Adventures in Statistics

Multiplicities, P-values, & Observational Studies

Donald A. Berry
dberry@mdanderson.org
Adventures in Statistics

• Clinical trials
  • Part 1: Bayesian bandits to platforms (Sept 28)
  • Part 2: Embracing adaptive Bayesian clinical trials (Oct 26)
  • Part 3: Regulators & CID, Complex Innovative Designs (today)

• Other adventures
  • Attributing ancestry: The corny but $B story of genetics of maize ... & Bayes
  • DNA profiling
  • The science of doping ... or lack thereof
  • BRCAPRO: Risk model for carrying a deleterious mutation of BRCA1 or BRCA2 based on family history ... a marriage of Mendel and Bayes
  • Attributing reductions in breast cancer mortality to treatment vs screening: A Bayesian CISNET model (early ABC—Approximate Bayesian Computation)
  • Multiplicities, p-values, & observational studies
OUTLINE

• My heroes
• John Oliver’s “Scientific Studies”
• Examples from PDQ: Cancer Screening and Prevention
• First responders on 9/11
• Dogs smelling cancer
• Study irreproducibility
• Multiplicities: Ubiquitous and necessary evils
• Case study of hidden multiplicities and biases galore: CYP2D6 and adjuvant tamoxifen
• Predicting future studies a la Bayes
• The Umbrella Man
My Heroes

• January’s Hero: Esther Tomljanovich
• February’s Hero: Mary Claire King
• March’s Hero: Fran Visco, President of NBCC
My Heroes

• September’s Hero: Mae West, on bandit strategies
• January’s Hero: Esther Tomljanovich
• February’s Hero: Mary Claire King
• March’s Hero: Fran Visco, President of NBCC
Prototypic Bandit Strategy—Frequently Optimal

“When choosing between two evils, I always like to try the one I’ve never tried before.”

Mae West

“Klondike Annie” (1936)
My Heroes

• September’s Hero: Mae West, on bandit strategies
• January’s Hero: Esther Tomljanovich
• February’s Hero: Mary Claire King
• March’s Hero: Fran Visco, President of NBCC
My Heroes

• September’s Hero: Mae West, on bandit strategies
• January’s Hero: Esther Tomljanovich
• February’s Hero: Mary Claire King
• March’s Hero: Fran Visco, President of NBCC
• April’s Hero: John Oliver: HBO, “Scientific Studies”
SCIENTIFIC STUDIES
RPT: GLASS OF RED WINE EQUIVALENT TO HOUR OF GYM TIME
Everything we eat both causes and cures cancer

- Wine
- Tomatoes
- Tea
- Milk
- Eggs
- Coffee
- Corn
- Butter
- Beef

Relative risk of cancer

SOURCE: Schoenfeld and Ioannidis, American Journal of Clinical Nutrition

Vox
FIGURE 1. Effect estimates reported in the literature by ingredient. Only ingredients with ≥10 studies are shown. Three outliers are not shown (effect estimates >10).
Cost of Health Care by Country | National Geographic
The Cost of Care

The United States spends more on medical care per person than any country, yet life expectancy is shorter than in most other developed nations and many developing ones. Lack of health insurance is a factor in life span and contributes to an estimated 48,000 deaths a year. Why the high cost? The U.S. has a fee-for-service system—paying medical providers piecemeal for appointments, surgery, and the like. That can lead to unneeded treatment that doesn’t reliably improve a patient’s health. Says Gerard Anderson, a professor at Johns Hopkins Bloomberg School of Public Health who studies health insurance worldwide, “More care does not necessarily mean better care.” —Michelle Andrews

Dollar figures reflect all public and private spending on care, from doctor visits to hospital infrastructures. Data are from 2007 or the most recent year available.
https://www.facebook.com/LastWeekTonight/videos/scientific-studies/896755337120143/

- (2:27-3:20)/19:46
- (3:28-5:00)/19:46
My Heroes

• September’s Hero: Mae West, on bandit strategies
• January’s Hero: Esther Tomljanovich
• February’s Hero: Mary Claire King
• March’s Hero: Fran Visco, President of NBCC
• April’s Hero: John Oliver: HBO, “Scientific Studies”
  • The hero behind the hero: Stuart Buck, Executive Director of the Good Science Project, a non-profit focused on improving science in the US.
Subjective Bayes—Applies to Multiplicities, Statistics Generally, ... and Life

“Information’s pretty thin stuff unless mixed with experience.”

Clarence Day
PDQ® Screening and Prevention Editorial Board

Members of this Board represent the fields of oncology, cancer prevention, cancer screening, hematology, radiology, urology, statistics, epidemiology, and economics. The Screening and Prevention Editorial Board currently meets six times a year to write and update information on cancer screening (early detection) and prevention.

- Jennifer Miller Croswell, M.D., M.P.H., Editor-in-Chief
  National Cancer Institute
  Rockville, MD
- Kevin Ault, M.D.
  University of Kansas Medical Center
  Kansas City, KS
- Donald Berry, Ph.D.
  M.D. Anderson Cancer Center
  Houston, TX
- Timothy Church, Ph.D.
  University of Minnesota
  Minneapolis, MN
- Kevin Dodd, Ph.D.
  National Cancer Institute
  Rockville, MD
- Joann Elmore, M.D., M.P.H.
  UCLA Medical Center
  Los Angeles, CA
- Michael Fiore, M.D., M.P.H., M.B.A.
  University of Wisconsin
  Madison, WI
- Mia Hashibe, Ph.D.
  University of Utah
  Salt Lake City, UT
- Ismail Jatoi, M.D., Ph.D., FACS
  University of Texas Health Science Center
  San Antonio, TX
- Barnett S. Kramer, M.D., M.P.H.
  National Cancer Institute
  Rockville, MD
- Leslie R. Laufman, M.D.
  Blood and Cancer Care of Ohio
  Columbus, OH
- Douglas A. Levine, M.D., FACOG, FACS
  New York University
  New York, NY
- Larissa A. Meyer, M.D., M.P.H.
  M.D. Anderson Cancer Center
  Houston, TX
- Paul Pinsky, Ph.D.
  National Cancer Institute
  Rockville, MD
- Philip Prorok, Ph.D.
  National Cancer Institute
  Rockville, MD
- David Ransohoff, M.D.
  University of North Carolina
  Chapel Hill, NC
- Ian Thompson, Jr., M.D.
  CHRISTUS Santa Rosa Hospital
  San Antonio, TX
- Jill Wruble, D.O.
  Veterans Affairs Medical Center
  West Haven, CT

Updated: July 13, 2020
Factors and Interventions with Adequate Evidence of Little or No Association

Abortion

Abortion has been proposed as a risk factor for breast cancer. Findings from observational studies have varied; some studies showed an association, while other studies did not. Observational studies that support this association were less rigorous and potentially biased because of differential recall by women on a socially sensitive issue.[141-144] For example, the impact of recall or reporting bias was demonstrated in a study that compared regions with different social attitudes on abortion.[145] The Committee on Gynecologic Practice of the American College of Obstetricians and Gynecologists has concluded that “more rigorous recent studies demonstrate no causal relationship between induced abortion and a subsequent increase in breast cancer risk.”[146] Studies that used prospectively recorded data regarding abortion, thereby avoiding recall bias, largely showed no association with the subsequent development of breast cancer.[147-152]
Factors and Interventions with Inadequate Evidence of an Association

Environmental factors
Occupational, environmental, or chemical exposures have all been proposed as causes of breast cancer. **Meta-analyses, describing up to 134 environmental chemicals, their sources, and biomarkers of their exposures, suggest that they may be associated with cancer.**[129,130] Some studies suggest that organochlorine exposures, such as those associated with insecticides, might be associated with an increase in breast cancer risk,[131,132] but other case-control and nested case-control studies do not.[133-138] Studies reporting positive associations have been inconsistent in the identification of responsible organochlorines. Some of these substances have weak estrogenic effects, but their effect on breast cancer risk remains unproven. The use of dichloro-diphenyl-trichloroethane was banned in the United States in 1972, and the production of polychlorinated biphenyls was stopped in 1977. Overall, the epidemiological and animal study evidence that support an association between breast cancer and specific environmental exposures is generally weak. Because so many factors must be considered, any associations with breast cancer or other cancers could be confounded by the analytical problems of multiplicities, measurement challenges, and recall and publication bias.[139,140]
My assessment:

- The data are equivocal.
- Evidence for causation is weak.
- Causation is unlikely.
- This is one more service to society’s penchant for body shaming.
9/11 and cancer
Experts say science lacking on 9/11 and cancer

by DAVID B. CARUSO and MICHAEL STOBBE Associated Press
Tuesday, June 19th 2012

NEW YORK (AP) - Call it compassionate, even political. But ... scientific? Several experts say there's no hard evidence to support the federal government's declaration this month that 50 kinds of cancer could be caused by exposure to World Trade Center dust.

The decision could help hundreds of people get payouts from a multibillion-dollar World Trade Center health fund to repay those ailing after they breathed in toxic dust created by the collapsing twin towers on Sept. 11, 2001.

But scientists say there is little research to prove that exposure to the toxic dust plume caused even one kind of cancer. And many acknowledge the payouts to cancer patients could take money away from those suffering from illnesses more definitively linked to Sept. 11, like asthma and some types of lung disease.

"To imagine that there is strong evidence about any cancer resulting from 9/11 is naive in the extreme," said Donald Berry, a biostatistics professor at the University of Texas MD Anderson Cancer Center in Houston.
9/11 cancer study won't settle debate over risks

Donald Berry, a biostatistics professor at the University of Texas MD Anderson Cancer Center in Houston, said the study has too many limitations to draw any definitive conclusions.

"There's no evidence that 9/11 caused any of these cancers," Berry said.

He pointed out that no increased risks were found for lung cancer — a cancer that might seem plausible after breathing lots of toxic dust and smoke.
Donald A. Berry, PhD, Professor of Biostatistics at The University of Texas MD Anderson Cancer Center, Houston, told The ASCO Post, “The reaction of patients is understandable. People want to explain their cancers. They want to attribute a cause, and not one involving some esoteric random mutation occurring in one of the body’s cells.”

Dr. Berry continued, “The recent media attention seems to be a reaction to the continuing reports of cancer among 9/11 workers. The question is whether the incidence of cancer among the workers is greater than it would be had 9/11 not occurred. I sympathize with the unfortunate people exposed to [the toxins created by] 9/11 who got cancer, just as I do with all cancer patients. But their cancers are about as likely to have been caused by 9/11 as were the cancers of Nebraskans who watched the aftermath of 9/11 on television.”
Dr. Berry, concerned that NIOSH caved to political pressure and public emotion commented, “The notion that 9/11 exposures played a role in causing 50 types of cancers is beyond the pale. With the exception of a heavy dose of radiation, it is difficult for any single event to cause cancer. Moreover, causes of cancer and factors that promote cancer growth are limited in their effects to a small number of cancers. This is true even for smoking, which is implicated in at most a dozen types of cancer.

Asked to comment on the JAMA study, Dr. Berry noted, “The authors highlighted three cancer types as having statistically significantly elevated rates of the 23 cancer types considered. In the context of multiple comparisons, 3 out of 23 is not unusual.”

He continued, “To reinforce the notion that the results are random, the three types—prostate, thyroid, and multiple myeloma—are unrelated, and it’s difficult to imagine a cause that would affect these three and not others. Besides, nine cancer types had a numerically lower incidence than expected. One of those was lung, a cancer that one would expect to be most susceptible to breathing all the bad substances in the air at Ground Zero.”
In the small world of people who train dogs to sniff cancer, a little-known Northern California clinic has made a big claim: that it has trained five dogs -- three Labradors and two Portuguese water dogs -- to detect lung cancer in the breath of cancer sufferers with 99 percent accuracy.

The study was based on well-established concepts. It has been known since the 80's that tumors exude tiny amounts of alkanes and benzene derivatives not found in healthy tissue.

Other researchers have shown that dogs, whose noses can pick up odors in the low parts-per-billion range, can be trained to detect skin cancers or react differently to dried urine from healthy people and those with bladder cancer, but never with such remarkable consistency.

The near-perfection in the clinic's study, as Dr. Donald Berry, the chairman of biostatistics at M. D. Anderson Cancer Center in Houston, put it, "is off the charts: there are no laboratory tests as good as this, not Pap tests, not diabetes tests, nothing."
But experts who read the study could not find any obvious fatal flaw in its methodology, and the idea that dogs can detect cancer is "not crazy at all," said Dr. Ted Gansler, director of medical content in health information for the American Cancer Society. "It's biologically plausible," he said, "but there has to be a lot more study and confirmation of effectiveness."

Dr. Berry, too, was interested but suspicious. "If true, it's huge," he said. "Which is one reason to be skeptical."

Dr. Berry noted, half-jokingly, that Gregor Mendel, the 19th-century discoverer of the laws of genetics, also reported data on his crossbreeding of green and yellow peas that was too good to be true: he repeatedly came up with the perfect 3-1 ratios he predicted. "But we've forgiven Mendel and his gardener," Dr. Berry added, "because his theory turned out to be right."
But experts who read the study could not find any obvious fatal flaw in its methodology, and the idea that dogs can detect cancer is "not crazy at all," said Dr. Ted Gansler, director of medical content in health information for the American Cancer Society. "It's biologically plausible," he said, "but there has to be a lot more study and confirmation of effectiveness."

Dr. Berry, too, was interested but suspicious. "If true, it's huge," he said. "Which is one reason to be skeptical."

Dr. Berry noted, half-jokingly, that Gregor Mendel, the 19th-century discoverer of the laws of genetics, also reported data on his crossbreeding of green and yellow peas that was too good to be true: he repeatedly came up with the perfect 3-1 ratios he predicted. "But we've forgiven Mendel and his gardener," Dr. Berry added, "because his theory turned out to be right."

Likely explanation: the OvaCheck phenomenon
https://www.nature.com/articles/429496a.pdf
Most of the effect is due to
-- multiplicities
-- regression to the mean
We learn by observation, but unless we use good filters most of what we learn is wrong!
ASCO newsletter (all 5 on a “random” day)

General Cancer News

- **High BMI Before Colorectal Cancer Diagnosis Linked to Lower Likelihood of Survival**
  - The Huffington Post (4/10, Chan, 11.54M) reports ...
  - HealthDay (4/10, Preidt, 5K) reports ...

- **Coffee Consumption May Be Linked to a Lower Risk of Hepatocellular Carcinoma**
  - The Huffington Post (4/10, Chan, 11.54M) reports ...
  - HealthDay (4/10, Goodman, 5K) points out ...

- **High-Fat Diet May Increase Risk of Certain Types of Breast Cancer**
  - HealthDay (4/10, Doheny, 5K) reports ...
  - MedPage Today (4/10, Raeburn, 205K) reports ...

- **Women With Irregular Menstrual Cycles May Have Higher Risk of Ovarian Cancer**
  - HealthDay (4/10, Gordon, 5K) reports ...

- **Individuals with Declining Mental Skills May Be Less Likely to Die from Cancer**
  - HealthDay (4/10, Salamon, 5K) reports ...
Researchers must avoid temptations based on observation ... like Ulysses avoiding the bewitching song of the sirens.

Statisticians can help.
Recommendations for Investigators (and Journals)

The following are my recommendations for authors and journals regarding publishing research. They are built upon two overarching principles. One is to have a protocol. The other is to report in detail everything you planned to do and everything you did. There is the need to address silent multiplicities because your report will make everything known!

1. **Need for a protocol.** Prepare a study protocol in advance of the study. The protocol should specify the study’s goals and methods. It should say precisely what analyses will be done. It should specify possible results of these analyses will lead to other analyses etc. It should address how the various multiplicities will be handled. These multiplicities include subset analyses, interim analyses, the various endpoints addressed, the various statistical calculations to be made, including which covariates will be used.

2. **Analyses not done.** Indicate in your report what analyses were proposed, endpoints or subsets were specified in the protocol but not included in the report, and why.

3. **Log of actual analyses.** Keep a log of all analyses, including:

1. **Need for a protocol.**
2. **Analyses not done.**
3. **Log of actual analyses.**
4. **Unspecified analyses.**
5. **To adjust for multiplicities?**
6. **Confirmation.**
7. **Piecemeal publication.**
8. **Biological rationale.**
9. **Frequentist versus Bayesian.**
10. **REMARK (for biomarkers).**
Recommendations for Investigators (and Journals)

5. *To adjust for multiplicities or not?* Adjusting is usually preferable. Assuming that items 1 through 3 above are satisfied, then adjusting is not essential. However, if you do not adjust for multiplicities, then you must provide a rationale. If part of the rationale is that there is a literature supporting your thesis and so your study is one of confirmation, then the need to have a complete literature survey raises additional multiplicity questions, such as how you conducted the search and how you chose which other studies to include, and these too should be described.
P-Values Are Not What Theyre Cracked Up to Be

Donald A. Berry

The ASA is to be congratulated for its “Statement on Statistical Significance and P-values.” Much has been written about p-values in the last 50 years. Many authors have been critical, with pointed warnings about misunderstanding and misinterpreting these strange but ubiquitous beasts. The cumulative impact of such criticisms on statistical practice and on empirical research has been minimal to none. Surprisingly, although statisticians can correctly define p-values and they properly struggle to not overestimate the extent of confidence one can have in a confidence interval, most statisticians do not really understand the issues in applied settings.

Recent attacks on p-values and the role of statistical significance in the “crisis of irreproducibility” has highlighted our lack of understanding. Our collective credibility in the science community is at risk. We cannot excuse ourselves by blaming non-statisticians for their failure to understand or heed what we tell them. The fault for widespread ignorance about statistical significance and for the misuses by substantive scientists of measures we promulgate is ours alone. We must communicate better even if we have to scream from the rooftops, which is exactly what the ASA is doing.

They do not consider applied problems with conclusions of statistical significance when the p-value is less than 0.05, say, but have no inferential content, are scientifically meaningless, and cannot be reproduced.

There is little controversy regarding interpreting p-values as summary statistics for a particular set of data. P-values are handy measures of extremity and serve to describe a set of numbers in a way similar to that of Z-scores and confidence intervals. Errors occur when attributing scientific import to a p-value. For instance, researchers may claim that a small p-value is evidence against the null hypothesis that a treatment is ineffective. The standard Bayesian non-informative-prior data-analytic approach is similar to using p-values for inference but is potentially more dangerous because it ostensibly concludes with a posterior probability of truth.

I will expand here on Principles 1 and 4 of the ASA statement.

The statement gives this “informal” definition: “a p-value is the probability under a specified statistical model that a statistical summary of the data...would be equal to or more extreme than its observed value.” Similarly, Principle 1 indicates that p-
My Black-box Warning

Donald A. BERRY

"Our study is exploratory and we make no claims for generalizability. Statistical calculations such as $p$-values and confidence intervals are descriptive only and have no inferential content."

Recent attacks on $p$-values and the role of statistical significance in the "crisis of irreproducibility" has highlighted our lack of understanding. Our collective credibility in the science community is at risk. We cannot excuse ourselves by blaming non-statisticians for their failure to understand or heed what we tell them. The fault for widespread ignorance about statistical significance and for the misuses by substantive scientists of measures we promulgate is ours alone. We must communicate better even if we have to scream from the rooftops, which is exactly what the ASA is doing.

$p$-values as descriptive statistics are less useful than $p$-values are set of evidence to a decision problem. The standard Bayesian non-informative-prior data-analytic approach is similar to using $p$-values for inference but is potentially more dangerous because it ostensibly concludes with a posterior probability of truth.

I will expand here on Principles 1 and 4 of the ASA statement.

The statement gives this "informal" definition: "a $p$-value is the probability under a specified statistical model that a statistical summary of the data...would be equal to or more extreme than its observed value." Similarly, Principle 1 indicates that $p$-
Recommendations for Investigators (and Journals)

8. *Biological rationale*. This is a knotty issue. Obviously, it is best to have and to give a biological explanation for any empirical observation, especially if the explanation occurred before the observation. The human mind is capable of retrospectively building a rationale for any observation, including when the observation is wrong! However, specifying a biological mechanism, even post hoc, opens the possibility of independent testing through other approaches by other groups.

9. *Frequentist versus Bayesian*. Both approaches are acceptable. The philosophical differences lead to some differences in handling multiplicities in the inferential process, but the basic principles of the need to “adjust” or “shrink” unusual observations are similar. And both have similar attitudes to confirmation and the advantages of prospective studies.
Example of multiplicities ... and biases run amok: The case of *CYP2D6* and adjuvant tamoxifen in breast cancer
Methods
We determined cytochrome P450 (CYP)2D6 (*4 and *6),
13 of the 223 patients were *4/*4
Using strict eligibility requirements (postmenopausal women with estrogen receptor–positive breast cancer, receiving 20 mg/day tamoxifen for 5 years, criterion 1), CYP2D6 poor metabolizer status was associated with poorer invasive disease–free survival (IDFS: hazard ratio = 1.25; 95% confidence interval = 1.06, 1.47; P = 0.009). However, CYP2D6 status was not statistically significant when tamoxifen duration, menopausal status, and annual follow-up were not specified (criterion 2, n = 2,443; P = 0.25) or when no exclusions were applied (criterion 3, n = 4,935; P = 0.38).
CYP2D6 Genotyping and the Use of Tamoxifen in Breast Cancer

Donald Berry

Correspondence to: Donald Berry, PhD, Department of Biostatistics, University of Texas MD Anderson Cancer Center, 1400 Pressler St, 4-5062 Pickens Academic Tower, Houston, TX 77030-1402 (e-mail: dberry@mdanderson.org).

Note

D. Berry led the initial statistical analysis of The International Tamoxifen Pharmacogenomics Consortium of the Pharmacogenetics and Pharmacogenomics Database—as yet unpublished—and dropped out of the consortium because I disagreed with the shift toward basing conclusions on ad hoc subset analyses.

Affiliation of author: Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX.
CYP2D6 Genotype and Adjuvant Tamoxifen

DA Berry

A study by Province et al. for the International Tamoxifen Pharmacogenomics Consortium (ITPC) appeared in a recent issue of this journal.1 In addition to discussing this study and biomarker studies in general, I provide an overall assessment of the evidence regarding testing for CYP2D6 in making clinical decisions regarding the use of tamoxifen in adjuvant breast cancer.

Credibility of biomarker studies

Advances in biomarker research have been hampered by poor understanding. Many biomarkers have been identified, but thorough follow-up has not been conducted, including their priorities. Consequent conclusions therefore lack credibility. Most protocols are inadequate because they are not sufficiently specific regarding analyses to be conducted, including their priorities.

The ITPC analysis1 is subject to multiplicities of every sort, some of which are described below. Post hoc subset analysis but several aspects are relevant. One is the subset’s size, both absolute and relative to the full set. Another is its complexity. A third aspect is the difficulty the investigators had in finding a positive subset, perhaps measured by the number of subsets considered. The more difficult it was to find something,
The credibility of the ITPC analysis

As a member of the ITPC during a two-year period in the middle of its existence starting in mid-2009, I led preliminary versions of the ITPC statistical analyses, including those presented at the San Antonio Breast Cancer Symposium in 2009 (ref. 5 in Province et al.\textsuperscript{1}). Being involved in 2010 when the investigators proposed their criteria 1, 2, and 3, I know why and under what circumstances the criteria were identified, as well as their \textit{post hoc} rationales. Having assumed the role of lead statistician after my resignation from the ITPC in 2011, Province was not exposed to the biases and multiplicities involved in the selection of the criteria and the final end points.
The ITPC analysis\(^1\) is subject to multiplicities of every sort, some of which are described below. *Post hoc* subset analyses are particularly relevant to this paper.

Business researchers understand well a problem called entrapment. “Knowing” a product’s innate beauty, remarkable effectiveness, or obvious usefulness, the entrapped developer dismisses evidence to the contrary and focuses only on the positive. An abundance of negative evidence is required to convince a developer to give up on a product. Sophisticated companies solve this problem by rotating product managers.

Entrapment is a human failing that is prevalent in medical research, in which the “product” is a hypothesis. A researcher who becomes convinced that a hypothesis is true can easily explain away experimental evidence to the contrary, perhaps identifying a subset of supportive data. The problem with such
The ITPC study criteria were based on 13 characteristics with the intention of including only patients who had all 13. In September 2010, I reported to the ITPC investigators that only 96 patients qualified and all 96 were from site 5 (Mayo). They decided to relax the criteria. In this process they knew the numbers of patients at each site who would qualify depending on various sets of the 13 characteristics, and they knew the CYP2D6 comparisons at each site. The final version of criterion 1 required having a particular 9 of the 13 characteristics. More than half of the patients in criterion 1 were from just 2 of the 12 sites: site 5, Mayo (ref. 10 in Province et al.), and site 8, Stuttgart (ref. 9 in Province et al.), which happened to be the two most positive sites regarding the utility of CYP2D6.
Many multiplicities. Just one example:

• “Goals and Eligibility” specified three outcomes: “time to recurrence, disease free survival, overall survival.”

• The ITPC paper\(^1\) does not even mention overall survival.

• Failing to address a prospectively defined end point is an egregious omission.

• I can report that overall survival was not statistically related to CYP2D6 status for any of the criteria considered. ... in the full data set, PMs had the best overall survival of the three metabolism groups ... hazard ratio of 0.92 in comparison with EMs.
Clinical relevance of CYP2D6 testing
In an editorial, I described the clinical implications of the current state of knowledge about CYP2D6 testing. There is no credible evidence that CYP2D6 testing is useful for determining treatment of breast cancer patients, and this is true regardless of how the test is carried out and the characteristics of the patient.
CYP2D6 Genotype and Tamoxifen: Considerations for Proper Nonprospective Studies

MP Goetz¹,² and JN Ingle¹

The International Tamoxifen Pharmacogenomics Consortium (ITPC) was formed to assess the relationship between CYP2D6 genotype and tamoxifen therapy outcomes because of discordant findings in the published literature. The ITPC analyses provided clear insights into the importance of quality control in considering the essential factors that are necessary to answer this pharmacogenetic question and, by extension, the precautions that must be considered for proper retrospective and “prospective-retrospective” studies.

ATAC trial less than 19% (n = 588) of the patients randomized to tamoxifen were analyzed. The reasons for the deviation in HWE were attributed in part to the use of nonstandard PCR techniques and the use of somatic DNA derived from breast tumor cores (instead of germline DNA), contraindicated given the frequent loss of heterozygosity known to occur at the CYP2D6 locus.⁵⁻⁷

By contrast, a secondary analysis of the Austrian Breast and Colorectal Cancer Study Group (ABCSG 8) trial, which compared 5 years of tamoxifen duration, menopausal status, and annual follow-up were not specified (criterion 2, n = 2,443; 49%; HR 1.17, 95% CI 0.90–1.52, P = 0.25) nor when no exclusions were applied (criterion 3, n = 4,935; 99%; HR 1.07; CI 0.92–1.26; P = 0.38). These three criteria were developed to allow analysis of a maximum number of samples but required a progressive loosening of requirements going from criterion 1, which approximates the eligibility requirements intended when the ITPC was originally formed (Supplementary Information online), to
Interpreting the CYP2D6 Results From the International Tamoxifen Pharmacogenetics Consortium

MA Province¹, RB Altman² and TE Klein²

Meta-analysis of the entire analyzable cohort of 4,935 tamoxifen-treated breast cancer patients by the International Tamoxifen Pharmacogenetics Consortium (ITPC) (criterion 3) revealed no CYP2D6 effect on outcomes but strong heterogeneity across sites.¹ However, a post hoc–defined subgroup (criterion 1: postmenopausal, estrogen receptor positive, receiving 20 mg/day tamoxifen for 5 years; n = 1,996) did find a statistically significant effect of CYP2D6 on both invasive disease–free survival as well as breast cancer–free interval, with little site heterogeneity. How should we interpret these discrepant findings?

If the ITPC investigators had been clever enough to have defined the various subgroup criteria a priori (i.e., before any data analysis), then it would be easier to interpret the results. In such an alternative universe, we would reason as follows. The full ITPC sample (criterion 3) shows no significant association between CYP2D6 and treatment response. However, in the (now a priori–defined) criterion 1 subset, CYP2D6 shows a significant effect on outcome. Either the subgroup analysis is a false-positive type I error or it is a true-positive finding and the lack of define the criterion 1 subgroup until after some preliminary analyses had been conducted that did not recapitulate previously published results in much the same data. This lack of replication prompted further investigation of the data, a change in analysis strategy from a mega-analysis to a meta-analysis, and several rounds of additional data cleaning that identified suspicious values, which in turn required going back to each of the 12 source sites to collect more data and either confirm or correct the data, followed by additional analyses. Such an iterative process is
If the ITPC investigators had been clever enough to have defined the various subgroup criteria *a priori* (i.e., before any data analysis), then it would be easier to interpret the results. In such an alternative universe, we would reason as follows.
Unfortunately, it appears impossible to formally adjust the significance levels to account for the "generation" process. But we believe that it is important to report those findings as best we can and to inform the readership as to exactly how they were generated. At the same time, we are extremely sensitive to the concerns of the CODATA-ICSTI Task Group on Data Citation Standards and Practices, which notes that "there is even evidence that researchers engage in data dredging, model fishing, or other methodological peccadilloes in search of results that are 'significant.'" This is why we made a full disclosure of the (admittedly flawed) process by which these subgroups were defined; furthermore, we have exceeded the Task Group's recommendations for data citation by making all the ITPC data and all analysis programs publicly available. To complete the hypothesis-generation experiment, our generated hypothesis about criterion 1 should be taken into an independent data set in which it can be formally tested—as we called for in the conclusion of our paper.
Failure in phase III due largely to extrapolating from a biased selection of phase II trials: Multiplicities & Regression to the Mean
Sometimes it’s easy to tell when a study is not reproducable.
Table 2. Efficacy of the Study Drug at Day 90 or at the Last Rating.

<table>
<thead>
<tr>
<th>Outcome Variable</th>
<th>Placebo Group</th>
<th>NXY-059 Group</th>
<th>Difference between NXY-059 and Placebo†</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% or score (95% CI)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Modified Rankin scale score (primary end point)**

<table>
<thead>
<tr>
<th>No. of patients</th>
<th>849</th>
<th>850</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score — no. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>93  (11.0)</td>
<td>131 (15.4)</td>
</tr>
<tr>
<td>1</td>
<td>170 (20.0)</td>
<td>153 (18.0)</td>
</tr>
<tr>
<td>2</td>
<td>99  (11.7)</td>
<td>97  (11.4)</td>
</tr>
<tr>
<td>3</td>
<td>108 (12.7)</td>
<td>121 (14.2)</td>
</tr>
<tr>
<td>4</td>
<td>175 (20.6)</td>
<td>144 (16.9)</td>
</tr>
<tr>
<td>5 (or death)</td>
<td>204 (24.0)</td>
<td>204 (24.0)</td>
</tr>
</tbody>
</table>

**Change from baseline in total NIHSS score (coprimary outcome)**

<table>
<thead>
<tr>
<th>No. of patients</th>
<th>851</th>
<th>851</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score — LSM ±SE</td>
<td>-1.7±0.5</td>
<td>-1.8±0.5</td>
</tr>
</tbody>
</table>

**Barthel index (dichotomized analysis)**

<table>
<thead>
<tr>
<th>No. of patients</th>
<th>848</th>
<th>850</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score, ≥95 — no. (%)</td>
<td>346 (40.8)</td>
<td>368 (43.3)</td>
</tr>
</tbody>
</table>

**Stroke Impact Scale**

<table>
<thead>
<tr>
<th>No. of patients</th>
<th>676</th>
<th>669</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score — LSM ±SE</td>
<td>63.4±1.1</td>
<td>66.2±1.1</td>
</tr>
</tbody>
</table>

**EuroQoL EQ-5D (weighted index)**

<table>
<thead>
<tr>
<th>No. of patients</th>
<th>816</th>
<th>819</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score — LSM ±SE</td>
<td>0.43±0.013</td>
<td>0.47±0.013</td>
</tr>
</tbody>
</table>

**EuroQoL EQ-5D (VAS)**

<table>
<thead>
<tr>
<th>No. of patients</th>
<th>671</th>
<th>670</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score — LSM ±SE</td>
<td>62.0±0.9</td>
<td>64.5±0.9</td>
</tr>
</tbody>
</table>

Odds ratio 1.20  p = 0.038
SAINT II:

- $N = 3200$ (up from 1700)
- Power 80% for Odds Ratio 1.20
- My predictive probability that SAINT II would be positive: 10%
Press Release

“SAINT II did not meet its primary outcome (p=0.33, odds ratio 0.94) compared to placebo.”

“The company plans no further development of NXY-059 in acute ischemic stroke.”
False-positives ubiquitous in BIG DATA!

How to separate the wheat from the chaff?
Building a prognostic -omic index

• 1550 node-positive breast cancer patients
• 20 markers
• Select markers with disease-free survival p-value < 0.10
• Build a Recurrence Score using multivariate Cox regression
Surviving DFS 10/04

Above median RS

Below median RS

\( p < 0.0001 \)
The punchline:

All 20 markers were white noise!
The good news:

I had a protocol!
OUTLINE

• My heroes
• John Oliver’s “Scientific Studies”
• Examples from PDQ: Cancer Screening and Prevention
• First responders on 9/11
• Dogs smelling cancer
• Study irreproducibility
• Multiplicities: Ubiquitous and necessary evils
• Case study of hidden multiplicities and biases galore: CYP2D6 and adjuvant tamoxifen
• Predicting future studies a la Bayes
• The Umbrella Man
https://www.youtube.com/watch?v=iuoZWb9gqv0

“The Umbrella Man”
Josiah “Tink” Thompson
Author of 6 Seconds in Dallas
Neville Chamberlain & Umbrella ... with Hitler
Chamberlain ... with Hitler ... with Joe Kennedy
APEASEMENT

BRITISH

PRESTIGE
"...and then I said to him " Neville, go back and tell them there'll be peace
Case Not Closed: The Umbrella Man

from Alex Cox PLUS