Supplementary materials for

Reporting bias when using real data sets to analyze classification performance

Mohammadmahdi R. Yousefi[1], Jianping Hua[2], Chao Sima[2], Edward R. Dougherty[1,2]

August 4, 2009

Real data

A collection of real data sets from twelve microarray experiments, with sample size larger than 150, are used for this study. We have tried to maintain the original labeling and also to follow the data preparing directions used in the papers reporting these data sets; however, in several cases we have re-labeled samples for reasons to be given in the following descriptions of the data sets:

- **Lung Cancer [Bhattacharjee *et al.*, 2001], 203 sample points, 12,600 features**

  This data set has been obtained from a total of 203 snap-frozen specimens composed of 186 lung tumors and 17 normal lung samples. Lung tumors include 139 adenocarcinomas, 21 squamous cell lung carcinomas, 20 pulmonary carcinoids, and 6 small-cell lung carcinomas (SCLC) (See Table 1). In this experiment, mRNA expression levels of 12,600 transcript sequences from samples are hybridized to human U95A oligonucleotide probe arrays (Affymetrix, Santa Clara, CA) for analysis and for providing evidence for biologically distinct subclasses of lung adenocarcinoma [Bhattacharjee *et al.*, 2001].

  We split the samples into two classes: adenocarcinomas and the remaining four groups. Since training sample size is 60, to preserve the sample ratio the training sample size of class adenocarcinoma equals 41.

Table 1: Lung cancer

Cancer Class	Number of Samples
Adenocarcinomas	139
Normal	17
Squamous Cell Lung Carcinomas	21
Pulmonary Carcinoids	20
Small-Cell Lung Carcinomas	6

- **Different Tumors [Su *et al.*, 2001], 174 sample points, 12,533 features**

  This data set contains samples from 11 different tumor cells: 27 serous papillary ovarian adenocarcinomas, 8 bladder/ureter carcinomas, 26 infiltrating ductal breast adenocarcinomas, 23 colorectal adenocarcinomas, 12 gastroesophageal adenocarcinomas, 11 clear cell carcinomas of the kidney, 7 hepatocellular carcinomas, 26 prostate adenocarcinomas, 6 pancreatic adenocarcinomas, 14 lung adenocarcinomas carcinomas, and 14 lung squamous carcinomas. Specimens in this set were assessed by H&E frozen section examination. Before doing RNA extraction, areas of specimens rich in tumor were cut from the frozen blocks. The samples consisted predominantly of neoplastic cells based on some special preparation policies. RNA extraction and hybridization on oligonucleotide microarrays (U95a GeneChip; Affymetrix Incorporated, Santa Clara, CA) was performed as described in [Su *et al.*, 2001].

The original paper [Su *et al.*, 2001] treated the problem as a multi-class classification problem. Since our study concerns binary classification, we aggregated 11 classes into two balanced classes: bladder/ureter, breast, colorectal and prostate in one class and remaining groups in the other class. Hence, the number of samples in the two classes are 83 and 91 (See Table 2).

Table 2: Different tumors

Cancer Class	Number of Samples
Bladder/ureter	8
Breast	26
Colorectal	23
Prostate	26
Ovary	27
Gastroesophagus	12
Kidney	11
Liver	7
Pancreas	6
Lung Adeno	14
Lung Squamous	14

- **Diffuse large-B-cell lymphoma [Rosenwald *et al.*, 2002], 203 sample points, 5013 features**

  This data set contains biopsy samples of diffuse large-B-cell lymphoma from 240 patients which were examined for gene expression with the use of DNA microarrays and analyzed for genomic abnormalities [Rosenwald *et al.*, 2002]. Lymphochip DNA microarrays were constructed from 12,196 clones of complementary DNA (i.e., microarray features) and were used to quantitate the expression of mRNA in the tumors. These arrays were composed of genes whose products are preferentially expressed in lymphoid cells and genes thought or confirmed to play a part in cancer or immune function [Rosenwald *et al.*, 2002].

  In the original study, germinal center B-cell like ($n = 115$), type 3 ($n = 52$), activated B-cell like ($n = 73$), were the results of clustering on data rather than a classification problem. Therefore, they were not the true clinical labels. It could be inferred from the paper that the label should be with survival outcome: alive/death. However, they were not quite sure about the censored data, i.e., patient was alive but the follow-up was less than 3 years. They were also worried about the effects of a previous study. In our study we just removed the censored data and used the following labels: alive for more than three years vs. death within three years. Consequently, we have 89 samples in the "alive" class, 114 in the "death" class, and 37 samples (censored data) not used.

  "Missing values" in the data set was another issue which we had to consider. The original paper approached this problem by excluding patients with missing values. Instead, we filled the missing values by the average value of each gene across all samples. Also, if the percentage of missing values in a particular feature was higher than a specific threshold (we considered 10 percent), we simply excluded that feature from the data set.

- **Primary breast carcinomas [van de Vijver *et al.*, 2002], 266 sample points, 5003 features**

  The breast cancer data from [van de Vijver *et al.*, 2002] contains 295 patients. All samples selected from the fresh-frozen-tissue bank of the Netherlands Cancer Institute were patients having stage I or II breast cancer (the tumor was primary invasive breast carcinoma being less than 5 cm in diameter at pathological examination) and they were younger than 53 years old; the only history of cancer in samples was nonmelanoma skin cancer; 151 had lymph-node–negative disease, and 144 had lymphnode–positive disease [van de Vijver *et al.*, 2002]. After isolation of RNA, labeling of complementary RNA (cRNA), hybridization of labeled cRNA to 25,000-gene arrays, and assessment of expression ratios, the slides were washed and scanned with a confocal laser scanner (Agilent Technologies). Quantification of fluorescence intensities on captured images was followed by correcting the values for the background level and also normalization. Arrays were combined from source files and were processed as described in [van't Veer *et al.*, 2002] which resulted in a final passing of 5003 genes.

  The patients were labeled into two groups: 180 patients for poor-prognosis signature group and 115 patients for good-prognosis signature, according to a classifier evaluating the correlation of each sample with the average profile of clinically "good" samples on 70 selected genes. Hence, the current class labels were based on the classifier of [van de Vijver *et al.*, 2002]. In our study, we could use the real clinical outcome as the patient label: patients selected either had distant metastases as a first event within five years or had remained free of disease for at least five years. Note that not all 295 patients would be included by this labeling method; there were patients having distant metastases but not as a first event, also patients free of disease but the follow up was shorter than five years to make any decision. Therefore, we would have 70 patients having distant metastases within five years, 196 patients disease-free for more than five years, and 29 patients which were excluded from labeling. Among 196 disease-free patients, 22 had distant metastases (18 first event) and 7 were metastases free but had other recurrences, later.

- **Pediatric acute lymphoblastic leukemia [Yeoh *et al.*, 2002], 248 sample points, 5077 features**

  The acute lymphoblastic leukemia (ALL) data set has been obtained from a study on pediatric acute lymphoblastic leukemia [Yeoh *et al.*, 2002]. ALL is a complicated disease containing several subtypes. Data points were labeled into six subtypes: T-ALL (43 sample points), E2A-PBX1 (27 sample points), TEL-AML1 (79 sample points), BCR-ABL (15 sample points), MLL (20 sample points), and hyperdiploid with >50 chromosomes (64 sample points). The data have been collected using Affymetrix's Human Genome HG_U95Av2 array (Santa Clara, CA) and are publicly available at http://www.stjuderesearch.org/data/ALL1. This microarray chip contains 12,000 probe sets (features).

  We have removed the features in which less than 1 percent of the sample points had a present call or more than 10 percent of the sample points had their values missing. This reduced the total feature size to 5077. The missing values were filled by averaging across all sample points. We labeled the data into two classes: one containing T-ALL, E2A-PBX1 and TEL-AML1 subtypes (149 sample points), the other containing BCR-ABL, MLL and hyperdiploid >50 subtypes (99 sample points). Hence, the total sample size is 248 [Hua *et al.*, 2009].

- **Hepatocellular carcinoma [Chen *et al.*, 2004], 157 sample points, 10,237 features**

  The original study was on hepatocellular carcinoma (HCC), which is the most common adult liver malignancy and ranks among the top five causes of cancer death in the world. The liver tissues (tumor and non-tumor) were obtained from surgical resections or transplants performed at Stanford University, CA, USA or Queen Mary Hospital, The University of Hong Kong, Hong Kong, China. Frozen liver samples were used for cDNA microarray study and reverse transcription-polymerase chain reaction (RT-PCR); and paraffin-embedded samples were used for IHC and in situ hybridization (ISH). Total RNA was extracted from tissues using Trizol Reagent (Invitrogen, Carlsbad, CA, USA). For microarray analysis, messenger RNA (mRNA) was isolated from total RNA using FastTract mRNA purification kit (Invitrogen) [Chen *et al.*, 2004].

In our study, the array data were obtained from the Stanford Microarray Database (http://genome-www.stanford.edu/microarray) which is labeled by 82 primary tumor tissues and 75 non-tumor tissues. First, we found well-measured genes in each sample (the fluorescent intensity in each channel was greater than 1.5 times the local background). This set might contain flagged genes and also genes which had missing values. We further assumed that flagged genes in this set were also missing values. Therefore, we had three different kinds of missing values: a) actual missing values in; b) bad-measured genes; c) flagged genes. Then, we removed genes which had 25 percent or more missing values across all samples and filled the missing values by the simple average of each gene in the reduced feature set. Doing this process resulted in reduction of the number of features from 24,168 to 10,237.

- **Acute myeloid leukemia [Valk *et al.*, 2004], 273 sample points, 22,215 features**

  Acute myeloid leukemia (AML) data set has been obtained from a study on the prognostic profiling of acute myeloid leukemia [Valk *et al.*, 2004]. The data and the associated clinical information are publicly available at the NIH GEO, under accession number GSE1159. The data have been collected using Affymetrix's Human Genome U133A Array (Santa Clara, CA), which contains 22,215 probe sets (features). The missing values were filled by averaging across all sample points. We labeled the data into two classes according to their karyotypes: one containing 116 normal karyotype samples, the other containing 157 abnormal karyotype samples. Hence, the total sample size is 273 [Hua *et al.*, 2009].

- **Drugs and toxicants response on rats [Natsoulis *et al.*, 2005], 181 sample points, 8491 features**

  Drugs and toxicants response on rats data set has been obtained from a study characterizing the gene expression of different drugs and toxicants on live rats [Natsoulis *et al.*, 2005]. Altogether, 22 drugs and toxicants have been fed to male Sprague–Dawley rats for several durations and up to 12 tissues have been harvested. The data are publicly available at the NIH GEO, under accession number GSE2187. In this paper, authors claimed that there were totally 597 arrays corresponding to 199 triplicate. But the data set at the NIH GEO contains 587 arrays (198 original sample points, some have only duplicates). In 198 sample points, there are 17 points with no label, so they were removed from the data set.

  The treatments correspond to four categories: fibrates (36 sample points), statins (31 sample points), azoles (53 sample points) and toxicants (61 sample points). We labeled the data into two classes: one containing 61 toxicants samples, the other containing 120 sample points from three the remaining categories. Hence, the total sample size is 181 [Hua *et al.*, 2009]. The data have been collected on cRNA microarray chips containing 8565 probes (features). We excluded the features in which more than 10 percent of the sample points had their values missing. This reduced the total feature size to 8491. The missing values were filled by averaging across all sample points.

- **Lymph-node-negative breast cancer [Wang *et al.*, 2005] 276 sample points, 22,215 features**

  This data set contains the expression of around 22,000 transcripts from total RNA of frozen tumor samples collected by Affymetrix Human U133a GeneChips from 286 lymph-node-negative patients who had not received adjuvant systemic treatment. The goal of the study was to identify genes that discriminated patients who developed distant metastases from those remaining metastasis-free for 5 years and to provide a better means than was currently available for individual risk assessment in patients with lymph-node-negative breast cancer [Wang *et al.*, 2005]. Therefore, we labeled sample points based on distant metastases vs. metastasis-free for 5 years. There were 103 patients with relapse $<$ 5 years, which 10 of them had brain relapses. Table 1 in [Wang *et al.*, 2005] shows that 93 patients had metastases within 5 years, 183 had metastases or follow-up $\geq$ 5 years, and 10 censored (metastases-free and follow-up $\leq$ 5 years).

- **Non–small-cell lung cancer [Potti *et al.*, 2006], 198 sample points, 22,215 features**

  The original study analyzed 198 tumor samples from three cohorts of patients with non–small-cell lung cancer (NSCLC). Total RNA was extracted from the tumor tissue with RNeasy Kits (Qiagen). The RNA quality was assessed with the use of a bioanalyzer (model 2100, Agilent). Hybridization targets were prepared from the total RNA according to standard Affymetrix protocols. The microarray assays were carried out with Affymetrix GeneChips U133 Plus 2.0 Array. Data sets are publicly available at the NIH GEO, under accession number GSE3593 [Potti *et al.*, 2006].

  They identified gene-expression profiles that predicted the risk of recurrence in a cohort of 89 patients with early-stage NSCLC (the lung metagene model); and they evaluated the constructed predictor on two independent groups of 25 and 84 patients from different places. Predictions of recurrence (with 0 representing 5-year disease-free survival and 1 representing death within 2.5 years after the initial diagnosis of NSCLC) were made in terms of the estimated relative probabilities [Potti *et al.*, 2006]. Therefore, we labeled sample points with 5-year disease-free survival vs. death within 2.5 years.

- **Multiple myeloma [Zhan *et al.*, 2006], 234 sample points, 54,613 features**

  Multiple myeloma (MM) data set has been obtained from a study on MM and monoclonal gammopathy of undetermined significance (MGUS) [Zhan *et al.*, 2006] and [Zhan *et al.*, 2007]. The data were collected using Affymetrix's Human Genome U133 Plus 2.0 Array (Santa Clara, CA) and are publicly available at the NIH Gene Expression Omnibus (GEO), under accession numbers GSE5900 and GSE2658. This microarray chip contains 54,613 probe sets (features) to cover all kinds of gene transcripts and variants.

  The original data set is consisted of four subtypes: MM (559 sample points), MGUS (44 sample points), smoldering MM (SMM, 12 sample points), and healthy donors with normal plasma-cell (NPC, 22 sample points). We labeled sample points based on MM sample points vs. the other containing MGUS, SMM, and NPC sample points (78 sample points). Since the number of MM patients is overwhelming and can have significant effects on the efficiency of feature selection and the accuracy of error estimation, we have randomly selected 156 sample points from among the 559 MM sample points and paired them with the 78 sample points of MGUS/SMM/NPC. Hence, the total sample size is 234 [Hua *et al.*, 2009].

- **Node-negative breast cancer [Desmedt *et al.*, 2007], 175 sample points, 22,215 features**

  The original study was carried out with frozen tumor samples from breast cancer patients. The authors tried to independently validate the results of a recent study reporting a 76-gene prognostic signature able to predict distant metastases in lymph node-negative breast cancer patients, and to compare the outcome with clinical risk assessment. The study was conducted by TRANSBIG and aimed to identify patients at high risk of early distant metastases [Desmedt *et al.*, 2007]. Extracted RNAs from samples went under the microarray analyses using the Affymetrix U133a GeneChip. The quality of the RNA obtained from each tumor sample was assessed via the RNA profile generated by the Agilent bioanalyzer. The gene expression data are publicly available at NIH GEO with accession number GSE7390.

As the authors mentioned, the end points considered in this study were time from diagnosis to distant metastases (TDM), which was the end point used to identify the gene signature, and overall survival, defined as time from diagnosis to death from any cause [Desmedt *et al.*, 2007]. They used two ways to label the data: 1. Distant metastasis within 5 years vs. distant metastasis free for 5 years; 2. Distant metastasis within 10 years vs. distant metastasis free for 10 years. These labellings would result in a very unbalanced data set. Therefore, we labeled sample points in the following way: disease-free death $<$ 10 years (77 sample points), disease-free survival $>$ 10 years or disease-free death $>$ 10 years (98 sample points), and 23 censoring sample points (disease-free survival $<$ 10 years).

References

[Hua *et al.*, 2009] Hua,J. *et al.* (2009) Performance of feature selection methods in the classification of high-dimensional data. *Pattern Recogn.*, **42**, 409-424.

[Bhattacharjee *et al.*, 2001] Bhattacharjee,A. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, **98**, 13790-13795.

[Su *et al.*, 2001] Su,A.I. *et al.* (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.*, **61**, 7388-7393.

[Rosenwald *et al.*, 2002] Rosenwald,A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Eng. J. Med.*, **346**, 1937-1947.

[van de Vijver *et al.*, 2002] van de Vijver,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Eng. J. Med.*, **347**, 1999-2009.

[van't Veer *et al.*, 2002] van't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530-536.

[Yeoh *et al.*, 2002] Yeoh,E.J. *et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133-143.

[Chen *et al.*, 2004] Chen,X. *et al.* (2004) Novel endothelial cell markers in hepatocellular carcinoma. *Modern Pathol*, **17**, 1198-1210.

[Valk *et al.*, 2004] Valk,P.J. *et al.* (2004) Prognostically useful gene-expression profiles in acute myeloid leukemia. *N. Eng. J. Med.*, **350**, 1617-1628.

[Natsoulis *et al.*, 2005] Natsoulis,G. *et al.* (2005) Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Res.*, **15**, 724-736.

[Wang *et al.*, 2005] Wang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671-679.

[Potti *et al.*, 2006] Potti,A. *et al.* (2006) A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N. Eng. J. Med.*, **355**, 570-80.

[Zhan *et al.*, 2006] Zhan,F. *et al.* (2006) The molecular classification of multiple myeloma. *Blood*, **108**, 2020-2028.

[Zhan *et al.*, 2007] Zhan,F. *et al.* (2007) Gene-expression signature of benign monoclonal gammopathy evident in multiple myeloma is linked to good prognosis, *Blood*, **109** 1692–1700.

[Desmedt *et al.*, 2007] Desmedt,C. *et al.* (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.*, **13**, 3207-3214.