

**Pomology**  
(fruits and nuts; excludes vegetables)

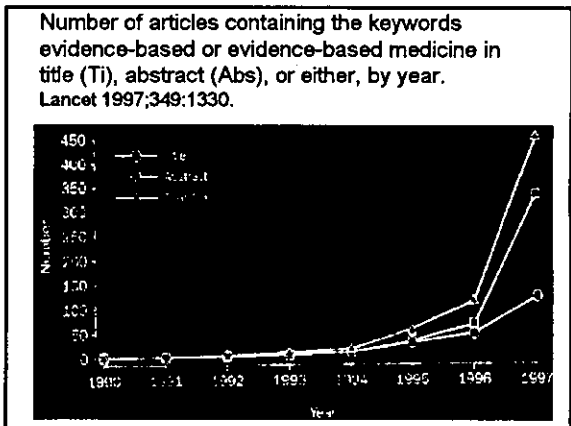
- value of agricultural products in California is 18 billion dollars
- pomology is 2.3 billion
- lots of small studies; each data point is expensive; there may be many unpublished studies
- perennial crops; influenced by environment and aging

**Evidence Based Medicine**

- Evidence Based Health Care is being used increasingly in the creation of health care and public policy
- Meta-analysis has a special appeal for decision-making in public health and clinical practice because it provides a method for using existing information.

**Pomology**  
(fruits and nuts; excludes vegetables)

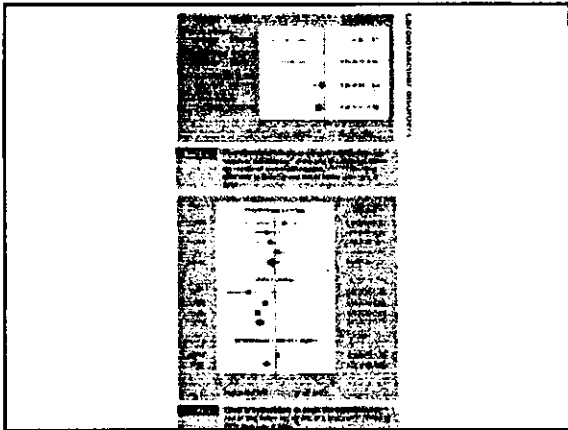
- **strawberries:**
  - appearance
  - firmness
  - yield
  - fruit size
- **other measures**
  - flavor
  - color
  - time to production
  - defect types (e.g., split fruit)
  - disease and pest resistance
  - shelf life



**Evidence Based Med**

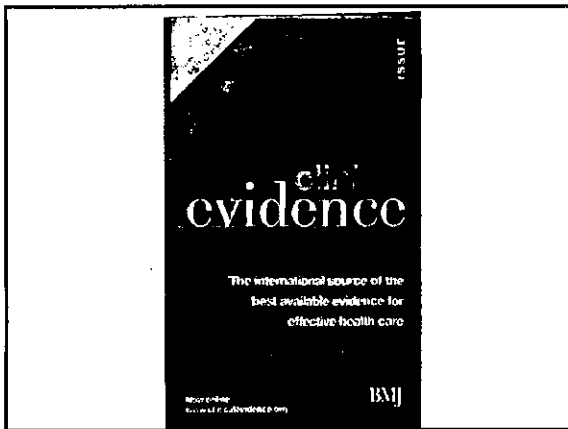
**Why do we need EBM now?**

- Better accountability
- Need to quantify outcomes
- Many effective treatments (alternatives)
- Costs
- Expectations
- Information overload/need for synthesis
- Changing patient-physician relationship



One common E.B.M. approach is meta-analysis: collating data from far-flung studies to come up with a definitive answer to a medical question. Such studies are overthrowing some conventional wisdom. Mammogram screenings? They don't save lives. Remember the placebo effect? It doesn't exist. E.B.M. is also credited with validating some simple cures. Most people know that if you have a heart attack, you should immediately take an aspirin. Thank an E.B.M. study for proving that this works.

After colds, the second-most-common reason for a visit to the doctor is lower-back pain. The "treatment" has always been bed rest. Why? Because, as a recent article explained, "The notion that rest is therapeutic and will relieve pain dates back to Hippocrates." But now that E.B.M. studies have used science instead of oral tradition to test this notion, they have found that bed rest "may delay return to functional status." What works better? Light exercise and getting back on your feet. This past June, the Agency for Healthcare Research and Quality integrated the no-bed-rest approach into its guidelines. This new standard of care, which will probably save billions of dollars in unnecessary sick leave, marks the end of 2,400 years of misguided treatment.



E.B.M. is yet another idea that can be credited to the computer revolution. Doctors have long known that they learn very little after med school when their exhausting schedules and the baffling profusion of 4,000 monthly professional journals make it nearly impossible to keep up with innovations in treatment. The E.B.M. movement began when six doctors in Canada came up with the idea of skimming the most dependable studies and crunching the results into an accessible, reliable database.

Indeed, in the wake of E.B.M., journals are filling with terms that sound almost anthropological to describe longstanding treatments: "local custom," "witch-doctoring," "myth." Or as one article this fall put it, "This process of examining beliefs that have been based primarily on teaching and empirical experience rather than evidence has been compared to stripping the curtain away from the Wizard of Oz to reveal an ordinary man." JACK HITT

**Evidence-Based Medicine** When visiting our family doctor, most of us feel secure in the belief that modern science has purged medicine of such practices as cupping and bloodletting. But according to a recent article in the journal *Patient Care*, "Some experts estimate that only 20 percent of medical practices are based on rigorous research evidence." The rest are based on what has been published in books repeatedly without independent testing — or what doctors have always said should work. In other words, it's a kind of folklore.

A revolution is erupting in the wards of practical medicine these days, one defined recently by *The British Medical Journal* as "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients." The revolution is called evidence-based medicine, or E.B.M., and many traditional treatments are being run through the machinery of the scientific method — and being found wanting.

## Experts

## Evaluation Alcohol – Cancer

### Criteria

(A) Degrees of evidence for carcinogenicity to humans and to experimental animals and supporting evidence

#### 1. Human carcinogenicity data

- Sufficient evidence
- Limited evidence
- Inadequate evidence
- Evidence suggesting lack of carcinogenicity

## Overall Evaluation Into Groups

- I: agent is carcinogenic to humans
- II: (a) agent is probably carcinogenic to humans  
(b) agent is possibly carcinogenic to humans
- III: agent is not classifiable as to its carcinogenicity to humans

#### 2. Experimental carcinogenicity data

- Sufficient evidence
- Limited evidence
- Inadequate evidence
- Evidence suggesting lack of carcinogenicity

#### 3. Supporting evidence of carcinogenicity

- Other relevant data

#### (B) Overall evaluations

## Overall Evaluations

- During the last ten years there have been 21 studies on the relation between alcohol consumption in relation to risk of breast cancer.
- There is strong evidence to support a dose-response relation in both the case-control and follow-up epidemiologic data.
- We interpret these findings not as proof of causality, but as strongly supportive of an association between alcohol consumption and risk of breast cancer.

## Sufficient evidence of carcinogenicity

The Working Group considers that a causal relationship has been established between the agent and increased incidence of malignant neoplasms or of an appropriate combination of benign and malignant neoplasms (as described on pp. 13-14) (a) in two or more species of animals or (b) in two or more independent studies in one species carried out at different times or in different laboratories or under different protocols

A significant positive association between alcohol intake and breast cancer incidence was seen in each of four large prospective studies and in seven of 13 case-control studies. In two of the case-control studies with a limited number of subjects, positive associations of similar magnitude were observed. A dose-response relationship generally with up to 1.5- to two-fold risks, has been observed. The consistency of this positive association makes it unlikely that the relationship is due to chance or methodological bias.

## Single Experiment

### Gordon, Loveland, Cureton - 1952

Duplicate studies, and studies bearing on the same essential problem but based on samples from non-identical populations, often lead to tests of significance whose associated probabilities are quite unequal. If each sample is small, moreover, the null hypothesis may not be rejected with any greater certainty on the basis of any one study. But if the results of the several studies are all or nearly all in the same direction, there is additional evidence against the null hypothesis. In such cases we may wish to base a new test of significance on the combined data from all the studies.

### Journal of Weather Modification Clement Todd - 1986

Some people are hoping for the day when an elegantly designed experiment, brilliantly evaluated, will make precipitation management credible to those who look at the field from the outside, but it doesn't appear as though credibility will come that way. Rather it will come from reporting on information assembled from a large number of imperfect experiments and operations, that show, even though there is diversity in the way different types of clouds respond to seeding, similar clouds respond similarly around the world.

### Egon Pearson - 1950

In statistical practice it often happens that we wish to combine the results of a number of independent experiments which have all been planned to test a common hypothesis. Thus, for example, several experiments comparing two treatments may have been carried out, but owing to differences in error variance or to other changes in conditions between experiments, it is not possible to pool all the data together. The overall test calls therefore for the combination of a number of independent tests of significance.

## A Single Experiment

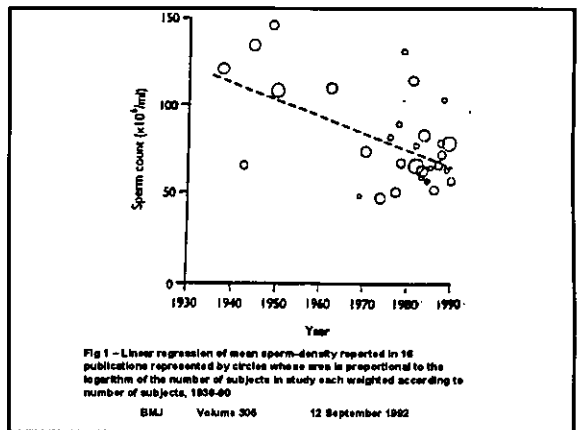
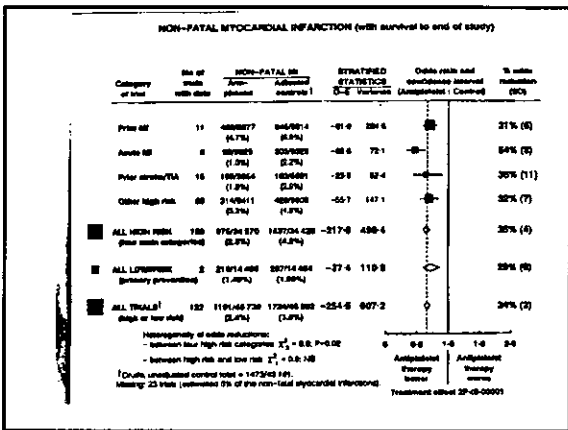
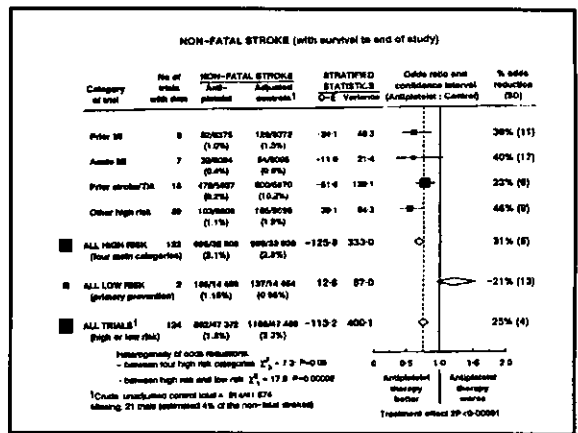
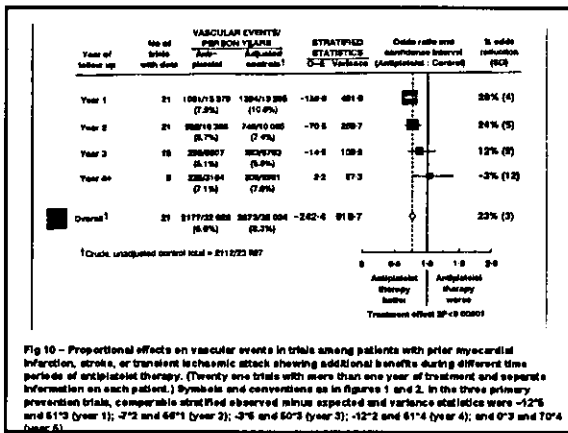
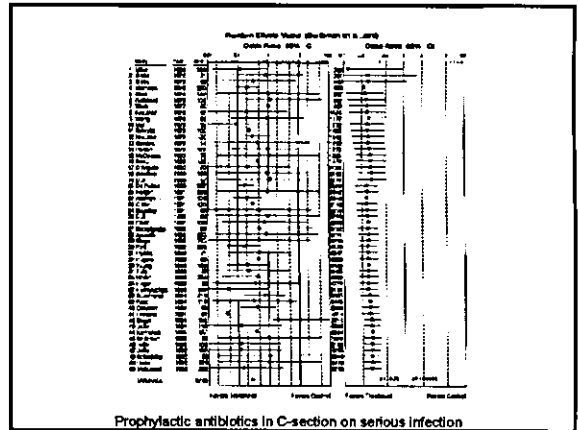
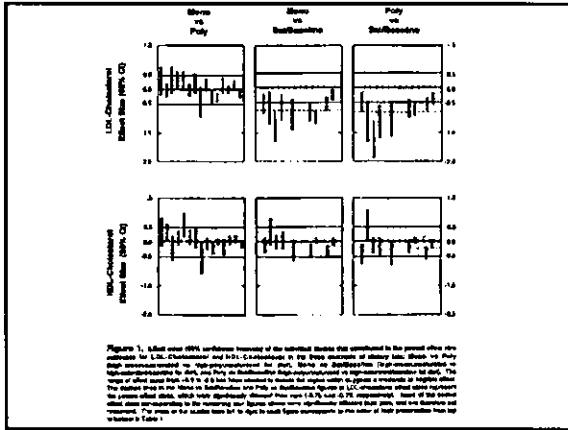
### Primary Analysis

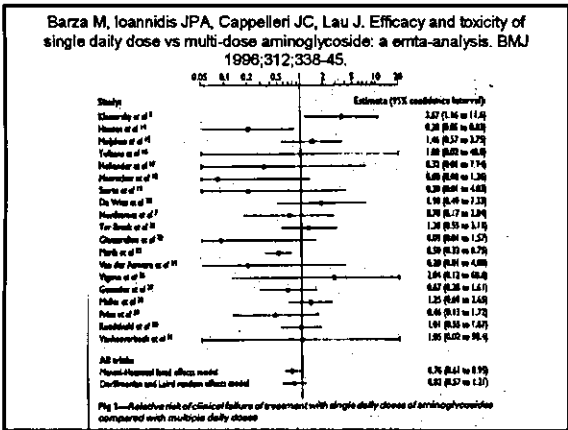
A study of 237 individuals showed a positive correlation between coffee drinking and pancreatic cancer.

### Secondary Analysis

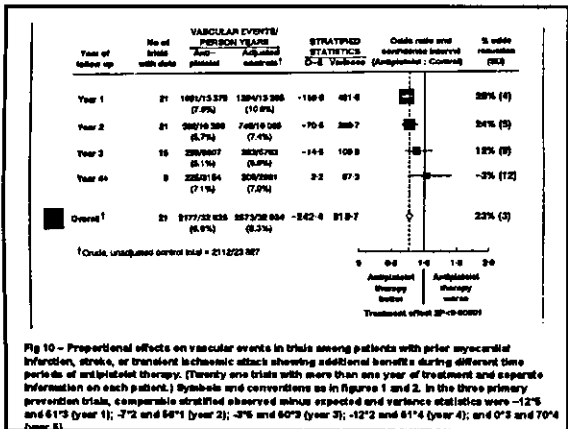
A re-examination of the data showed that 82 patients in the study had a history of gastrointestinal problems, which contaminated the results. Of the remaining 155 patients, there was no indication of a correlation between coffee and pancreatic cancer.

## Meta Examples





- ### Problem Areas & Research Areas
- Publication bias
  - Selection bias
  - Blinded vs nonblinded
  - RCT's vs observational studies
  - Missing observations
  - Equating measures
  - More general linear models
  - Reporting of data in journals



- ### Empirical evidence that trial results are related to methodological quality
- Sacks et al. (1982)** compared 50 RCTs and 56 HCTs.
- 20% of RCTs but 79% of HCTs had significant treatment benefit

### Biases in MA

- Schulz et al. (1995)** examined 250 RCTs from 33 meta-analyses.
- treatment effect (odds ratio) was 30% - 41% larger in trials which did not report adequate concealment of treatment allocation
  - 17% exaggeration in trials that were not double blind

**Khan *et al.* (1996)** looked at 9 randomized trials in infertility.

- Seven trials with low methodological quality found a benefit of treatment (OR=2.6)
- Two high-quality trials found treatment was associated with harm (OR=0.5)

**Do certain countries produce only positive results? A systematic review of controlled trials.**

Vickers A, *et al.* *Control Clin Trials* 1998;19:159-166

**Controversy that may arise from quality assessment: the case of screening mammography RCTs**

- Gotszche and Olsen. *Lancet* 2000;355:129
- A 1999 study found no decrease in breast cancer mortality in Sweden, where screening has been recommended since 1985
- Revised methodological quality of mammography trials and repeated a meta-analysis

**Do certain countries produce only positive results? Results of controlled clinical trials of interventions other than acupuncture by country**

Country	Abstracts screened	Abstracts included	Favoring Test Treatment	
			Number	Percentage
China	196	109	108	99
England	329	107	80	75
Japan	317	120	107	89
Russia	180	29	28	97
Taiwan	78	40	38	95
<b>Total</b>	<b>110</b>	<b>405</b>	<b>361</b>	<b>89</b>

**Relative risk of death from breast cancer in screening versus control groups**

	Number Randomized		# of deaths from breast CA		Relative Risk (95% CI)
	Screening	Control	Screening	Control	
<b>Randomization adequate</b>					
Malmö	21084	21195	63	66	0.96 (0.68-1.35)
Canada	44925	44910	120	111	1.08 (0.84-1.40)
<b>Total</b>	<b>66013</b>	<b>66105</b>	<b>183</b>	<b>177</b>	<b>1.04 (0.84-1.27)</b>
<b>Randomization not adequate</b>					
Göteborg	11724	1421	18	40	0.55 (0.31-0.95)
Stockholm	40318	19943	96	45	0.73 (0.50-1.06)
Köppenberg	36588	18582	123	104	0.59 (0.45-0.78)
Ostergötland	38491	37403	135	175	0.76 (0.61-0.95)
New York	30151	30565	153	196	0.79 (0.64-0.98)
Edinburgh	22926	21342	158	187	0.87 (0.70-1.08)
<b>Total</b>	<b>182178</b>	<b>142052</b>	<b>654</b>	<b>725</b>	<b>0.78 (0.67-0.93)</b>

**Do certain countries produce only positive results? Results of controlled clinical trials of acupuncture by country of research**

Country	Total Trials Analyzed	Favoring Test Treatment	
		Number	Percentage
USA	47	25	53
China	36	36	100
Sweden	27	16	59
UK	20	12	60
Denmark	16	8	50
Germany	16	10	63
Canada	11	3	27
Russia/USSR	11	10	91
Taiwan	6	6	100
Australia	6	1	17
Japan	5	5	100
Hong Kong	3	3	100

### Negative Results: A Positive Viewpoint

It verges on the scandalous, in the minds of many critics of modern science, that negative results are so seldom published.... But pride and careers in science rarely benefit much from negative results.

The Nobel prize is rarely awarded for producing crucial negative information, no matter how skillful the scientists or how time-consuming the work.

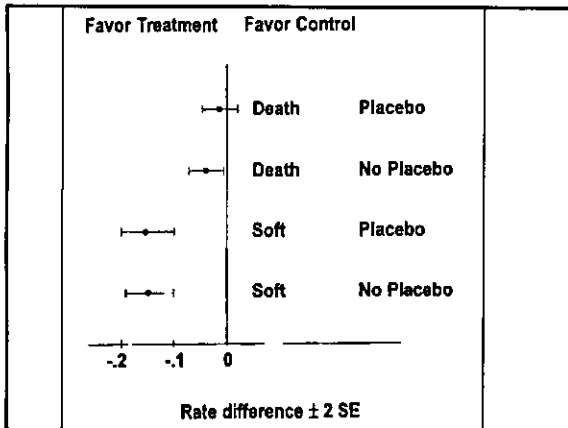
(The Doctor's World – NY Times 1986)

### An early Clinical Trial (N = 2)

In the late 18<sup>th</sup> century, King Gustav III of Sweden decided that coffee was poison and ordered a clinical trial.

J Int Med. October 1991; 289 – Introduction to editorial from Nordic School of Public Health, Goteborg Sweden

Reprinted in Ann Intern Med 1992; 117:30



### Study design

- The king condemned a convicted murderer to drink coffee every day.
- Control: another murderer was condemned to drink tea daily.
- Outcome: death.
- Two physicians were appointed to determine the outcome.

### Early Clinical Trials

### Results

- The two doctors died first.
- The king was murdered.
- Both convicts enjoyed long life until the tea drinker died at age 83. (No age was given for the coffee drinker.)



# Non-Parametric p-Values

Nonparametric Measures of Treatment vs. Control				
Study :	1	2	...	k
p-values:	$p_1$	$p_2$	...	$p_k$
Fisher:	$-2 \sum \log p_i$			$\chi^2_{2k}$
Stouffer :	$\frac{\sum z_i}{\sqrt{k}}$ , $z_i = \Phi^{-1}(p_i)$ ,			N(0,1)
Tippett :	Min( $p_1, \dots, p_k$ )		uniform distribution	

- ## Methods for Combining
- Nonparametric methods
  - Parametric Models
    - Normal distributions
    - Binomial distributions
    - Correlations

## Use of p-values

Fisher  
 Use of anti-coagulants in myocardial infarction  
 Chalmers, Motta, Smith, Kunzer (1977)

low-level lead exposure & IQ of children  
 Needleman & Gatsonis (1990)

Stouffer  
 toxic shock syndrome  
 Michel & Hall (1982)

## Fisher on Combining Probabilities From Tests of Significance

When a number of quite independent tests of significance have been made, it sometimes happens that, although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance. It is sometimes desired, taking account only of these probabilities, and not of the detailed composition of the data from which they are derived, which may be of very different kinds, to obtain a single test of the significance of the aggregate, based on the product of the probabilities individually preserved.

Statistical Methods For Research Workers, R.A. Fisher  
 Section 21.1, page 99, 10th edition, 1932

University Division	Availability Proportion of Women with Ph.D's	Total Hires	Female Hires	Level of Significance
Arts-Humanities	.383	48	14	.097
Human Development	.385	32	12	.452
Management Administration	.043	26	0	.319
Natural Sciences	.138	38	1	.025
Social Sciences	.209	34	6	.319
		178	33	

Court Case		
p-value	-2 log p	normal
.097	1.013	1.30
.452	.345	.12
.319	.496	.47
.025	1.602	1.96 *
.319	.496	.47

---

$\min p = .025 = 1 - (1 - (\alpha))^{1/6}$   
 yields  $(\alpha) = .12$

---

$-\sum 2 \log p = 3.952 \quad \chi^2_{10} = .06 = \alpha$

---

normal  $\frac{z_1 + \dots + z_k}{\sqrt{k}} = \frac{4.32}{\sqrt{5}} = 1.932 \quad \alpha = .03$

---

\* is only single significant value  
 could use other analyses, e.g., Mantel-Haenszel