Chaos in the brickyard: Translational research in 2007

David F. Ransohoff MD
Departments of Medicine and Epidemiology
University of North Carolina at Chapel Hill

‘Medicine: Mind the Gap’
NIH seminar series - November 5, 2007
Chaos in the Brickyard

Once upon a time, among the activities and occupations of man there was

And then it came to pass that a misunderstanding spread among the brickmakers (there are some who say that this misunderstanding developed as

BERNARD K. FORSCHER
Mayo Clinic, Rochester, Minnesota

Forscher BK. Science 1963;142:339
Chaos in the brickyard

Forscher BK. Science 1963;142:339

Once upon a time, among the activities and occupations of man there was an activity called scientific research and the performers of this activity were called scientists. In reality, however, these men were builders who constructed edifices, called explanations or laws, by assembling bricks, called facts.
When the bricks were sound and were assembled properly, the edifice was useful and durable and brought pleasure, and sometimes reward, to the builder.
If the bricks were faulty or if they were assembled badly, the edifice would crumble, and this kind of disaster could be very dangerous to innocent users of the edifice as well as to the builder who sometimes was destroyed by the collapse.
And then it came to pass that a misunderstanding spread among the brickmakers.... The brickmakers became obsessed with the making of bricks. When reminded that the ultimate goal was edifices, not bricks, they replied that, if enough bricks were available, the builders would be able to select what was necessary and still continue to construct edifices.
It became difficult to complete a useful edifice because, as soon as the foundations were discernible, they were buried under an avalanche of random bricks.
And, saddest of all, sometimes no effort was made even to maintain the distinction between a pile of bricks and a true edifice.
Topics discussed by Forscher

1. bricks
2. edifices
3. builders vs brickmakers
4. training
Topics discussed by Forscher

1. bricks
2. edifices
3. builders vs brickmakers
4. training

“If the bricks [facts] were faulty... the edifice would crumble....”
Translational Research

Type 1: Apply lab discoveries to studies in humans (bench to bedside)
Type 2: Adopt practices in community

from RFA-RM-06-002:
Institutional Clinical and Translational Science Award (CTSA)
How do we know: Is brick faulty?

Challenge:
Translational research involves different disciplines; ‘rules of evidence’ (to decide ‘when is brick strong’) may vary in different fields.

Nat Rev Cancer 2005; 5;142-9
How do we know: Is brick faulty?

Use *fundamental principles of science.*
Fundamental principles

Cargo Cult Science

by RICHARD P. FEYNMAN

Some remarks on science, pseudoscience, and learning how to not fool yourself. Caltech's 1974 commencement address.

Feynman RP. Engineering and Science 1974(June):10-13
Fundamental principles

“Details that could throw doubt on your interpretation must be given, if you know them.... [I]f you know anything at all wrong, or possibly wrong--to explain it.”

Feynman 1974

Ask: “What could be wrong.”
Fundamental principles

Ask: “*What could be wrong.*”

“... in analyzing theories of antibody formation, Joshua Lederberg (8) gives a list of nine propositions “subject to denial,” discussing which ones would be “most vulnerable to experimental test.”

Chaos in the brickyard:  
Translational research in 2007  

to illustrate problems: ‘markers for cancer’
New York Times, 2.3.04

New Cancer Test Stirs Hope and Concern

By ANDREW POLLACK

Jill Doimer's mother died in 2002 from ovarian cancer, detected too late to be effectively treated.

So Ms. Doimer is eagerly awaiting the introduction of a new test that holds the promise of detecting early-stage ovarian cancer far more accurately than any test available now, using only blood from a finger prick.

Not only does she plan to be tested, but an advocacy group she helped found, Ovarian Awareness of Kentucky, also intends to spread the word to women and doctors.

"If it's going to happen to me or anyone I know, I want it to be caught at an early stage," said Ms. Doimer, who lives in Louisville.

The new test, expected to be available in the next few months, could have a big effect on public health if it works as advertised. That is because when ovarian cancer is caught early, when it is treatable by surgery, more than 90 percent of women live five years or longer. But right now, about three-quarters of cases are detected after the cancer has advanced, and then only 35 percent of women survive five years.

The test is also the first to use a new technology that some believers say could revolutionize diagnostics. It looks not for a single telltale protein — like the prostate-specific antigen, or P.S.A., used to diagnose prostate cancer — but rather for a complex fingerprint formed by all the proteins in the blood. Similar tests are being developed for prostate, pancreatic, breast and other cancers. The technique may work for other diseases as well.

"I've been in cancer research for 40 years and I think it's the most important breakthrough in those years," said Dr.

Continued on Page 6
Running before we can walk?

Two years ago, a new proteomic test was heralded as the future of cancer diagnostics. But since then, doubts about its effectiveness have begun to grow. Erika Check reports.

Seldom does a single piece of research prompt the US Congress to pass a resolution. In November 2002, Correlogic granted licences to two larger firms, Quest Diagnostics and Bioinformatics. They had reason to believe that Liotta and Petricoin’s work was groundbreaking.
Getting the Noise Out of Gene Arrays

Thousands of papers have reported results obtained using gene arrays, which track the activity of multiple genes simultaneously. But are these results reproducible?

When Margaret Cam began hunting for genes that are turned up or down in stressed-out pancreas cells a couple of years ago, she wasn't looking for a scientific breakthrough. She was shopping. As director of a support lab at the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), she wanted to test-drive manufactured devices called microarrays or gene arrays that measure gene expression; she had her eye on three different brands. These devices are hot, as they provide panoramic views of the genes that are active in a particular cell or tissue at a particular time.

The disharmony appears in a striking illustration in Cam's 2003 paper in *Nucleic Acids Research*. It shows a Venn diagram of overlapping circles representing the number of genes that were the most or least active on each device. From a set of 185 common genes that Cam selected, only four behaved consistently on all three platforms—"very low concordance," she said at an August forum in Washington, D.C., run by the Cambridge Healthtech Institute, based in Newton Upper Falls, Massachusetts. Using less rigorous criteria, she found about 30% agreement—but never more than 52% between two brands. "It was nowhere near what we would expect if the probes were assaying for the same genes," she said. She then went on to compare how well the devices worked producing kidney tumor cells, the less significant it seemed.

But those who have persevered with gene expression arrays attribute such problems to early growing pains. They claim that experienced labs are already delivering useful clinical information—such as whether a breast cancer patient is likely to require strong chemotherapy—and that new analytical methods will make it possible to combine results from different experiments and devices. Francis Barany of Cornell University's Weill Medical College in New York City insists that arrays work well—if one digs deeply into the underlying biology.

**Imperfections**

Digging into the biology is just what Cam did after her experiments produced reams of discordant data. She and colleagues in Marvin Gershengorn's group at NIDDK wanted to pick out a set of key genes active in pancreatic tumor cells undergoing differentiation. From there, they meant to go on to examine how islet cells develop.
Markers for cancer

Bricks can be faulty

Past
  • History
  • ‘Threats to validity’ of clinical research

Present
  • RNA expression genomics for cancer prognosis.
  • Serum proteomics for cancer diagnosis.

Future
  • Lessons: making research reliable, efficient
History: Validation of cancer markers is ‘disappointing’ (not reproducible)

1. Non-invasive markers: holy grail of cancer diagnosis
   • carcinoembryonic antigen (CEA)
   • CA-125
   • magnetic resonance spectroscopy of plasma
DETECTION OF MALIGNANT TUMORS

Water-Suppressed Proton Nuclear Magnetic Resonance Spectroscopy of Plasma

ERIC T. FOSSEL, PH.D., JUXTINE M. CARR, M.D., AND JAN McDONAGH, PH.D.

Abstract  A sensitive and specific blood test for cancer has long been sought. The water-suppressed proton nuclear magnetic resonance (NMR) spectrum of plasma is dominated by the resonances of plasma lipoprotein lipids. We measured the mean line widths of the methyl and methylene resonances, which were found to be correlated with the presence or absence of malignant tumors. Values for the average line width were lower in patients with cancer. We analyzed plasma from 331 people (normal controls, patients with malignant and benign tumors, patients without tumors, and pregnant patients); NMR analysis and measurement of line widths were blinded to diagnosis or patient group. The mean line width for 44 normal controls (±SD) was 38.5 ± 1.6 Hz. For 81 patients with untreated cancer, demonstrated by biopsy, the line width was 29.9 ± 2.5 Hz. Patients with malignant tumors were reliably distinguished from normal controls by this method (P < 0.0001). and differed from patients with diseases that did not involve tumors (line width, 36.1 ± 2.6 Hz; P < 0.0001). Patients with benign tumors (e.g., those of the breast, ovary, uterus, and colon) had line widths of 36.7 ± 2.0 Hz and were different from those with malignant tumors (P < 0.0001). However, pregnant patients and those with benign prostatic hyperplasia had line widths consistent with the presence of malignant tumors. The narrowing of lipoprotein-lipid resonances with cancer is consistent with the response of a host to tumor growth.

We conclude that these preliminary results demonstrate that water-suppressed proton NMR spectroscopy is a potentially valuable approach to the detection of cancer and the monitoring of therapy. (N Engl J Med 1986, 315:1369-76.)
History: Validation of cancer markers is ‘disappointing’ (not reproducible)

1. Non-invasive markers: holy grail of cancer diagnosis
   • carcinoembryonic antigen (CEA)
   • CA-125
   • magnetic resonance spectroscopy of plasma

2. Lessons from CEA
   • initial results (PNAS): ~100% sensitivity, specificity for colon cancer
   • high expectations…followed by disappointment
History: Validation of cancer markers is ‘disappointing’ (not reproducible)

1. Non-invasive markers: holy grail of cancer diagnosis
   • carcinoembryonic antigen (CEA)
   • CA-125
   • magnetic resonance spectroscopy of plasma

2. Lessons from CEA
   • initial results (PNAS): ~100% sensitivity, specificity for colon cancer
   • high expectations…followed by disappointment
   • experience led to lessons, ‘rules of evidence’ to evaluate diagnostic tests  
     (Ransohoff and Feinstein. NEJM 1978)
SPECIAL ARTICLE

PROBLEMS OF SPECTRUM AND BIAS IN EVALUATING THE EFFICACY OF DIAGNOSTIC TESTS

DAVID F. RANSOHOF, M.D., AND ALVAN R. FEINSTEIN, M.D.

Abstract  To determine why many diagnostic tests have proved to be valueless after optimistic introduction into medical practice, we reviewed a series of investigations and identified two major problems that can cause erroneous statistical results for the "sensitivity" and "specificity" indexes of diagnostic efficacy. Unless an appropriately broad spectrum is chosen for the diseased and nondiseased patients who comprise the study population, the diagnostic test may receive falsely high values for its "rule-in" and "rule-out" performances. Unless the interpretation of the test and the establishment of the true diagnosis are done independently, bias may falsely elevate the test's efficacy. Avoidance of these problems might have prevented the early optimism and subsequent disillusionment with the diagnostic value of two selected examples: the carcinoembryonic antigen and nitro-blue tetrazolium tests. (N Engl J Med 299:926-930, 1978)
Now, cancer markers are promising

Knowledge of molecular biology provides targets to measure
  • past: knew little about what to target
  • now: know DNA ‘path’ from normal.. adenoma.. cancer

Assays can measure targets
  • past: assays ‘one dimensional,’ like CEA, fecal occult blood testing (FOBT); prostate-specific antigen (PSA)
  • now: assays multi-dimensional; can measure any target
    -DNA - primers and probes, PCR
    -protein - mass spectrometry
Now, cancer markers promising, but..

Mother Nature guards her secrets closely. New reductionist methods mean more data, but not necessarily more knowledge. 

*Rules of evidence have not changed.*

Our job:

- to explore new technologies/fields efficiently
- to avoid predictable mistakes, inflated expectations
- make effort interdisciplinary, translational: molecular biology, biochemistry, etc...
  ... and clinical epidemiology, biostatistics.

*Culture clash may hinder exploration.*
Markers for cancer

Bricks can be faulty

Past
• History
• ‘Threats to validity’ of clinical research

Present
• RNA expression genomics for cancer prognosis.
• Serum proteomics for cancer diagnosis.

Future
• Lessons: making research reliable, efficient
“Validity”

Meaning of “validity” is broad (Lat: “strong”) and confusing; meaning must be clarified.

Two critical threats to validity

1. Chance
   Does chance explain ‘discrimination’?

2. Bias
   Does bias explain ‘discrimination’?

Nat Rev Cancer 2005;5:142-9
A GENE-EXPRESSION SIGNATURE AS A PREDICTOR OF SURVIVAL IN BREAST CANCER

Marc J. van de Vijver, M.D., Ph.D., Yudong D. He, Ph.D., Laura J. van 't Veer, Ph.D., Hongyue Dai, Ph.D., Augustinus A.M. Hart, M.Sc., Dorien W. Voskuil, Ph.D., George J. Schreiber, M.Sc., Johannes L. Peterse, M.D., Chris Roberts, Ph.D., Matthew J. Marton, Ph.D., Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Ph.D., Leonie Delahaye, Tony van der Velde, Harry Bartelink, M.D., Ph.D., Sjoerd Rodenhuis, M.D., Ph.D., Emiel T. Rutgers, M.D., Ph.D., Stephen H. Friend, M.D., Ph.D., and Rene Bernards, Ph.D.

ABSTRACT

Background A more accurate means of prognostication in breast cancer will improve the selection of patients for adjuvant systemic therapy.

Methods Using microarray analysis to evaluate our previously established 70-gene prognosis profile, we classified a series of 295 consecutive patients with primary breast carcinomas as having a gene-expression signature associated with either a poor prognosis or a good prognosis. All patients had stage I or II breast cancer and were younger than 53 years old; 151 had

ADJUVANT systemic therapy substantially improves disease-free and overall survival in both premenopausal and postmenopausal women up to the age of 70 years with lymph-node-negative or lymph-node-positive breast cancer. It is generally agreed that patients with poor prognostic features benefit the most from adjuvant therapy. The main prognostic factors in breast cancer are age, tumor size, status of axillary lymph nodes, histologic type of the tumor, pathological grade, and hormone-receptor status. A large number of other
Gene Expression Signature as a Predictor of Survival in Breast Cancer

Ratio (log scale)

- Metastasis
- Total follow-up
- Lymph node status
- Metastasis
- Death
B  All Patients

Overall Survival

Years

No. AT RISK

Low risk  115  114  112  91  65  43  23
High risk  180  167  134  100  62  40  19

P < 0.001
Strong discrimination led to interpretation as “definitive”

for clinical practice
“... gene-expression patterns of primary tumours are better than available clinicopathological methods for determining the prognosis of individual patients.\textsuperscript{6,10,11}”

Ramaswamy and Perou, Lancet 2003;361:1576-7

for biological research
“... compelling evidence... genetic program of a cancer cell at diagnosis defines its biologic behavior many years later, refuting a competing hypothesis....”

Wooster and Weber, NEJM 2003;348:2339-47
Can chance explain results?

Definition: In multivariable predictive models, overfitting (a problem of ‘chance’) occurs when large N of predictor variables is fit to a small N of subjects. A model may ‘fit’ perfectly by chance, even if no real relationship.

(Simon, JNCI 2003)

Consequence: Results not reproducible in independent group.

Method to check for: Assess reproducibility in independent group.
Can chance explain results?

to the editor:

“In research to validate a prognostic system, the inclusion of 61 patients from the… [training group in the validation group (N=295) means] the validation group is not independent.... [and] the degree of prognostic discrimination may have been inflated....”

(NEJM 2003;348:1716)
How much discrimination when different, independent subjects are assessed?

Degree of discrimination: not as large

Buyse, JNCI 2006
If less discrimination, would interpretation be so strong?

for clinical practice
“... gene-expression patterns of primary tumours are better than available clinicopathological methods for determining the prognosis of individual patients."\textsuperscript{6,10,11}

Ramaswamy and Perou, Lancet 2003;361:1576-7

for biological research
“... compelling evidence... genetic program of a cancer cell at diagnosis defines its biologic behavior many years later, refuting a competing hypothesis....”

Wooster and Weber, NEJM 2003;348:2339-47
To check for overfitting, assess reproducibility in independent group


Figure 1 | Method of dividing original sample to assess reproducibility and overfitting.
Overfitting is not addressed in many studies of RNA expression

Lancet. Feb 5, 2005

Michiels et al.
When studies of RNA expression and prognosis of cancer were ‘reanalyzed’, using original data, in 5 of 7, results were ‘no better than chance.’

Ioannidis, in editorial ("Microarrays and molecular research: noise discovery?") suggests: “validation” groups were not independent.
Overfitting is not addressed in many studies of RNA expression

Lancet. Feb 5, 2005

Michiels et al.
When studies of RNA expression and prognosis of cancer were ‘reanalyzed’, using original data, in 5 of 7, results were ‘no better than chance.’

Ioannidis, in editorial ("Microarrays and molecular research: noise discovery?"), suggests: “validation” groups were not independent.

This problem is readily avoidable.
Chance/overfitting is addressed in study of RNA expression

A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer

Soonmyung Paik, M.D., Steven Shak, M.D., Gong Tang, Ph.D., Chungyeul Kim, M.D., Joffre Baker, Ph.D., Maureen Cronin, Ph.D., Frederick L. Baehner, M.D., Michael G. Walker, Ph.D., Drew Watson, Ph.D., Taesung Park, Ph.D., William Hiller, H.T., Edwin R. Fisher, M.D., D. Lawrence Wickerham, M.D., John Bryant, Ph.D., and Norman Wolmark, M.D.

Chance/overfitting is addressed in study of RNA expression

... because Methods showed ‘independent validation’:

“The prospectively defined assay methods and end points were finalized in a protocol signed on August 27, 2003. RT-PCR analysis was initiated on September 5, 2003, and... data were transferred... for analysis on September 29, 2003.”

Chance as a threat to validity


Opinion

Rules of evidence for cancer molecular-marker discovery and validation

David F. Ransohoff

According to some claims, molecular markers are set to revolutionize the process of evaluating prognosis and diagnosis for cancer. Research about cancer markers has, however, been characterized by inflated expectations, followed by disappointment when original results cannot be reproduced. Even now, disappointment might be expected, in part because rules of evidence to assess the validity of studies about diagnosis and prognosis are both underdeveloped and not routinely applied. What challenges are involved in assessing studies and how might problems be avoided so as to realize the full potential of this emerging technology?

Molecular markers

Reasons for optimism. Molecular markers hold great promise for refining our ability to establish early diagnosis and prognosis, and to predict response to therapy. Optimism about molecular markers is based on exciting new knowledge and new technology. Knowledge about the molecular biology of cancer allows the identification of candidate target markers, such as the mutations that occur during the evolution of colon tissue from normal to adenoma to invasive cancer. Powerful technologies including polymerase chain reaction, serial analysis of gene expression, single-nucleotide polymorphism analysis and microarrays can described briefly, should be considered in similar depth.

Reasons for caution. Although molecular markers will undoubtedly provide advances in diagnosis and prognosis, the degree of success claimed at present is extraordinary. Will we look back in 10 years and find that initial results were not reproducible? In an example from a generation ago, carcinoembryonic antigen (CEA) was purported to be nearly 100% sensitive and specific for colorectal cancer screening in initial research, whereas subsequent research had very different results. History might not necessarily repeat itself, but it indicates caution before making claims of success. The non-reproducibility of the CEA results was due, in large part, to the fact that individuals who were initially studied had extensive cancer, whereas individuals who were later studied had less extensive asymptomatic cancer in which CEA might not have been increased.

The fact that test results vary with the ‘spectrum’ of disease might seem obvious now, but there was little understanding in that era of the concept of spectrum and of the biases that affect research about diagnostic tests. Development of the methods and
Two critical threats to validity

1. **Chance**
   Does chance explain ‘discrimination’?
   (illustration: genomics)

2. **Bias**
   Does bias explain ‘discrimination’?
   (illustration: proteomics)

Nat Rev Cancer 2005;5:142-9
Bias

Definition
Systematic difference between compared groups, so that comparison is ‘erroneous.’

Bias is Serious
• Biases are common in observational research.
• Even one bias can be “fatal.”
Strong claims that serum proteomics can diagnose cancer

Claims:

• for multiple cancers (ovary, prostate, breast)
  • sensitivity: 95-100%
  • specificity: 95-100%

• appeared in Lancet, Clin Chem, WSJ, NBC, PBS, etc.

• led to plans for commercial test, Ovacheck, in 2003;
  but plans delayed by FDA

• led researchers to redirect effort, grant proposals.
Use of proteomic patterns in serum to identify ovarian cancer


Summary

Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Methods Proteomic spectra were generated by mass spectrometry (surface-enhanced laser desorption and ionisation). A preliminary "training" set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analysed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders.

Findings The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 66 cases of non-malignant disease, 63 were recognised as not cancer. This result yielded a sensitivity of 100% (95% CI 93–100), specificity of 95% (87–99), and positive predictive value of 94% (84–99).

Interpretation These findings justify a prospective population-based assessment of proteomic pattern technology as a screening tool for all stages of ovarian cancer in high-risk and general populations.

Lancet 2002; 359: 572–77
Use of proteomic patterns in serum to identify ovarian cancer


Purpose
to diagnose ovarian cancer vs no cancer

Methods
• ovarian cancer, controls
• serum assessed by mass spectroscopy (SELDI-TOF)
• spectra analyzed by ‘genetic algorithm’ (Correlogic)

Lancet 2002; 359: 572–77
A mass analyzer
(Glish, Nat Rev 2003)
Results:

“The discriminatory pattern correctly identified all 50 ovarian cancer cases… [for] a sensitivity of 100%… specificity of 95%…”
Does bias explain some serum proteomics results for ovarian cancer?
(Keith Baggerly’s proposal, as reported in Nature news 2004)

Was bias introduced by ‘run order’ of specimens?

If cancers and non-cancers are run on different days and if the mass spec ‘drifts’ over time, then non-biologic ‘signal,’ associated with Ca vs no-Ca, is hard-wired into results. Bias (signal) is not detected or removed by ‘splitting’ sample into training and validation groups; signal is actually present.
Recent example:

*Bias may explain ‘discrimination’*

Differential exoprotease activities confer tumor-specific serum peptidome patterns


*J Clin Invest 2006;116:271*
Bias may explain ‘discrimination’

Promise
Peptide pattern in serum discriminates PrCa vs control:
~100% sensitive, specific

Interpretation
• Exoprotease activities should be focus of “future peptide biomarker discovery efforts” (authors)
• Low molecular weight “biomarker pipeline is surging with potential” (editorialists)
Bias may explain ‘discrimination’

Compared groups are different:
• Cancer: mean age 67 y.o.; 100% men
Bias may explain ‘discrimination’

Compared groups are different:
• Cancer: mean age 67 y.o.; 100% men
• Control: mean age 35 y.o.; 58% women
Bias may explain ‘discrimination’

Compared groups are different:
- Cancer: mean age 67 y.o.; 100% men
- Control: mean age 35 y.o.; 58% women

Are there other differences (biases)?
Bias is **the** challenge in observational research

Bias is not ‘icing on the cake’; it *is* the cake.

Bias is a large topic, difficult:
- multiple biases; require different methods to address (e.g., randomization, blinding, uniform handling, etc)
- some methods not available in observational research
- some biases may be impossible to identify
- even ONE bias may be fatal

*The ‘process’ to deal with bias is routinely ignored by authors, reviewers, editors in ‘omics’ research.*
Bias is so serious that results are guilty (of bias) until proven innocent.

Innocence is proven by:

**Doing process**
- design - to avoid bias
- conduct - to measure if bias occurred
- interpretation - to determine if important

**Reporting process**
Methods, Results, Discussion

Process to deal with bias of ‘baseline inequality’ in RCT: Report ‘results of randomization’, in ‘Table 1’

<table>
<thead>
<tr>
<th>Variable</th>
<th>Exemestane (N=2362)</th>
<th>Tamoxifen (N=2380)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographic characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age — yr</td>
<td>64.3±8.1</td>
<td>64.2±8.2</td>
</tr>
<tr>
<td>White race — no. (%)</td>
<td>2308 (97.7)</td>
<td>2325 (97.7)</td>
</tr>
<tr>
<td><strong>Nodal status — no. (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>1211 (51.3)</td>
<td>1211 (50.9)</td>
</tr>
<tr>
<td>1–3 Positive nodes</td>
<td>715 (30.3)</td>
<td>706 (29.7)</td>
</tr>
<tr>
<td>≥4 Positive nodes</td>
<td>321 (13.8)</td>
<td>330 (13.9)</td>
</tr>
<tr>
<td>Positive, but no. of nodes missing</td>
<td>5 (0.2)</td>
<td>9 (0.4)</td>
</tr>
<tr>
<td>Unknown</td>
<td>84 (3.6)</td>
<td>96 (4.0)</td>
</tr>
<tr>
<td><strong>Histologic type — no. (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infiltrating ductal</td>
<td>1814 (76.8)</td>
<td>1871 (78.6)</td>
</tr>
<tr>
<td>Infiltrating lobular</td>
<td>346 (14.6)</td>
<td>327 (13.7)</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td>172 (7.3)</td>
<td>156 (6.6)</td>
</tr>
<tr>
<td>Unknown</td>
<td>3 (0.1)</td>
<td>1 (&lt;0.1)</td>
</tr>
<tr>
<td>Missing data</td>
<td>26 (1.1)</td>
<td>28 (1.2)</td>
</tr>
<tr>
<td><strong>Estrogen receptor status — no. (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>1917 (81.2)</td>
<td>1936 (81.3)</td>
</tr>
<tr>
<td>Progesterone-receptor positive</td>
<td>1312 (55.6)</td>
<td>1307 (54.9)</td>
</tr>
<tr>
<td>Progesterone-receptor negative</td>
<td>351 (14.9)</td>
<td>384 (16.1)</td>
</tr>
<tr>
<td>Progesterone-receptor status unknown or missing</td>
<td>254 (10.8)</td>
<td>245 (10.3)</td>
</tr>
<tr>
<td>Negative</td>
<td>26 (1.1)</td>
<td>33 (1.4)</td>
</tr>
<tr>
<td>Unknown</td>
<td>398 (16.9)</td>
<td>392 (16.5)</td>
</tr>
<tr>
<td>Missing data</td>
<td>21 (0.9)</td>
<td>19 (0.8)</td>
</tr>
<tr>
<td><strong>Progesterone-receptor status — no. (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>1320 (55.9)</td>
<td>1333 (55.2)</td>
</tr>
<tr>
<td>Negative</td>
<td>360 (15.2)</td>
<td>395 (16.6)</td>
</tr>
<tr>
<td>Unknown</td>
<td>659 (27.9)</td>
<td>653 (27.4)</td>
</tr>
<tr>
<td>Missing data</td>
<td>23 (1.0)</td>
<td>19 (0.8)</td>
</tr>
<tr>
<td><strong>Type of surgery — no. (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mastectomy</td>
<td>1222 (51.7)</td>
<td>1235 (51.9)</td>
</tr>
<tr>
<td>Breast-conserving</td>
<td>1116 (47.2)</td>
<td>1123 (47.2)</td>
</tr>
</tbody>
</table>
169 specimens were included in this study. The cancer group consisted of 103 serum samples from breast cancer patients at different clinical stages: stage 0 (n = 4), stage I (n = 38), stage II (n = 37) and stage III (n = 24). Diagnoses were pathologically confirmed, and specimens were obtained before treatment. Age information was not available on six of these patients. The median age of the remaining 97 patients was 56 years (range, 34–87 years). The noncancer control group included serum from 25 patients with benign breast diseases (BN) and 41 healthy women (HC). Exact age information was not available from 21 healthy women. The median age of the remaining 20 healthy women was 45 years (range, 39–57 years). The median age of the BN group was 48 years (range, 21–78 years). All samples were stored at −80 °C until use.
“All samples were stored at -80ºC until use.”

OK… but were specimens handled equally in all steps, e.g.,
- time from blood draw to spin/freeze
- number of thaw-freeze cycles
- duration of storage
- type of blood collection tube (red/purple)
- time from thawing to assay
- etc….

Any step is possible source of fatal bias in a proteomics study.
Does serum proteomics diagnose cancer?

**Question:** ‘Where is the landmark study in NEJM, or Science, that shows proof of principle?’ (i.e., convincingly avoids chance, bias)

**Answer:**
- Nov 2007: None exists.
- > Nov 2007: Who knows?
Many sources of bias to explain ‘discrimination’

Before specimens are received, differences occur in demographics, collection methods, etc.

After specimens are received, differences occur in handling: time, ‘place,’ etc.

Specimens received in lab

Cancer

Control

Discipline: study design; clinical epidemiology

Discipline: quality control in laboratory science
Design of ‘experiment’: Bias may be avoided by randomization, other methods

Nat Rev Cancer 2005;5:142-9
Lessons for ‘translational’ research

What kinds of translational research do problems of ‘illustrative example’ apply to?
• ‘omics’ research (genomics, proteomics, etc.)
• discovery-based research?
• research about diagnosis or prognosis in people?
• other?
Lessons for ‘translational’ research

What kinds of translational research do problems of ‘illustrative example’ apply to?
- ’omics’ research (genomics, proteomics, etc.)
- discovery-based research?
- research about diagnosis or prognosis in people?
- other?

Answer:
*Whenever study design or method is observational, not experimental.*
Many sources of bias to explain ‘discrimination’

Before specimens are received, differences occur in demographics, collection methods, etc.

After specimens are received, differences occur in handling: time, ‘place,’ etc.

Cancer

Control

Discipline: study design; clinical epidemiology

Discipline: quality control in laboratory science
Bias as a threat to validity

Bias as a threat to the validity of cancer molecular-marker research

David F. Ransohoff

Abstract: Claims that molecular markers can accurately diagnose cancer have recently been disputed; some prominent results have not been reproduced and bias has been proposed to explain the original observations. As new "omics" fields are explored to assess molecular markers for cancer, bias will increasingly be recognized as the most important "threat to validity" that must be addressed in the design, conduct and interpretation of such research.

specifically, the problem of overfitting (Box 1) — can threaten the validity of molecular-marker research. This Perspectives article considers the even more important problems caused by bias.

Experimental and observational design
As summarized by Hulley and colleagues, a fundamental decision when designing studies for scientific research is "... whether to take a passive role in the events taking place in the study subjects in an observational study, or to apply an intervention and examine its effects on those events in a randomized clinical trial."

The experimental (intervention) method, provides more effective ways to deal with bias than the observational method. In clinical research, the heterogeneity of groups studied might provide particularly problematic sources of bias when groups of participants differ in ways that can affect outcome. By contrast, in a laboratory setting, the subjects might be genetically-identical cell lines.
Topics discussed by Forscher

1. bricks
2. edifices
3. builders vs brickmakers
4. training

“If the bricks were faulty... the edifice would crumble....”
Lessons for ‘translational’ research

In 2007:
1. Faulty bricks exist in many areas of current translational research.
2. ‘Rules of evidence’ can help us test strength of bricks and, ultimately, to build stronger edifices.
3. Future efforts should emphasize ‘faulty brick prevention.’
For future...

1. Role of journals
2. Role of specimens
3. ‘Conceptual framework’: research v development
4. Incentives
5. ‘Attitude’
6. etc.
COMMENTARY

How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design

David F. Ransohoff

@EDRN website (Early Detection Research Network):
http://www.compass.fhcrc.org/edrnncl/bin/studyDesign/guide.asp
For future…

1. Role of journals
2. Role of specimens
3. ‘Conceptual framework’: research v development
4. Incentives
5. ‘Attitude’
6. etc.

Over time we will learn strategies for doing ‘translational research’ that are more reliable, efficient.
Special thanks

**Support:** BRG/DCP/NCI
CPTI/NCI

**Collaborators:** UNC-CH
EDRN/DCP/NCI
CPTI/NCI

**Colleagues:** UNC-CH
NCI
References


Thank You!

Please visit the Medicine: Mind the Gap homepage to view upcoming seminars as they become finalized.

http://consensus.nih.gov/mindthegap

Contact Us:
NIH Office of Medical Applications of Research
Kelli Marciel
marcielk@od.nih.gov | 301.496.4819

Presented by the NIH Consensus Development Program
Celebrating 30 Years of Service