Πρόγραμμα Μεταπτυχιακών Σπουδών:
«Πληροφορική και Υπολογιστική Βιοϊατρική»

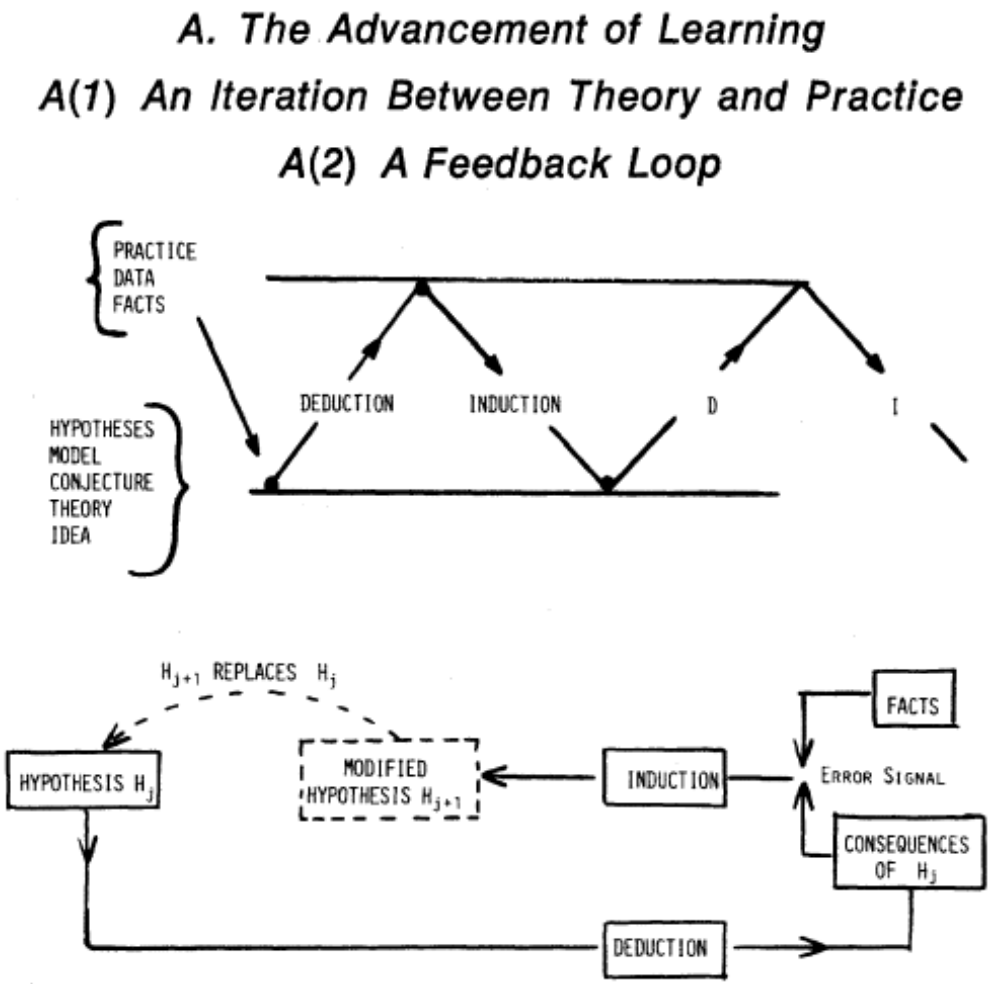
ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Μάθημα: Μεθοδολογία της έρευνας

2020

Παντελής Μπάγκος
Καθηγητής

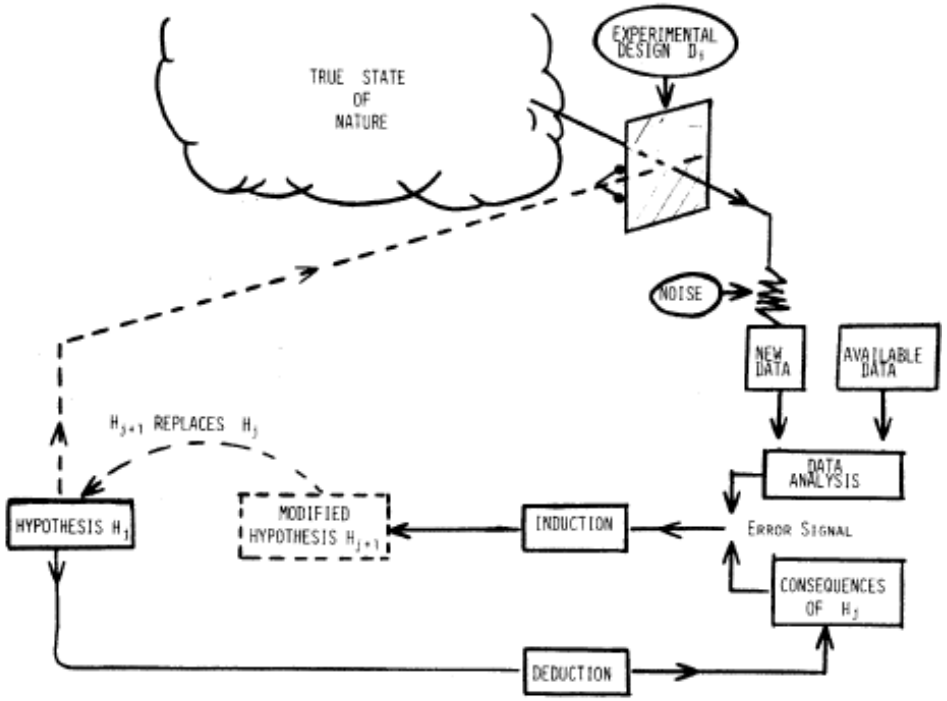
Επιστήμη και Στατιστική (1)



George E. P. Box **Science and Statistics**. Journal of the American Statistical Association, Vol. 71, No. 356 (Dec., 1976), pp. 791-799

Επιστήμη και Στατιστική (2)

B. Data Analysis and Data Getting in the Process of Scientific Investigation[a]



[a] The experimental design is here shown as a movable window looking onto th true state of nature. Its positioning at each stage is motivated by current beliefs hopes, and fears.

George E. P. Box **Science and Statistics**. Journal of the American Statistical Association, Vol. 71, No. 356 (Dec., 1976), pp. 791-799

Frequentist or Bayesian???

Τι είναι η στατιστική συμπερασματολογία;

Στατιστική συμπερασματολογία είναι η επιστήμη η οποία έχει στόχο να εξάγει συμπεράσματα για έναν πληθυσμό μελετώντας ένα δείγμα που προέρχεται από τον πληθυσμό αυτό.

Η στατιστική συμπερασματολογία οδηγεί σε συμπεράσματα για την παράμετρο θ του πληθυσμού μέσω της παρατήρησης της μεταβλητής X, και τα βασικά συμπεράσματα βασίζονται στο ότι οι τιμές του θ που δίνουν **μεγάλη πιθανότητα** στην τιμή του x που παρατηρήθηκε, είναι πιο πιθανές απ'ότι εκείνες που δίνουν στο x μικρή πιθανότητα (**αρχή της μέγιστης πιθανοφάνειας**).

Πέτρος Δελλαπόρτας – Παναγιώτης Τσιαμυρτζής, Σημειώσεις μαθήματος "**Στατιστική κατά Bayes**", Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Στατιστικής (2004)

Κλασική vs Μπεϋζιανή Στατιστική

Κλασική Στατιστική	Μπεϋσιανή Στατιστική
Η παράμετρος θ (η οποία δεν είναι γνωστή), χρησιμοποιείται περισσότερο σαν να είναι μία σταθερά και όχι σαν τυχαία μεταβλητή.	Η παράμετρος θ (η οποία δεν είναι γνωστή), είναι τυχαία μεταβλητή.
Τα δεδομένα που παρατηρήθηκαν είναι τυχαία.	Τα δεδομένα που παρατηρήθηκαν είναι σταθερά.
Ο υπολογισμός των διαστημάτων εμπιστοσύνης και γενικότερα η συμπερασματολογία δεν είναι πιθανοθεωρητική.	Η συμπερασματολογία για την παράμετρο θ βασίζεται σε μια κατανομή πιθανότητας.

Πέτρος Δελλαπόρτας – Παναγιώτης Τσιαμυρτζής, Σημειώσεις μαθήματος "**Στατιστική κατά Bayes**", Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Στατιστικής (2004)

Ερμηνεία του 95% Δ.Ε. στην κλασική στατιστική

Έστω το 95% διάστημα εμπιστοσύνης του [0.08, 0.12], εννοούμε πως υπάρχει 95% πιθανότητα η παράμετρος θ να βρίσκεται μεταξύ 0.08 και 0.12.

Όμως σε αυτήν την περίπτωση υπάρχει πρόβλημα ερμηνείας, διότι το θ δεν είναι τυχαίο: είτε θα ανήκει στο διάστημα, είτε δεν θα ανήκει σε αυτό, και άρα η πιθανότητα δεν μπορεί και δεν πρέπει να υπεισέρχεται σαν παράγοντας στην ερμηνεία του.

Το μόνο τυχαίο στοιχείο στο μοντέλο πιθανότητας είναι τα δεδομένα, οπότε η σωστή ερμηνεία του διαστήματος είναι πως αν επαναλάβουμε την διαδικασία πολλές φορές, τότε τα διαστήματα που θα κατασκευάσουμε θα περιλαμβάνουν την παράμετρο θ στο 95% των περιπτώσεων.

Χαρακτηριστικά της Μπεϋζιανής προσέγγισης (1)

- **A-priori Πληροφορία (Prior Information)**: Κάθε πρόβλημα είναι μοναδικό και έχει το δικό του περιεχόμενο. Από αυτό ακριβώς το περιεχόμενο εξάγονται a-priori πληροφορίες και είναι η διατύπωση και η εκμετάλλευση της προηγούμενης γνώσης που διαχωρίζουν την Μπεϋζιανή θεωρία από αυτήν της κλασικής στατιστικής.

- **Υποκειμενική Πιθανότητα (Subjective Probability):** Η κλασική στατιστική εξαρτάται από μία μακροχρόνια συχνότητα καθορισμού των πιθανοτήτων. Αν και αυτό είναι επιθυμητό, οδηγεί σε «δυσκίνητα» συμπεράσματα. Αντίθετα, η Μπεϋζιανή στατιστική θέτει με σαφήνεια την ιδέα ότι όλες οι πιθανότητες είναι υποκειμενικές και εξαρτώνται από τις πεποιθήσεις του κάθε ατόμου και τις γνώσεις που μπορεί να έχει καθένας από μας για μια δεδομένη «κατάσταση». Η συμπερασματολογία της βασίζεται στην **a-posteriori κατανομή (posterior distribution)** $f(\theta|x)$, η μορφή της οποίας εξαρτάται (μέσω του θεωρήματος του Bayes) από τον τρόπο καθορισμού της a-priori κατανομής $f(\theta)$.

Πέτρος Δελλαπόρτας – Παναγιώτης Τσιαμυρτζής, Σημειώσεις μαθήματος "**Στατιστική κατά Bayes**", Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Στατιστικής (2004)

Χαρακτηριστικά της Μπεϋζιανής προσέγγισης (2)

- **Συνέπεια (Self-Consistency)**: Χρησιμοποιώντας την παράμετρο θ σαν τυχαία, όλη η ανάπτυξη της Μπευζιανής συμπερασματολογίας πηγάζει και εξαρτάται μόνο από την θεωρία πιθανοτήτων. Αυτό έχει πολλά πλεονεκτήματα και σημαίνει πως όλα τα συμπεράσματα μπορούν να παρουσιαστούν με την μορφή πιθανοτήτων για την παράμετρο θ, πράγματι προκύπτουν άμεσα από την a-posteriori κατανομή.

- **Μη προσκόλληση σε «συνταγές»**: Επειδή η κλασική στατιστική δεν είναι σε θέση να χρησιμοποιήσει όρους πιθανοτήτων για την παράμετρο θ, έχουν αναπτυχθεί αρκετά κριτήρια με σκοπό να καθορίσουν πότε ένας συγκεκριμένος εκτιμητής θα μπορούσε να χαρακτηριστεί ως «καλός».

Το Θεώρημα του Bayes (1)

Στην βασική του μορφή το θεώρημα του Bayes είναι απλό και αφορά υπό συνθήκη πιθανότητες. Αν Α και Β είναι δύο ενδεχόμενα με P(A)>0, τότε:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

Η χρησιμότητα του θεωρήματος του Bayes σε εφαρμογές πιθανοτήτων είναι ότι παρέχει την δυνατότητα αντιστροφής της «θέσης» των ενδεχομένων. Έτσι, γίνεται εμφανές πώς η πιθανότητα του B|A σχετίζεται με την πιθανότητα του A|B. Μια μικρή προέκταση του θεωρήματος του Bayes μπορεί να γίνει, αν θεωρήσουμε τα ενδεχόμενα $C_1,\ldots,C_k$, τα οποία διαμερίζουν ένα δειγματικό χώρο Ω, έτσι ώστε τα $C_i \cap C_j = \emptyset$ για κάθε $i \neq j$ και $C_1 \cup \ldots \cup C_k = \Omega$. Σε αυτήν την περίπτωση θα έχουμε:

$$P(C_i \mid A) = \frac{P(A \mid C_i)P(C_i)}{\sum_{j=1}^{k} P(A \mid C_J)P(C_j)} \quad ,i = 1,\ldots,k.$$

Πέτρος Δελλαπόρτας – Παναγιώτης Τσιαμυρτζής, Σημειώσεις μαθήματος "**Στατιστική κατά Bayes**", Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Στατιστικής (2004)

Το Θεώρημα του Bayes (2)

Το θεώρημα του Bayes σε όρους τυχαίων μεταβλητών με πυκνότητες που συμβολίζονται γενικά με *f* , παίρνει την εξής μορφή:

$$f(\theta \mid x) = \frac{f(\theta) f(x \mid \theta)}{\int f(\theta) f(x \mid \theta) d\theta}$$

Θα πρέπει να προσέξουμε ιδιαίτερα το γεγονός ότι από την στιγμή που ολοκληρώνουμε ως προς *θ*, ο παρανομαστής στο θεώρημα του Bayes είναι συνάρτηση μόνο ως προς *x*. Συνεπώς, για δεδομένες παρατηρήσεις *x*, ο παρανομαστής είναι σταθερά και ονομάζεται **σταθερά κανονικοποίησης**. Με βάσει αυτά ένας εναλλακτικός τρόπος παρουσίασης του θεωρήματος του Bayes είναι ο εξής:

$$f(\theta|x) \propto f(\theta) f(x|\theta)$$

ή αλλιώς θα λέγαμε ότι **η a-posteriori κατανομή (posterior distribution) είναι ανάλογη της a-priori κατανομής (prior distribution) πολλαπλασιαζόμενης με την συνάρτηση πιθανοφάνειας (likelihood function)**.

Πέτρος Δελλαπόρτας – Παναγιώτης Τσιαμυρτζής, Σημειώσεις μαθήματος "**Στατιστική κατά Bayes**", Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Στατιστικής (2004)

Παράδειγμα εφαρμογής του θεωρήματος του Bayes

"A patient goes to see a doctor. The doctor performs a test with 99 percent reliability--that is, 99 percent of people who are sick test positive and 99 percent of the healthy people test negative. The doctor knows that only 1 percent of the people in the country are sick. Now the question is: if the patient tests positive, what are the chances the patient is sick?

The intuitive answer is 99 percent, but the correct answer is 50 percent....«

The solution to this question can easily be calculated using Bayes's theorem. Bayes, who was a reverend who lived from 1702 to 1761 stated that the probability you test positive AND are sick is the product of the likelihood that you test positive GIVEN that you are sick and the "prior" probability that you are sick (the prevalence in the population). Bayes's theorem allows one to compute a conditional probability based on the available information.

Bayes's Theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

	Diseased	Not Diseased	
Test +	99	99	198
Test -	1	9,801	9,802
	100	9,900	10,000

Παράδειγμα εφαρμογής του θεωρήματος του Bayes

What we want to know is P (A | B), i.e., the probability of disease (A), given that the patient has a positive test (B).

We know that prevalence of disease (the unconditional probability of disease) is 1% or 0.01; this is represented by P(A). Therefore, in a population of 10,000 there will be 100 diseased people and 9,900 non-diseased people.

We also know the sensitivity of the test is 99%, i.e., P(B | A) = 0.99; therefore, among the 100 diseased people, 99 will test positive.

We also know that the specificity is also 99%, or that there is a 1% error rate in non-diseased people. Therefore, among the 9,900 non-diseased people, 99 will have a positive test.

And from these numbers, it follows that the unconditional probability of a positive test is 198/10,000 = 0.0198; this is P(B).

Thus, P(A | B) = (0.99 x 0.01) / 0.0198 = 0.50 = 50%.

From the table above, we can also see that given a positive test (subjects in the Test + row), the probability of disease is 99/198 = 0.05 = 50%

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\text{not } A)P(\text{not } A)}$$

$$P(A|B) = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.001 \times 0.999} \approx 0.5,$$

Αντίθετες απόψεις για την Μπεϋζιανή Θεωρία

✗ Τα συμπεράσματα εξαρτώνται από την επιλογή της a-priori κατανομής.

✓ Η κλασική συμπερασματολογία χρησιμοποιεί επίσης κάποιες προηγούμενες γνώσεις όπως η κατασκευή του κατάλληλου μοντέλου πιθανοφάνειας.

✗ Στην κλασική στατιστική, οι εκτιμητές μέγιστης πιθανοφάνειας προκύπτουν από την επιλογή εκείνης της τιμής που μεγιστοποιεί την πιθανοφάνεια. Αντίθετα, όσον αφορά την Μπεϋζιανή συμπερασματολογία, χρησιμοποιεί έναν μέσο όρο της πιθανοφάνειας. Η αμφισβήτηση για την όλη συμπερασματολογία του Bayes, προκαλείται εξαιτίας του ότι ο μέσος όρος αυτός σταθμίζεται με βάση την a-priori κατανομή.

✓ Στην κλασική στατιστική, είναι επίσης αρκετά σύνηθες να δίνονται διαφορετικά βάρη σε διαφορετικά κομμάτια πληροφορίας, όπως γίνεται για παράδειγμα στην σταθμισμένη παλινδρόμηση.

Πέτρος Δελλαπόρτας – Παναγιώτης Τσιαμυρτζής, Σημειώσεις μαθήματος "**Στατιστική κατά Bayes**", Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Στατιστικής (2004)

Γιατί η κλασική στατιστική χρησιμοποιείται περισσότερο στην πράξη;

1. Ease of use: Fisher's theory in particular is well set up to yield answers on an easy and almost automatic basis.
2. Model building: Both Fisherian and NPW theory pay more attention to the preinferential aspects of statistics.
3. Division of labor: The NPW school in particular allows interesting parts of a complicated problem to be broken off and solved separately. These partial solutions often make use of aspects of the situation, for example, the sampling plan, which do not seem to help the Bayesian.
4. Objectivity: The high ground of scientific objectivity has been seized by the frequentists.

Neyman-
Pearson-
Wald

Efron B., **Why Isn't Everyone a Bayesian** The American Statistician, 1986. **40**(1).

Απόψεις για τη χρήση του p-value ως μέτρο εκτίμησης της στατιστικής σημαντικότητας

P-value

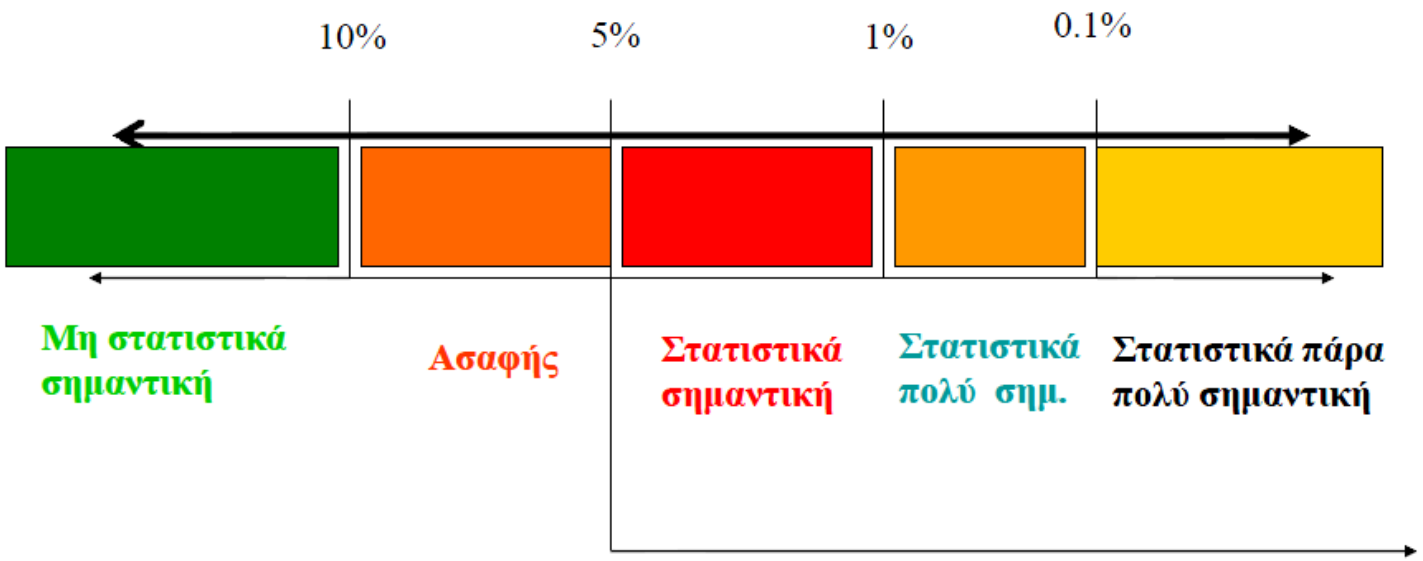
p-value: είναι η πιθανότητα να παρατηρήσουμε ένα αποτέλεσμα τόσο ή περισσότερο ακραίο όσο το αποτέλεσμα ενός συγκεκριμένου δείγματος δεδομένου ότι ισχύει η μηδενική υπόθεση

Μικρό p-value: τα αποτελέσματα από το δείγμα δεν είναι πιθανά δεδομένου της H_0

Χρησιμοποιούμε ως επίπεδο σημαντικότητας a=0,05 (5%) (αυθαίρετο)

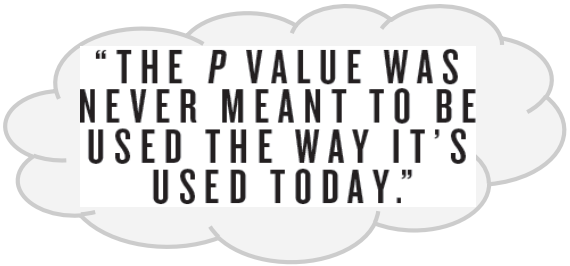
Αν p-value < a, τότε απορρίπτουμε την H_0 και αποδεχόμαστε την H_a.

Αν p-value > a, τότε αποτυγχάνουμε να απορρίψουμε την H_0



Προβλήματα που προκύπτουν από τον ορισμό του p-value (1)

- Το p-value δεν έχει πληροφορίες για την εναλλακτική υπόθεση. Υπολογίζεται μόνο υπό τη μηδενική υπόθεση.
- Μια μεγάλη επίδραση μικρής κλινικής δοκιμής ή μια μικρή επίδραση μεγάλης κλινικής δοκιμής μπορεί να οδηγήσουν σε ίδια p-values.
- Στον υπολογισμό των p-values συμμετέχουν και οι πιο "ακραίες" τιμές οι οποίες όμως δεν έχουν παρατηρηθεί.
- Οι τιμές που θεωρούνται πιο "ακραίες" εξαρτώνται από το πως έχει πραγματοποιηθεί το πείραμα.
- Τα p-values καθορίζονται από υποκειμενικά κριτήρια του ερευνητή. Για παράδειγμα η απόφαση να εφαρμόσουμε μονόπλευρους ή αμφίπλευρους στατιστικούς ελέγχους οδηγεί σε τελείως διαφορετικά p-values.



Goodman, S.N. and R. Royall, **Evidence and scientific research**. Am J Public Health, 1988. **78**(12): p. 1568-74.

Προβλήματα που προκύπτουν από τον ορισμό του p-value (2)

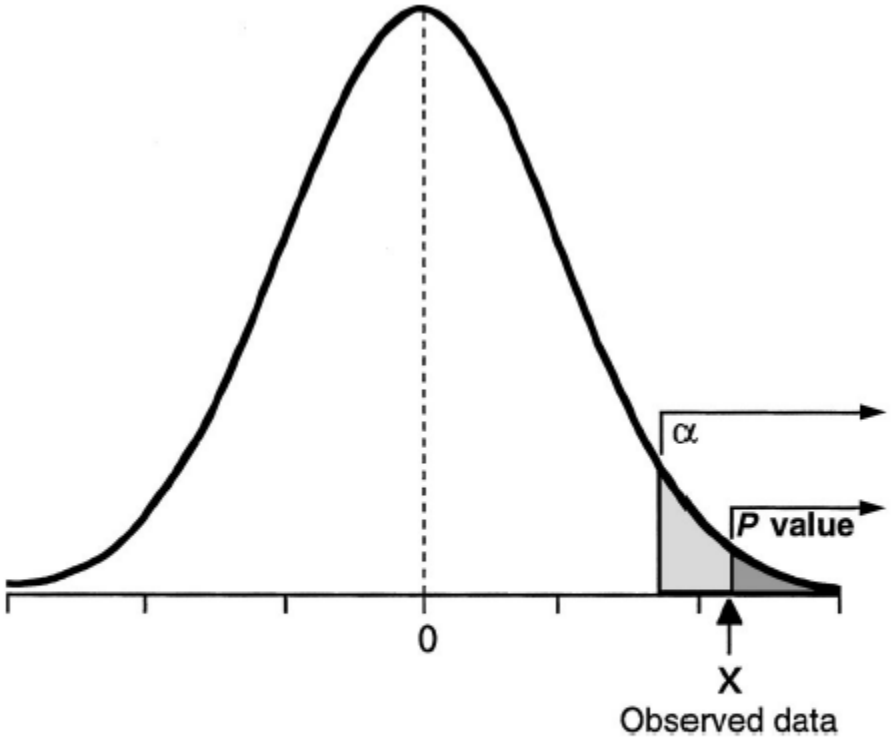


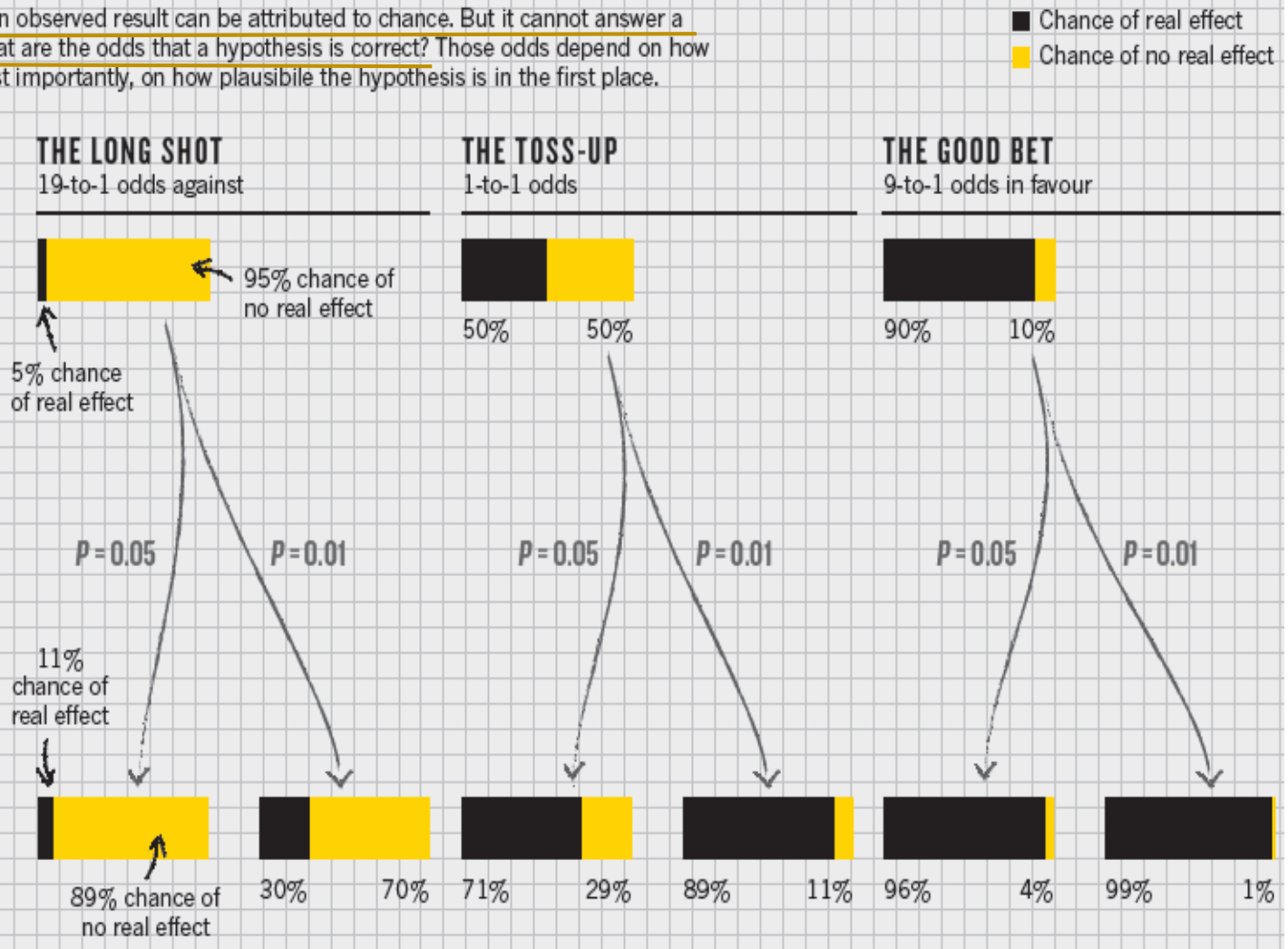
Figure 3. The bell-shaped curve represents the probability of every possible outcome under the null hypothesis. Both α (the type I error rate) and the *P* value are "tail areas" under this curve. The tail area for α is set before the experiment, and a result can fall anywhere within it. The *P* value tail area is known only after a result is observed, and, by definition, the result will always lie on the border of that area.

Goodman, S.N. **Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy.** Ann Intern Med. 1999;130:995-1004

R. NUZZO; SOURCE: T. SELLKE *ET AL. AM. STAT.* **55**, 62–71 (2001)

PROBABLE CAUSE

A *P* value measures whether an observed result can be attributed to chance. But it cannot answer a researcher's real question: what are the odds that a hypothesis is correct? Those odds depend on how strong the result was and, most importantly, on how plausibile the hypothesis is in the first place.



Before the experiment
The plausibility of the hypothesis — the odds of it being true — can be estimated from previous experiments, conjectured mechanisms and other expert knowledge. Three examples are shown here.

The measured *P* value
A value of 0.05 is conventionally deemed 'statistically significant'; a value of 0.01 is considered 'very significant'.

After the experiment
A small *P* value can make a hypothesis more plausible, but the difference may not be dramatic.

Εναλλακτικές προσεγγίσεις-Πηλίκο Πιθανοφάνειας (1)

- Το πηλίκο πιθανοφάνειας δεν επηρεάζεται από τα αίτια τερματισμού μιας κλινικής δοκιμής ούτε από τον αριθμό των πολλαπλών ελέγχων.
- Η ερμηνεία του πηλίκου πιθανοφάνειας δεν εξαρτάται από το σχεδιασμό της μελέτης ή το μέγεθος του δείγματος.

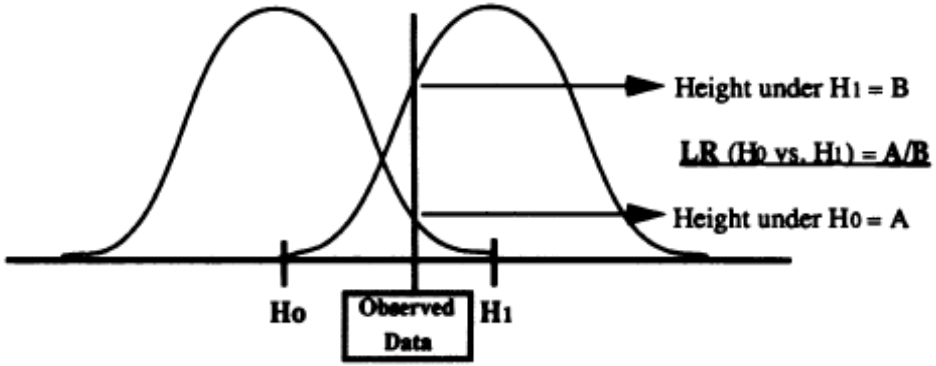


FIGURE 1—A Graphical Representation of the Calculation of the Likelihood Ratio (LR) for Two Simple Hypotheses Given Experimental Data under a Single Statistical Model
Even though these curves represent probability distributions, the likelihoods are defined only at the observed data point, and this has arbitrary scale, hence the y axis has no units. The one-sided p-value corresponding to these data would be the proportion of the area under the H_0 curve to the right of the data point. Different stopping rules can affect the shape of these curves and hence change that area, but the ratio of heights at the data point will remain the same.

Goodman, S.N. and R. Royall, **Evidence and scientific research**. Am J Public Health, 1988. **78**(12): p. 1568-74.

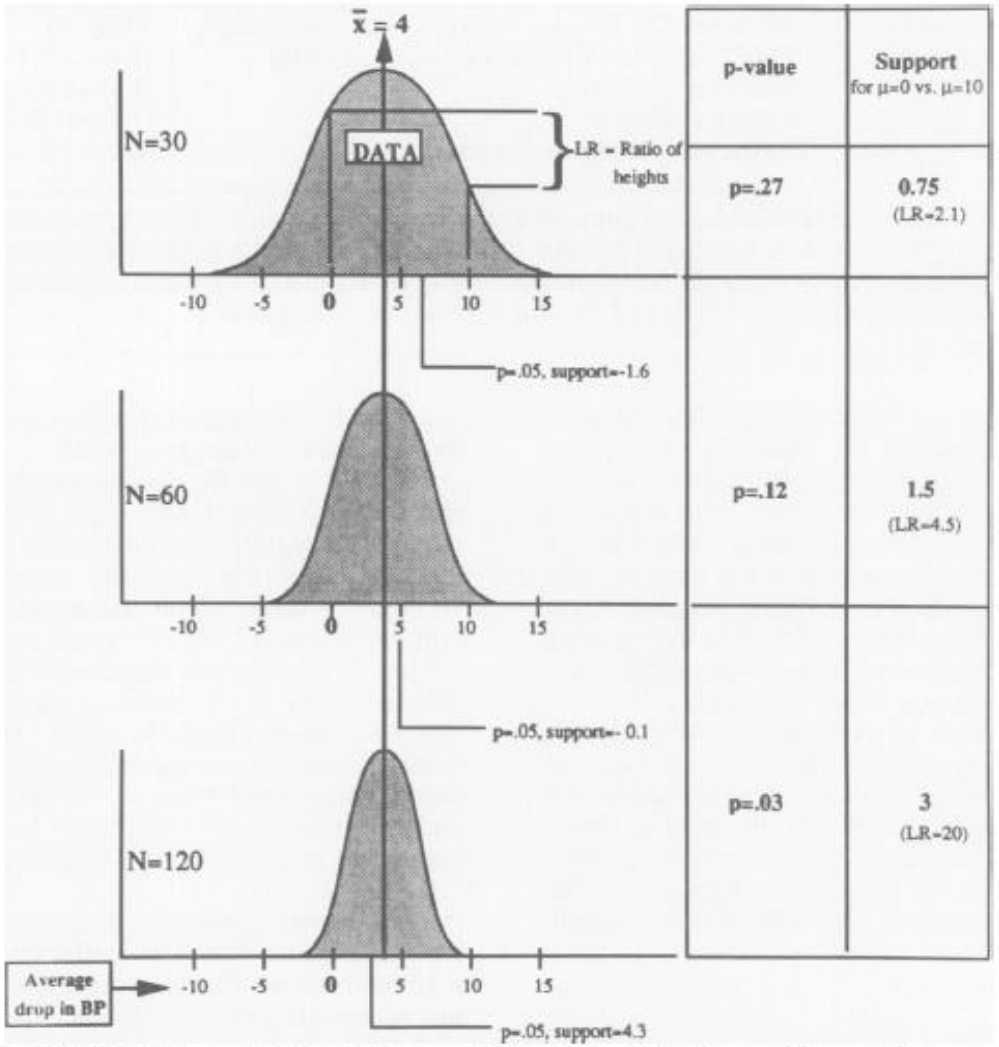


FIGURE 2—P-Value and Likelihood Measures of Evidence Provided by the same Observed Difference in Experiments of Three Different Sizes
The curves are the likelihood functions of the average fall in blood pressure from a drug given an observed four-point average drop. The LR for any pair of hypotheses is the ratio of curve heights at those hypotheses (it is calculated here for $\mu = 0$ vs $\mu = 10$). Positive support is evidence for no effect, a negative one is evidence for a ten point effect. Support corresponding to the $p = .05$ point is also marked on each curve. See text and Appendix for more details.

Goodman, S.N. and R. Royall, **Evidence and scientific research**. Am J Public Health, 1988. **78**(12): p. 1568-74.

Εναλλακτικές προσεγγίσεις-Πηλίκο Πιθανοφάνειας (2)

0 units support	= No evidence for alternative vs null hypothesis	(LR=1)
-1 unit support	= **Weak** evidence for the alternative vs null	(LR=1/2.7=0.37)
-2 units support	= **Moderate** evidence	(LR=1/7.4=0.14)
-3 units support	= **Strong** evidence	(LR=1/20=0.05)
-4 units support	= **Extremely strong** evidence	(LR=1/55=0.02)

Informal guide to interpretation of support of the null over the alternative hypothesis. The range of negative units is displayed because this corresponds to the familiar situation when we are measuring increasing statistical distance away from the null. Positive units of support would indicate evidence for the null vs. the alternative hypothesis.

Goodman, S.N. and R. Royall, **Evidence and scientific research**. Am J Public Health, 1988. **78**(12): p. 1568-74.

Εναλλακτικές προσεγγίσεις-Το θεώρημα του Bayes

Final log odds (Final Beliefs)	=	Initial log odds (Initial Beliefs)	+	Log Likelihood Ratio (Weight of Evidence, Support)

Bayes Theorem

Prior odds of hypothesis being true (before seeing data)	**Subjective component**
×	
Bayes Factor	**Data component** *("evidence")*
⇓	
Final ("posterior") odds of null hypothesis being true	**Probability of truth**

Figure 2. Bayes theorem, in words.

Έχει υποστεί κρητική καθώς απαιτείται να οριστεί μια prior πιθανότητα η οποία είναι υποκειμενική.

Goodman, S.N. and R. Royall, **Evidence and scientific research**. Am J Public Health, 1988. **78**(12): p. 1568-74.
Goodman, S.N. **Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy.** Ann Intern Med. 1999;130:995-1004

Τι είναι το P-Hacking;

Συλλογή ή επιλογή δεδομένων ή στατιστικών αναλύσεων μέχρι μη στατιστικά σημαντικά αποτελέσματα να γίνουν σημαντικά.

Table 1. Tests for evidential value and p-hacking across disciplines, using *p*-values obtained from the Results section.

Discipline	Number of *p*-values between 0 and 0.025	Number of *p*-values between 0.025 and 0.05	Binomial test for evidential value	Number of *p*-values between 0.04 and 0.045	Number of *p*-values between 0.045 and 0.05	Binomial test forp-hacking
Agricultural and veterinary sciences	375	125	<0.001	10	16	0.163
Biological sciences	11,074	3,562	<0.001	350	423	0.005
Chemical sciences	380	110	<0.001	14	17	0.360
Earth sciences	76	25	<0.001	0	4	0.063
Education	280	101	<0.001	9	8	0.685
Engineering	471	183	<0.001	16	12	0.828
Environmental sciences	657	190	<0.001	10	19	0.068
Information and computing sciences	790	266	<0.001	20	30	0.101
Mathematical sciences	72	22	<0.001	3	0	1.000
Medical and health sciences	45,460	16,537	<0.001	1,477	1,785	<0.001
Multidisciplinary	21,209	6,793	<0.001	638	750	0.001
Psychology and cognitive sciences	1,355	487	<0.001	29	50	0.012
Studies in human society	139	45	<0.001	8	3	0.967
Technology	94	37	<0.001	3	3	0.656

Number of *p*-values in each bin is the mean number based on 1,000 bootstraps of one *p*-value per Results section, rounded to the nearest whole number. Disciplines (n = 8) for which we found fewer than 50 *p*-values below 0.05 in the Results section were excluded.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) **The Extent and Consequences of P-Hacking in Science**. PLoS Biol 13(3): e1002106

Τι είναι το P-Hacking;

Table 2. Tests for evidential value and p-hacking across disciplines, using *p*-values obtained from the Abstract.

Discipline	Number of *p*-values between 0 and 0.025	Number of *p*-values between 0.025 and 0.05	Binomial test for evidential value	Number of *p*-values between 0.04 and 0.045	Number of *p*-values between 0.045 and 0.05	Binomial test for p-hacking
Agricultural and veterinary sciences	96	35	<0.001	3	2	0.813
Biological sciences	1,787	632	<0.001	54	66	0.158
Chemical sciences	76	31	<0.001	3	4	0.500
Education	88	22	<0.001	2	0	1.000
Engineering	121	52	<0.001	2	1	0.875
Environmental sciences	42	15	<0.001	2	2	0.688
Information and computing sciences	251	105	<0.001	5	15	0.021
Medical and health sciences	18,428	6,692	<0.001	633	692	0.056
Multidisciplinary	5,056	1,621	<0.001	123	174	0.002
Psychology and cognitive sciences	98	37	<0.001	1	5	0.109

Number of *p*-values in each bin is the mean number based on 1,000 bootstraps of one *p*-value per Abstract, rounded to the nearest whole number. Disciplines (n = 12) for which we found fewer than 50 *p*-values below 0.05 in the Abstract were excluded.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) **The Extent and Consequences of P-Hacking in Science.** PLoS Biol 13(3): e1002106

Τι είναι το P-Hacking;

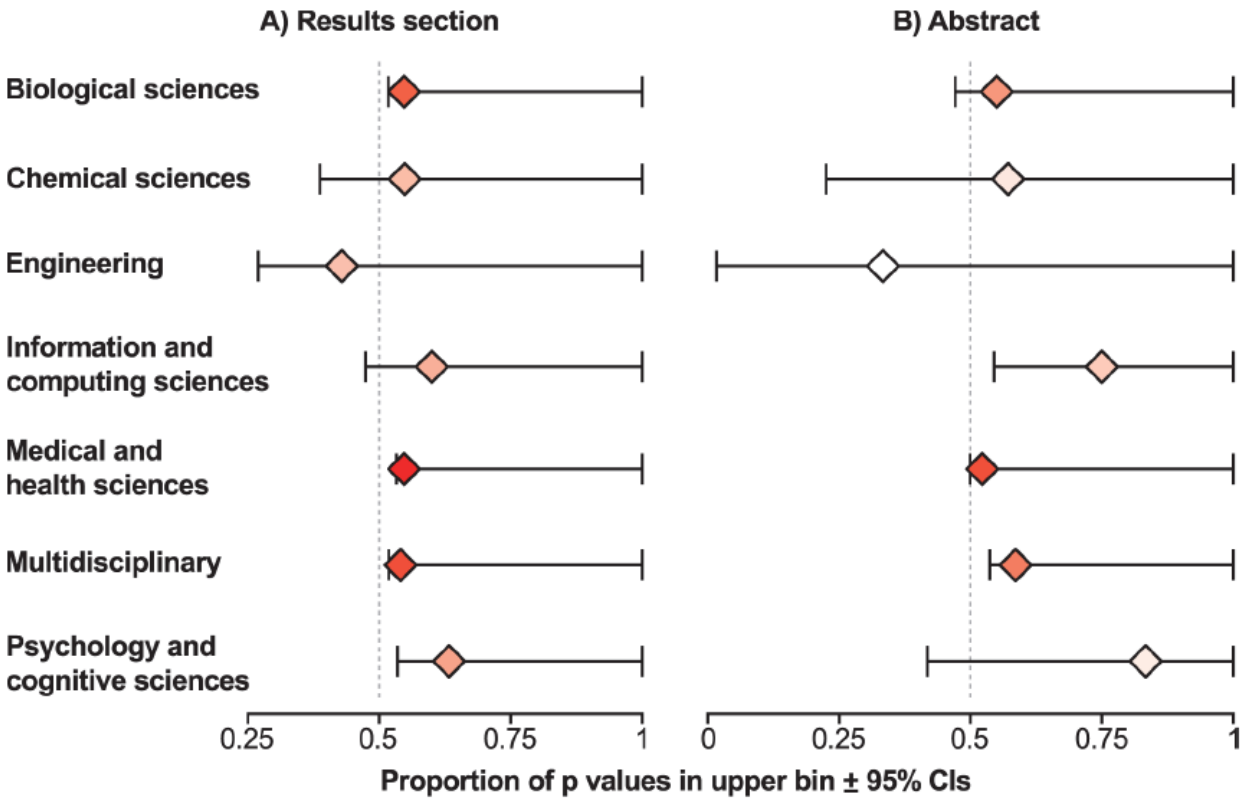


Fig 3. Evidence for p-hacking across scientific disciplines. A) Evidence for p-hacking from *p*-values obtained from Results sections. B) Evidence for p-hacking from *p*-values obtained from Abstracts. The strength of p-hacking is presented as the proportion of p-values in the upper bin ($0.045 < p < 0.05$) with one-tailed 95% confidence intervals (calculated following Clopper and Pearson [47] using the *binom.test* function in R). Only disciplines where text-mining of the Results sections returned more than 25 *p*-values between 0.04 and 0.05 are presented. Marker colour is shaded according to the sample size: with white indicating low samples sizes and red indicating larger sample sizes.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) **The Extent and Consequences of P-Hacking in Science.** PLoS Biol 13(3): e1002106

Γιατί τα περισσότερα ευρήματα δημοσιευμένων μελετών είναι λάθος;

- The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.
- The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.
- The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.
- The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.
- The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.
- The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

Ioannidis, J.P., **Why most published research findings are false**. PLoS Med, 2005. **2**(8): p. e124.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημιοσιευμένων μελετών (10)

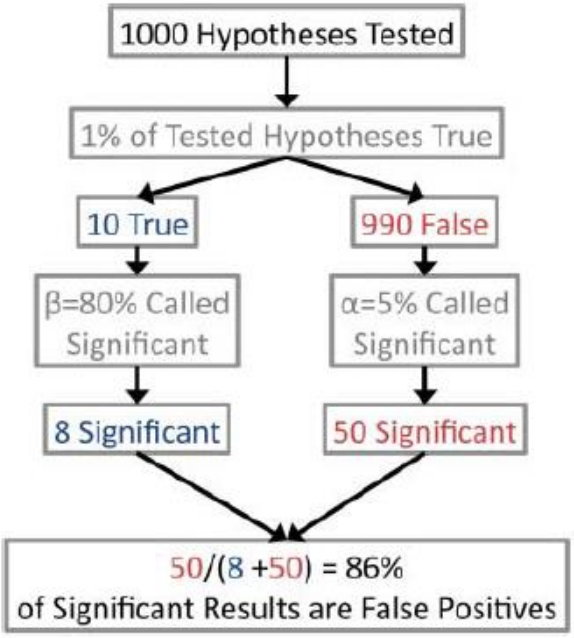
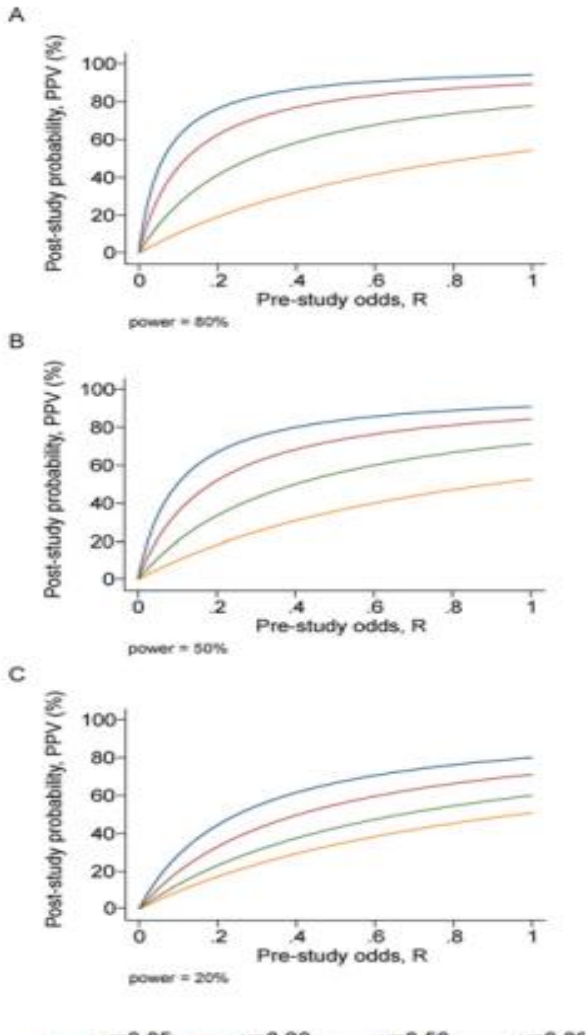


Fig. 1. The theoretical argument suggests that most published research is false. If the probability a research hypothesis is true is low, then most tested hypotheses will be false. The definition of P-values and customary significance cutoffs mean that $(\alpha \cdot 100)\%$ of false-positive hypotheses and $(\beta \cdot 100)\%$ of true-positive hypotheses will be called significant. If only 1% of tested hypotheses are true and the customary values of $\alpha = 0.05$ and $\beta = 0.8$ are used, then 86% of reported significant results will be false positives. This final percent of published results corresponding to false positives is the quantity that we estimate. A version of this figure appeared on the blog Marginal Revolution and is reproduced with permission (Tabarrok, 1989).

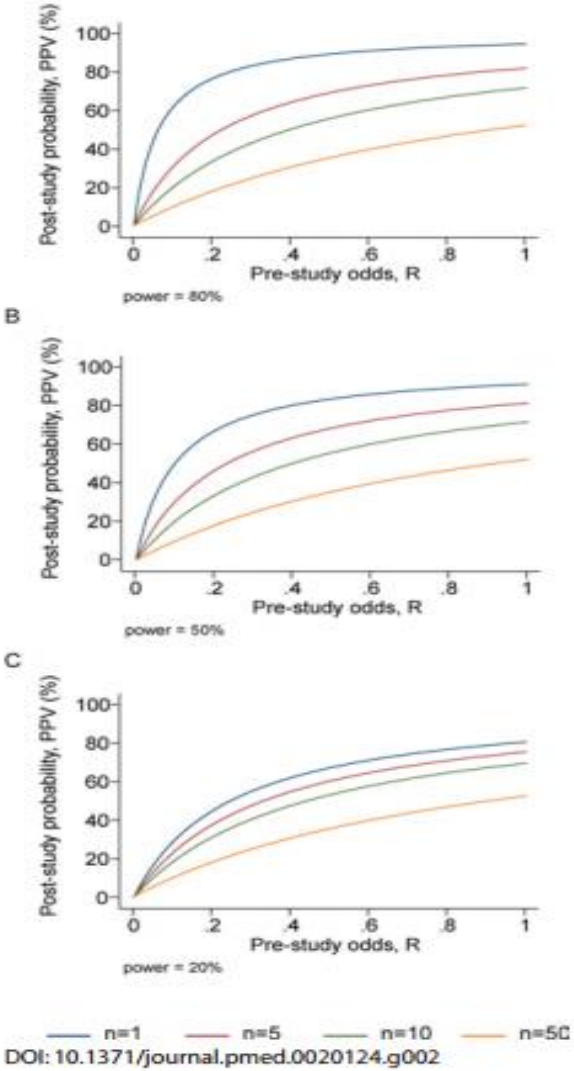
Jager, L.R. and J.T. Leek, **An estimate of the science-wise false discovery rate and application to the top medical literature.** Biostatistics, 2014. **15**(1): p. 1-12.



DOI: 10.1371/journal.pmed.0020124.g001

Figure 1. PPV (Probability That a Research Finding Is True) as a Function of the Pre-Study Odds for Various Levels of Bias, *u*

Panels correspond to power of 0.20, 0.50, and 0.80.



DOI: 10.1371/journal.pmed.0020124.g002

Figure 2. PPV (Probability That a Research Finding Is True) as a Function of the Pre-Study Odds for Various Numbers of Conducted Studies, *n*

Panels correspond to power of 0.20, 0.50, and 0.80.

Ioannidis, J.P., **Why most published research findings are false**. PLoS Med, 2005. **2**(8): p. e124.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημιοσιευμένων μελετών (1)

- Η επιλογή του p-value=0.05 ως κατώφλι στατιστικής σημαντικότητας οδηγεί σε λάθος συμπεράσματα στο 30% των περιπτώσεων.
- Εαν τα πειράματα έχουν χαμηλή ισχύ οδηγούμαστε σε λάθος συμπεράσματα τις περισσότερες φορές.

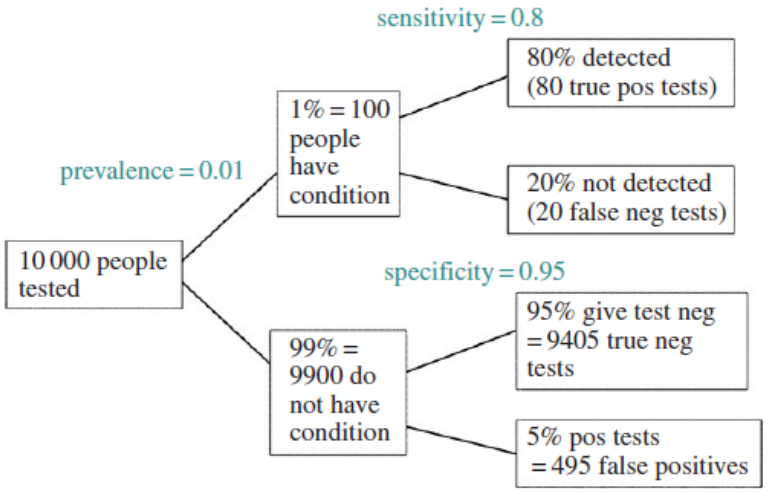


Figure 1. Tree diagram to illustrate the false discovery rate in screening tests. This example is for a prevalence of 1%, specificity 95% and sensitivity 80%. Out of 10 000 people screened, 495 + 80 = 575 give positive tests. Of these, 495 are false positives so the false discovery rate is 86%.

Colquhoun D. 2014 **An investigation of the false discovery rate and the misinterpretation of p-values.** R. Soc. open sci. 1: 140216.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημοσιευμένων μελετών (2)

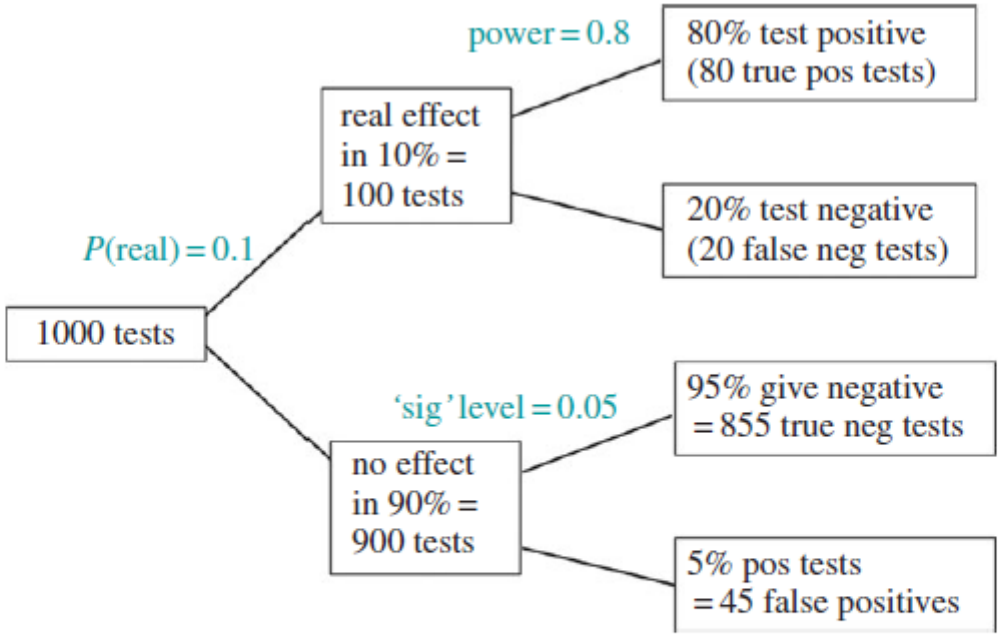


Figure 2. Tree diagram to illustrate the false discovery rate in significance tests. This example considers 1000 tests, in which the prevalence of real effects is 10%. The lower limb shows that with the conventional significance level, $p = 0.05$, there will be 45 false positives. The upper limb shows that there will be 80 true positive tests. The false discovery rate is therefore $45/(45 + 80) = 36\%$, far bigger than 5%.

Colquhoun D. 2014 **An investigation of the false discovery rate and the misinterpretation of p-values.** R. Soc. open sci. 1: 140216.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημοσιευμένων μελετών (3)

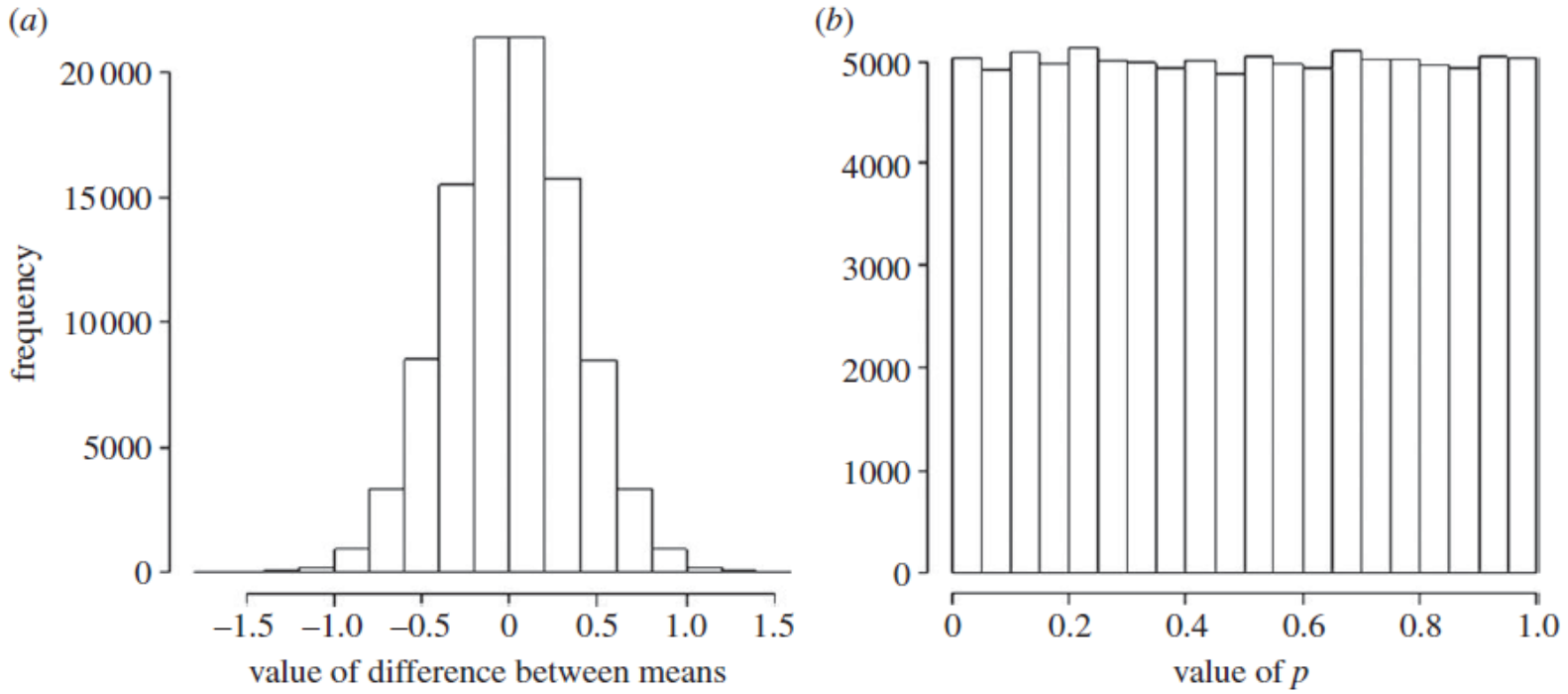


Figure 3. Results of 100 000 simulated *t*-tests, when the null hypothesis is true. The test looks at the difference between the means of two groups of observations which have identical true means, and a standard deviation of 1. (*a*) The distribution of the 100 000 'observed' differences between means (it is centred on zero and has a standard deviation of 0.354). (*b*) The distribution of the 100 000 *p*-values. As expected, 5% of the tests give (false) positives ($p \leq 0.05$), but the distribution is flat (uniform).

Colquhoun D. 2014 **An investigation of the false discovery rate and the misinterpretation of p-values.** R. Soc. open sci. 1: 140216.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημοσιευμένων μελετών (4)

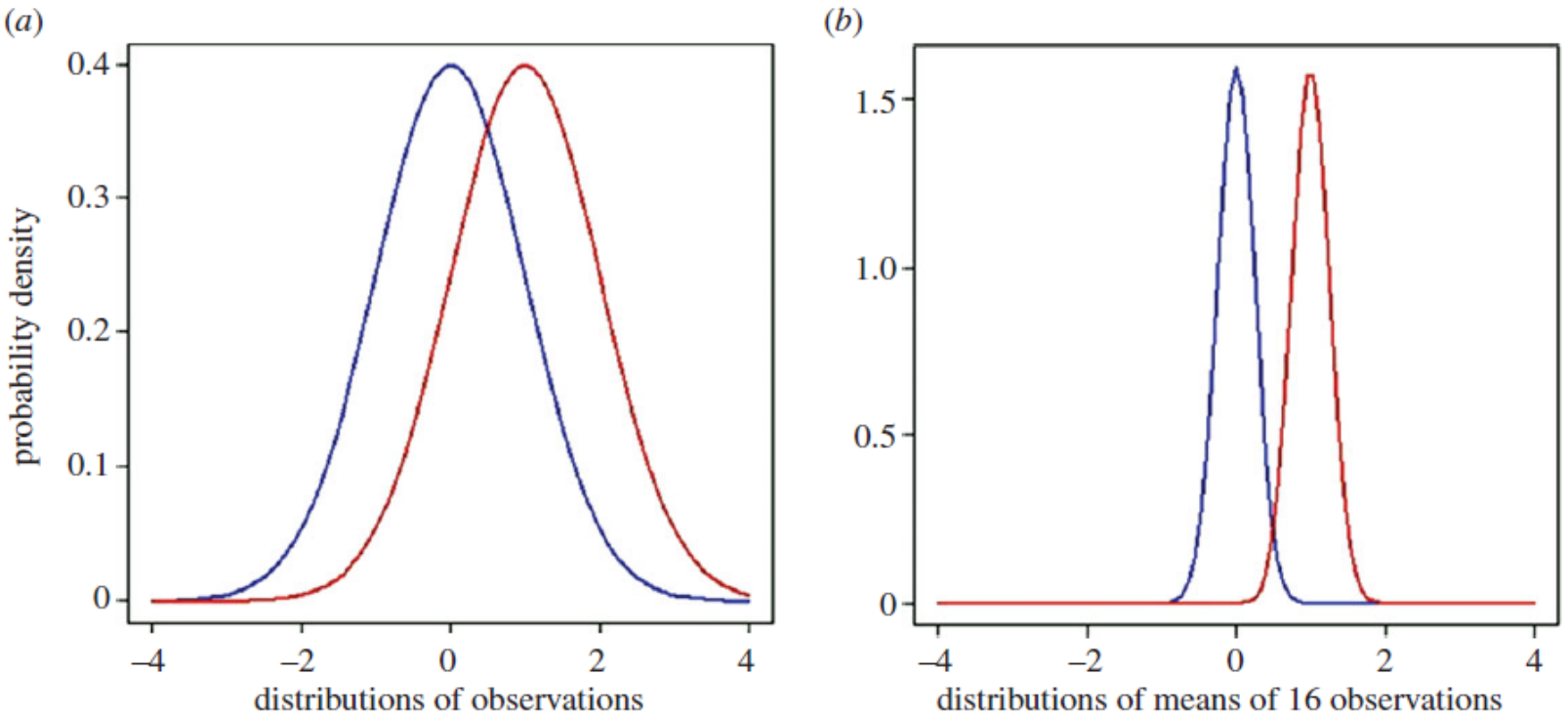


Figure 4. The case where the null hypothesis is *not* true. Simulated *t*-tests are based on samples from the postulated true distributions shown: blue, control group; red, treatment group. The observations are supposed to be normally distributed with means that differ by 1 s.d., as shown in (*a*). The distributions of the means of 16 observations are shown in (*b*).

Colquhoun D. 2014 **An investigation of the false discovery rate and the misinterpretation of p-values.** R. Soc. open sci. 1: 140216.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημιοσιευμένων μελετών (5)

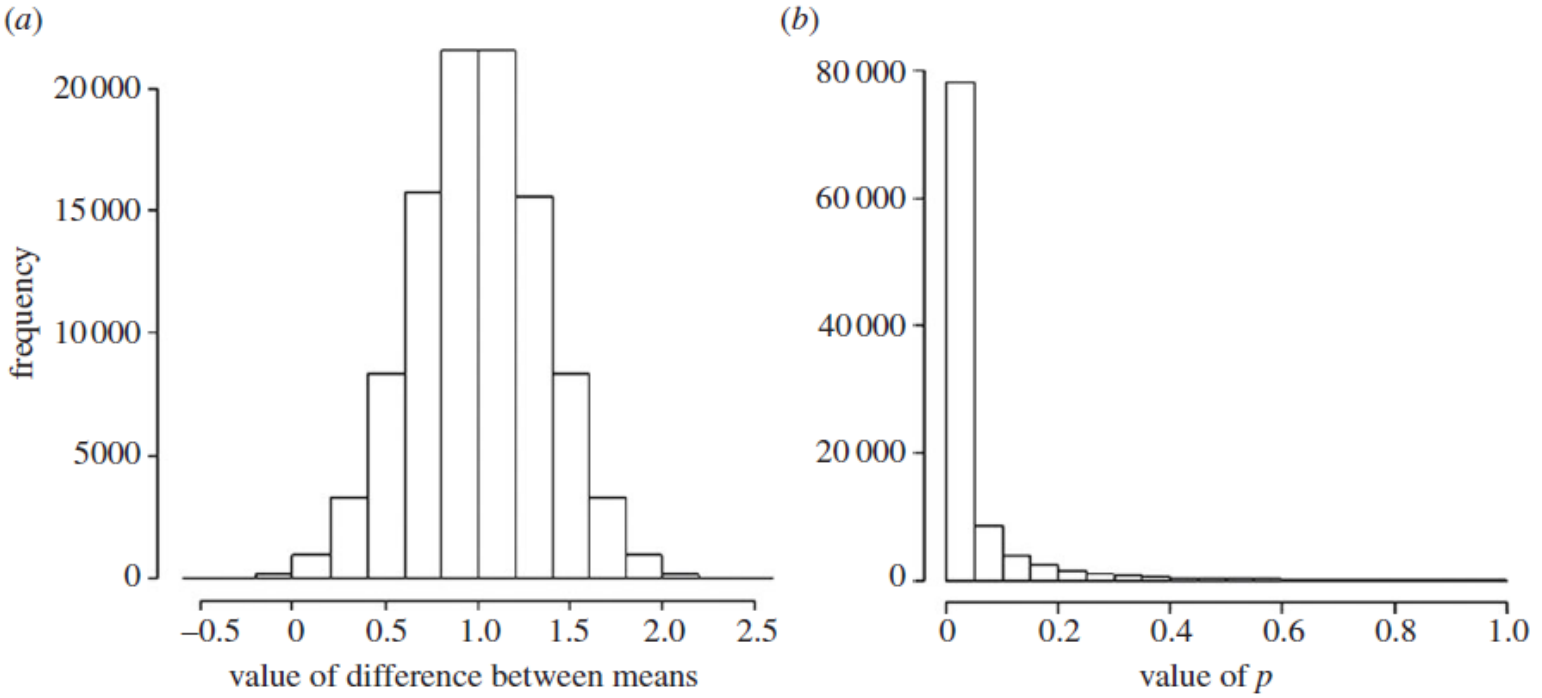


Figure 5. Results of 100 000 simulated *t*-tests in the case where the null hypothesis is *not* true, but as shown in figure 4. (*a*) The distribution of the 100 000 'observed' values for the differences between means of 16 observations. It has a mean of 1, and a standard deviation of 0.354. (*b*) The distribution of the 100 000 *p*-values: 78% of them are equal to or less than 0.05 (as expected from the power of the tests).

Colquhoun D. 2014 **An investigation of the false discovery rate and the misinterpretation of p-values.** R. Soc. open sci. 1: 140216.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημιοσιευμένων μελετών (6)

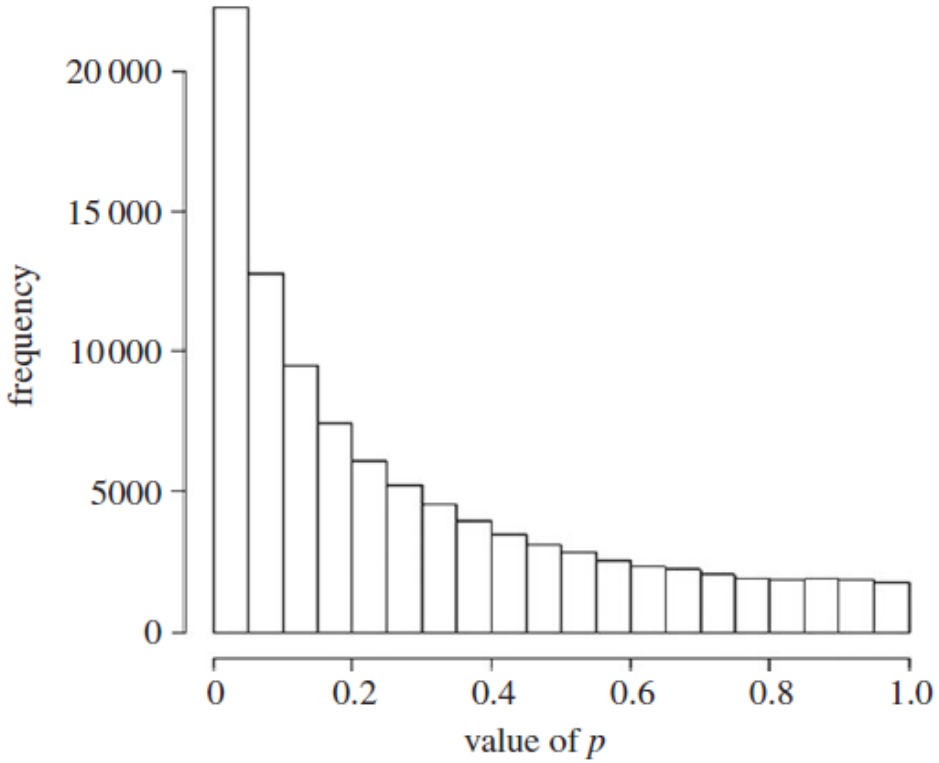


Figure 6. Distribution of 100 000 *p*-values from tests like those in figure 5, but with only four observations in each group, rather than 16. The calculated power of the tests is only 0.22 in this case, and it is found, as expected, that 22% are $p \leq 0.05$.

Colquhoun D. 2014 **An investigation of the false discovery rate and the misinterpretation of p-values.** R. Soc. open sci. 1: 140216.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημιοσιευμένων μελετών (7)

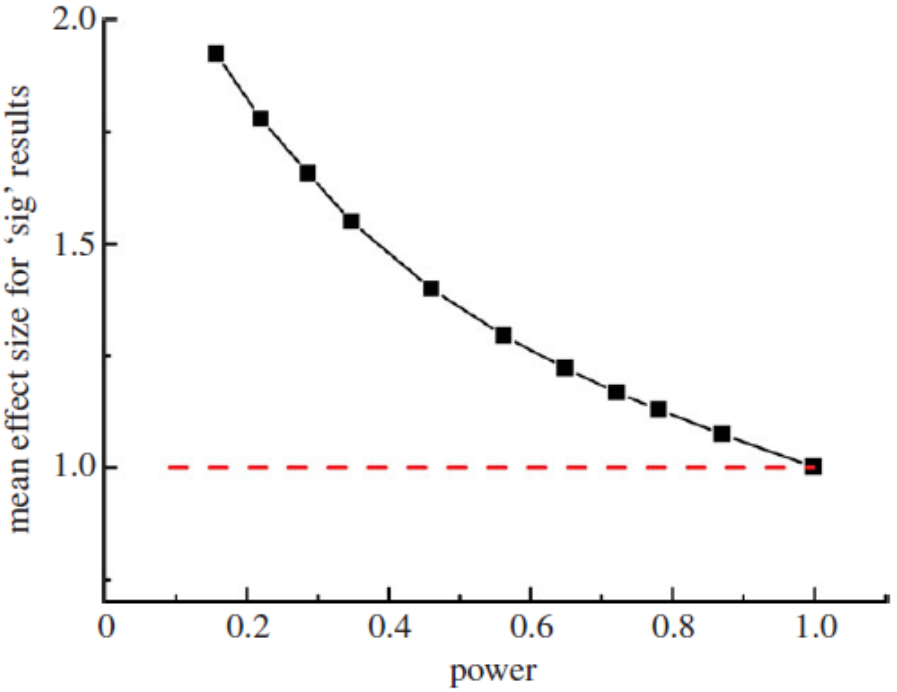


Figure 7. The average difference between means for all tests that came out with $p \leq 0.05$. Each point was found from 100 000 simulated *t*-tests, with data as in figure 4. The power of the tests was varied by changing the number, *n*, of 'observations' that were averaged for each mean. This varied from $n = 3$ (power $= 0.157$) for the leftmost point, to $n = 50$ (power=0.9986) for the rightmost point. Intermediate points were calculated with $n = 4, 5, 6, 8, 10, 12, 14, 16$ and 20.

Colquhoun D. 2014 **An investigation of the false discovery rate and the misinterpretation of p-values.** R. Soc. open sci. 1: 140216.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημιοσιευμένων μελετών (8)

What can be done:

❖Note that all statistical tests of significance assume that the treatments have been allocated at random. This means that application of significance tests to observational data, e.g. epidemiological surveys of diet and health, is not valid.

❖Never, ever, use the word ‘significant’ in a paper.

❖If you do a significance test, just state the p-value and give the effect size and confidence intervals.

❖Observation of a p~0.05 means nothing more than ‘worth another look’.

❖Do some rough calculations of the sample size that might be needed to show a worthwhile effect.

❖If you want to avoid making a fool of yourself very often, do not regard anything greater than $p<0.001$ as a demonstration that you have discovered something.

Colquhoun D. 2014 **An investigation of the false discovery rate and the misinterpretation of p-values.** R. Soc. open sci. 1: 140216.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημιοσιευμένων μελετών (9)

- Υιοθετήθηκαν μέθοδοι εκτίμησης από το χώρο της γονιδιωματικής για τον υπολογισμό του ποσοστού των ψευδών θετικών ευρημάτων στην ιατρική έρευνα.
- Από ένα σύνολο 77,430 δημοσιευμένων μελετών στα περιοδικά The Lancet, The Journal of the American Medical Association, The New England Journal of Medicine, The British Medical Journal, and The American Journal of Epidemiology από το 2000-2010 έγινε καταγραφή 5322 p-values.
- Του ποσοστό των ψευδών θετικών ευρημάτων ήταν 14% (s.d. 1%) κάτι που έρχεται σε αντίθεση με προηγούμενες εκτιμήσεις (Ioannidis, J.P., **Why most published research findings are false**. PLoS Med, 2005. **2**(8): p. e124.).
- Επίπλεον, βρέθηκε ότι δεν υπάρχει σημαντική αύξηση των ψευδώς θετικών ευρημάτων με το πέρασμα του χρόνου.

Jager, L.R. and J.T. Leek, **An estimate of the science-wise false discovery rate and application to the top medical literature.** Biostatistics, 2014. **15**(1): p. 1-12.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημιοσιευμένων μελετών (10)

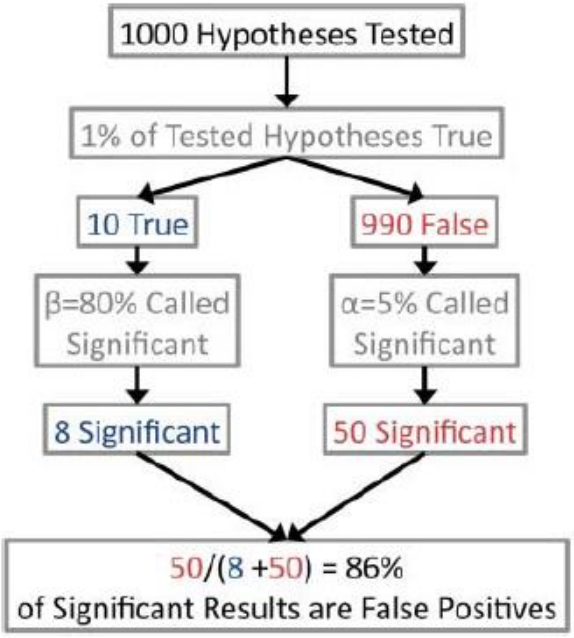


Fig. 1. The theoretical argument suggests that most published research is false. If the probability a research hypothesis is true is low, then most tested hypotheses will be false. The definition of P-values and customary significance cutoffs mean that $(\alpha \cdot 100)\%$ of false-positive hypotheses and $(\beta \cdot 100)\%$ of true-positive hypotheses will be called significant. If only 1% of tested hypotheses are true and the customary values of $\alpha = 0.05$ and $\beta = 0.8$ are used, then 86% of reported significant results will be false positives. This final percent of published results corresponding to false positives is the quantity that we estimate. A version of this figure appeared on the blog Marginal Revolution and is reproduced with permission (Tabarrok, 1989).

Jager, L.R. and J.T. Leek, **An estimate of the science-wise false discovery rate and application to the top medical literature.** Biostatistics, 2014. **15**(1): p. 1-12.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημοσιευμένων μελετών (11)

Jager, L.R. and J.T. Leek, **An estimate of the science-wise false discovery rate and application to the top medical literature.** Biostatistics, 2014. **15**(1): p. 1-12.

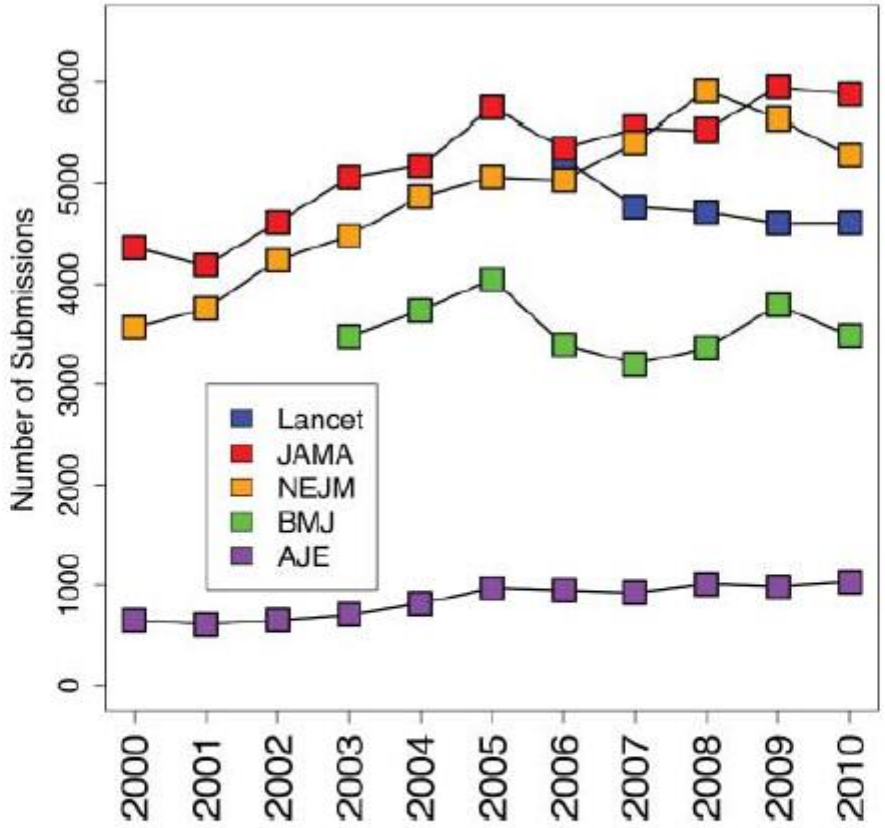


Fig. 2. Major medical journal submissions are increasing over time. A plot of the number of submissions to the major medical journals *The Lancet*, *The Journal of the American Medical Association* (*JAMA*), *The New England Journal of Medicine* (*NEJM*), *The British Medical Journal* (*BMJ*) and the flagship epidemiological journal *The American Journal of Epidemiology* (*AJE*) between the years 2000 and 2010. Submission data are available only for the years 2006–2010 for *The Lancet* and the years 2003–2010 for *The BMJ*.

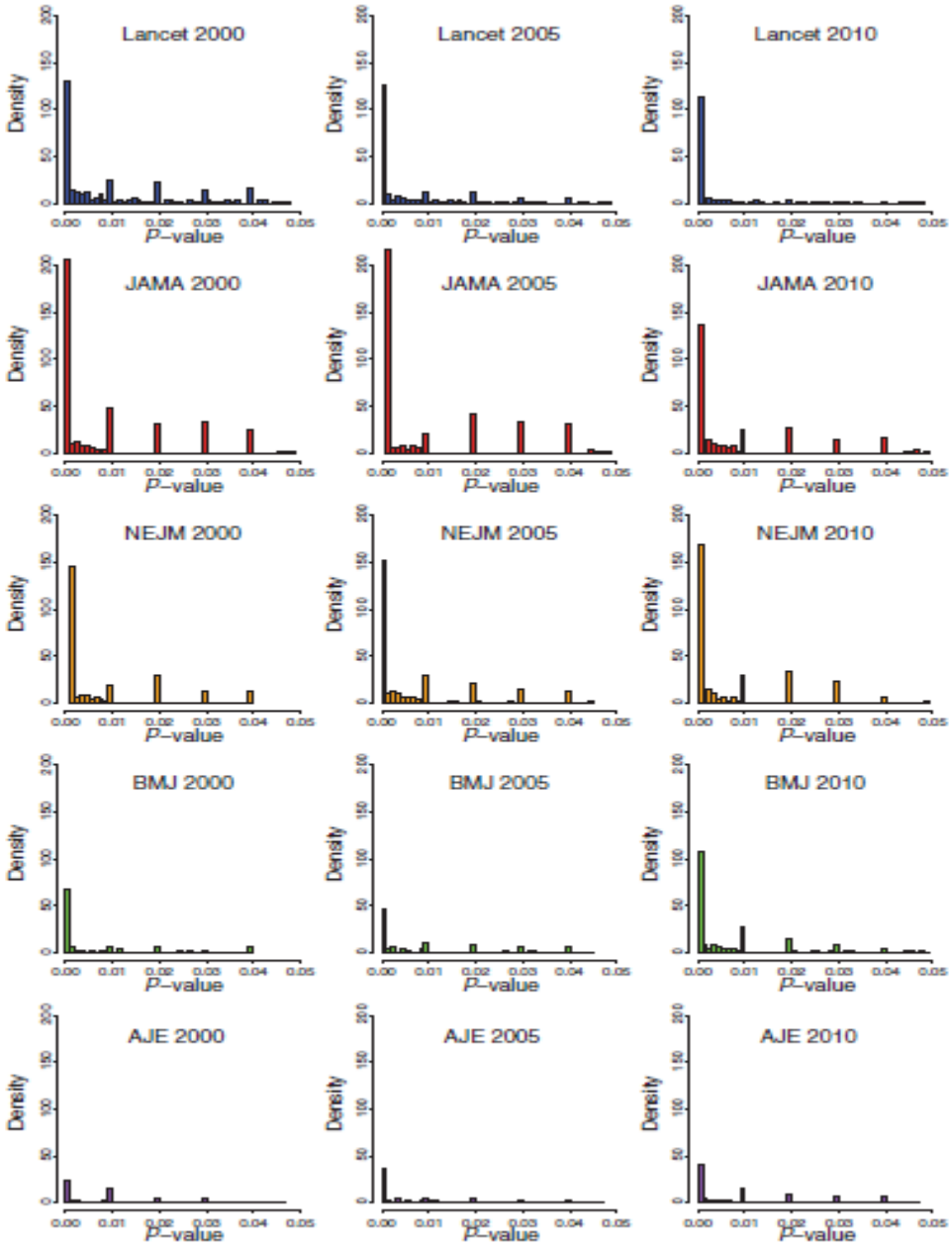


Fig. 3. Histogram of P-values < 0.05. The observed P-value distributions for all $P < 0.05$ scraped from PubMed for *AJE*, *JAMA*, *NEJM*, *BMJ*, and *The Lancet* in the years 2000, 2005, and 2010.

Διερεύνηση των ψευδώς θετικών ευρημάτων δημιοσιευμένων μελετών (12)

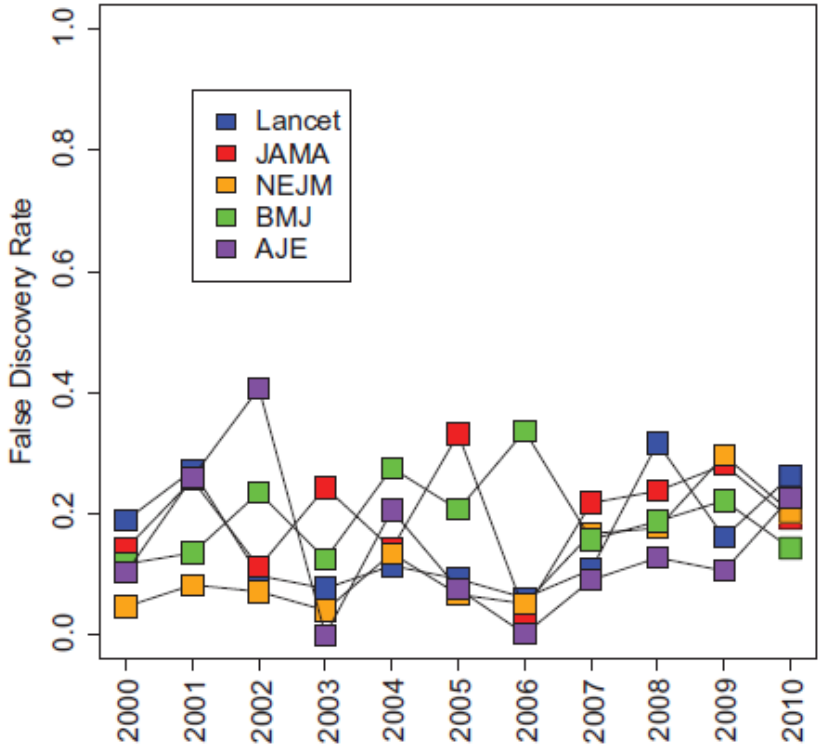


Fig. 4. Estimated false discovery rates for the years 2000–2010 by journal. The estimated false discovery rates for *AJE*, *JAMA*, *NEJM*, *BMJ*, and *The Lancet* in the years 2000, 2005, and 2010. There is no significant trend in false discovery rates over time or with increasing numbers of submissions.

Jager, L.R. and J.T. Leek, **An estimate of the science-wise false discovery rate and application to the top medical literature.** Biostatistics, 2014. **15**(1): p. 1-12.

Γιατί οι περισσότερες συσχετίσεις είναι υπερεκτιμημένες; (1)

TABLE 1. Selected Evaluations Suggesting That Early Discovered Effects Are Inflated

Research Field	Theoretical Work or Empirical Evidence and References
Highly cited clinical research	A quarter of most-cited clinical trials and 5/6 most-cited epidemiological studies were either fully contradicted or found to have exaggerated results[2]
Early stopped clinical trials	Early stopping results in inflated effects in theory[3,4] and shown also in practice[5]
Clinical trials of mental health interventions	More likely for effect sizes of pharmacotherapies to diminish than to increase over time[6]
Clinical trials on heart failure interventions	"Regression to the truth" in phase III trials for interventions with early promising results[7]
Clinical trials on diverse interventions	Effectiveness shown to fade over time[8]
Multiple meta-analyses on effectiveness	Eleven independent meta-analyses on acetylcysteine show decreasing effects over time[9]
Epidemiologic associations	Expected to be inflated in multiple testing with significance threshold; empirical demonstration for occupational carcinogens[10]
Pharmacoepidemiology	"Phantom ship" associations that do not stand upon further evaluation[11]
Gene-disease associations	Several empirical evaluations showing dissipation of effect sizes over time[12–15]
Linkage studies in humans	Theory anticipates large upward bias ("winner's curse") in effects of discovered loci[16–18]
Genetic traits in experimental crosses	As above (actually literature on the "Beavis effect" precedes literature on humans)[19–22]
Genome-wide associations	Large winner's curse anticipated for discovered effects in underpowered conditions[23,24]
Ecology and evolution	Empirical demonstration that relationships fade over time[25,26]
Psychology	Replication studies in psychology failing to confirm true effects because the new studies were underpowered due to reliance on the estimate of effect from the original positive study[27]
Early repeated data peaking in general	Simulations to model inflation of effects with repeated data peaking[28]
Prognostic models	Overestimated prognostic performance with stepwise selection of variables based on significance thresholds[29–32]
Regression models in general	Exaggerated effects (coefficients) with stepwise selection based on significance thresholds and small datasets[32–34]; may correct substantially if a very lenient alpha = 0.20 is used for selection[34] [thus having enough power]

Ioannidis, J.P., **Why most discovered true associations are inflated.** Epidemiology, 2008. **19**(5): p. 640-8.

Γιατί οι περισσότερες συσχετίσεις είναι υπερεκτιμημένες; (2)

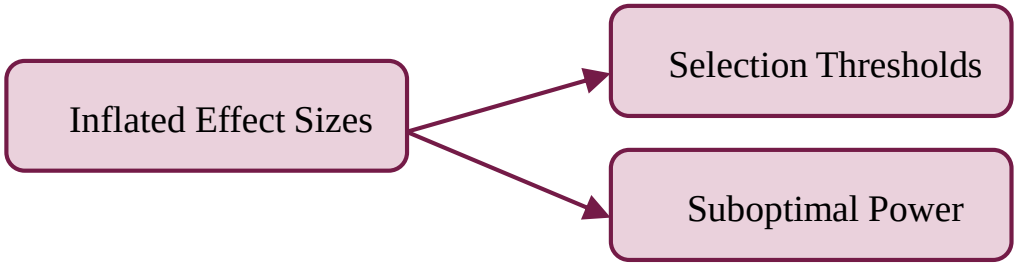


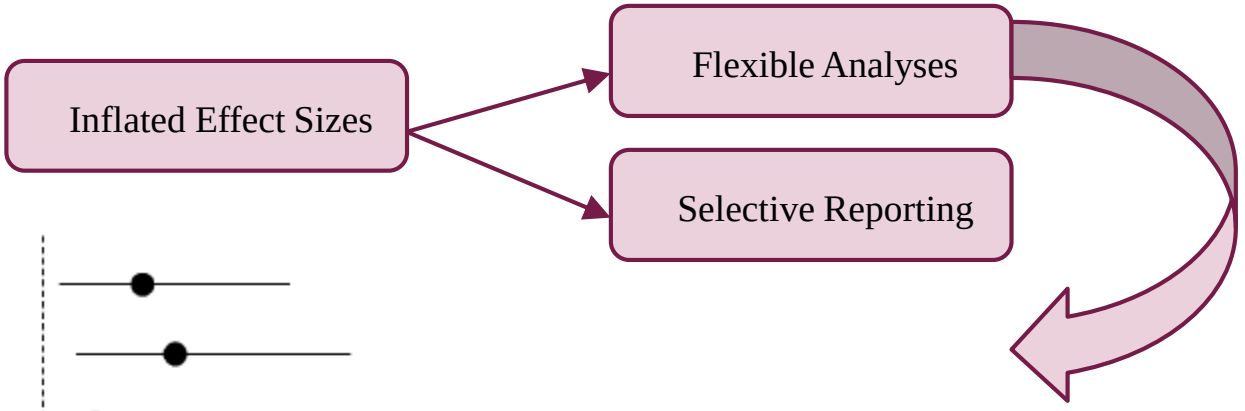
TABLE 2. Simulations for Effect Sizes Passing the Threshold of Formal Statistical Significance ($P = 0.05$)

			Observed OR in Significant Associations	
True OR	Control Group Rate (%)	Sample n Per Group	Median (IQR)	Median Fold Inflation
1.10	30	1000	1.23 (1.23–1.29)	1.11
1.10	30	250	1.51 (1.49–1.55)	1.37
1.25	30	1000	1.29 (1.26–1.39)	1.03
1.25	30	250	1.60 (1.50–1.67)	1.28
1.25	30	50	2.73 (2.60–3.16)	2.18

IQR indicates interquartile range.

Ioannidis, J.P., **Why most discovered true associations are inflated.** Epidemiology, 2008. **19**(5): p. 640-8.

Γιατί οι περισσότερες συσχετίσεις είναι υπερεκτιμημένες; (3)



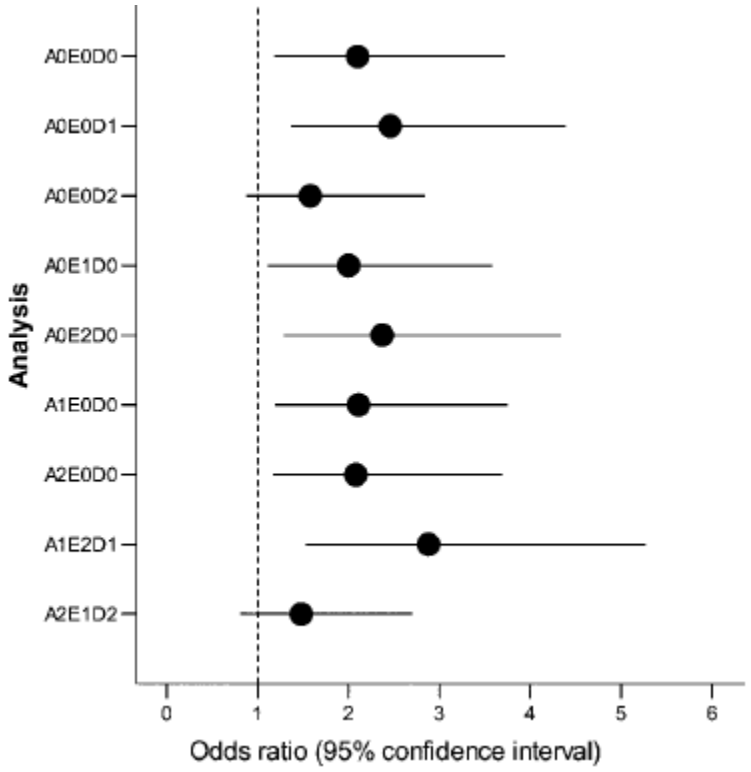


FIGURE 1. Vibration of an effect size: the odds ratio with 95% confidence interval is obtained for a simulated study, with or without adjustment (A0 indicates no adjustment; A1, adjustment for one randomly generated variable; A2, adjustment for another randomly generated variable), application of various eligibility criteria (E0 indicates all participants included; E1, excluding 6% of the participants according to high values on a random variable; E2, excluding 10% of the participants according to high values on a random variable) and arbitration of 2% of the data on the exposure based on knowledge of outcome (D0 indicates no arbitration; D1, 2% of the exposure data [5 observations] changed to be consistent with the direction of the association; D2, 2% of the exposure data [5 observations] changed to be against the direction of the postulated association).

Ioannidis, J.P., **Why most discovered true associations are inflated.** Epidemiology, 2008. **19**(5): p. 640-8.

Επιλεκτική παρουσίαση αποτελεσμάτων

- Αξιολογήθηκαν 389 επιδημιολογικές μελέτες οι οποίες ανέφεραν στην περίληψή τους τουλάχιστον ένα μέτρο σχετικού κινδύνου για κάποιον συνεχή παράγοντα.
- Σε 342 μελέτες (87.9%) ένα ή περισσότερα στατιστικά σημαντικά αποτελέσματα αναφέρονταν στην περίληψη ενώ μόλις σε 169 μελέτες παρουσιαζόταν κάποιο μη σημαντικό αποτέλεσμα.
- Ανάμεσα σε 50 τυχαία επιλεγμένες μελέτες όπου εξετάστηκε το πλήρες κείμενο, η διάμεσος των στατιστικά σημαντικών σχετικών κινδύνων ήταν 9 (IQR=5–16) και 6 των μη σημαντικών (IQR=3–16) (p= 0.25).
- Παρατηρήθηκε επιλεκτική παρουσίαση των αποτελεσμάτων μεταξύ των πιο ακραίων ομάδων στις περιπτώσεις όπου ο σχετικός κίνδυνος ήταν εγγενώς μειωμένος.

Kavvoura, F.K., G. Liberopoulos, and J.P. Ioannidis, **Selection in reported epidemiological risks: an empirical assessment.** PLoS Med, 2007. **4**(3): p. e79.

Table 1. Characteristics of Analyzed Studies

Characteristic	Journal or Category	Articles (*n* [%])
Most frequent journals	*Cancer Epidemiology, Biomarkers, and Prevention*	29 (7.5)
	American Journal of Clinical Nutrition	23 (5.9)
	American Journal of Epidemiology	21 (5.4)
	International Journal of Cancer. Journal International du Cancer	17 (4.4)
	Diabetes Care	16 (4.1)
	Journal of the National Cancer Institute	12 (3.1)
	Archives of Internal Medicine	12 (3.1)
	Stroke	12 (3.1)
	Circulation	10 (2.6)
	Neurology	9 (2.3)
Impact factor > 7		90 (23.1)
US affiliation		199 (51.2)
More than one publication from same cohort[a]		152 (39.1)
Structured abstract		268 (68.9)
Design and metric	Case-control, OR	72 (18.5)
	Other, OR	82 (21.1)
	Cohort, other than OR	235 (60.4)
Any significant relative risk		342 (87.9)
Any nonsignificant relative risk		169 (43.4)
First presented percentile contrasts[b]	Median	11 (2.8)
	Extreme tertiles	75 (19.3)
	Extreme quartiles	167 (42.9)
	Extreme quintiles	110 (28.3)
	Extreme tertile versus other	7 (1.8)
	Extreme quartile versus other	16 (4.1)
	Extreme quintile versus other	3 (0.8)

Kavvoura, F.K., G. Liberopoulos, and J.P. Ioannidis, **Selection in reported epidemiological risks: an empirical assessment.** PLoS Med, 2007. **4**(3): p. e79.

Table 3. Logistic Regressions for Presence of Significant Relative Risks and Nonsignificant Relative Risks

Variable	Group	Presence of Statistically Significant Relative Risks in the Abstract			Presence of Statistically Nonsignificant Relative Risks in the Abstract		
		n/N (%)	Univariate OR (95% CI)	Multivariate OR (95% CI)	*n/N* (%)	Univariate OR (95% CI)	Multivariate OR (95% CI)
Impact factor	>7	83/90 (92.2)	1.83 (0.79–4.24)	Not selected	38/90 (42.5)	0.94 (0.58–1.51)	Not selected
	<7	259/299 (86.6)	Reference	Not selected	131/299 (43.8)	Reference	Not selected
Country	United States	167/199 (83.9)	0.45 (0.26–0.86)	0.41 (0.20–0.86)	107/199 (53.8)	2.40 (1.59–3.63)	3.10 (1.84–5.24)
	Other	175/190 (92.1)	Reference	Reference	62/190 (32.6)	Reference	Reference
Cohort with more than one article	Yes	126/152 (82.9)	0.47 (0.26–0.87)	Not selected	86/152 (50.0)	2.42 (1.59–3.67)	Not selected
	No	216/237 (91.1)	Reference	Not selected	83/237 (35.0)	Reference	Not selected
Design and metric	Cohort, not OR	204/235 (86.8)	Reference	Not selected	109/235 (46.4)	Reference	Not selected
	Case-control, OR	60/72 (83.3)	0.72 (0.36–1.45)	Not selected	37/72 (51.4)	1.22 (0.72–2.07)	Not selected
	Other, OR	78/82 (95.1)	2.94 (1.01–8.58)	Not selected	23/82 (28.0)	0.45 (0.26–0.78)	Not selected
Structured abstract	Yes	249/268 (92.9)	4.08 (2.17–7.66)	2.25 (1.06–4.80)	98/268 (36.6)	0.41 (0.26–0.63)	Not selected
	No	93/121 (76.9)	Reference	Reference	71/121 (58.7)	Reference	Not selected
Tested risk factor[a]	Dietary	103/122 (84.4)	Reference	Not selected	71/122 (58.2)	Reference	Reference
	Toxic exposures	10/12 (83.3)	0.92 (0.19–4.55)	Not selected	8/12 (66.7)	1.44 (0.41–5.03)	2.03 (0.51–8.10)
	Biological markers	134/154 (87.0)	1.24 (0.63–2.44)	Not selected	65/154 (42.2)	0.53 (0.32–0.85)	0.78 (0.44–1.41)
	Psychosocial	28/31 (90.3)	1.72 (0.48–6.24)	Not selected	9/31 (29.0)	0.29 (0.13–0.69)	0.91 (0.34–2.40)
	Physical activity	8/8 (100.0)	Undefined	Not selected	4/8 (50.0)	0.72 (0.17–3.01)	1.73 (0.32–9.51)
	Body composition	24/25 (96.0)	4.43 (0.57–34.7)	Not selected	7/25 (4.1)	0.28 (0.11–0.72)	0.31 (0.09–1.06)
	Other	35/37 (94.6)	3.23 (0.72–14.56)	Not selected	5/37 (13.5)	0.11 (0.04–0.31)	0.12 (0.03–0.45)
Tested outcome[a]	Mortality	40/45 (88.9)	0.35 (0.10–1.21)	0.38 (0.10–1.47)	12/45 (26.7)	0.87 (0.41–1.86)	1.41 (0.59–3.69)
	Non-mortality Malignancies	79/109 (72.5)	0.12 (0.05–0.29)	0.19 (0.07–0.51)	83/109 (76.1)	7.68 (4.35–13.56)	8.16 (4.17–15.9)
	Vascular	86/92 (93.5)	0.63 (0.20–2.01)	0.56 (0.16–1.94)	32/92 (34.8)	1.28 (0.73–2.25)	2.32 (1.16–4.66)
	Other	137/143 (95.8)	Reference	Reference	42/143 (29.4)	Reference	Reference
SE of lnRR (per 1)			4.66 (0.47–46.8)	Not selected		0.54 (0.14–2.07)	Not selected

$n = 389$ studies.
OR, odds ratio; SE, standard error; lnRR, natural logarithm of relative risk.
[a]For the first presented relative risk in the abstract.
doi:10.1371/journal.pmed.0040079.t003

Kavvoura, F.K., G. Liberopoulos, and J.P. Ioannidis, **Selection in reported epidemiological risks: an empirical assessment.** PLoS Med, 2007. **4**(3): p. e79.

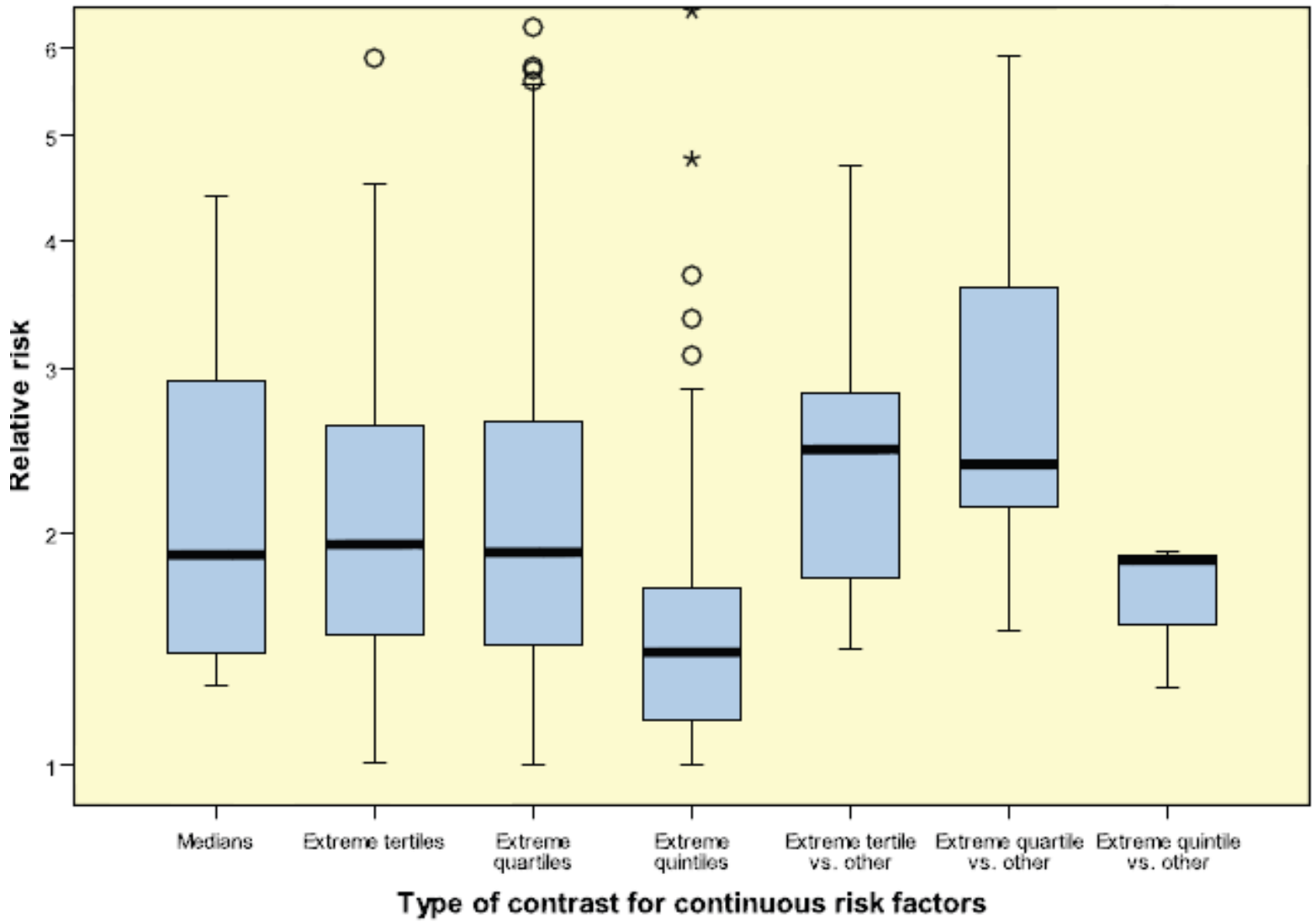


Figure 2. Box Plots for Relative Risks for Different Contrasts of the Values of the Postulated Risk Factor

All relative risks have been coined to be ≥1.00 for consistency.

doi:10.1371/journal.pmed.0040079.g002

Kavvoura, F.K., G. Liberopoulos, and J.P. Ioannidis, **Selection in reported epidemiological risks: an empirical assessment.** PLoS Med, 2007. **4**(3): p. e79.

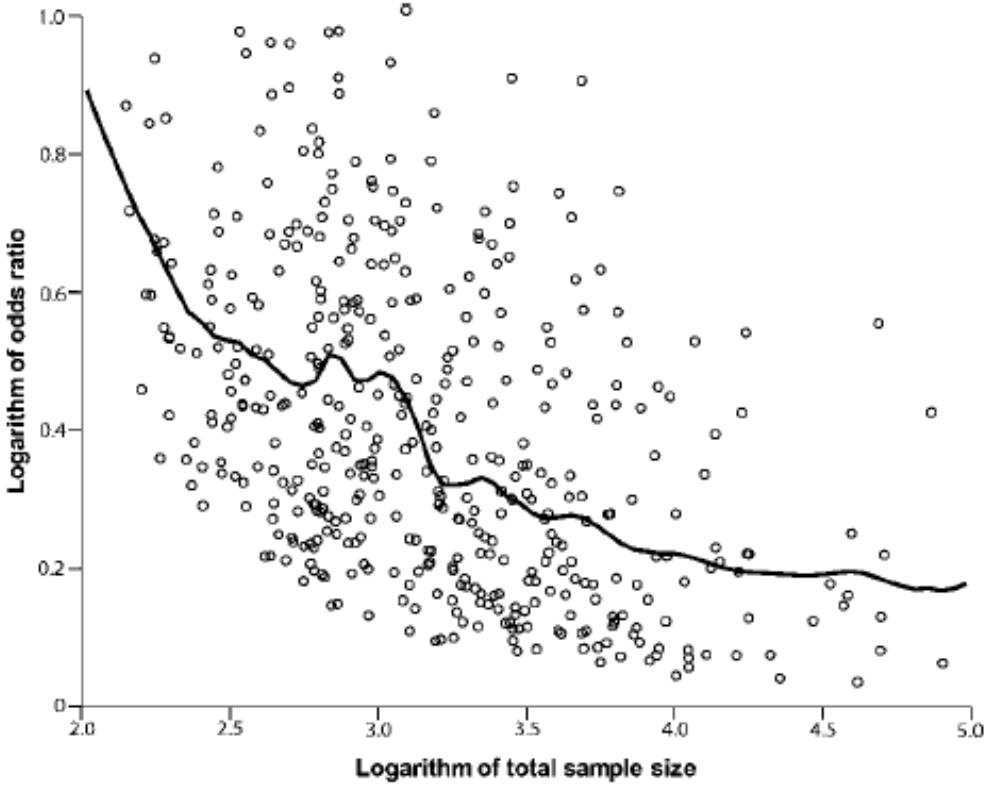


FIGURE 2. Relationship between total sample size and the effect size (odds ratio) for 256 Cochrane meta-analyses with formally statistically significant results ($P < 0.05$ according to random effects calculations) and at least 4 included studies. Both axes are in log10 scale. Also shown is a fit LOESS line. All odds ratios have been coined to be >1.00 for consistency. The median effect size for the 40 meta-analyses with at least 10,000 subjects is 1.53. Not shown are 5 outliers with extreme sample size or effect size.

Ioannidis, J.P., **Why most discovered true associations are inflated.** Epidemiology, 2008. **19**(5): p. 640-8.

Προτάσεις για την ορθή εκτίμηση συσχετίσεων (1)

TABLE 3. Avoiding Being Misled on Effect Sizes of True Associations in Early Discovery

Be cautious about effect sizes (and even about the mere presence of any effect in new discoveries)
Consider rational down-adjustment of effect sizes
Consider analytical methods that correct for anticipated inflation
Ignore effect sizes arising from discovery research
Conduct large studies in discovery phase
Use strict protocols for analyses
Adopt complete and transparent reporting of all results
Use methodologically rigorous, unbiased replication (potentially ad infinitum)
Be fair with interpretation

TABLE 4. Two Stances in Hunting Associations

	Aggressive Discoverer	Reflective Replicator
What matters is …	Discovery	Replication
Databases are …	Private goldmines not to be shared	Public commodity
A good epidemiologist …	Can think of more exploratory analyses	Is robust about design and analysis plan
One should report …	What is interesting	Everything
Publication mode	Publish each association as a separate paper	Publish everything as single paper
After reporting …	Push your findings forward	Be critical/cautious

Ioannidis, J.P., **Why most discovered true associations are inflated.** Epidemiology, 2008. **19**(5): p. 640-8.

Προτάσεις για την ορθή εκτίμηση συσχετίσεων (2)

Box 1. Some Research Practices that May Help Increase the Proportion of True Research Findings

- Large-scale collaborative research
- Adoption of replication culture
- Registration (of studies, protocols, analysis codes, datasets, raw data, and results)
- Sharing (of data, protocols, materials, software, and other tools)
- Reproducibility practices
- Containment of conflicted sponsors and authors
- More appropriate statistical methods
- Standardization of definitions and analyses
- More stringent thresholds for claiming discoveries or "successes"
- Improvement of study design standards
- Improvements in peer review, reporting, and dissemination of research
- Better training of scientific workforce in methods and statistical literacy

Ioannidis, J.P., **How to make more published research true.** PLoS Med, 2014. **11**(10): p. e1001747.

Προτάσεις για την ορθή εκτίμηση συσχετίσεων (3)

Table 1. Some major stakeholders in science and their extent of interest in research and its results from various perspectives; typical patterns are presented (exceptions do occur).

	Extent of interest in research results			
	Publishable	**Fundable**	**Translatable**	**Profitable**
Scientists	+++	+++	+	
Industry – sales and marketing				+++
Industry – R & D			+++	+++
Private investors, including hedge funds			++	+++
Public funders – open (e.g. NIH, NSF)	++		+	
Public funders – closed (e.g. military)			+++	
Not-for-profit funders/philanthropists	++		+++	
Journal editors	+++			+
For-profit publishers	+			+++
Professional and scientific societies	+			
Universities	+	+++		+
Not-for-profit research institutions	+++	+++	+	+
Supporting non-scientific staff		+++		
Hospitals and other professional facilities offering services related to science			+	+++
Other financial entities that are affected by these services (e.g. insurance)				+++
Governments and state/federal authorities				++
Consumers of products and services			+++	

doi:10.1371/journal.pmed.1001747.t001

Ioannidis, J.P., **How to make more published research true.** PLoS Med, 2014. **11**(10): p. e1001747.

Προτάσεις για την ορθή εκτίμηση συσχετίσεων (4)

Table 2. An illustration of different exchange rates for various currencies and wealth items in research.

	Different examples of reward systems		
	Current	**Change 1**	**Change 2**
CURRENCIES			
Publication (per unit)	Win 1	No value	No value
Replicated publication (per unit)	Win 1	Win 2	Win 2
Successfully translated publication (per unit)	Win 1	Win 5	Win 5
Refuted publication (per unit)	Win 1	Lose 1	Lose 1
Sharing data, protocols, analysis codes (per unit)	No value	Win 2	Win 2
Contribution to peer-review (per unit)	No value	Win 2	Win 2
Contribution to education/training (per unit)	No value	Win 1	Win 1
Grant funding (per one R01)	Win 5	Win 5	Lose 5
OTHER WEALTH ITEMS			
Assistant professor, title in good university	Win 3	Win 3	No value
Associate professor, title in good university	Win 10	Win 10	No value
Tenured professor, title in good university	Win 20	Win 20	No value
Team leader/director			
Per 1 doctoral student/post-doc	Win 2	Win 2	Lose 2
Administrative power, networking, lobbying	Win up to 200	No value	Lose up to 200

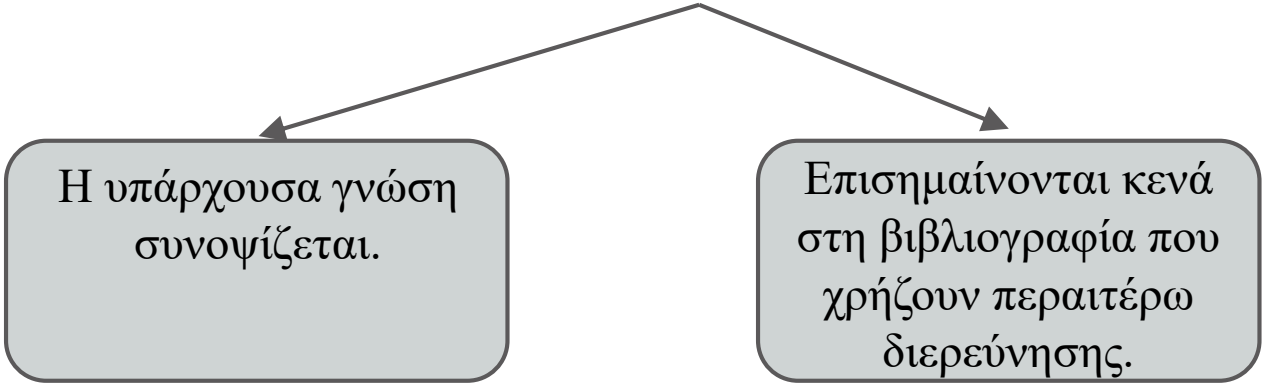
doi:10.1371/journal.pmed.1001747.t002

Ioannidis, J.P., **How to make more published research true.** PLoS Med, 2014. **11**(10): p. e1001747.

Συστηματική ανασκόπηση-Μετα-ανάλυση

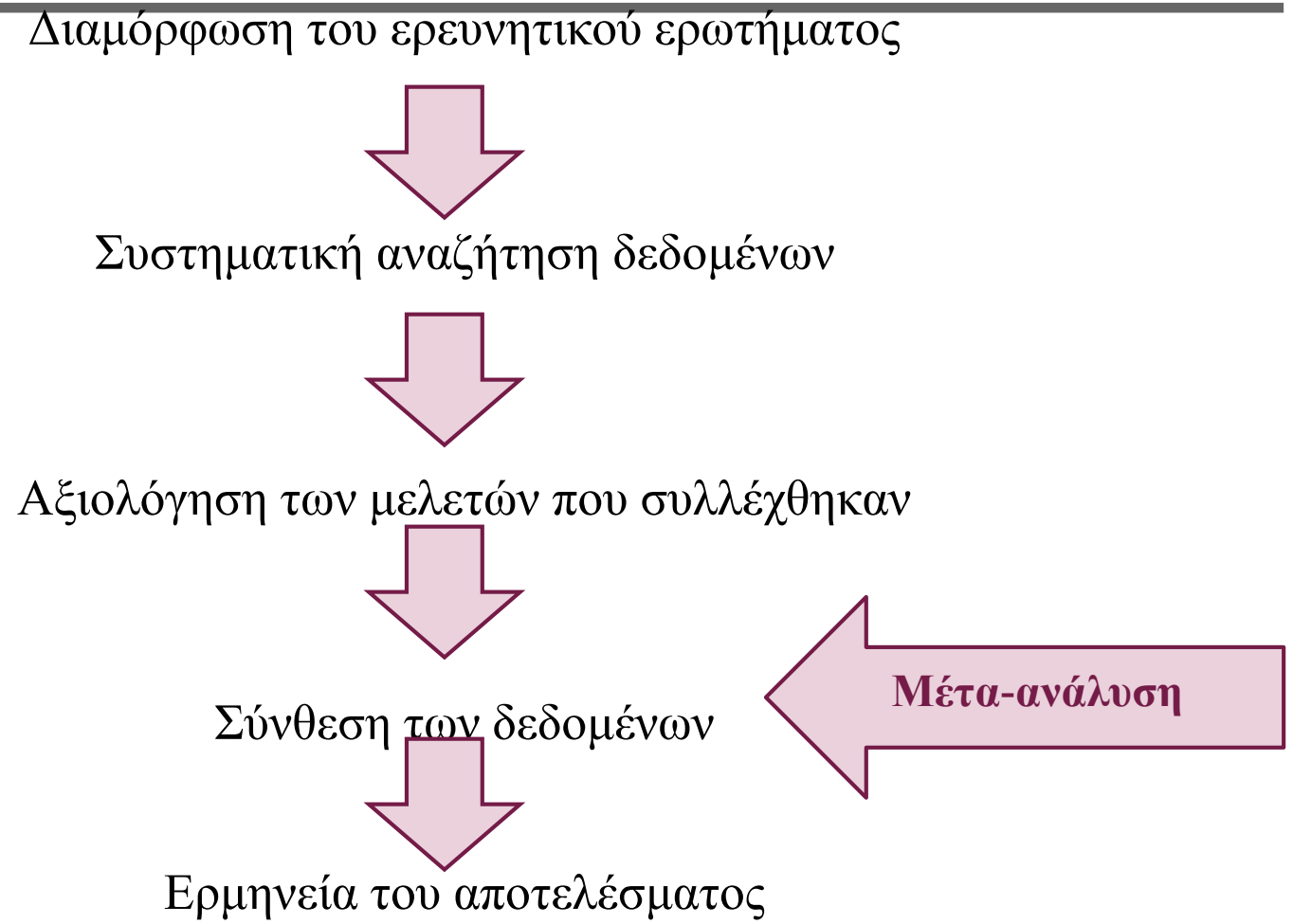
Συστηματική ανασκόπηση

Ο τεράστιος όγκος δεδομένων και πληροφοριών που έχει συσσωρευτεί καθιστά τα άρθρα **ανασκόπησης** (reviews) απαραίτητα:



Μια ανασκόπηση καλείται "**συστηματική**" εάν βασίζεται σε ένα καλά διατυπωμένο ερευνητικό ερώτημα, εντοπίζει σχετικές με το θέμα μελέτες, αξιολογεί την ποιότητά τους και συνοψίζει τα αποτελέσματα χρησιμοποιώντας σαφώς καθορισμένη μεθοδολογία.

Βήματα συστηματικής ανασκόπησης



Khan, K.S., et al., **Five steps to conducting a systematic review**. J R Soc Med, 2003. **96**(3): p. 118-21.

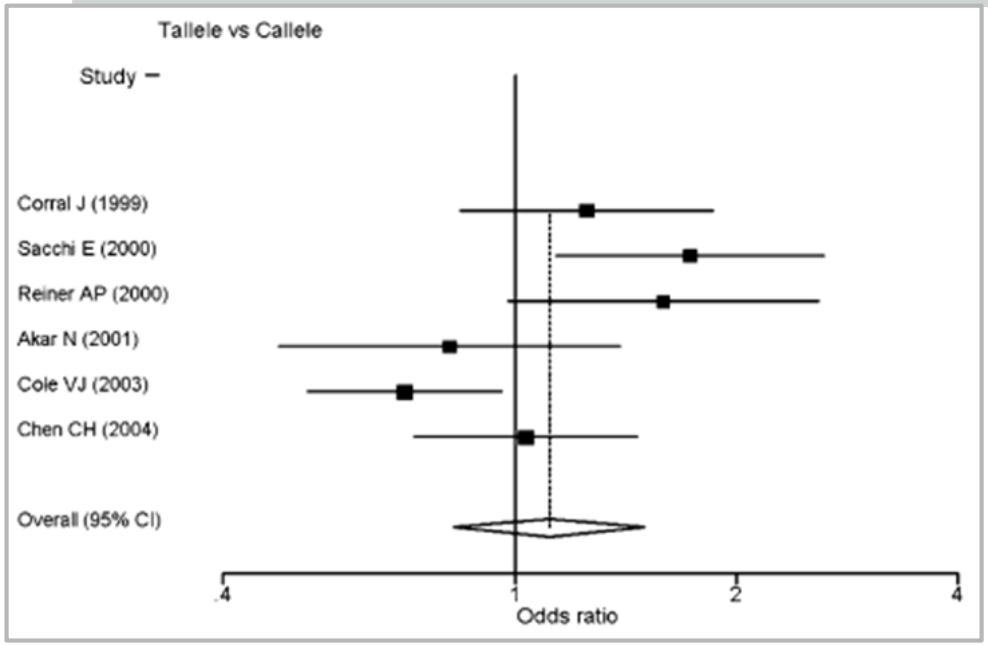
Μετα-ανάλυση

- Η **μετα-ανάλυση** μπορεί γενικώς να οριστεί ως η ποσοτική ανασκόπηση και σύνθεση των αποτελεσμάτων σχετιζομένων αλλά και ανεξάρτητων μελετών.
- Η μεθοδολογία τοποθετείται χρονολογικά στην εποχή του Fisher.
- Σαν όρος εμφανίστηκε για πρώτη φορά στην Ψυχολογία (Glass, 1976).
- Στην απλούστερη μορφή πρόκειται για ένα σταθμισμένο μέσο όρο των εκτιμήσεων των αποτελεσμάτων των επιμέρους μελετών.

$$Y_i = \log OR_i = \log\left(\frac{\alpha\delta}{\beta\gamma}\right), \quad s_i = \sqrt{\frac{1}{\alpha}+\frac{1}{\beta}+\frac{1}{\gamma}+\frac{1}{\delta}}$$

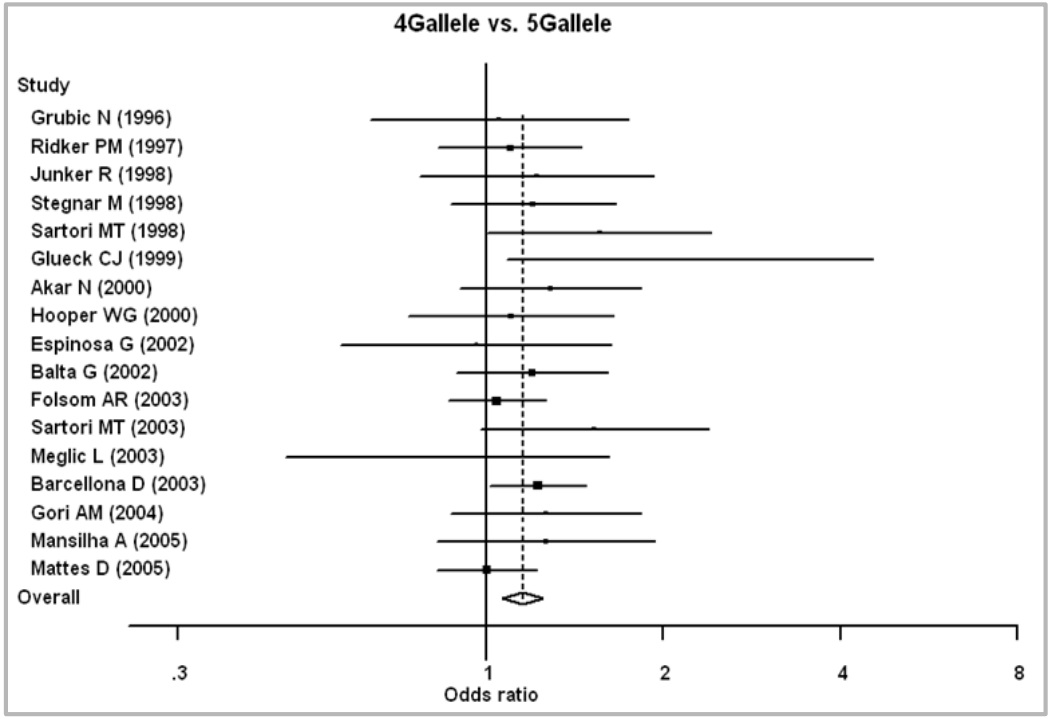
- Αυξάνει τη στατισ $\hat{\theta}_{\mathrm{MLE}} = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}$ with $W_i = \frac{1}{s_i^2}$ επιδράσεων.

Glass GV. **Primary, secondary, and meta-analysis of research**. Educational Researcher, 1976; 5: 3-8



Nikolopoulos GK *et al.* **Integrin, alpha 2 gene C807T Polymorphism and Risk of Ischemic Stroke: a Meta-Analysis.** 2007, *Thrombosis Research*; 119 **(4)**: 501-510

Tsantes AE *et al.* **Association between the Plasminogen Activator Inhibitor-1 4G/5G Polymorphism and Venous Thrombosis: a Meta-Analysis.** 2007, *Thrombosis and Haemostasis*; 97(6):907-13

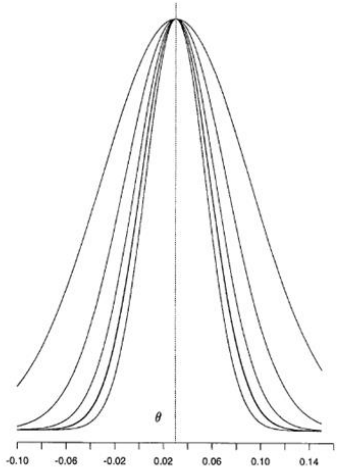


Στατιστικά μοντέλα

Μοντέλα σταθερών επιδράσεων:

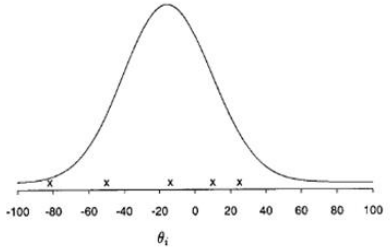
$$Y_i \overset{\text{indep.}}{\sim} \mathrm{N}(\theta, s_i^2) \quad \text{for } i = 1, 2, \ldots, k$$

$$\hat{\theta}_{\mathrm{MLE}} = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i} \quad \text{with } W_i = \frac{1}{s_i^2}$$



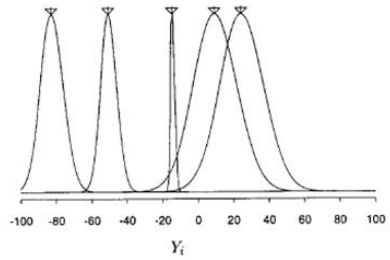
Μοντέλα τυχαίων επιδράσεων:

$$Y_i \mid \theta_i, s_i^2 \overset{\text{indep.}}{\sim} \mathrm{N}(\theta_i, s_i^2).$$



$$\theta_i \mid \theta, \tau^2 \overset{\text{indep.}}{\sim} \mathrm{N}(\theta, \tau^2).$$

$$\hat{\theta}(\tau)_{\mathrm{MLE}} = \frac{\sum_i W_i(\tau) Y_i}{\sum_k W_i(\tau)} \quad \text{with} \quad W_i(\tau) = \frac{1}{s_i^2 + \tau^2}.$$



Γιατί όλες οι μετα-αναλύσεις δεν οδηγούν στα ίδια αποτελέσματα; (1)

Example: **corticosteroids for acute bacterial meningitis**

- 1994: no question about benefits, but beware of harms;
- 1997: definite benefit only for some bacteria, limit to 2 days to avoid harm;
- 2003: definite benefit only for children, no increase in harm;
- 2003 correction: actually benefit is seen also in adults;
- 2007: benefit in high-income countries, but not in low-income countries;
- 2009: clear benefit, give it to all, this is it;
- 2010: no benefit at all.

1. steroids are modestly beneficial;
2. steroids are modestly harmful;
3. steroids are both modestly beneficial and modestly harmful.

Ioannidis, J.P., **Meta-research: The art of getting it wrong.** Res. Syn. Meth. 2010, 1 169-184

Γιατί όλες οι μετα-αναλύσεις δεν οδηγούν στα ίδια αποτελέσματα; (2)

Table 1. Some reasons for discrepant meta-analyses on the same topic.

- Different study questions
- Different data sources
- Different search strategies
- Different timing (evolution/accumulation of evidence)
- Different inclusion/exclusion criteria for study designs
- Different inclusion/exclusion criteria for outcomes
- Different inclusion/exclusion criteria for eligible populations
- Different inclusion/exclusion criteria for settings, co-interventions, other features
- Errors in the primary data
- Errors in the data extraction
- Different disambiguation and arbitration processes in study selection and data cleaning.
- Differential retrieval of unpublished data
- Different definitions of outcomes
- Differential criteria for performing quantitative synthesis
- Different models of statistical analysis
- Differential use of subgroup analyses, meta-regressions, and other exploratory analyses
- Differential selective reporting of meta-analysis results
- Differential use and interpretation of heterogeneity metrics
- Differential use and interpretation of bias tests
- Differential qualitative interpretation of results
- Inappropriate emphasis on seeming discrepancies that are not necessarily discrepancies (not beyond chance)

Ioannidis, J.P., **Meta-research: The art of getting it wrong.** Res. Syn. Meth. 2010, 1 169-184

Μεροληψίες στη μετα-ανάλυση (1)

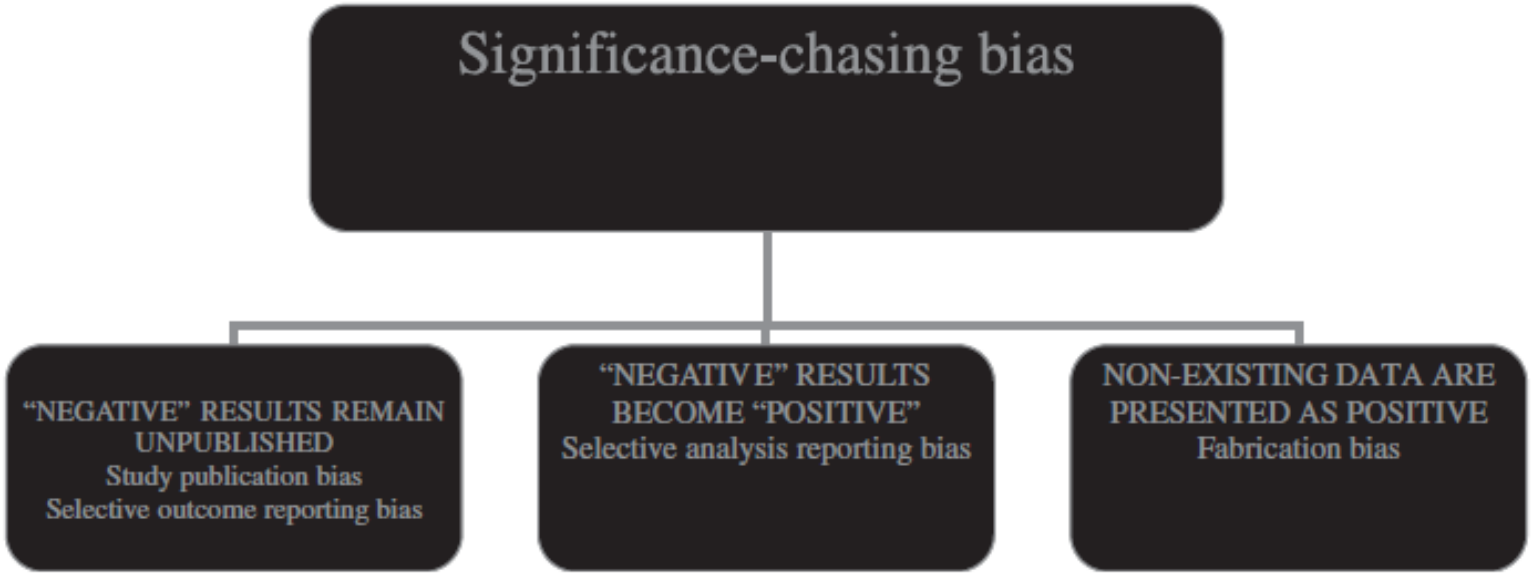


Figure 1. Significance-chasing biases.

Ioannidis, J.P., **Meta-research: The art of getting it wrong.** Res. Syn. Meth. 2010, 1 169-184

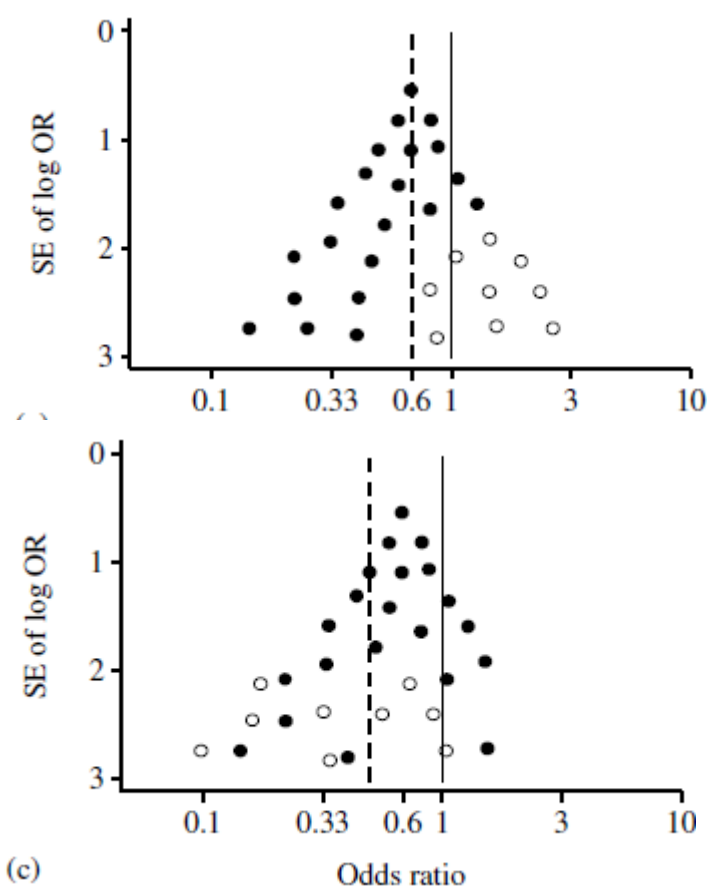
Μεροληψίες στη μετα-ανάλυση (2)

Table 4.1 Sources of bias affecting the search process.

Type of bias	Definition
Publication bias	Studies with statistically significant results are more likely to be published than those with statistically non-significant or null results.
Time-lag bias	Studies with statistically significant results are more likely to be stopped earlier than originally planned and published quicker.
Language bias	Studies with statistically significant results are more likely to be published in English.
Duplication bias	Studies with statistically significant results are more likely to be published more than once.
Citation bias	Studies with statistically significant results are more likely to be cited by others.

Hannah R. Rothstein, Alexander J. Sutton , Michael Borenstein, (2005) **Publication Bias in Meta-Analysis Prevention, Assessment and Adjustments (**ISBN: 978-0-470-87014-3**)**

Συστηματικό σφάλμα δημοσίευσης (Publication bias)



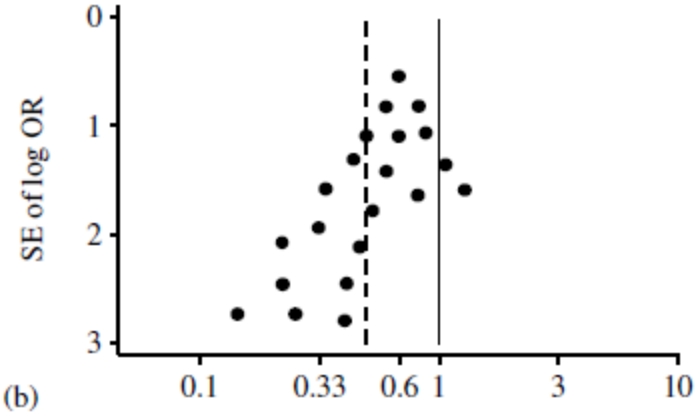


Figure 5.1 Hypothetical funnel plots: (a) symmetrical plot in the absence of bias (open circles indicate smaller studies showing no beneficial effects); (b) asymmetrical plot in the presence of publication bias (smaller studies showing no beneficial effects are missing); (c) asymmetrical plot in the presence of bias due to low methodological quality of smaller studies (open circles indicate small studies of inadequate quality whose results are biased towards larger beneficial effects). (Based on figure from Sterne *et al.*, 2001).

Hannah R. Rothstein, Alexander J. Sutton , Michael Borenstein, (2005) **Publication Bias in Meta-Analysis Prevention, Assessment and Adjustments (**ISBN: 978-0-470-87014-3**)**

Προτάσεις μείωσης της μεροληψίας μέσω προσεκτικής αναζήτησης της βιβλιογραφίας

- Searching the Cochrane Central Register of Controlled Trials and C2-SPECTR
- Searching electronic databases and the merits of handsearching
- Searching conference proceedings
- Contact with researchers
- Searching research registers
- Searching the Internet
- Taking account of differences in searching between the health and social sciences
- Including studies found in the grey literature

Hannah R. Rothstein, Alexander J. Sutton , Michael Borenstein, (2005) **Publication Bias in Meta-Analysis Prevention, Assessment and Adjustments (**ISBN: 978-0-470-87014-3**)**

Φαινόμενο «γκρίζας» βιβλιογραφίας (Grey literature bias)

No librarian who takes his job seriously can today deny that careful attention has also to be paid to the 'little literature' and the numerous publications not available in normal bookshops, if one hopes to avoid seriously damaging science by neglecting these.

(Minde-Pouet, 1920, cited by Schmidmaier, 1986)

Table 4.2 Citations for grey literature in 513 Cochrane reviews (a total of 6266 trials).

Grey literature source	Number of trial references[a]
Unpublished information	1259 (55 %)
Conference abstracts	805 (35 %)
Government reports	78 (4 %)
Company reports	66 (3%)
Theses/dissertations	63 (3 %)
Total grey citations (in 1446 trials)	2271 (100 %)

Source: Mallett *et al*. (2000).

[a] Trials can be referenced by more than one grey literature source; 4820/6266 trials were referenced only by published journal articles. Seventeen trials had missing citations.

Hannah R. Rothstein, Alexander J. Sutton , Michael Borenstein, (2005) **Publication Bias in Meta-Analysis Prevention, Assessment and Adjustments (**ISBN: 978-0-470-87014-3**)**

Φαινόμενο «ξένης» βιβλιογραφίας (Local literature bias) (1)

- Διερευνήθηκε εάν μελέτες γενετικής συσχέτισης που είχαν δημοσιευτεί στα Κινέζικα παρουσιάζουν επιλεκτικά αποτελέσματα ή υπόκεινται σε μεροληψίες γλώσσας δημοσίευσης πέραν της Αγγλικής.
- Επιλέχθηκαν 13 γονίδια που έχουν συσχετιστεί με κάποια ασθένεια και για τα οποία έχουν γίνει μετα-αναλύσεις στις οποίες περιλαμβάνονται τουλάχιστον 15 μη Κινέζικες μελέτες.
- Πραγματοποιήθηκε αναζήτηση στην Chinese Journal Full-Text Database για επιπλέον μελέτες.
- Βρέθηκαν 161 Κινέζικες μελέτες για 12 από τα γονίδια. Μόνο οι 20 από αυτές είχαν καταχωρηθεί στην Pubmed.
- Με μια μόνο εξαίρεση, η πρώτη Κινέζικη μελέτη εμφανίστηκε με μια χρονοκαθυστέρηση (2-21 χρόνια) μετά τη δημοσίευση της πρώτης μη Κινέζικης μελέτης.
- Οι Κινέζικες μελέτες παρουσίαζαν αποτελέσματα με μεγαλύτερη στατιστική σημαντικότητα έναντι των μη Κινέζικων, με το 48% αυτών να είναι από μόνες τους στατιστικά σημαντικές παρότι είχαν μικρότερο μέγεθος δείγματος (διάμεσος 146 έναντι 268).

Pan, Z., et al., **Local literature bias in genetic epidemiology: an empirical evaluation of the Chinese literature.** PLoS Med, 2005. **2**(12): p. e334.

Φαινόμενο «ξένης» βιβλιογραφίας (Local literature bias) (2)

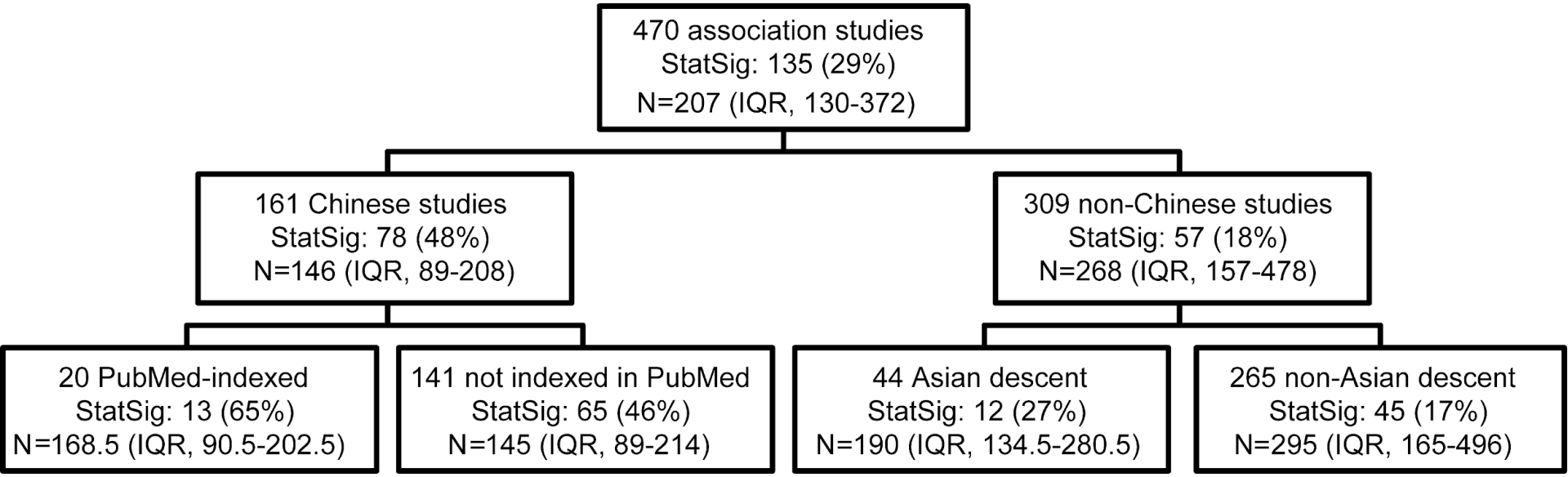


Figure 1. Categorization of the Examined Genetic Association Studies

Φαινόμενο «ξένης» βιβλιογραφίας (Local literature bias) (3)

Table 1. Eligible Meta-Analyses

ID	Disease/Outcome	Gene (Polymorphism)	Genetic Contrast	Studies (Total Sample)		Studies with $p < 0.05$ (%)		First Year Published	
				Chinese	Non-Chinese	Chinese	Non-Chinese	Chinese	Non-Chinese
1	Myocardial infarction	*ACE* (insertion/deletion)	DD versus DI + II	27 (4,514)	15 (18,664)	17 (63)	6 (40)	1996	1992
2	Ischemic heart disease	*ACE* (insertion/deletion)	DD versus DI + II	35 (6,586)	17 (21,876)	20 (57)	4 (23)	1998	1994
3	Cancer	*HRAS/HRAS1* (rare alleles)	Rare versus common alleles	23 (1,559)	24 (8,542)	7 (30)	8 (33)	1994	1985
4	Bladder cancer	*NAT2* (slow acetylation alleles)	Slow/slow versus others	3 (417)	20 (5,836)	1 (33)	6 (30)	2000	1979
5	Diabetic nephropathy	*ACE* (insertion/deletion)	II versus ID + DD	25 (3,857)	20 (5,393)	8 (32)	6 (30)	1997	1994
6	Coronary artery disease	*ITGB3* (L33P)	A2A2 versus A1A2 + A1A1	1 (152)	31 (17,315)	0 (0)	4 (13)	1999	1996
7	Bladder cancer	*GSTM1* (gene deletion)	Null/null versus others	2 (400)	20 (5,795)	1 (50)	7 (35)	2002	1992
8	SLE nephritis	*FCGR2A* (R131H)	RR versus RH + HH	1 (86)	24 (2,801)	0 (0)	2 (8)	2003	1995
9	SLE	*FCGR2A* (R131H)	RR versus RH + HH	2 (261)	21 (4,708)	2 (100)	4 (19)	2000	1995
10	Coronary heart disease	*MTHFR* (677C/T)	TT versus CC	14 (1,778)	40 (23,922)	8 (57)	7 (18)	1998	1996
11	Schizophrenia	*DRD3* (Bal1)	Ser/Ser + Gly/Gly versus Ser/Gly	4 (1,527)	39 (8,556)	0 (0)	1 (2)	1993	1993
12	Lung cancer	*GSTM1* (gene deletion)	Null/null versus others	24 (5,909)	38 (16,119)	14 (58)	3 (6)	1997	1991

ACE, angiotensin converting enzyme; *DRD2/DRD3*, dopamine receptor D2/D3; *FCGR2A*, low-affinity receptor of the Fc domain of immunoglobulin G; *GSTM1*, glutathione S-transferase M1; *HRAS*, Harvey rat sarcoma viral oncogene homolog; *ITGB3*, platelet glycoprotein receptor IIIa; *MTHFR*, methylenetetrahydrofolate reductase; SLE, systemic lupus erythematosus.

DOI: 10.1371/journal.pmed.0020334.t001

Φαινόμενο «ξένης» βιβλιογραφίας (Local literature bias) (4)

Table 2. Genetic Effects in Chinese and Non-Chinese Studies

ID	Random Effects Odds Ratio (95% CI)		Discrepancy in Effect (*p*-Value)	I^2 for Heterogeneity (%)	
	Chinese	Non-Chinese		Chinese	Non-Chinese
1	2.21 (1.84–2.66)[a]	1.28 (1.09–1.50)[a]	−4.42 (<0.01)	35	65
2	2.02 (1.66–2.47)[a]	1.20 (1.06–1.36)[a]	−4.36 (<0.01)	60	54
3	7.66 (4.51–13.0)	1.84 (1.54–2.21)	−4.99 (<0.01)	0	7
4	3.06 (0.95–9.87)[a]	1.43 (1.20–1.71)[a]	−1.26 (0.21)	81	48
5	0.53 (0.41–0.69)	0.68 (0.55–0.84)[a]	1.43 (0.15)	66	50
6	0.83 (0.02–44.4)	1.10 (0.99–1.21)[a]	0.14 (0.89)	NP	45
7	1.58 (1.02–2.43)	1.44 (1.25–1.67)	−0.38 (0.70)	0	27
8	0.83 (0.34–2.03)	1.11 (0.88–1.41)	0.63 (0.53)	NP	26
9	2.87 (1.50–5.50)	1.29 (1.10–1.52)	−2.33 (0.02)	0	19
10	2.03 (1.44–2.86)[a]	1.14 (1.01–1.30)[a]	−3.09 (<0.01)	59	50
11	0.95 (0.76–1.17)	1.11 (1.01–1.22)	1.33 (0.18)	0	8
12	1.63 (1.38–1.92)[a]	1.14 (1.07–1.22)	−3.54 (<0.01)	52	0

The ID numbers correspond to Table 1. The discrepancy between the Chinese and non-Chinese studies is expressed as a *z*-score and the corresponding *p*-value.

[a]Significant between-study heterogeneity ($p < 0.10$ for the *Q* statistic)

CI, confidence interval; NP, not pertinent (only one study available).

DOI: 10.1371/journal.pmed.0020334.t002

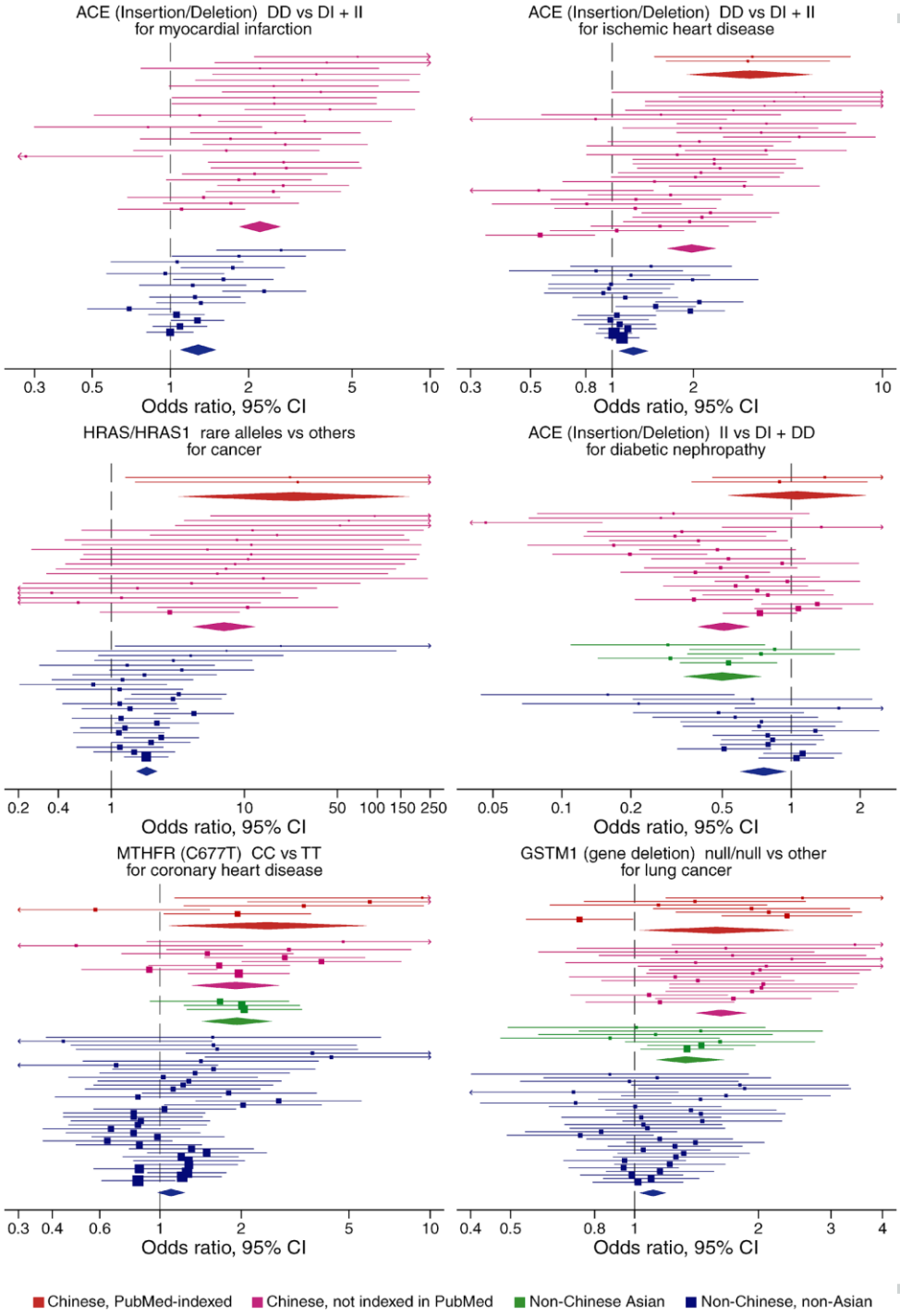
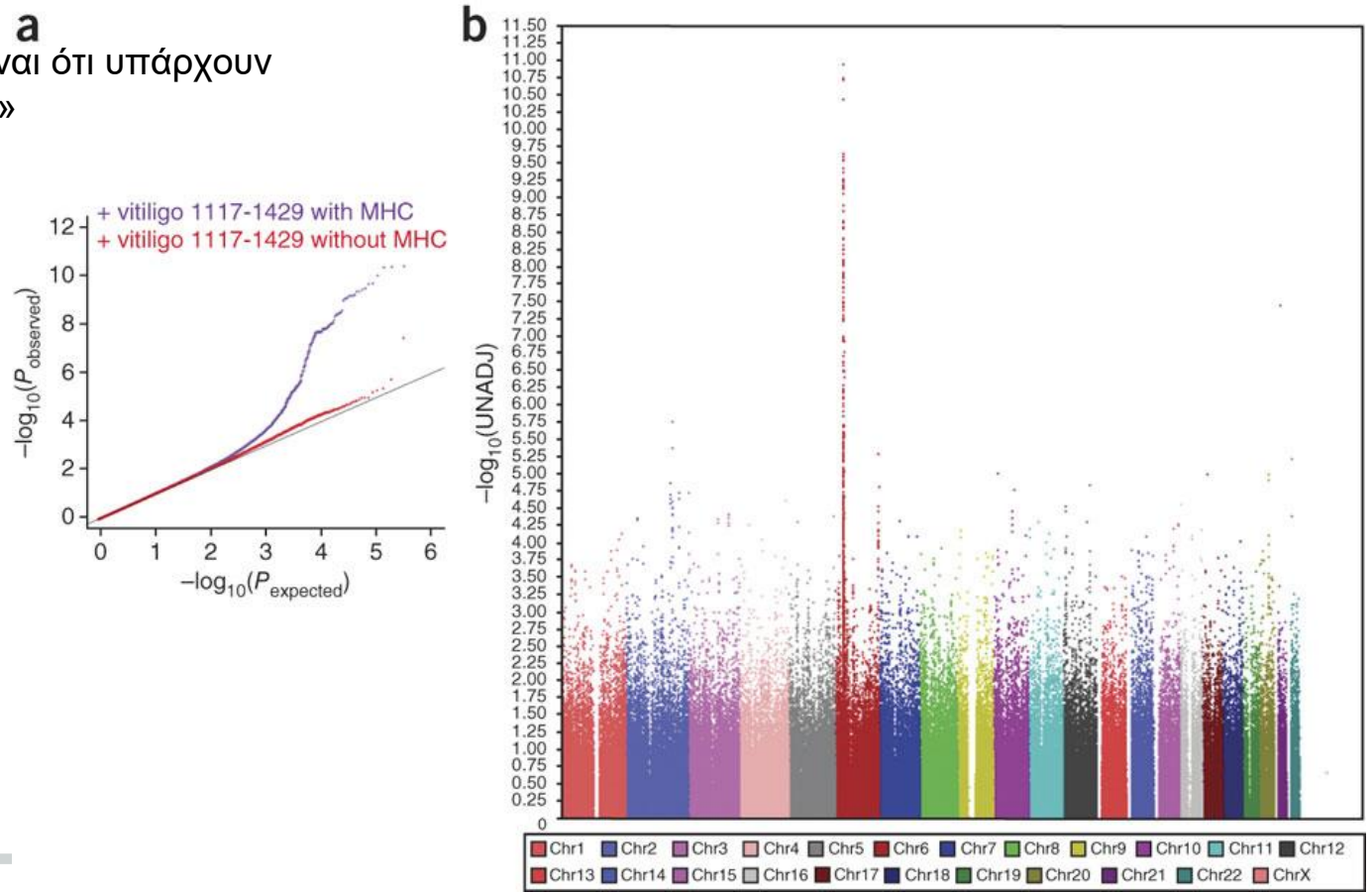


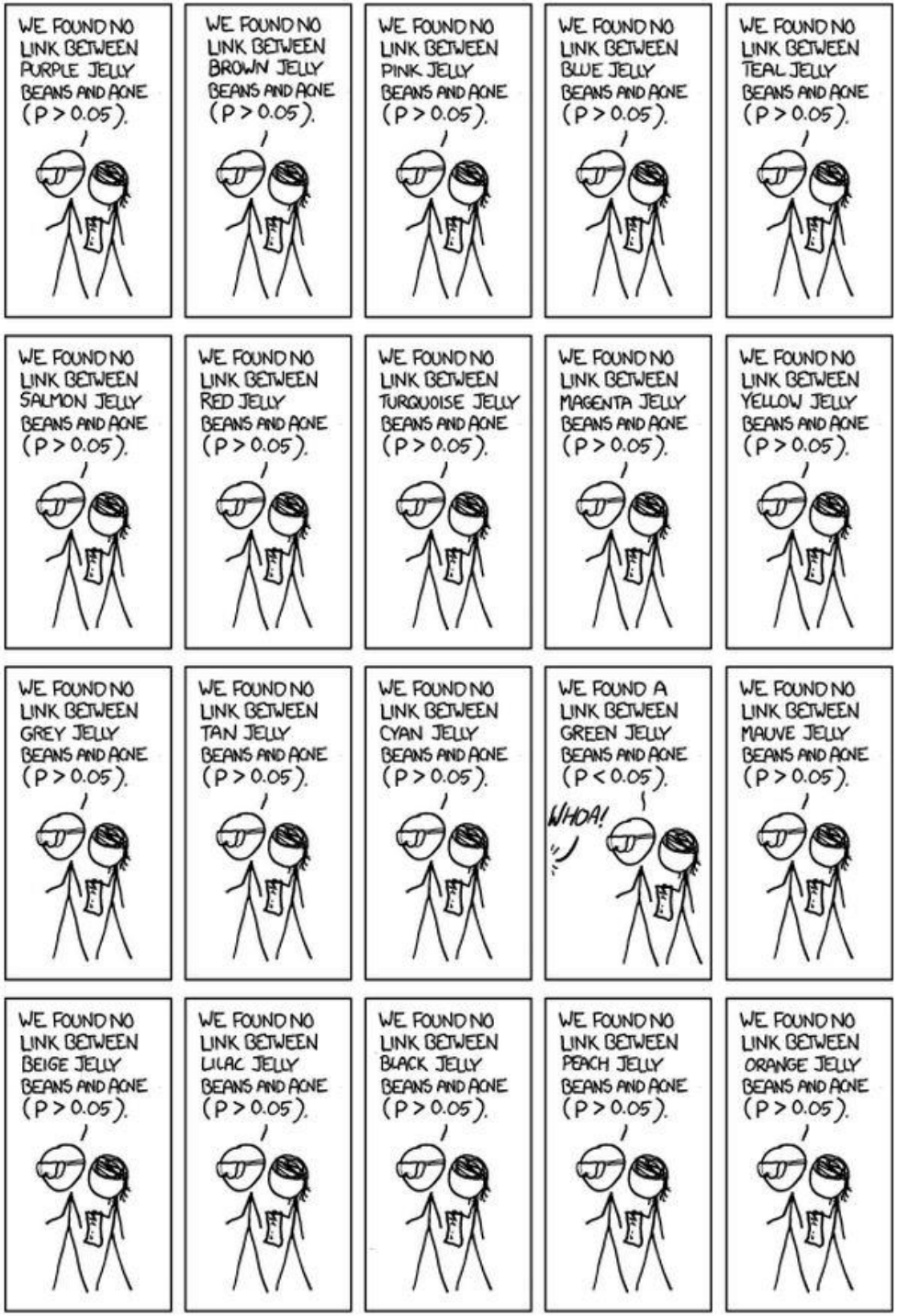
Figure 2. Meta-Analyses of Gene-Disease Associations in a Large Number of Both Non-Chinese and Chinese Studies Each study is shown by its odds ratio and 95% confidence intervals (CIs). The box of the point estimate is proportional to the study weight. Also shown are summary estimates by random effects calculations (diamonds). Summary estimates are obtained separately for Chinese studies indexed in PubMed (red), Chinese studies not indexed in PubMed (pink), non-Chinese studies of Asian descent populations (green), and studies of persons of non-Asian descent (blue). An odds ratio of 1 means no genetic effect, odds ratios larger than 1 mean genetic predisposition, and odds ratios less than 1 mean genetic protection.

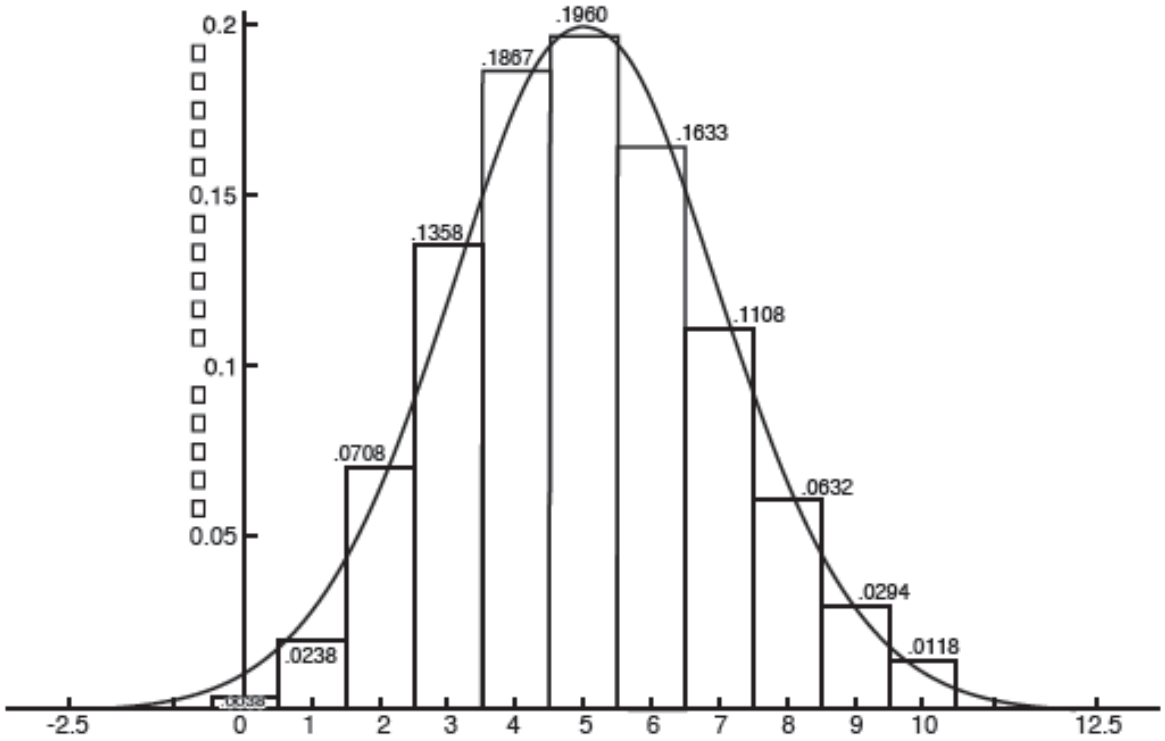
Πολλαπλές συγκρίσεις

Φαινόμενο ιδιαίτερα έντονο στη μοριακή εποχή
-Genomewide association studies
-gene expression studies
-biological sequence comparisons

Κοινό σημείο σε όλα, είναι ότι υπάρχουν πολλοί πιθανοί «στόχοι»




WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).



Figuring the odds—How probability is Calculated

BY ARTHUR BENJAMIN

Probability is a more precise mathamatical tool for determining if results are random.
The exact probability that a person guesses the correct ESP symbol exactly x times out of 25 is:

$$Px = \binom{25}{x} (.2)^x (.8)^{25-x}$$

where $\binom{25}{x} = \frac{25!}{x!(25-x)!}$

is the number of ways to choose x different numbers between 1 and 25 (order not important).
A very good approximation can be obtained by approximating the histogram by a normal distribution with a mean 25(.2) + 5 and standard deviation $\sqrt{25(.2)(.8)} = 2$. The probability of guessing 3 to 7 correct is approximately the area under the normal curve betweeen 2.5 and 7.5 … about 79%.

Probability predicts these test results for a test of 25 questions with five possible answers if chance is operating:

Most people (79%) will get between 3 and 7 correct (probability is a more precise calculation).

The probability of guessing 8 or more correctly is 10.9% (in a group of 25, you can always expect several scores in this range purely by chance.)

The chances of getting 15 correct is about 1 in 90,000.

Guessing 20 out of 25 has a probability of about 1 in 5 billion.

Guessing all 25 correct has a chance of (.2) = 3.3 x 10, or about 1 in 300 quadrillion! (A wager against such an unusual occurrence would be a safe bet.)

Παράδειγμα από τις μικροσυστοιχίες

Παράδειγμα: Ας υποθέσουμε ότι εξετάζονται 10000 γονίδια τότε με p-value<0.05, 500 γονίδια αναμένεται να βρεθούν στατιστικά σημαντικά κατά τύχη (by chance)

Ανάγκη χρησιμοποίησης των μεθόδων διόρθωσης για πολλαπλές συγκρίσεις

Bonferroni: $p_{cor(i)} = p_{(i)} * n$

Sidak: $p_{cor(i)} = 1-(1- p_{(i)})^{\frac{1}{n}}$

Holm: $p_{cor(i)} = (n-i) * p_{(i)}$

Holland: $p_{cor(i)} = (n-i+1) * p_{(i)}$

FDR: $p_{cor(i)} = \frac{n}{n-i} * p_{(i)}$

Φαινόμενο του «Πρωτέα» (Proteus phenomenon-Molecular bias) (1)

- Συχνά παρατηρείται το φαινόμενο πρώιμες δημοσιεύσεις που αφορούν τη συσχέτιση ενός γενετικού παράγοντα με μια ασθένεια ή την επίδραση ενός φαρμάκου σε μια κλινική δοκιμή να δείχνουν μια ισχυρή θετική ή αρνητική συσχέτιση, πυροδοτώντας την επακόλουθη αύξηση μελετών στο ίδιο θέμα προκειμένου να επαληθευτούν και να αναπαραχθούν παρόμοια αποτελέσματα.

- Ορισμένες φορές όμως, η πρώιμη ισχυρή συσχέτιση φαίνεται να μην ισχύει (οπότε η μελέτη χαρακτηρίζεται από σφάλμα τύπου Ι) και συνεπώς, το συνολικό αποτέλεσμα της μέτα-ανάλυσης αλλάζει με το χρόνο, καταδεικνύοντας την πραγματική συσχέτιση του παράγοντα με την πιθανότητα εμφάνισης της νόσου. Το φαινόμενο αυτό καλείται φαινόμενο του Πρωτέα (‘Proteus phenomenon’) και εμφανίζεται συχνά στη Μοριακή Επιδημιολογία.

Φαινόμενο του «Πρωτέα» (Proteus phenomenon-Molecular bias) (2)

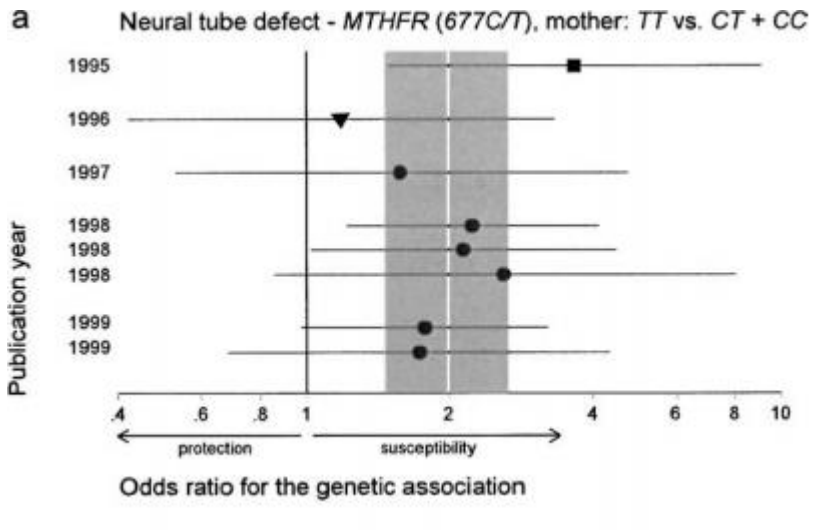


Fig. 1. Extreme differences in the results of a first study vs. a study published shortly thereafter. In both presented examples, studies are ordered chronologically and their results are shown by the odds ratio and 95% confidence intervals. All studies published in the same calendar year are packed together, unless one was clearly the first study. The study with the most favorable-ever results for the presence of an association is shown by a square, and the study with the least favorable-ever results is shown by a triangle, while all other studies are shown by circles. The white line corresponds to the summary odds ratio and the shaded area shows the 95% confidence interval. Also shown is the vertical line of no association (odds ratio=1). (a) The first published study on the relationship between the methylenetetrahydrofolate reductase (MTHFR) TT genotype in the mother and the risk of neural tube defects in the child found a very strong, statistically significant association (odds ratio 3.67, 95% confidence interval 1.47– 9.07) and was published in The Lancet. The following year, data reported in the same journal showed only a minor nonsignificant trend. Subsequent studies provided intermediate results between these two extremes.

Ioannidis, J.P. and T.A. Trikalinos, **Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials.** J Clin Epidemiol, 2005. 58(6): p. 543-9.

Φαινόμενο του «Πρωτέα» (Proteus phenomenon-Molecular bias) (3)

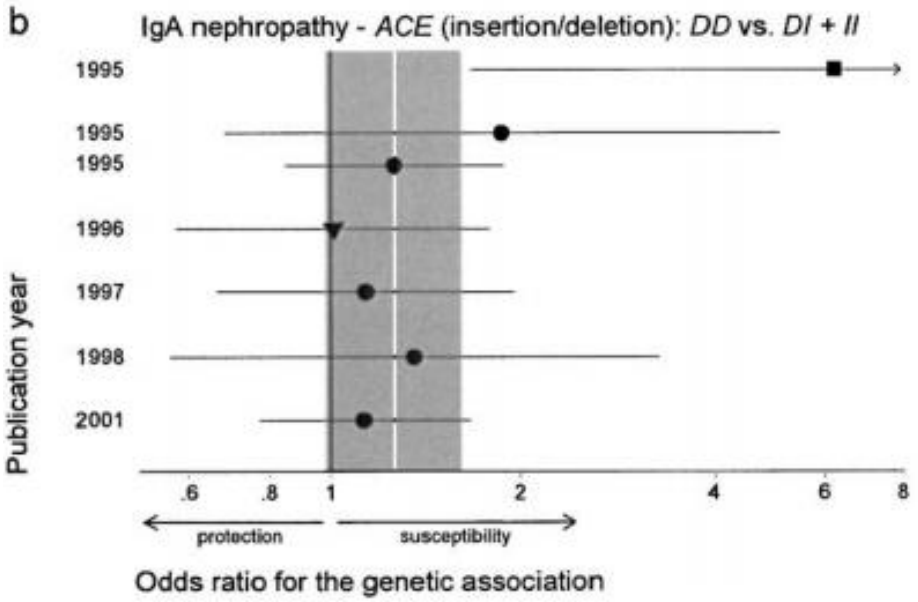


Fig. 1.(b) The first conducted study on the relationship between the angiotensin converting enzyme (ACE) DD genotype and IgA nephropathy showed a highly statistically significant association and was published in the Journal of Clinical Investigation. Two other studies published in the same year in nephrology journals found no significant association; a study published the following year found no association at all and between-study variance was maximized. Subsequent studies had intermediate results. The overall data are still inconclusive, but exclude the effect observed in the first study.

Ioannidis, J.P. and T.A. Trikalinos, **Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials.** J Clin Epidemiol, 2005. 58(6): p. 543-9.

Φαινόμενο του «Πρωτέα» (Proteus phenomenon-Molecular bias) (4)

- Πραγματοποιήθηκε μετα-ανάλυση 370 μελετών οι οποίες αξιολογούν 36 συσχετίσεις γονιδίων με διάφορες ασθένειες.
- Η ετερογένεια μεταξύ των μελετών ήταν συχνή ενώ τα αποτελέσματα της πρώτης μελέτης συσχετίζονται ελαφρώς με εκείνα των επόμενων.
- Τόσο η μεροληψία όσο και η πληθυσμιακή ποικιλομορφία μπορούν να εξηγήσουν γιατί οι πρώιμες μελέτες τείνουν να εκτιμούν στατιστικά σημαντικές συσχετίσεις γονιδίων με ασθένειες.

Ioannidis, J.P., *et al.*, **Replication validity of genetic association studies**. Nat Genet, 2001. 29(3): p. 306-9.

Φαινόμενο του «Πρωτέα» (Proteus phenomenon-Molecular bias) (5)

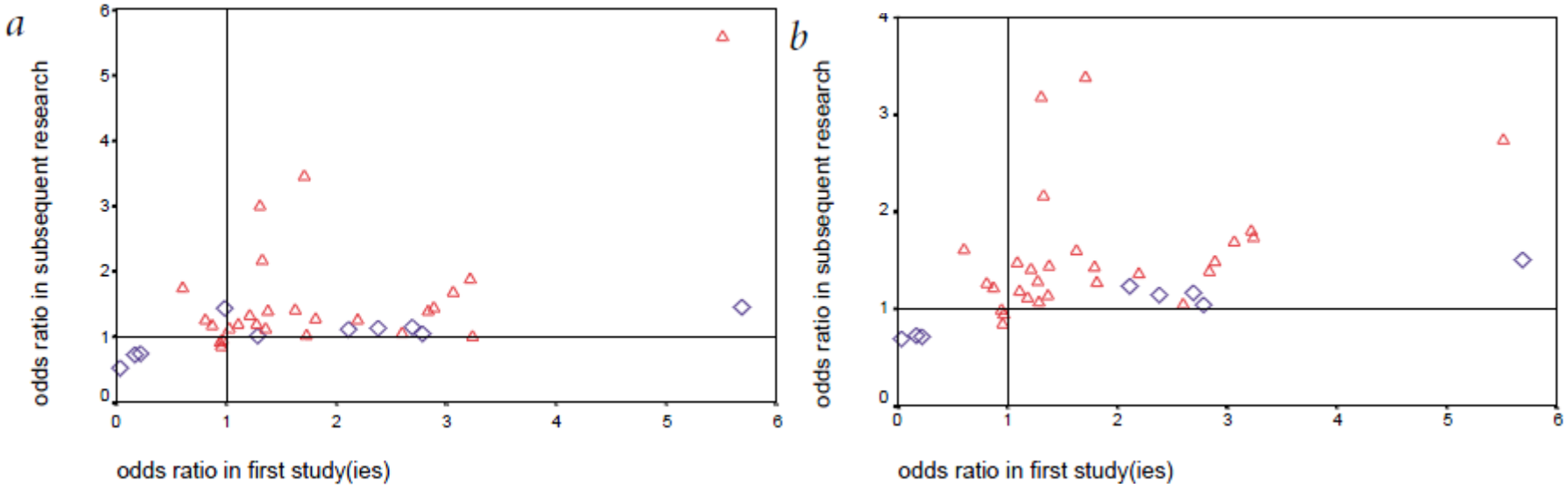


Fig. 1 Correlation between the odds ratio (OR) in the first study/studies and in subsequent research. OR>1 suggests predisposition towards the disease, whereas OR<1 suggests protection from the disease. Blue diamonds denote statistically significant discrepancies beyond chance between first and subsequent studies (a, fixed effects; b, random effects).

Ioannidis, J.P., *et al.*, **Replication validity of genetic association studies**. Nat Genet, 2001. 29(3): p. 306-9.

Φαινόμενο του «Πρωτέα» (Proteus phenomenon-Molecular bias) (6)

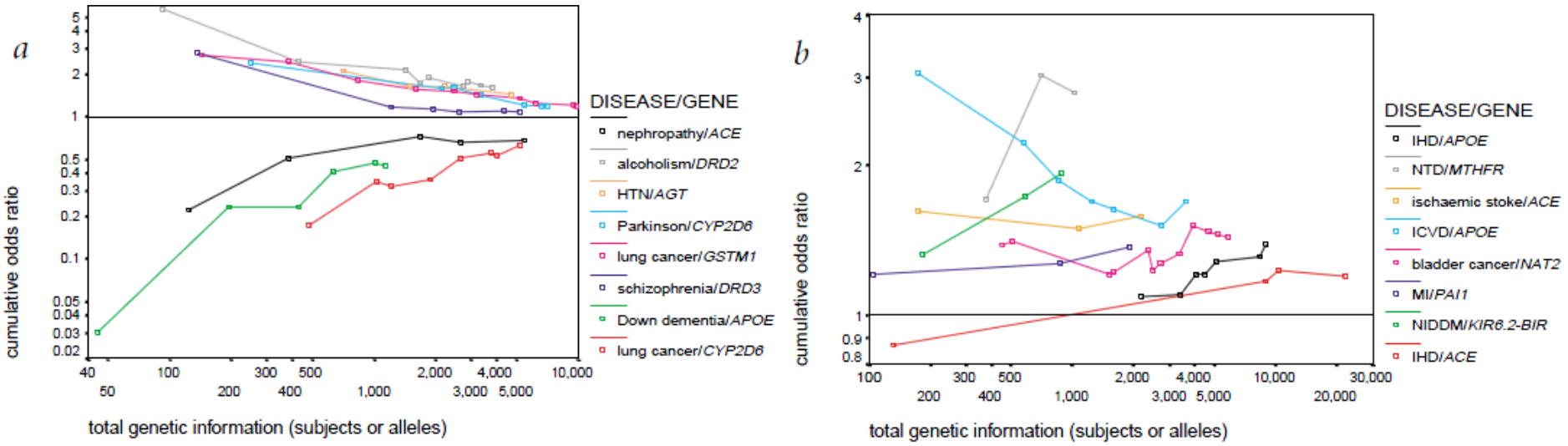


Fig. 2 Evolution of the strength of an association as more information is accumulated. The strength of the association is shown as an estimate of the odds ratio (OR) without confidence intervals. a, Eight topics in which the results of the first study or studies differed beyond chance ($P<0.05$) when compared with the results of the subsequent studies. b, Eight topics in which the first study or studies did not claim formal statistical significance for the genetic association but formal significance was reached by the end of the meta-analysis. Each trajectory starts at the OR of the first study or studies. Updated cumulative OR estimates are obtained at the end of each subsequent year, summarizing all information to that time (random effects). The horizontal axis (total genetic information) shows the total number of subjects genotyped with one of the contrasted genotypes, or the total number of typed alleles when specificallele frequencies are compared between disease cases and controls.

Ioannidis, J.P., *et al.*, **Replication validity of genetic association studies**. Nat Genet, 2001. 29(3): p. 306-9.

Φαινόμενο του «Πρωτέα» (Proteus phenomenon-Molecular bias) (7)

- Πραγματοποιήθηκαν 55 αθροιστικές μετα-αναλύσεις μελετών γενετικής συσχέτισης (με 579 μελέτες) για να διερευνηθεί εάν τα πρώιμα στατιστικά σημαντικά αποτελέσματα μπορούν να αποτελέσουν προγνωστικό παράγοντα της εγκαθίδρυσης της συσχέτισης ενός γονιδίου με μια ασθένεια.
- Σε 35 μετα-αναλύσεις η πρώτη μελέτη είχε στατιστικά σημαντικά αποτελέσματα ενώ για αυτές ο ρυθμός δημοσίευσης των μελετών που ακολούθησαν αυξήθηκε κατά 1,71 φορές.
- Από τη σύγκριση των αποτελεσμάτων της πρώτης μελέτης έναντι των υπολοίπων υπολογίστηκε ευαισθησία και ειδικότητα 0.65 και 0.38 αντίστοιχα.

Trikalinos, T.A., et al., **Establishment of genetic associations for complex diseases is independent of early study findings.** Eur J Hum Genet, 2004. **12**(9): p. 762-9.

Φαινόμενο του «Πρωτέα» (Proteus phenomenon-Molecular bias) (8)

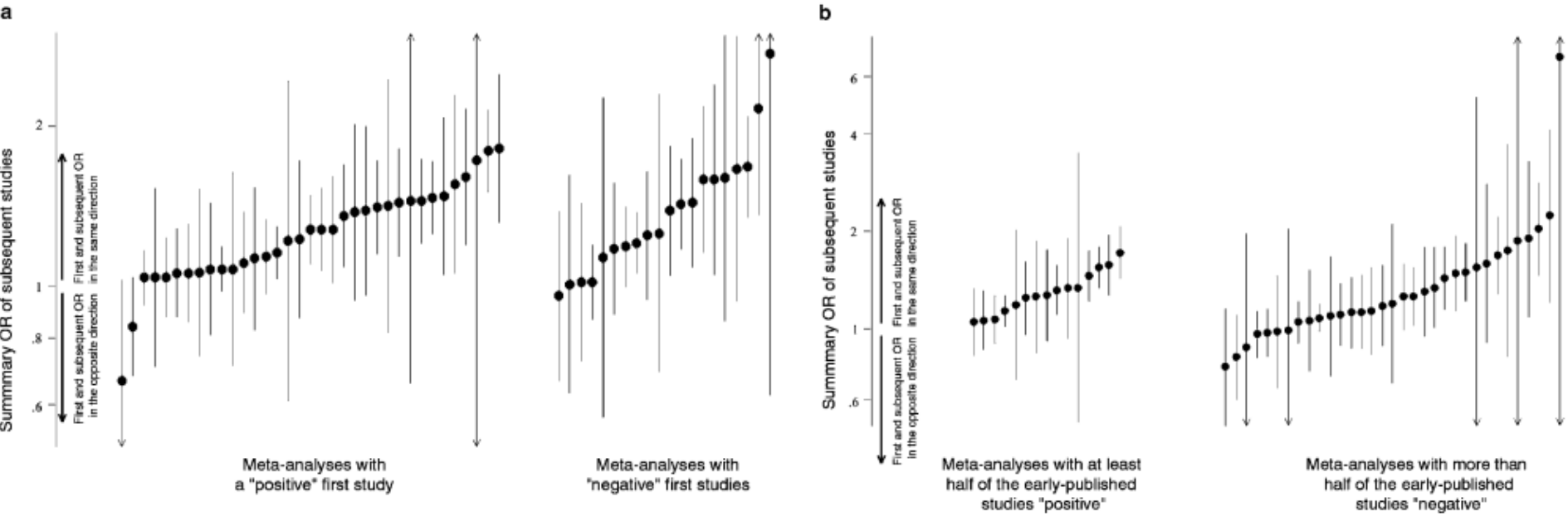


Figure 1 Summary ORs (dots) and corresponding 95% CIs (vertical bars) for research published after the first studies (a) or after at least 3 early-published studies (b) for each one of the eligible meta-analyses (55 for (a) and 48 for (b)). Arrowheads imply that the upper or lower boundary of the 95% CI extend beyond the edges of the graph. Meta-analyses with ORs greater than 1.00 are showing effects in the direction proposed by the first studies, or the synthesis of at least three early-published studies. Meta-analyses with ORs less than 1.00 are showing effects in a direction opposite to that of the first studies, or the synthesis of at least three early-published studies. Ordering is by ascending OR values. Inclusion of the first or early-published studies in the summary OR calculations yielded largely similar results (not shown). OR: odds ratio.

Trikalinos, T.A., et al., **Establishment of genetic associations for complex diseases is independent of early study findings.** Eur J Hum Genet, 2004. **12**(9): p. 762-9.

Φαινόμενο του «Πρωτέα» (Proteus phenomenon-Molecular bias) (9)

Table 3 Diagnostic performances of first and early-published studies against the statistical significance of the meta-analysis

Assessment	*Sensitivity (95% CI)*	*Specificity (95% CI)*	*LR+*	*LR−*
First or early-published studies excluded from the meta-analysis				
First study with $P<0.05$ (35/55)	0.65 (0.43, 0.84)	0.38 (0.21, 0.56)	1.1	0.92
At least half of early-published studies with $P<0.05$ (15/48)	0.40 (0.16, 0.68)	0.73 (0.54, 0.87)	1.5	0.82
Very low *P*-values in early-published studies[a] (9/48)	0.20 (0.04, 0.48)	0.82 (0.65, 0.93)	1.1	0.98
Attributable fraction in first study $\geqslant$2% based on 95% CI coverage (19/55)	0.39 (0.20, 0.61)	0.69 (0.50, 0.84)	1.3	0.88
All studies included in the meta-analysis				
First study with $P<0.05$ (35/55)	0.67 (0.46, 0.83)	0.39 (0.22, 0.59)	1.1	0.85
At least half of early-published studies with $P<0.05$ (15/48)	0.52 (0.31, 0.72)	0.91 (0.72, 0.99)	5.8	0.53
Very low *P*-values in early-published studies[a] (9/48)	0.47 (0.23, 0.72)	0.96 (0.78, 1.00)	11	0.55
Attributable fraction in first study $\geqslant$2% based on 95% CI coverage (19/55)	0.44 (0.25, 0.65)	0.75 (0.55, 0.89)	1.8	0.75

[a]Refers to the presence of $\geqslant$2 studies with *P*-values <0.01, or the presence of $\geqslant$2 statistically significant studies, one of which has $P<0.001$ among the early-published studies.
CI: confidence interval; LR+: positive likelihood ratio; LR−: negative likelihood ratio.
CIs were derived using exact methods.

Trikalinos, T.A., et al., **Establishment of genetic associations for complex diseases is independent of early study findings.** Eur J Hum Genet, 2004. **12**(9): p. 762-9.

Επαναληψιμότητα (1)

Table 1. Examples of Some Reported Reproducibility Concerns in Preclinical Studies

Author	Field	Reported Concerns
Ioannidis et al (2009)[22]	Microarray data	16/18 studies unable to be reproduced in principle from raw data
Baggerly et al (2009)[23]	Microarray data	Multiple; insufficient data/poor documentation
Sena et al (2010)[24]	Stroke animal studies	Overt publication bias: only 2% of the studies were negative
Prinz (2011)[1]	General biology	75% to 80% of 67 studies were not reproduced
Begley & Ellis (2012)[2]	Oncology	90% of 53 studies were not reproduced
Nekrutenko & Taylor(2012)[25]	NGS data access	26/50 no access to primary data sets/software
Perrin (2014)[26]	Mouse, in-vivo	0/100 reported treatments repeated positive in studies of ALS
Tsilidis et al (2013)[27]	Neurological studies	Too many significant results, overt selective reporting bias
Lazic & Essioux (2013)[28]	Mouse VPA model	Only 3/34 used correct experimental measure
Haibe-Kains et al (2013)[29]	Genomics/cell line analysis	Direct comparison of 15 drugs and 471 cell lines from 2 groups revealed little/no concordant data
Witwer (2013)[30]	Microarray data	93/127 articles were not MIAME compliant
Elliott et al (2006)[31]	Commercial antibodies	Commercial antibodies detect wrong antigens
Prassas et al (2013)[32]	Commercial ELISA	ELISA Kit identified wrong antigen
Stodden et al (2013)[33]	Journals	Computational biology: 105/170 journals noncompliant with National Academies recommendations
Baker et al (2014)[34]	Journals	Top tier fail to comply with agreed standards for animal studies
Vaux (2012)[35]	Journals	Failure to comply with their own statistical guidelines

ALS indicates amyotrophic lateral sclerosis; MIAME, minimum information about a microarray experiment; NGS, next generation sequencing; and VPA, valproic acid (model of autism).

Begley, C.G. and J.P. Ioannidis, **Reproducibility in science: improving the standard for basic and preclinical research.** Circ Res, 2015. **116**(1): p. 116-26.

Επαναληψιμότητα (2)

Table 2. Additional Basic Science Fields Where Concerns Regarding Reproducibility Have Been Raised

Discipline	Issues Raised	Author
Neuroscience	Low statistical power; small sample size	Button et al (2013)[36]
Pharmacology	Lack of training, lack of statistical power, blinding, hypothesis, requisite PK studies, randomization, dose-response, controls, prospective plan, validation, independent replication, and selection of doses that are not tolerable in humans	Henderson et al(2013)[37]; Kenakin et al (2014)[38]; McGonigie et al (2014)[39]; Winquist et al (2014)[40]; Marino (2014)[41]
Genomics/bioinformatics	Irreproducibility of high-profile studies	Sugden et al (2013)[42]
Stem cell biology	Lack of reliable, quality data	Plant & Parker (2013)[43]
Oncology, in vitro testing	Use of clinically unachievable concentrations	Smith & Houghton(2013)[44]
Chemistry lead-discovery	Artifacts; false positives and negatives	Davis & Erlanson (2013)[45]
Computational biology	10 common errors	Sandve et al (2013)[46]
Pathology/Biomarkers	Biospecimen quality	Simeon-Dubach et al (2012)[47]
Organizational psychology	Suppression of negative studies	Kepes & McDonald (2103)[48]
Observational research	0/52 hypotheses confirmed in randomized Trials	Young & Karr (2011)[11]

Begley, C.G. and J.P. Ioannidis, **Reproducibility in science: improving the standard for basic and preclinical research.** Circ Res, 2015. **116**(1): p. 116-26.

Table 3. Some Proposals to Improve Experimental Rigor and Quality in Preclinical Research

Proposal	Author
Editors solicit replication bids	Wagenmakers and Forstman (2014)[74]
Plea to improve editorial standards	Multiple, eg, Kraus (2014),[75] and Refs. 56-72
Reward quality rather than quantity	Kraus (2014)[75]
Emphasis on hypothesis testing research	Winquist et al (2014)[40]
Prospective, rigorous experimental plan	Kenakin et al (2014)[38]
Improved understanding of statistics	Marino (2014)[41]; Vaux (2012)[35]
Improved experimental design	Henderson et al (2013)[37]
Systematic reviews of animal studies	Hooijmans & Ritskes-Hoitinga (2013)[76]
Use clinically relevant concentrations	Smith & Houghton (2013)[44]
Consider litter effects	Lazic & Essioux (2013)[28]
Recommendations to improve computational biology	Sandve et al (2013)[46]
Focus on reproducibility in training, grants, journals	LeVeque et al (2012)[61]
Pathology: Biospecimen quality control	Simeon-Dubach et al (2012)[47]
Microarray analyses: Provide data access	Witwer (2013)[30]
Psychology: open data, methods and workflow	Nosek et al (2012)[72]
Meta-analyses of animal data	Macleod et al (2004)[77]
Judge academics on quality, reproducibility, sharing	Ioannidis et al (2014)[6]
Greater institutional responsibility	Chan et al (2014)[9]
Apply greater skepticism to new technologies	Glaeser (2006)[78]

Begley, C.G. and J.P. Ioannidis, **Reproducibility in science: improving the standard for basic and preclinical research.** Circ Res, 2015. **116**(1): p. 116-26.

Επαναληψιμότητα (4)

Table 4. Issues That Could Be Addressed by a Policy of Good Institutional Practice for Basic Research

Focus	Proposal
Students/post-doctoral fellows	Core training in experimental methods and experimental design; data selection; data analysis; blinding; inclusion of controls; statistical interpretation; reagent validation; experimental replicates and repeats
	Mentoring provided by senior colleague from independent department
Investigator	Requirement that subjective end points are assessed by blinded investigators
	Compulsory refresher courses on experimental design; data selection; inclusion of controls; data analysis; statistical interpretation; reagent validation; issues in emerging technologies
	Requirement to comply with Federal and Scientific community guidelines and recommendations
Institution	Guidelines for dealing with fraud
	Independent committee to review compliance
	Requirement that raw data will be made available on request
	Guidelines for recording of laboratory notebooks
	Random reviews of laboratory notebooks
	Transparent promotion process that weighs quality above flashy, nonreproducible research; rewards mentoring and training

Begley, C.G. and J.P. Ioannidis, **Reproducibility in science: improving the standard for basic and preclinical research.** Circ Res, 2015. **116**(1): p. 116-26.

Table 5. Some Potential Recommendations

Funding Agencies, Investigators, Institutions, Journals

- Routine application of good scientific method (blinding, controls, repeats, presentation of representative data, reagent validation, adequate powering, etc)
- Demand and monitor compliance with consensus-based, peer-endorsed guidelines (eg, recommendations for animal pharmacology; MIAME; neuroscience studies, etc)
- Demand and monitor compliance with National Science Foundation and the National Institutes of Health requirements regarding data access

Funding agencies

- Provide more longer-term funding (people rather than projects)
- Fund projects to evaluate effect of compliance interventions
- Support reagent validation projects (antibodies; small molecules; siRNA, etc)
- Provide courses on scientific method for training junior investigators
- Monitor and reward reproducible, robust rather than flashy studies

Institutions

- Monitor and reward investigator-compliance with peer-generated guidelines and funding-agency requirements
- Monitor and reward reproducible, robust rather than flashy studies
- Support studies to evaluate effect of compliance interventions
- Provide compulsory courses on scientific method for junior and senior investigators

Journals

- Label exploratory investigations for what they are (ie, the equivalent of a phase 1 clinical study)
- Give greater weight to hypothesis-testing studies (the equivalent of a phase 2,3 clinical study)
- Encourage prepublication (eg via arXiv.org)

MIAME indicates minimum information about a microarray experiment.

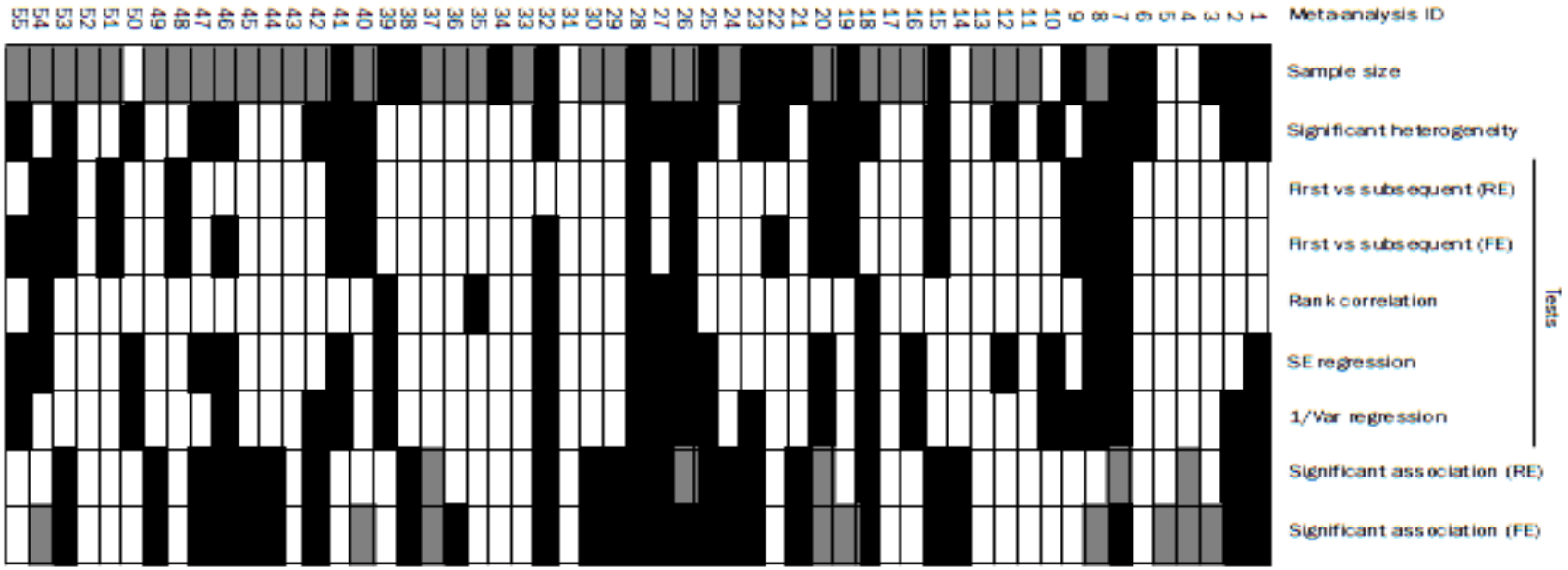
Begley, C.G. and J.P. Ioannidis, **Reproducibility in science: improving the standard for basic and preclinical research.** Circ Res, 2015. **116**(1): p. 116-26.

Επίδραση μελετών μικρού μεγέθους δείγματος (1)

- Από ένα σύνολο 55 μετα-αναλύσεων μελετών γενετικής συσχέτισης διερευνήθηκε εάν η επίδραση του γενετικού παράγοντα άλλαζε ανάλογα με το μέγεθος του δείγματος κάθε μελέτης.
- Στατιστικά σημαντική ετερογένεια μεταξύ των μελετών εκτιμήθηκε για 26 μετα-αναλύσεις (47%).
- Οι μεγάλες μελέτες σε γενικές γραμμές είχαν πιο συντηρητικά αποτελέσματα σε σχέση με τις μετα-αναλύσεις που περιελάμβαναν το σύνολο των μελετών.
- Σε 14 μετα-αναλύσεις η πρώτη μελέτη υποδείκνυε αυξημένη συσχέτιση του υπό μελέτη γονιδίου με την ασθένεια σε σχέση με τις μελέτες που έπονταν.
- Μόνο σε 9 μετα-αναλύσεις παρατηρήθηκε επαναληψιμότητα των αποτελεσμάτων των επιμέρους μελετών και απουσία ετερογένειας-μεροληψίας.

Ioannidis, J.P., et al., **Genetic associations in large versus small studies: an empirical assessment.** Lancet, 2003. 361(9357): p. 567-71.

Επίδραση μελετών μικρού μεγέθους δείγματος (2)



Heterogeneity and bias

Total sample size (participants or alleles): 0–1000 (white), 1001–5000 (grey), and >5000 (black). Heterogeneity and tests of bias and heterogeneity are black if significant. Also shown are presence of significant association in the overall meta-analysis by random effects (RE) and by fixed effects (FE): black=significance both in the overall meta-analysis and when the first studies are excluded; grey=significance in the overall meta-analysis that is lost when the first studies are excluded. Var=variance.

Ioannidis, J.P., et al., **Genetic associations in large versus small studies: an empirical assessment.** Lancet, 2003. 361(9357): p. 567-71.

Ιεράρχηση της αξιοπιστίας μοριακών ευρημάτων σε πολυπαραγοντικές ασθένειες (1)

Table 1 Effect sizes in the pre-molecular era and in the molecular era[a]

Effect sizes	Putative frequency	Typical examples of postulated risk factors	
		Pre-molecular era	Molecular era
Large (RR > 5)	Rare	Smoking and lung cancer	APOE and Alzheimer's disease[31]
			BRCA1 and breast cancer[32]
Moderate (RR 2–5)	Uncommon	Moderate obesity and cholesterol gallstones	NOD2 and Crohn's disease[33]
			HLA shared epitopes and rheumatoid arthritis[34]
Small (RR 1.2–2)	Common	Racial descent and hypertension	FcγRIIa and SLE[35]
			GSTM1 and bladder cancer[36]
Very small (RR 1–1.2)	Unclear frequency[a]	Passive smoking and lung cancer	GSTM1 and lung cancer[37]
			MTHFR and ischaemic stroke[38]

RR: relative risk.

[a] Presented examples reflect current state of knowledge and are subject to possible refutation in the future; for small and very small effect sizes, it is uncertain whether these risk factors are true, even when evidence is based on large sample sizes from several studies.

Ioannidis, J.P., **Commentary: grading the credibility of molecular evidence for complex diseases.** Int J Epidemiol, 2006. **35**(3): p. 572-8; discussion 593-6.

Ιεράρχηση της αξιοπιστίας μοριακών ευρημάτων σε πολυπαραγοντικές ασθένειες (2)

Table 2 Typical credibility of research findings according to effect size and extent of replication

Effect size (relative risk)	Replication	Typical credibility (%)
Large (>5)	None	10–60
	Limited	30–80
	Extensive	70–95
Moderate (2–5)	None	5–20
	Limited	10–40
	Extensive	50–90
Small (1.2–2)	None	<5
	Limited	2–20
	Extensive	10–70
Very small (1–1.2)	None	<1
	Limited	1–5
	Extensive	2–30

Ioannidis, J.P., **Commentary: grading the credibility of molecular evidence for complex diseases.** Int J Epidemiol, 2006. **35**(3): p. 572-8; discussion 593-6.

Table 3 Proposed grading of credibility in molecular evidence

First axis: Effect size	
1.1	Very small or small effect size (relative risk < 2)
1.2	Moderate effect size (relative risk 2–5)
1.3	Large effect size (relative risk > 5)
Second axis: Amount and replication of evidence	
2.1	Single or few scattered studies
2.2	Meta-analyses of group data
2.3	Large-scale evidence from inclusive networks
Third axis: Protection from bias	
3.1	Clear presence of strong bias in the evidence
3.2	Uncertain about the presence of bias
3.3	Clear strong protection from bias
Fourth axis: Biological credibility	
4.1	No functional/biological data or negative data
4.2	Limited or controversial functional/biological data
4.3	Convincing functional/biological data
Fifth axis: Relevance	
5.1	No clinical or public health applicability
5.2	Limited clinical or public health applicability
5.3	Considerable clinical/public health applicability

Ioannidis, J.P., **Commentary: grading the credibility of molecular evidence for complex diseases.** Int J Epidemiol, 2006. **35**(3): p. 572-8; discussion 593-6.