined below is defined in the Competition Official Rules. Please consult the Competition Official Rules for complete details.

DO read the Official Rules carefully before you register DONE

DO provide only accurate and truthful information when you register DONE

DO respect the privacy, confidentiality and security of the Competition’s Data Sets OKAY

DO observe customary and prudent medical data security and privacy practices (including but not limited to data confidentiality, storage and encryption practices) with respect to all data that you use in connection with the Competition OKAY

DO check the Website regularly for updates OKAY

DON'T attempt to determine the identity of the individuals and/or health care providers reflected in any data that you receive in connection with the Competition OKAY

DON'T register if you do not accept all of the Official Rules OKAY

DON'T register or attempt to register under more than name or with more than one account. OKAY

DON'T try to be a member of more than one Team OKAY

DON'T disclose or share the Data Sets with anyone who isn’t on your Team or use the Data Sets for any purpose other than the Competition OKAY

DON'T attempt to reverse engineer, "crack", "hack", manipulate or gain unauthorized access to the Leaderboard OKAY
DON'T post anything in the Forum that others might find harassing, abusive, threatening, harmful, vulgar, profane, obscene, excessively violent, racially, ethnically or otherwise objectionable or offensive in way OKAY

DON'T attempt to deliberately damage the Website or undermine the legitimate operation of this Competition OKAY

FAQs

Source: https://www.heritagehealthprize.com/c/hhp/details/FAQ

IMPORTANT NOTE: The information provided below is intended only to provide general guidance to participants in the Heritage Health Prize Competition and is subject to the Competition Official Rules. Any capitalized term not defined below is defined in the Competition Official Rules. Please consult the Competition Official Rules for complete details.

Updated as of May 4, 2011

General

Why the Competition?

Dr. Richard Merkin, President and CEO of Heritage Provider Network, Inc. ("HPN") developed the U.S. $3 million Heritage Health Prize Competition (the "Competition") to promote innovation through predictive modeling, challenging the brightest minds in the world to build an algorithm to find a solution to the approximately U.S. $2 trillion dollars we're spending on health care in the U.S. We believe it's currently the world's largest health prize for predictive modeling, larger than The Nobel Prize for Medicine or The Gates Prize for International Health.

It is truly a visionary approach to saving the estimated US $30 billion spent by Americans in avoidable hospitalizations, changing the approach to healthcare in a global manner. Doing so would improve the health care of all Americans, while at the same time lowering the cost of care for the average American family. If the Competition is successful, HPN expects it will be able to use the winning algorithm to both keep people healthy and out of the hospital and lower the cost of care.

What is the Heritage Provider Network?

Heritage Provider Network is one of the largest integrated physicians networks in the world, providing health care for approximately 700,000 Californians. HPN connects patients with physicians, hospitals and specialists to provide quality care in a cost effective manner. HPN is a group of physicians, not an insurance company. HPN does not sell insurance.

How long will the Competition last?

The Competition begins on April 4, 2011 and ends at 6:59:59 UTC on April 4, 2013.

How do I enter the Competition?
To register, visit [http://www.heritagehealthprize.com](http://www.heritagehealthprize.com) (the "Website") or [www.kaggle.com](http://www.kaggle.com) and follow the onscreen instructions to complete and submit your registration. You may register at any time prior to 6:59:59 UTC on October 4, 2012.

What am I being asked to predict?

You're being asked to predict the number of days HPN members will spend in a hospital in a given year (Y4 or Y5, as applicable). For the purposes of the Competition, a hospitalization includes both inpatient admissions and emergency room visits.

Eligibility

Can anyone enter the Competition?

You must be of at least the age of majority or older in your place of residence as of the time that you register and you cannot be a resident of Cuba, Sudan, Iran, North Korea, Syria or any other country designated by the United States Treasury's Office of Foreign Assets Control. See [http://www.treasury.gov/resource-center/sanctions/Programs/Pages/Programs.aspx](http://www.treasury.gov/resource-center/sanctions/Programs/Pages/Programs.aspx) for a list of sanction programs by country.

Also, if you are an officer, director, employee or advisory board member of Heritage Provider Network, Inc., Kaggle Pty Ltd, Children's Hospital of Eastern Ontario Research Institute, Inc., University of Ottawa, University of Maryland, Baltimore County or Privacy Analytics Inc. or one of their respective affiliates, agents, judges and advertising and promotion agencies -- or are an immediate family member or member of the same household of someone who is -- you are not eligible to participate in the Competition.

Am I eligible to enter the Competition if I have/had a business relationship with HPN?

Yes. If you are or were a third party service provider, consultant or contractor of HPN, you may participate in the Competition; however, you may not use in the Competition any data or information received in connection with your work for or on behalf of HPN.

Teams

Do I need to be part of a Team to participate?

No. **You may participate individually** or as a member of a Team but not both.

Is there a limit on the number of people who may be on a Team?

Yes. A Team can have up to eight (8) members. Each Team member must register individually.

Can I be part of more than one Team?

No, you can't. We limit you to participating individually or as a member of only one Team because we have a strict one Entry per day limit. If you are part of several Teams or register with several accounts, you can circumvent this limit. If we
learn that a participant is part of more than one Team, we reserve the right to disqualify him/her and remove his/her Entries. We feel strongly that every Entrant should compete on a level playing field.

How do I form a Team?

One Team member must serve as the Team leader. Initially, the Team leader will be the Team member who first completes the ‘Team Wizard’ on the Competition Website. You will be prompted to complete the Team Wizard just before you submit your first Entry. You will not be considered a member of a Team unless and until you confirm your Team membership by responding to the Team notification message available through your Account.

Can the Team leader change?

Yes. To change the Team leader, please use the Website's Contact Us Form.

What are Team Mergers and why are you limiting them?

Team mergers allow two separate Teams to combine to participate in the Competition as a single Team. They are a great feature of the Competition because they permit participants to collaborate with others with whom they wouldn't otherwise work and to learn new techniques.

However, we must balance the benefits of Team mergers with the need to keep the Competition as fair as possible. We want everybody to compete on a level playing field so Team mergers will be considered on a case-by-case basis only when requested. They generally will be permitted when we determine that the merger will not result in the merged Teams getting more Entries to win one of the Competition's seven cash prizes than other participants.

Data

When will I get the Competition Data Sets?

You will get the Data Sets after completing the registration process. Data Sets will be released to Entrants on the Website as follows:

- April 4, 2011  Claims Table - Y1 and DaysInHospital Table - Y2
- May 4, 2011  All other Data Sets except Labs Table and Rx Table
- June 4, 2011  Labs Table and Rx Table

Why has certain identifying information been removed from the Data Sets?

The identity of individual patients and health care providers, as well as other individual identifiable information, has been removed from the Data Sets to protect the privacy of those involved and to comply with applicable law.

How should I use the Data Sets?
The Data Sets should be used only to participate in the Competition, including but not limited to developing your Competition Entries and participating in the forum discussions on the Website.

**How should I maintain the Data Sets?**

You must maintain customary and prudent medical data security and privacy practices (including but not limited to data confidentiality, storage and encryption practices) with respect to the Data Sets and other data used in connection with the Competition. If we determine that you aren't, you may be disqualified.

**Can I use data other than the Data Sets?**

Yes. You may use data other than the Data Sets to develop and test your/your Team’s Prediction Algorithm and Entries until 11:59:59 UTC on April 4, 2012 if (a) the data are freely available to all other Entrants and (b) you publish the data or a link to the data in the External Data portion of the Forum section of the Website within one (1) week of your submission of an Entry using the data.

You may not, however, link the Data Sets to records in other external databases such that new demographic, socioeconomic or clinical information about the members in the Data Sets is gained. HPN reserves the right in its sole discretion to disqualify any Entrant who it discovers has undertaken or attempted to undertake such linking of the Data Sets.

You may not use any data other than the Data Sets after 11:59:59 UTC on April 4, 2012 without HPN's prior approval.

**Can I use the Data Sets for research purposes?**

HPN does not permit use of the Data Sets for any purpose other than participation in the Competition without HPN's prior written approval. If you would like to request permission to use the Data Sets for research purposes, please use the Website’s Contact Us Form and include a reasonably detailed description of the proposed research that would involve the Data Sets. All such requests will be given careful consideration.

**Why do the DaysInHospital_Yx Tables contain only a subset of all members?**

The DaysInHospital_Yx Tables only contain the members who are eligible to make a claim in Yx and the 12-month period prior to Yx. For these purposes, "eligible to make a claim" refers to individuals who were HPN members in both those years. For example, the DaysInHospital_Y4 Table contains HPN members who each were eligible to make a claim in Y4 and Y3. A member can become ineligible to make a claim if he/she ceases to be a HPN member.

**Entries**

**What is an “Entry”??**

"Entry" means the data submitted via the Website in the manner and format specified on the Website on an Entry form.

**How do I submit Entries?**

Each Entry must be uploaded to the Website in the manner and format specified on the Website.
How many Entries can I submit?

You or your Team may submit one Entry each day during the Competition period, beginning on May 4, 2011.

Are there any fees for submitting Entries?

No. You can participate in the Competition without paying any fee.

Will all my Entries be considered for all of the prizes?

No. You must designate which Entries you want considered for each prize. You may (but are not required to) designate one Entry for consideration for each Milestone Prize round and up to five Entries for consideration for the Grand Prize. If you don't designate five Entries for consideration for the Grand Prize, your five Entries with the lowest Prediction Error Rates (as described in Rule 12 of the Competition Official Rules) on the Leaderboard will be automatically designated for Grand Prize consideration.

Can I change my Entry designations?

Yes. Entry designations for consideration for a prize can be changed any time prior to the applicable designation deadline.

What are the deadlines for designating Entries for prizes?

All designations for Milestone Prizes Entries must be made by 06:59:59 (UTC) on the following dates:

- Round 1 Milestone Prizes: 06:59:59 UTC on August 31, 2011
- Round 2 Milestone Prizes: 06:59:59 UTC on February 13, 2012
- Round 3 Milestone Prizes: 06:59:59 UTC on September 4, 2012

All Grand Prize Entry designations must be made by 06:59:59 UTC on April 4, 2013.

What is a "Prediction Algorithm"?

A "Prediction Algorithm" is the algorithm used to produce the data in an Entry taken as a whole (i.e., its particular total configuration) but does not include individual components of the Prediction Algorithm or tools used for analysis or development of the Prediction Algorithm.

Must my Prediction Algorithm and Entries be my (or my Team’s) original work?

Yes. Your Prediction Algorithm and Entries must be the exclusive original work of you (or if applicable, your Team), must not have been previously published or won any other prize and must not violate any laws or third party rights.

Should I use a version control system in developing my Prediction Algorithm?

We recommend that you use some kind of version control system (like git) to help ensure that, if you are a conditional prize winner, the judges will be able to use your Prediction Algorithm to reproduce your Entries.
What is the Prediction Error Threshold?

The Prediction Error Threshold is the maximum Prediction Error Rate required to win the Grand Prize. The Prediction Error Threshold will be published on the Website on May 4, 2011. The Grand Prize will be awarded to the Entrant that develops the Prediction Algorithm that produces the eligible Grand Prize Entry with the lowest Prediction Error Rate that is at least as low as the Prediction Error Threshold as of 06:59:59 UTC on April 4, 2013.

What if no Grand Prize Entry has a Prediction Error Rate lower than the Prediction Error Threshold?

If no Grand Prize Entry has a Prediction Error Rate that is lower than the Prediction Error Threshold as of 6:59:59 UTC on April 4, 2013, the US $3 million Grand Prize will not be awarded and the US $500,000 Consolation Prize will be awarded to the eligible Grand Prize Entry having the lowest Prediction Error Rate.

If no Grand Prize Entry meets the Accuracy Threshold as of 6:59:59 UTC on April 4, 2013, the U.S. $3 million Grand Prize will not be awarded and the US $500,000 Consolation Prize will be awarded to the eligible Grand Prize Entry having the most accurate prediction score.

Do Milestone Prize Entries need to meet the Prediction Error Threshold?

No.

If my Entry is selected as a conditional winner, what else will I need to provide to verify my Entry?

If your Entry is selected as a conditional winner, you must provide the code for the Prediction Algorithm that produced your Entry and any requested back-up documentation. You also may be asked to submit computer hardware or virtual machine instance that runs the code.

What can I/we use to develop my/our Prediction Algorithm?

You may use any algorithm and/or analytical and developmental tool, including open source tools, pre-existing algorithms and other items of your choosing.

What should I do if a component of my/our Prediction Algorithm belongs to someone else?

If any component of your Prediction Algorithm is legally available only through a fee-based license and/or for non-commercial use, then you must contact us for approval before including the component in your Prediction Algorithm. You do not need to obtain our permission for use of any component available pursuant to an OSI-approved license listed at http://www.opensource.org/licenses/alphabetical.

If I'm not (or my Team isn't) a conditional prize winner, might I still need to submit the Prediction Algorithm to HPN?

We may request that you submit your (or your Team’s) Prediction Algorithm to us at any time. This will help us to preserve the integrity of the Leaderboard and otherwise prevent fraud so that the Competition is as fair as possible for all Entrants.

Will HPN have the right to use my Entries and Prediction Algorithm?
Yes. By registering in the Competition, you grant HPN an exclusive, worldwide, perpetual, royalty free license to use and grant others the right to use your Entries (see description above) and the Prediction Algorithm (see description above) used to produce your Entries. Such use may be for any commercial or other purpose and won’t require your further approval or any payment to you.

*Will I have the right to use my Entries and Prediction Algorithm?*

Yes. You may use your Entries and the Prediction Algorithm used to produce your Entries but such use must be solely for your own patient management and internal business purposes. You may not transfer or grant any rights to your Entries or Prediction Algorithm to any other entity or person.

*Why is HPN seeking to control my Prediction Algorithm?*

HPN wants to help ensure that Prediction Algorithms are used to promote better patient health care and not for improper purposes.

**Judging**

*Who is judging the Competition?*

Judging will be conducted by a panel of independent, qualified experts.

*When will the Leaderboard be available?*

From May 4, 2011 through the end of the Competition.

*Do Entrants winning on the Leaderboard receive any prize?*

The Leaderboard is intended only to give you an indication of how you are performing relative to other participants who have submitted Entries. Standings on the Leaderboard don’t determine prize winners.

*What if there are ties?*

If any two or more Prediction Algorithms produce eligible Entries with the same lowest prediction score, the Entry received first will be deemed the winner of the applicable prize.

**Prizes**

*What can I win?*

The Competition offers a Grand Prize of US $3,000,000 (or if no Grand Prize Entry produces a Prediction Error Rate lower than the Prediction Error Threshold, a Consolation Prize of US $500,000) and six Milestone Prizes:

- **Round 1 Milestone Prizes**
  - First Prize: U$30,000
  - Second Prize: US $20,000

- **Round 2 Milestone Prizes**
  - First Prize: US $50,000
  - Second Prize: US $30,000
Round 3 Milestone Prizes

First Prize: US $60,000
Second Prize: US $40,000

What if my Team wins?

If a Team wins a prize, all Team members must submit a single written statement describing how the prize is to be allocated among the Team members. If the Team fails to submit the statement within 30 days after HPN requests it, HPN will distribute the prize to Team members in equal shares.

Forum

Who is responsible for Forum posts?

You are responsible for your own Forum posts, although HPN reserves the right to remove posts which contain inappropriate, offensive or illegal material.

Timeline

Source: https://www.heritagehealthprize.com/c/hhp/details/Timeline

April 4, 2011 at 19:00 UTC: Registration opens; Claims Table - Y1 and DaysInHospital Table - Y2 released

May 4, 2011: All other Data Sets (except Labs Table and Rx Table) released; Entries accepted and Leaderboard go-live; Accuracy Threshold released

June 4, 2011: Labs Table and Rx Table released

August 31, 2011 at 06:59:59 UTC: Round 1 Milestone Prize Entry Deadline

February 13, 2012 at 06:59:59 UTC: Round 2 Milestone Prize Entry Deadline

April 4, 2012 at 11:59:59 UTC: Deadline for Use of Other Data

September 4, 2012 at 06:59:59 UTC: Round 3 Milestone Prize Entry Deadline

October 4, 2012 at 06:59:59 UTC: Registration Deadline / Team Merger Deadline

April 4, 2013 at 06:59:59 UTC: Grand Prize Entry Deadline; Competition ends
Team

Source: https://www.heritagehealthprize.com/teams/13360/brand-niemann

Delete/disband team (since it has made no submissions)

Brand Niemann  Change team name

• BrandNiemann (you're the leader) Brand Niemann

Save team member additions

Search for a team member from all users who are not already on a team

NOTE: If a user does not appear in the search results but they are a Kaggle member, then they're probably already part of a team. If they have not made any submissions yet to that team and they're the leader, they can go into the "Team" page and click "Disband team" to delete the team and thus make them eligible to be a part of your team.

Make a submission

Source: https://www.heritagehealthprize.com/c/hhp/submissions/attach

Attach Files & description
You will be making a submission on behalf of "Brand Niemann."
You have 1 entry remaining today. This competition allows your team to make 1 entry in a day (resets at midnight, UTC: 4 hours from now)

Submission * Browse...

Your entry must:

• be in CSV format
• have your prediction in column 3
• have exactly 70,943 rows

Each predicted value must be:

• Total number of days spent in the hospital. That is, a real-valued number in the interval [0, 15].

Description of your technique

1. Select Team
2. Attachments

Your Submissions

Source: https://www.heritagehealthprize.com/c/hhp/submissions
You are submitting as part of team Brand Niemann.

**Note:** You can select up to 1 submission that will be used to calculate your final leaderboard score. If you do not select them, up to 1 entry will be chosen for you based on your most recent submissions. Your final score will not be based on the same exact subset data as the public leaderboard, but rather a different private data subset of your full submission. Your public score is only a rough indication of what your final score might be. You should choose an entry that will most likely be best overall, and not necessarily just on the public subset.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Files</th>
<th>Public Score</th>
<th>Selected?</th>
</tr>
</thead>
</table>

Submit Selection Changes

---

**Forum**

Source: [https://www.heritagehealthprize.com/c/hhp/forums](https://www.heritagehealthprize.com/c/hhp/forums)

355 topics, 2,379 posts

---

**Leaderboard**

Source: [https://www.heritagehealthprize.com/c/hhp/leaderboard](https://www.heritagehealthprize.com/c/hhp/leaderboard)

941 teams as of Tue, 13 Mar 2012 19:46:12

---

**Round 2 Milestone Leaderboard**

Source: [https://www.heritagehealthprize.com/c/hhp/leaderboard/milestone2](https://www.heritagehealthprize.com/c/hhp/leaderboard/milestone2)

830 teams as of Mon, 13 Feb 2012 06:59:59

This leaderboard represents the private leaderboard rankings as of the Round 2 Milestone progress prize deadline. The top two teams were required to describe their techniques to be eligible for the progress prizes:

1. [Market Makers - Milestone 2 Description](https://www.heritagehealthprize.com/c/hhp/leaderboard/milestone2)
2. [Edward and Willem - Milestone 2 Description](https://www.heritagehealthprize.com/c/hhp/leaderboard/milestone2)

---

**In the Media**

Source: [https://www.heritagehealthprize.com/news](https://www.heritagehealthprize.com/news)

---

**In the News**

• **Wall Street Journal Blog** - 23 Sep 2011
• **New York Times Bits Blog** - 21 May 2011
• **New York Times** - 21 May 2011
• **The Healthcare Blog** - 26 April 2011
• **The Economist** - 10 April 2011
• **Slate** - 4 April 2011
• **Forbes** - 4 April 2011
• **IBM Smarter Planet** - 4 April 2011
• **Popular Science** - 1 April 2011
• **Computer World** - 30 March 2011
• **Network World** - 29 March 2011
• **Revolution Analytics** - 17 March 2011
• **Visualising Data** - 17 March 2011
• **The Wall Street Journal** - 16 March 2011
• **The Hospital Leader** - 16 March 2011
• **The Drucker Exchange** - 16 March 2011
• **Silicon Angle** - 16 March 2011
• **iHealthBeat** - 16 March 2011
• **eHealthCentral** - 16 March 2011
• **IBM Systems Magazine** - 15 March 2011
• **The New Scientist** - 3 February 2011
• **O'Reilly Radar** - 3 February 2011
• **eHealth Insider** - 1 February 2011
• **Fast Company** - 31 January 2011
• **Washington Post** - 28 January 2011
• **The Daily Caller** - 27 January 2011
• **KDNuggets** - 25 January 2011
• **Business Insider** - January 21 2011
• **eHealthCentral** - January 21 2011
• **Health Leader Media** - December 30 2010
• **Read Write Web** - December 16 2010
• **Health Data Management** - December 6 2010

**Video Coverage**

Media, iframe, embed and object tags are not supported inside of a PDF.
Judging Panel

Source: https://www.heritagehealthprize.com/team

Who's Behind the Prize?

Advisory Board/Judges

Yehuda Koren joined Yahoo! Research in September 2008. Prior to that, he was a member of AT&T Labs-Research. He obtained his PhD in Computer Science from The Weizmann Institute in 2003. He was awarded best paper award at INFOVIS 2005 for the work on directed graph layout through constrained energy minimization. More recently he won KDD 2009 best research paper award for his work on collaborative filtering with temporal dynamics. Yehuda led the team that won first awards in the Netflix Prize competition, and was part of the team winning the Netflix Grand Prize. He co-organized the KDD-Cup’11 contest.
Charles Elkan joined the UCSD faculty in 1990, after earning his PhD that same year in computer science at Cornell University. He did his undergraduate degree at Cambridge University. Elkan was a postdoctoral fellow at the University of Toronto. In 1998-99, he was a visiting Associate Professor in computer science at Harvard University. While at Harvard, he was Senior Scientist at the software firm Knowledge Stream Partners. Elkan has consulted for Hewlett-Packard, SAIC, Sony, IBM, and Alcoa. He has won numerous best-paper awards, including first-place at the CoIL Challenge 2000 data mining competition. Elkan is the co-founder of UCSD's Artificial Intelligence Laboratory.

Claudia Perlich is currently working as Chief Scientist at m6d. She is working on targeted online display advertising. Prior to that she spent five years at the Predictive Modeling Group in the IBM T.J. Watson Research Center. Her research interests were the practical applications of machine learning. She graduated in 2004 from the Information Systems Department at Stern, NYU under the supervision of Foster Provost.

Saharon Rosset has been a Senior Lecturer in the Statistics Department at Tel Aviv University since 2007. Prior to that he spent four years at IBM Research, in the Data Analytics Research Group. He graduated with a PhD in statistics from Stanford University in 2003, where he worked with Jerry Friedman and Trevor Hastie, researching regularization and boosting. His research interests are statistical genetics, statistical learning theory and applications of data mining.
Khaled El Emam is an Associate Professor at the University of Ottawa, Faculty of Medicine, a senior investigator at the Children’s Hospital of Eastern Ontario Research Institute, and a Canada Research Chair in Electronic Health Information at the University of Ottawa. His main area of research is developing techniques for health data anonymization and secure disease surveillance for public health purposes. Previously Khaled was a Senior Research Officer at the National Research Council of Canada, and prior to that he was head of the Quantitative Methods Group at the Fraunhofer Institute in Kaiserslautern, Germany. He has co-founded two companies to commercialize the results of his research work. In 2003 and 2004, he was ranked as the top systems and software engineering scholar worldwide by the Journal of Systems and Software based on his research on measurement and quality evaluation and improvement, and ranked second in 2002 and 2005. He holds a Ph.D. from the Department of Electrical and Electronics, King’s College, at the University of London (UK).

Arvind Narayanan (PhD 2009) is a post-doctoral fellow at the Stanford Computer Science department and a Junior Affiliate Scholar at the Stanford Law School Center for Internet and Society. He studies privacy from a multidisciplinary perspective, focusing on the intersection between technology, law and policy. His research has shown that data anonymization is broken in fundamental ways, for which he jointly received the 2008 Privacy Enhancing Technologies Award. He is one of the researchers behind the “Do Not Track” proposal.
Dr Richard Merkin has more than 30 years of experience in the health care field. He has specific expertise in the development and administration of integrated physician systems. As the founder of Heritage Provider Network, which was established in 1979, Dr. Merkin develops clinically focused networks to bring efficient and quality driven systems to the communities in which it operates by working with physicians and physician organizations, hospitals and integrated delivery systems, health plans, public and community-based health care entities, and other health care professionals.

Dr. Merkin is a visionary and a sought-after healthcare expert who encourages innovation and challenge. Responding to our country’s 2 trillion dollar health care crises, Dr. Merkin created, developed and sponsored the 3 million dollar Heritage Health Prize for predictive modeling to save more than 30 billion in avoidable hospitalizations. It is the largest predictive modeling prize in the world, larger than the Nobel Prize for Medicine and the Gates Prize for Health. Dr. Merkin is genuinely excited to bring new minds to the healthcare table with the prize and believes data miners hold great potential for not only bringing a winning algorithm, but also to grab the attention of data miners globally and raise awareness about competitive innovation. As a Board member and core contributor to the X Prize Foundation, Dr. Merkin is well aware of the need for the private sector to step up to the plate and assist in a globally transformative way, specifically solving one of our country’s biggest problems - keeping patients healthier and out of the hospital. Dr. Merkin was named Healthcare CEO of the Year for 2011 by the Los Angeles Business Journal.

Dr. Merkin established the Richard Merkin Foundation for stem cell research at the Broad Institute at Harvard and established the Richard Merkin Initiative at the Johns Hopkins Brain Sciences Institute and The Richard Merkin Foundation for Neural Regeneration at UCLA.

Outside of healthcare, Dr. Merkin is involved in educational, economic and financial endeavors. He is on the board of advisors for the Asia Society Southern California, on the board of advisors for Tennenbaum Capital and has been on the Mayor’s Executive Committee for Los Angeles. A strong advocate for education, Dr. Merkin is on the boards for the Sierra Nevada College, The Alliance for College-Ready Public Schools, The United Friends of Children and EdVoice, a team dedicated to education reform. The Richard Merkin Middle School opened its doors September 2006 in the city of Los Angeles, committed to bringing first-rate education to an under-served area.

Dr. Merkin earned his M.D. at the University of Miami, School of Medicine.
Anthony Goldbloom is the founder and CEO of Kaggle. He assists companies with framing modeling tasks as data prediction competitions, ensuring that competitions reflect real-life projects.

Before founding Kaggle, Anthony worked in the macroeconomic modelling areas of the Reserve Bank of Australia and before that the Australian Treasury. In these roles, Anthony was responsible for building macroeconomic models, generating economic forecasts and simulating the impact of changes in interest rates and fiscal policy on the Australian economy.

Anthony holds a first class honours degree in economics and econometrics from the University of Melbourne and has published in The Economist magazine and the Australian Economic Review.

Jeremy Howard is Kaggle's Chief Data Scientist. He worked as a management consultant at McKinsey & Company and at AT Kearney for nine years before founding FastMail.FM (an email provider), which he later sold to Opera Software, and Optimal Decisions Group (an insurance pricing optimization specialist), which he later sold to ChoicePoint.

Jeremy regularly appears as an IT expert on television shows such as Sunrise, Midday, Evening News and the Morning Show. He joined Kaggle after prize-winning performances in a number of the site's early data prediction competitions.
Jeff Moser is Kaggle's Chief Technology Officer. He has been fascinated with creating software since he was a kid, starting programming on an Apple II at age eight. Before joining Kaggle, Jeff developed software in the defense and enterprise telephony sectors.

Jeff majored in computer science and mathematics at Purdue University, where he graduated in the top of his class. He loves learning new things and occasionally blogs at moserware.com, covering the ins and outs of everything from Microsoft's Trueskill rating system to the Advanced Encryption Standard using stick figures.