Can we trust claims from clinical trial reports?

A statistician's perspective

Stuart Pocock

London School of Hygiene and Tropical Medicine

Reporting of a Major Randomized Trial

company press release?

⇓

conference presentation

⇓

journal publication

⇓

regulatory submission to FDA, EMA

⇓

company advertising

Assurance re Quality of Reports

journals: peer review, CONSORT guidelines

regulators: totality of evidence, ICH guidelines

Internal validity

Randomisation OK? often inadequately reported

Masking (blinding) implemented OK?

Size: were enough patients included?

Patient follow-up: problems of drop-outs, non-compliance

Results: correct analysis? emphasis on pre-defined aims?
clear and informative presentation?

Conclusions: compatible with results?
balanced account of efficacy **and side-effects**
limitations assessed

External validity

Relevant patients

beware of inappropriate extrapolations

Treatment regimens: implementable and cost effective

appropriate control group

Outcome measures: relevant to patient well-being

treatment and follow-up long enough

Other evidence: adequate account of other trials

biological rationale

constructive critical appraisal

no trial is perfect!

Statistics in trial reports

beware of **positive spin**: post hoc emphasis on good news

Andrew Lang 1844-1912

"He uses statistics as a drunken man uses a lamp-post: for support rather than illumination"

obsession with $P<.05 \Rightarrow$ a "positive" trial

lots of data to play with $\Rightarrow$ data dredging

disappointing result: try data dredging?!

BEAUTIFUL trial [Lancet 2008; 372 p 807]
ivabradine in 10,917 patients with stable coronary disease

primary composite outcome: CV death, MI, heart failure
median 19 months follow-up

	ivabradine	**placebo**	
primary	15.4%	15.3%	P=.94
subgroup with heart rate ≥ 70:			
primary	17.2%	18.5%	P=.17
myocardial infarction	**3.1%**	**4.9%**	**P=.001**

hypothesis generating, but highlighted in abstract

a short history of the P-value

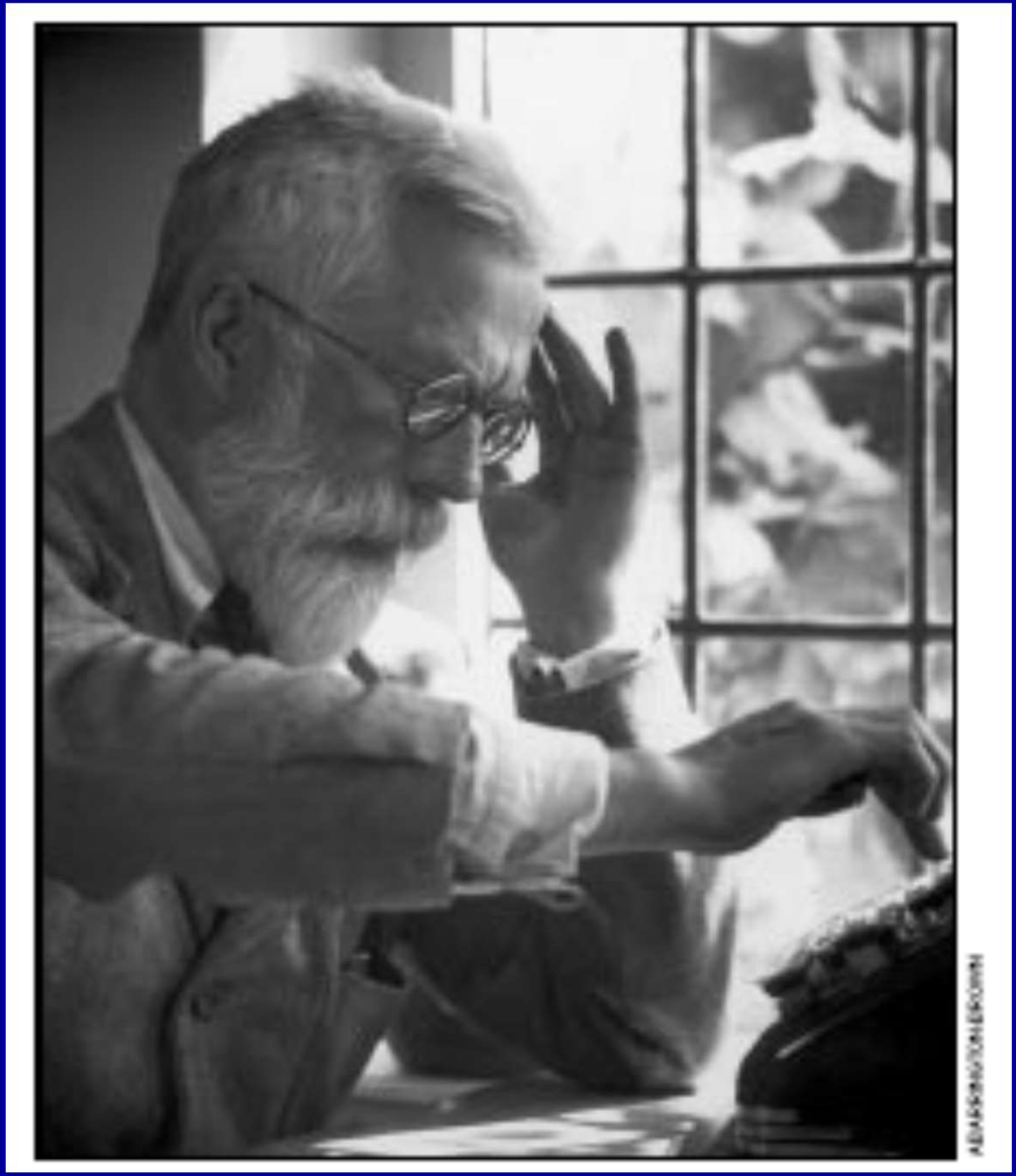
R A Fisher (1925)
Statistical Methods for Research Workers

introduced hypothesis testing and P-values

tables for specially selected values of P
eg. P=.1 .05 .02 .01 .001
but no particular emphasis on .05

"when P is between .02 and .05 the result must be judged significant but barely so. The data do not demonstrate this point beyond reasonable doubt"

P<.05 is **not** strong evidence, P<.001 **is**



R A Fisher, the founder of statistical inference, working on a mechanical calculator

Jerzy Neyman and Egon Pearson (1933)

Philosophical Transactions of the Royal Society

hypothesis testing formulation

null hypothesis versus alternative hypothesis

P<.05 **reject** null hypothesis

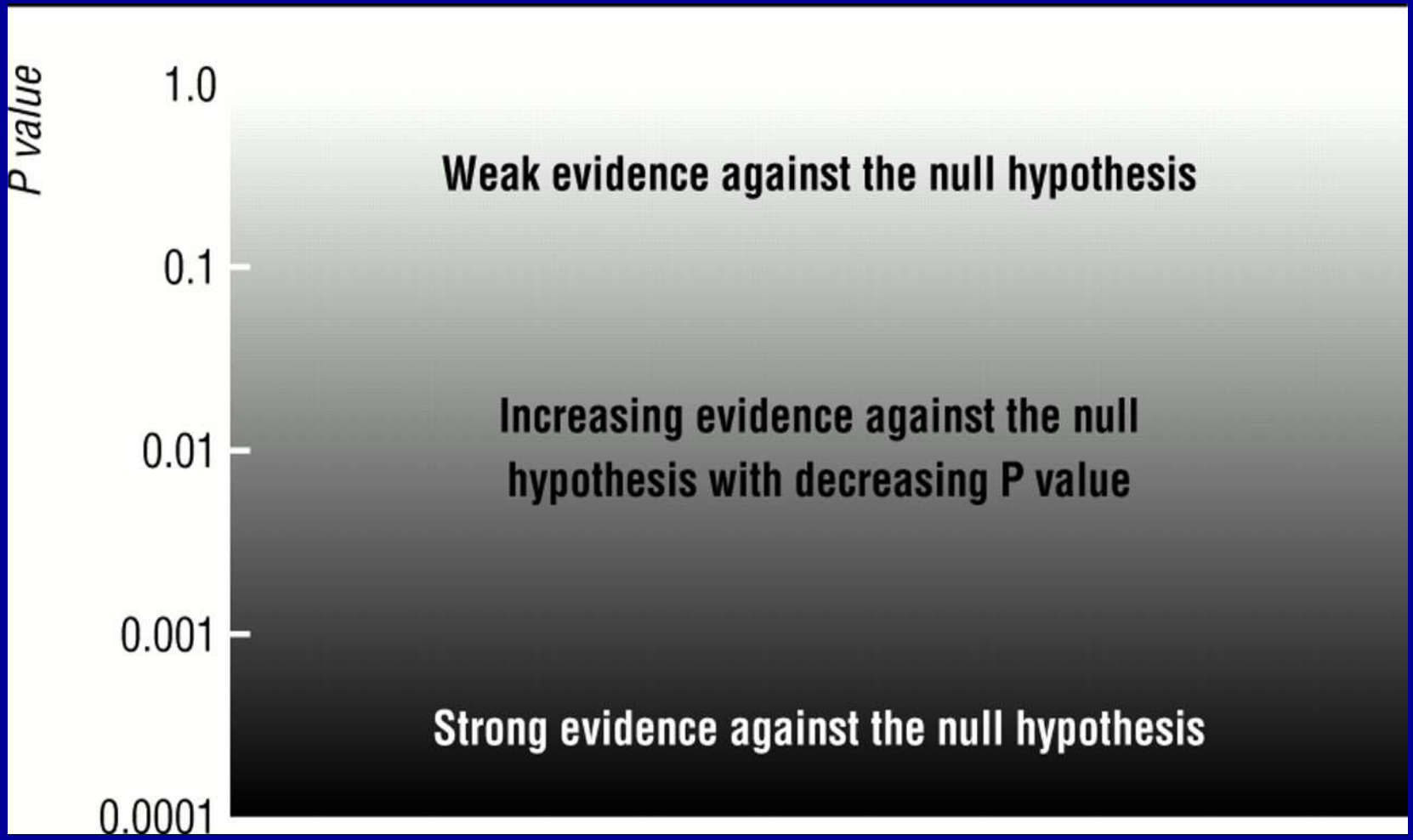
P>.05 **accept** null hypothesis

Fisher's strength of evidence interpretation
versus
Neyman & Pearson's decision approach

introduced Type II error, power calculations

Fisher: "the calculation is absurdly academic, for no scientific worker has a fixed level of significance at which he rejects hypotheses"

P-values are about “Shades of Grey”



Sterne, J. A C et al. BMJ 2001;322:226-231

“A P-value is no substitute for a brain”

BMJ

beware of big effects in small trials

Perioperative beta-blocker use in non-cardiac surgery

DECREASE trial of bisoprolol [NEJM 1999;341 p 1789-]

	bisoprolol	**control**		
N	59	53		"too good to be true"?
death	2	9	P=0.02	**scientific misconduct**
myocardial infarction	0	9	P<0.001	

POISE trial of metoprolol [Lancet 2008;371 p 1839-]

	metropolol	**placebo**	
N	4174	4177	
death	129	97	P=0.03
myocardial infarction	176	239	P<0.002

ESC/ESA Guidelines 2014: evidence inconclusive

beware of small effects in big trials

IMPROVE-IT trial [NEJM 2015;372 p2387-]

18,144 acute coronary syndrome patients on simvastatin

ezetimibe vs placebo

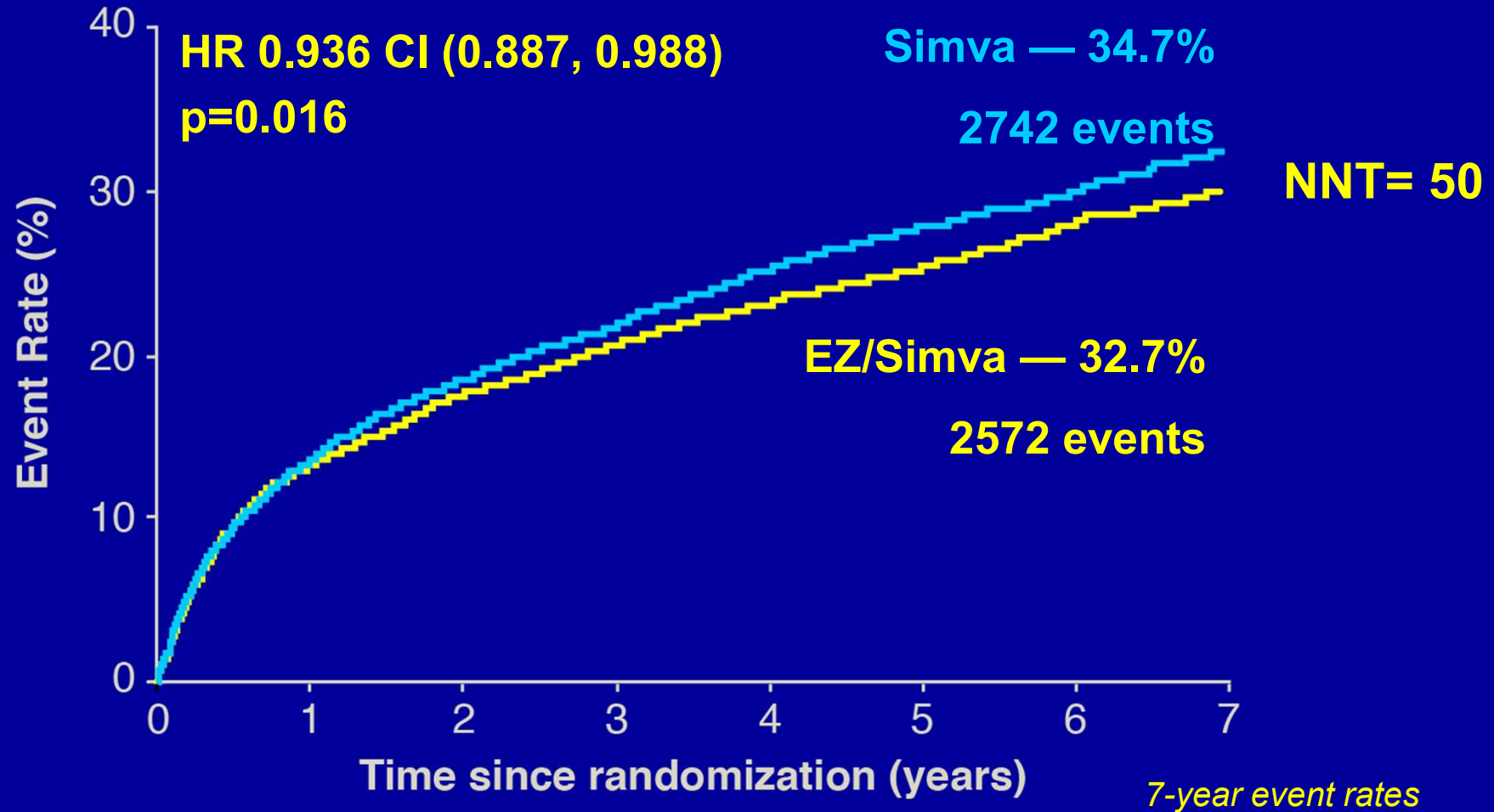
composite primary endpoint:
CV death, MI, stroke, unstable angina, coronary revasc.

5314 primary events over mean 5.4 years follow-up

the definitive study of ezetimibe?

Primary Endpoint — ITT

Cardiovascular death, MI, documented unstable angina requiring rehospitalization, coronary revascularization (≥30 days), or stroke



7-year event rates

on top of simvastatin, ezetimibe had a modest mean reduction in LDL-C (16.7 mg/dl)

modest impact on cardiovascular primary events:

relative risk reduction 6.4% (95% CI 2.2% to 11.3%)

absolute risk reduction 2.0% over 7 years

is this a worthwhile benefit?

ODYSSEY OUTCOMES Trial [ACC March 2018]

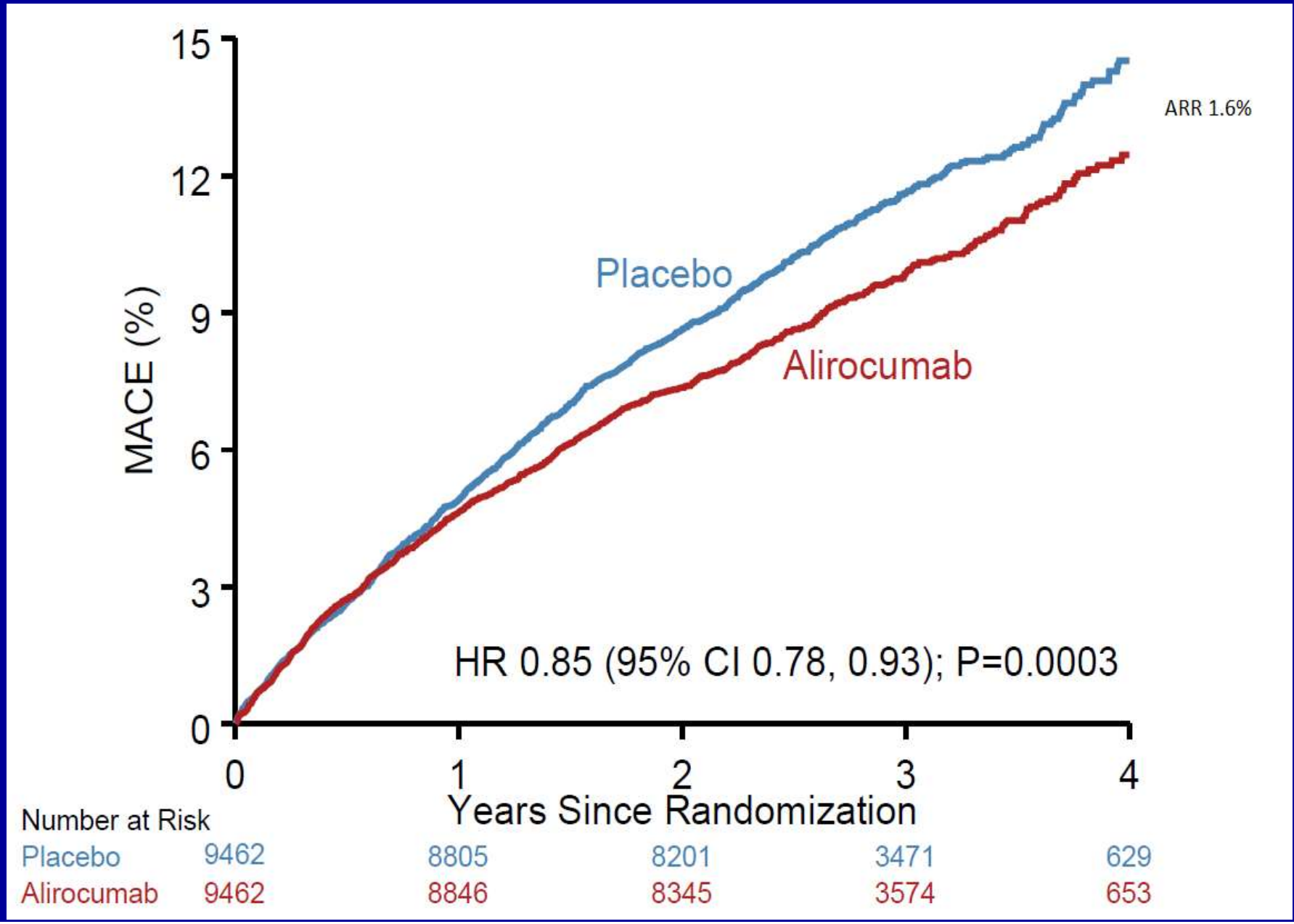
alirocumab vs placebo in 18,924 ACS patients
with LDL-C ≥70 mg/dl and on high-dose statin

this PCSK9 inhibitor reduces LDL-C by ~60%

composite primary outcome:
CHD death, MI, ischemic stroke, hospitalized unstable angina

median 2.8 years follow-up

ODYSSEY trial: primary efficacy endpoint MACE



ODYSSEY Primary Endpoint and it Components

Endpoint, n (%)	Alirocumab (N=9462)	Placebo (N=9462)	Hazard Ratio (95% CI)	Log-rank P-value
MACE (primary)	**903 (9.5)**	**1052 (11.1)**	**0.85 (0.78, 0.93)**	**0.0003**
CHD death	205 (2.2)	222 (2.3)	0.92 (0.76, 1.11)	0.38
Non-fatal MI	626 (6.6)	722 (7.6)	0.86 (0.77, 0.96)	0.006
Ischemic stroke	111 (1.2)	152 (1.6)	0.73 (0.57, 0.93)	0.01
Unstable angina	37 (0.4)	60 (0.6)	0.61 (0.41, 0.92)	0.02

absolute reduction in first MACE event:
5.62 per 1000 patient years (95% CI 2.35 to 8.89)

no. needed to treat (NNT):
63 patients for median 2.8 years (95% CI 41 to 141)

no apparent benefit in first 12 months

ODYSSEY Secondary Endpoints in order of hierarchical testing

Endpoint, n (%)	Alirocumab (N=9462)	Placebo (N=9462)	Hazard Ratio (95% CI)	Log-rank P-value	
CHD event	1199 (12.7)	1349 (14.3)	0.88 (0.81, 0.95)	0.001	✓
Major CHD event	793 (8.4)	899 (9.5)	0.88 (0.80, 0.96)	0.006	✓
CV event	1301 (13.7)	1474 (15.6)	0.87 (0.81, 0.94)	0.0003	✓
Death, MI, ischemic stroke	973 (10.3)	1126 (11.9)	0.86 (0.79, 0.93)	0.0003	✓
CHD death	205 (2.2)	222 (2.3)	0.92 (0.76, 1.11)	0.38	X
CV death	240 (2.5)	271 (2.9)	0.88 (0.74, 1.05)	0.15	(X)
All-cause death	334 (3.5)	392 (4.1)	0.85 (0.73, 0.98)	0.026*	(✓)

all-cause death not formally significant, it's exploratory

no mortality signal in FOURIER trial of evolocumab

Subgroup Claim

alirocumab more effective if LDL-C ≥100 mg/dl:
24% reduction in MACE, 29% reduction in death

but no significant interactions (P=0.09 and P=0.12 respectively)

beware of such “positive spin”

more plausible is that absolute benefit for MACE is greater in higher risk patients

need to stratify patients by overall risk (not just LDL-C)
and concentrate on absolute (not relative) reduction

General Issues arising from ODYSSEY

Estimating the magnitude of treatment effect

relative risk reduction useful, but quantify uncertainty:

eg risk ratio, odds ratio, hazard ratio
plus 95% confidence interval

absolute risk reduction more useful

eg. difference in %, difference in rates, NNT
again plus 95% CI

absolute risk reduction greater in higher-risk patients

The meaning of LIFE [Lancet March 23, 2002]

losartan vs atenolol, 9193 patients with hypertension
primary endpoint: death, MI or stroke over 4.8 years

no of events $\frac{508}{4605}$ losartan vs $\frac{588}{4588}$ atenolol

relative risk reduction 13% 95% CI 2% to 23% P=.021

“losartan prevents more cardiovascular events than atenolol”

but not overwhelming evidence
also what about absolute risk reduction?

Absolute Risk Reduction in the LIFE trial

primary endpoint: death, MI or stroke

	losartan	**atenolol**
it occurred in	508	588 patients
rate per 1000 patient years	23.8	27.9

difference 4.1 per 1000 patient years
with 95% CI 1.1 to 7.1 per 1000 patient years

No. Needed to Treat (NNT)

244 patient years of treatment to prevent one event
95% CI 141 to 909 patient years
a small gain imprecisely estimated

Subgroup analyses: interpret with caution

1) **patients are not homogeneous**
   response to treatment may well vary
   legitimate to explore in subgroup analyses

2) **trials usually not large enough**
   lack power to detect subgroup effects

3) **many possible subgroups**
   guard against data dredging/false positive

4) **do not rely on subgroup P-values**
   use interaction tests instead

A second **LIFE study** report on **diabetic subgroup**

[*Lancet* 23 March 2002]

"the benefits of losartan were more marked in this group"

	primary events (CV death, MI, stroke)		
	losartan	atenolol	relative risk
diabetics N=1195	103	139	0.77
non-diabetics N=7798	405	449	0.89

is relative risk greater in diabetics?

interaction test (test for heterogeneity) P=0.22

insufficient evidence that more effective in diabetic subgroup

Can we be more sensibly cautious re subgroup claims

CURRENT OASIS 7

Standard vs double dose clopidogrel in 25,087 ACS patients

NEJM 2 Sept 2010

"in patients referred for an invasive strategy, there was no significant difference between double-dose and standard dose"

Lancet on-line 1 Sept 2010

"In patients undergoing PCI double dose was associated with a reduction in CV events"

Confused?!

CURRENT OASIS 7

Primary Composite Outcome: CV Death, MI, stroke at 30 days

Overall 4.4% vs 4.2% P=0.37

no evidence that double dose is beneficial?

Subgroup analysis

	clopidogrel dose				
	standard	**double**	**hazard ratio**		**interaction test**
PCI (N=17232)	4.5%	3.9%	0.85	P=.039	P=.016
no PCI (N=7855)	4.2%	4.9%	1.17	P=.14	

"double dose clopidogrel reduced CV events in PCI patients"?!

beware: 1) PCI an improper subgroup
2) qualitative interactions are rare/implausible
3) such secondary evidence is weak

Analysis by Intention to Treat (ITT)

Analyse **all** randomised patients in their **allocated** groups

An unbiased comparison of strategies

Dilution bias for “pure” treatment effect

Per protocol analysis of compliers

Potential bias

Nearer to “pure” treatment effects

Often do both, but **emphasize ITT**

ROCKET-AF trial [NEJM 2011 365 p883-]

rivaroxaban vs warfarin in atrial fibrilation

14264 patients with 1.9 years median follow-up

primary endpoint: stroke or systemic embolism

Analysis by Intention to Treat

269 vs 306 hazard ratio 0.88 95% CI 0.74 to 1.03, P=0.12

Per Protocol Analysis

188 vs 241 hazard ratio 0.79 95% CI 0.66 to 0.96, P=0.02

Conclusions concentrated on the former

The Journal did not permit a claim of superiority

When can per protocol analysis be of value?

CABANA trial

2204 patients with atrial fibrillation

Ablation therapy **vs** **Drug therapy**

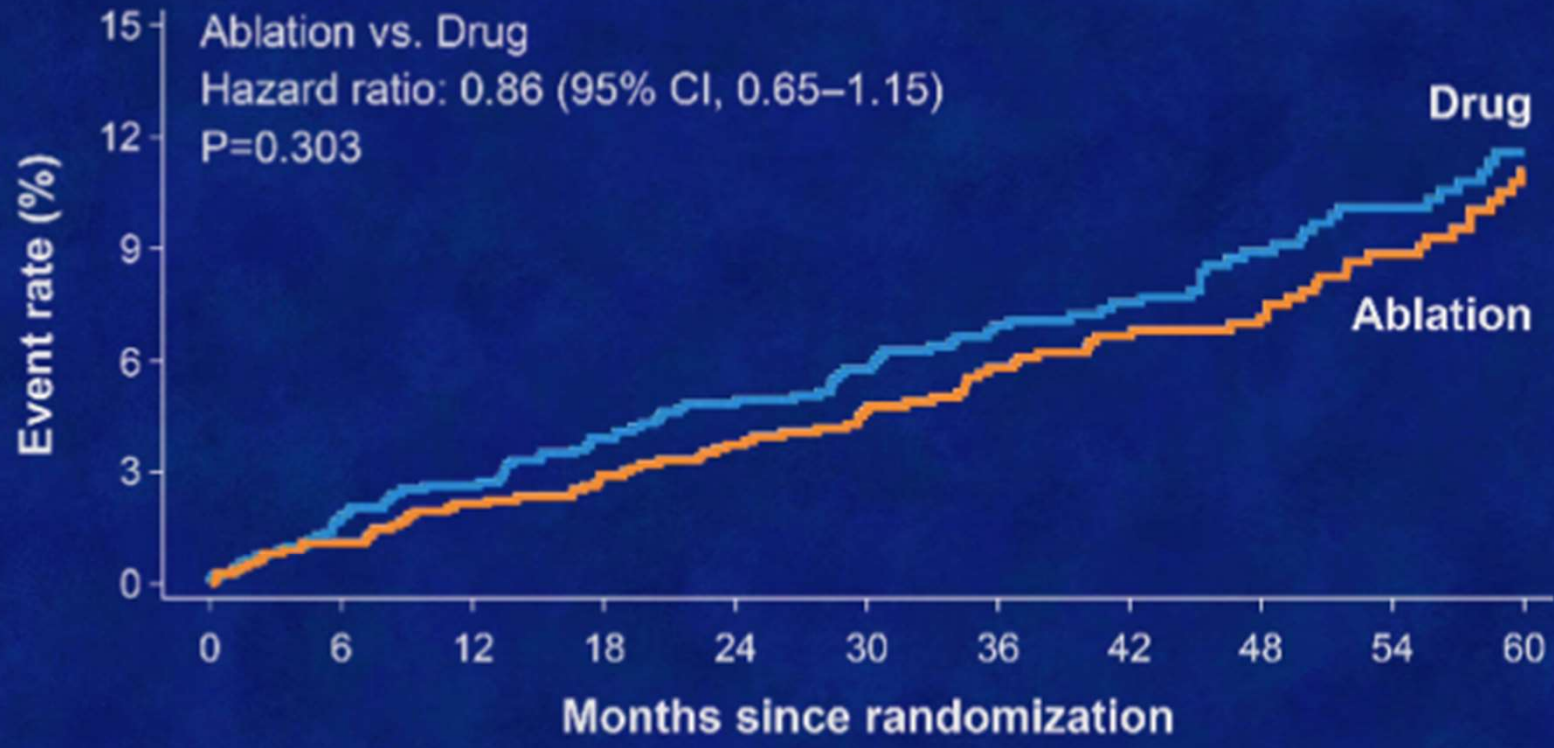
⇓ ⇓

9.2% not ablated 27.5% ablated

Primary endpoint: death, stroke, serious bleed, cardiac arrest over mean 4 years follow-up

ITT analysis “diluted” by crossovers

Primary Endpoint (Death, Disabling Stroke, Serious Bleeding, or Cardiac Arrest) (ITT)

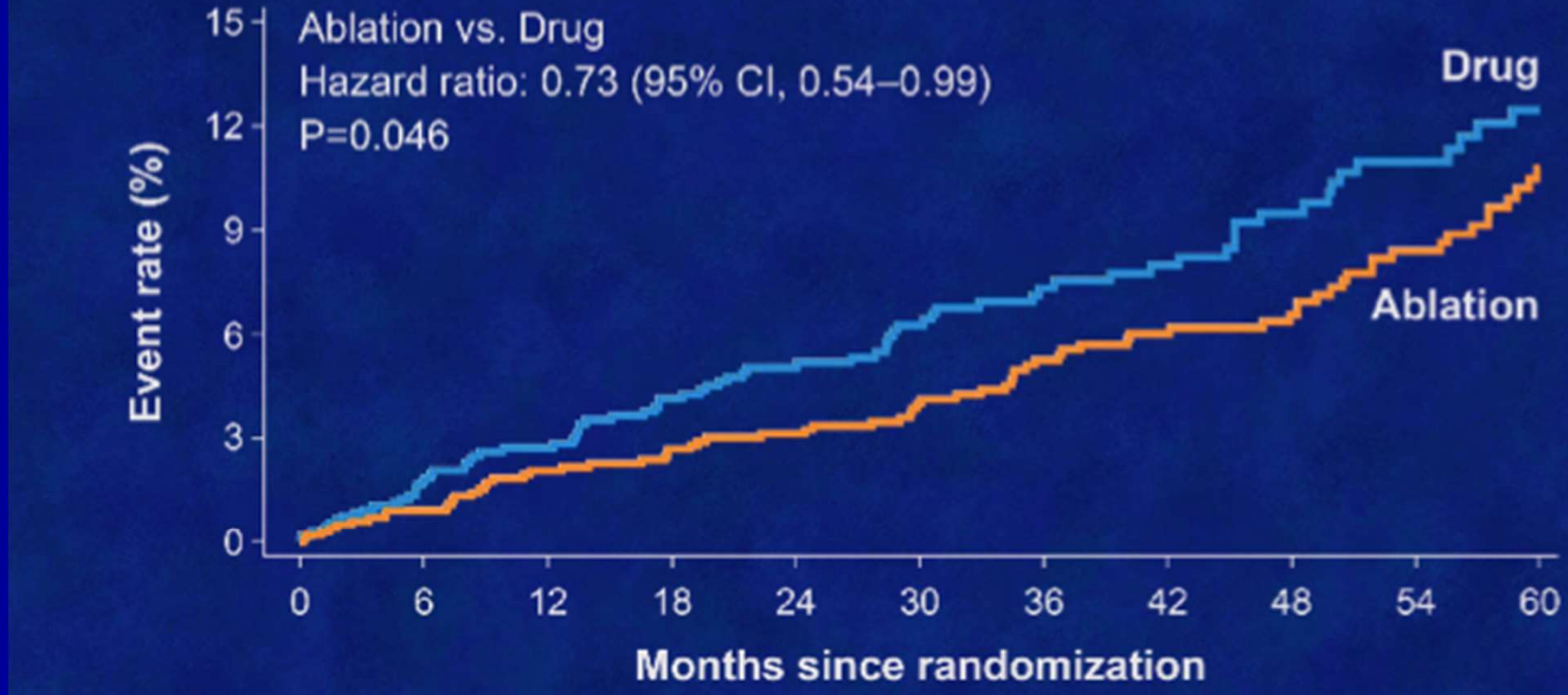


Number at risk

Drug	1096	1036	1006	970	880	763	652	578	499	418	312
Ablation	1108	1045	1021	996	915	793	700	614	535	432	309

MAYO CLINIC Duke Clinical Research Institute NIH National Heart, Lung, and Blood Institute

Primary Endpoint (Death, Disabling Stroke, Serious Bleeding, or Cardiac Arrest (Per Protocol)



Number at risk

Drug	1096	954	860	778	680	566	464	396	330	275	204
Ablation	987	958	937	918	849	735	648	566	494	404	291

MAYO CLINIC Duke Clinical Research Institute NIH National Heart, Lung, and Blood Institute

My Conclusions re CABANA

Analysis by Intention to Treat

no significant reduction on primary endpoint

results affected by cross-overs and low event rates

ablation reduced mortality or CV hospitalization by 17% (P=0.002)

Per Protocol Analysis

33% reduction in primary endpoint (P=0.046)
40% reduction in mortality (P=0.005)

but not clear what method was used

potential bias, controversy remains

Strategies for handling Secondary Endpoints

pre-define a limited set of key secondary endpoints

other endpoints become exploratory

pre-define either:

1) a hierarchy of secondary testing

or 2) correction for multiple testing (eg Bonferroni)

strict control of type I error: good or bad?

it matters to FDA, less to scientific knowledge

flexibility without cheating

all-cause death is different?

Interpretation of Secondary Endpoint Surprises

safety concern: heart failure in SAVOR trial

efficacy bonus: mortality in EMPA-REG trial

my prior statistical perspective

any unexpected finding (good or bad) is prone
to be an exaggeration

note one of multiple hypotheses across secondary endpoints

collect more data (if you can) and expect
regression to the truth

is it a real effect or just due to chance?
often impossible to tell

SAVOR-TIMI 53 trial [NEJM 2013; 369 p 1317-]

Saxagliptin vs Placebo in 16,492 high risk type II diabetics

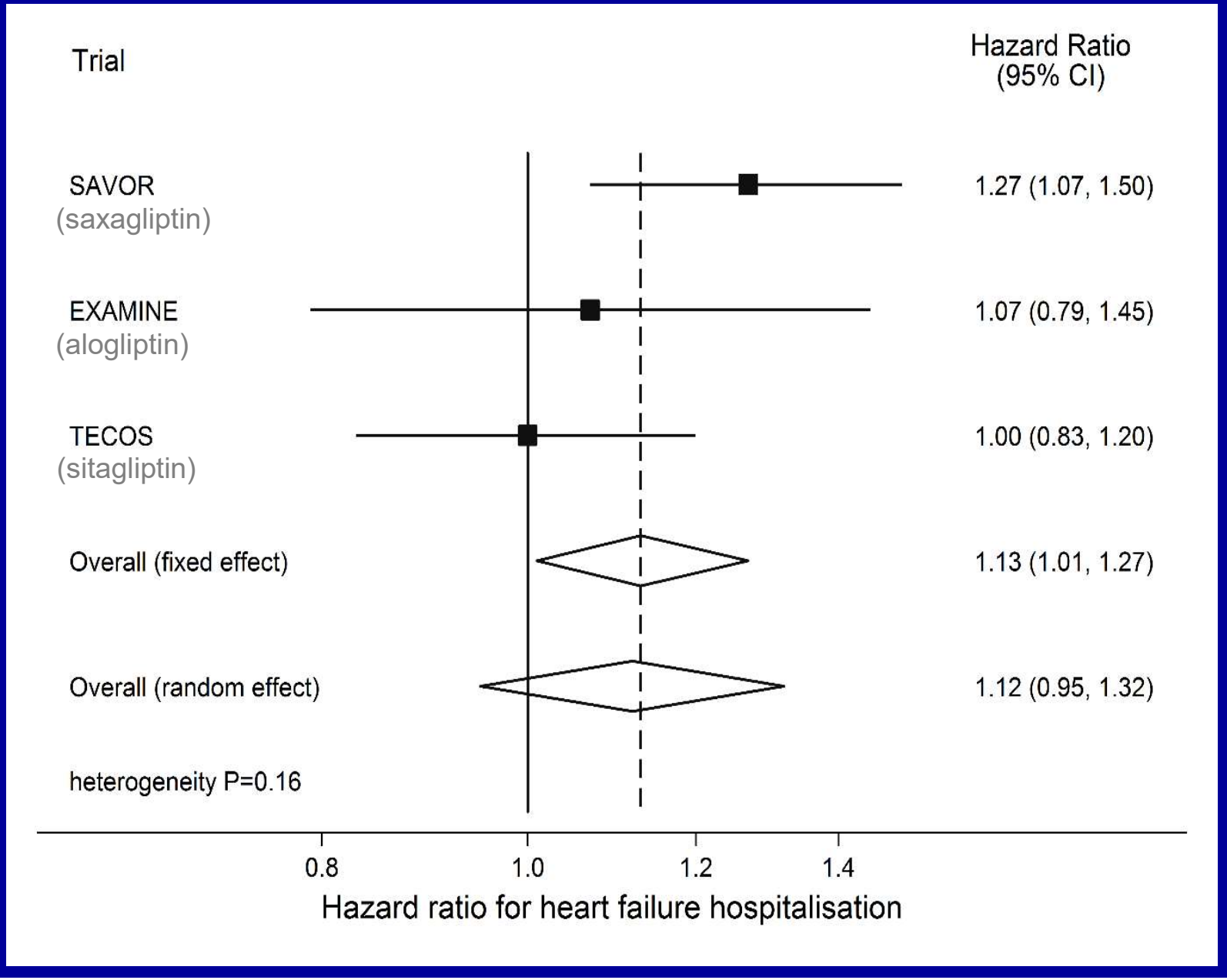
788 sites in 26 countries, median 2.1 years follow-up

	saxagliptin [N=8280]	**placebo** [N=8212]	**hazard ratio (95% CI)**
primary endpoint (CV death, MI, stroke)	613	609	1.00 (0.89 to **1.12**)
heart failure hosp^{n}.	289	228	1.27 (1.07 to 1.51) ⇓ P=.007

primary endpoint: non-inferiority established, but no benefit

heart failure: given multiple testing, a false positive?

Combining Evidence from 3 Related Trials



EMPA-REG OUTCOME trial [NEJM 17 Sept 2015]

empagliflozin 10mg vs 25mg vs placebo in 7020 type 2 diabetics

primary endpoint: CV death, MI, stroke over median 3.1 years

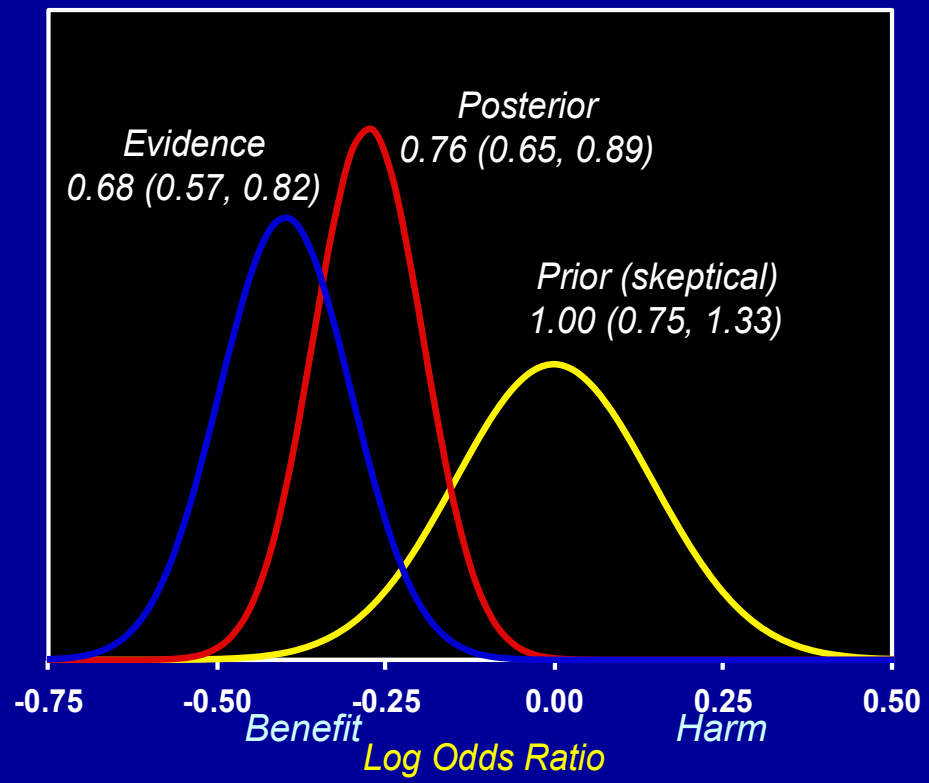
	empagliflozin combined [N=4687]	**Placebo** [N=2333]	
primary	10.5%	12.1%	P=.04
heart failure hosp^{n}.	2.7%	4.1%	P=.002
all-cause death	5.7%	8.3%	P<.0001

how do we interpret such impressive secondary findings?

Moderating "Too Good To Be True" Results

Bayesian Analysis of All Cause Mortality
EMPA REG OUTCOME Trial (Empagliflozin in T2 DM)

Study Group	OUTCOME +	EVENT -	Total
Ept	269	4418	4687
Con	194	2139	2333
Total	463	6557	7020



	N	Event Rate (%)
Empagliflozin	4687	5.7
Control	2333	8.3

RRR	95% CI	P (2-tailed)
32%	18%-43%	<0.001

Prior	Pb≥0	Pb≥10%	Pb≥15%	Pb≥20%
Noninformative	0.999	0.998	0.992	0.963
Skeptical	0.999	0.982	0.918	0.743

courtesy of Sanjay Kaul

COMPASS trial [NEJM 2017:377 p1319-]

27,395 patients with stable cardiovascular disease

Rivaroxaban 2.5 mg bd + Aspirin 100mg
vs
Rivaroxaban 5 mg bd alone
vs
Aspirin 100 mg alone

primary outcome:
composite of death, myocardial infarction, stroke

trial stopped early for superiority of R + A

mean follow-up 23 months

COMPASS trial Primary Outcome

	R + A N=9152	R alone N=9117	A alone N=9126	P-value for R + A v Alone
Primary Outcome (CV death, MI, stroke)	379	448	496	P<0.001
CV death	160	195	203	P=0.01
Stroke	83	117	142	P<0.001
Myocardial Infarction	178	182	205	P=0.12
Major Bleeding	288	255	170	P<0.001

strong evidence that R + A has superior efficacy
and R + A increases bleeding risk

Feb 6 2017: 1st formal interim analysis

for primary efficacy outcome:

R + A vs A alone has z = 4.592, exceeds boundary

R alone vs A alone has z = 2.44, P=.015

DSMB recommends stopping

Aug 27 2017: results published in NEJM

R + A vs A alone z = 4.126

hazard ratio 0.76 (95% CI 0.66 to 0.86) P<0.0001

R alone vs A alone z = 1.575

hazard ratio 0.90 (95% CI 0.79 to 1.03) P=0.12

some regression to the truth

Issues re Stopping Early

extreme boundary means superiority of R + A believable

but may be exaggerated somewhat (random high)

mean follow-up restricted to 23 months

less assurance re long-term benefit

balancing efficacy and safety (bleeding): less data

regulatory consequences with FDA and EMA

a negative trial: can it be rescued?
TRUE-AHF trial [NEJM 2017; 376 p1956-]

Ularitide vs Placebo in 2157 acute heart failure patients

Co-Primary Endpoints	**Ularitide**	**Placebo**	
Cardiovascular Death (median 15 months)	21.7%	21.0%	P=0.75
Clinical Composite over 48 hrs:			
improved	48.6%	47.5%	P=0.82
unchanged	44.8%	44.2%	
worse	6.6%	8.3%	

seemingly, a “negative” trial

but 16.6% of patients were ineligible

also 3.7% were missing clinical composite

TRUE-AHF Results in 1799 Eligible Patients

	Ularitide	Placebo	
Cardiovascular Death	20.7%	20.8%	P=0.87
Clinical Composite over 48 hrs:			
improved	49.8%	45.8%	P=0.035
unchanged	43.9%	45.6%	
worse	6.3%	8.6%	

a weak signal of short-term benefit in eligible patients

such post hoc findings need cautious interpretation

concerns about quality of trial conduct

Were some countries deficient in trial conduct?

TOPCAT trial [Circulation 2015; 131 p34-]

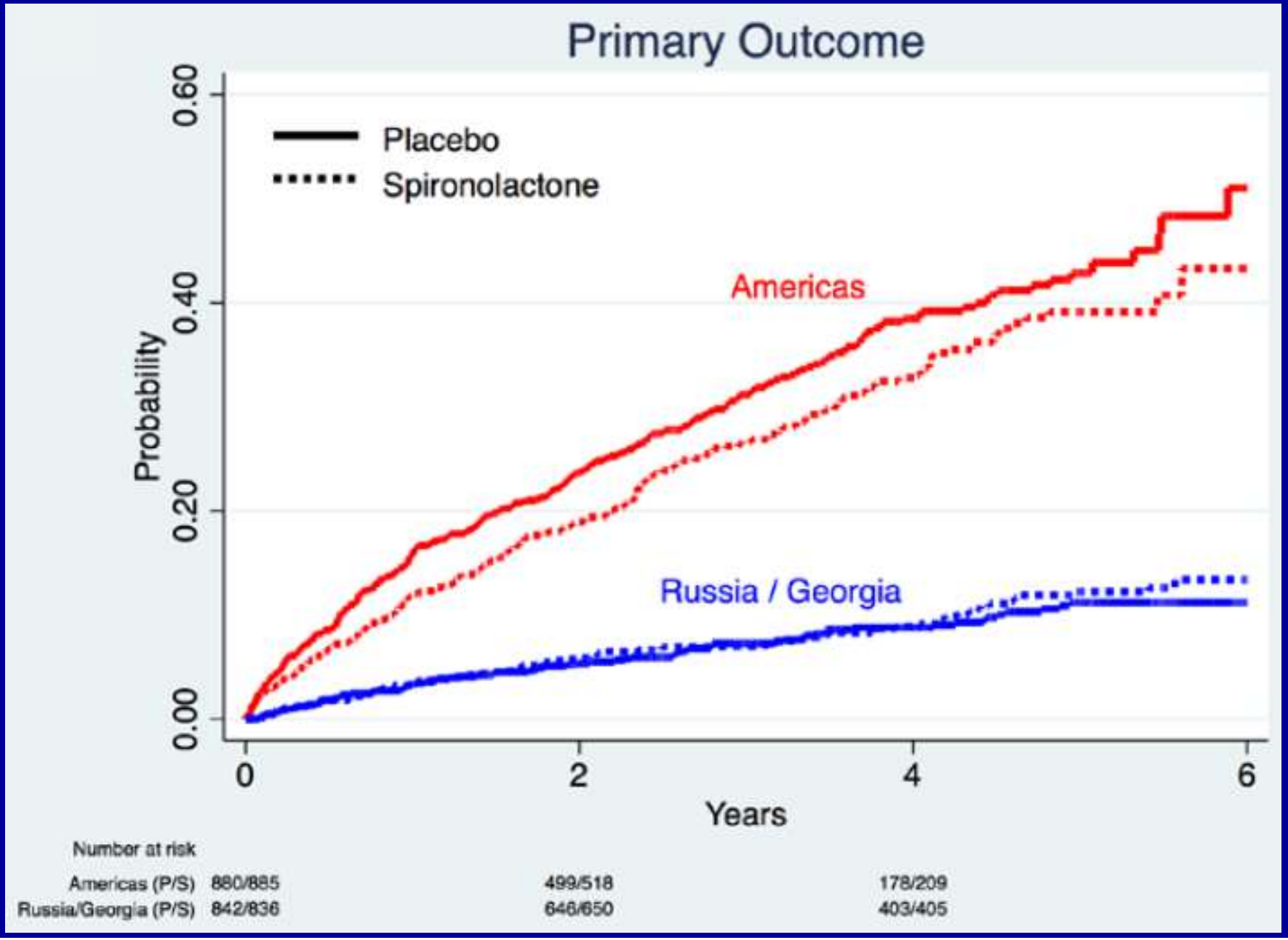
spironolactone vs placebo in preserved EF heart failure
primary outcome: CV death, cardiac arrest or heart failure hosp^{n}.
hazard ratio 0.89 (95% CI 0.77 to 1.04) P=0.14

patients in Russia and Georgia (N=1678)
few events, not representative

in Americas subgroup (N=1767)
hazard ratio 0.82 (95% CI 0.69 to 0.98) P=0.026

is this convincing enough to recommend spironolactone?

Regional Variation in TOPCAT



Meta-analysis: the pros and cons

any one trial is 1) too small and 2) lacks generalisability

combining evidence from related trials is good in principle

but can meta-analyses be trusted?

3 key concerns re any meta-analysis:

breadth how similar are the trials?
re patients, treatments, outcomes

quality which trials are good enough to include?
use of individual patient data is better

representativeness can one identify all eligible studies?
risk of publication bias

meta-analysis is a "growth industry"

Journals are wary: publish the good, ignore the bad!

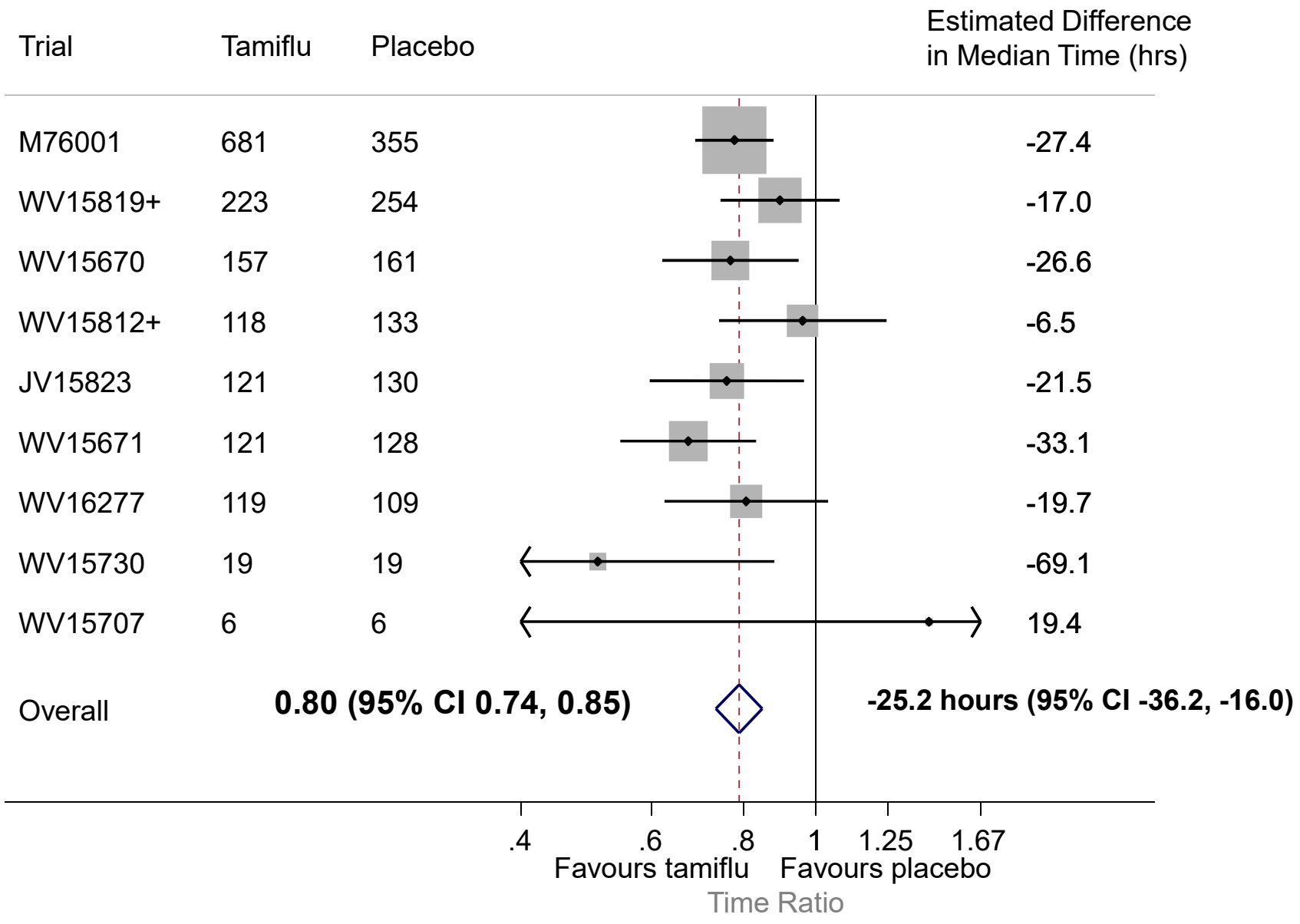
Oseltamivir (tamiflu) treatment for influenza in adults

meta-analysis of individual patient data

from 9 randomised trials (all sponsored by Roche)

[Lancet 2015; 385 p 1729-]

Time to Symptom Alleviation: Accelerated Failure Time Meta-Analysis



Estimated Benefits and Risks of Tamiflu

median time to alleviation of symptoms ↓ 25 hours

risk of lower respiratory infection ↓ 3.8%

risk of hospital admission (rare) ↓ 0.6%

incidence of nausea ↑ 3.7% vomiting ↑ 4.7%

Controversy: validity of findings questioned
our integrity challenged

Conclusions

clinical trial reports cannot be automatically trusted

constructive critical appraisal is always required

positive spin (being economical with the truth) is common

trialists (and sponsors) struggle with

1) the search for truth (honest science)

versus

2) the desire for a “positive” trial

journal articles more reliable than conference presentations

only regulators see the full story

Further Reading

JACC 2018; 71: 2957-69

NEJM 2016; 375: 861-70 and 971-9

JACC 2015; 66: 2536-49, 2648-62, 2757-66 and 2886-98

JACC 2014; 64; 1615-28