Advice to statistical consultants contributing and evaluating evidence for a nutrition intervention

GGB Seminar

November 17, 2015

Daniel J. Tancredi, PhD

Associate Professor in Residence of Pediatrics, UC Davis

Objectives

Review statistical concepts, principles and methods relevant to evaluating evidentiary quality of nutrition studies

1. STATISTICAL PRELIMINARIES

1A. P VALUES & CONFIDENCE INTERVALS

Which of the following things does a report of $P < 0.05$ allow you to know?

1. The probability that the null hypothesis is true.
2. The probability that the alternative hypothesis is true.
3. The probability that the observed effect is real.
4. The probability that a claim of a positive result is a false positive claim.
5. The probability that the result can be replicated.
6. The strength of the evidence in the data against the null hypothesis.

Sources: Lew MJ. (2012) “Bad statistical practice in pharmacology…: you probably don’t know P”. BJP 166: 1559-1567

Haller H, Krauss S (2002) “Misinterpretation of significance: a problem students share with their teachers.” Methods Psych Res 7: 1-20

Definition of a P-value

Suppose you have

a null hypothesis and

a method for converting sample data into a test statistic that has the property that extreme values constitute evidence against the null hypothesis.

Example

H_o: Mean of outcome is equal in two comparison groups

Test statistic is between-group difference in sample means

Definition of a P-value *(cont.)*

Only then can you define the p-value associated with the value for the test statistic observed in the given sample.

The p-value is the conditional probability, under (the data generating model associated with) the **null hypothesis**, of obtaining a value for the test statistic that is as least as extreme as **the observed value in the sample**.

P-values and “significance testing”

- R.A. Fisher promoted the P-value as a measure of the strength of the evidence within the observed data against a null hypothesis and introduced the word “significant”
- Fisher’s rivals Jerzy Neyman and Egon Pearson introduce an alternative inferential approach that uses
  - long-term error rates,
  - appropriately powered experiments
  - binary decision making

P-values and "significance testing"

Unfortunately, both approaches use term *"significant"*, leading to confusing hybrid approaches seen in practice (i.e. same paper using $p < 0.05$, $p < 0.01$, etc.)

Sources: Fisher RA (1925): Statistical Methods for Research Workers. Oliver and Boyd: Edinburgh. http://pscychclassics.yorku.ca/Fisher/Methods/

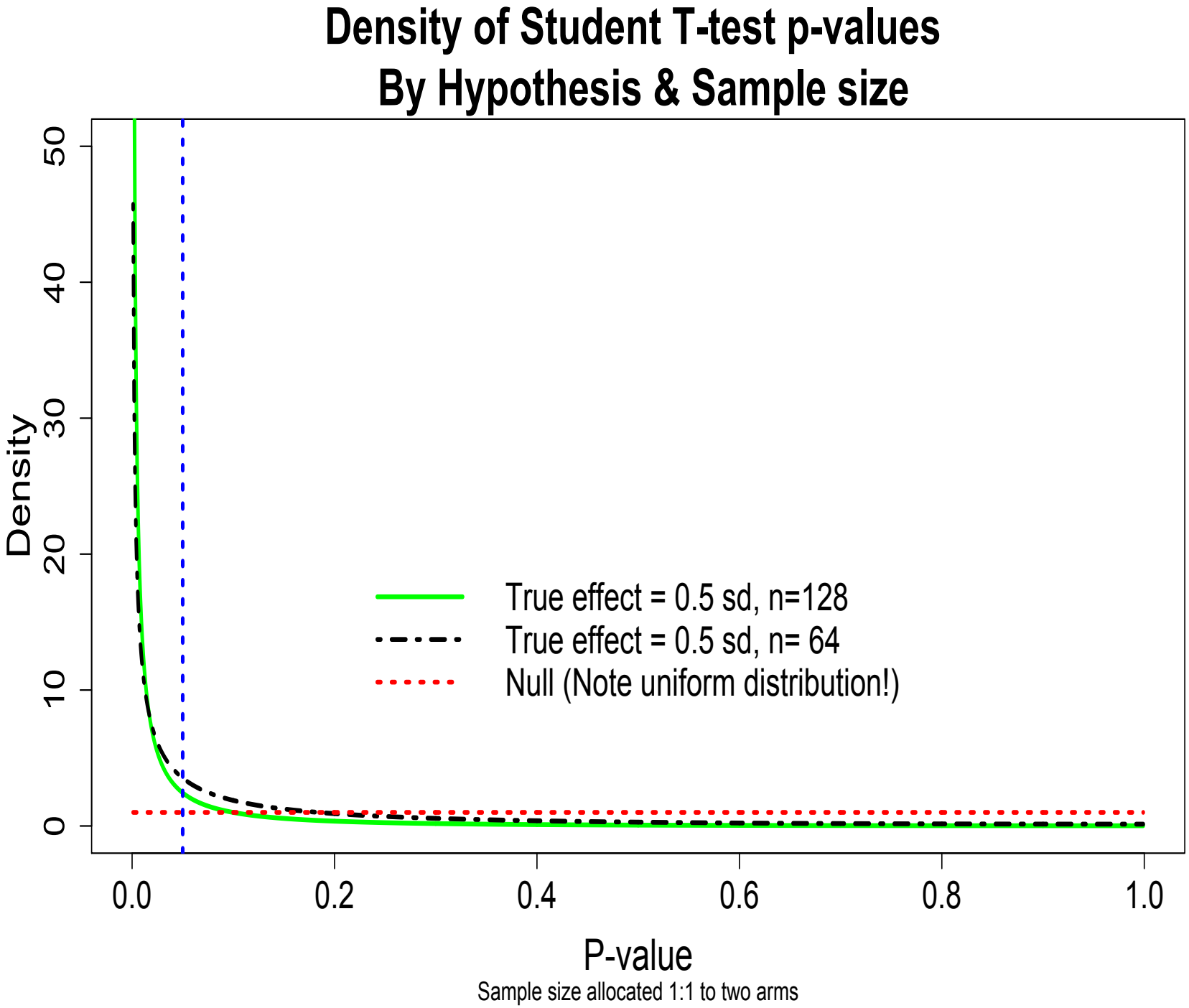
Neyman J, Pearson ES (1933): On the problem of the most efficient test of statistical hypotheses. Philos Trans R Soc Long A 231: 289-337.

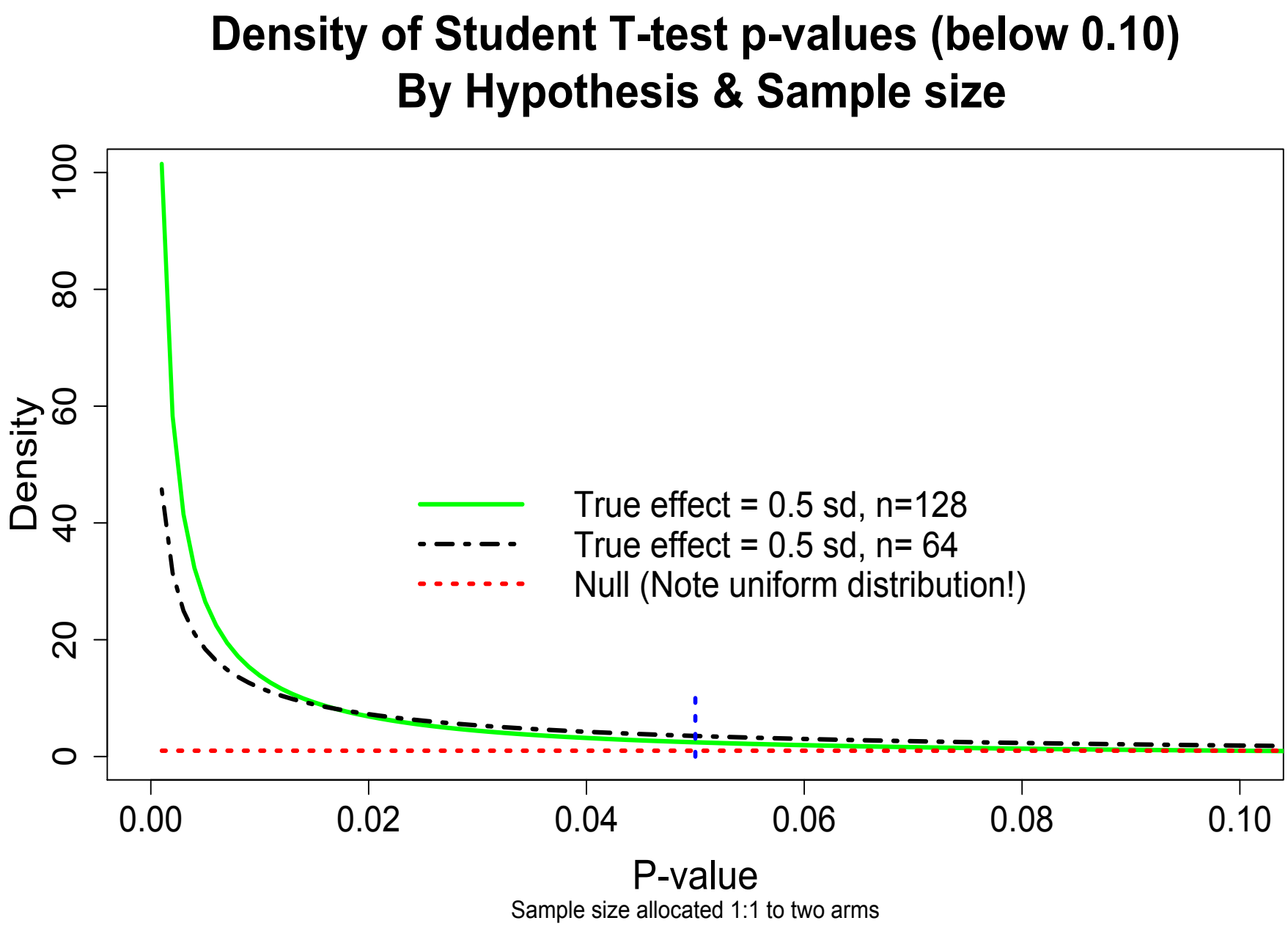
Distribution of p values

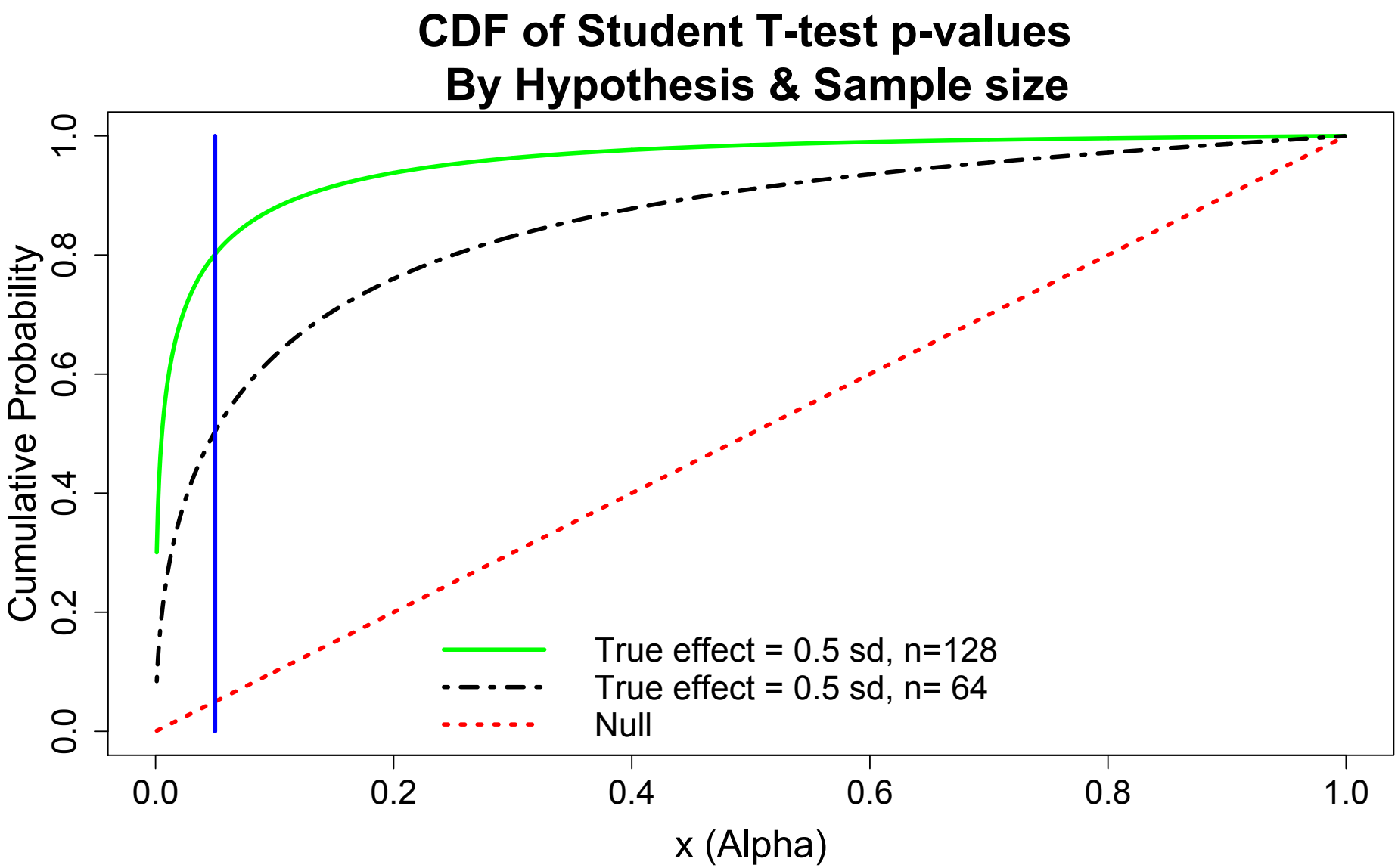
Suppose you want to perform a two-group comparison of means using Student's t-test.

What's the shape of the (theoretical) distribution of the p-values under

- Null hypothesis?
- When true effect size (difference in means) is 0.5 standard deviation and power is
  - 50%?
  - 80%?


Density of Student T-test p-values
By Hypothesis & Sample size
Density
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
P-value
True effect = 0.5 sd, n=128
True effect = 0.5 sd, n= 64
Null (Note uniform distribution!)
Sample size allocated 1:1 to two arms


Density of Student T-test p-values (below 0.10)
By Hypothesis & Sample size
Density
0
20
40
60
80
100
0.00
0.02
0.04
0.06
0.08
0.10
P-value
Sample size allocated 1:1 to two arms
True effect = 0.5 sd, n=128
True effect = 0.5 sd, n= 64
Null (Note uniform distribution!)



Power is value of CDF at given alpha (5%, typically).

So, which of the following things does a report of P < 0.05 allow you to know?

1. The probability that the null hypothesis was true.
2. The probability that the alternative hypothesis was true.
3. The probability that the observed effect was real.
4. The probability that a claim of a positive result is a false positive claim.
5. The probability that the result can be replicated.
6. The strength of the evidence in the data against the null hypothesis.

Sources: Lew MJ. (2012) “Bad statistical practice in pharmacology…: you probably don’t know P”. BJP 166: 1559-1567

Haller H, Krauss S (2002) “Misinterpretation of significance: a problem students share with their teachers.” Methods Psych Res 7: 1-20

Problems with p-values and so-called Null Hypothesis Significance Testing

1. Failing to reject H0 is not proof that H0 is true ("absence of evidence is not evidence of absence").
2. P value is very likely to be quite different if experiment is repeated, particularly for underpowered (most!) studies
3. H0 is almost never true (strictly), anyway. As n grows, so does probability of rejecting H0.
4. **P value does not give an estimate of the effect size.**
5. **P value does not give information on precision.**

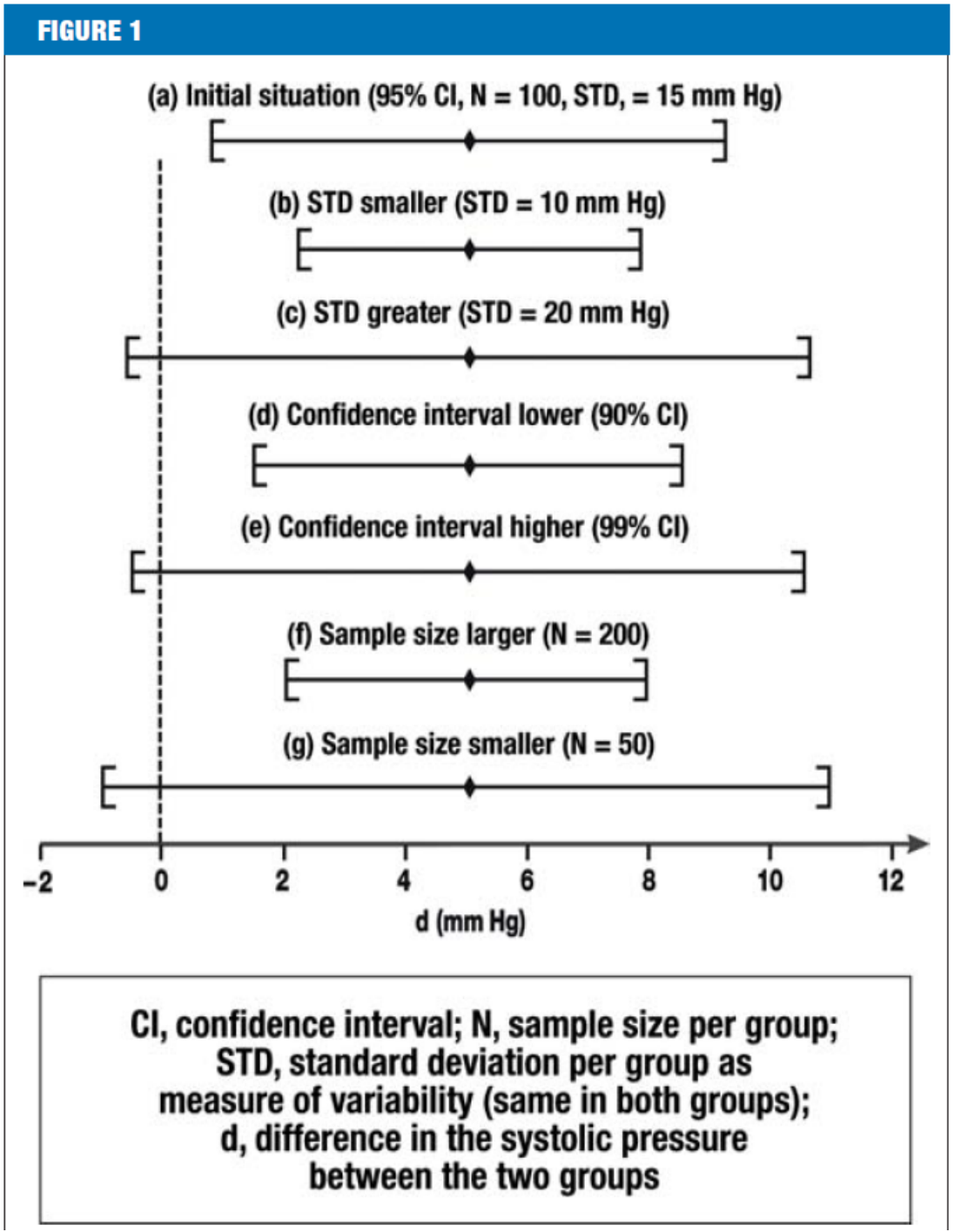
Sources: Cumming G (2008) "Replication and p Intervals: p Values predict the future Only Vaguely but Confidence Intervals do Much Better". Persp of Bio Sci 3:286-300

Tressoldi PE et al (2013) "High Impact = High Statistical Standards? Not Necessarily So". PLOS ONE 8:e56180

Confidence intervals rather than P values: estimation rather than hypothesis testing

Well known article by Martin Gardner and Doug Altman [Br Med J (Clin Res Ed). 1986 Mar 15;292(6522):746-50] led to recent high-profile recommendations from CONSORT, APA, & ICMJE (http://www.icmje.org/manuscript_1prepare.html), like this:

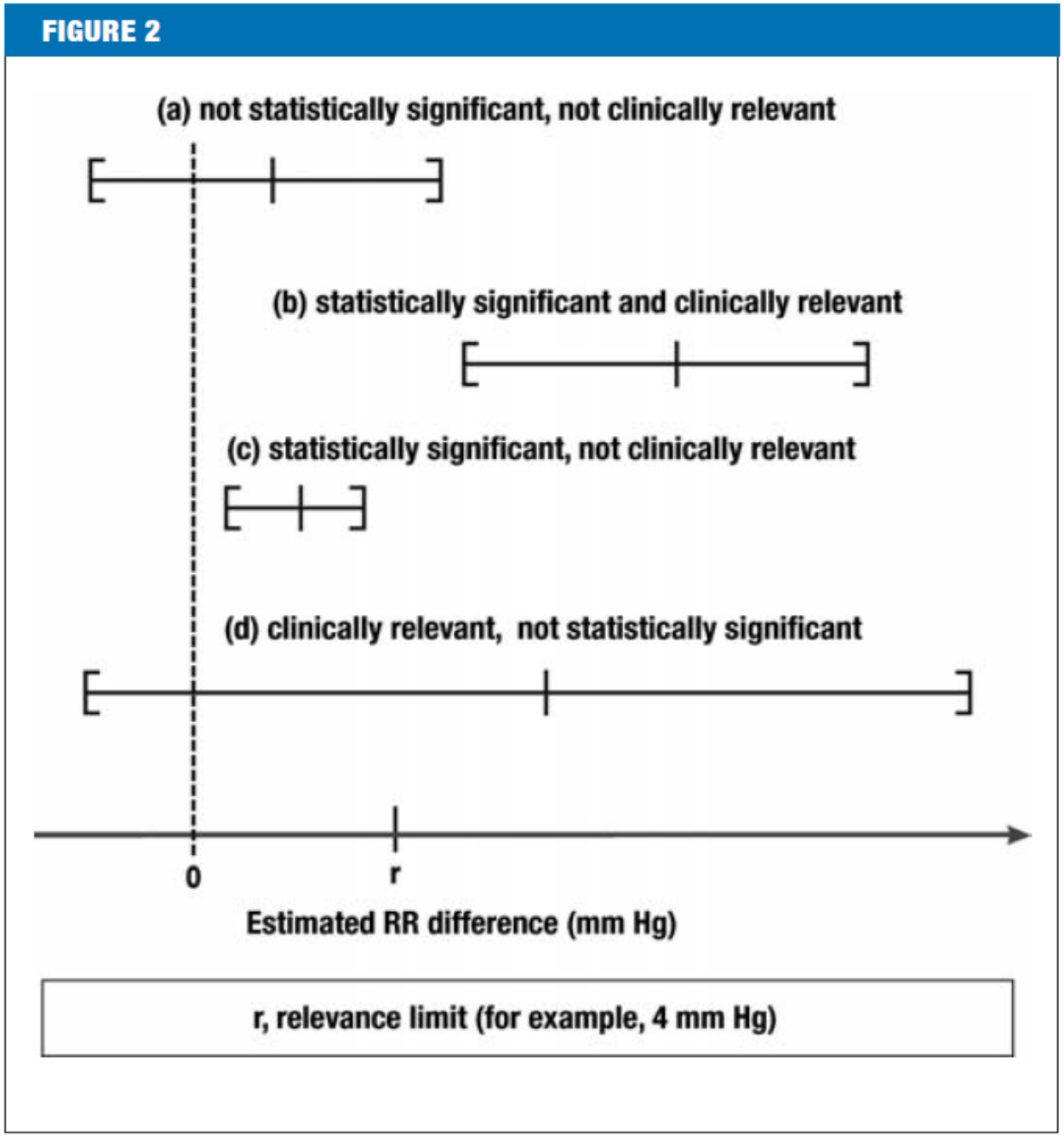
"...When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as *P* values, which fail to convey important information about effect size..."



The confidence interval is a range of values with the property that it includes the true value of the parameter with a probability defined in advance. (The probability is a property of the procedure used to convert sample data into interval estimates.)

Describes hypothesized values of the true parameter that would be considered "plausible".

Source: (see next slide)



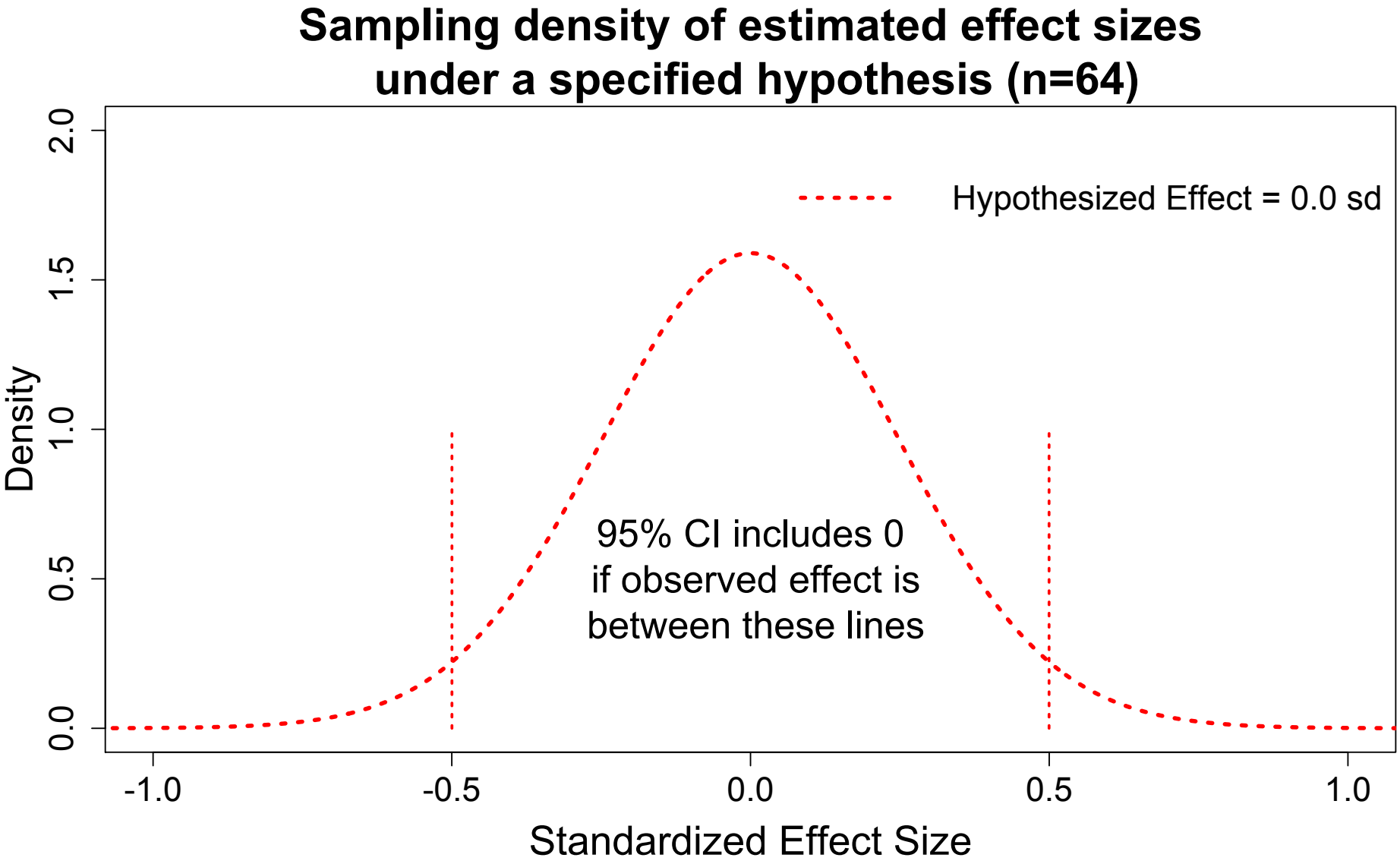
Statistical significance and clinical relevance

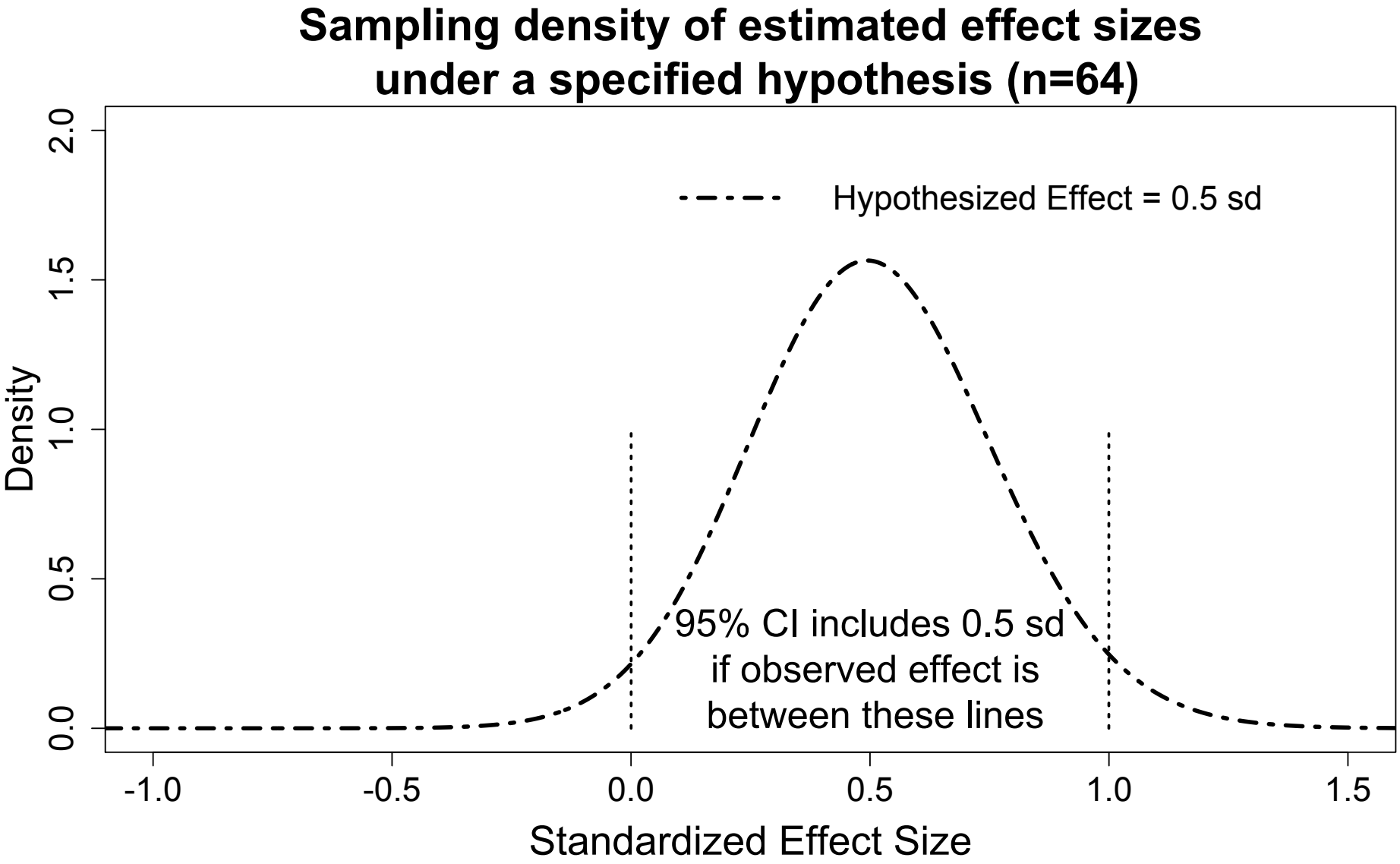
Confidence Intervals are harder for lay readers to misinterpret disastrously!

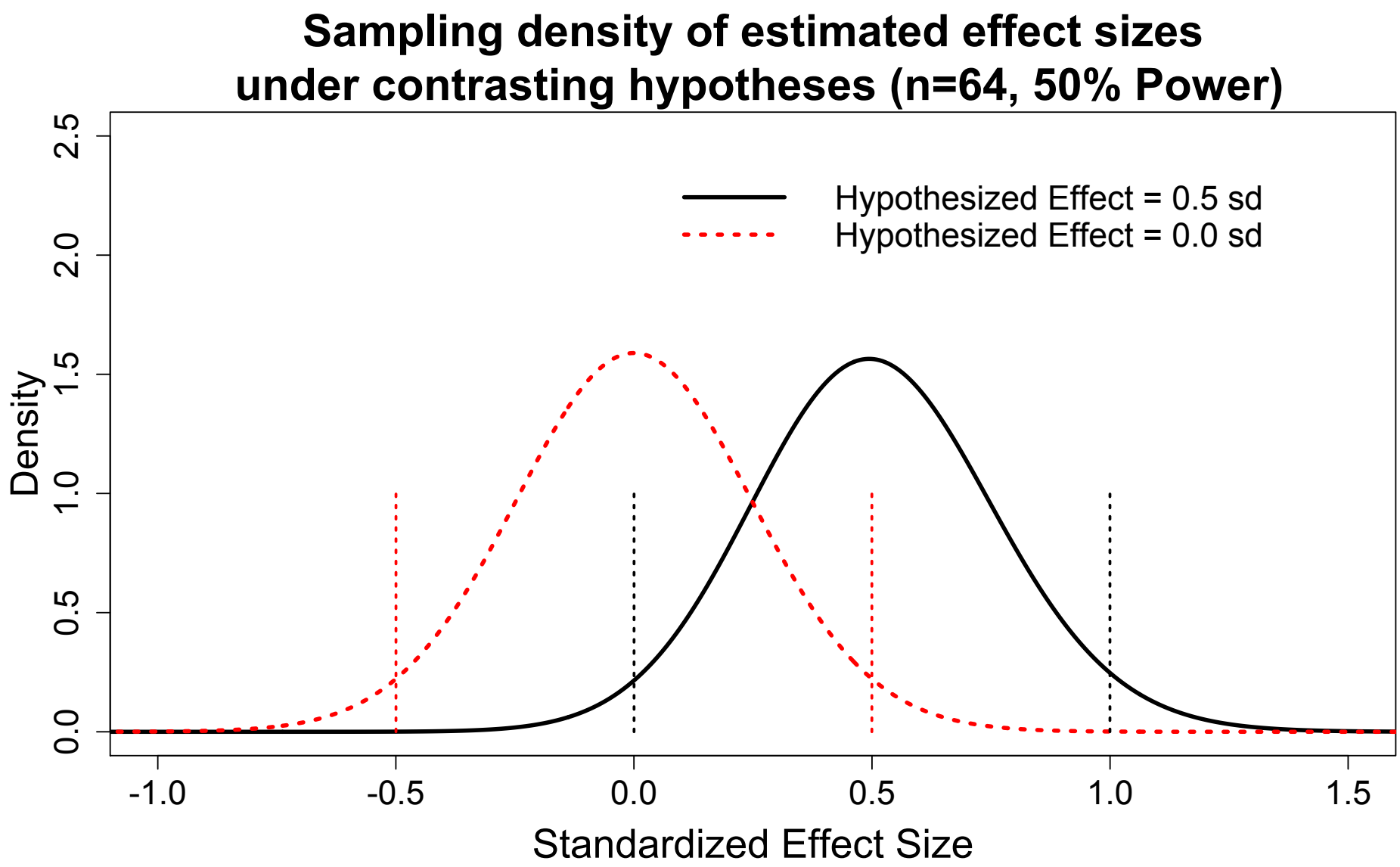
Source for Figures 1-2: du Prel J-B, Hommel G, et al. (2009) "Confidence Interval of P-value?" Dtsch Arztebl Int. May; 106(19): 335–339.

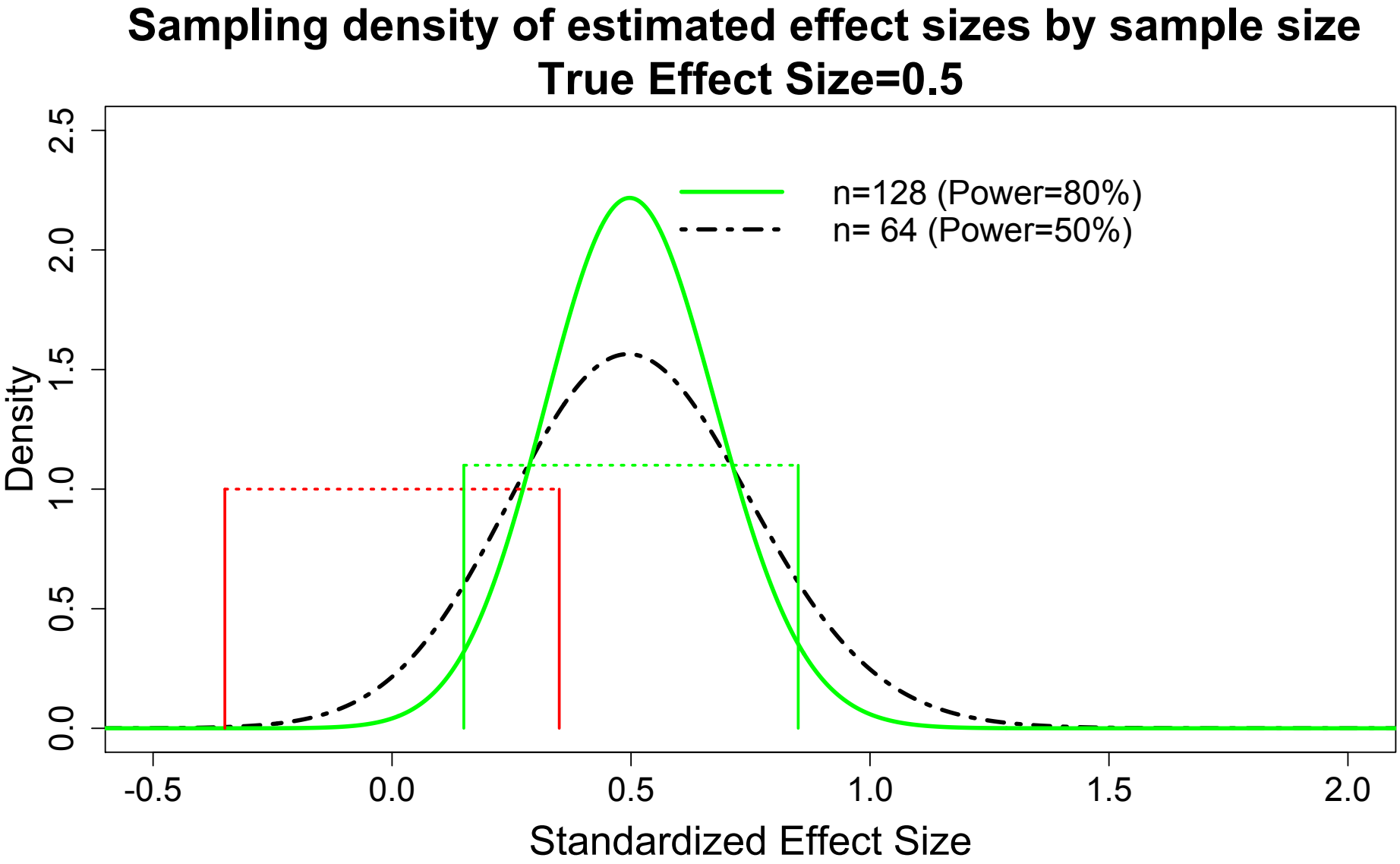
Also see: Hoenig JM and Heisey (2001). The abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. Am Stat 55: 1-6.

1B. DUALITY BETWEEN CONFIDENCE INTERVALS AND SIGNIFICANCE TESTING

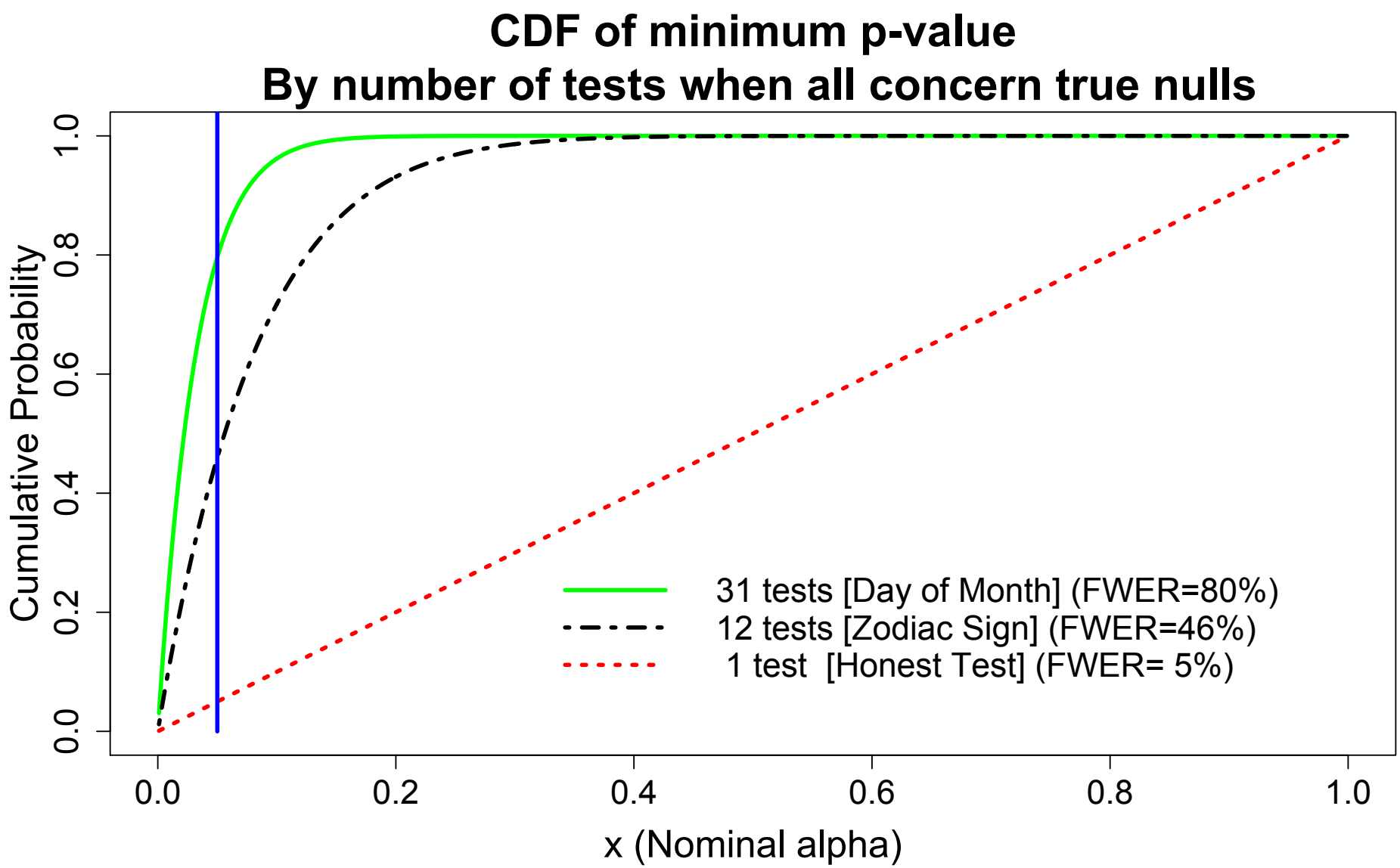

Sampling density of estimated effect sizes
under a specified hypothesis (n=64)
Hypothesized Effect = 0.0 sd
95% CI includes 0
if observed effect is
between these lines
Density
0.0
0.5
1.0
1.5
2.0
-1.0
-0.5
0.0
0.5
1.0
Standardized Effect Size


Sampling density of estimated effect sizes
under a specified hypothesis (n=64)
Hypothesized Effect = 0.5 sd
95% CI includes 0.5 sd
if observed effect is
between these lines
Density
0.0
0.5
1.0
1.5
2.0
-1.0
-0.5
0.0
0.5
1.0
1.5
Standardized Effect Size


Sampling density of estimated effect sizes
under contrasting hypotheses (n=64, 50% Power)
Hypothesized Effect = 0.5 sd
Hypothesized Effect = 0.0 sd
Density
0.0
0.5
1.0
1.5
2.0
2.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Standardized Effect Size


Sampling density of estimated effect sizes by sample size
True Effect Size=0.5
n=128 (Power=80%)
n= 64 (Power=50%)
Density
0.0
0.5
1.0
1.5
2.0
2.5
-0.5
0.0
0.5
1.0
1.5
2.0
Standardized Effect Size

1C. MULTIPLE TESTING


CDF of minimum p-value
By number of tests when all concern true nulls
Cumulative Probability
0.0
0.2
0.4
0.6
0.8
1.0
x (Nominal alpha)
31 tests [Day of Month] (FWER=80%)
12 tests [Zodiac Sign] (FWER=46%)
1 test [Honest Test] (FWER= 5%)

Bonferonni & Holm-Bonferonni Adjustments

Suppose 3 tests were performed and order p-values from smallest to largest, P1, P2, P3

	Bonferroni		Holm-Bonferroni	
Ordered	Adj. alpha	Adj. p	Adj. alpha	Adj. p
P1	0.05 / 3	3*P1	0.05 / 3	3*P1
P2	0.05 / 3	3*P2	0.05 / 2	2*P2
P3	0.05 / 3	3*P3	0.05 / 1	1*P3

1D. EFFECT SIZES

Important effect sizes

(Standardized) differences in means:

Differences in proportion (aka Risk Reduction)

Number Needed to Treat (aka NNT)

= 1 / Risk Reduction

Risk ratios (or variants involving Odds, Hazards)

[Difference in group mean log-transformed values is a log geometric mean ratio]

[Regression-based estimates of above]

2. QUALITY OF EVIDENCE

What do/should we mean when we talk about the <u>quality</u> of a study or a group of studies in how it addresses <u>a research question</u>?

Bias

A single study can provide an estimate of the true effect of an intervention (on average, in the sampled population):

Study estimate = True Effect + Study error

Study error = Systematic error + sampling error

Bias = **Long run average***(Study error),*

over hypothetical repetitions of the study.

Hence, bias is essentially **systematic error** arising from such features as subject recruitment & retention, treatment assignment, measurement procedures and analysis

Internal Validity

The extent to which the observed results of a clinical research study are not biased.

- "Were the comparison groups similar in all important characteristics that may affect the measurements?"
- "Were the data measured and compared using accurate methods?"

For causal claims, an internally valid study would:

- Show association
- Show temporal precedence
- **Rule out plausible alternative explanations**

Source for definition of *Internal Validity*:

http://www.effectivehealthcare.ahrq.gov/index.cfm/glossary-of-terms/

Internal validity & research designs

Quality of evidence depends crucially on level of internal validity associated with study

True experiments typically the preferred (primary) study design

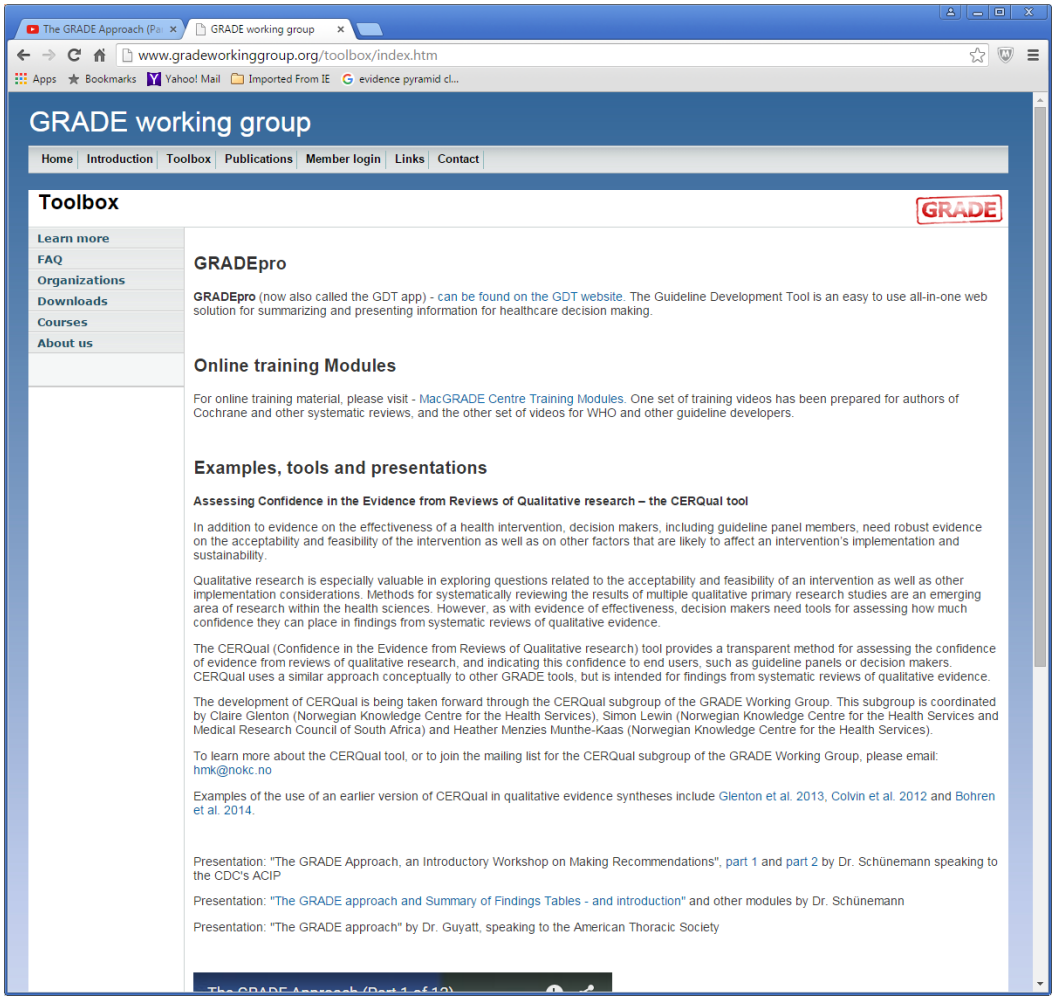
See, for example,

Puddy, R. W. & Wilkins, N. (2011). *Understanding Evidence Part 1: Best Available Research Evidence. A Guide to the Continuum of Evidence of Effectiveness*. Atlanta, GA: Centers for Disease Control and Prevention(http://www.cdc.gov/violenceprevention/pdf/understanding_evidence-a.pdf)

2. GRADE WORKING GROUP'S APPROACH TO QUALITY OF EVIDENCE

Key sources of material

gradeWorkingGroup.org, especially Dr. Guyatt's presentation to American Thoracic Society and the series of BMJ papers



Frame research question

Explicit specification of *PICO(TS)*

- *Population, including settings/locations*
- *Intervention(s), including vehicles/matrices*
- *Comparison*
- *Outcome(s), including Timing & how measured*
- *Study types (designs & methodological quality)*

Source: Counsell, Carl. Formulating questions and locating primary studies for inclusion in systematic reviews. *Ann Intern Med*. 1997 Sep 1;127(5):380-7.

PICOTS specify key elements for reviewing efficacy claims

Population: Condition(s), comorbidities, patient demographics, diet, physical activity levels, etc.

Intervention: Dosage, frequency, and method of administration.

Comparator: Placebo, usual diet, or active control.

Outcome: Health outcomes: morbidity, mortality, quality of life. **T**iming: Duration of follow-up.

Setting: Lab, home; co-interventions.

Explicitly address each outcome's importance

Score each outcome:

7-9) Critical for decision making

4-6) Important but not critical

1-3) Limited importance

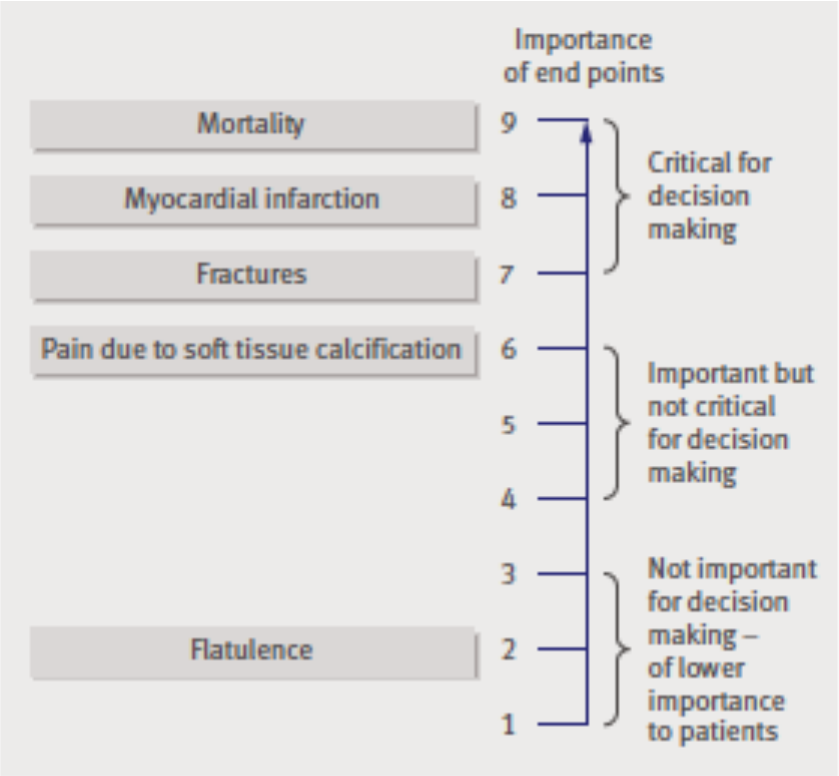


Fig 1 Hierarchy of outcomes according to importance to patients to assess effect of phosphate lowering drugs in patients with renal failure and hyperphosphataemia

Quality of evidence and context for recommendation

Quality of evidence in a study is confidence that estimated effect size is close to true parameter

For decision making, quality is extent to which confidence in estimated effect is adequate to support decision

GRADE Quality of evidence definitions

High: Further research (FR) unlikely to change confidence in estimated effect size (EES)

Moderate: FR can impact confidence in and may change EES

Low: FR very likely to impact confidence and likely to change EES

Very low: Any estimate of EES is uncertain

- Grading done for each important outcome!

Classification of QoE

GRADE classifies QoE according to

- Study limitations
- Inconsistency of results
- Indirectness of evidence
- Imprecision
- Reporting bias

Study design & limitations

RCTs presumed best, observational studies lower

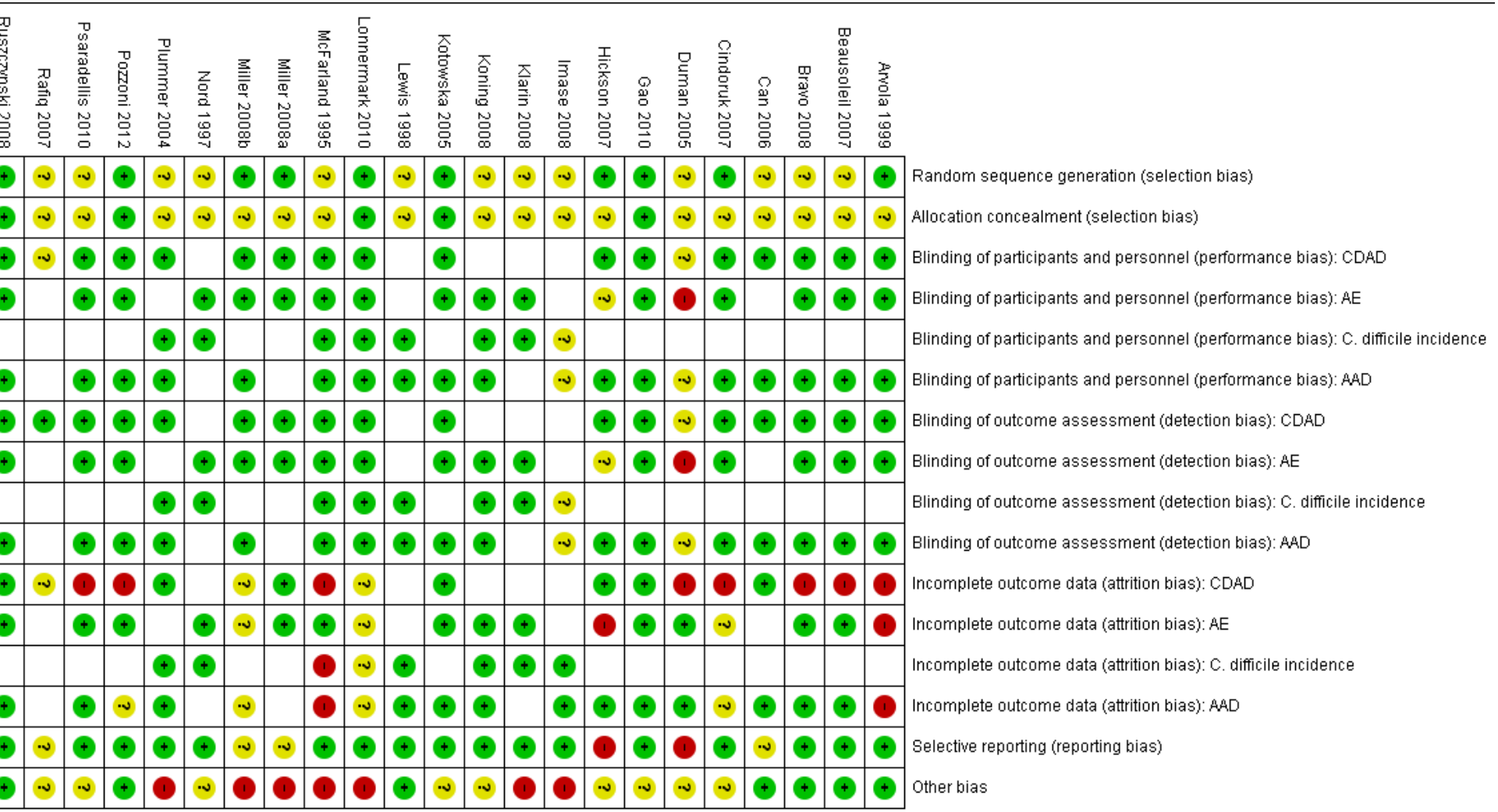
For RCTs, assess (to lower quality rating)

- Random sequence generation/concealment
- Blinding
- Incomplete outcome data
- Selective reporting and other biases

For Observational studies, assess (to increase quality rating)

- Large magnitude of effect
- Size & direction of plausible confounding
- Dose-response gradient

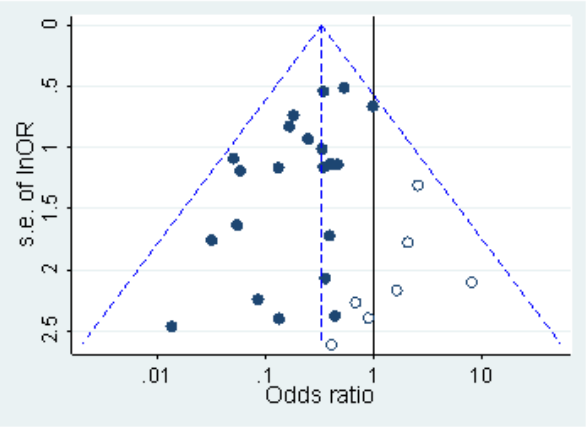
Probiotics for the prevention of Clostridium difficile-associated diarrhea in adults and children



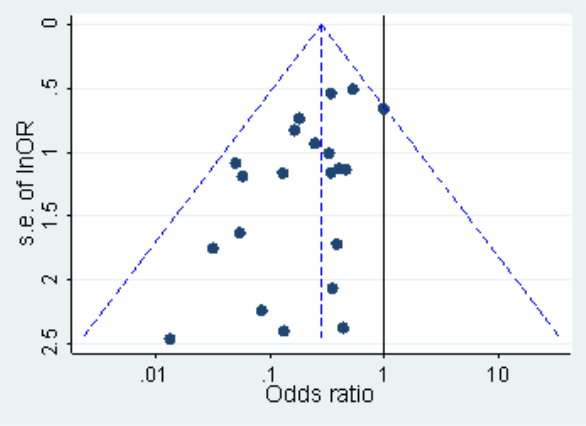
Cochrane Database of Systematic Reviews
31 MAY 2013 DOI: 10.1002/14651858.CD006095.pub3
http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD006095.pub3/full#CD006095-fig-0002

Funnel Plots Example

Symmetrical plot in the absence of reporting bias



Asymmetrical plot in the presence of reporting bias



http://handbook.cochrane.org/chapter_10/figure_10_4_a_hypothetical_funnel_plots.htm

TRIAL Registration

Registration of a clinical trial in a recognized trial registry is a crucial protection against reporting biases!

http://www.ICMJE.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html

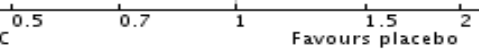
Meta-Analysis

- Meta-analysis may reduce imprecision, but it can't reduce biases
- Heuristically, a pooled effect size is estimated as a weighted averages of sample effect sizes, resulting in a more precise estimate (with a smaller uncertainty interval)
- Assess explainable/unexplainable heterogeneity in effect sizes, including subgroups
- Sensitivity analyses to assess robustness

Sackett DL, Glasziou P, Chalmers I. *Meta-analysis may reduce imprecision, but it can't reduce bias*. Unpublished commentary commissioned by the New England Journal of Medicine, 1997. (see SR in Health Care, p. xiv)

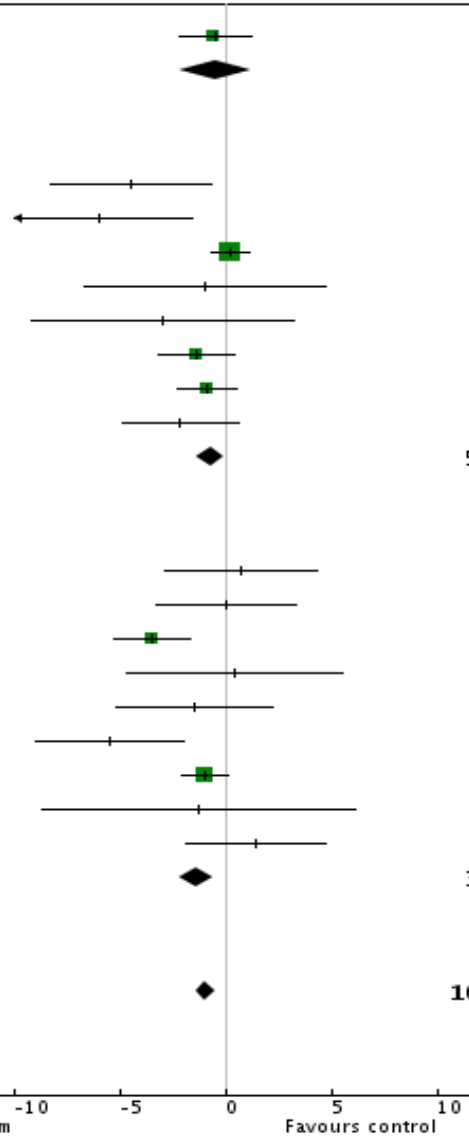
Review: Vitamin C for preventing and treating the common cold
Comparison: 1 Incidence of colds while taking ≥ 0.2 g/day vitamin C regularly
Outcome: 1 Proportion of participants developing ≥ 1 cold episodes during the trial

Study or subgroup	Vitamin C n/N	Placebo n/N	Risk Ratio M-H,Fixed,95% CI	Weight	Risk Ratio M-H,Fixed,95% CI
1 All eligible trials with exception of subgroup removed below					
Peters 1996b	5/41	11/45		0.4 %	0.50 [0.19, 1.31]
Moolla 1996b	5/11	12/19		0.3 %	0.72 [0.35, 1.50]
Charleston 1972	31/47	37/43		1.5 %	0.77 [0.60, 0.97]
Coulehan 1974a	19/190	23/192		0.9 %	0.83 [0.47, 1.48]
Anderson 1972	302/407	335/411		12.7 %	0.91 [0.85, 0.98]
Coulehan 1974b	16/131	17/128		0.7 %	0.92 [0.49, 1.74]
Dahlberg 1944	131/1259	142/1266		5.4 %	0.93 [0.74, 1.16]
Bancalari 1984	21/32	21/30		0.8 %	0.94 [0.67, 1.32]
Anderson 1974a	922/1191	233/285		14.4 %	0.95 [0.89, 1.01]
Franz 1956	14/44	15/45		0.6 %	0.95 [0.52, 1.74]
Sasazuki 2006	68/140	67/133		2.6 %	0.96 [0.76, 1.23]
Cowan 1942	184/208	142/155		6.2 %	0.97 [0.90, 1.03]
Ludvigsson 1977b	230/304	240/311		9.1 %	0.98 [0.90, 1.07]
Pitt 1979	298/331	309/343		11.6 %	1.00 [0.95, 1.05]
Coulehan 1976	98/428	98/428		3.7 %	1.00 [0.78, 1.28]
Clegg 1975	48/67	50/70		1.9 %	1.00 [0.81, 1.24]
Elwood 1976	296/339	298/349		11.2 %	1.02 [0.96, 1.09]
Briggs 1984	125/265	121/263		4.6 %	1.03 [0.85, 1.23]
Carson 1975	85/121	84/123		3.2 %	1.03 [0.87, 1.22]
Van Straten 2002	35/84	34/84		1.3 %	1.03 [0.72, 1.48]
Ludvigsson 1977a	49/80	44/78		1.7 %	1.09 [0.84, 1.41]
Liljefors 1972	10/33	9/33		0.3 %	1.11 [0.52, 2.38]
Peters 1993b	18/34	18/39		0.6 %	1.15 [0.72, 1.82]
Himmelstein 1998a	10/23	8/25		0.3 %	1.36 [0.65, 2.84]
Subtotal (95% CI)	**5810**	**4898**		**96.1 %**	**0.97 [0.94, 1.00]**
Total events: 3020 (Vitamin C), 2368 (Placebo)					
Heterogeneity: Chi² = 17.75, df = 23 (P = 0.77); I² =0.0%					
Test for overall effect: Z = 1.90 (P = 0.057)					
2 Short-term exposure to severe physical stress and/or cold					
Peters 1996a	7/44	19/47		0.7 %	0.39 [0.18, 0.84]
Sabiston 1974	6/56	14/56		0.5 %	0.43 [0.18, 1.04]
Moolla 1996a	4/13	13/19		0.4 %	0.45 [0.19, 1.07]
Peters 1993a	14/43	28/41		1.1 %	0.48 [0.30, 0.77]
Ritzel 1961	17/139	31/140		1.2 %	0.55 [0.32, 0.95]
Subtotal (95% CI)	**295**	**303**		**3.9 %**	**0.48 [0.35, 0.64]**
Total events: 48 (Vitamin C), 105 (Placebo)					
Heterogeneity: Chi² = 0.60, df = 4 (P = 0.96); I² =0.0%					
Test for overall effect: Z = 4.99 (P < 0.00001)					
Total (95% CI)	**6105**	**5201**		**100.0 %**	**0.95 [0.92, 0.98]**
Total events: 3068 (Vitamin C), 2473 (Placebo)					
Heterogeneity: Chi² = 44.85, df = 28 (P = 0.02); I² =38%					
Test for overall effect: Z = 3.18 (P = 0.0015)					
Test for subgroup differences: Chi² = 22.74, df = 1 (P = 0.00), I² =96%					



Review: Calcium supplementation for prevention of primary hypertension
Comparison: 1 Calcium supplementation/fortification vs control
Outcome: 20 Effect mean difference of diastolic blood pressure by dose

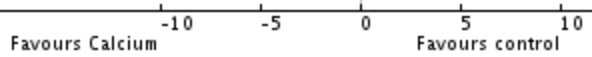
Study or subgroup	Calcium N	Mean(SD)	Control N	Mean(SD)	Mean Difference IV,Fixed,95% CI	Weight	Mean Difference IV,Fixed,95% CI
1 Diary calcium intake < less than 1000 mg							
Reid 2010 (1)	108	-0.71 (6.36)	54	-0.17 (4.45)		8.2 %	-0.54 [-2.23, 1.15]
Subtotal (95% CI)	**108**		**54**			**8.2 %**	**-0.54 [-2.23, 1.15]**
Heterogeneity: not applicable							
Test for overall effect: Z = 0.63 (P = 0.53)							
2 Diary calcium intake 1000 - 1250 mg							
Belizan 1983	15	-3.89 (5.8)	14	0.61 (4.71)		1.6 %	-4.50 [-8.33, -0.67]
Belizan 1983	15	-6.71 (6.15)	13	-0.69 (5.72)		1.2 %	-6.02 [-10.42, -1.62]
Cutler 1992	237	-2.75 (4.87)	234	-2.95 (5.21)		28.1 %	0.20 [-0.71, 1.11]
Hilary Green 2000	19	75 (9)	19	76 (9)		0.7 %	-1.00 [-6.72, 4.72]
McCarron 1985	16	75 (9)	16	78 (9)		0.6 %	-3.00 [-9.24, 3.24]
Reid 2010 (2)	108	-1.57 (7.3)	53	-0.17 (4.45)		7.0 %	-1.40 [-3.23, 0.43]
Sacks 1998	53	-0.6 (3.8)	103	0.3 (4.8)		12.2 %	-0.90 [-2.28, 0.48]
Shidfar 2010	24	-4.3 (3.4)	25	-2.1 (6.1)		3.1 %	-2.20 [-4.95, 0.55]
Subtotal (95% CI)	**487**		**477**			**54.5 %**	**-0.71 [-1.37, -0.06]**
Heterogeneity: Chi² = 15.46, df = 7 (P = 0.03); I² =55%							
Test for overall effect: Z = 2.14 (P = 0.033)							
3 Diary calcium intake 1500 mg or more							
Davis 1996	17	91.3 (4.7)	17	90.6 (6)		1.8 %	0.70 [-2.92, 4.32]
Johnson 1985	41	78 (8)	40	78 (7)		2.2 %	0.0 [-3.27, 3.27]
Lijnen 1995	16	-2.6 (2.5804)	16	0.9 (2.5804)		7.3 %	-3.50 [-5.29, -1.71]
Lyle 1987 (3)	10	77.1 (4.5)	11	76.7 (7.3)		0.9 %	0.40 [-4.74, 5.54]
Lyle 1987 (4)	27	72.8 (4.8)	27	74.3 (8.6)		1.7 %	-1.50 [-5.21, 2.21]
Lyle 1992	21	81.8 (4.8)	21	87.3 (6.7)		1.9 %	-5.50 [-9.03, -1.97]
Reid 2005	732	-0.2 (10.8222)	739	0.8 (10.8738)		19.0 %	-1.00 [-2.11, 0.11]
Thomsen 1987	14	77.3 (10.1)	14	78.6 (9.9)		0.4 %	-1.30 [-8.71, 6.11]
Van Beresteyn 1986	29	63.4 (4.8)	29	62 (7.7)		2.1 %	1.40 [-1.90, 4.70]
Subtotal (95% CI)	**907**		**914**			**37.3 %**	**-1.43 [-2.22, -0.64]**
Heterogeneity: Chi² = 16.22, df = 8 (P = 0.04); I² =51%							
Test for overall effect: Z = 3.55 (P = 0.00039)							
Total (95% CI)	**1502**		**1445**			**100.0 %**	**-0.97 [-1.45, -0.48]**
Heterogeneity: Chi² = 33.83, df = 17 (P = 0.01); I² =50%							
Test for overall effect: Z = 3.92 (P = 0.000088)							
Test for subgroup differences: Chi² = 2.15, df = 2 (P = 0.34), I² =7%							



(1) Intervention: elemental calcium 600 mg daily
(2) Intervention: elemental calcium 1200 mg daily
(3) Black men
(4) White men

Review: Calcium supplementation for prevention of primary hypertension
Comparison: 1 Calcium supplementation/fortification vs control
Outcome: 24 Final value in diastolic blood pressure by dose

Study or subgroup	Calcium N	Mean(SD)	Control N	Mean(SD)	Weight	Mean Difference IV,Fixed,95% CI
1 Diary calcium intake less than 1000 mg						
Subtotal (95% CI)	**0**		**0**			**Not estimable**
Heterogeneity: not applicable						
Test for overall effect: not applicable						
2 Diary calcium intake 1000 - 1250 mg						
Hilary Green 2000	19	75 (9)	19	76 (9)	5.8 %	-1.00 [-6.72, 4.72]
McCarron 1985	16	75 (9)	16	78 (9)	4.9 %	-3.00 [-9.24, 3.24]
Subtotal (95% CI)	**35**		**35**		**10.7 %**	**-1.91 [-6.13, 2.30]**
Heterogeneity: Chi² = 0.21, df = 1 (P = 0.64); I² =0.0%						
Test for overall effect: Z = 0.89 (P = 0.37)						
3 Diary calcium intake 1500 mg or more						
Davis 1996	17	91.3 (4.7)	17	90.6 (6)	14.5 %	0.70 [-2.92, 4.32]
Johnson 1985	41	78 (8)	40	78 (7)	17.7 %	0.0 [-3.27, 3.27]
Lyle 1987 (1)	27	72.8 (4.8)	27	74.3 (8.6)	13.8 %	-1.50 [-5.21, 2.21]
Lyle 1987 (2)	10	77.1 (4.5)	11	76.7 (7.3)	7.2 %	0.40 [-4.74, 5.54]
Lyle 1992	21	81.8 (4.8)	21	87.3 (6.7)	15.3 %	-5.50 [-9.03, -1.97]
Thomsen 1987	14	77.3 (10.1)	14	78.6 (9.9)	3.5 %	-1.30 [-8.71, 6.11]
Van Beresteyn 1986	29	63.4 (4.8)	29	62 (7.7)	17.4 %	1.40 [-1.90, 4.70]
Subtotal (95% CI)	**159**		**159**		**89.3 %**	**-0.80 [-2.26, 0.65]**
Heterogeneity: Chi² = 9.79, df = 6 (P = 0.13); I² =39%						
Test for overall effect: Z = 1.08 (P = 0.28)						
Total (95% CI)	**194**		**194**		**100.0 %**	**-0.92 [-2.30, 0.46]**
Heterogeneity: Chi² = 10.24, df = 8 (P = 0.25); I² =22%						
Test for overall effect: Z = 1.31 (P = 0.19)						
Test for subgroup differences: Chi² = 0.24, df = 1 (P = 0.63), I² =0.0%						



(1) White men
(2) Black men

Factors that affect strength of a recommendation

GRADE considers

- Quality of evidence
- Uncertainty about the balance of desirable & undesirable effects
- Uncertainty or variability in values & preferences
- Uncertainty whether intervention represents a wise use of resources

Choices for recommendation: *weak* or *strong*

Table 3. Quality assessment (GRADE evidence profile).

No. studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Intervention group (n)	Control group (n)	Effect	Quality	Importance
Emergency department visits for asthma											
1	RCT	No serious limitations	n/a	No serious indirectness	No serious imprecision	n/a	50	50	P = 0.015[g]	Moderate	Critical
Asthma exacerbations (outcomes are number of participants experiencing asthma attacks and wheezing requiring beta2-agonists, use of beta2-agonists (puffs/day), and undefined asthma exacerbation attack)											
6	RCT	Serious limitations[b]	No serious inconsistency	Serious indirectness[c]	No serious imprecision	None	257	250	0.41 (0.27 to 0.63)[f]; P<0.05[g]; no effect[d]	Low	Critical
Asthma symptoms (outcomes are scores (in points) based on ACT, ATAQ for children, daily diary card, ACQ and undefined asthma symptom scores)											
6	RCT	Serious limitations[a]	No serious inconsistency	Serious indirectness[c]	No serious imprecision	None	117	114	No effect[d]; P = 0.01(6mo follow-up)[g]	Low	Critical
Lung function (outcomes are FEV1 (L in 1 sec or % of predicted value) and PEF (mL/min)											
7	RCT	Serious limitations[a]	No serious inconsistency	No serious indirectness	No serious imprecision	None	167	164	0.00 (-3.17 to 3.18)[e]; P<0.001[g]; no effect[d]	Low	Critical
Serum 25(OH)D (nmol/L)											
6	RCT	Serious limitations[a]	Serious inconsistency[h]	No serious indirectness	No serious imprecision	None	117	114	19.66 (5.96 to 33.37)[i]; no effect[d]	Low	Important

Abbreviations: ACT, Asthma Control Test; ATAQ, Asthma Therapy Assessment Questionnaire; ACQ, Asthma Control Questionnaire; FEV1, forced expiratory volume in 1 second; PEF, peak expiratory flow rate.

[a]Unclear allocation concealment, blinding of participants and outcome assessors, accounting of patients and outcome events, and other risk of bias (carryover effects in crossover trial).

[b]Unclear allocation concealment, blinding of participants and outcome assessors, accounting of patients and outcome events.

[c]Differences in interventions and outcomes measured across studies.

[d]Non-significant effect across studies not included in the meta-analysis.

[e]Weighted difference in mean (WMD) change between intervention and control group.

[f]Risk ratio (RR): risk of experiencing asthma exacerbation in the intervention group as compared to the control group.

[g]Not included in the meta-analysis; favours intervention group.

[h]Significant statistical heterogeneity observed based on random effects meta-analysis.

[i]Weighted mean difference (WMD) at end of intervention between intervention and control group.

doi:10.1371/journal.pone.0136841.t003

Recommendations

- ✓ Frame research questions meaningfully
- ✓ Consider whether and how study will contribute to evidence synthesis
- ✓ Emphasize key determinants of study quality (adequate sample size, randomization, allocation concealment, objective measurement, complete follow-up and honest reporting) and of quality of evidence synthesis (study limitations, inconsistency, indirectness, imprecision, reporting biases)
- ✓ Report effect sizes and 95% CI
- ✓ Have analysis and interpretation strategies to account for multiple outcomes

Questions and Comments

djtancredi@ucdavis.edu