Biomarker Classifiers for Identifying Susceptible Subpopulation for Treatment Decision

James J. Chen, PhD

China Medical University
Taichung, Taiwan
May 11, 2011

* The views expressed in this presentation do not represent those of the U.S. Food and Drug Administration
Outline

Background

Biomarkers of Susceptibility and Biomarkers of Response

Statistical Model and Procedure

Simulation Example

Summary
Motivation

A drug is developed intended to treat the entire population of patients with certain disease.

- The drug will be unable to show efficacy if only a small fraction of population is effective.
- An approved drug is sometimes removed after the post-marketing discovery of unexpected toxicity.

An explanation for such events is an underlying heterogeneity of the general population.

To develop biomarker classifiers that can identify patients ahead of time as either susceptible or non-susceptible so that drugs could be approved selectively for the patients who can be beneficial from the drug.
Biomarkers

Biomarker: A biological indicator of status of an organism of a particular health condition or disease state.

- a set of measures associated with gene, diet, lifestyle, and disease as an indicator of health status or disease risk
- Gene expression levels, pathways, SNPs, CNV, …

Nutritional biomarker: an indicator of dietary intake and nutritional status (exposure), an index of nutrient metabolism or biological consequences of dietary intake (response).

Pharmacogenomic biomarker: an indicator of biologic processes, pathogenic processes, or pharmacological responses to therapeutic interventions.
A biomarker classifier is a mathematical function that maps the biomarker values to a set of categories. These categories correspond to levels of predicted outcome:

- **Disease prognosis**: classify individuals into risk categories for specific diseases or other health consequences.
- **Treatment assignment**: assign patients according to:
  - genomic profile
  - disease type
Typical Design and Approach

- Statistical and data mining techniques have been used to develop prediction model for identifying target patients to determine appropriate treatment, under population homogeneity.

- The study designs were dose-response experiments or a control and a dose group of many compounds including both “toxic” and “non-toxic” drugs.

- These algorithms are generally developed to classify subjects into positives and negatives with respective to response.

- Biomarker classifiers developed from biomarkers of response are not useful for identifying susceptible patients.
Population Heterogeneity

Individual patients response to a drug can be categorized into three groups:

1. Drug is efficacious – responders
2. Drug has no effect
3. Drug is harmful – identifying sensitive subpopulation

Aim: Identifying the small proportion of susceptible patients

Efficacy testing: identifying patients in 2 and 3
Safety testing: identifying patient in 3
Assumptions*

- There exists a set of biomarkers can discriminate susceptible subjects from non-susceptible subjects.

- Susceptible subjects will more likely result in toxicity positive than non-susceptible subjects from the drug treatment.

- There are endpoints or surrogate endpoints which can be used as an indicator of drug-induced toxicity.

- An observed positive response, if any, in the untreated group is considered to be not related to drug treatment.

* Safety assessment
Basic Model

A two-group design with $n$ subjects per group

$p$: Proportion of the susceptible subpopulation

$1-p$: Proportion of the non-susceptible subpopulation.

$u_1$: Probability of adverse reaction for the susceptible.

$u_2$: Probability of adverse reaction for the non-susceptible

The expected number of susceptible subjects in each group is $pn$.

In the treated group:

- The expected number of positives is $pnu_1 + (1-p)nu_2$
- The expected number of negatives is $n - [pnu_1 + (1-p)nu_2]$. 

Example – Basic Model

To develop a prediction model based on the observed positives and negatives in the treated group to identify the $pn$ susceptible subjects in the control group.

Example: $n = 200$, $p = 0.05$, $u_1 = 0.7$ and $u_2 = 0.0$

The susceptible number in each group is $pn = 10$. The positive number is 7 and the negative number is 193.

Given the 7 positives and 193 negatives in the treated group, the goal is to identify the 10 susceptible and 190 non-susceptible in the control group.
Key Steps

Two step to development of a biomarker classifier for identifying a fraction of susceptible patients:

Step 1: Identify a set of biomarkers of susceptibility from a mixture of biomarkers of susceptibility and biomarkers of response.

Step 2: Develop an imbalanced-data classifier, as the number of susceptible patients is generally much smaller than the number of non-susceptible patients, to identify susceptible patients.
Biomarker: A biological indicator (DNA or RNA characteristics) of status of an organism of a particular health condition and disease state:

Biomarkers of response (exposure): Indicators of individual response to a treatment - Levels of expression of gene, gene function, or RNA which were altered by the presence of a disease or drug treatment.

Biomarkers of susceptibility: indicators of individual differences prior to the treatment with respect to the response to the treatment or to the disease prognosis after initiation.
Two subpopulations with respect to drug exposure: susceptible and non-susceptible patients

Three sets of biomarkers with respect to drug exposure and individual susceptibility:

S: Genes express differently between the susceptible and non-susceptible individuals before the drug exposure – biomarkers of susceptibility

T: Genes express differently between the toxicity and non-toxicity subjects after exposure - biomarkers of response

D: Remaining genes
Model: Biomarkers of Susceptibility

**S:** Genes express differently between the susceptible and non-susceptible individuals before the drug exposure

**S = A ∪ B,** where

**A = S ∩ T:** Common genes in S and T *(susceptibility and response)*- Different between susceptible and non-susceptible subjects before treatment, changed by treatment.

**B = S \ A:** Genes in S not in T *(susceptibility and not response)*- Different between susceptible and non-susceptible subjects before treatment, and unchanged by treatment.
**Model: Biomarkers of Response**

\[ T = A \cup C_1 \cup C_2, \text{ where} \]

**A**: Common markers for the sets S and T, \( A = S \cap T \).

**C_1**: Genes express no different between susceptible and non-susceptible subjects before exposure, but express different changes after exposure, due to different effects.

**C_2**: Genes express no different between susceptible and non-susceptible subjects before exposure, but express same change after exposure.

The set of biomarkers of exposure T can be identified by comparing the control and treated groups.
The three biomarker sets $S^*$, $T$, and $B = S^* - T$
Identifying Biomarker of Susceptibility

- $S = A \cup B$ is not identifiable without exposure.

- A larger set $S^* = A \cup B \cup C_1$ can be identified by comparing the positive and negative subjects in the treated group.
  - $S^*$ is inefficient to develop a classifier if the number of biomarkers in $A$ or $C_1$ is large.
  - $A \cup C_1 = S^* \cap T$: Common genes in $S^*$ and $T$.

- $B$ is most useful to develop biomarkers classifier and can be indentified by subtracting the common genes in $S^*$ and $T$, $B = S^* \setminus (A \cup C_1)$. 
### Statistical model for biomarker of susceptibility for gene sets A, B, and C

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Susceptible</th>
<th>Non-susceptible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Exposure</td>
<td>$\delta$</td>
<td>0</td>
</tr>
<tr>
<td>Exposure</td>
<td>$\delta + \delta_s$</td>
<td>$\delta$</td>
</tr>
</tbody>
</table>

- **A**: Biomarkers of susceptibility and exposure
- **B**: Biomarkers of susceptibility-only
- **C = C1 ∪ C2**: Biomarkers of exposure-only.
- $\delta$: effect size of susceptibility.
- $\delta_s$: effect sizes of exposure for susceptible subpopulation
- $\delta_{NS}$: effect sizes of exposure for non-susceptible subpopulation.
Power: Proportion of Detections

For a given sample size \( n \), the expected fraction of detections of the genes in \( B \) (\( \lambda_B \)) depends on the model parameters in the two comparisons at the fixed type I error \( \alpha \), (and conversely).

In the first comparison, the expected fraction of detections of \( \lambda_{S^*} \) (genes in \( S^* = A \cup B \cup C_1 \)) depends on \( n, \delta, \alpha_1, \) and the proportion \( p \) and adverse effect probabilities \( u_1 \) and \( u_2 \).

In the second comparison, the expected fraction of detections of \( \lambda_T \) (genes in \( T = A \cup C \)) depends on \( n, \delta_S, \delta_{NS}, \alpha_2 \) and \( p, u_1, u_2 \).
- \( \lambda_C \) should be high in order to identify as much \( C \) as possible.

The expected number of detection is \( \lambda_B \).
Identifying the biomarker sets $S^*$ and $B$. Given $\lambda=0.8$, $\alpha=0.005$, $m=2000$, $A=50$, $B=50$ and $C1=0$ or 150, $C2=150-C1$ with $n = 200$ (for $\delta = 1.5$), the estimates of $\lambda_B$ and $\alpha$, and the numbers of markers in $A$, $B$, and $C1+C2+D$, based on 1,000 repetitions.

<table>
<thead>
<tr>
<th>Biomarker Set</th>
<th>($\delta_S$, $\delta_{NS}$)</th>
<th>$\lambda_B$</th>
<th>$\alpha$</th>
<th>A</th>
<th>B</th>
<th>C1+C2+D</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^*$</td>
<td>(0,0)</td>
<td>0.788</td>
<td>0.005</td>
<td>39.40</td>
<td>39.43</td>
<td>9.09</td>
</tr>
<tr>
<td></td>
<td>(1,0)</td>
<td>0.890</td>
<td>0.033</td>
<td>49.59</td>
<td>39.43</td>
<td>63.41</td>
</tr>
<tr>
<td></td>
<td>(1,1)</td>
<td>0.788</td>
<td>0.005</td>
<td>39.40</td>
<td>39.43</td>
<td>9.09</td>
</tr>
<tr>
<td></td>
<td>(1.5,1)</td>
<td>0.873</td>
<td>0.009</td>
<td>47.90</td>
<td>39.43</td>
<td>16.98</td>
</tr>
<tr>
<td></td>
<td>(1.5,1.5)</td>
<td>0.788</td>
<td>0.005</td>
<td>39.40</td>
<td>39.43</td>
<td>9.09</td>
</tr>
<tr>
<td>$B$</td>
<td>(0,0)</td>
<td>0.786</td>
<td>0.025</td>
<td>39.26</td>
<td>39.29</td>
<td>9.05</td>
</tr>
<tr>
<td></td>
<td>(1,0)</td>
<td>0.786</td>
<td>0.057</td>
<td>49.33</td>
<td>39.29</td>
<td>62.81</td>
</tr>
<tr>
<td></td>
<td>(1,1)</td>
<td>0.786</td>
<td>0.004</td>
<td>0</td>
<td>39.29</td>
<td>8.33</td>
</tr>
<tr>
<td></td>
<td>(1.5,1)</td>
<td>0.786</td>
<td>0.004</td>
<td>0</td>
<td>39.29</td>
<td>8.33</td>
</tr>
<tr>
<td></td>
<td>(1.5,1.5)</td>
<td>0.786</td>
<td>0.004</td>
<td>0</td>
<td>39.29</td>
<td>8.33</td>
</tr>
</tbody>
</table>
The needed sample size for B

- The needed sample size depends on a sufficient sample size for $S^*$ and a sufficient sample size for $T$
  - The sample size of 200 per group was calculated based on the effect size of 1.5 for the gene set $S$.
  - The effect size for identifying the biomarkers $T$ is $\Delta = p \delta_S + (1-p) \delta_{NS}$.
  - Since $p$ is small, $\Delta$ will be small if $\delta_{NS}$ is small.
- For $(\delta_S, \delta_{NS}) = (1, 0)$ the needed sample size
  - To eliminate 95% of markers in A and C1 from $S^*$ is 16235
  - To eliminate 80% is 10906 per group
- The use of 200 per group reduced the performance accuracy.
Class-Imbalanced Prediction

For sample size of 200 per group at the given $p = 0.05$, $\delta = 1.5$, $u_1 = 0.7$ and $u_2 = 0.001$, $\lambda = 0.8$, $\alpha = 0.005$

- The expected number of the positive subjects is about 7 and the expected number of the negative subjects is about 193.
- The number of negatives is much larger than the number of positives.

Development of a classifier, in which class of interest is relatively small as compared to the other classes, is known as the class-imbalance prediction.
Imbalanced Data

Imbalanced data: the class sizes are different substantially.

Ex: clinical diagnostic test of rare diseases, fraud detection. (Much more negative data than positive data)

If P:N has the 1:9 ratio, a procedure predicts all negatives will have 90% accuracy, but 0% sensitivity.

A procedure with a high specificity but very low sensitivity, or vice versa, is not useful in some applications.

- Clinical diagnostic tests: high sensitivity.
- Epidemiology screening tests: high specificity.
Problems in Class-imbalance Data

Fundamental issues:

1. Imbalance ratio
2. Domain complexity
3. Small disjuncts
4. Feature selection

The standard classifiers can result in a low accuracy on the minority class prediction with a high accuracy on the majority class.
Separation of the Distributions of Two Classes
Sufficient Number of Minority Data

Separating Hyperplane
Classification of imbalanced data using the SVM

<table>
<thead>
<tr>
<th>Data</th>
<th># + : # -</th>
<th>ACC</th>
<th>SN</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene imprinted</td>
<td>43 : 88</td>
<td>84.4</td>
<td>69.8</td>
<td>91.6</td>
</tr>
<tr>
<td>Liver tumor</td>
<td>282 : 714</td>
<td>72.3</td>
<td>8.5</td>
<td>97.4</td>
</tr>
<tr>
<td>Estrogen activity</td>
<td>131 : 101</td>
<td>81.1</td>
<td>87.0</td>
<td>73.4</td>
</tr>
</tbody>
</table>

**Correction Strategies:**

*Algorithm-based approach* modifies the standard classification algorithm to account for class imbalance. e.g, logistic model.

*Data-based approach* uses the sampling technique to account for class imbalance without modifying a classification algorithm.
If \( f(x) \) is the decision function of SVM for given a sample \( X \).

A sample \( x \) will be assigned to the positive class if \( f(x) \geq 0 \), to the negative class if \( f(x) < 0 \).

An adjusted threshold:

\[
\theta = \frac{n_+ - n_-}{n_+ + n_- + 2a}
\]

where \( n_+ \) and \( n_- \) are the class sizes, and \( a \) is a constant.
An Ensemble Classifier

Re-sampling subsets of majority class

Chen et al. (2005) SAR and QSAR
Three Classification Algorithms

- Diagonal linear discriminant analysis with between-within variance ratio (DLDA-BW) - The BW ratio is used as part of DLDA for feature selection to develop classifiers
- Random forest with mean decrease in accuracy (RF-MDA)
- Support vector machine with recursive feature elimination (SVM-RFE)
  (Both RFE and MDA are built in feature selection algorithm).

- Each algorithm represents a unique characteristic for classifying high-dimensional class imbalanced data.
- The three classifiers are used with an ensemble data-based correction strategy.
Performance for the biomarkers $S^*$ from DLDA, SVM, and RF.

<table>
<thead>
<tr>
<th></th>
<th>(δₘ,δₙₛ)</th>
<th>2000 markers</th>
<th>Standard S*</th>
<th>Ensemble S*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SN</td>
<td>SP</td>
<td>SN</td>
</tr>
<tr>
<td>DLDA</td>
<td>(0,0)</td>
<td>0.47</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>(1,0)</td>
<td>0.20</td>
<td>0.99</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>(1,1)</td>
<td>0.25</td>
<td>0.98</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>(1.5,1)</td>
<td>0.08</td>
<td>0.99</td>
<td>0.27</td>
</tr>
<tr>
<td>SVM</td>
<td>(0,0)</td>
<td>0.67</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>(1,0)</td>
<td>0.03</td>
<td>1.00</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>(1,1)</td>
<td>0.15</td>
<td>1.00</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>(1.5,1)</td>
<td>0.01</td>
<td>1.00</td>
<td>0.03</td>
</tr>
<tr>
<td>RF</td>
<td>(0,0)</td>
<td>0.01</td>
<td>1.00</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>(1,0)</td>
<td>0.01</td>
<td>1.00</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(1,1)</td>
<td>0.01</td>
<td>1.00</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(1.5,1)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Performance for the biomarkers B from DLDA, SVM, and RF.

<table>
<thead>
<tr>
<th>(δ_S, δ_NS)</th>
<th>Standard with B</th>
<th>Ensemble with B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SN</td>
<td>SP</td>
</tr>
<tr>
<td>DLDA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0,0)</td>
<td>0.96</td>
<td>0.99</td>
</tr>
<tr>
<td>(1,0)</td>
<td>0.62</td>
<td>1.00</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0.73</td>
<td>1.00</td>
</tr>
<tr>
<td>(1.5,1)</td>
<td>0.27</td>
<td>1.00</td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0,0)</td>
<td>0.69</td>
<td>1.00</td>
</tr>
<tr>
<td>(1,0)</td>
<td>0.08</td>
<td>1.00</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0.68</td>
<td>0.99</td>
</tr>
<tr>
<td>(1.5,1)</td>
<td>0.68</td>
<td>0.99</td>
</tr>
<tr>
<td>RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0,0)</td>
<td>0.15</td>
<td>1.00</td>
</tr>
<tr>
<td>(1,0)</td>
<td>0.01</td>
<td>1.00</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0.18</td>
<td>1.00</td>
</tr>
<tr>
<td>(1.5,1)</td>
<td>0.18</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Simulation - Summary

- The number of negatives is much larger than the number of positives; it is expected to have high sensitivity and low specificity, and can result in high accuracy in all models.

- Feature selection does not improve the performance of DLDA substantially, compared to the estimates of using all 2000 variables without selection.

- There was little improvement for SVM and RF since the SVM and RF algorithms implicitly performed variable selection; but, many selected variables were not useful for improvement because of imbalanced class sizes.

- The sensitivities of the models developed from B were much higher than the sensitivities developed from S*.
Summary – Pharmacogenomics

- Populations of patients are heterogeneous due to differences in genetic pre-dispositions, lifestyle, or disease characteristics.

- An aim of pharmacogenomics is to identify inter-individual variability with respect to drug or treatment response (both efficacy and toxicity) to guide personalized medicine.

- An initial step to address the population heterogeneity is to assume an existence of a small fraction of susceptible population that can be represented by a set of genomic characteristics – biomarker of susceptibility.
Summary – Biomarker of Susceptibility

- Biomarkers of susceptibility indicate individual differences prior to the treatment.

- Biomarkers of susceptibility can be of two sets:
  1. different before treatment and changed by treatment
  2. different before treatment, and unchanged by treatment. These are the most useful biomarkers.
Identification of susceptible patients involves two steps:

1. To identify a set of biomarkers of susceptibility from a mixture of biomarkers of susceptibility and response.

2. To develop a classifier to discriminate the susceptible patients from the non-susceptible patients.

The number of susceptible individuals is much smaller than the number of non-susceptible; the classifier should take imbalanced class sizes into consideration.

- The ensemble method was shown to perform well for identifying the susceptible individuals.
Thank You
Paired Design

- In the paired design, the genomic samples of subjects, such as blood or cell tissues, can be tested prior to the drug treatment (and tested again after the treatment, if needed).
  - In paired design $S^*$ can be directly used to identify susceptible patients since the genomic data are collected before the treatment ($S^* = S$).

- The paired designs can be used to determine if cancer patients should receive postoperative adjuvant chemotherapy or not, where the tumor issues are obtained from the surgery.
Accuracy, Sensitivity, & Specificity from the logistic regression with the threshold $\tau$ (0.1-0.9).

As the decision threshold increases, sensitivity decreases and specificity increases. However, the accuracy stays more or less constant.