Hindawi Publishing Corporation
Disease Markers
Volume 35 (2013), Issue 6, Pages 661–667
http://dx.doi.org/10.1155/2013/393020

Research Article

Sex-Specific Genomic Biomarkers for Individualized Treatment of Life-Threatening Diseases

Hojin Moon,[1] Karen L. Lopez,[1] Grace I. Lin,[2] and James J. Chen[3,4]

[1] *Department of Mathematics and Statistics, California State University, 1250 Bellflower Boulevard, Long Beach, CA 90840-1001, USA*
[2] *Department of Computer Science, University of California, Santa Cruz, CA 95064, USA*
[3] *Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, FDA, Jefferson, AR 72079, USA*
[4] *Graduate Institute of Biostatistics and Biostatistics Center, China Medical University, Taichung, Taiwan*

Correspondence should be addressed to Hojin Moon; hojin.moon@csulb.edu

Received 30 June 2013; Revised 7 October 2013; Accepted 20 October 2013

Academic Editor: Kishore Chaudhry

Numerous studies have demonstrated sex differences in drug reactions to the same drug treatment, steering away from the traditional view of one-size-fits-all medicine. A premise of this study is that the sex of a patient influences difference in disease characteristics and risk factors. In this study, we intend to exploit and to obtain better sex-specific biomarkers from gene-expression data. We propose a procedure to isolate a set of important genes as sex-specific genomic biomarkers, which may enable more effective patient treatment. A set of sex-specific genes is obtained by a variable importance ranking using a combination of cross-validation methods. The proposed procedure is applied to three gene-expression datasets.

1. Introduction

Personalized medicine is defined by the use of genomic signatures of patients to assign effective therapies in order to achieve the best medical outcomes for individual patients, thus improving public health. Despite the variety of clinical, morphological, and molecular parameters used to classify human malignancies, patients receiving the same diagnosis can have markedly different clinical courses and treatment responses. Since there is no simple way to determine who will have an adverse reaction, the current system of "one-size-fits-all-" diagnoses is simply not good enough.

An increasing number of studies have demonstrated sex differences in drug reactions to the same drug treatment. Migeon [1] implied that males and females responded differently to drug treatments and that sex plays a key role in cancer. In addition, females are historically less studied subjects due to the complication of estrous cycle, and therefore such studies would further benefit women's health and promote public health.

Recent advancements in biotechnology have accelerated the search for molecular biomarkers useful in the diagnosis and treatment of disease. Molecular biomarkers of disease risk and status are critical to an accurate treatment by identifying patients most likely to benefit from particular drugs or experience adverse reactions. Because medicine is always practiced on individuals rather than populations, the goal is to change the assignment of therapies from a population-based approach to an individualized approach.

Gene-expression data can be used to identify patients with a good disease prognosis, thereby preventing some patients from unnecessary therapies and toxicity. For example, gene-expression profiling was used to predict clinical outcomes in pediatric patients with acute myeloid leukemia and to find genes whose aberrant expression leads to a poor prognosis [2]. Thus, accurate classification of prognosis of patients leads to efficient cancer treatment and prolonged survival of patients.

Classification algorithms are needed for biomedical decision making in clinical assignment of patients to treatment therapies based on individual risk factors and disease characteristics. Since many of those genes are not relevant, feature selection is a commonly addressed problem in classification [3, 4]. The goal of gene selection is to identify a set of genes which plays an important role in the classification.

A common approach is to select a fixed number of the highest ranked genes based on t-test-like statistics [5], some discrimination measures [6, 7], or classification algorithms including support vector machines (SVM) [8, 9] and random forest (RF) [10, 11].

Development of a biomarker classifier involves two distinct components: (1) a procedure for building a classifier and (2) validation/evaluation procedure for estimating the error rate of biomarker. The most important consideration is the validation/evaluation of a biomarker classifier to assess whether it can accurately and unbiasedly predict new samples based on a set of selected features. Two methods are commonly used to develop and assess performance of a classifier: the split-sample procedure and cross-validation procedure. In the split-sample procedure, the sample dataset is randomly split into two subsets: a training set for model building and a test set for performance assessment. That is, the training dataset is used for building the biomarker classifiers and the test set is used for evaluating the classifier. However, samples are often insufficient to split into two sample sets of approximately equal size for model building and model testing, and a cross-validation (CV) procedure is commonly used for performance assessment [12, 13]. A CV can be regarded as a generalization of the split-sample method. It involves repeatedly splitting the data into a training set containing most of the samples and applying the prediction rule to the test set of the remaining samples to estimate the prediction accuracy rate. The prediction accuracy is the average accuracy of the numerous training-test partitions.

For model building, highly accurate classification algorithms can be used to find sex-specific biomarkers. Given that differences in the biology of lung cancer and other diseases exist between men and women [14, 15], investigations to identify genomic biomarkers for clinical assignment of therapies on an individual patient basis are crucial. In this paper, the ratio of between-group to within-group sums of squares (BW ratio, [6]) gene selection algorithm is used to obtain a feasible set of influential genes via variable importance ranking procedure [12] in the training phases of 20 trials of 10-fold CV within leave-one-out cross-validation (LOOCV) procedure as illustrated below. The genes with largest BW ratios are ranked high as significant genes.

In the development of biomarker classifier, the predictive accuracy of the classifier must be evaluated on a separate set of data. To derive an unbiased accuracy estimate, a nested cross-validation procedure, 20 trials of 10-fold CV within LOOCV, is used in this paper. In other words, in each LOOCV, 90% of the $(n - 1)$ observations instead of 90% of n observations are randomly selected without replacement as a learning set. Wherefore genes are selected only using learning data sets of size $(n - 1)$ each, and in the evaluation/validation step a never-touched and left-out single observation in LOOCV is used to assess the predictive performance of selected genes. In this way, each test case is never used for gene selection. To avoid a bias due to partitions of data, twenty trials are used instead of a single trial. On the other hand, if the performance is assessed using the very same data that are used for developing the classifier, this obviously leads to a biased down estimate of classification error. In other words, if one applies a method to the original data with 20 trials of 10-fold CV only as a way of building the classifier and selects a set of genes and then on that very same data calculates the classification error, it clearly leads to a biased upward estimate of classification accuracy.

Three publicly available data sets of interest in this paper are pediatric acute myeloid leukemia (AML) [2], B-cell chronic lymphocytic leukemia (B-CLL) [15], and primary cutaneous melanoma [16]. The data sets are downloaded from the BRB-ArrayTools Data Archive for Human Cancer Gene Expression located at the website http://linus.nci.nih.gov/~brb/DataArchive_New.html.

2. Method for Identifying Sex-Specific Predictive/Prognostic Genomic Biomarkers

Sex differences in disease rates or in rates of adverse reactions to treatment are common, which we intend to exploit to obtain sex-specific biomarkers from gene-expression data. We hypothesize that genomic biomarkers developed from the sex-specific application of classification algorithms will further improve class prediction accuracy.

The summary of our algorithm to find sex-specific predictive/prognostic genomic biomarkers for efficacy or toxicity in individualized treatment of patients for serious diseases is as follows. In each LOOCV trial, firstly, the data is partitioned into a test data set with one observation and the remaining data as the learning data set. The learning data set is further separated into male and female patients' learning data sets. Thus, this process will be applied n times, where n represents the total number of patients. Secondly, within each trial of LOOCV, twenty trials of 10-fold CV are conducted for each set of male and female patients in the learning data set. At each 10-fold CV step, two sets of top-ranked genes, one for male patients (S_{M_i}) and one for female patients (S_{F_i}), are obtained via the BW ratio to differentiate types of patients such as diseased versus nondiseased. This can be done by applying variable importance ranking approach in order to extract most influential genes and by combining a measure of importance in each gene. A score taking the average of the measure in cross-validation (CV) is prioritized in the list of genes according to variable importance. Thirdly, the mutually exclusive sets of male-specific genes and female-specific genes are obtained. Fourthly, within each LOOCV trial, after the sets of potential sex-specific genes are determined in the second and the third steps, they are fitted to a model with diagonal linear discriminant analysis (DLDA; [6]) using each learning data set for male and female patients. Finally, the performance is measured with the test data with one observation in each LOOCV trial. This method is implemented in R. The R code is available upon request.

Within each trial of LOOCV, each set of top-ranked genes for male and female patients is obtained in the second and the third steps as follows. First, in order to build genomic biomarkers, 20 trials of 10-fold CV are applied to each learning data set for males and females, where 90% were randomly selected without replacement as a set for each trial of CV. Next, for each trial of 10-fold CV, the BW ratio was

applied to this 90 percent learning set and the top 25 ranked genes were selected in each process with the target endpoint of a dataset. The BW ratio for a gene j is defined as

$$\mathrm{BW}_j = \frac{\sum_i \sum_k I\left(y_i = k\right)\left(\overline{x}_{kj} - \overline{x}_{\cdot j}\right)^2}{\sum_i \sum_k I\left(y_i = k\right)\left(x_{ij} - \overline{x}_{kj}\right)^2}, \quad (1)$$

where $\overline{x}_{\cdot j} = \sum_i x_{ij}/N$ indicates the average value of a gene j across all the training samples, $\overline{x}_{kj} = \sum_{I(y_i=k)} x_{ij}/N_k$ indicates the average value of a gene j for a class k, and i indicates an observation. Here, $I(\cdot)$ is the indicator function. This criterion has been shown to be reliable for variable selection from high-dimensional data sets [6, 12].

Next, to avoid selection bias from a pattern of selection of learning samples, we repeat the entire process 20 times by shuffling samples at every 10-fold CV. In order to obtain the variable importance ranking, 200 sets of top 25 ranked genes were combined so that the maximum possible rank score of a gene would be 200. A set of genes that has been selected at least once (rank score > 0) was obtained separately for males and females. These most influential genes used in the classification process are identified in order to extract a feasible set of sex-specific genes. For both male and female learning data sets, a set of genes scored greater than 0 are selected for males and females. The final product of the 20 trials of 10-fold CV described in each LOOCV is the sets of genes selected for male and female patients in the learning data set $S_{M_j} = \{g_1, \ldots, g_{x_{M_j}}\}$ and $S_{F_j} = \{g_1, \ldots, g_{x_{F_j}}\}$, $j = 1, 2, \ldots, n$, respectively. The sets S_{M_j} and S_{F_j} contain prognostic/predictive genes for male and female patients, respectively. There would be n sets of selected genes. Within each trial of LOOCV, genes that are commonly identified between male-specific and female-specific genes are removed for each set of sex-specific genes. Finally, the test sample with one observation is tested using these sex-specific genes in each LOOCV. At the end of the entire LOOCV, each selected gene has a combined variable importance score that is less than or equal to n.

To verify sex-specific biomarkers, we consider the following four different cases: we classify the outcome of (1) male patients with a set of male-specific genes S_{M_j},(2) female patients with a set of male-specific genes S_{M_j},(3) male patients with a set of female-specific genes S_{F_j}, and (4) female patients with a set of female-specific genes S_{F_j}. We anticipate that data with a set of male-specific genes have higher predictable power to predict male patients than the data with female-specific genes. Similarly, data with female-specific genes have higher predictable power to predict female patients than the data with male-specific genes. Upon completion of LOOCV, the performance of sex-specific biomarkers is obtained.

In order to validly evaluate the performance of a gene set selected by the proposed method, twenty trials of 10-fold CV within LOOCV are used. CV utilizes resampling without replacement of the entire data set to repeatedly develop classifiers on a training set and to evaluate these classifiers on a separate test set and then averages the results over the resamples.

3. Results: Identification and Evaluation of Predictive/Prognostic Genomic Biomarkers

In this section, we apply the proposed algorithm to the following genomic data sets to find the most meaningful sex-specific predictive/prognostic genomic biomarkers for improving individualized treatment of patients and for evaluating the biomarkers from the proposed algorithm.

3.1. Prognostic Biomarkers Associated with Pediatric Acute Myeloid Leukemia (AML) Prognosis. Current chemotherapy enables a high percentage of pediatric patients with AML to enter complete remission (CR), but a large number of them experience relapse (R) with resistant disease [2]. Because of the wide heterogeneity of AML, predicting a patient's risk for treatment failure or relapse at the time of diagnosis is critical for the optimal treatment.

This gene-expression data set consists of 54 AML pediatric patients (<15 years old) with an oligonucleotide microarray containing 12,566 probe sets and it is also available at ftp://ftp.ncbi.nih.gov/pub/geo/DATA/SOFT/GDS/GDS1059.soft.gz. Patients with CR for more than 3 years are classified as having a good prognosis, while patients experienced relapse within 1 year of the first CR are considered as having a poor prognosis. In this data set, there are 28 patients with CR and 25 patients with R. Since a five-month male patient experienced induction failure (no CR achievement) within 3 months of the start of treatment is considered neither CR nor R, we exclude this subject. Therefore, there are 32 male patients and 21 female patients in this data set.

With the prognostic endpoint (R/CR), the average accuracy of 66% (sd 3.0%) for pediatric patient classification was obtained when no gene selection was introduced. When a set of 200 genes was selected in a learning phase of each CV, the average accuracy of 68.0% (sd 3.0%) was achieved. When a set of 20 genes was selected in a learning phase of each CV, the average accuracy of 71.0% (sd 5.0%) was obtained. Since it appeared to be no substantial difference between accuracy and the number of genes selected, a feasible set of 25 genes in the learning phase of each CV was collectively ranked by the BW ratio to find sex-specific biomarkers.

At the end of LOOCV trials, a set of male-specific genes that were selected at least 75% of the time in the entire LOOCV were 1882_g_at (MECOM: MDS1 and EVI1 complex locus), 37902_at (CRYZ: crystallin, zeta (quinone reductase)), 38789_at (TKT: transketolase (Wernicke-Korsakoff syndrome)), 39105_at (VASP: vasodilator-stimulated phosphoprotein), 40844_at (CTR9: Ctr9, Paf1/RNA polymerase II complex component, homolog (*S. cerevisiae*)), 36981_at (SRP9: signal recognition particle 9 kDa), 36338_at (LUZP1: leucine zipper protein 1), 31870_at (CD37: CD37 antigen), 1624_at (RAP1GDS1: RAP1, GTP-GDP dissociation stimulator 1), and 39142_at (NUDT21: cleavage and polyadenylation specific factor 5, 25 kDa). These ten top-ranked male-specific genes were selected as potential male-specific genomic biomarkers to classify male patients into R/CR. Among them

38789_at (TKT: Transketolase) and 36338_at (EST) were also included in the thirty-five genes associated with prognosis of pediatric AML identified by Yagi et al. [2].

In order to select a feasible set of sex-specific biomarkers a cut-off criterion of 75 percent is used. Since every dataset may have a different sample size, the number of genes selected is given by the percentage, which is a rank score of the selected genes greater than 75% of the sample size in our case. For example, if a sample size is 100, then genes that have rank scores greater than 75 have been selected. The LOOCV depends on the sample size n and repeats n times.

Similarly, nine top-ranked genes for classifying female patients into R/CR were 40601_at (TM2D1: TM2 domain containing 1), 36330_at (CCBL1: cysteine conjugate beta-lyase; cytoplasmic (glutamine transaminase K, kynurenine aminotransferase)), 40586_at (EEF1E1: eukaryotic translation elongation factor 1 epsilon 1), 36648_at (CRSP9: cofactor required for Sp1 transcriptional activation, subunit 9, 33 kDa), 32351_at (GPR20: G protein-coupled receptor 20), 1718_at (ARPC2: actin-related protein 2/3 complex, subunit 2, 34 kDa), 38622_at (MTG1: mitochondrial GTPase 1 homolog (*S. cerevisiae*)), 36496_at (IMPA2: inositol(myo)-1(or 4)-monophosphatase 2), and 38337_at (ZNF193: zinc finger protein 193). These nine top-ranked genes which were selected at least 75% of the time in the entire LOOCV were considered as potential female-specific genomic biomarkers to classify female patients into R/CR.

For the verification of sex-specific genes we applied DLDA classification algorithm. The result showed the following outcomes as expected. Data with male-specific genes showed higher prediction accuracy (71.9%) to classify male patients than the accuracy (43.8%) to classify male patients from data with female-specific genes. Similarly, data with female-specific genes showed higher prediction accuracy (76.2%) to classify female patients than the accuracy (61.9%) to classify female patients from data with male-specific genes. As shown in Table 1, sensitivity and specificity were also higher in using sex-specific genes for each sex.

3.2. Prognostic Biomarkers Associated with B-Cell Chronic Lymphocytic Leukemia (B-CLL). Genomic aberrations and mutational status of the immunoglobulin variable heavy chain (VH) gene have been shown to be among the most important predictors for outcome in patients with B-CLL [15]. B-CLL is the most common leukemia in the Western world, and due to its clinical heterogeneity (wide range of life expectancy) and correlations to genomic aberrations, it is important to predict a patient's VH mutation status at the time of diagnosis for optimal treatment. In addition, the study presented by Haslinger et al. [15] suggested that the genomic signature for VH mutational status might be sex related.

The gene-expression data consists of 100 B-CLL patients with an oligonucleotide microarray containing around 12,000 probe sets, and it is available at http://linus.nci.nih.gov/~brb/DataArchive_New.html. Patients are classified as either VH-mutated or unmutated (M/NM). There are 62 males (33 M and 29 NM) and 38 females (18 M and 20 NM), with a total of 51 mutated and 49 unmutated patients.

In Step 1, for each data set with male only and female only patients using the target endpoint (i.e., VH-mutated (M) versus unmutated (NM)), we separately selected and ranked 25 potential prognostic genes for males and for females in every CV trial and separately combined ranks of these genes for males and females during the learning phase of 20 trials of 10-fold CV within each LOOCV trial. In every LOOCV trial, we prioritized and combined the final top-ranked 200 genes from the male patients result (S_M) and 200 genes from the female patients result (S_F) as explained in Section 3.1. At the end of the entire LOOCV trials, a set of potential sex-specific genes are obtained for each sex by prioritizing and combining n sets of top-ranked 200 genes.

After deletion of overlapped genes in both males and females, eleven potential male-specific genes were obtained to classify male patients into M/NM. They were 41209_at (LPL: lipoprotein lipase), 41755_at (COBLL1: COBL-like 1), 39878_at (PCDH9: protocadherin 9), 38211_at (ZBTB20: zinc finger and BTB domain containing 20), 39488_at (PCDH9: protocadherin 9), 36886_f_at (KIR2DL3: killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 3), 32140_at (SORL1: sortilin-related receptor, L(DLR class) A repeats-containing), 33535_at (P2RX1: purinergic receptor P2X, ligand-gated ion channel, 1), 39967_at (LDOC1: leucine zipper, downregulated in cancer 1), 32842_at (BCL7A: B-cell CLL/lymphoma 7A), and 36899_at (SATB1: special AT-rich sequence binding protein 1 (binds to nuclear matrix/scaffold-associating DNA's)). For female patients, only five genes were selected at least 75% of the time in the entire LOOCV trial as potential female-specific genomic biomarkers to classify female patients into M/NM. They were 33745_at (PHKG2: phosphorylase kinase, gamma 2 (testis)), 38152_at (LOH11CR2A: loss of heterozygosity, 11, chromosomal region 2, gene A), 34142_at (PDE8A: phosphodiesterase 8A), 39593_at (FGL2: fibrinogen-like 2), and 217_at (KLK2: kallikrein 2, prostatic).

To verify the sex-specific genomic biomarkers, we considered the performance of four different cases in the LOOCV trials as described in Section 2. Data with male-specific genomic biomarkers showed higher accuracy (67.7%) to classify male patients into M/NM than the accuracy to classify male patients into M/NM from data with female-specific genomic biomarkers (40.3%). However, female data with female-specific genes showed prediction accuracy similar to random guess close to 50%. Therefore, there is insufficient evidence to support that the selected female-specific genes were female-specific genes (see Table 2) with 38 female patients in this data set.

3.3. Predictive Genomic Biomarkers Associated with a Classification of Primary Cutaneous Melanoma. Skin cancer is one of the most common malignancies in the United States. Although cutaneous melanoma represents a small subset, it is the most life-threatening neoplasm of the skin, and its incidence and mortality have been increasing worldwide. The key underlying molecular events have not been clearly elucidated, which may explain why no target has been developed and why almost no clinical benefits from new therapies have been

Table 1: Performance (%) of the sex-specific genes in pediatric AML data set. (ACC: accuracy; SEN: sensitivity; SPC: specificity; PPV: positive predictive value; NPV: negative predictive value).

	Patients	ACC	SEN	SPC	PPV	NPV
Male data with male genes	32	71.9	86.7	58.8	65.0	83.3
Male data with female genes	32	43.8	40.0	47.1	40.0	47.1
Female data with male genes	21	61.9	70.0	54.5	58.3	66.7
Female data with female genes	21	76.2	70.0	81.8	77.8	75.0

Table 2: Performance (%) of overlapped genes in the B-CLL data set. (ACC: accuracy; SEN: sensitivity; SPC: specificity; PPV: positive predictive value; NPV: negative predictive value).

	Patients	ACC	SEN	SPC	PPV	NPV
Male data with male genes	62	67.7	72.7	62.1	68.6	66.7
Male data with female genes	62	40.3	36.4	44.8	42.9	38.2
Female data with male genes	38	50.0	50.0	50.0	47.4	52.6
Female data with female genes	38	47.4	61.1	35.0	45.8	50.0

clearly demonstrated in patients with melanoma since the late 1970s [16]. Additionally, gene-expression profiling data for human primary cutaneous melanomas are scarce because of the lack of retrospective collections of frozen tumors [16].

This gene-expression data set was collected from 83 patients corresponding to the training data set and 17 patients corresponding to the validation data set. The data consists of approximately 37,000 probe sets with dual-channel oligonucleotide microarrays. The probes are from tumor tissue and from reference tissue that are differentially labeled by the incorporation of cyanine 3 (Cy3) and cyanine 5 (Cy5), respectively. The data is available at: http://linus.nci.nih.gov/data_archive/MelanomaEBI-E-TABM12-Project.zip.

In this data set, the endpoint was patient prognosis and survival along with tumor stages, defined as follows. In Stage I, cure rates are excellent with surgical removal, since they are the least likely to spread. In Stage II, melanomas can be cured, but the success rate lags behind that of Stage I because a small number of cancer cells may have spread to distant sites. In Stage III, since the tumor has started to metastasize (the spreading of a disease from one organ or part to another nonadjacent organ or part), the survival rate for these stages is lower than the earlier ones. Stage IV is associated with metastasis beyond the regional lymph nodes to distant sites in the body, such as the lung, liver, or brain, or to distant areas of the skin. Based on the tumor size, descriptions, and number of lymph nodes the stages are categorized in two classes. A class of high survival and small tumor size (HS/ST) is defined and composed of Stages 1 and 2. The second is defined as low survival and non-small tumor size (LS/NST), which is composed of Stages 3 and 4.

Tables 3 and 4 show the distribution of the primary cutaneous melanoma data based on sex, the defined clinical endpoint HS/ST and LS/NST for the training data and validation data. Among 83 patients in the training dataset, there are 27 males (12 HS/ST and 15 LS/NST) and 56 females (30 HS/ST and 26 LS/NST). There are 42 cases of HS/ST and 41 cases of LS/NST in total. After preprocessing the data the final gene count is 4641 genes. Among 17 patients in the validation data set, there are 8 males (1 HS/ST and 7 LS/NST) and 9 females (1 HS/ST and 8 LS/NST). There are 2 cases of HS/ST and 15 cases of LS/NST in total.

Since the validation set was separately provided, the sex-specific genes were selected via 20 trials of 10-fold CV in the learning set. The following ten male-specific genes were selected at least 75% of the time during 20 trials of 10-fold CV: A_23_P128263 (PRB1: proline-rich protein BstNI subfamily 1), A_23_P83838 (CA8: carbonic anhydrase VIII), A_24_P212990 (MGC70863: similar to RPL23AP7 protein), A_23_P333650 (RAD9B: RAD9 homolog B (*S. cerevisiae*)), A_32_P125251 (N/A), A_23_P108835 (YPEL5: yippee-like 5 (*Drosophila*)), A_32_P150856 (LOC407835: mitogen-activated protein kinase kinase 2 pseudogene), A_23_P11936 (UBXN11: UBX domain protein 11), A_24_P169976 (N/A), and A_32_P3998 (ZNF600: zinc finger protein 600). For female patients, the following eight female-specific genes were selected from 20 trials of 10-fold CV: A_23_P69497 (CLEC3B: C-type lectin domain family 3, member B), A_24_P118884 (N/A), A_23_P152420 (KIAA0182), A_24_P265177 (PHC3: polyhomeotic-like 3 (*Drosophila*)), A_23_P251421 (CDCA7: Cell division cycle associated 7), A_23_P211738 (UBP1: upstream binding protein 1 (LBP-1a)), A_23_P47377 (HSD17B12: hydroxysteroid (17-beta) dehydrogenase 12), and A_32_P113646 (CDNA FLJ45341 fis, clone BRHIP3009672).

Using a given validation set, the sex-specific genes were verified. As shown in the confusion matrix in Table 5, there was no difference in the performance from female data with female genes compared with the performance of female data with male genes. However, there was only one misclassification for male patients with male genes as shown in Table 6, while there were four misclassifications for male patients with female genes as shown in Table 7. Therefore, we conclude that there exists evidence of male-specific genes in a classification of primary cutaneous melanoma.

4. Discussion and Conclusions

Large inter individual differences in benefit from chemotherapy highlight the need to develop predictive genomic

Table 3: Melanoma training dataset distribution.

Gender (class)	Melanoma training set Clinical endpoint (abbrev., class)		Total patients	Final gene count
	High survival and small tumor size (HS/ST, "0")	Low survival and nonsmall tumor size (LS/NST, "1")		
Male ("0")	12	15	27	4641
Female ("1")	30	26	56	
Total	**42**	**41**	**83**	

Table 4: Melanoma validation dataset distribution.

Gender (class)	Melanoma validation set Clinical endpoint (abbrev., class)		Total patients
	High survival and small tumor size (HS/ST, "0")	Low survival and nonsmall tumor size (LS/NST, "1")	
Male ("0")	1	7	8
Female ("1")	1	8	9
Total	2	15	17

Table 5: Confusion matrix from female data with female genes and from female data with male genes using the validation set.

	Predicted	
	(HS/ST, "0")	(LS/NST, "1")
True class		
(HS/ST, "0")	0	1
(LS/NST, "1")	0	8

Table 6: Confusion matrix from male data with male genes using the validation set.

	Predicted	
	(HS/ST, "0")	(LS/NST, "1")
True class		
(HS/ST, "0")	0	1
(LS/NST, "1")	0	7

Table 7: Confusion matrix from male data with female genes using the validation set.

	Predicted	
	(HS/ST, "0")	(LS/NST, "1")
True class		
(HS/ST, "0")	0	1
(LS/NST, "1")	3	4

biomarkers for selecting the right treatment for the right patient. Inappropriate chemotherapy can result in the selection of more resistant and aggressive tumor cells. To date, no reliable genomic biomarkers have been developed to provide the physician with prechemotherapy information to accurately predict the efficacy of a specific therapy.

We proposed a procedure to find sex-specific prognostic and predictive genomic biomarkers in order to assign individualized treatments in a personalized paradigm using variable importance ranking via combination of 20 trials of 10-fold CV and LOOCV. The proposed procedure was applied to data sets obtained from the BRB ArrayTools Data Human Cancer Archive [17]. However, the issue arose out of there being not enough samples, and the data was unbalanced (i.e., more males than females or more positives (disease patients) than negatives (nondisease patients)). While patient's sex information was not specified for many of the publicly available data sets adding to that the difficulty of searching for data, we found the following three genomic data sets that had sex information: (1) pediatric patients with AML, (2) B-cell chronic lymphocytic leukemia (B-CLL), and (3) primary cutaneous melanoma.

In one application, pediatric patients with AML were classified by the algorithms as having either a good or poor prognosis, in terms of the likelihood of induction failure or relapse within one year of the first complete remission, based on gene-expression profiles. If this were brought into clinical application, a patient with a confidently predicted good prognosis might want to elect out of adjuvant chemotherapy and its associated debilitating side effects. With current rule-based decisions, almost all patients are subjected to chemotherapy. The overall average accuracy of this data set with a variable selection from pooled patients (males and females) was about 71.0%. However, using male-specific genes found by the proposed procedure, the accuracy was improved to about 72% as we found in the model validation studies. Similarly, using female-specific genes found by the proposed procedure, the average accuracy was improved to about 76% (see Table 1).

In the B-cell chronic lymphocytic leukemia (B-CLL) dataset we found male-specific prognostic genomic biomarkers associated with B-cell chronic lymphocytic leukemia and its average classification accuracy was improved to about 68%. There was no substantial evidence to find female-specific prognostic genomic biomarkers in this data set. Similarly, male-specific predictive genomic biomarkers associated with a classification of primary cutaneous melanoma were found with the classification accuracy of about 88%.

The scope of our paper was to find sex-specific genomic biomarkers, if any, imbedded in the data instead of finding genomic biomarkers from the data. If commonly identified genes were kept in the proposed procedure, our procedure was not sex-specific genomic biomarker classifier involving two populations (males and females) any more but rather it became genomic biomarker classifier involving one combined population. In fact, even though commonly identified genes were kept, it did not necessarily improve the classification accuracy. For a counterexample, for the pediatric AML data of Yagi et al. [2], the accuracy for female patients with female-specific genes by keeping the commonly identified genes was 62%; however, the accuracy without commonly identified genes was 76%. For the male patients with male-specific genes by keeping the commonly identified genes the accuracy was 59%, but the accuracy without the commonly identified genes was 72%. For the related note, the accuracy of this data can be found anywhere between 59% and 66% with gene preprocessing and between 58% and 71% with gene selection [12].

It is not an easy task to find sex-specific genes, let alone verifying and proving that they are indeed sexspecific. We have presented a procedure for finding sex-specific prognostic and predictive genomic biomarkers in order to assign individualized treatments in a personalized paradigm. The procedure is shown to have good "sensitivity" and "specificity" in the sense that the sex-specific genes obtained can improve prediction accuracy in classification of individual patient's prognosis. The proposed procedure to discover predictive and prognostic sex-specific genomic biomarkers for individualized treatment of diseases can play a critical role in developing safer and more effective therapies that replace one-size-fits-all drugs with treatments that focus on specific patient needs.

Acknowledgments

Hojin Moon's research was partially supported by the Research, Scholarship, and Creative Activity (RSCA) Award from California State University, Long Beach, and was partially supported by the Faculty Research Participation Program at the NCTR administered by the Oak Ridge Institute for Science and Education through an interagency agreement between USDOE and USFDA. The views presented in this paper are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration.

References

[1] B. R. Migeon, "Why females are mosaics, x-chromosome inactivation, and sex differences in disease," *Gender Medicine*, vol. 4, no. 2, pp. 97–105, 2007.

[2] T. Yagi, A. Morimoto, M. Eguchi et al., "Identification of a gene expression signature associated with pediatric AML prognosis," *Blood*, vol. 102, no. 5, pp. 1849–1856, 2003.

[3] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[4] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[5] H. Ahn and H. Moon, "Classification: supervised learning with high-dimensional biological data," in *Statistical Bioinformatics, A Guide for Life and Biomedical Science Researchers*, J. K. Lee, Ed., Chapter 8, John Wiley & Sons, Chichester, UK, 2009.

[6] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–86, 2002.

[7] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Informatics Series*, vol. 13, pp. 51–60, 2002.

[8] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[9] C.-A. N. Tsai, C.-H. Chen, T.-C. Lee, I.-C. Ho, U.-C. Yang, and J. J. Chen, "Gene selection for sample classifications in microarray experiments," *DNA and Cell Biology*, vol. 23, no. 10, pp. 607–614, 2004.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, article 3, 2006.

[12] S. Baek, H. Moon, H. Ahn, R. L. Kodell, C.-J. Lin, and J. J. Chen, "Identifying high-dimensional biomarkers for personalized medicine via variable importance ranking," *Journal of Biopharmaceutical Statistics*, vol. 18, no. 5, pp. 853–868, 2008.

[13] S. Baek, C.-A. Tsai, and J. J. Chen, "Development of biomarker classifiers from high-dimensional data," *Briefings in Bioinformatics*, vol. 10, no. 5, pp. 537–546, 2009.

[14] J. D. Patel, P. B. Bach, and M. G. Kris, "Lung cancer in us women: a contemporary epidemic," *Journal of the American Medical Association*, vol. 291, no. 14, pp. 1763–1768, 2004.

[15] C. Haslinger, N. Schweifer, S. Stilgenbauer et al., "Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status," *Journal of Clinical Oncology*, vol. 22, no. 19, pp. 3937–3949, 2004.

[16] J. Homsi, M. Kashani-Sabet, J. L. Messina, and A. Daud, "Cutaneous melanoma: prognostic factors," *Cancer Control*, vol. 12, no. 4, pp. 223–229, 2005.

[17] Y. Zhao and R. Simon, "BRB-ArrayTools Data Archive for human cancer gene expression: a unique and efficient data sharing resource," *Cancer Informatics*, vol. 6, pp. 9–15, 2008.