EBM: a useful instrument to verify and evaluate uncertainty

John P.A. Ioannidis

Professor and Chairman, Department of Hygiene and Epidemiology, University of Ioannina
School of Medicine, Ioannina, Greece

Professor of Medicine (adjunct), Tufts University School of Medicine, Boston, USA

Roma, November 2006

Why is there uncertainty?

- There is bias
- There is random error
- There is both
- We are not in paradise yet

A key aim of EBM is to evaluate and if possible minimize the impact of bias

- Bias is any non-random deviation from the truth
- Conscious, subconscious, or unconscious
- One may create theory (or theories) about bias or may study its consequences
- The former seem more robust, but it is the latter that we measure, witness, and eventually suffer

Tackling bias in systematic review of the evidence

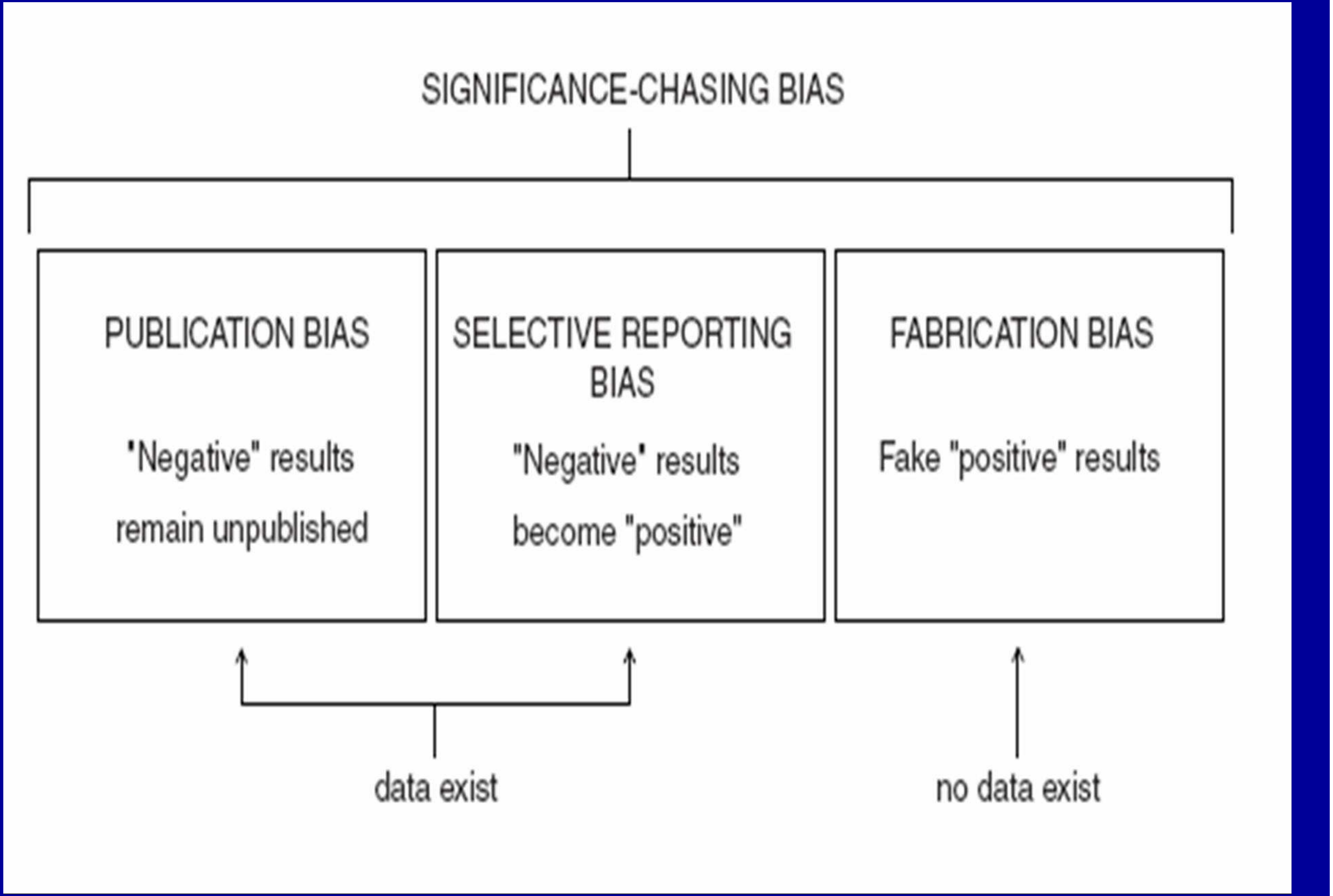
- Systematic review of the evidence is the prime opportunity to detect and discuss biases in the constituent studies that form the evidence
- It is also a primary opportunity for addressing the potential bias that affects the specific scientific field at large

Field-wide issues in evidence synthesis

- Selection biases
- Early vs. late evidence
- Large vs. small studies
- Different study design effects
- “Quality” effects
- Heterogeneity and subgroups
- Overall validity of the field research findings

Selection biases

- Publication bias
- Time lag bias
- Selective outcome reporting bias



SIGNIFICANCE-CHASING BIAS

PUBLICATION BIAS

"Negative" results remain unpublished

SELECTIVE REPORTING BIAS

"Negative" results become "positive"

FABRICATION BIAS

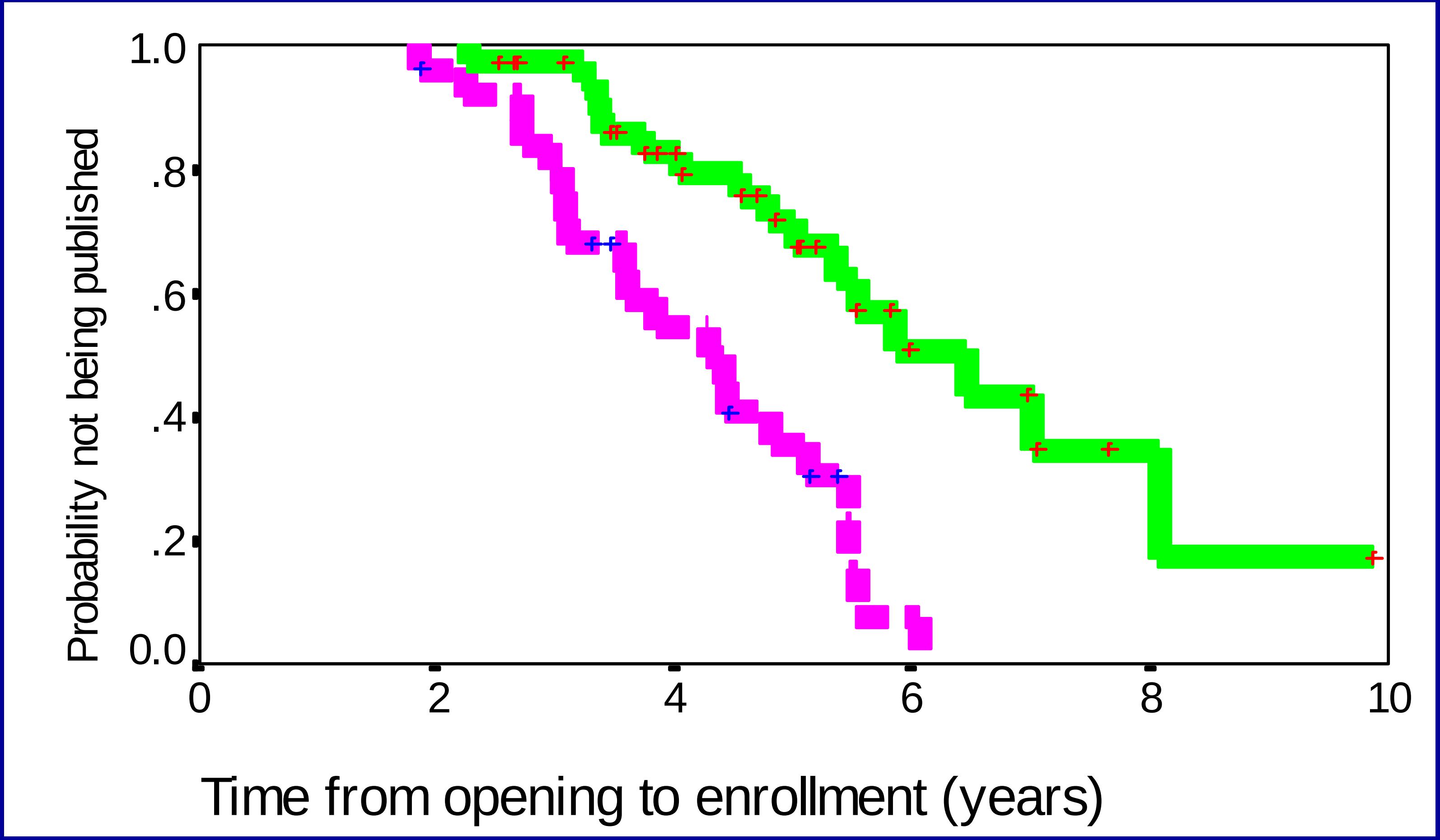
Fake "positive" results

data exist

no data exist

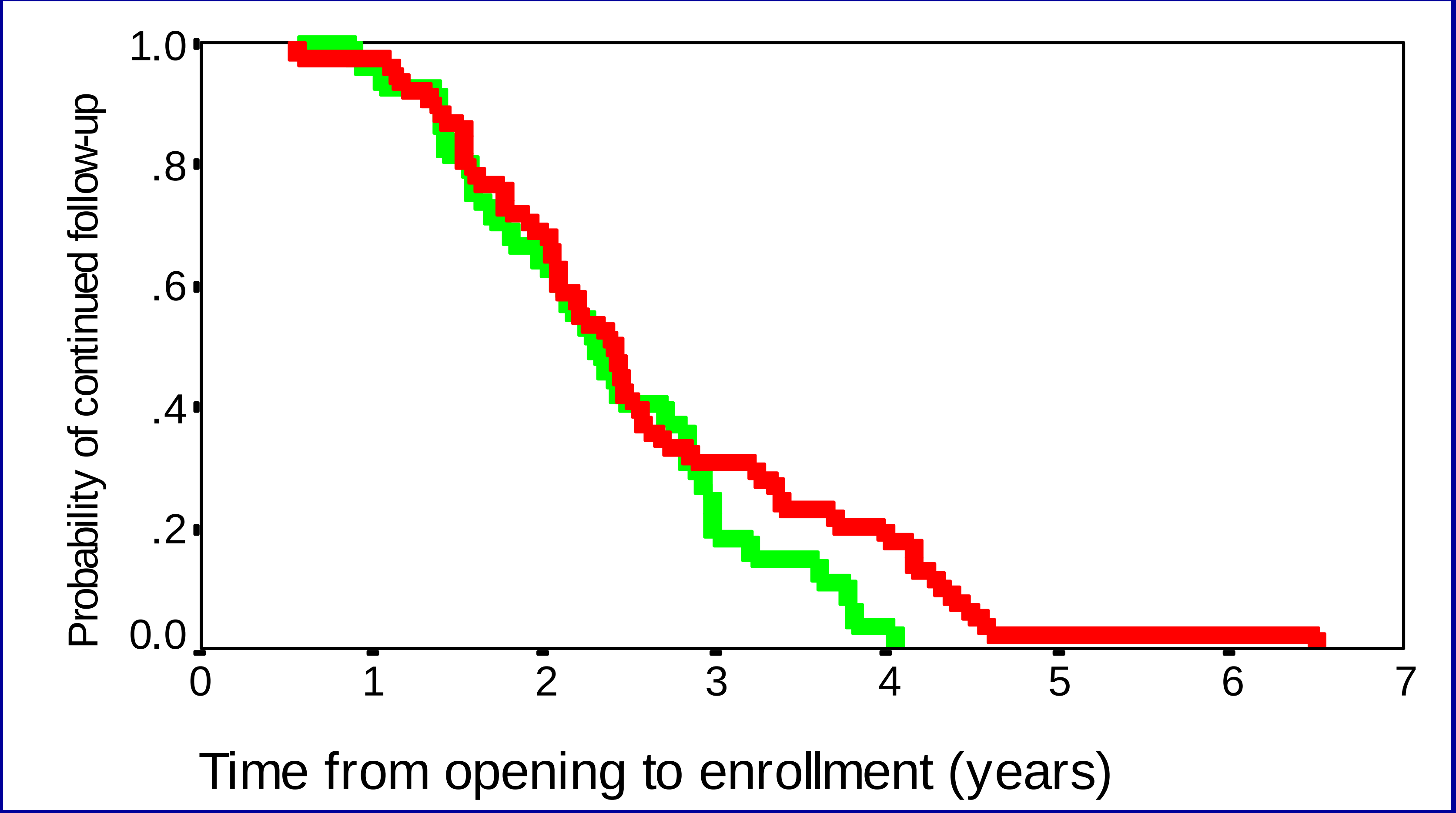
Ioannidis PLoS Clinical Trials 2006 and Clinical Trials (in press)

Time lag: bad news take longer to appear

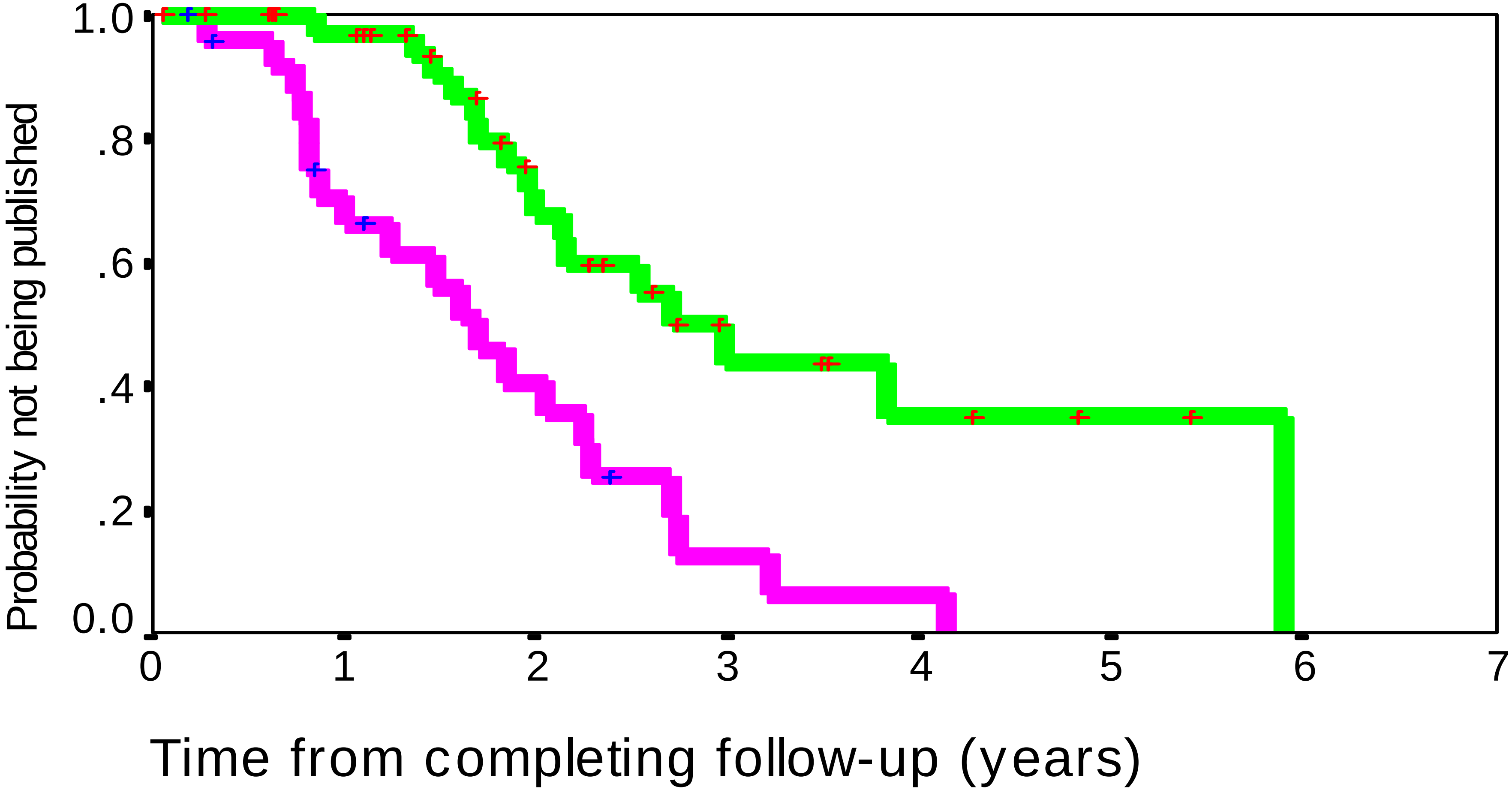


Ioannidis JP. JAMA 1998

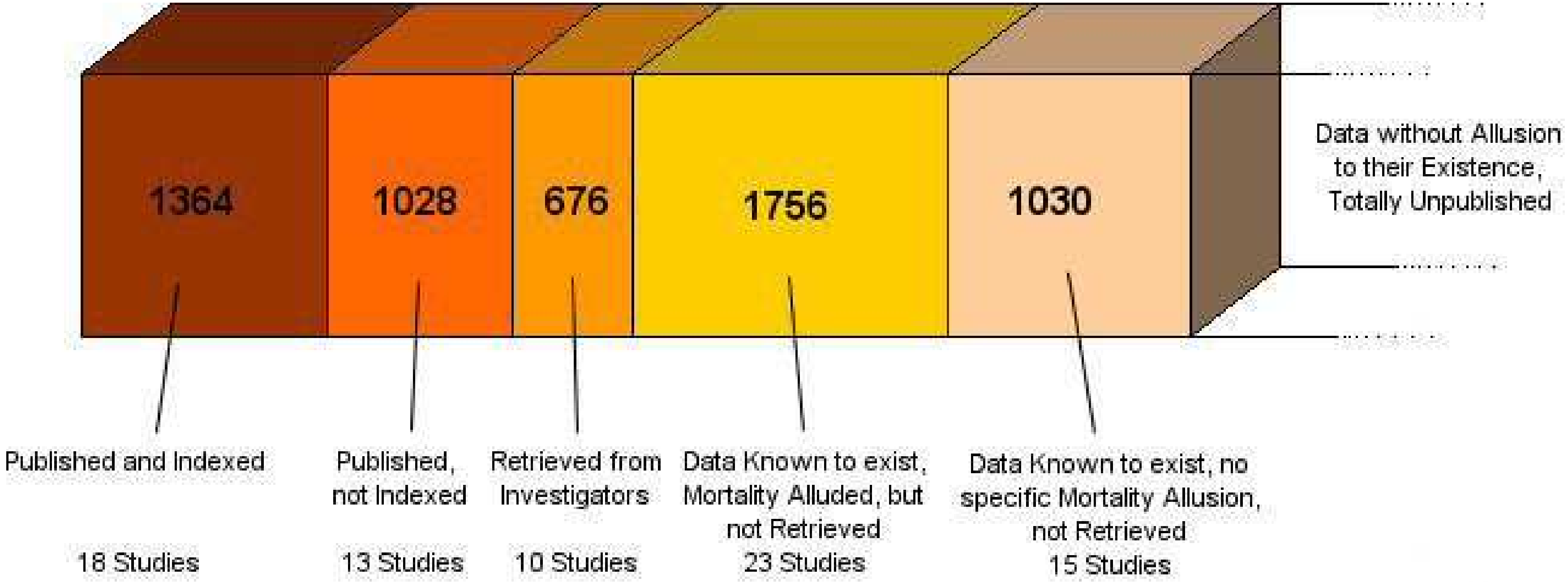
… even though they are obtained as fast..



…but publication is delayed



Prognostic factor meta-analysis: Readily available, available, hidden, and very well hidden data

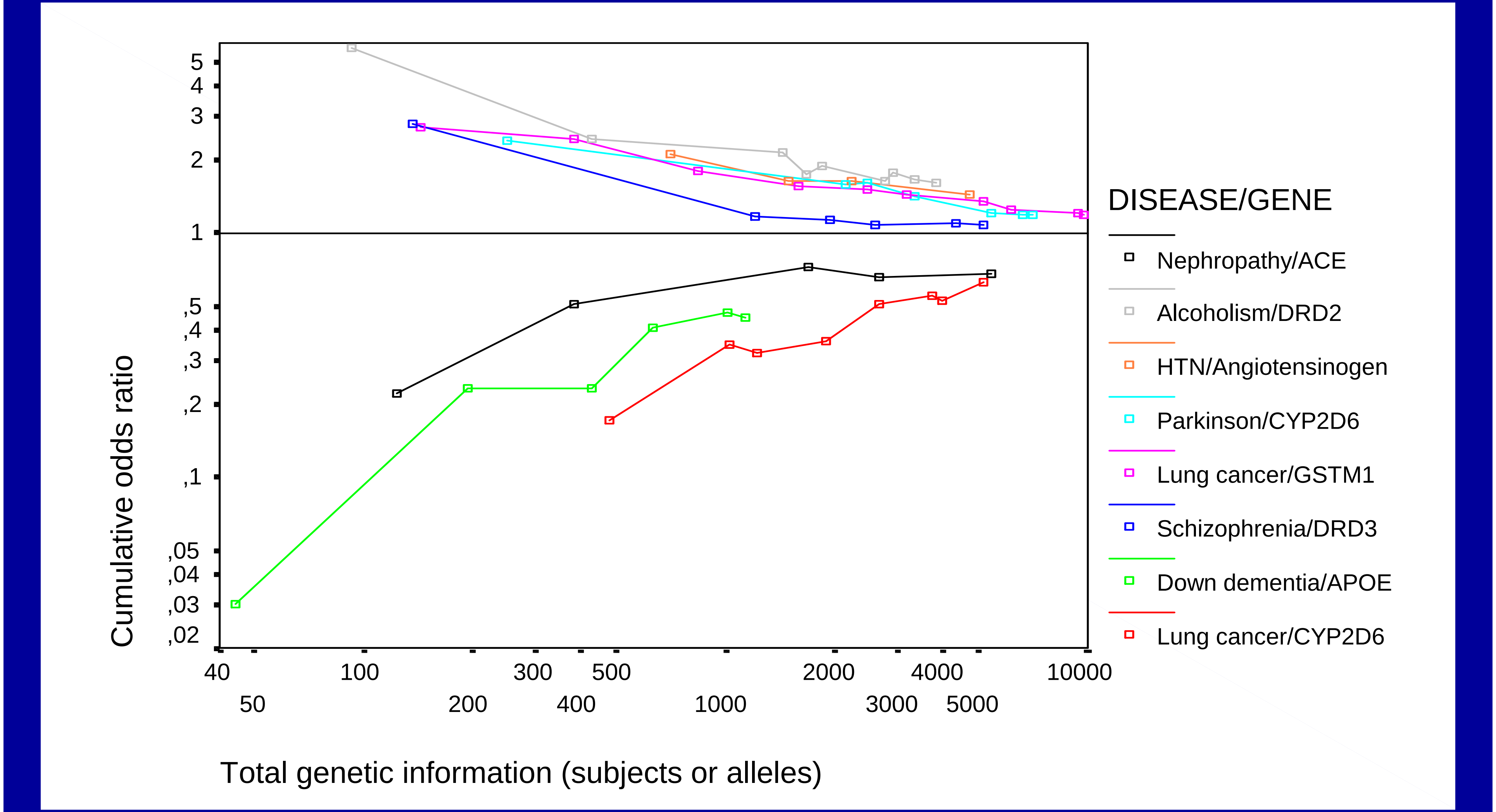


Kyzas, Loizou, Ioannidis. JNCI 2005

Early vs. late evidence

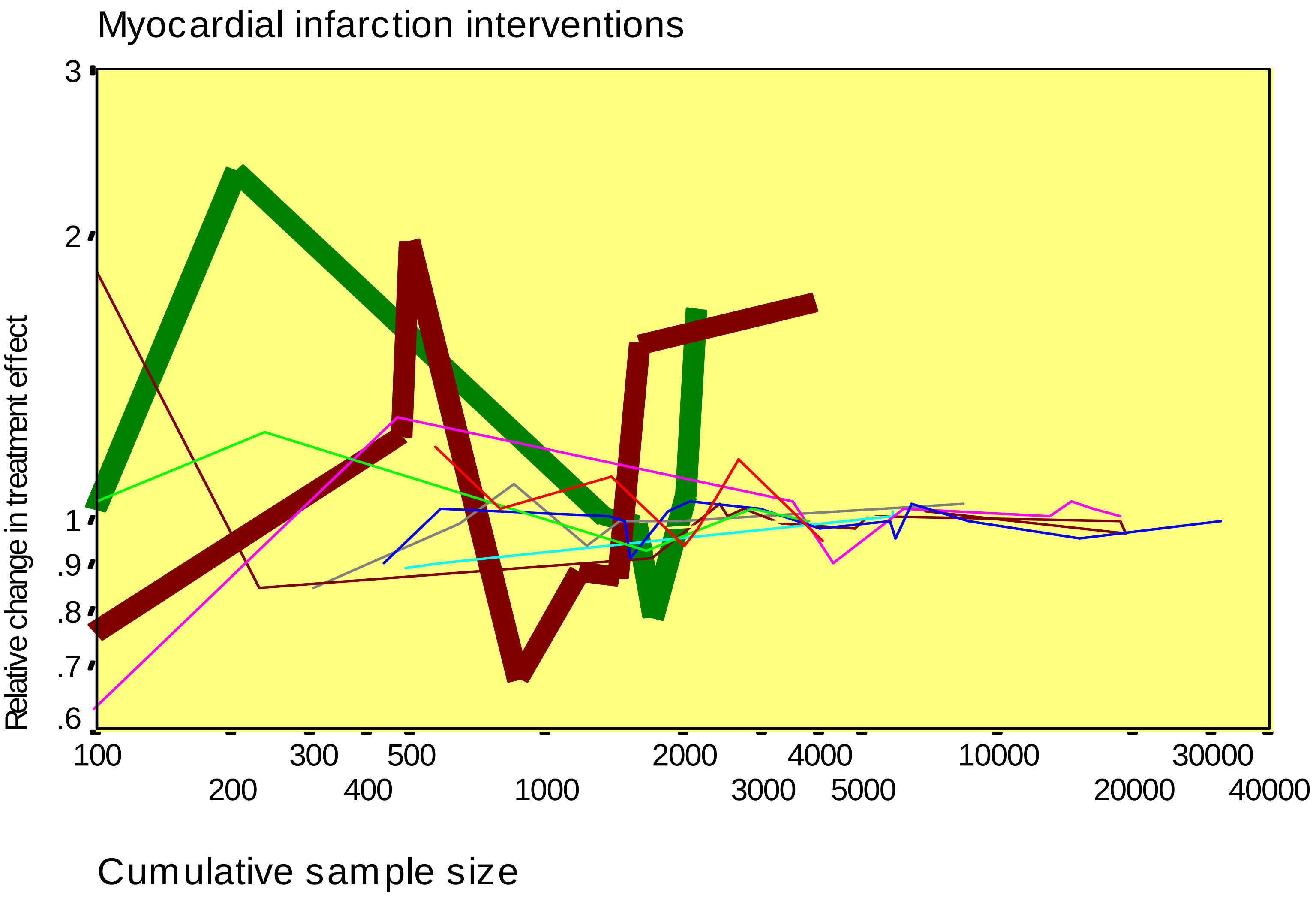
- Evidence evolves over time, it is never constant
- Evolution may change effect sizes
- Opposing effects may occasionally succeed each other in rapid sequence

Non-replicated diminishing effects



Ioannidis et al, Nature Genetics 2001

Discrepancies over time occur even in randomized trials



Ioannidis and Lau, PNAS 2001

Diminishing effects are common in clinical medicine

- Across 100 meta-analyses of mental health related interventions, when it comes to pharmacotherapies, it was far more likely for effect sizes to diminish rather than increase with the appearance of newer trials

Trikalinos et al. J Clin Epidemiol 2004

Large vs. small studies

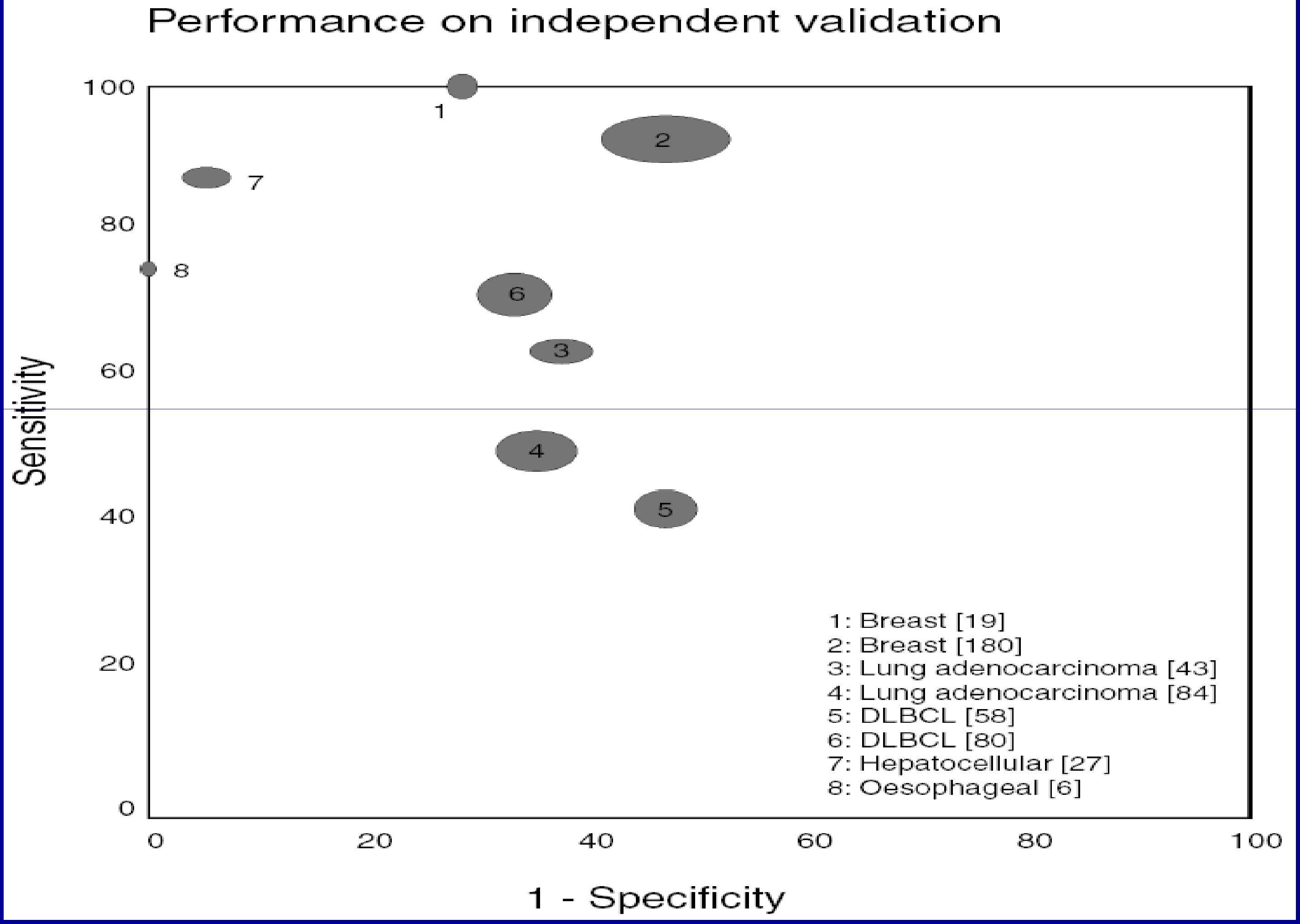
- Theoretically they should not get different results
- Differences reflect both within study issues and field issues

Large vs. small studies in RCTs

- Empirical evidence shows that usually their results agree, but discrepancies may occur beyond change in 10-30% of the cases
- In these situations, large studies tend to give more conservative results, but this is not always the case
- Discrepancies tend to be more frequent for secondary than for primary endpoints

Ioannidis, Cappelleri and Lau, JAMA 1998

An example of small, over-optimistic studies: microarrays



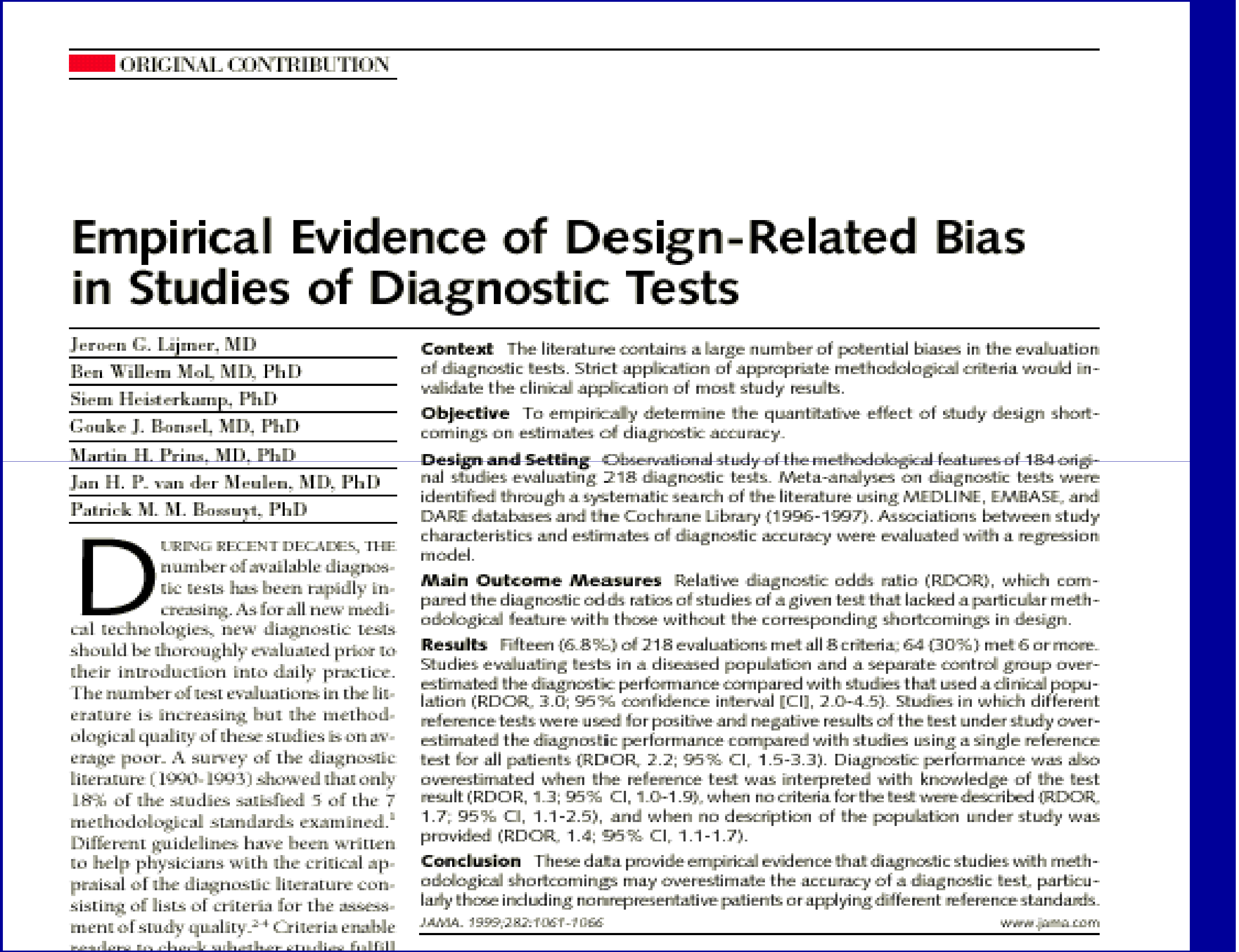
Ntzani and Ioannidis Lancet 2003

“Quality” of studies

- Early empirical evaluations suggested that effect sizes may depend on aggregate quality scores; this has been dismissed, since there are so many quality scores, that inferences are widely different
- Other empirical evaluations suggested that specific quality items such as lack of blinding and lack of allocation concealment in RCTs may inflate treatment effects (e.g. Shultz et al. JAMA 1995)
- Now it seems more likely that such quality deficits may be associated either with inflated or with deflated treatment effects (e.g. Balk et al. JAMA 2002)

The two kinds of bad quality

- Quality is bad on (evil) purpose = the effect sizes are almost always inflated
- Quality is bad because of stupidity = the effect sizes may be anything; usually, but not always, they are deflated



Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests

Jeroen G. Lijmer, MD
Ben Willem Mol, MD, PhD
Siem Heisterkamp, PhD
Gouke J. Bonsel, MD, PhD
Martin H. Prins, MD, PhD
Jan H. P. van der Meulen, MD, PhD
Patrick M. M. Bossuyt, PhD

Context The literature contains a large number of potential biases in the evaluation of diagnostic tests. Strict application of appropriate methodological criteria would invalidate the clinical application of most study results.

Objective To empirically determine the quantitative effect of study design shortcomings on estimates of diagnostic accuracy.

Design and Setting Observational study of the methodological features of 184 original studies evaluating 218 diagnostic tests. Meta-analyses on diagnostic tests were identified through a systematic search of the literature using MEDLINE, EMBASE, and DARE databases and the Cochrane Library (1996-1997). Associations between study characteristics and estimates of diagnostic accuracy were evaluated with a regression model.

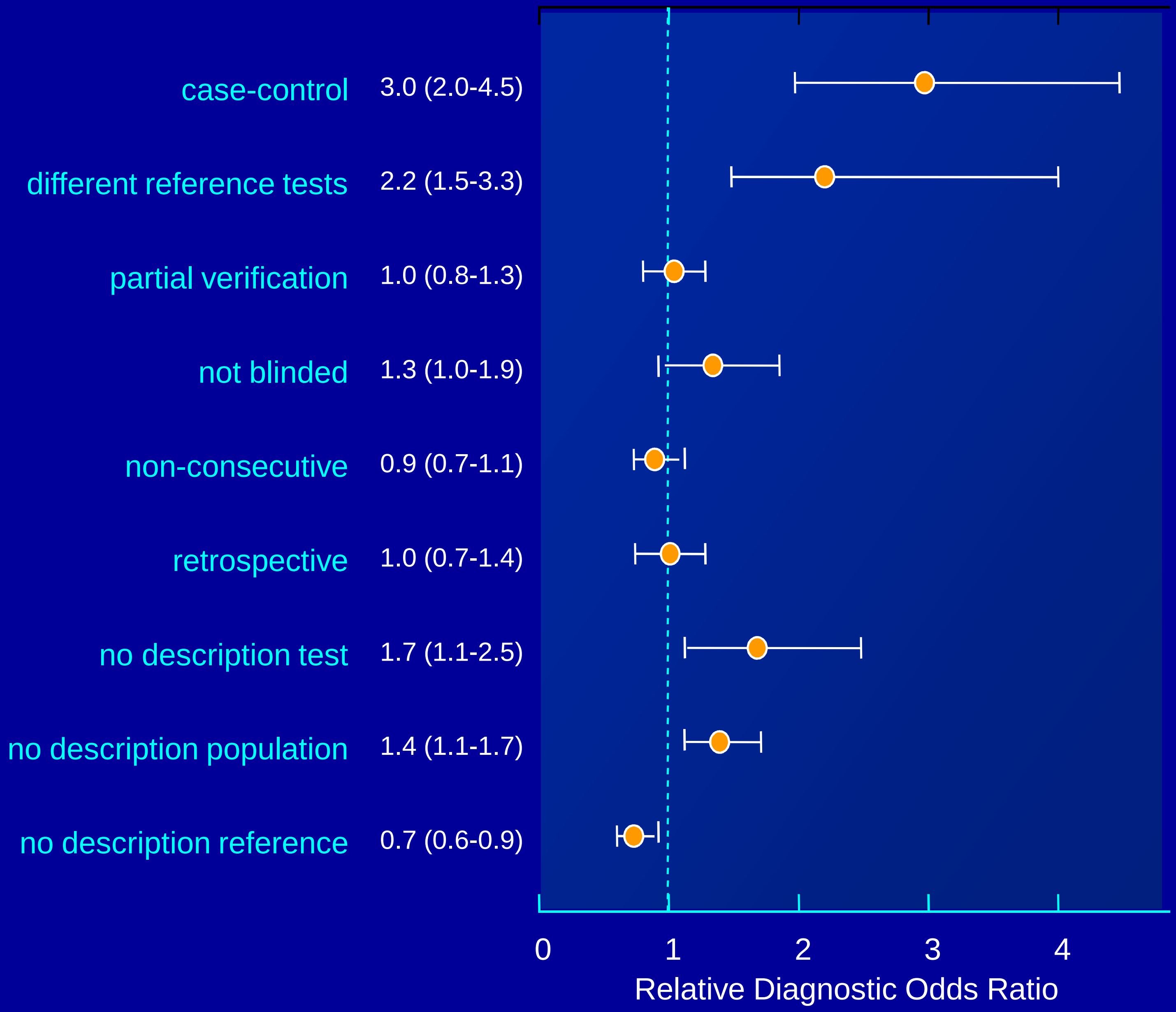
Main Outcome Measures Relative diagnostic odds ratio (RDOR), which compared the diagnostic odds ratios of studies of a given test that lacked a particular methodological feature with those without the corresponding shortcomings in design.

Results Fifteen (6.8%) of 218 evaluations met all 8 criteria; 64 (30%) met 6 or more. Studies evaluating tests in a diseased population and a separate control group overestimated the diagnostic performance compared with studies that used a clinical population (RDOR, 3.0; 95% confidence interval [CI], 2.0-4.5). Studies in which different reference tests were used for positive and negative results of the test under study overestimated the diagnostic performance compared with studies using a single reference test for all patients (RDOR, 2.2; 95% CI, 1.5-3.3). Diagnostic performance was also overestimated when the reference test was interpreted with knowledge of the test result (RDOR, 1.3; 95% CI, 1.0-1.9), when no criteria for the test were described (RDOR, 1.7; 95% CI, 1.1-2.5), and when no description of the population under study was provided (RDOR, 1.4; 95% CI, 1.1-1.7).

Conclusion These data provide empirical evidence that diagnostic studies with methodological shortcomings may overestimate the accuracy of a diagnostic test, particularly those including nonrepresentative patients or applying different reference standards.

JAMA. 1999;282:1061-1066 www.jama.com

DURING RECENT DECADES, THE number of available diagnostic tests has been rapidly increasing. As for all new medical technologies, new diagnostic tests should be thoroughly evaluated prior to their introduction into daily practice. The number of test evaluations in the literature is increasing but the methodological quality of these studies is on average poor. A survey of the diagnostic literature (1990-1993) showed that only 18% of the studies satisfied 5 of the 7 methodological standards examined.[1] Different guidelines have been written to help physicians with the critical appraisal of the diagnostic literature consisting of lists of criteria for the assessment of study quality.[2-4] Criteria enable readers to check whether studies fulfill


case-control 3.0 (2.0-4.5)
different reference tests 2.2 (1.5-3.3)
partial verification 1.0 (0.8-1.3)
not blinded 1.3 (1.0-1.9)
non-consecutive 0.9 (0.7-1.1)
retrospective 1.0 (0.7-1.4)
no description test 1.7 (1.1-2.5)
no description population 1.4 (1.1-1.7)
no description reference 0.7 (0.6-0.9)
0
1
2
3
4
Relative Diagnostic Odds Ratio

Heterogeneity and subgroups

- Heterogeneity is very interesting: it may hint to both genuine diversity and bias
- Too much heterogeneity is suspect
- Too little heterogeneity may also be suspect
- Some heterogeneity is almost ubiquitous
- Over-interpretation through postulated subgroup differences can be dangerous

What can I believe after all? Overall credibility…

- Depends on the pre-evidence odds
- Depends on the evidence
- Depends on bias
- Depends on the field
- All of these may depend on each other

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

Context Controversy and uncertainty ensue when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most prominent when high-impact research is involved.

Objectives To understand how frequently highly cited studies are contradicted or find effects that are stronger than in other similar studies and to discern whether specific characteristics are associated with such refutation over time.

Design All original clinical research studies published in 3 major general clinical journals or high-impact-factor specialty journals in 1990-2003 and cited more than 1000 times in the literature were examined.

Main Outcome Measure The results of highly cited articles were compared against subsequent studies of comparable or larger sample size and similar or better controlled designs. The same analysis was also performed comparatively for matched studies that were not so highly cited.

Results Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these, 7 (16%) were contradicted by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged. Five of 6 highly-cited nonrandomized studies had been contradicted or had found stronger effects vs 9 of 39 randomized controlled trials ($P=.008$). Among randomized trials, studies with contradicted or stronger effects were smaller ($P=.009$) than replicated or unchallenged studies although there was no statistically significant difference in their early or overall citation impact. Matched control studies did not have a significantly different share of refuted results than highly cited studies, but they included more studies with "negative" results.

Conclusions Contradiction and initially stronger effects are not unusual in highly cited research of clinical interventions and their outcomes. The extent to which high citations may provoke contradictions and vice versa needs more study. Controversies are most common with highly cited nonrandomized studies, but even the most highly cited randomized trials may be challenged and refuted over time, especially small ones.

JAMA. 2005;294:218-228 www.jama.com

CLINICAL RESEARCH ON IMPORTANT questions about the efficacy of medical interventions is sometimes followed by subsequent studies that either reach opposite conclusions or suggest that the original claims were too strong. Such disagreements may upset clinical practice and acquire publicity in both scientific circles and in the lay press. Several empirical investigations have tried to address whether specific types of studies are more likely to be contradicted and to explain observed controversies. For example, evidence exists that small studies may sometimes be refuted by larger ones.[1,2]

Similarly, there is some evidence on disagreements between epidemiological studies and randomized trials.[3-5] Prior investigations have focused on a variety of studies without any particular attention to their relative importance and scientific impact. Yet, most research publications have little impact while a small minority receives most attention and dominates scien-

Contradiction in highly-cited clinical research on interventions

- Analyzed 115 articles published in 1990-2003 in the 3 major general medical journals (NEJM, JAMA, Lancet) and the top specialty journals that had received over 1000 citations each by august 2004
- Of those, 49 pertained to original assessments of interventions for therapy or prevention and 45 claimed effectiveness.
- Five of the 6 efficacy findings based on non-randomized trials were already contradicted or found to be exaggerated by 2004
- Even among highly-cited randomized trials, efficacy findings were already contradicted or found to be exaggerated in 9 of 39 interventions

Ioannidis JP. JAMA 2005; July 13

Highly-cited contradicted findings

- Vitamin E and cardiovascular mortality (two large prospective cohorts and one trial of 2,002 subjects claimed large decreases in mortality)
- Hormone replacement therapy and coronary artery disease (major benefits claimed by the Nurses' Health Study and the PEPI trial [on surrogates])
- HA-1A antibody to endotoxin initially found to halve mortality in patients with sepsis
- Nitric oxide found initially to markedly improve outcomes in respiratory distress syndrome

Science at various pre-study odds of true findings

Ioannidis. Why most published research findings are false? PLoS Medicine (in press)

Positive predictive value (PPV) of research findings for various combinations of power (1-β), ratio of true to no relationships (R) and bias (u)

1-β	R	u	Practical example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	.85
0.95	2:1	0.30	Confirmatory meta-analysis of good quality RCTs	.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	.41
0.20	1:5	0.20	Underpowered, phase I/II well-performed RCT	.23
0.20	1:5	0.80	Underpowered, phase I/II poorly performed RCT	.17
0.80	1:10	0.30	Adequately powered, exploratory epidemiological study	.20
0.20	1:10	0.30	Underpowered, exploratory epidemiological study	.12
0.20	1:1000	0.80	Discovery-oriented exploratory research with massive testing	.0010
0.20	1:1000	0.20	As above, but with more limited bias (more standardized)	.0015

Effect size = bias

- In several scientific disciplines, the effect sizes observed in different studies are, on average, accurate estimates of the extent of net bias operating in the field
- Thus, disciplines that find larger effect sizes (=are scientifically considered more successful) are simply more biased than others that find smaller effect sizes
- In the same scientific discipline, the most successful and appreciated studies are simply the ones that suffer the worst net bias

Post-study odds of a true finding are small

- When effect sizes are small
- When studies are small
- When field are “hot” (many teams work on them)
- When there is strong interest in the results
- When databases are large
- When analyses are more flexible

Ioannidis JP. PLoS Medicine 2005

Conclusions

- The whole gives more information than the parts
- EBM focuses both on the constituent studies and in the composite picture
- Bias and heterogeneity are almost ubiquitous
- EBM may offer the best opportunity to understand how heterogeneity and bias work and manifest themselves