WASHINGTON UNIVERSITY GRANTS PERMISSION TO USE AND REPRODUCE THE *LIES, DAMN LIES, AND STATISTICS* AS IT APPEARS IN THE PDF AVAILABLE HERE WITHOUT MODIFICATION OR EDITING OF ANY KIND SOLELY FOR END USER USE IN INVESTIGATING RHINOSINUSITIS IN CLINICAL CARE OR RESEARCH (THE "PURPOSE"). FOR THE AVOIDANCE OF DOUBT, THE PURPOSE DOES NOT INCLUDE THE (I) SALE, DISTRIBUTION OR TRANSFER OF THE *LIES, DAMN LIES, AND STATISTICS* OR COPIES THEREOF FOR ANY CONSIDERATION OR COMMERCIAL VALUE; (II) THE CREATION OF ANY DERIVATIVE WORKS, INCLUDING TRANSLATIONS; AND/OR (III) USE OF THE *LIES, DAMN LIES, AND STATISTICS* AS A MARKETING TOOL FOR THE SALE OF ANY DRUG. ALL COPIES OF THE *LIES, DAMN LIES, AND STATISTICS* SHALL INCLUDE THE FOLLOWING NOTICE: "ALL RIGHTS RESERVED. COPYRIGHT 2013. WASHINGTON UNIVERSITY IN ST. LOUIS, MISSOURI." PLEASE CONTACT JAY PICCIRILLO (314-362-8641) FOR USE OF THE *LIES, DAMN LIES, AND STATISTICS* FOR ANY OTHER INTENDED PURPOSE.

"ALL RIGHTS RESERVED. COPYRIGHT 2013. WASHINGTON UNIVERSITY IN ST. LOUIS, MISSOURI."
Lies, Damn Lies, and Statistics
Tools for Critically Evaluating the Medical Literature

Jay F. Piccirillo, MD, FACS

June 21, 2013

WASHINGTON・UNIVERSITY・IN・ST・LOUIS
“Why Most Published Research Findings Are False”

- Published research findings are often refuted by subsequent evidence
- Not surprising, this creates confusion and disappointment
- Refutation and controversy is seen across the range of research designs – from clinical trials and traditional epidemiological studies to the most modern molecular research
- For many current scientific fields, claimed research findings may often simply be accurate measures of the prevailing bias

Additional Reasons

• Study Design
• Poor methodology and greater flexibility in designs, definitions, outcomes, and analytical modes
• Spurious – Chance & Bias
• Small sample size and low study power
• Prior to conducting the study the probability of a true relationship is small
• When effect sizes are small
• When there is a greater number of tested relationships
• When there is greater financial interests and prejudice
• More teams are involved in a scientific field in chase of statistical significance
Methodological Problems
Group Comparability?

- Randomization attempts to control for unknown prognostic factors. In fact, there may be significant differences between groups based purely on chance.

- Most observational and comparative clinical trials include either a table or a paragraph in the text showing the baseline characteristics of the groups being studied.

- One should always inspect these tables.
Recruitment Information

*Significant events following enrollment, but prior to assignment*

400 participants were screened. 175 did not meet criteria.

**Participant Flow: Treatment/Initial Randomization (Period 1)**

<table>
<thead>
<tr>
<th>Milestone</th>
<th>Placebo</th>
<th>Drug X, Low Dose</th>
<th>Drug X, High Dose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STARTED</strong></td>
<td>115</td>
<td>114</td>
<td>110</td>
</tr>
<tr>
<td>Received Intervention</td>
<td>114</td>
<td>110</td>
<td>0</td>
</tr>
<tr>
<td><strong>COMPLETED</strong></td>
<td>104</td>
<td>101</td>
<td>0</td>
</tr>
<tr>
<td>Not Completed</td>
<td>11</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Adverse Event</td>
<td>4</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Lost to Follow-up</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Neutropenia</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

**Participant Flow: Treatment Re-Randomization (Period 2)**

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Drug X, Low Dose</th>
<th>Drug X, High Dose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STARTED</strong></td>
<td>104</td>
<td>51</td>
<td>50</td>
</tr>
<tr>
<td><strong>COMPLETED</strong></td>
<td>99</td>
<td>48</td>
<td>45</td>
</tr>
<tr>
<td>Not Completed</td>
<td>5</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Lost to Follow-up</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Neutropenia</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

"Customized" Reason for Non-completion

400 Assessed for Eligibility

Excluded (n=175)

225 Randomized

115 Placebo
1 Excluded

110 Drug X, Low Dose

104 Analyzed
11 Excluded

101 Analyzed
9 Excluded

51 Drug X, Low Dose

45 Analyzed
5 Excluded

50 Drug X, High Dose

48 Analyzed
3 Excluded
Baseline characteristics display.

<table>
<thead>
<tr>
<th>Measure Type and Dispersion</th>
<th>Arm/Group</th>
<th>Placebo</th>
<th>Drug X</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Participants</strong></td>
<td></td>
<td>195</td>
<td>210</td>
<td>405</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td>54.4 ± 10.2</td>
<td>57.1 ± 12.5</td>
<td>55.5 ± 9.9</td>
</tr>
<tr>
<td>[units: years] Mean ± Standard Deviation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[units: participants]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>101</td>
<td>103</td>
<td>204</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>94</td>
<td>107</td>
<td>201</td>
</tr>
<tr>
<td><strong>Systolic Blood Pressure</strong></td>
<td></td>
<td>128 ± 18.6</td>
<td>126 ± 21.3</td>
<td>127 ± 19.1</td>
</tr>
<tr>
<td>[mm Hg] Mean ± Standard Deviation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Diastolic Blood Pressure</strong></td>
<td></td>
<td>82 ± 9.3</td>
<td>80 ± 8.1</td>
<td>81 ± 8.5</td>
</tr>
<tr>
<td>[mm Hg] Mean ± Standard Deviation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nausea Severity</strong>[^1]**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[participants]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td></td>
<td>52</td>
<td>50</td>
<td>102</td>
</tr>
<tr>
<td>Mild</td>
<td></td>
<td>143</td>
<td>160</td>
<td>303</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

[^1]Zarin Nausea Scale range: 1 (severe) to 10 (none). Severe = 1-3; Mild = 4-9.
Baseline Participant Characteristics*

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Simvastatin (n = 4449)</th>
<th>Atorvastatin (n = 4439)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD), y</td>
<td>61.6 (0.5)</td>
<td>61.8 (0.5)</td>
</tr>
<tr>
<td>Male sex</td>
<td>3597 (80.8)</td>
<td>3590 (80.9)</td>
</tr>
<tr>
<td>Blood pressure, mean (SD), mm Hg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systolic</td>
<td>137.0 (19.9)</td>
<td>136.7 (20.2)</td>
</tr>
<tr>
<td>Diastolic</td>
<td>80.6 (10.2)</td>
<td>80.1 (10.3)</td>
</tr>
<tr>
<td>Body mass index, mean (SD)†</td>
<td>27.3 (3.8)</td>
<td>27.3 (3.9)</td>
</tr>
<tr>
<td>Cardiovascular history</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;1 Previous MIs</td>
<td>756 (17.0)</td>
<td>736 (16.6)</td>
</tr>
<tr>
<td>≤2 mo Since last MI</td>
<td>506 (11.4)</td>
<td>493 (11.1)</td>
</tr>
<tr>
<td>Coronary angioplasty only</td>
<td>877 (19.7)</td>
<td>885 (19.9)</td>
</tr>
<tr>
<td>CABG surgery only</td>
<td>747 (16.8)</td>
<td>732 (16.5)</td>
</tr>
<tr>
<td>Both angioplasty and CABG surgery</td>
<td>163 (3.7)</td>
<td>127 (2.9)</td>
</tr>
<tr>
<td>Cerebrovascular disease</td>
<td>376 (8.5)</td>
<td>353 (8.0)</td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td>195 (4.4)</td>
<td>182 (4.1)</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>244 (5.5)</td>
<td>233 (5.6)</td>
</tr>
<tr>
<td>Atrial fibrillation or flutter</td>
<td>336 (7.9)</td>
<td>347 (7.8)</td>
</tr>
<tr>
<td>Risk factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current smoker</td>
<td>943 (21.2)</td>
<td>892 (20.1)</td>
</tr>
<tr>
<td>Former smoker</td>
<td>2614 (58.8)</td>
<td>2577 (58.1)</td>
</tr>
<tr>
<td>Systemic hypertension</td>
<td>1469 (33.0)</td>
<td>1461 (32.9)</td>
</tr>
<tr>
<td>History of diabetes mellitus</td>
<td>537 (12.1)</td>
<td>532 (12.0)</td>
</tr>
<tr>
<td>Per-randomization statin therapy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simvastatin</td>
<td>2230 (50.1)</td>
<td>2233 (50.3)</td>
</tr>
<tr>
<td>Atorvastatin</td>
<td>512 (11.5)</td>
<td>499 (11.2)</td>
</tr>
<tr>
<td>Pravastatin</td>
<td>431 (9.7)</td>
<td>419 (9.4)</td>
</tr>
<tr>
<td>Other statins</td>
<td>202 (4.5)</td>
<td>157 (4.2)</td>
</tr>
<tr>
<td>Concomitant therapy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspirin</td>
<td>3536 (79.5)</td>
<td>3494 (78.7)</td>
</tr>
<tr>
<td>Warfarin or dicoumarol</td>
<td>559 (12.6)</td>
<td>558 (12.6)</td>
</tr>
<tr>
<td>β-Blockers</td>
<td>3281 (73.7)</td>
<td>3377 (76.1)</td>
</tr>
<tr>
<td>Calcium antagonists</td>
<td>840 (18.9)</td>
<td>882 (19.9)</td>
</tr>
<tr>
<td>ACE inhibitors</td>
<td>1367 (30.7)</td>
<td>1296 (29.2)</td>
</tr>
<tr>
<td>Angiotensin II blockers</td>
<td>270 (6.1)</td>
<td>263 (5.9)</td>
</tr>
</tbody>
</table>

Abbreviations: ACE, angiotensin-converting enzyme; CABG, coronary artery bypass graft; MI, myocardial infarction.
*Data are expressed as No. (%) unless otherwise noted.
†Body mass index was calculated as weight in kilograms divided by the square of height in meters.

### Table 1. Baseline Characteristics of Elderly Persons by Completion Category and Treatment Groups

<table>
<thead>
<tr>
<th></th>
<th>All Randomized</th>
<th></th>
<th>Completers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Vitamin E (n = 311)</td>
<td>Placebo (n = 306)</td>
<td>Vitamin E (n = 231)</td>
<td>Placebo (n = 220)</td>
</tr>
<tr>
<td>Age, mean (range), y</td>
<td>84.9 (65-102)</td>
<td>84.5 (66-103)</td>
<td>84.7 (65-102)</td>
<td>84.3 (66-99)</td>
</tr>
<tr>
<td>Women, No. (%)</td>
<td>228 (73)</td>
<td>220 (72)</td>
<td>176 (76)</td>
<td>162 (74)</td>
</tr>
<tr>
<td>Whites, No. (%)</td>
<td>293 (94)</td>
<td>290 (95)</td>
<td>219 (95)</td>
<td>208 (95)</td>
</tr>
<tr>
<td>Current smoker, No. (%)</td>
<td>17 (6)†</td>
<td>28 (9)</td>
<td>12 (5)†</td>
<td>22 (10)</td>
</tr>
<tr>
<td>Medical history, No. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COPD*</td>
<td>86 (28)</td>
<td>74 (24)</td>
<td>60 (26)</td>
<td>48 (22)</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>116 (37)</td>
<td>97 (32)</td>
<td>80 (35)</td>
<td>63 (29)</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>66 (21)</td>
<td>64 (21)</td>
<td>42 (18)</td>
<td>40 (18)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>151 (49)</td>
<td>166 (54)</td>
<td>122 (53)</td>
<td>117 (53)</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>54 (17)†</td>
<td>71 (23)</td>
<td>39 (17)‡</td>
<td>54 (25)</td>
</tr>
<tr>
<td>Malignancy</td>
<td>26 (8)</td>
<td>32 (11)</td>
<td>22 (10)</td>
<td>24 (11)</td>
</tr>
<tr>
<td>Dementia</td>
<td>164 (53)</td>
<td>142 (46)</td>
<td>127 (55)†</td>
<td>101 (46)</td>
</tr>
<tr>
<td>Alzheimer disease (% of dementia)</td>
<td>33 (20)</td>
<td>39 (27)</td>
<td>29 (23)</td>
<td>26 (26)</td>
</tr>
<tr>
<td>Participants taking NSAIDs, No. (%)</td>
<td>120 (39)</td>
<td>106 (35)</td>
<td>93 (40)</td>
<td>72 (33)</td>
</tr>
<tr>
<td>C-reactive protein, mean (SD), mg/L</td>
<td>6.8 (10)</td>
<td>8.4 (17)</td>
<td>6.2 (8)</td>
<td>7.2 (12)</td>
</tr>
<tr>
<td>Total No. of medications, mean (SD)</td>
<td>7.4 (4.0)</td>
<td>7.4 (4.0)</td>
<td>7.2 (3.8)</td>
<td>7.3 (4.0)</td>
</tr>
</tbody>
</table>

Abbreviations: COPD, chronic obstructive pulmonary disease; NSAID, nonsteroidal anti-inflammatory drug.

*Includes COPD, chronic bronchitis, and asthma.

†P<.10 compared with placebo.

‡P = .04 compared with placebo.
- Significant differences, even in a randomized trial, may exist between groups

- Did authors acknowledge these differences and make adjustments?
Compliance

- After randomization, subjects may not comply with the assigned treatment
- Noncompliance may be overt or covert
  - Overt – dropouts
  - Coverts – failure to admit not taking medication
- Noncompliance reduces any true differences
Compliance -- Example

• VA Study of Treatment of Hypertension
• RCT Clofibrate vs. placebo for lowering cholesterol
• Primary endpoint was mortality
• Compliers and noncompliers were identified

Coronary Drug Project: Five-Year Mortality in Patients Given Clofibrate or Placebo

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mortality, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clofibrate</td>
<td>1,065</td>
<td>18.2</td>
</tr>
<tr>
<td>Placebo</td>
<td>2,695</td>
<td>19.4</td>
</tr>
<tr>
<td></td>
<td>Clofibrate</td>
<td>Placebo</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Mortality, %</td>
</tr>
<tr>
<td>Total</td>
<td>1,065</td>
<td>18.2</td>
</tr>
<tr>
<td>Poor Complier</td>
<td>357</td>
<td>24.6</td>
</tr>
<tr>
<td>Good Complier</td>
<td>708</td>
<td>15.0</td>
</tr>
<tr>
<td></td>
<td>Clofibrate</td>
<td>Placebo</td>
</tr>
<tr>
<td>----------------------</td>
<td>------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Mortality, %</td>
</tr>
<tr>
<td>Total</td>
<td>1,065</td>
<td>18.2</td>
</tr>
<tr>
<td>Poor Complier</td>
<td>357</td>
<td>24.6</td>
</tr>
<tr>
<td>Good Complier</td>
<td>708</td>
<td>15.0</td>
</tr>
</tbody>
</table>
Conclusions Regarding Compliance

- People who do not comply differ from those who do comply.
- Differ in terms of demographic, social, psychological, and cultural ways.
- These differences may have important roles in determining outcome.
- Randomization that reduces bias is essential.
Value of Control Group

Control groups are necessary to be able to draw a causal inference regarding the relationship between treatment and subsequent outcome and to accurately assess treatment effectiveness.
Value of Control Group

• 30 Ménière’s patients were selected for endolymphatic shunt operation
  – 15 active operation
  – 15 placebo operation

• Patients’ symptoms were evaluated each month for next 12 months after intervention

*Ann N Y Acad Sci 1981;374:820-830*
Active Arm Only
Active vs. Placebo

Figure 1. Total pre- and postoperative scores in both groups for all parameters. Broken lines indicate patients with higher scores postoperatively compared with preoperatively. No significant difference between active and placebo.
Second Endolymphatic Shunt Study

- 26 Ménière’s patients were selected for endolymphatic shunt operation
  - 26 active operation
  - 0 placebo operation
  - 19 (73%) completed study

- Mean number of vertigo episodes was significantly reduced from 8.3 to 2.6 times per month ($p=0.006$)

- “ELS significantly improved perception of physical health”

*Laryngoscope 2005;115:1454-1457*
Active Arm Only
Inquisitive Medical Student

Story as told by Dr. Ear Peacock, Chairman of the Department of Surgery at the University of Arizona.

- One day when I was a junior medical student, a very important Boston surgeon visited the school and delivered a great treatise on a large number of patients who had undergone successful operations for vascular reconstruction. At the end of the lecture, a young student at the back of the room timidly asked, “Do you have any controls?” Well, the great surgeon drew himself up to his full height, hit the desk, and said, “Do you mean did I not operate on half of my patients?” The hall grew very quiet then. The voice at the back of the room very hesitantly replied, “Yes, that’s what I had in mind.” Then the visitor's fist really came down as he thundered, “Of course not. That would have doomed half of them to their death.” God, it was quiet then, and one could scarcely hear the small voice ask, “Which half?”
Results can always be improved by omitting controls

Hugo Muensch

Biometrics, 1974
Masking (Blinding)

- Purpose of blinding is to prevent bias associated with subjects' and investigators' expectations
- If interventions are compared with no intervention, an identical placebo may be used
- Compared interventions must be identical in taste, smell, appearance, and mode of administration
- Any difference may destroy the blinding
- Involves several components
  - Subjects not to know which group they are assigned to
  - Observers or data collectors not to know
Example

• 311 employees of the National Institutes of Health volunteered to take 1 gm of Vitamin C or placebo capsules TID for 9 months
• At the onset of a cold, the volunteers were given an additional 3 gm daily of either Vitamin C or a placebo
• 190 volunteers completed the study

JAMA 1975;231:1038
• Dropouts (34% Vitamin C and 44% placebo) were defined as those who missed at least one month of drug ingestion

• Analysis of these data showed that Vitamin C had only a minor influence on the duration and severity of colds

• Effects might be explained by a break in the double blind
<table>
<thead>
<tr>
<th></th>
<th>Suspected Drug</th>
<th>Suspected Drug</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Actual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vitamin C</td>
<td>40 (77%)</td>
<td>12 (23%)</td>
<td>52</td>
</tr>
<tr>
<td>Placebo</td>
<td>11 (22%)</td>
<td>39 (78%)</td>
<td>50</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>51</td>
<td>51</td>
<td>102</td>
</tr>
</tbody>
</table>
The data suggest that the rate of colds was higher in subjects who received Vitamin C but thought they were receiving placebo than in subjects who received placebo but thought they were receiving Vitamin C.

We must be very concerned about the lack of masking or blinding of the subjects and its potential effects on the results of the study, particularly when dealing with subjective endpoints.
Data Analyzed According to Original Study Protocol?

• Terminating a trial early, extending a trial, or reporting interim results may be examples of not analyzing the data according to original protocol and may lead to misleading conclusions.


• Subgroup analyses -- If results are uninteresting, ask the computer to go back and see if any particular subgroups behaved differently.
Celebrex (celecoxib)

• In September 2000, celecoxib long term (1 year) arthritis safety study (CLASS) published in JAMA

• CLASS was reported as a three arm trial comparing ulcer-related complications between celecoxib, ibuprofen, and diclofenac in osteoarthritis or rheumatoid arthritis

• The trial was funded by celecoxib's manufacturer Pharmacia

• Investigators reported 6-month results

JAMA. 2000;284:1247-1255

Months of follow up

Vertical axis

Celecoxib
Diclofenac
Ibuprofen
• 6-month analysis showed significant reduction in ulcer-related complications

• Analysis according to pre-specified protocol at one year showed similar numbers of ulcer-related complications in the comparison groups

• Almost all the ulcer complications that had occurred during the second half of the trial were in users of celecoxib

• These one-year results clearly contradict the published conclusions

• One-year results were available when the manuscript was submitted, but were neither referred to in the article nor reported to *JAMA*
Vioxx Gastrointestinal Outcomes Research (VIGOR)

• Designed primarily to compare gastrointestinal events in patients with rheumatoid arthritis randomly assigned to treatment with rofecoxib (Vioxx) or naproxen (Naprosyn)
• Data on cardiovascular events were also monitored
• Three MIs, all in the rofecoxib group, were not included in the submitted data
• NEJM editors first became aware of the additional MIs in 2001 when updated data were made public by the FDA

*N Engl J Med 2000;343:1520-1528*
• Editors originally believed late events not known to the authors in time to be included in published article
• Yet a memorandum obtained by subpoena in the Vioxx litigation, shows that at least two of the authors knew about 3 additional MIs at least 2 weeks before the authors submitted the first of 2 revisions and 4\(^{1/2}\) months before publication of the article
• It appears that there was ample time to include the data on these 3 additional MIs in the article
• Lack of inclusion resulted in an understatement of the difference in MI risk between rofecoxib and naproxen
The cardiovascular analysis was performed according to a predefined plan developed by Merck prior to the close of the VIGOR study.

Merck indicated that they chose the study termination date to allow sufficient time to adjudicate these events.

Additional MIs in the rofecoxib group and one additional stroke in the naproxen group occurred after the prespecified cardiovascular cutoff date.

Changing the analysis post hoc and after unblinding would not have been appropriate.
Intention-to-Treat vs. On-Treatment Analysis

- Selective reporting of results can affect the risk-benefit profile

- Mortality findings associated with Vioxx (rofecoxib) in clinical trials of patients with Alzheimer’s disease

*JAMA. 2008;299(15):1813-1817*
## Intention-to-Treat

<table>
<thead>
<tr>
<th>Drug</th>
<th>N</th>
<th>Deaths</th>
<th>Rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rofecoxib</td>
<td>1069</td>
<td>34</td>
<td>3.2</td>
</tr>
<tr>
<td>Placebo</td>
<td>1078</td>
<td>12</td>
<td>1.1</td>
</tr>
</tbody>
</table>

ARI=2.1%
RRI=191%
NNH=48
On-Treatment

<table>
<thead>
<tr>
<th>Drug</th>
<th>N</th>
<th>Deaths</th>
<th>Rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rofecoxib</td>
<td>1067</td>
<td>29</td>
<td>2.7</td>
</tr>
<tr>
<td>Placebo</td>
<td>1075</td>
<td>17</td>
<td>1.6</td>
</tr>
</tbody>
</table>

ARI = 1.1%
RRI= 69%
NNH=91
Solutions

To help ensure confidence in industry-sponsored studies, *JAMA* requires

- an author who is independent of the sponsor to take responsibility for the completeness and integrity of all the data in the study
- require an independent statistical analysis for studies in which the data analysis was conducted only by statisticians employed by the sponsor
Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write

H.G. Wells
“Are you just pissing and moaning, or can you verify what you’re saying with data?”
Interpreting Studies

What are the Results?
- Data Display
- Appropriate Statistical Test
- Precision of the Results
- Positive Studies
- Negative Studies
Fig. 1-Interrelations of defects in 232 cases.
For explanation, see text. HSPD = hypospadias; ANEN = anencephaly; EXOS = exomphalos; SP> BIF. = spina bifida; SUA = single umbilical artery; IMP. AN. = imperforate anus; OATEF = oesophageal atresia/tracheal fistula; D.H. = diaphragmatic hernia; O.B.D. = other brain defects.
FIG. 1—Individual plasma glucose concentrations during insulin infusion.
The Visual Display of Quantitative Information

EDWARD R. TUFTE
Beware the Orphan “±” Citation

• A customary method of summarizing data in medical literature is to cite an expression as

   37 ± 8

• Is the 8 standard deviation or standard error of the mean?
% Increases, Changes in % Difference, and Absolute Differences

- If 300 is 50% bigger than 200, is 200 50% less than 300?

- In 2000, Product A had 10% of market share and now it has 15%.

Is that a 50% increase or a 5% increase?
Cost of Imported Drugs

“Importing drugs would save Americans 30 to 300 percent of the cost, but industry sources say that with this discount come fewer safety controls and a risk to the development of new life-saving medicines”
Cost of Imported Drugs

“Importing drugs would save Americans 30 to 300 percent of the cost, but industry sources say that with this discount come fewer safety controls and a risk to the development of new life-saving medicines”

Please explain how it is possible to save more than 100% !
Abstruse Comparisons

• Confidence Intervals and P values denote probabilistic variability or “statistical significance”
• No standard mechanism to express magnitude of comparison or evaluate its impressiveness
• Contrast of two numbers, A and B, can be compared the following ways
  – Direct Increment: A-B
  – Relative Change: \([A-B/B]\)
  – Ratio: A/B

Abstruse Comparisons

• Adverse cardiac events following treatment
  – A: 4.1%
  – B: 2.7%

• Absolute Difference: 1.4%

• Relative Reduction: 34% (1.4%/4.1%)

• Ratio: 1.5

• Number Needed to Treat: 71
Oral Contraceptive Pill (OCP) Scare

- 1995 UK Committee on Safety of Medicine issued warning about 3rd generation OCP
- Two-fold increase risk of life-threatening blood clot in legs or lungs compared to 2nd generation OCP
- News caused great anxiety
- Many women stopped taking OCP

Furedi Human Reproductive Update 1999
How Big Was the Risk?

• “Two-fold increase” or “100% increase”
• Risk with 2\textsuperscript{nd} generation OCP: 1 woman out of 7000
• Risk with 3\textsuperscript{rd} generation OCP: 2 women out of 7000

• Absolute Risk Increase: 1 woman out of 7000
• Relative Risk Increase: 100%
Be Aware

• Reporting of benefits in relative risk (big numbers)
• Reporting of harms in absolute risks (small numbers)
• This asymmetry magnifies benefits and minimizes harm
• Solution?
  Present both benefits and harms in the same format – absolute risks
Interpreting Results of Positive Studies
5 Possible Explanations

• “True-positive” -- Cause-effect

• Real Effects, Just Not Cause-Effect
  – Effect-Cause
  – Confounding

• “False-Positive” – Spurious
  – Chance
  – Bias
Relationship - Causality

An association between two variables is likely to be causal if it is strong, consistent, specific, plausible, follows a logical time sequence, and shows a dose-response gradient.
Cause-Effect

• If the association or effect is real ... 

• Is it clinically meaningful?

• “Who cares?”
Statistical Significance ≠ Clinical Significance

– Give me a sample size large enough and I will obtain a significant $p$ value

– Real, non-random effects may, nevertheless, be very small

– The results of intervention trials should be expressed in terms of the likely benefit an individual could expect
Statistical Significance ≠ Clinical Significance

Overall success rate for 29,102 adults with sinusitis was 90.4% (95% CI 90.0% to 90.8%)

Piccirillo et al. JAMA 2001;286:1849-1856
Statistical Significance ≠ Clinical Significance

Overall success rate for 29,102 adults with sinusitis was 90.4% (95% CI 90.0%-90.8%)
  – Success rate for 17,329 patients who received first-line antibiotic was 90.1%
Statistical Significance ≠ Clinical Significance

Overall success rate for 29,102 adults with sinusitis was 90.4% (95% CI 90.0%-90.8%)
  – Success rate for 17,329 patients who received first-line antibiotic was 90.1%
  – Success rate for 11,773 patients who received second-line antibiotic was 90.8%
Statistical Significance ≠ Clinical Significance

Overall success rate for 29,102 adults with sinusitis was 90.4% (95% CI 90.0%-90.8%)

– Success rate for 17,329 patients who received first-line antibiotic was 90.1%
– Success rate for 11,773 patients who received second-line antibiotic was 90.8%
– Difference of 0.7% (95% CI 0.01%-1.40%; P <.05)

Difference was statistically significant, but clinically meaningless
False-Positive - Chance

• The $p$-value tells you the probability of obtaining the result you obtained (or more extreme) by chance alone

• If the value is small, you conclude that there is a small probability that the differences are due to chance alone

• But, the probability is never zero and so the results may be due to chance
Chance Effects

• UK MRC 12th Acute Myeloid Leukemia (AML12) trial
• Recruited 3459 patients between November 1994 and May 2002
• Research Question – Would an extra course of consolidation therapy confer additional benefit beyond standard four course therapy?

*Controlled Clinical Trials 2003; 24(1): 66-70*
Power and Trial Monitoring

• 1000 patients, 90% power to detect 20% proportional improvement 10% absolute improvement (50% to 60%) in survival

• Yearly trial data monitoring – continue, stop, modify

• Proof beyond reasonable doubt (3 sd; \( p = 0.002 \))
Monitoring Board Decision

- March 1998 review – Data favored five course
  HR 0.47 (CI 0.29 to 0.77, p = 0.003)
- Seek additional follow-up data and recommend not to close study
- September 1998 review – Data favored five course
  HR 0.55 (CI 0.38 – 0.80, p = 0.002)
- Recommend continued enrollment
## Interim Results

<table>
<thead>
<tr>
<th>Timepoint</th>
<th>Deaths/Patients</th>
<th>Odds Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>7/102 v. 15/100</td>
<td>57%</td>
</tr>
<tr>
<td>1998 (1)</td>
<td>23/171 v. 42/169</td>
<td>53%</td>
</tr>
<tr>
<td>1998 (2)</td>
<td>41/240 v. 66/240</td>
<td>45%</td>
</tr>
<tr>
<td>1999</td>
<td>51/312 v. 69/309</td>
<td>33%</td>
</tr>
<tr>
<td>2000</td>
<td>79/349 v. 91/345</td>
<td>20%</td>
</tr>
<tr>
<td>2001</td>
<td>106/431 v. 113/432</td>
<td>11%</td>
</tr>
<tr>
<td>2002</td>
<td>157/537 v. 140/541</td>
<td>-9%</td>
</tr>
</tbody>
</table>
HR plot of mortality in five vs. four courses randomization in the MRC AML 12 trials

—Wheatley at al. Controlled Clinical Trials 24 (2003) 66-70
Why Did Monitoring Board Not Stop Study?

- Observed treatment effect was much greater than that which would be plausibly anticipated for the addition of a single extra course.

- The observed proportional reductions of 53% and 45% were considered to be implausible.

- Chance Effects – “A remarkable fluke”
False Positive Study

- Give me enough studies and I will obtain a positive study.

- Beware of positive study with small sample size – results may be spurious.

- Multiple-Site Pharmaceutical Study example.

- Multiple comparisons.
Multiple-Site Pharmaceutical Study

- Now recruiting medical students, interns, and junior residents
- Money to be earned
Selective Publication

• Reviews from the FDA for studies of 12 antidepressant agents involving 12,564 patients

• Systematic literature search to identify matching publications

• For trials that were reported in the literature, compared the published outcomes with the FDA outcomes

Published, agrees with FDA decision
Published, conflicts with FDA decision
Not published

Positive
N = (38)

Questionable
N = (12)

Negative
N = (24)

Number of studies

Trials Favoring Test Drug More Likely to Be Published

Antidepressant trials conducted by drug companies (n = 74) are more likely to be published in medical journals if their primary outcomes are positive as judged by the FDA. Trials with negative or questionable results are much less likely to appear in literature and, if published, may emphasize positive findings rather than negative conclusions.

According to an internal company document obtained by the Canadian Medical Association Journal, GlaxoSmithKline company officials decided to suppress negative results from one study because, in their words, “It would be commercially unacceptable to include a statement that efficacy had not been demonstrated, as this would undermine the profile of paroxetine.”

CMAJ. 2004;170(5):783
Multiple Comparisons

• Arachinbutyrophobia – fear of peanut butter sticking to the roof of your mouth
• Investigator has discovered a gold-based product for the treatment
• Conducts a three-arm study
  – Placebo
  – Gold Formula #1
  – Gold Formula #2

*Richard Rosenfeld*  Instructional Course on Biostatistics
Difference in mean consumption = 29 grams

$t = 2.44, P = .022$

60 grams of peanut butter consumed

Placebo
$N = 15$

Gold #2
$N = 15$

89 grams of peanut butter consumed
Difference in mean consumption:

- 29 grams
  - t = 2.44, P = .022

- 14 grams
  - t = 1.36, P = .19

- 15 grams
  - t = 1.80, P = .087

- 89 grams of peanut butter consumed

- 75 grams of peanut butter consumed

- 60 grams of peanut butter consumed

Placebo  
N = 15

Gold #1  
N = 15

Gold #2  
N = 15
"Fooling some of the people some of the time is good enough for me."

Bias – Systematic Error

• Systematic difference between the research question and the actual question answered by the study
• May cause the study to give the wrong answer to the research question
• Reader should check consistency with other studies, especially those with different methods
Survival
Advanced Head and Neck Cancer

Treatment options for advanced head and neck cancer include:

– Radiation alone
– Surgery alone
– Chemotherapy alone
– Combined radiation and surgery
– Combined radiation and chemotherapy
– Combined radiation, chemotherapy, and surgery
## Advanced Head and Neck Cancer

<table>
<thead>
<tr>
<th>Prognostic Comorbidity</th>
<th>Initial Treatment</th>
<th>Risk Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>100/534 (19%)</td>
<td>1.0</td>
</tr>
<tr>
<td>Present</td>
<td>38/74 (51%)</td>
<td>3.60 (2.38-5.44)</td>
</tr>
<tr>
<td>Total</td>
<td>138/608 (22%)</td>
<td></td>
</tr>
</tbody>
</table>
Example
Survivor Treatment Selection Bias

• Cohort study of 1000 patients with AIDS

• New treatment, Placebovir, becomes available after the first month of observation

• Placebovir provides no survival benefit
• Overall, 50 of the 1000 patients (5%) die each month regardless of Placebovir use, so that all patients are dead by 20 months of observation

• Beginning after the first month, 10% of patients per month who are alive and not yet receiving treatment initiate treatment with Placebovir

• Thus, at the end of the second month of observation, 100 patients will have died
  – The first 50 died during the first month and were untreated
  – The subsequent 50 patients who died, 5 (10%) had begun Placebovir and 45 were untreated

• As time passes, more patients who are still living will be using Placebovir because, as they survive longer, more of them gain access to the drug
Hypothetical example of survivor treatment selection bias


Annals of Internal Medicine
Effect-Cause

• The outcome has caused the predictor
• Problem in case-control or cross-sectional studies; not so with observational studies
Confounding

• An extrinsic factor involved in the association that is the real cause of the outcome

• A confounding variable is one that is associated with the predictor variable and is a cause of the outcome variable
BCG Vaccination Example I

- Study to examine the effects of bacillus Calmette-Guérin (BCG) vaccination against TB in children from tuberculous families in NYC
- Physicians were told to divide the group of eligible children into a group to be immunized and a control group
- TB mortality almost 5 times higher among controls
- Vaccinations were selectively performed in children from families that were more likely to be conscious of health and related issues

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Deaths, n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vaccinated</td>
<td>445</td>
<td>3</td>
<td>0.67</td>
</tr>
<tr>
<td>Controls</td>
<td>545</td>
<td>18</td>
<td>3.30</td>
</tr>
</tbody>
</table>

*Am Rev Tuberculosis 1946;53:517-532*
BCG Vaccination Example II

- Alternate children were vaccinated and the remainder served as controls
- TB mortality no different between vaccinated and controls

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Deaths, n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vaccinated</td>
<td>556</td>
<td>8</td>
<td>1.44</td>
</tr>
<tr>
<td>Controls</td>
<td>528</td>
<td>8</td>
<td>1.52</td>
</tr>
</tbody>
</table>

*Am Rev Tuberculosis 1946;53:517-532*
Interpreting Results of Negative Studies
Beware Small Sample Size

• Low Power  =  Low Precision
• It’s hard to find significant differences and no difference means nothing
• Big effects may not be significant if sample size is low
• Results are imprecise and would likely vary if study were repeated
Solution

• Look for discussion by the authors of a priori sample size calculation and power determination
  – The importance of beta, the type II error and sample size in the design and interpretation of the randomized clinical trial. Survey of 71 negative trials. *NEJM* 1978;299:690-4

• Look for confidence intervals to determine the range of results consistent with the data
Today's Random Medical News

According to a report released today...
Inferences Concerning Individuals vs. Groups

• Difference between the significance of a particular change for an individual and a change of the same magnitude for a group of patients

• Mean change in BP of 2 mm Hg
  – Individual – trivial difference
  – Group - large reduction in number of strokes

Why is this?

• Individual BP difference of 2 mm Hg is within the error of measurement

• At the Group level, there is a distribution of BP differences— not every individual experiences the same degree of difference
  – Some patients experience much greater reduction
  – Some patients experience much less reduction
  – Some even experience increase
• Variability in individual responses highlights the problem with summarizing treatment effects as a difference in means

• Depending on the distribution of individual differences, the same mean difference can have different implications
• Assume there is threshold below which any change has no clinical implications and mean change for group is below this threshold (ie., “negative” study)

• Distribution of individual change scores
  • Narrow – possible that no patient received benefit
  • Large – possible that substantial number received benefit
Solution

- A clinically meaningful difference is determined for each individual
- The results of two treatments are compared with the clinically meaningful difference as endpoint
- Results reported as % of subjects in each treatment arm achieving minimal clinically important difference rather than the mean values of the continuous variable
Delusions about Cause-Specific Mortality and Disease-Specific Survival

- No standard criteria for making diagnoses
- No provision for the impact of secular changes in diagnostic technology and in nosologic concepts of disease
- No standard criteria for clinically deciding which of several diagnoses is *the* cause of death
- No instruction to clinicians in preparing death certificate
- Allows for biased judgement of cause of death
Higher Survival Does Not Mean Longer Life


*Psychological Science in the Public Interest 2008;8(2):53-96*
• Data from 2000
  – 49 British men per 100,000 diagnosed with prostate cancer
  – 21 or 43% survived 5 years

  – USA 82% survived 5 years

• These numbers are meaningless when comparing groups of people that differ dramatically in how the diagnosis is made
How Prostate Cancer is Diagnosed

• Britain – symptoms
• USA – PSA screening

• To learn which country is doing better you need to look at mortality rates
• 5-Year Survival Rate = number of patients diagnosed with cancer still alive 5 years after diagnosis/number of patients diagnosed with cancer

• Annual Mortality Rate = number of people who die from cancer over 1 year/number of people in the group

• Key difference are the words *diagnosis* and *diagnosed*
Screening Profoundly Biases Survival

- Affects the timing of diagnosis
- Affects the nature of diagnosis by including people with nonprogressive cancer
Without screening

Cancer diagnosed because of symptoms at age 67

Cancer starts -> Dead at age 70

5-year survival = 0%

With screening

Cancer diagnosed because of screening at age 60

Cancer starts -> Dead at age 70

5-year survival = 100%
Without screening

1,000 people with progressive prostate cancer
5 years later
5 year survival = \( \frac{440}{1,000} = 44\% \)
440 alive
560 dead

With screening

2,000 people with nonprogressive cancer
1,000 people with progressive prostate cancer
5 years later
5 year survival = \( \frac{2,440}{3,000} = 81\% \)
2,000 alive
440 alive
560 dead
• Due to overdiagnosis and lead-time bias, changes in 5-year survival rates have no reliable relationship to changes in mortality.

• Correlation between changes in 5-year survival and mortality for the 20 most common tumors over the past 50 years?
• Due to overdiagnosis and lead-time bias, changes in 5-year survival rates have no reliable relationship to changes in mortality

• Correlation between changes in 5-year survival and mortality for the 20 most common tumors over the past 50 years? \( r = 0.0 \)

*JAMA 2000;283:2975-2978*
\[ r = 0.0 \]
• Due to overdiagnosis and lead-time bias, changes in 5-year survival rates have no reliable relationship to changes in mortality.

• Correlation between changes in 5-year survival and mortality for the 20 most common tumors over the past 50 years -- 0.0

  *JAMA 2000;283:2975-2978*

• Knowing about changes in survival tells you nothing about changes in mortality.
What is the Real Story?

- Mortality Risk
  Number of Prostate Cancer Deaths/100,000
  - Britain 27
  - USA 26

- Looking at the incidence and mortality data together suggests that many American men have been unnecessarily diagnosed with prostate cancer due to PSA and have undergone unnecessary surgery & radiation therapy
How Prostate Cancer is Diagnosed

• Britain – symptoms
• USA – PSA screening

• To learn which country is doing better you need to look at mortality rates
Today's Epidemiologist

June 2001 Issue
Vol. 1, No. 1

Confessions Of An Index Case
Time, Place, and Person Revealed!

Don't Ask, Don't Tell
Double Blinding In Military Studies

Boosting Your Confidence Intervals

Do You Have Survey Phobia?
Take Our Quiz And Find Out...

Gay Epi Lifestyles: Cross-Over Designs

Absent Sex Life?
Lucky You! 38 Sexually Transmitted Diseases You Won't Get

NEW DIET
Lowers P Values

Do's & Don'ts of Epi-Quette

Adjusting For Sex In 10 Easy Lessons
What The Kama Sutra Failed To Mention

Sensitivity Analysis
Are You Really A New Man?
Scott's parabola: the rise and fall of a surgical technique

PROMISING IDEA

Strong media pressure for universal acceptance

Widespread enthusiasm

”Of possible value—but only as a research tool”

Standard treatment

General introduction

Doubts creep in

Damaging survey reported

Condemned by several authorities

Falls into disuse

Operating theatre staff ponder possible uses for large quantities of expensive, obsolete equipment

Widely publicised medicolegal case

Used only in highly specialised circumstances

Very old surgeons amaze their juniors with rollicking stories of the old days

J W Scott consultant gynaecologist, Poole Hospital NHS Trust, Poole, Dorset BH15 2JB

Thank you

Jay F. Piccirillo, MD
http://otooutcomes.wustl.edu/Pages/index.aspx
piccirilloj@ent.wustl.edu
314.362.8641