

Harnessing Biotechnology, Biostatistics and Bioinformatics for 21st Century Biomedical Research

Richard Simon, D.Sc.,
Chief Biometric Research Branch
National Cancer Institute
<http://linus.nci.nih.gov/brb>



DNA Microarray Technology

- Powerful tool for understanding mechanisms and enabling predictive medicine
- Challenges ability of biomedical scientists to analyze and interpret data
- Challenges mathematical scientists with new problems for which existing analysis paradigms are often inapplicable

Microarray Environment

- Hype
- Excessive skepticism
- Mis-information

Problems Illustrated by Microarray Development

- Inefficient utilization of technology for advancing medicine
- Substantial waste of resources on development of ineffective software systems
- Contributions of specialists limited by inability to bridge knowledge boundaries
- Organizations searching for ways of fostering interdisciplinary research
- Insufficient opportunities for attracting the best and brightest undergraduates to quantitative medicine

EDITORIAL

www.nature.com/clinicalpractice/onc

Bioinformatics in cancer therapeutics—hype or hope?

Richard Simon

Bioinformatics should be viewed broadly as the use of mathematical, statistical, and computational methods for the processing and analysis of biologic data. The genomic revolution would not be possible without the sophisticated statistical algorithms on which DNA sequencing, microarray expression profiling and genomic sequence analysis rest. Although data management and integration are important, data analysis and interpretation are the rate-limiting steps for achieving biological understanding and therapeutic progress. Effective integration of scientific bioinformatics into biology and the training of a new generation of biologists and statistical bioinformaticists will require leadership with a vision of biology as an information science.

Development and use of bioinformatics is essential for the future of cancer therapeutics. Most cancer treatments work for only a subset of patients and this is likely to remain true for many molecularly-targeted drugs. This results in a large proportion of patients receiving ineffective treatments and is a huge financial burden on our health care system. It is essential that we develop accurate tools for delivering the right treatment to the right patient based on biological characterization of each patient's tumor.

Gene-expression profiling of tumors using DNA microarrays is a powerful tool for pharmacogenomic targeting of treatments. A good example is the Oncotype DX™ assay (Genomic Health) recently described for identifying the subset of node-negative estrogen-receptor-positive breast cancer patients who do not require adjuvant chemotherapy.¹ Development of genomic tests that are sufficiently validated for broad clinical application requires the sustained effort of a team that includes clinical investigators, biologic scientists and biostatisticians. Accurate, reproducible, predictive diagnostics rarely result from the unstructured retrospective studies of heterogeneous groups of patients that are commonly deposited in the oncology literature, but never independently validated or broadly applied.² With proper focus and support,

The tools to achieve rapid advances in cancer therapeutics are available today. Rapid progress requires wisdom to establish innovative multidisciplinary approaches for the delivery of a new generation of truly effective cancer treatments

gene-expression-based diagnostic tests could be developed today to assist patients and physicians with a wide range of difficult decisions regarding the use of currently existing treatments. Development of such tests should be part of a new paradigm for future therapeutics.

Bioinformatics is also essential for enhancing the discovery of new drugs. Many tumors consist of mixtures of subclones containing different sets of mutated, overexpressed and silenced genes. This heterogeneity makes the process of identifying good molecular targets very challenging. Most overexpressed genes and mutated genes may not represent good molecular targets because resistant subclones are present. The best target is a 'red dot' gene whose mutation occurs early in oncogenesis and dysregulates a key pathway that drives tumor growth in all of the subclones. Examples include mutations in the genes *ABL*, *HER-2*, *KIT*, *EGFR* and probably *BRAF* in chronic myelogenous leukemia, breast cancer, gastrointestinal stromal tumors, non-small-cell lung cancer and melanoma, respectively. Effective development of therapeutics requires identification of red-dot targets, development of drugs that inhibit the red-dot targets, and diagnostic classification of the pathways driving the growth of each patient's tumor. Development and application of bioinformatics by multidisciplinary teams conducting focused translational research is essential for all steps of this process.

Taking advantage of genomic technologies to develop drugs effectively and target them to the right patients depends on the use of bioinformatics, in its broadest sense. The tools to achieve rapid advances in cancer therapeutics are available today. Rapid progress requires wisdom to establish innovative multidisciplinary approaches to focus our technologies and organize our talents for the delivery of a new generation of truly effective cancer treatments.

Supplementary information, in the form of a reference list, is available on the Nature Clinical Practice Oncology website.

R. Simon is an Advisory Board member of Nature Clinical Practice Oncology

Competing interests
The author declared he has no competing interests.
www.nature.com/clinicalpractice
doi:10.1038/ncponc0176

BRB Website Resources for Education of Biomedical Scientists

<http://linus.nci.nih.gov/brb>

- Reprints & Technical Reports
- Powerpoint presentations
 - Audio files
- BRB-ArrayTools software
 - Message board
- BRB-ArrayTools Data Archive
- Sample Size Planning for Targeted Clinical Trials

Using Genomic Classifiers In Clinical Trials

- Simon R and Maitnourim A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* 10:6759-63, 2004.
- Maitnourim A and Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine* 24:329-339, 2005.
- Simon R. When is a genomic classifier ready for prime time? *Nature Clinical Practice – Oncology* 1:4-5, 2004.
- Simon R. An agenda for Clinical Trials: clinical trials in the genomic era. *Clinical Trials* 1:468-470, 2004.
- Simon R. Development and Validation of Therapeutically Relevant Multi-gene Biomarker Classifiers. *Journal of the National Cancer Institute* 97:866-867, 2005.
- Simon R. A roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of Clinical Oncology* 23:7332-41,2005.
- Freidlin B and Simon R. Adaptive signature design. *Clinical Cancer Research* 11:7872-78, 2005.
- Simon R. Validation of pharmacogenomic biomarker classifiers for treatment selection. *Disease Markers* (In Press).
- Simon R. Guidelines for the design of clinical studies for development and validation of therapeutically relevant biomarkers and biomarker classification systems. In *Biomarkers in Breast Cancer*, Hayes DF and Gasparini G, pp 3-15, Humana Press, 2006.

Using Genomic Classifiers In Clinical Trials

- Simon R. and Wang SJ. Use of genomic signatures in therapeutics development in oncology and other diseases, *The Pharmacogenomics Journal* 6:166-73, 2006.
- Simon R. A checklist for evaluating reports of expression profiling for treatment selection. *Clinical Advances in Hematology and Oncology* 4:219-224, 2006.
- Trepicchio WL, Essayan D, Hall ST, Schechter G, Tezak Z, Wang SJ, Weinreich D, Simon R. Designing prospective clinical pharmacogenomic trials- Effective use of genomic biomarkers for use in clinical decision-making. *The Pharmacogenomics Journal* 6:89-94,2006.
- Dupuy A and Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting, *Journal of the National Cancer Institute*, In Press.

BRB Website

<http://linus.nci.nih.gov/brb>

- 15,000 hits per month
- 702 hits to Technical Reports & Talks
- 1985 hits to BRB-ArrayTools home page

BRB-ArrayTools

- Statistically state-of-the-art integrated software for DNA microarray data analysis
 - Architecture and statistical content by R Simon
 - Programming by contractor
- User interface for use and education of biomedical scientists
- Publicly available for non-commercial use
- Active user list-serve and message board

<http://linus.nci.nih.gov/brb>

BRB-ArrayTools

June 2006

- 6283 Registered users
- 2000+ Distinct institutions
- 62 Countries
- 245 Citations

- Registered users
 - 3528 in US
 - 456 at NIH
 - 246 at NCI

My Documents FileMaker Pro 4.1

My Computer GoogleTool...

My Network Places Hawaii Conference...

Recycle Bin Mozilla

Internet Explorer Nal

Microsoft Office Outlook PowerPoint

Adobe Acrobat 6.0 Profession... R. 2.2.0

Desktop Manager R. 2.3.0

Free Meeting! Reflection X 7.20

Java Web Start tr20-01

Mozilla Firefox Word

Fujitsu Documents

StatPro

eed-report

Excel

Biometric Research Branch home page - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

Address <http://linus.nci.nih.gov/BRB-ArrayTools.html>

Google G Go 2 blocked Check AutoLink AutoFill Settings

BRB ArrayTools

Developed by: Richard Simon & Amy Peng Lam

BRB ArrayTools is an integrated package for the visualization and statistical analysis of DNA microarray gene expression data. It was developed by professional statisticians experienced in the analysis of microarray data and involved in the development of improved methods for the design and analysis of microarray based experiments. The array tools package utilizes an Excel front end. Scientists are familiar with Excel and utilizing Excel as the front end makes the system portable and not tied to any database. The input data is assumed to be in the form of Excel spreadsheets describing the expression values and a spreadsheet providing user-specified phenotypes for the samples arrayed. The analytic and visualization tools are integrated into Excel as an add-in. The analytic and visualization tools themselves are developed in the powerful R statistical system, in C and Fortran programs and in Java applications. Visual Basic for Applications is the glue that integrates the components and hides the complexity of the analytic methods from the user. The system incorporates a variety of powerful analytic and visualization tools developed specifically for microarray data analysis.

 Download version 3.5.0 beta1 (Released on September 6, 2006)	 Download version 3.4.1 Stable Release (Released on October 18, 2006)
 BRB ArrayTools Message Board <small>Questions and Answers</small>	 Book for DNA Microarray Analysis
 BRB-ArrayTools Data Archive for Human Cancer Gene Expression	 Publications Based on BRB ArrayTools Analyses <small>(Report generated from scholar.google.com on 09/20/2006)</small>

 [Email BRB-ArrayTools Support](#)

Done Internet



Inbox - Microsoft Outlook

File Edit View Go Tools Actions Help

Type a question for help

Why do I really need to get Array tools - Message (Plain Text)

File Edit View Insert Format Tools Actions Help

Reply Reply to All Forward

You replied on 10/25/2006 3:08 PM.

From: Olga [fe0@mail.ru] Sent: Wed 10/25/2006 2:27 PM

To: Simon, Richard (NIH/NCI) [E]

Cc:

Subject: Why do I really need to get Array tools

Dear Dr.Simon!

I was asked to contact you to get password for downloading BRB-ArrayTools. My name is Bogatova Olga, I'm on my third year in MIPT (Moscow). During this year i'm selecting a laboratory, where i'd like to start my future researchings. And i'm very intrested in one of the laboratories of Institute of Bioorganic Chemistry (Moscow). Microarray method was developed there quite recently (the group developed it is studied human leukemia and some specific factors associated with it), so there are no people there who can execute the statistical analysis of data. Now i'm trying to study out this aspect and i really need the special software to get the final results! So, I hope my attempts will be successful (and i'd like to continue my work in this direction)

S.Y., Olga

So,please, send me a password!

Calendar

Contacts

Tasks

Dear Dr. Simon,
The conference room has 320 seats. To avoid too crowded, we limit each workshop on 250 seats so far. Now we have already drawn over 300 people to join these 2 workshops. Workshop I has been full for local registers, but we remain som...

Paper

Dear Rich,
Please find attached the word file, the figures and the supplementary materials for the paper on confidence intervals for prediction error.
Wenyu

78 Items

All folders are up to date. Connected

- My Documents
- FileMaker Pro 4.1
- My Computer
- GoogleTool...
- My Network Places
- Hawaii Conference...
- Recycle Bin
- Mozilla
- Internet Explorer
- Nal
- Microsoft Office Outlook
- PowerPoint
- Adobe Acrobat 6.0 Profession...
- R 2.2.0
- Desktop Manager
- R 2.3.0
- Free Meeting!
- Reflection X 7.20
- Java Web Start
- tr20-01
- Mozilla Firefox
- Word
- Fujitsu Documents
- StatPro
- eed-report
- Excel

BRB ArrayTools Message Board - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail News RSS Feeds

Address <http://linus.nci.nih.gov/cgi-bin/brb/board1.cgi>

Google Go Bookmarks 2 blocked Check AutoLink AutoFill Send to Settings

BRB ArrayTools Message Board

Importing of over 200 CEL files

Posted by: Soon Paik (soon.paik@nsabp.org) 09/25/2006 11:10

Dear BRB team,

When we compiled U133_2.0 plus CEL files, with 200 cases it ran fine. However, the program sent error message when we tried to compile 238 CEL files.

The error message is:

Error cannot allocate vector of size 1123452 Kb

09/25/2006 11:10 By: Soon Paik [E-Mail](#)







Message Board By Biometric Research Branch/NCI/NIH

[What's New](#) | [Guest Book](#) | [Message Board](#) | [Download](#) | [Feedback](#) | [Licenses](#) | [Technical Reports](#)

Done Internet

	A	B	C	D	E	F	G	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1																									
2																									
3																									
4																									
5																									
6																									
7																									
8																									
9																									
10																									
11																									
12																									
13																									
14																									
15																									
16																									
17																									
18																									
19																									
20																									
21																									
22																									
23																									
24																									
25																									
26																									
27																									
28																									
29																									
30																									
31																									
32																									
33																									
34																									
35																									
36																									
37																									
38																									
39																									
40																									
41																									
42																									
43																									
44																									
45																									
46																									
47																									
48																									
49																									
50																									
51																									
52																									
53																									
54																									
55																									
56																									
57																									
58																									

- ArrayTools
- Getting started
- Data import wizard
- Import data
- Analysis wizard
 - Scatterplot
 - Clustering
 - Multidimensional scaling of samples
 - Class comparison
 - Class prediction**
 - Survival analysis
 - Quantitative trait analysis
- Filter and subset the data
- Plugins
- Utilities
- Support
- About BRB-ArrayTools
- About R-COM
- License agreement

- Class prediction**
- Binary tree prediction
- Prediction analysis of Microarrays (PAM)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	EXP	PatientID	BeforeAfter																										
2	svcc77		10 AF																										
3	svcc78		10 BE																										
4	svcc86		100 AF																										
5	svcc104		100 BE																										
6	svcc85		101 AF																										
7	svcc84		101 BE																										
8	svcc82		102 AF																										
9	svcc101		102 BE																										
10	svcc65		104 AF																										
11	svcc120		104 BE																										
12	svcc121		109 AF																										
13	svcc105		109 BE																										
14	svcc126		112 AF																										
15	svcc124		112 BE																										
16	svcc97		12 AF																										
17	svcc53		12 BE																										
18	svcc52		14 AF																										
19	svcc81		14 BE																										
20	svcc64		15 AF																										
21	svcc106		15 BE																										
22	svcc112		18 AF																										
23	svcc134		18 BE																										
24	svcc91		26 AF																										
25	svcc89		26 BE																										
26	svcc109		27 AF																										
27	svcc111		27 BE																										
28	svcc135		39 AF																										
29	svcc93		39 BE																										
30	svcc125		41 AF																										
31	svcc107		41 BE																										
32	svcc79		47 AF																										
33	svcc76		47 BE																										
34	svcc103		48 AF																										
35	svcc92		48 BE																										
36	svcc58		53 AF																										
37	svcc131		53 BE																										
38	svcc66		61 AF																										
39	svcc83		61 BE																										
40	svcc63		7 AF																										
41	svcc98		7 BE																										

Class prediction

This procedure computes a classifier which can be used for predicting the class of a new sample.

Class Prediction Options

Cross-validation method:

- Leave-one-out validation
- 10 - fold validation
Repeated 1 times
- 0.632 bootstrap validation

Do statistical significance test of cross-validated mis-classification rate.

Number of permutations for significance test of cross-validated mis-classification rate: 100

Probability:

- Bayesian Analog of Compound Covariate Predictor
- Prior Probabilities that reflect possible imbalances in class membership

Name to use for output files: ClassPrediction

Use separate test set:
Column containing "training", "predict", "exclude" labels:

OK Cancel Options Reset Help

Support vector machines

OK Cancel Options Reset Help



BRB-ArrayTools Data Archive - Microsoft Internet Explorer

Address: <http://linus.nci.nih.gov/~brb/DataArchive.html>

BRB-ArrayTools Data Archive for Human Cancer Gene Expression

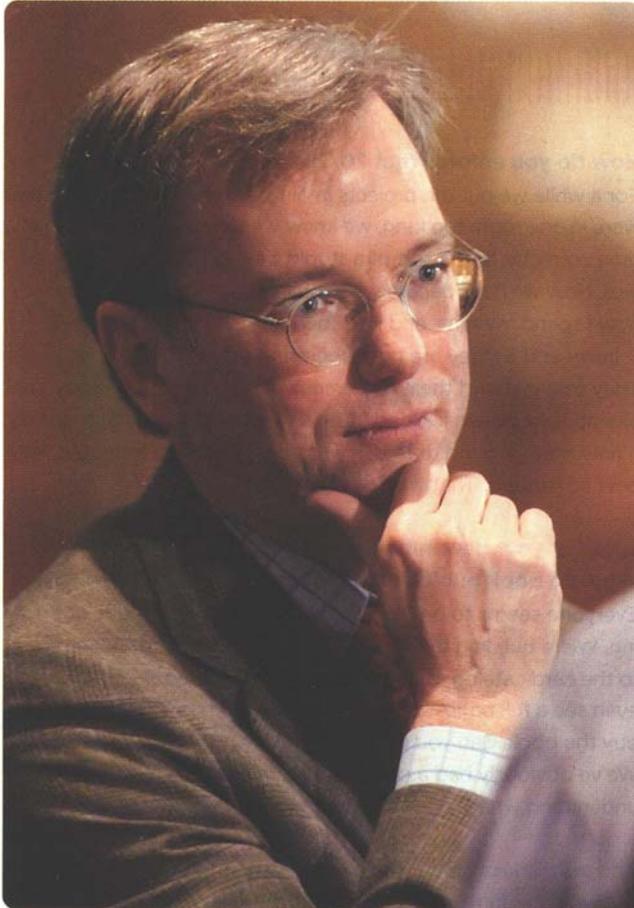
Expression array data and clinical descriptors are stored as BRB-ArrayTools project folders. To analyze a dataset, simply download and unzip the file and open the project worksheet in excel on a Microsoft Windows PC. BRB-ArrayTools must have been installed on the machine.

BRB-ArrayTools is a comprehensive state-of-the-art statistical analysis system for the analysis of microarray gene expression data. It is free for non-commercial purposes and can be licenced for commercial purposes from the NIH. It is easily installed as an excel plug-in using its self installer. To download BRB-ArrayTools, go to <http://linus.nci.nih.gov/~brb/BRB-ArrayTools.html>

Tumor Type	Citation	Paper	Transformed Zip File	Readme File
Brain	Pomeroy et al (2002) Nature 415:436-42	Prediction of central nervous system embryonal tumour outcome based on gene expression	6.6 MB	
Brain	Liang et al. (2005) PNAS 102:5814-9	Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme	18.1 MB	
Breast	Perou CM et al (2000) Nature 406:747-52	Molecular portraits of human breast tumours	14.9 MB	
Breast	Sorlie et al (2001) PNAS 98:10869-74	Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications	18.3 MB	
Breast	van't Veer et al (2002) Nature 415:530-6	Gene expression profiling predicts clinical outcome of breast cancer	21.4 MB	
Breast	Sotiriou et al (2003) PNAS 100:10393-8	Breast cancer classification and prognosis based on gene expression profiles from a population-based study	9.9 MB	
Breast	Sorlie et al (2003) PNAS 100:8418-2398	Repeated observation of breast tumor subtypes in independent gene expression data sets	48.7 MB	
Breast	Yang et al (2005) Clin Cancer Res 11:6226-32	Gene expression patterns and profile changes pre- and post-erlotinib treatment in patients with metastatic breast cancer	10.0 MB	
Colon	Alon et al. (1999) PNAS 96:6745-50	Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays	1.4 MB	
Gastric	Chen et al (2003) Mol Biol Cell 14:3208-15	Variation in gene expression patterns in human gastric cancers	112.6 MB	
GIST	Subramanian et al (2004)	Gastrointestinal stromal tumors (GISTs) with KIT and	80.2 MB	

BRB Array Tools

- Wasn't designed by committee
- Wasn't a response to users dictating what should be developed.
 - We encourage our users to have clear objectives; they are not experts on how to achieve those objectives
- Offers substance, not flashy displays
 - State of the art statistical advice in a form useable by non-statisticians



“Everything new here comes out of the time engineers spend on side projects. **It certainly doesn’t come from management.”**



So is Google building a computing platform? A Web-based operating system, if you will?

The problem I have with that question is that “operating system” and “platform” and “Web OS” are very generic terms, so I prefer not to engage in those discussions. There is this presumption that Google has to go build its own OS, its own browser, when those technologies are quite mature and well valued. There is a great deal of strategic leverage for us in building an ecosystem around content and advertising that is an extension of our search mission.

Moving from Correlative Studies to Predictive Medicine

“Biomarkers”

- Surrogate endpoints
 - A measurement made on a patient before, during and after treatment to determine whether the treatment is working
- Predictive classifier
 - A measurement made before treatment to predict whether a particular treatment is likely to be beneficial

Surrogate Endpoints

- It is extremely difficult to properly validate a biomarker as a surrogate for clinical outcome. It requires a series of randomized trials with both the candidate biomarker and clinical outcome measured

Biomarkers for Treatment Selection

- Oncologists need improved tools for selecting treatment for individual patients
- Most cancer treatments benefit only a minority of patients to whom they are administered
- Being able to predict which patients are likely to benefit would save patients from unnecessary toxicity, inconvenience and enhance their chance of receiving a drug that helps them
- The current over treatment of patients results in a major expense for individuals and society

Oncology Needs Predictive Markers not Prognostic Factors

- Most prognostic factors are not used because they are not therapeutically relevant
- Most prognostic factor studies use a convenience sample of patients for whom tissue is available. Often the patients are too heterogeneous to support therapeutically relevant conclusions

Pusztai et al. The Oncologist 8:252-8, 2003

- 939 articles on “prognostic markers” or “prognostic factors” in breast cancer in past 20 years
- ASCO guidelines only recommend routine testing for ER, PR and HER-2 in breast cancer
- “With the exception of ER or progesterone receptor expression and HER-2 gene amplification, there are no clinically useful molecular predictors of response to any form of anticancer therapy.”

- Targeted clinical trials can be much more efficient than untargeted clinical trials, if we know who to target

- In new drug development, the role of a classifier is to select a target population for treatment
 - The focus should be on evaluating the new drug in a population defined by a predictive classifier, not on “validating” the classifier

Developmental Strategy (I)

- **Develop** a diagnostic classifier that identifies the patients likely to benefit from the new drug
- Develop a reproducible assay for the classifier
- **Use** the diagnostic to restrict eligibility to a prospectively planned evaluation of the new drug
- Demonstrate that the new drug is effective in the prospectively defined set of patients determined by the diagnostic

Develop Predictor of Response to New Drug

```
graph TD; A[Develop Predictor of Response to New Drug] --> B[Patient Predicted Responsive]; A --> C[Patient Predicted Non-Responsive]; B --> D[New Drug]; B --> E[Control]; C --> F[Off Study];
```

Patient Predicted Responsive

Patient Predicted Non-Responsive

New Drug

Control

Off Study

Evaluating the Efficiency of Strategy (I)

- Simon R and Maitnourim A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* 10:6759-63, 2004; Correction & Supplement 12:3229, 2006
- Maitnourim A and Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine* 24:329-339, 2005.
- reprints and interactive sample size calculations at <http://linus.nci.nih.gov/brb>

Two Clinical Trial Designs

- Un-targeted design
 - Randomized comparison of E to C without screening for expression of molecular target
- Targeted design
 - Assay patients for expression of target
 - Randomize only patients expressing target

Pharmacogenomic Model for Two Treatments With Binary Response

- Molecularly targeted treatment E
- Control treatment C
- γ Proportion of patients that express target
- p_c control response probability
- response probability for E patients who express target is $(p_c + \delta_1)$
- Response probability for E patients who do not express target is $(p_c + \delta_0)$

Randomized Ratio

(approximation)

- $\text{RandRat} = n_{\text{untargeted}}/n_{\text{targeted}}$

$$\text{RandRat} \approx \left(\frac{\delta_1}{\gamma\delta_1 + (1-\gamma)\delta_0} \right)^2$$

- δ_1 = rx effect in marker + patients
- δ_0 = rx effect in marker - patients
- γ = proportion of marker + patients
- If $\delta_0=0$, $\text{RandRat} = 1/\gamma^2$
- If $\delta_0 = \delta_1/2$, $\text{RandRat} = 4/(\gamma + 1)^2$

Randomized Ratio

$$n_{\text{untargeted}}/n_{\text{targeted}}$$

Proportion Assay Positive	No Treatment Benefit for Assay Negative Patients	Treatment Benefit for Assay Negative Patients is Half That for Assay Positive Patients
0.75	1.78	1.31
0.5	4	1.78
0.25	16	2.56

Screened Ratio

Proportion Assay Positive	No Treatment Benefit for Assay Negative Patients	Treatment Benefit for Assay Negative Patients is Half That for Assay Positive Patients
0.75	1.33	0.98
0.5	2	0.89
0.25	4	0.64

Imperfect Assay Sensitivity & Specificity

- λ_{sens} = sensitivity
 - $\Pr[\text{assay+} \mid \text{target expressed}]$
- λ_{spec} = specificity
 - $\Pr[\text{assay-} \mid \text{target not expressed}]$

Proportion of Assay Positive Patients That Express Target

$$w_1 = \frac{\gamma \lambda_{sens}}{\gamma \lambda_{sens} + (1 - \gamma)(1 - \lambda_{spec})}$$

Randomized Ratio

- $\text{RandRat} = n_{\text{untargeted}}/n_{\text{targeted}}$

$$\text{RandRat} = \left(\frac{w_1 \delta_1 + (1 - w_1) \delta_0}{\gamma \delta_1 + (1 - \gamma) \delta_0} \right)^2$$

Randomized Ratio

sensitivity=specificity=0.9

γ Express target	$\delta_0=0$	$\delta_0= \delta_1/2$
0.75	1.29	1.26
0.5	1.8	1.6
0.25	3.0	1.96

- For Trastuzumab, even a relatively poor assay enabled conduct of a targeted phase III trial which was crucial for establishing effectiveness
- Recent results with Trastuzumab in early stage breast cancer show dramatic benefits for patients selected to express Her-2

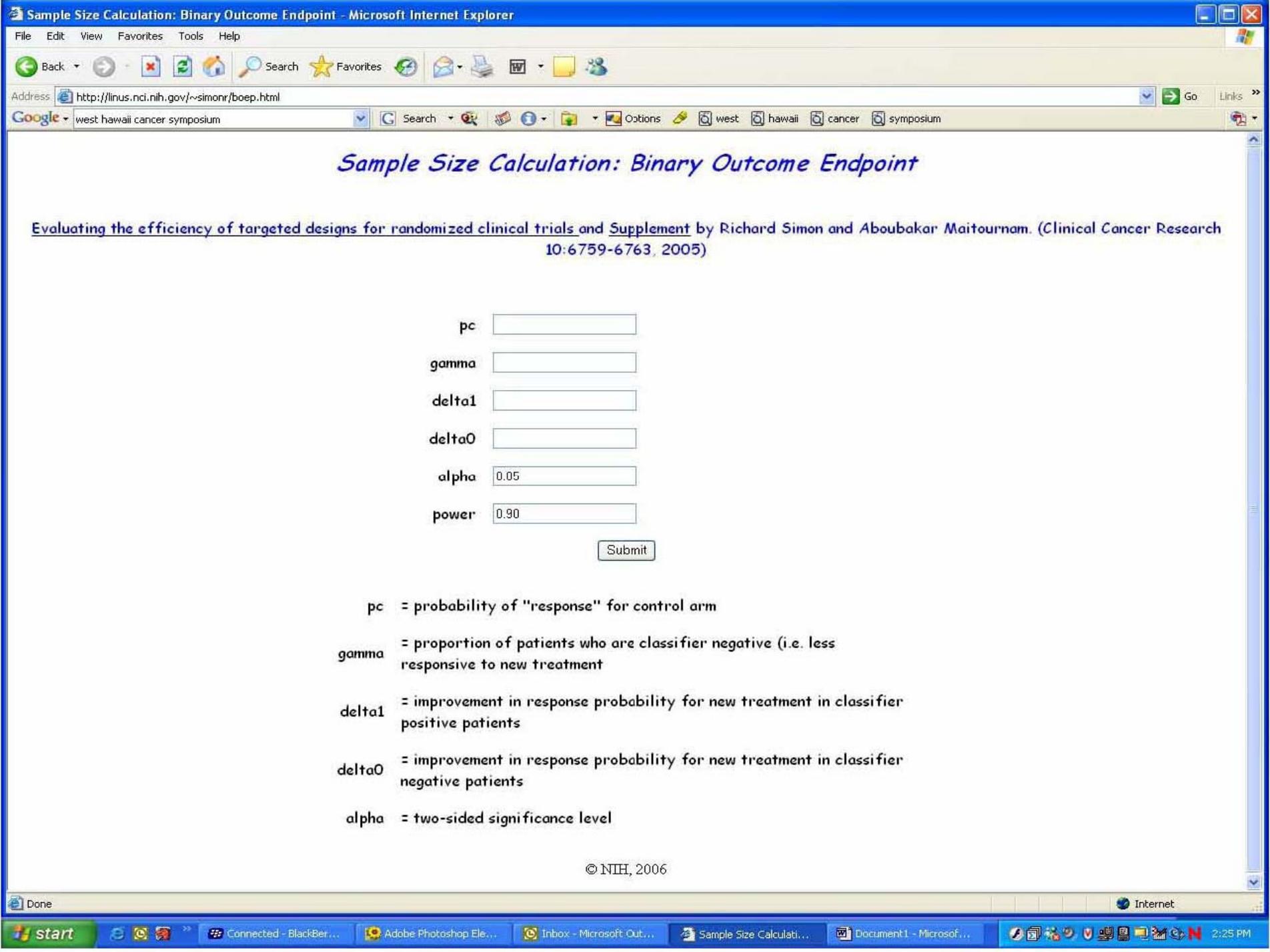
Comparison of Targeted to Untargeted Design

Simon R, Development and Validation of Biomarker Classifiers for Treatment Selection, JSPI

Treatment Hazard Ratio for Marker Positive Patients	Number of Events for Targeted Design	Number of Events for Traditional Design		
		Percent of Patients Marker Positive		
		20%	33%	50%
0.5	74	2040	720	316

Interactive Software for Evaluating a Targeted Design

- <http://linus.nci.nih.gov/brb/>



Sample Size Calculation: Binary Outcome Endpoint

Evaluating the efficiency of targeted designs for randomized clinical trials and Supplement by Richard Simon and Aboubakar Maitournam. (Clinical Cancer Research 10:6759-6763, 2005)

pc

gamma

delta1

delta0

alpha

power

pc = probability of "response" for control arm

gamma = proportion of patients who are classifier negative (i.e. less responsive to new treatment)

delta1 = improvement in response probability for new treatment in classifier positive patients

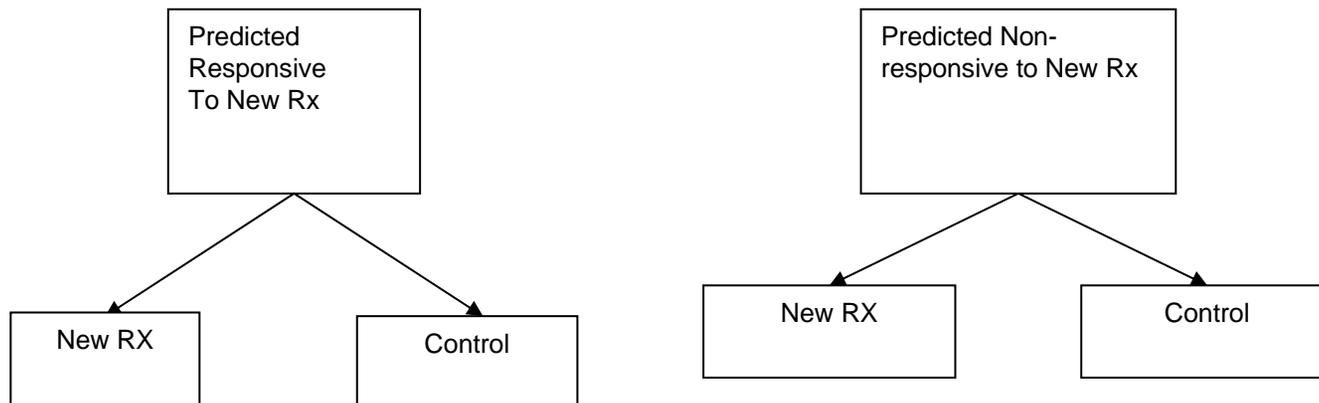
delta0 = improvement in response probability for new treatment in classifier negative patients

alpha = two-sided significance level

© NIH, 2006

Developmental Strategy (II)

Develop Predictor of
Response to New Rx



Developmental Strategy (II)

- Do not use the diagnostic to restrict eligibility, but to structure a prospective analysis plan.
- Compare the new drug to the control overall for all patients ignoring the classifier.
 - If $p_{\text{overall}} \leq 0.04$ claim effectiveness for the eligible population as a whole
- Otherwise perform a single subset analysis evaluating the new drug in the classifier + patients
 - If $p_{\text{subset}} \leq 0.01$ claim effectiveness for the classifier + patients.

Key Features of Design (II)

- The purpose of the RCT is to evaluate treatment T vs C overall and for the pre-defined subset; not to re-evaluate the components of the classifier, or to modify or refine the classifier

Developmental Strategy (IIb)

- Do not use the diagnostic to restrict eligibility, but to structure a prospective analysis plan.
- Compare the new drug to the control for classifier positive patients
 - If $p_+ > 0.05$ make no claim of effectiveness
 - If $p_+ \leq 0.05$ claim effectiveness for the classifier positive patients and
 - Continue accrual of classifier negative patients and eventually test treatment effect at 0.05 level

The Roadmap

1. Develop a completely specified genomic classifier of the patients likely to benefit from a new drug
2. Establish reproducibility of measurement of the classifier
3. Use the completely specified classifier to design and analyze a new clinical trial to evaluate effectiveness of the new treatment with a pre-defined analysis plan.

Guiding Principle

- The data used to develop the classifier must be distinct from the data used to test hypotheses about treatment effect in subsets determined by the classifier
 - Developmental studies are exploratory
 - Studies on which treatment effectiveness claims are to be based should be definitive studies that test a treatment hypothesis in a patient population completely pre-specified by the classifier

Development of Genomic Classifiers

- Single gene or protein based on knowledge of therapeutic target
- Single gene or protein culled from set of candidate genes identified based on imperfect knowledge of therapeutic target
- Empirically determined based on correlating gene expression to patient outcome after treatment

Development of Genomic Classifiers

- During phase II development or
- After failed phase III trial using archived specimens.
- Adaptively during early portion of phase III trial.

Development of Empirical Gene Expression Based Classifier

- 20-30 phase II responders are needed to compare to non-responders in order to develop signature for predicting response
 - Dobbin KK, Simon RM. Sample size planning for developing classifiers using high dimensional DNA microarray data, Biostatistics (In Press)

Adaptive Signature Design

An adaptive design for generating and prospectively testing a gene expression signature for sensitive patients

Boris Freidlin and Richard Simon

Clinical Cancer Research 11:7872-8, 2005

Adaptive Signature Design

End of Trial Analysis

- Compare E to C for **all patients** at significance level 0.04
 - If overall H_0 is rejected, then claim effectiveness of E for eligible patients
 - Otherwise

- Otherwise:
 - Using only the first half of patients accrued during the trial, develop a binary classifier that predicts the subset of patients most likely to benefit from the new treatment E compared to control C
 - Compare E to C for patients accrued in second stage who are predicted responsive to E based on classifier
 - Perform test at significance level 0.01
 - If H_0 is rejected, claim effectiveness of E for subset defined by classifier

**Treatment effect restricted to subset.
10% of patients sensitive, 10 sensitivity genes, 10,000 genes, 400 patients.**

Test	Power
Overall .05 level test	46.7
Overall .04 level test	43.1
Sensitive subset .01 level test (performed only when overall .04 level test is negative)	42.2
Overall adaptive signature design	85.3

**Overall treatment effect, no subset effect.
10,000 genes, 400 patients.**

Test	Power
Overall .05 level test	74.2
Overall .04 level test	70.9
Sensitive subset .01 level test	1.0
Overall adaptive signature design	70.9

Validation of Predictive Classifiers for Use with Available Treatments

- Should establish that the classifier is robust, reproducibly measurable and has clinical utility
- Studies of predictive classifiers should be viewed as either *developmental* or *validation* studies

- Developmental studies should develop classifiers for homogeneously treated patients and provide split-sample or cross-validated estimates of prediction accuracy
- Validation studies should establish whether patient outcome is improved using the pre-specified new classifier for treatment selection compared to using current practice standards

COMMENTARY

Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification

Richard Simon, Michael D. Radmacher, Kevin Dobbin, Lisa M. McShane

DNA microarrays have made it possible to estimate the level of expression of thousands of genes for a sample of cells. Although biomedical investigators have been quick to adopt this powerful new research tool, accurate analysis and interpretation of the data have provided unique challenges. Indeed, many investigators are not experienced in the analytical steps needed to convert tens of thousands of noisy data points into reliable and interpretable biologic information. Although some investigators recognize the importance of collaborating with experienced biostatisticians to analyze microarray data, the number and availability of experienced biostatisticians is inadequate. Consequently, investigators are using available software to analyze their data, many seemingly without knowledge of potential pitfalls. Because of serious problems associated with the analysis and reporting of some DNA microarray studies, there is great interest in guidance on valid and effective methods for analysis of DNA microarray data.

The design and analysis strategy for a DNA microarray experiment should be determined in light of the overall objectives of the study. Because DNA microarrays are used for a wide variety of objectives, it is not feasible to address the entire range of design and analysis issues in this commentary. Here, we address statistical issues that arise from the use of DNA microarrays for an important group of objectives that has been called "class prediction" (1). Class prediction includes derivation of predictors of prognosis, response to therapy, or any phenotype or genotype defined independently of the gene expression profile.

EXPERIMENTAL OBJECTIVES DRIVE DESIGN AND ANALYSIS

Good DNA microarray experiments, although not based on gene-specific mechanistic hypotheses, should be planned and conducted with clear objectives. Three commonly encountered types of study objectives are "class comparison," "class prediction," and "class discovery" (1).

Class comparison is the comparison of gene expression in different groups of specimens. The major characteristic of class comparison studies is that the classes being compared are defined independently of the expression profiles. The specific objectives of such a study are to determine whether the expression profiles are different between the classes and, if so, to identify the differentially expressed genes. One example of a class comparison study is the comparison of gene expression profiles of stage I breast cancer patients who are long-term survivors with the gene expression profiles of those who have recurrent disease. Another example is the comparison between gene expression profiles in breast cancer patients with and without germline BRCA1 mutations (2).

Class prediction studies are similar to class comparison studies in that the classes are predefined. In class prediction studies,

however, the emphasis is on developing a gene expression-based multivariate function (referred to as the predictor) that accurately predicts the class membership of a new sample on the basis of the expression levels of key genes. Such predictors can be used for many types of clinical management decisions, including risk assessment, diagnostic testing, prognostic stratification, and treatment selection. Many studies include both class comparison and class prediction objectives.

Class discovery is fundamentally different from class comparison or class prediction in that no classes are predefined. Usually the purpose of class discovery in cancer studies is to determine whether discrete subsets of a disease entity can be defined on the basis of gene expression profiles. This purpose is different from determining whether the gene expression profiles correlate with some already known diagnostic classification. Examples of class discovery are the studies by Bitiner et al. (3) that examined gene expression profiles for advanced melanomas and by Alizadeh et al. (4) that examined the gene expression profiles of patients with diffuse large B-cell lymphoma. Often the purpose of class discovery is to identify clues regarding the heterogeneity of disease pathogenesis.

LIMITATIONS OF CLUSTER ANALYSIS FOR CLASS PREDICTION

One of the most common errors in the analysis of DNA microarray data is the use of cluster analysis and simple fold change statistics for problems of class comparison and class prediction. Although cluster analysis is appropriate for class discovery, it is often not effective for class comparison or class prediction. Cluster analysis refers to an extensive set of methods for partitioning samples into groups on the basis of the similarities and differences (referred to as distances) among their gene expression profiles. Because there are many ways of measuring distances among gene expression profiles involving thousands of genes and because there are many algorithms for partitioning, cluster analysis is a very subjective analysis strategy.

Cluster analysis is considered an unsupervised method of analysis because no information about sample grouping is used. The distance measures are generally computed with regard to the complete set of genes represented on the array that are measured with sufficiently high signals, or with regard to all the genes that

Affiliations of authors: R. Simon, K. Dobbin, L. M. McShane, Biometric Research Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD; M. D. Radmacher, Departments of Biology and Mathematics, Kenyon College, Gambier, OH.

Correspondence to: Richard Simon, D.Sc., National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892-7434 (e-mail: rsimon@nih.gov).

See "Note" following "References."

© Oxford University Press

show meaningful variation across the sample set. Because relatively few genes may distinguish any particular class, the distances used in cluster analysis will often not reflect the influence of these relevant genes. This feature accounts for the poor results often obtained in attempting to use cluster analysis for class prediction studies.

Cluster analysis also does not provide statistically valid quantitative information about which genes are differentially expressed between classes. Investigators often use simple average fold change measures or visual inspection of a cluster image display to identify differentially expressed genes. However, average fold change indices do not account for variability in gene expression across samples within the same class; some twofold average effects represent statistically significant differences and some do not. Neither fold change indices nor visual inspection of cluster image displays enable the investigator to deal with multiple comparison issues in a statistically valid manner. For example, in examining expression levels of thousands of randomly varying genes, there may be many genes that spuriously appear to be differentially expressed between two classes on the basis of visual inspection or fold change thresholds.

CLASS PREDICTION USING SUPERVISED METHODS

For class prediction studies, it is more appropriate to use a supervised method (i.e., one that makes distinctions among the specimens on the basis of predefined class label information) than an unsupervised method, such as cluster analysis. Supervised class prediction is usually based on the assumption that a collection of differentially expressed genes is associated with class distinction.

The first step toward constructing the class predictor (sometimes called the classifier) is to select the subset of informative genes. The second step is often to assign weights related to the individual predictive strengths of these informative genes. Predictors based on linear combinations of the weighted intensity measurements of the informative genes have been proposed (1,5). One alternative method is to use a dimension reduction technique such as principal components analysis or partial least squares on the informative genes and to base the prediction on the resulting factors (6–8). Many other methods for defining a multivariate predictor have been described (9,10). The final step in constructing the classifier is to define the prediction rule. For example, in a two-group classification where a single predictor is computed, the classification rule may simply be a threshold value; a specimen is classified as being in one group if the derived predictor value is less than the threshold and classified as being in the other group if the derived predictor value is more than the threshold.

One major limitation of supervised methods is overfitting the predictor. Overfitting means that the number of parameters of the model is too large relative to the number of cases or specimens available. Because the model parameters are optimized for the data, the model will fit the original data but may predict poorly for independent data. This happens because the model fits random variations within the original data that do not represent true relationships that hold for independent data. Consequently, it is essential to obtain an unbiased estimate of the true error rate of the predictor (i.e., the probability of incorrectly classifying a randomly selected future case).

Methods for obtaining unbiased estimates of a predictor's error rate include leave-one-out cross-validation or application

of the prediction rule developed from a supervised analysis of one dataset to an independent dataset. By using these techniques, it is possible not only to evaluate overfitting the predictor, but also to compare various prediction methods and assess which ones are less prone to overfitting. The appropriate use of leave-one-out cross-validation and validation of independent datasets is discussed in the next two sections.

CROSS-VALIDATION OF PREDICTION ACCURACY

The performance of a class prediction rule is best assessed by applying the rule created on one set of data (the training set) to an independent set of data (the validation set). Because most clinical research laboratories have access to only a limited number of tumor samples, withholding a substantial proportion of the samples from the training set for the sake of creating a validation set may considerably reduce the performance of the prediction rule. Cross-validation procedures use the data more efficiently. A small number of specimens are withheld, and most of the specimens are used to build a predictor. The predictor is used to predict class membership for the withheld specimens. This process is iterated, leaving out a new set of specimens at each step, until all specimens have been classified. In leave-one-out cross-validation, for example, each specimen is excluded from the training set one at a time and then classified on the basis of the predictor built from the data for all of the other specimens. The leave-one-out cross-validation procedure provides a nearly unbiased estimate of the true error rate of the classification procedure. The estimated error rate applies to the procedure used to build the classifier rather than to the specific prediction model based on all the data, because there is a different classifier for each leave-one-out training set (11,12). Other cross-validation methods omit more than one specimen at a time (13) and also produce nearly unbiased estimates.

In the previous section, three common components of class prediction methods were listed: 1) selection of informative genes, 2) computation of weights for selected informative genes, and 3) creation of a prediction rule. It is important that all three steps undergo the cross-validation procedure. Failure to cross-validate all steps may lead to substantial bias in the estimated error rate.

We performed a simulation to examine the bias in estimated error rates for a class prediction study with various levels of cross-validation (see supplemental information at <http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue/1/index.shtml> and at <http://lmus.nci.nih.gov/~brb> for a full description of the simulation). We considered two types of leave-one-out cross-validation: one with removal of the left-out specimen before selection of differentially expressed genes and one with removal of the left-out specimen after gene selection but before computation of gene weights and application of the prediction rule. We also computed the resubstitution estimate of the error rate (this estimate results from building the predictor on the full dataset and then reapplying it to each specimen for classification purposes). In each simulated dataset, 20 expression profiles of 6000 genes were randomly generated from the same distribution. Ten profiles were arbitrarily assigned to class 1 and the other 10 profiles to class 2, creating an artificial separation of the profiles into two classes. Because no true underlying difference existed between the two classes, the class prediction should perform no better than a random guess, with

estimated error rates for simulated datasets centering around 0.5 (i.e., 10 misclassifications of 20).

The observed number of misclassifications resulting from each level of cross-validation for 2000 simulated datasets is shown in Fig. 1. The resubstitution estimate of the error is biased for small datasets (11,12); our simulation confirms this, with 98.2% of the simulated datasets resulting in zero misclassifications, even though no true underlying difference existed between the two groups. Moreover, the maximum number of misclassified specimens obtained using the resubstitution method was one. Cross-validating the prediction rule after selecting differentially expressed genes from the full dataset does little to correct the bias, with 90.2% of simulated datasets still resulting in zero misclassifications. However, when gene selection is also subjected to cross-validation, we observed results in line with our expectation: the median number of misclassified profiles was 11, although the range was large (0–20).

The simulation results underscore the importance of cross-validating all steps of predictor construction in estimating a true error rate. A recently published study (14) also illustrates the point: van't Veer et al. (14) predicted clinical outcome of patients with axillary lymph node-negative breast cancer (metastatic disease within 5 years versus disease-free at 5 years) from gene expression profiles, first by using an incomplete cross-validation method and then by using a fully cross-validated method. Their incomplete cross-validation method did not include reselection of the differentially expressed genes. The investigators controlled for the number of misclassified recurrent patients (i.e., the sensitivity of the test) in both situations. To illustrate the importance of proper cross-validating, we focus attention on the difference in estimated error rates for the disease-free patients. The incomplete cross-validation method and the fully cross-validated leave-one-out method result in estimated error rates of 27% (12 of 44) and 41% (18 of 44), respectively. The incomplete method results in a seriously biased underestimate of the error rate, probably largely from overfitting the predictor to the specific dataset.

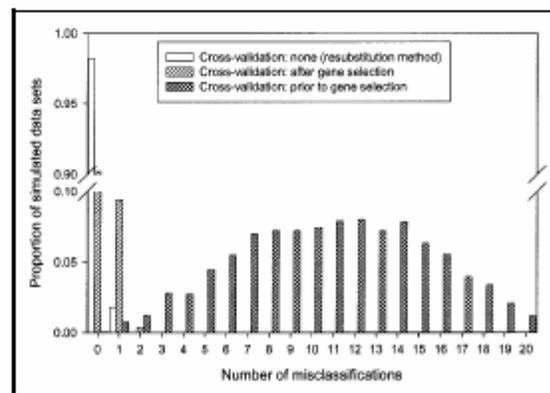
Although van't Veer et al. (14) report both the partially and fully cross-validated estimates of the error rate, it is the smaller and invalid partially cross-validated estimate that has received more attention (15). When cross-validation methods are improperly performed, i.e., without repeating all steps of gene selection and predictor construction within each stage of the cross-validation, the results can be almost as biased as if cross-validation had not been used. Unfortunately, this error is common. In a recent study (16) where decision trees were built from gene expression data to classify specimens as normal colon or colon cancer, the authors used a procedure that cross-validated only the steps that occurred after selection of the informative genes. The full dataset was used to identify the informative genes.

VALIDATION ON INDEPENDENT DATA

A class predictor that results in a small, properly cross-validated error rate for a collection of tumor specimens is a potentially important finding but one that still requires further validation. This is especially true for class prediction studies based on fewer than 50 tumor specimens. The relatively small sample sizes necessitate validation of the predictor on independently collected data for at least two reasons. First, although cross-validated error estimates are nearly unbiased, they have a large variance for small sample sizes (17). For example, a cross-validated error rate of 0.10 derived from a set of 20 tumors may have a large associated standard error and does not guarantee a true error rate of 0.10 for the predictor. The standard error can be reduced somewhat by using more complex versions of cross-validation (17). Second, the tumors used to build the predictor in the original study may not accurately reflect all characteristics of the underlying populations of interest; the predictor may ignore important properties of the larger population or heavily weight peculiarities of the training set.

An independent validation dataset should ideally be large enough to demonstrate statistically that predictions are accurate.

Fig. 1. Effect of various levels of cross-validation on the estimated error rate of a predictor derived from 2000 simulated datasets. Class labels were arbitrarily assigned to the specimens within each dataset, so poor classification accuracy is expected. Class prediction was performed on each dataset as described in the supplemental information (<http://jnci.aphispubs.nrc.gov/journals.org/jnci/content/vol95/issue1/index.html> and <http://www.aphis.gov/ehb>), varying the level of leave-one-out cross-validation used in the prediction. Vertical bars indicate the proportion of simulated datasets (of 2000) resulting in a given number of misclassifications for a specified cross-validation strategy.



Prediction Error Estimation: A Comparison of Resampling Methods

Annette M. Molinaro^{ab*}, Richard Simon^c, Ruth M. Pfeiffer^a

^aBiostatistics Branch, Division of Cancer Epidemiology and Genetics, NCI, NIH, Rockville, MD 20852, ^bDepartment of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, ^cBiometric Research Branch, Division of Cancer Treatment and Diagnostics, NCI, NIH, Rockville, MD 20852

ABSTRACT

Motivation: In genomic studies, thousands of features are collected on relatively few samples. One of the goals of these studies is to build classifiers to predict the outcome of future observations. There are three inherent steps to this process: feature selection, model selection, and prediction assessment. With a focus on prediction assessment, we compare several methods for estimating the 'true' prediction error of a prediction model in the presence of feature selection.

Results: For small studies where features are selected from thousands of candidates, the resubstitution and simple split-sample estimates are seriously biased. In these small samples, leave-one-out (LOOCV), 10-fold cross-validation (CV), and the .632+ bootstrap have the smallest bias for diagonal discriminant analysis, nearest neighbor, and classification trees. LOOCV and 10-fold CV have the smallest bias for linear discriminant analysis. Additionally, LOOCV, 5- and 10-fold CV, and the .632+ bootstrap have the lowest mean square error. The .632+ bootstrap is quite biased in small sample sizes with strong signal to noise ratios. Differences in performance among resampling methods are reduced as the number of specimens available increase.

Availability: A complete compilation of results in tables and figures is available in Molinaro *et al.* (2005). R code for simulations and analyses is available from the authors.

Contact: annette.molinaro@yale.edu

1 INTRODUCTION

In genomic experiments one frequently encounters high dimensional data and small sample sizes. Microarrays simultaneously monitor expression levels for several thousands of genes. Proteomic profiling studies using SELDI-TOF (surface-enhanced laser desorption and ionization time-of-flight) measure size and charge of proteins and protein fragments by mass spectroscopy, and result in up to 15,000 intensity levels at prespecified mass values for each spectrum. Sample sizes in such experiments are typically less than 100.

In many studies observations are known to belong to predetermined classes and the task is to build predictors or classifiers for new observations whose class is unknown. Deciding which genes or proteomic measurements to include in the prediction is called *feature selection* and is a crucial step in developing a class predictor. Including too many noisy variables reduces accuracy of the prediction and may lead to over-fitting of data, resulting in promising but often non-reproducible results (Ransohoff, 2004).

Another difficulty is model selection with numerous classification models available. An important step in reporting results is assessing the chosen model's error rate, or generalizability. In the absence of independent validation data, a common approach to estimating predictive accuracy is based on some form of resampling the original data, e.g., cross-validation. These techniques divide the data into a learning set and a test set and range in complexity from the popular learning-test split to *v*-fold cross-validation, Monte-Carlo *v*-fold cross-validation, and bootstrap resampling. Few comparisons of standard resampling methods have been performed to date, and all of them exhibit limitations that make their conclusions inapplicable to most genomic settings. Early comparisons of resampling techniques in the literature are focussed on model selection as opposed to prediction error estimation (Breiman and Spector, 1992; Burman, 1989). In two recent assessments of resampling techniques for error estimation (Braga-Neto and Dougherty, 2004; Efron, 2004), feature selection was not included as part of the resampling procedures, causing the conclusions to be inappropriate for the high-dimensional setting.

We have performed an extensive comparison of resampling methods to estimate prediction error using simulated (large signal to noise ratio), microarray (intermediate signal to noise ratio) and proteomic data (low signal to noise ratio), encompassing increasing sample sizes with large numbers of features. The impact of feature selection on the performance of various cross validation methods is highlighted. The results elucidate the 'best' resampling techniques for

*to whom correspondence should be addressed

- Much of the conventional wisdom of statistical analysis is focused on inference, not on prediction
- Demonstrating statistical significance of prognostic factors is not the same as demonstrating predictive accuracy
- Predictive models should predict accurately for independent data; the model itself need not be reproducibly derivable on independent data
- Most statistical methods were not developed for prediction problems and particularly not for prediction problems with $>10,000$ variables and <100 cases

Myth

- Development of good predictive classifiers is not possible with >1000 genes and <100 cases or requires huge sample sizes

Sample Size Planning

References

- K Dobbin, R Simon. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics* 6:27-38, 2005
- K Dobbin, R Simon. Sample size planning for developing classifiers using high dimensional DNA microarray data. *Biostatistics* (In Press)

Conclusions

- New technology and biological knowledge make it increasingly feasible to identify which patients are most likely to benefit from a specified treatment
- “Predictive medicine” is feasible but does not mean a different treatment for each patient
- Targeting treatment can greatly improve the therapeutic ratio of benefit to adverse effects
 - Smaller clinical trials needed
 - Treated patients benefit
 - Economic benefit for society

Conclusions

- Achieving the potential of new technology requires paradigm changes in focus and methods of “correlative science.”
- Achieving the potential of new technology requires paradigm changes in partnerships among industry, academia, and government.
- Effective interdisciplinary research requires increased emphasis on cross education of laboratory, clinical and statistical scientists

Acknowledgements

- Post-Docs
 - Kevin Dobbin
 - Aboubakar Maitournam
 - Annette Molinaro
 - Michael Radmacher
 - Sudhir Varma
 - Wenyu Jiang
- Boris Freidlin
- Yingdong Zhao
- BRB-ArrayTools Development Team
 - Amy Lam
 - Ming-Chung Li
 - Supriya Menenzes
 - Michael Ngan