ASA 2016 Statement on Statistical Significance – The ASA Six

1. P-values can indicate how incompatible the data are with a specified statistical model.

2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

Ronald Wasserstein & Nicole Lazar, "The ASA's Statement on p-Values: Context, Process, and Purpose," 70 *Am. Stat.* 129 (2016)

4. Proper inference requires full reporting and transparency.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Ronald L. Wasserstein & Nicole A. Lazar, “The ASA's Statement on p-Values: Context, Process, and Purpose,” 70 *Am. Statistician* 129 (2016)

ASA Statement in Litigation

Interpretation in LITIGATION: p-values do not measure anything of importance or interest

State v. Gregory, 427 P.3d 621 (Wash. 2018)
http://schachtmanlaw.com/the-american-statistical-association-statement-on-significance-testing-goes-to-court-part-i/

In re Zoloft Products Liability Litigation, 858 F.3d 787 (3d Cir. 2017)
http://schachtmanlaw.com/expert-witnesses-who-dont-mean-what-they-say/

In re Lipitor Marketing, Sales Pract. & Prods. Liab. Litig., 892 F.3d 624 (4th Cir. 2018)

ASA 2016 – Invitation to “Other Approaches”

“In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. These include methods that emphasize estimation over testing, such as * Bayesian methods… .**

All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct.

Prevalence of Misuses

Is prevalence of misuse a function of prevalence of use?

How often are Bayesian analyses used? Misused?

Consider the Debate over Virus Susceptibility in Microsoft versus Apple OS

[Resolved: Use Linux]

Bayes’ Rule – “Odds” Version

Posterior Odds = Prior Odds x Likelihood Ratio (LR)

$$\frac{P(A|B)}{P(\overline{A}|B)} = \frac{P(A)}{P(\overline{A})} \text{ x } \frac{P(B|A)}{P(B|\overline{A})}$$

A = event of interest

B = new piece of evidence

P = Odds/(1 + Odds).

Odds on the event = p/(1-p) = p/~p

Problems with Bayesian Inference in the Law

Different pieces of evidence are incommensurate in terms of quantitative probabilities.

We have no agreed upon starting point in terms of prior odds.

Lawyers (like scientists) frequently have situations in which we are ignorant of all the evidence and we cannot formulate all the hypotheses that may explain our evidence.

> ***e.g.*****, $P(A) + P(\sim A) << 1$ because we are largely ignorant of all the evidence.**

Risk (or Rate) Ratios

Relative Risks, Odds Ratios, Hazard Ratios, Mortality Ratios, Standardized Mortality Ratio, Standardized Incidence Ratios, Prevalence Ratios, etc.

Risk Ratio = 1.0 = No Difference in Rates Between
Exposed and Unexposed

Risk Ratio > 1.0: Greater Risk Among Exposed

Risk Ratio < 1.0: Less Risk Among Exposed

Relative Risk (Cohort Study)
Measure of association for a 2 x 2 table

EXPOSED	DISEASE	
	YES	NO
YES	a	b
NO	c	d

$$\textbf{Relative Risk} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{\textbf{Incidence among exposed}}{\textbf{Incidence among unexposed}}$$

Odds Ratio (Case-Control Study)
Measure of association for 2 x 2 table

EXPOSED	DISEASE	
	YES (Cases)	NO (Controls)
YES	a	b
NO	c	d

$$\text{Odds Ratio} = \frac{a}{c} \bigg/ \frac{b}{d} = ad / bc$$

Cross-product ratio

Odds Ratios Approximate Relative Risks

$$\mathrm{RR} = \frac{a/(a+b)}{c/(c+d)} \approx \frac{a/b}{c/d} = \frac{ad}{bc} = \mathrm{OR}$$

- When “a” is very small in relation to “b”; and “c” is small compared with “d”
- Odds ratios are always larger than relative risks
- When the risks are small, the odds ratios and relative risks will be almost identical
- Case-control studies are more “efficient”

How do we summarize information?

- **Traditional Approach**
  - Expert opinion (authority based)
  - Narrative review articles, textbook chapters
  - Consensus statements (group expert opinion)
  - GOBSAT

- **New Approach (Systematic reviews)**
  - Evidence-based reviews; comprehensive
  - Explicit quantitative synthesis of ALL evidence

Evolution of Meta-analysis

- 1980-90s – critics become more vociferous; Al Feinstein: “statistical alchemy”
- 1993
- 1999-2000 – consensus statements on procedure & publication QUORUM, PRISMA for RCTs; MOOSE for observational studies
- 2000 forward – used by I.O.M., regulatory agencies
- 1980, 0; 2011 , > 2,000 meta-analyses in PubMed

Goals of Meta-Analysis

- More objective, quantitative summary of evidence
- Enhance precision
- Enhance power
- Answer questions single studies cannot
- Causal inference - Evaluate Hill factors, including strength, consistency, and exposure-response
- Evaluate bias, confounding by stratifying studies, and by subgroup analyses

Product Liability & Envt'l Health Effects Litigation and Regulatory Controversies

- 1980s: PCBs, Bendectin
- 1990s: silicone gel breast implants
- 2000s: fenfluramine, anti-depressants (SSRIs), Baycol, Bextra, Celebrex, Vioxx, HRT, Fosamax, Avandia, Actos, Neurontin, Seroquel, Zyprexa, gadolinium, Trasylol, etc.
- isotretinoin (Accutane)
- testosterone

Weighting Individual Study Results by Inverse Variance

Basic approach: We weight each study included in the meta-analysis by multiplying it by a weight that is the inverse of that study's variance, and then average the weighted studies.

The AVANDIA Debacle – Story About Zeroes & Other Methodological Questions

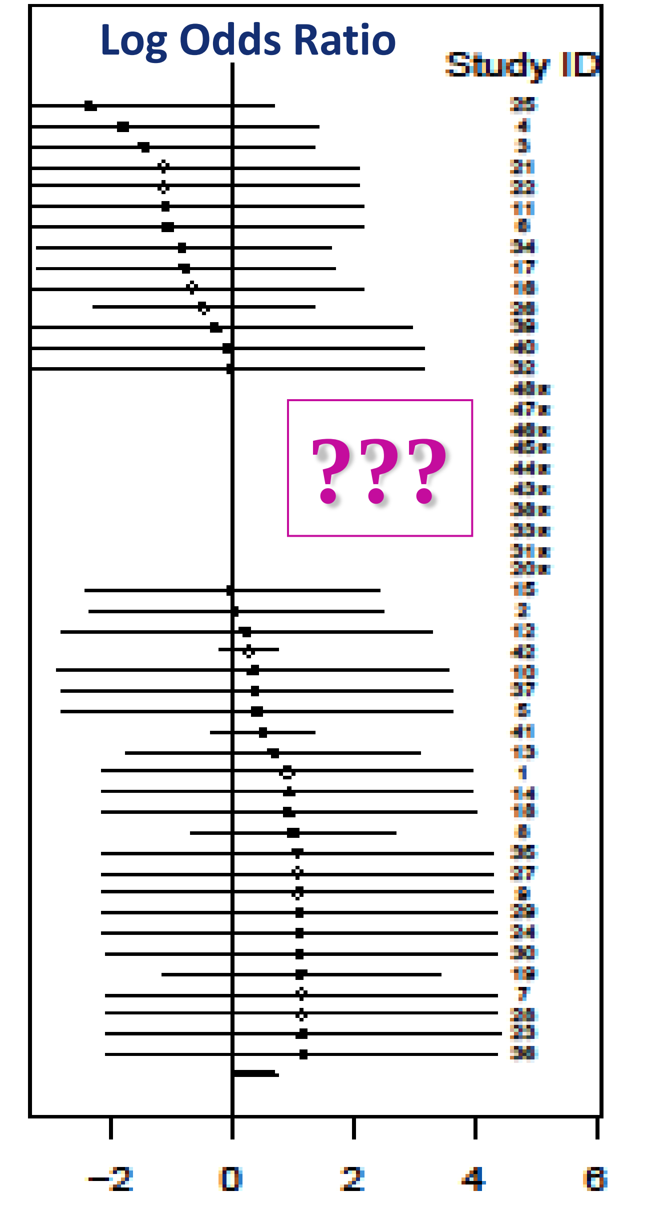
- Rare events
  - How best to incorporate studies with 0/0 events? And 0/1 events?
  - Validity of asymptotic inference?
    - Should an exact method be used?
- Choice of effect measure?
- Fixed versus Random Effects Models
  - How best to capture between-study heterogeneity?

Did Exclusion of Zero-Event Trials Biased Nissen's Study?

- "[T]he exclusion of zero total event trials from meta-analyses increases the effect size compared to meta-analyses that include these trials."
- Zero-event trials "provide relevant data by showing that event rates for both the intervention and control groups are low and relatively equal."

Friedrich, et al., "Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data," 7 *BMC Med. Research Methodology* 5 (2007)

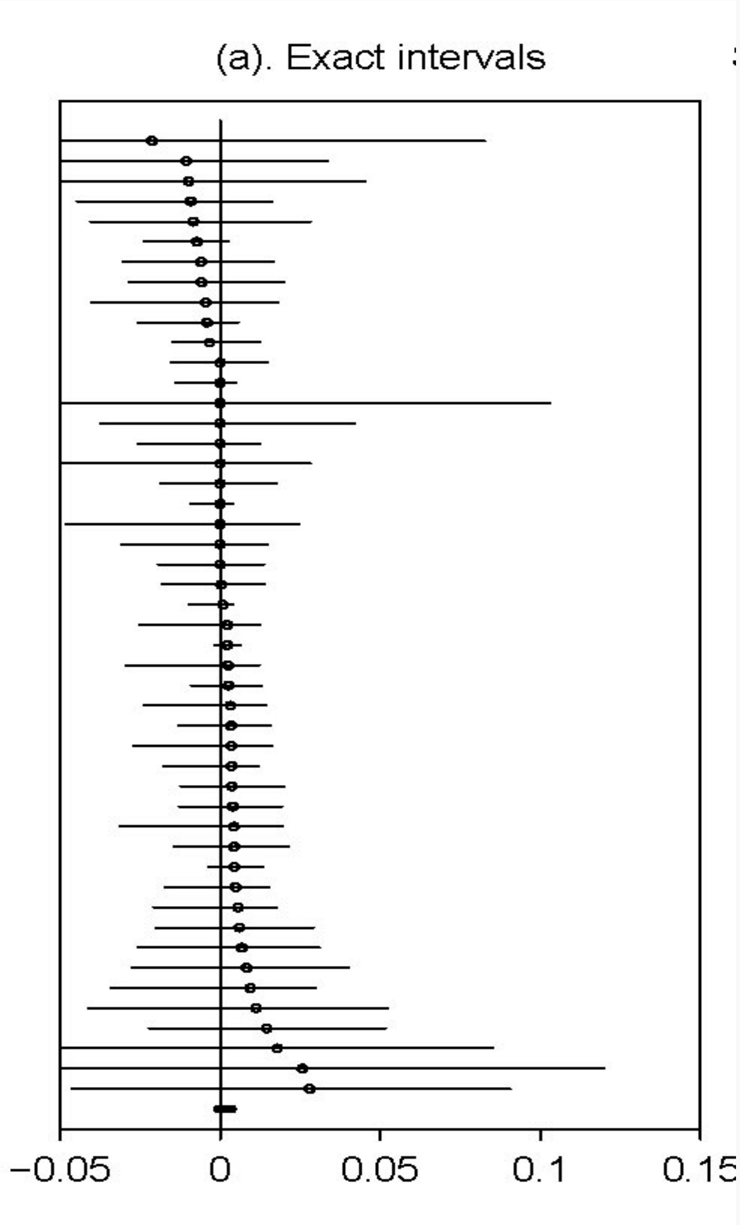
Myocardial Infarction Σ Peto OR in the “Nissen RCTs”



What is the story about nothing?

Σ OR = 1.43 (95% CI, 1.03, 1.98); p-value = 0.03

Lee-Jen Wei's Risk Difference – M.I.



95% CIs of risk difference for heart attack (RSG – Control)

48 studies

(Small circles are the observed risk differences)

$\hat{\Delta} = 0.18\%$

Exact 95% CI (-0.08, 0.38), p = 0.27

Tian, *et al*., "Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 × 2 tables with all available data but without artificial continuity correction" 10 *Biostatistics* 275 (2008)

Methodology Matters

"As rosiglitazone case demonstrates, minor modifications of the meta-analysis protocol can change the statistical significance of the result. For small effects, even the direction of the treatment effect estimate may change."

Adrian V. Hernandez, Esteban Walker, John P.A. Ioannidis, and Michael W. Kattan, "Challenges in meta-analysis of randomized clinical trials for rare harmful cardiovascular events: the case of rosiglitazone," 156 *Am. Heart J.* 23, 28 (2008)

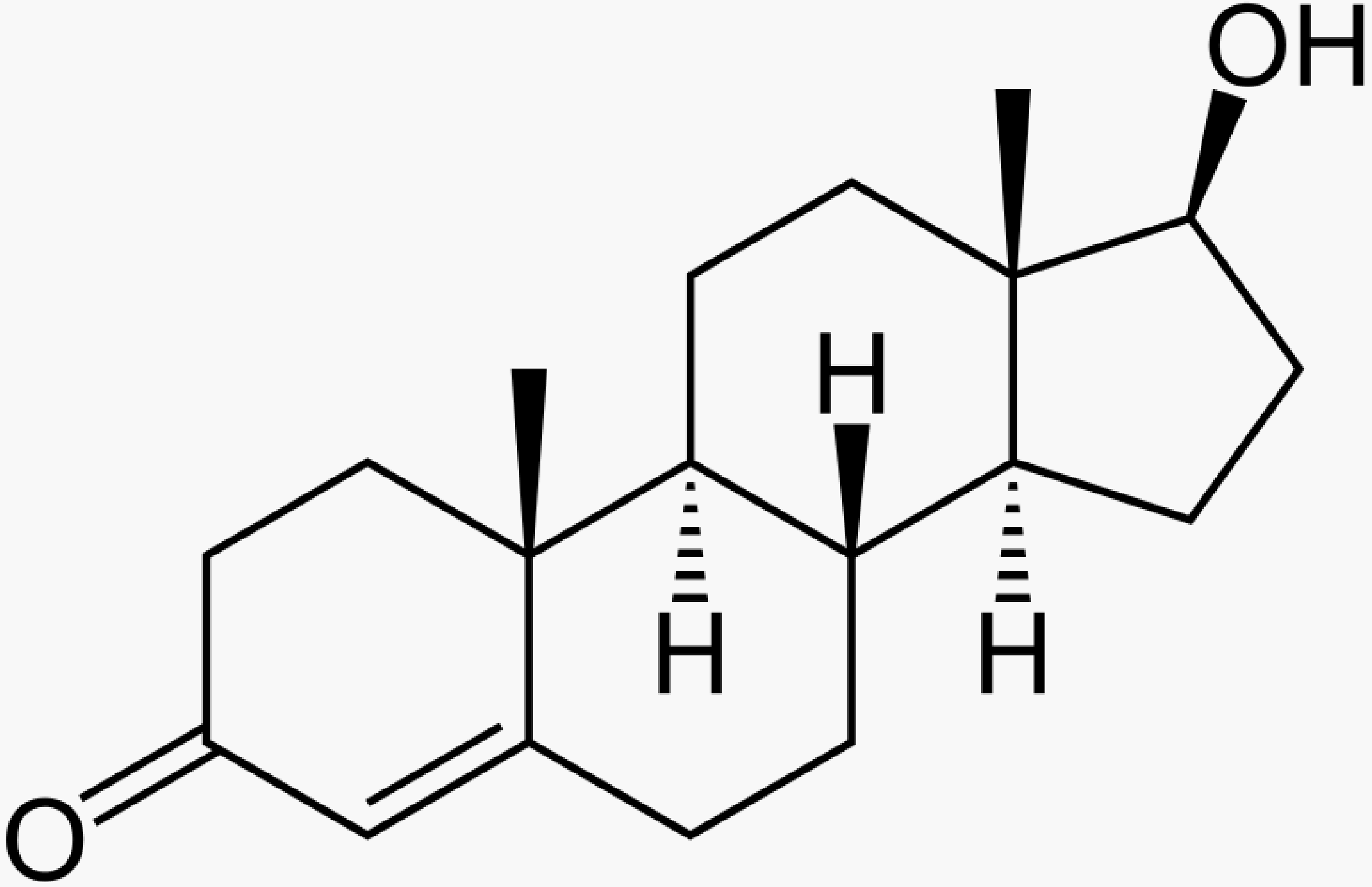
The RECORD Trial – Mega-Trial

- In progress when FDA imposed heightened warnings and an REMS
- RECORD was a randomized CV outcome RCT in over 4,400 patients followed for 5.5 yrs (avg.)
- non-inferiority trial to rule out a 20% increased risk when compared with standard-of-care diabetes treatment
- study met its primary objective:

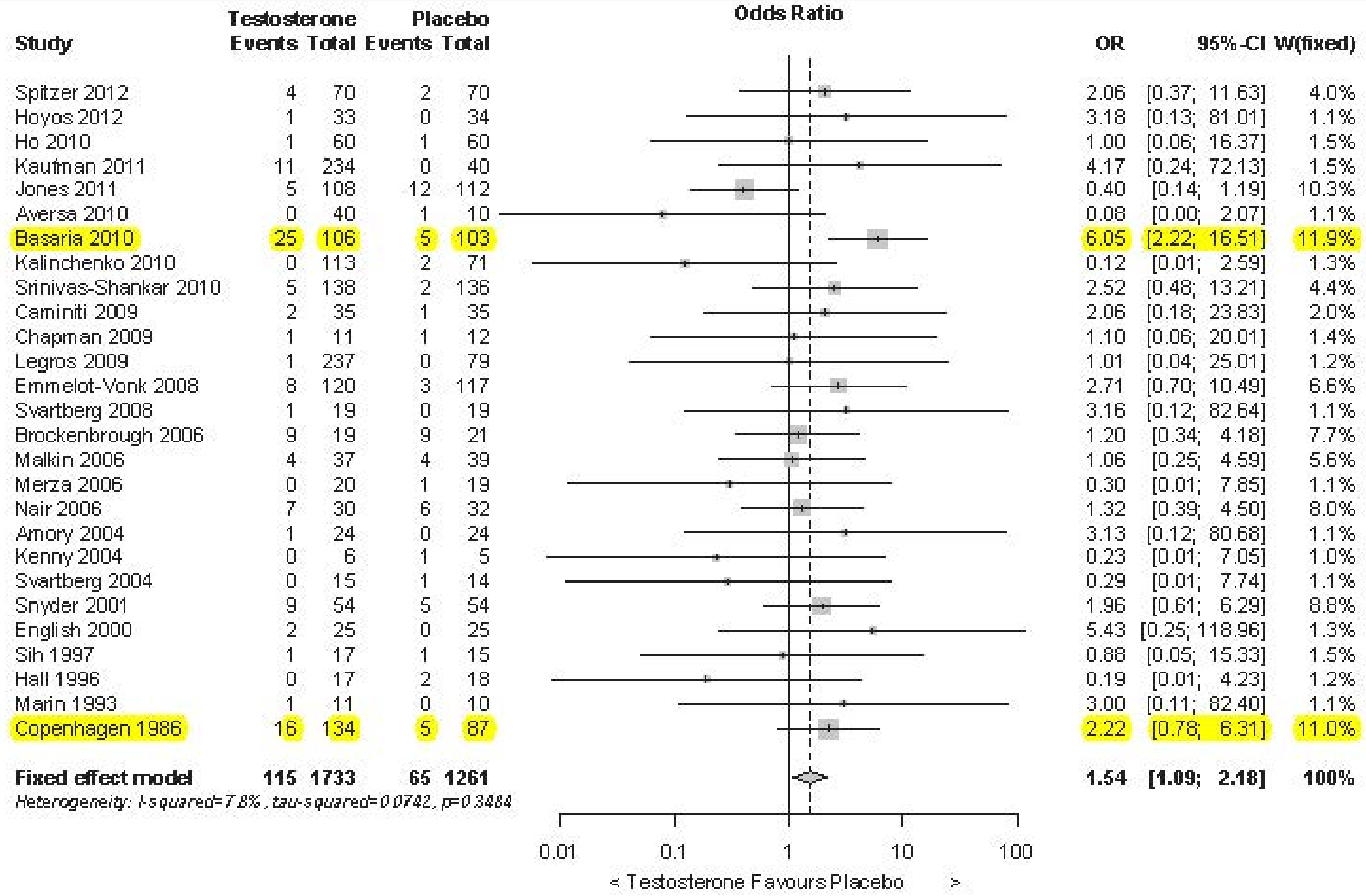
  CV Death / CV Hospitalization: 0.99 (0.85-1.16)

http://schachtmanlaw.com/pharmacovigilantism-avandia-litigation/
http://schachtmanlaw.com/learning-to-embrace-flawed-evidence-the-avandia-mdls-daubert-opinion/

Testosterone Therapy



Xu’s Meta-Analysis of Composite CV Events in Testosterone RCTs (2013)



Study	Testosterone Events	Testosterone Total	Placebo Events	Placebo Total	OR	95%-CI	W(fixed)
Spitzer 2012	4	70	2	70	2.06	[0.37; 11.63]	4.0%
Hoyos 2012	1	33	0	34	3.18	[0.13; 81.01]	1.1%
Ho 2010	1	60	1	60	1.00	[0.06; 16.37]	1.5%
Kaufman 2011	11	234	0	40	4.17	[0.24; 72.13]	1.5%
Jones 2011	5	108	12	112	0.40	[0.14; 1.19]	10.3%
Aversa 2010	0	40	1	10	0.08	[0.00; 2.07]	1.1%
Basaria 2010	25	106	5	103	6.05	[2.22; 16.51]	11.9%
Kalinchenko 2010	0	113	2	71	0.12	[0.01; 2.59]	1.3%
Srinivas-Shankar 2010	5	138	2	136	2.52	[0.48; 13.21]	4.4%
Caminiti 2009	2	35	1	35	2.06	[0.18; 23.83]	2.0%
Chapman 2009	1	11	1	12	1.10	[0.06; 20.01]	1.4%
Legros 2009	1	237	0	79	1.01	[0.04; 25.01]	1.2%
Emmelot-Vonk 2008	8	120	3	117	2.71	[0.70; 10.49]	6.6%
Svartberg 2008	1	19	0	19	3.16	[0.12; 82.64]	1.1%
Brockenbrough 2006	9	19	9	21	1.20	[0.34; 4.18]	7.7%
Malkin 2006	4	37	4	39	1.06	[0.25; 4.59]	5.6%
Merza 2006	0	20	1	19	0.30	[0.01; 7.85]	1.1%
Nair 2006	7	30	6	32	1.32	[0.39; 4.50]	8.0%
Amory 2004	1	24	0	24	3.13	[0.12; 80.68]	1.1%
Kenny 2004	0	6	1	5	0.23	[0.01; 7.05]	1.0%
Svartberg 2004	0	15	1	14	0.29	[0.01; 7.74]	1.1%
Snyder 2001	9	54	5	54	1.96	[0.61; 6.29]	8.8%
English 2000	2	25	0	25	5.43	[0.25; 118.96]	1.3%
Sih 1997	1	17	1	15	0.88	[0.05; 15.33]	1.5%
Hall 1996	0	17	2	18	0.19	[0.01; 4.23]	1.2%
Marin 1993	1	11	0	10	3.00	[0.11; 82.40]	1.1%
Copenhagen 1986	16	134	5	87	2.22	[0.78; 6.31]	11.0%
Fixed effect model	**115**	**1733**	**65**	**1261**	**1.54**	**[1.09; 2.18]**	**100%**

Heterogeneity: I-squared=7.8%, tau-squared=0.0742, p=0.3484

Figure 3 Forest plots of placebo-controlled randomized trials examining the pooled effect of testosterone therapy on cardiovascular-related events.

Rent-Seeking Lobbies

Citing 2 observational studies (Vigen; Finkle), and the Xu meta-analysis, Public Citizen petitioned for a boxed warning.

FDA rejected the petition, but called AdComm meeting in 2014.

AdComm made no finding of CV harm, but…..

“The benefit and safety of these medications have not been established for the treatment of low testosterone levels due to aging, even if a man’s symptoms seem related to low testosterone.”

FDA required rewording of the indication and a warning “about a *possible* increased risk of heart attacks and strokes in patients taking testosterone.”

FDA called for large, long-term RCT.

And then came the litigation…

And with it, courtroom science.

Welcome:

the Ill-Informed Vague & Non-Informative Prior Probability Distribution

Carlin’s Bayesian Meta-Analysis

Applied DerSimonian & Laird’s “random effects” approach to a Bayesian model

Chose a prior distribution centered on ln [OR 1.0], inverse gamma (0.0001, 0.0001) with variance of 1,000

Deployed MCMC algorithm for to arrive at posterior probability distribution

Carlin’s Bayesian Meta-Analysis

Carlin used “vague” priors;

Assumed normal sampling variance for measurements.

“The assumption of normally distributed observed values with known variance is likely to be reasonable in most situations, as long as the studies are large and *observed counts are not too small*.”

“The validity of an assumption of a normal prior distribution for the true effects is more difficult to assess … .”

“A study of the sensitivity of conclusions to the choice of prior would be important.”

Carlin at 157 (emphasis added)

Carlin’s Use of Vague Priors

Carlin believed non-informative or locally uniform priors were reasonable assumptions for meta-analyses, given the belief that the likelihood function has “appreciable magnitude.”

-- Even a small number of studies, say ~10 would combine to “become relatively informative.”

Carlin deployed two sample meta-analyses, both of which involved a large number of studies, with each study containing large event sizes for contributing 2x2 tables

Carlin at 146

Plaintiffs' Application of Carlin's Method to TRT Meta-Analysis

Xu (2013) discredited and out of date; so Plaintiffs selected later (but not latest) meta-analysis: Morley & Albert (2016) (data supplement with count data provided)

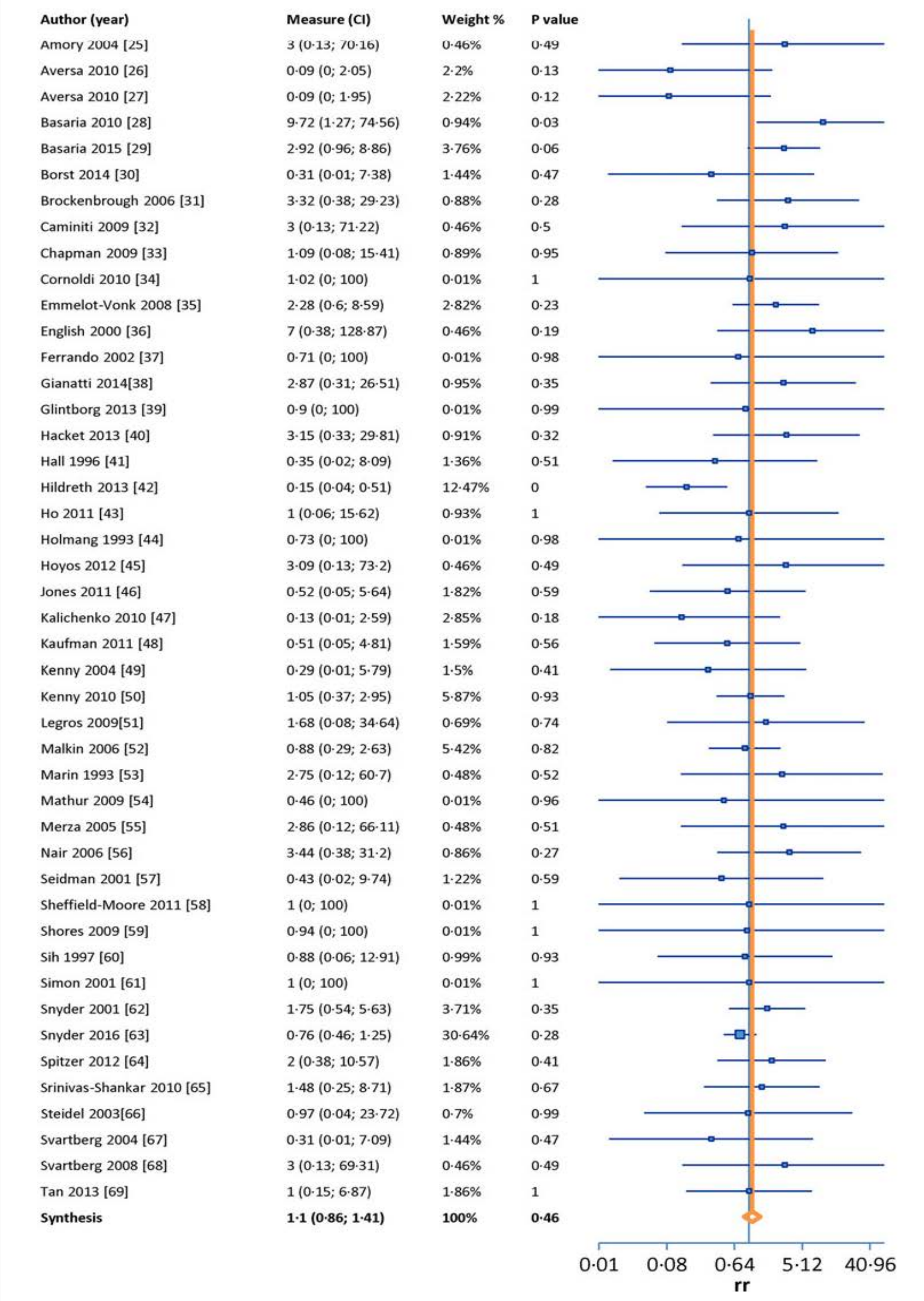
Recalculated an OR 1.4, 95% confidence interval 0.78 - 2.72 for a composite end point of "stroke or heart attack"

Selected prior of 0 (ln OR) with wide, flat distribution - variance = 1,000, inverse gamma (.001, .001)

assumed normal prior for super population mean, and normal distribution of the effect sizes

Using Carlin's Bayesian meta-analysis methodology, the posterior probability of the O.R. > 1.0 is 0.8551

Albert & Morley (2016)



Author (year)	Measure (CI)	Weight %	P value
Amory 2004 [25]	3 (0·13; 70·16)	0·46%	0·49
Aversa 2010 [26]	0·09 (0; 2·05)	2·2%	0·13
Aversa 2010 [27]	0·09 (0; 1·95)	2·22%	0·12
Basaria 2010 [28]	9·72 (1·27; 74·56)	0·94%	0·03
Basaria 2015 [29]	2·92 (0·96; 8·86)	3·76%	0·06
Borst 2014 [30]	0·31 (0·01; 7·38)	1·44%	0·47
Brockenbrough 2006 [31]	3·32 (0·38; 29·23)	0·88%	0·28
Caminiti 2009 [32]	3 (0·13; 71·22)	0·46%	0·5
Chapman 2009 [33]	1·09 (0·08; 15·41)	0·89%	0·95
Cornoldi 2010 [34]	1·02 (0; 100)	0·01%	1
Emmelot-Vonk 2008 [35]	2·28 (0·6; 8·59)	2·82%	0·23
English 2000 [36]	7 (0·38; 128·87)	0·46%	0·19
Ferrando 2002 [37]	0·71 (0; 100)	0·01%	0·98
Gianatti 2014[38]	2·87 (0·31; 26·51)	0·95%	0·35
Glintborg 2013 [39]	0·9 (0; 100)	0·01%	0·99
Hacket 2013 [40]	3·15 (0·33; 29·81)	0·91%	0·32
Hall 1996 [41]	0·35 (0·02; 8·09)	1·36%	0·51
Hildreth 2013 [42]	0·15 (0·04; 0·51)	12·47%	0
Ho 2011 [43]	1 (0·06; 15·62)	0·93%	1
Holmang 1993 [44]	0·73 (0; 100)	0·01%	0·98
Hoyos 2012 [45]	3·09 (0·13; 73·2)	0·46%	0·49
Jones 2011 [46]	0·52 (0·05; 5·64)	1·82%	0·59
Kalichenko 2010 [47]	0·13 (0·01; 2·59)	2·85%	0·18
Kaufman 2011 [48]	0·51 (0·05; 4·81)	1·59%	0·56
Kenny 2004 [49]	0·29 (0·01; 5·79)	1·5%	0·41
Kenny 2010 [50]	1·05 (0·37; 2·95)	5·87%	0·93
Legros 2009[51]	1·68 (0·08; 34·64)	0·69%	0·74
Malkin 2006 [52]	0·88 (0·29; 2·63)	5·42%	0·82
Marin 1993 [53]	2·75 (0·12; 60·7)	0·48%	0·52
Mathur 2009 [54]	0·46 (0; 100)	0·01%	0·96
Merza 2005 [55]	2·86 (0·12; 66·11)	0·48%	0·51
Nair 2006 [56]	3·44 (0·38; 31·2)	0·86%	0·27
Seidman 2001 [57]	0·43 (0·02; 9·74)	1·22%	0·59
Sheffield-Moore 2011 [58]	1 (0; 100)	0·01%	1
Shores 2009 [59]	0·94 (0; 100)	0·01%	1
Sih 1997 [60]	0·88 (0·06; 12·91)	0·99%	0·93
Simon 2001 [61]	1 (0; 100)	0·01%	1
Snyder 2001 [62]	1·75 (0·54; 5·63)	3·71%	0·35
Snyder 2016 [63]	0·76 (0·46; 1·25)	30·64%	0·28
Spitzer 2012 [64]	2 (0·38; 10·57)	1·86%	0·41
Srinivas-Shankar 2010 [65]	1·48 (0·25; 8·71)	1·87%	0·67
Steidel 2003[66]	0·97 (0·04; 23·72)	0·7%	0·99
Svartberg 2004 [67]	0·31 (0·01; 7·09)	1·44%	0·47
Svartberg 2008 [68]	3 (0·13; 69·31)	0·46%	0·49
Tan 2013 [69]	1 (0·15; 6·87)	1·86%	1
Synthesis	**1·1 (0·86; 1·41)**	**100%**	**0·46**

Fig. 2 Forest plot of relative risk for cardiovascular events of trials of testosterone supplementation.

RR =1.10 (95% CI 0.86; 1.41, P = 0.45)

Stewart G. Albert & John E. Morley, Testosterone therapy, association with age, initiation and mode of therapy with cardiovascular events: a systematic review, 85 *Clin Endocrinol* 436, 438 (2016)

Rule 702: Testimony by Expert Witnesses

A witness ***qualified * may testify *** *if:***

(a) the expert's * knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;**

(b) the testimony is based on sufficient facts or data;

(c) the testimony is the product of reliable principles and methods; and

(d) the expert has reliably applied the principles and methods to the facts of the case.

http://www.law.cornell.edu/rules/fre/rule_702 (effective Dec. 2011)

Rule 702: Rx for Pathology of Science & Knowledge

(Patho-epistemology – Specious Claiming)

- Analytical Gaps
- Ipse Dixit
- Logical Lacunae
- Cherry Picking
- Fallacies and Non-Sequiturs
- Misapplied methodologies

AbbVie's "*Daubert*" Challenge

- plaintiffs' witnesses' failed to publish their analysis;
- one witness's never published Bayesian analysis before;
- absence of Bayesian analyses in the relevant RCTs on TRT;
- rarity of Bayesian analyses in product liability cases;
- witnesses' failure to state what the actual risk was, as opposed to the probability that it exceeded 1.0; and
- **defense witness's calculation that the claimed CV risk met "only a 70% level of evidence, which is far below the 95% level required."**

Judicial Reaction – Motion Denied

Witnesses had adequate expertise in Bayesian methods.

RMSE states that Bayesian methods are used by a "well-established minority"

AbbVie failed to show how the Bayesian methods were subjective in a way that the frequentist methods were not

AbbVie had conflated 95% probability of coefficient of confidence with posterior probability – citing Richard Morey

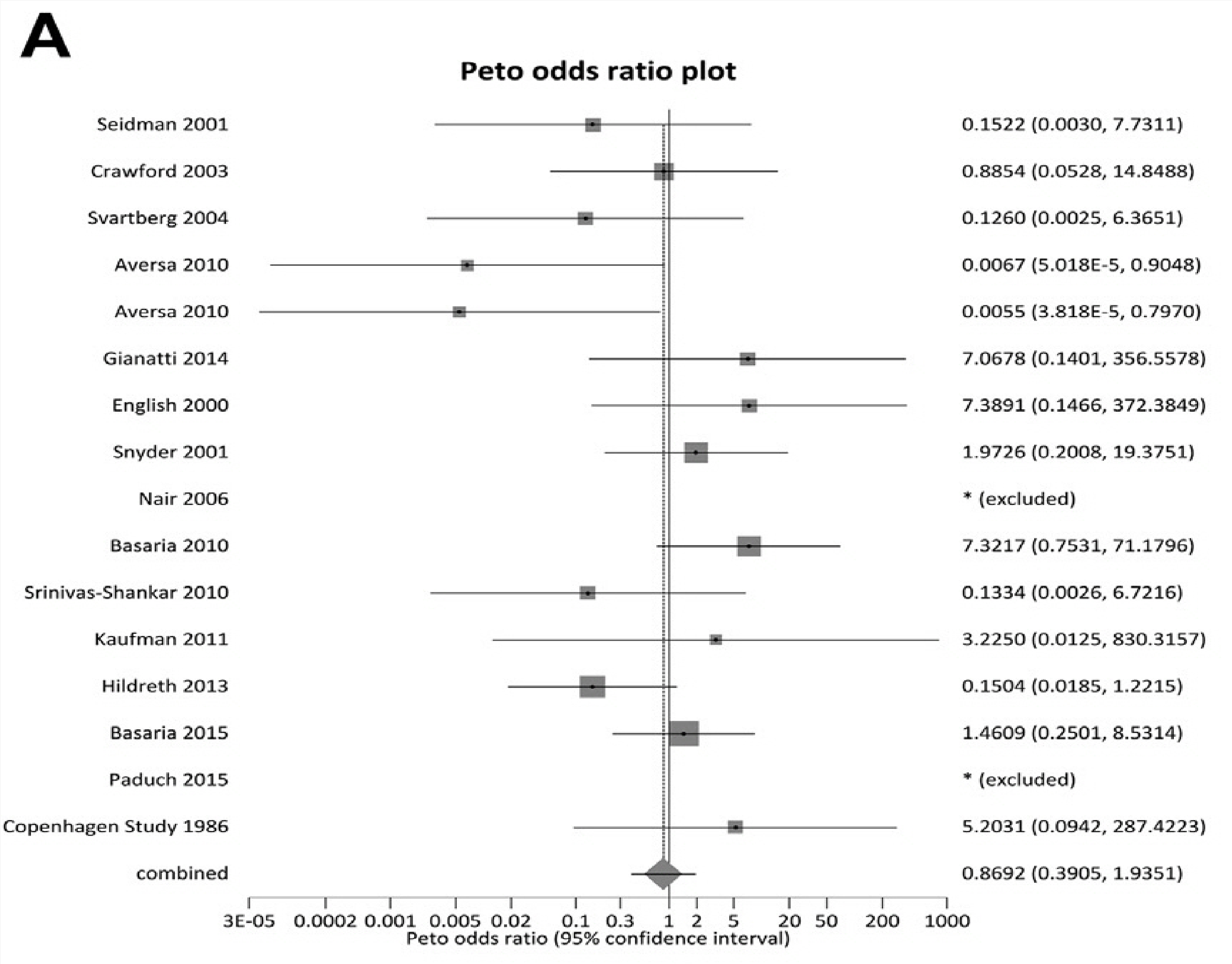
In re Testosterone Replacement Therapy Prods. Liab. Litig., MDL No. 2545, C.M.O. No. 46, 2017 WL 1833173 at *11-12 (N.D. Ill. May 8, 2017); *id*. at MDL No. 2545, 2018 WL 4030585, *8 (N.D.Ill. Aug. 23, 2018) (reaffirming prior ruling for Dr. Martin Wells)

Problems Ignored by AbbVie

1. Composite End Point

Plaintiffs’ expert witnesses ignored standard epidemiologic practice in failing to present constituent end point results.

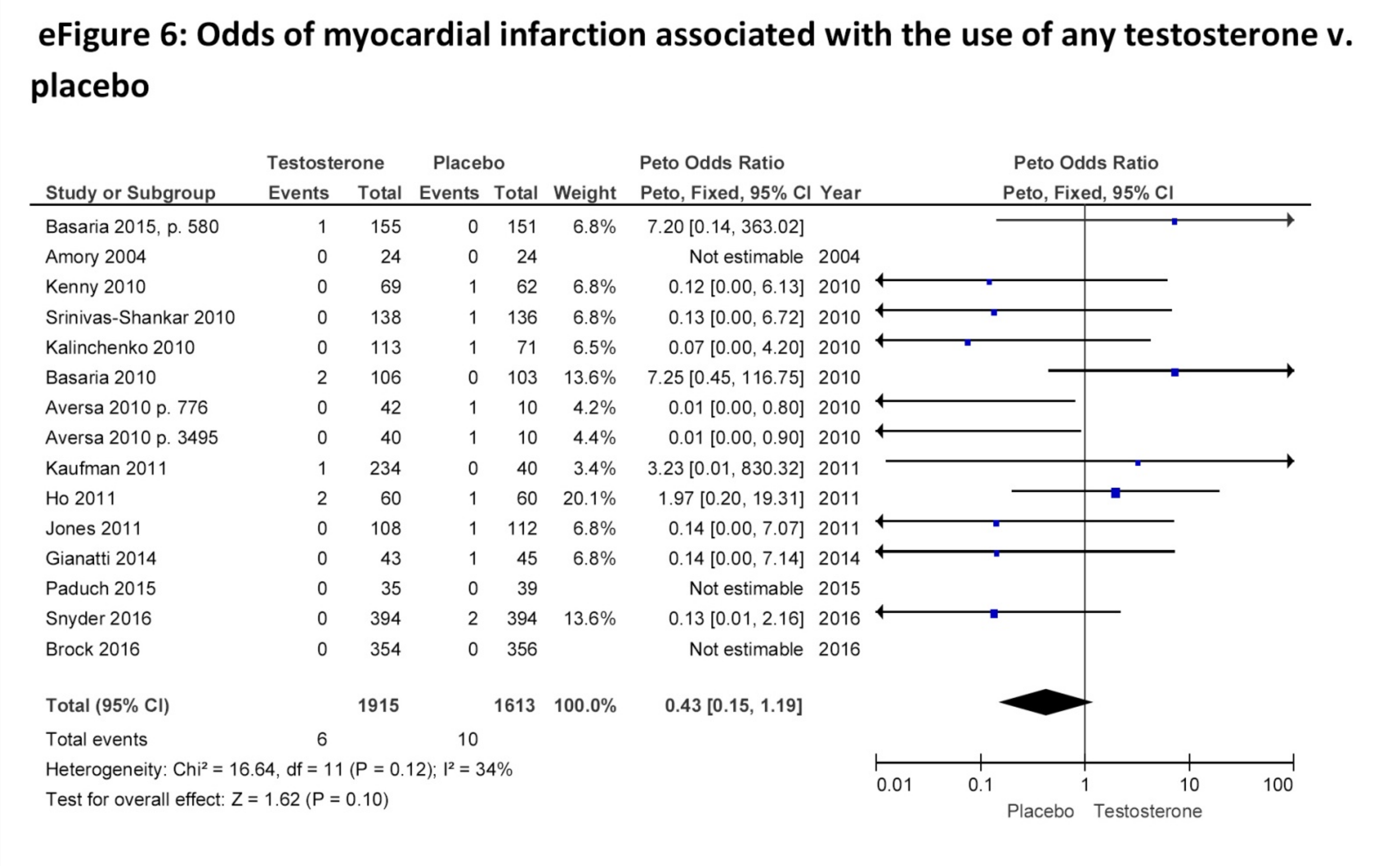
Alexander (2016) – Heart Attack Only



M.I. ΣOR = 0.87 (95% C.I., 0.39 – 1.93)

G. Caleb, et al., “Cardiovascular Risks of Exogenous Testosterone Use Among Men: A Systematic Review and Meta-Analysis,” 130 *Am. J. Med*. 293 (2017)

Elliott (2017) meta-analysis in hypogonadal men
M.I. ΣO.R. = (0.43, 95% C.I. 0.15-1.19)



eFigure 6: Odds of myocardial infarction associated with the use of any testosterone v. placebo

	Testosterone		Placebo			Peto Odds Ratio	
Study or Subgroup	Events	Total	Events	Total	Weight	Peto, Fixed, 95% CI	Year
Basaria 2015, p. 580	1	155	0	151	6.8%	7.20 [0.14, 363.02]	
Amory 2004	0	24	0	24		Not estimable	2004
Kenny 2010	0	69	1	62	6.8%	0.12 [0.00, 6.13]	2010
Srinivas-Shankar 2010	0	138	1	136	6.8%	0.13 [0.00, 6.72]	2010
Kalinchenko 2010	0	113	1	71	6.5%	0.07 [0.00, 4.20]	2010
Basaria 2010	2	106	0	103	13.6%	7.25 [0.45, 116.75]	2010
Aversa 2010 p. 776	0	42	1	10	4.2%	0.01 [0.00, 0.80]	2010
Aversa 2010 p. 3495	0	40	1	10	4.4%	0.01 [0.00, 0.90]	2010
Kaufman 2011	1	234	0	40	3.4%	3.23 [0.01, 830.32]	2011
Ho 2011	2	60	1	60	20.1%	1.97 [0.20, 19.31]	2011
Jones 2011	0	108	1	112	6.8%	0.14 [0.00, 7.07]	2011
Gianatti 2014	0	43	1	45	6.8%	0.14 [0.00, 7.14]	2014
Paduch 2015	0	35	0	39		Not estimable	2015
Snyder 2016	0	394	2	394	13.6%	0.13 [0.01, 2.16]	2016
Brock 2016	0	354	0	356		Not estimable	2016
Total (95% CI)		**1915**		**1613**	**100.0%**	**0.43 [0.15, 1.19]**	
Total events	6		10				

Heterogeneity: Chi² = 16.64, df = 11 (P = 0.12); I² = 34%

Test for overall effect: Z = 1.62 (P = 0.10)

Jesse Elliott, Shannon E Kelly, Adam C Millar, Joan Peterson, Li Chen, Amy Johnston, Ahmed Kotb, Becky Skidmore, Zemin Bai, Muhammad Mamdani & George A Wells, "Testosterone therapy in hypogonadal men: a systematic review and network meta-analysis," 7 *BMJ Open* e015284 (2017)

Problems Ignored by AbbVie

2. Ignores Carlin’s caveats about sparse events.

Mathematical model, selection of prior, failed to represent reality and our pre-analysis state of knowledge

Carlin’s method assumes: sampling variability of individual effect estimates is normally distributed

Holds asymptotically for estimated log odds ratio

Actual sampling distribution can depart substantially from normal when event counts are low
{0-0; 0-1; 1-0; 2-1; 2-1, etc.}

Ill-Informed Non-Informative Prior

3. The Vague Prior

Does vagueness represent our current state of knowledge?

Or is it merely a mathematical convenience to start the Bayesian analysis?

Before starting the analysis, 0.3 < OR < 3.0

But variance of 1,000 around OR = 1.0, with wide, flat distribution suggests that:

TRT prevents heart attacks, to TRT causes all heart attacks.

Ill-Informed Non-Informative Prior

Vague prior with sparse data virtually eliminates "shrinkage"

Conveniently yields a 95% credible [sic] interval that closely resembles the 95% confidence
Interval

Posterior probability distribution was not credible

FDA and other “experts” had called for a large RCT, which could not be undertaken if the Plaintiffs’ posterior probability was correct, or even approximately correct

“Weak Priors Are Most Absurd…”

“Weak priors are most absurd when all the values of θ being debated (and thus all serious candidates for a prior median) are so close to the null that it is difficult to distinguish among them or distinguish them from the null. Many, if not most, modern controversies (eg, long-term nutrient effects and drug side effects) involve debates within ranges like 1⁄4 < RR < 4.”

Sander Greenland & Charles Poole, “Living with P Values: Resurrecting a Bayesian Perspective on Frequentist Statistics,” 24 *Epidem*. 62, 65 (2013) (internal citations omitted)

Ill-Informed Non-Informative Prior

“The default conclusion from a noninformative prior analysis will almost invariably put too much probability on extreme values. A vague prior distribution assigns much of its probability on values that are never going to be plausible, and this disturbs the posterior probabilities more than we tend to expect—something that we probably do not think about enough in our routine applications of standard statistical methods.”

Andrew Gelman, “P-Values and Statistical Practice,” 24 Epidem. 69, 72 (2013)

Ill-Informed Non-Informative Prior

“consider what would happen if we routinely interpreted one-sided P values as posterior probabilities. In that case, an experimental result that is 1 standard error from zero—that is, exactly what one might expect from chance alone—would imply an 83% posterior probability that the true effect in the population has the same direction as the observed pattern in the data at hand. It does not make sense to me to claim 83% certainty—5-to-1 odds—based on data that not only could occur by chance alone but in fact represent an expected level of discrepancy.”

Andrew Gelman, “P-Values and Statistical Practice,” 24 *Epidem*. 69, 72 (2013)

Deconstruction

Q. Let me ask you some hypotheticals.· If you had done a Bayesian meta-analysis in the way you did and you got a summary point estimate of 1.0 -- it came in just amazingly … at exactly at 1.0 -- and the distribution was roughly symmetrical. All right? --

A. All right.

Q. -- then, there would be a 50 percent [probability] of an increased risk, correct? --

A. Right.

Q. -- and there would be a 50 percent probability of a decreased risk?

A. Yes.

Martin Wells Deposition in *Martin v. Actavis, Inc*., 178:9-15, 178:20-25 (N.D. Ill. April 2, 2018)

Q. Well -- let me change the hypothetical. Same hypothetical with respect to the 1.0 for the first·round of your meta-analysis, and now, you have to add a crummy little clinical trial and there’s an imbalance of events; there’s one in the placebo group and two in the testosterone group.

If you add those numbers into your prior Bayesian meta-analysis – I’ve given you a summary point estimate of 1.0 -- now, the new summary point estimate will be north of 1.0, correct?
A. Slightly north.
****
Q. And it will now be greater than 50 percent probability in the Bayesian hypothesis test that there is an increased risk?
A. Right.

Deposition at 179:22 - 180:23

Q. [I]f you did a Bayesian analysis of just the myocardial infarction endpoint in Elliott [meta-analysis with ΣOR = 0.43], you almost certainly would have gotten a summary point estimate below one, correct?

A. Yes.

***

Q. And if you then did the analysis that you did in terms of a Bayesian hypothesis test, you would get a probability greater than 50 percent that myocardial infarction risk was decreased in men who got testosterone if they were hypogonadal, correct?

A. Yes.

Dep. at 182:2; 182:15-21

Discrepancy between Calculated Posterior and Actual Probability of Claim

- myriad potential biases (systematic errors) in data collection
- departures from random sampling
- protocol violations
- sampling the wrong population
- measuring the wrong variables
- other threats to study's internal validity (improper design or implementation)
- credibility of study participants
- credibility of study researchers
- errors in measurements or in data collection

More Factors

- errors in data reporting, categorization, cleaning, and handling
- failure to validate statistical models
- error in selecting appropriate statistical test
- external validity of the study
- errors by authors in transcribing, describing data and analyses
- other available studies, and their respective data and analysis factors

If these largely independent factors each had a probability or 95%, which would be high, the conjunction of their probabilities would put the probability of the "true value" well below 50%.