Assessing patient eligibility in clinical trials with predictive analytics
Introduction

In the healthcare industry, especially considering the large amounts of historical transactional data available, data mining leads to finding interesting patterns that can be exploited by predictive models to identify risks and opportunities in assessing the patient eligibility criteria. Predictive models capture relationships, among other factors, allowing the assessment of risk or potential risk associated with a particular set of conditions that guide decision-making for candidate transactions.
Randomized clinical trials have been a key for the development of a reliable evidence-based medicine. The trials generally evaluate a treatment relative to a control regimen for a broadly defined population of patients based on primary site, diagnosis, stage and number of prior treatments. Placebo test control studies—a method of research in which an inactive substance (a placebo) is given to one group of participants while the treatment (usually a drug or vaccine) being tested is given to another group—are also used to assess patient eligibility.

One limitation of randomized clinical trials and Placebo test control studies is that they can lead to the over-treatment of patients, most being those who don’t benefit from the drugs. Some procedures have resulted in statistically average treatment effects.

**Traditional Methods for Clinical Trials:**

Randomized clinical trials have been a key for the development of a reliable evidence-based medicine. The trials generally evaluate a treatment relative to a control regimen for a broadly defined population of patients based on primary site, diagnosis, stage and number of prior treatments. Placebo test control studies—a method of research in which an inactive substance (a placebo) is given to one group of participants while the treatment (usually a drug or vaccine) being tested is given to another group—are also used to assess patient eligibility.

One limitation of randomized clinical trials and Placebo test control studies is that they can lead to the over-treatment of patients, most being those who don’t benefit from the drugs. Some procedures have resulted in statistically average treatment effects.

**Clinical Trials in a Nut Shell**

- Approved Protocol
- Investigator Selection
- Approval Process
- Statistical Analysis
- Data Entered and Reviewed
- Patient Recruitment and Participation
- Presentation and Publication of Report
- Data file and Registration obtained
Present Challenges:

There are challenges when scientists try to determine patient eligibility for studies, but come across incomprehensible and ambiguous criteria, as well as under-specified criteria requiring clinical judgment or assessments. The development of electronic health records (EHR), and the increasing volume of medical data contained within them, has raised the question of how this data can be reused to improve the recruitment process of clinical trials.

Clinical studies account for almost half of the cost of drug development. More than 80 percent of all clinical trials experience significant delays, the cost of which can exceed $35,000 per day. Patient recruitment and retention in clinical trials are widely recognized as the leading clog in the new drug development pipeline. The unfortunate result of this problem is that pharmaceutical, biotech and medical device companies spend millions in additional development costs and lose hundreds of millions in revenue. When discrepancies between planned and actual recruitment rates occur, trials need to be prolonged at a considerable cost or may be aborted.
New Methods for Clinical Trials Using Clinical Trials Simulators:

One of the best methods commonly approached is the Clinical Trials Simulators (CTRSS) method. It is a user-friendly decision support technology that takes into account all of the inherent variability in clinical trials and generates realistic data on how patient recruitment will perform, resulting in faster and more effective decision-making.

Clinical Trials Simulator’s design and easy to use interface gives you the power to:
• Experiment with and optimize the number of sites any trial will require.
• Mitigate risk around recruitment by accurately projecting “true” patient enrollment and retention.
• Accurately project trial milestone dates such as first subject first visit (FSFV), last subject last visit (LSLV) or close out.
• Accurately forecast the cost of the optimal trial solution.
NOTE: To support free text handling problems, Zhang et al. recently reported on their subtree match method, which finds structural patterns in free text sentences and thus allows for the ability to find similar sentences in other documents. Both the list of keywords and their grammatical structure can be automatically derived from the text. We believe this approach could be further enhanced to allow the inclusion of small text fragments into the predictive modeling process. (Ref: http://www.biomedcentral.com/1472-6947/13/134)

Though CTRSS offers many advantages—namely it’s independence from the trial protocol’s definition of eligibility criteria and the terminology of the clinical database, there exists need of further studies to compare the robustness of case-based and rule-based CTRSS for missing data and incorrect data. We hypothesize that case-based systems are more insensitive to missing data, because they are based on a broader base of data elements and more sensitive to incorrect data as these cause imprecise prediction models.

Furthermore, recent literature suggests that an increase in recruited patients and improvements in recruitment efficiency can be expected, although the number of recruited patients will depend on the error rate of the recruitment process being replaced. Finally, to increase the quality of future CTRSS reports, there is a proposal for a checklist of items that should be included.

Predictive analytics has long held the promise of solving the issue of improving patient recruitment efficiency. Only recently has data become available in sufficient quality and quantity to bring predictive analytics closer to fulfilling that promise.Powered by vast quantities of high fidelity clinical and claims data, predictive analytics can identify patients’ eligibility with greater speed and accuracy than ever before.
Methodology for Conventional Methods vs. Predictive Modeling:

<table>
<thead>
<tr>
<th>Conventional Methods of Modeling</th>
<th>Predictive Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less accurate predictions</td>
<td>More accurate predictions</td>
</tr>
<tr>
<td>Less feasibility in clinical trials as reporting on prototypes is less</td>
<td>Feasibility in clinical trials because of the ability to report on a prototype’s performance for different system configurations</td>
</tr>
<tr>
<td>Fails to handle huge data appropriately</td>
<td>Handles big data appropriately</td>
</tr>
<tr>
<td>Biased handling of large number of variables</td>
<td>Unbiased handling of large number of variables</td>
</tr>
<tr>
<td>Fails to show the correct direction of implication of the variables</td>
<td>Shows correct direction of significance of variable attributes</td>
</tr>
<tr>
<td>Heterogeneous samples gets difficult to handle</td>
<td>Can easily handle the heterogeneous sample of data</td>
</tr>
<tr>
<td>Works on explicit rules</td>
<td>Works on historical data making case-based reasoning</td>
</tr>
<tr>
<td>Goodness of fit (distances between predicted and observed outcomes. Lesser the distance more accurate is the model) is bad here</td>
<td>Goodness of fit is good here with lesser distance between the predicted and observed outcomes</td>
</tr>
</tbody>
</table>
Explained the meaning of goodness of fit and the difference between goodness of fit in conventional and predictive modeling method.

Predictive modeling applies mathematical and statistical techniques to predict future outcomes and improve the overall ability to segment a population on the basis of a future probability or outcome. Predictive analytics are composed of a variety of statistical techniques from modeling, machine learning and data mining that analyze current and historical facts to make predictions about future events.

The feasibility of predictive modeling to assess the eligibility of patients for clinical trials and report on a prototype’s performance for different system configurations help to solve this challenge. Also, the importance is growing in terms of comparing the studies between the control and test samples of various clinical trials.

Biomarker studies with a combination of predictive modeling could help to address this issue. Predictive biomarker assessments are developed and validated either internally or externally by partner companies with expertise in molecular analysis. The recent U.S. Food and Drug Administration (FDA) co-approvals of several therapeutic compounds and their companion diagnostic devices (FDA News Release, 2011, 2013) to identify patients who would benefit from treatment have led to considerable interest in incorporating predictive biomarkers in clinical studies.
Why Today is the Tipping Point for Predictive Analytics in Assessing Patient Eligibility:

In past years, the challenge with using predictive analytics in assessing patient eligibility has been the underlying data. Until recently, the main source of digital health care data was claims data. But claims data does not get at a patient’s overall health or disease-specific functioning. This clinical data was often handwritten, dictated or incomplete. Hence, predictive modeling relied on relatively small data sets, often of poor quality and with limited variables. The result was marginally predictive models. Because of the recent and rapid adoption of EMRs, large and diverse digital health care data sets are now available. Technology also exists that can aggregate this clinical data with other data sources, from across care settings and organizations. This data can then be better structured to enable analytics. Natural language processing (NLP) can be used to access unstructured data. The result is “bigger and better” data with higher predictive ability.
Biomarkers are biological measures of a biological state. By definition, a biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention."

Biomarkers are the measures used to perform a clinical assessment such as blood pressure or cholesterol level and are used to monitor and predict health states in individuals or across populations so that appropriate therapeutic intervention can be planned.

**Types of Biomarkers:**

- Early detection biomarker
- Endpoint biomarker
- Prognostic biomarkers
- Predictive biomarkers
Major Problems with Biomarker Studies Without a Statistical Approach:

- Inadequate focus on intended use
- Cases selected based on availability of specimens rather than for relevance to intended use
- Heterogeneous sample of patients with mixed stages and treatments.
- Too great a focus on which marker is prognostic or independently prognostic, not whether the marker is effective for intended use
- Develop Predictor of Response to New Drug

Applying predictive analysis, along with biomarkers, addresses the major issues challenging the assessment of eligibility of patients.

At the individual level, predictive analytics leverages the fact that key outcomes and outputs for biomarkers are often correlated with the patient’s prior behaviors, circumstances and characteristics, as well as those of the client’s family, associates, service providers and surroundings. By examining these correlations, predictive analytics methods can be used to rank biomarker-tested patients based on the likelihood that an outcome, whether positive or negative, will occur. For example, analysis predicting which biomarker is likely to be effective might leverage existing information about the patient’s characteristics. The model might tap these factors and other information to rank patients on the likelihood that they are eligible.

Below we describe two key uses of predictive analytics for biomarkers:

1) Identifying patients’ reactions based on the biomarkers
2) Predicting the optimal eligibility of the patients
Biomarker Selection Design (Using Predictive Modeling):

Randomize n patients to T or C.

Measure K candidate markers on each patient

- Select best marker
- Compute P1…Pk
- P1 is treatment effect in patients positive for markers 1 using optimized cut point

Perform final analysis
- Assess the p value

Declare effectiveness of T for patient positive to marker K

Computation of P:
P* = min(P1…Pk)

Where k is marker with minimum p value
Compute permutation significance level for P*
Compute whether the value of $P^*$ is statistically significant when adjusted for multiple testing.

• Adjust for multiple testing by permuting the treatment labels and re-calculating $p_1...p_K$ and $p^*$ for the permuted treatment labels
• Repeat for several random permutations to approximate the null distribution of $p^*$

A biomarker predicts the differential outcome of a particular therapy or treatment (e.g., only biomarker-positive patients will respond to the specific treatment or to a greater degree than those who are biomarker negative).

A predictive biomarker is a baseline characteristic which categorizes patients by their degree of response to a particular treatment. In this case, biomarker-positive patients perform moderately better than biomarker-negative patients when standard treatment is administered, thus exhibiting the likelihood that test treatment may be more effective in the biomarker-positive group.
Designs when there are many candidate markers and too much patient heterogeneity for any single marker:

Measure Candidate Markers

Randomize n Patients to T or C

Follow-up

Perform Final Analysis

\[ P < .01 \] Declare Broad Effectiveness

\[ P > .01 \] Partition into training set T and Validation set V

Develop Indication Classifier in T

Compare new treatment to Control in Classifier + Patients of V

\[ P > .04 \text{ in subset} \]

Declare Effectiveness for Classifier Positive Patients
Clinical Validation of Predictors:

Assessment of predictive accuracy includes internal and external validation. The internal validation is to assess predictive accuracy for the study population from which the predictor was built, typically using validation techniques such as split-sample or cross-validation. On the other hand, the external validation is performed using an independent set of samples, possibly from a more relevant population for clinical application of the predictor.

For the assessment of internal validity in high dimensions, re-sampling techniques such as cross-validation and bootstrap are useful, particularly when the sample size is small. When using these techniques, it is critical that all aspects of model building including gene selection are re-performed for each round in re-sampling.

When selection of genes and prediction models are optimized based on cross-validated predictive accuracy, the optimization process should be included in the cross-validation procedure or an independent validation set is needed to have an unbiased estimate of the predictive accuracy. If the cross-validated predictive accuracy measures without incorporating the optimization process then it turns relatively insensitive to selection of the tuning parameters used in the optimization, Confidence intervals for cross-validated prediction error can also be calculated.

Unbiased estimation of the predictive accuracy is particularly important when the number of candidate variables (genes) available for use in the predictor is much greater than the number of samples available for analysis. For class prediction problems, the proportion of correct classification, sensitivity, and specificity are common measures of predictive accuracy.

In high-dimensional situations, one must focus clearly on the objective of accurate prediction and not confuse this objective with that of achieving biological insight or ensuring that all variables included are essential, or that the model is "correct". For example, a prognostic genomic signature might contain a gene that is only representative of a group of highly correlated prognostic genes.
With slightly different data, a different gene from that group might be selected. Therefore, the signature will be rather unstable with different interpretations, while prediction performance may not be affected much. In other words, there might exist many “solutions” of predictor with comparable predictive accuracy under high dimension. For example, several prognostic signatures developed for breast cancer had little overlap of the component genes, but showed comparable prediction accuracy. Reproducibly of the gene list reported among similar correlative studies, which can be critical in elucidating the underlying biological mechanisms, can mislead in the assessment of genomic signatures for predictive medicine.

For assessment of predictive accuracy, a completely specified genomic signature is needed. Complete specification of the signature includes not only the list of component genes, but also the mathematical form used to combine genomic data for the genes used in the signature, weights for the relative importance of the genes, and cut-off values when making classification.

It is also important to establish that the predictive accuracy is statistically higher than that expected when there is no relationship between genomic data and the clinical variable. A permutation procedure is proposed to assess the statistical significance of cross-validated predictive accuracy.

When the model building process is complex and not easily specified in an algorithmic manner, an independent validation set would be needed.
Supporting Examples:

Conventional methods of statistical approaches versus predictive modeling in biomarker studies:

1) Study of Oncotype-Dx score to analyze the benefits of node negative Estrogen+ breast cancer patients from cytotoxic chemotherapy. The study started off with conventional methods of statistical analysis such as t tests, ANOVA and regression models. The results went erroneous in identifying the patients who are likely/unlikely to go for the cytotoxic therapy due to the following reasons from traditional methods:

   - The problem which exists in analytical difficulty when number of genes is greater than the number of cases.
   - There were pains in transitioning from an over-dependence on inference.
   - Conventional methods are based on inference problems and are not applicable to prediction problems.
   - Goodness of fit is not a proper measure of predictive accuracy.
   - Odds ratios and hazards ratios are not proper measures of prediction accuracy.
   - Statistical significance of regression coefficients are not proper measures of predictive accuracy.
   - Fitting of a model to the same data used to develop will give no evidence of prediction accuracy for independent data. Future prediction turns out difficult.

Predictive modeling method added the following advantages:

   - Predictive model was developed from a compound covariate predictor built from the log-ratios of the 10 most differentially expressed genes and can concentrate on heterogeneous group of diverse molecular entities which vary fundamentally with regard to the oncogenic mutations that cause them their responsiveness to specific drugs.

2) Determination of patient’s risk of developing a malignancy breast or ovarian cancer especially on women carrying germline mutation, such as BRCA1. Approaches like sample tests were used for reproducibility and validation.
However, if the samples in the test and validate groups are not independent, sample tests will lead to overfitting which in turn can lead to the appearance that a test has excellent discriminatory ability, which cannot be reproduced when independently validated. Here, the usage of empirical predictive models based on surveillance, epidemiology end results incidence and relative hazards from a prospective cohort, would have helped to overcome the problem.

3) Galectin3 effect on patients in determining the heart failure. Log rank test, survival distribution, multivariable cox regression model were used. The results were not appropriate because there was only a single point time concentration and there was no speculation on the importance of time. Also, it could not assess the severity of disease. It was just the hypothesis generation and the methods could not resist large sample sizes. Predictive modeling will help to cater to all these loopholes especially on predictions with large samples and over a period of time.

4) Assessment of long-term benefits to patients with ER, HER2, KRAS for the treatment of OncotypeDx. Initial application of regression modeling Fisher’s LDA vs. diagonal LDA and conventional wisdom of statistics were flawed and based on inference problems and not applicable to prediction problems. In this situation, prediction modeling helped accurately predict independent data. It also was able to divide the samples to equal-sized sets and predict in an unbiased manner with regard to their value to patients using the actual RCT data.

5) Although disease biomarkers are used widely in medicine, very few biomarkers are useful for cancer diagnosis and monitoring. This is partly because the biomarkers were not validated with the proper statistical approach which lead to discoveries. In the standard statistical modeling, the initial model typically starts with main effects and adds interaction terms as appropriate. Therefore, it would not test interactions without the main effects present in the model. However, it is well-known that the power for assessing interaction effects is often poor. A primary reason is that the sample size is calculated to address the main effects. Models should have more power to detect an interaction effect and therefore, is useful for identification of predictive biomarkers. This is where predictive modeling plays its vital role.
Conclusion:

The use of predictive analytics in clinical trials can prove useful when determining patient eligibility at a pace that is both quick and thorough for participants. Recent advances in biotechnology and genomics have stimulated further research of bio statistical and bioinformatics methodologies for the development and validation of new genomic biomarkers or diagnostic tests that are useful for selecting the right treatments for patients. The established heterogeneity of disease based on genomic biomarkers then warrants the development of new paradigms of design and analysis of clinical trials for assessing the patient eligibility criteria and the companion biomarkers toward reliable personalized or predictive medicine.
References:


4) Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data

5) Biomarkers and Treatments designing Trials


7) Prognostic and predictive markers in recurrent high grade glioma; results from the BR12 randomised trial

8) Biomarker-based predictive models for prognosis in amyotrophic lateral sclerosis

9) Gene Expression and Benefit of Chemotherapy in Women With Node-Negative, Estrogen Receptor–Positive Breast Cancer

10) The development of a predictive model to estimate cardiotoxic risk for patients with metastatic breast cancer receiving anthracyclines

11) Galectin-3 is an independent marker for ventricular remodeling and mortality in patients with chronic heart failure
12) Prognostic value of galectin-3, a novel marker of fibrosis, in patients with chronic heart failure: data from the DEAL-HF study

13) Use of Prognostic & Predictive Biomarkers in Clinical Trial Design

14) Use of Prognostic & Predictive Biomarkers in Clinical Trial Design

15) Statistical Challenges for Predictive Oncology

16) Clinical Trials of Predictive Medicine New Challenges and Paradigms


18) BIG DATA IN BIOMARKER DISCOVERY AND DRUG DEVELOPMENT

19) Validation of Biomarker-Based Risk Prediction Models

20) Identification of Gene Expression Biomarkers for Predicting Radiation Exposure

21) Translating Science to Health Care: the Use of Predictive Models in Decision Making

22) Study Designs and Statistical Analyses for Biomarker Research

23) Biomarker adaptive designs in clinical trials
Author:

Kavitha Hanumanthaiah
Analytics Lead,
Fractal Analytics

Kavitha has over 8 years of experience in pharma, finance, risk, market, IT, contact center, sales force and telecom verticals with specialization in analytics consulting and strategic planning.
About Fractal Analytics:

Fractal Analytics is a global analytics firm that serves Fortune 500 companies to gain a competitive advantage by providing them a deep understanding of consumers and tools to improve business efficiency. Producing accelerated analytics that generate data driven decisions, Fractal Analytics delivers insight, innovation and impact through predictive analytics and visual story-telling.

Fractal Analytics was founded in 2000 and has 800 people in 13 offices around the world serving clients in over 100 countries.

The company has earned recognition by industry analysts and has been named one of the top five “Cool Vendors in Analytics” by research advisor Gartner. Fractal Analytics has also been recognized for its rapid growth, being ranked on the exclusive Inc. 5000 list for the past three years and also being named among the USPAACC’s Fast 50 Asian-American owned businesses for the past two years.

Learn more at www.fractalanalytics.com

For more information, contact us at:
+1 650 378 1284
info@fractalanalytics.com

Follow us: 📱💻🌐