

Institut de
recherche
Centre universitaire
de santé McGill



Research
Institute
McGill University
Health Centre



McGill

A Bayesian view of contemporary randomized clinical trials – New reflections from an old lens

James (Jay) Brophy MD PhD

Professor

Departments of Medicine and of Epidemiology,
Biostatistics, and Occupational Health
McGill University

Conflicts of Interest

- None

Learning Objectives

1. Understand the general principles of Bayesian reasoning
2. Bayesian reanalysis of some recent “positive” and “negative” randomized clinical trials (RCTs)
3. Bayesian analysis of a recently completed local RCT

Background

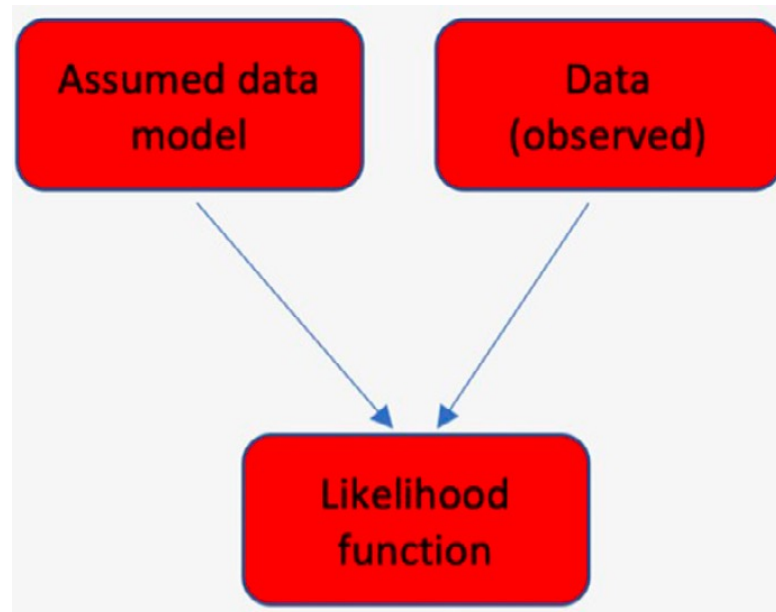
The need for statistical models

Some data

- 8 events in a group of 97 patients
- 29 events in a control group of 97 patients
- Risk difference is $(8/97) - (29/97) = -22\%$
- Risk ratio is $(8/97) / (29/97) = .28$
- What inferential statements concerning the risk ratios or risk differences can be made?

Statistical inference (frequentist)

- Statistical inference requires a fundamental concept - Likelihood function
- Indicates how likely a particular population is to produce an observed sample (\mathbf{X}).
- ie $\mathbf{P}(\mathbf{X}|\Theta)$ or $\mathbf{L}(\mathbf{X}|\Theta)$ be the distribution of the data \mathbf{X} , where Θ is assumed to be fixed but unknown model parameter



Statistical inference

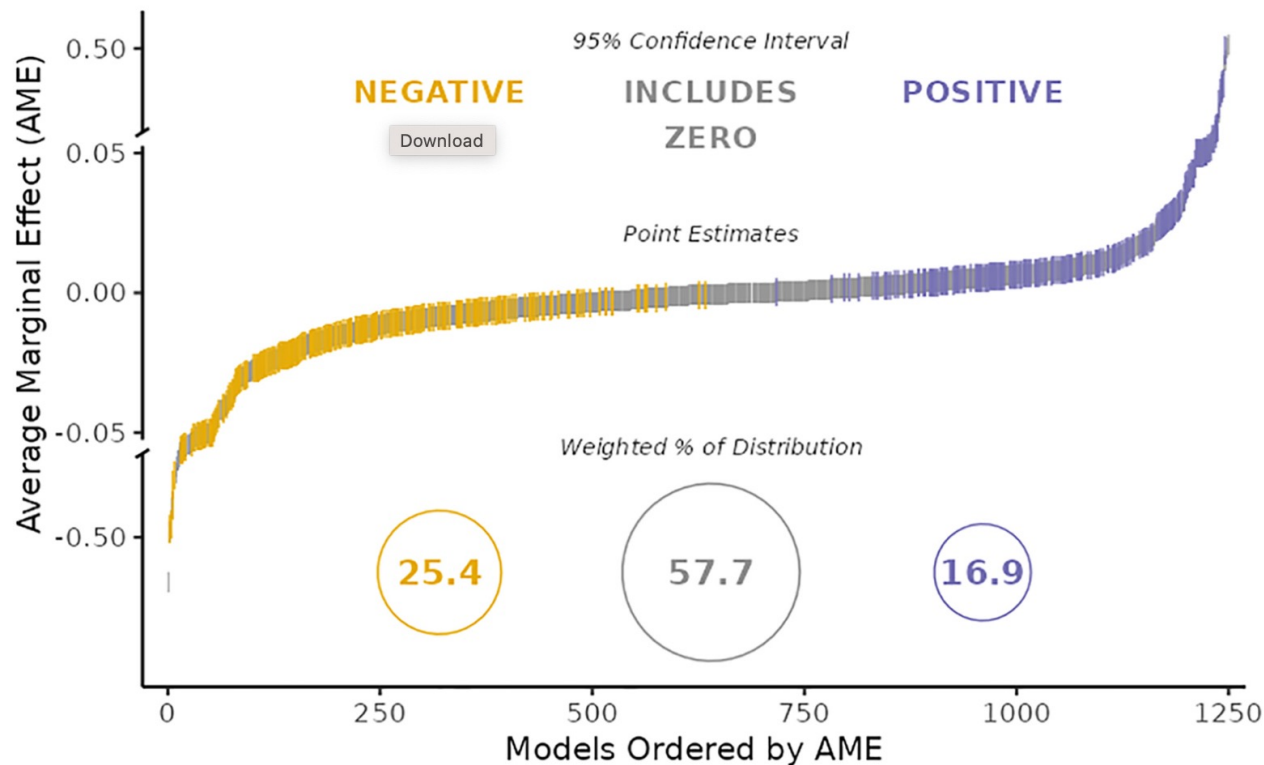
- `fisher.test(matrix(c(8,29,89,68), nrow=2))`
- Fisher's Exact Test for Count Data
- data: `matrix(c(8, 29, 89, 68), nrow = 2)`
- p-value = $2e-04$
- alternative hypothesis: true odds ratio is not equal to 1
- 95 percent confidence interval: 0.079 - 0.514
- sample estimates: odds ratio 0.212

What is the underlying model?

Model is two independent binomial random variables
 $X_1 \sim \text{Bin}(n_1, \theta_1)$ and $X_2 \sim \text{Bin}(n_2, \theta_2)$ with null hypothesis
 $\theta_1 = \theta_2$.

It's not only about the data

- Statistical inferences require a **mathematical model**
- A mathematical model aims to explain a complex phenomenon -> better understanding & decision making



Broad variation in the findings from 73 teams testing the same hypothesis with the same data. The distribution of estimated AMEs across all converged models ($n = 1,253$) includes results that are negative (yellow; in the direction

doi:<https://doi.org/10.1073/pnas.2203150119>

Modelling – beyond the data

- Models act as a mediating tool btw what we observe, and what we believe is the data generating mechanism
- Modelling is not an objective enterprise, **assumptions are always present!**
- A mathematical model should be the beginning of a discussion, not the (definitive) end
- Blind model acceptance is not correct at best, dangerous at worst, and disastrous at worst.
- All models are wrong, but some are useful. — Box ([1979](#))

Example # 1

A very “positive” trial

CASTLE-HF- Too good to be true?

October 12, 2023

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Catheter Ablation in End-Stage Heart Failure with Atrial Fibrillation

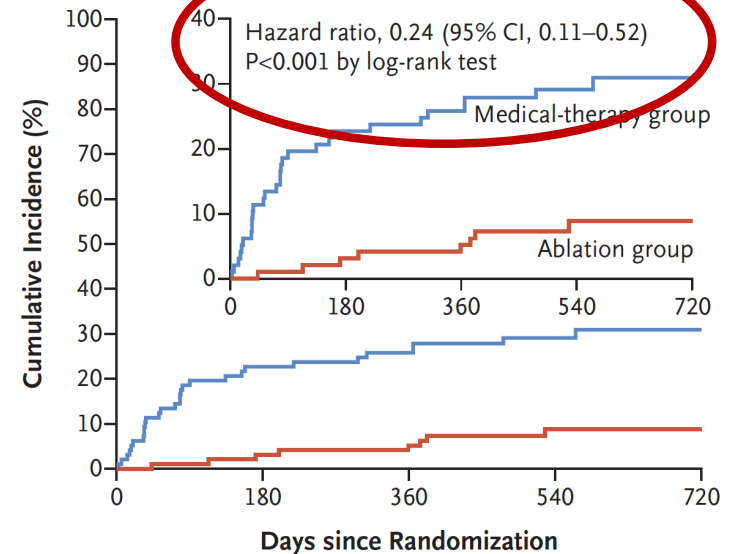
RESULTS

A total of 97 patients were assigned to the ablation group and 97 to the medical-therapy group. The trial was stopped for efficacy by the data and safety monitoring board 1 year after randomization was completed. Catheter ablation was performed

CONCLUSIONS

Among patients with atrial fibrillation and end-stage heart failure, the combination of catheter ablation and guideline-directed medical therapy was associated with a lower likelihood of a composite of death from any cause, implantation of a left ventricular assist device, or urgent heart transplantation than medical therapy alone.

A Primary End Point

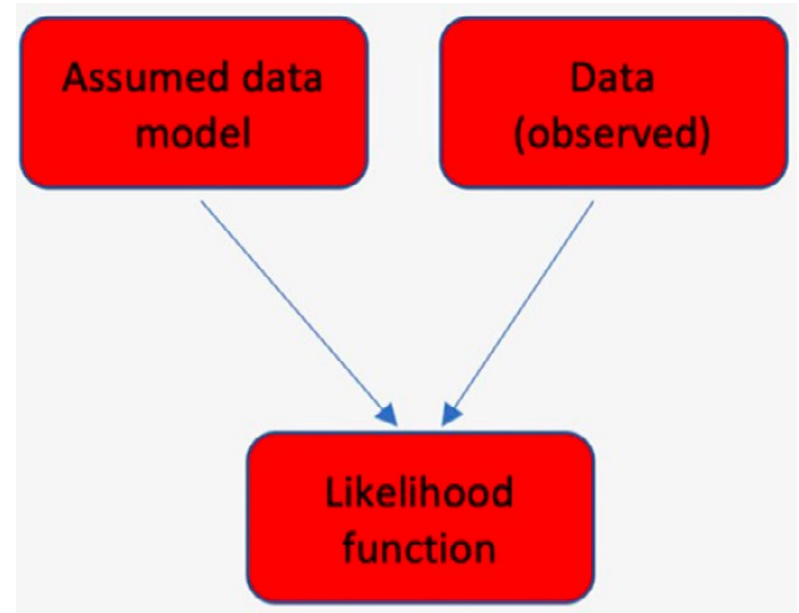


No. at Risk

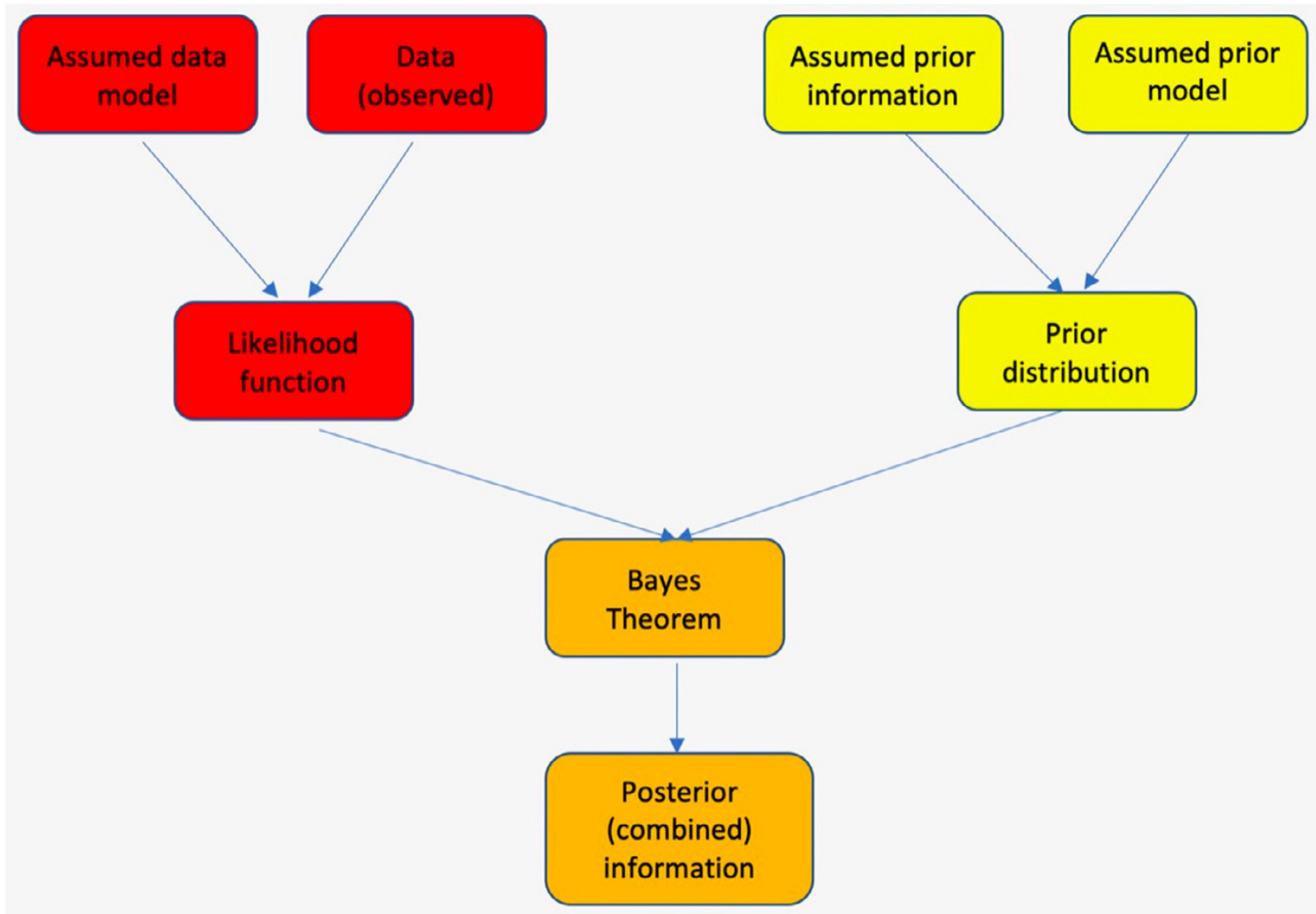
Medical-therapy group	97	75	72	41	12
Ablation group	97	94	88	50	20

Frequentist analysis

- `fisher.test(matrix(c(8,29,89,68), nrow=2))`
- Fisher's Exact Test for Count Data
- data: `matrix(c(8, 29, 89, 68), nrow = 2)`
- p-value = $2e-04$
- alternative hypothesis: true odds ratio is not equal to 1
- 95 percent confidence interval: 0.079 - 0.514
- sample estimates: odds ratio 0.212



Bayesian paradigm



Reproduce results

- Frequentist

```
glm(formula = I(1 - prop_success) ~ Tx, family = binomial(link = "logit"),
     data = castle_hf, weights = total)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-0.8522	0.2218	-3.843	0.000122	***
Tx1	-1.5570	0.4306	-3.616	0.000299	***

- $OR = \exp(-1.557) = 0.21$ (95%CI 0.09 – 0.49)

- Bayesian – vague prior

```
Formula: I(total - success) | trials(total) ~ Tx
Data: castle_hf (Number of observations: 2)
Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
       total post-warmup draws = 4000
```

Population-Level Effects:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-0.85	0.22	-1.28	-0.41	1.00	3746	2915
Tx1	-1.60	0.43	-2.49	-0.79	1.00	2051	2130

- $OR = \exp(-1.60) = 0.20$ (95%CrI 0.09 – 0.47)

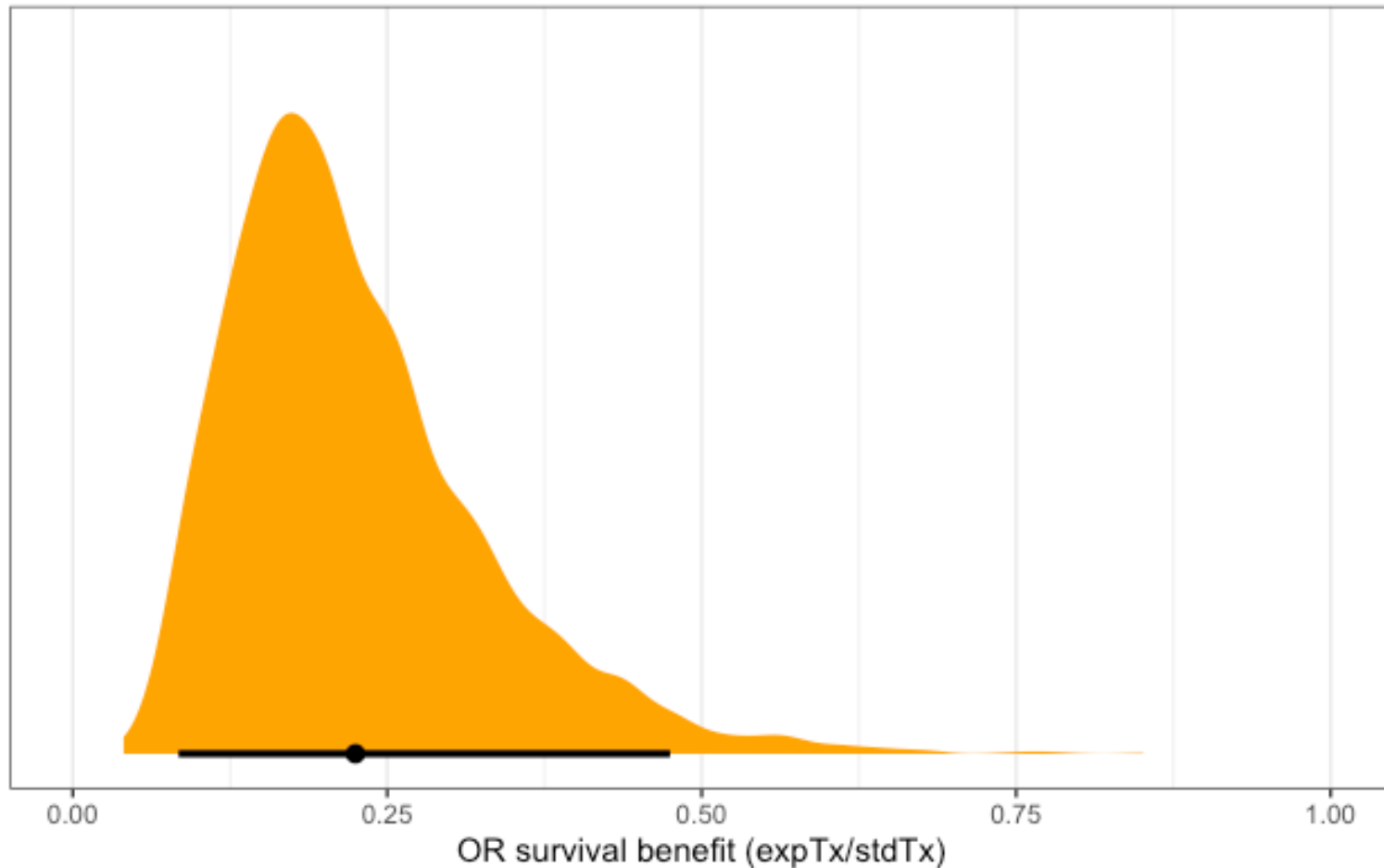
If the Bayesian model gives the same results, why might we be still interested in it?

1. Easier to understand the results – probabilities of hypotheses being true, $P(H_0|\text{data})$, as opposed to probability of observing more extreme data than was actually observed, $P(\text{data}|H_0)$
2. Concentrate on parameter estimation and not NHST
3. Can examine complete probability distribution and not limited to examining one artificial cutpoint (@ the null)
4. Main reason can look at more complex models, including incorporating prior knowledge

Bayesian – vague prior

CASTLE HTx odds ratio (expTx/stdTx) with vague prior

composite outcome = all cause death, LVAD implantation, or urgent heart Tx



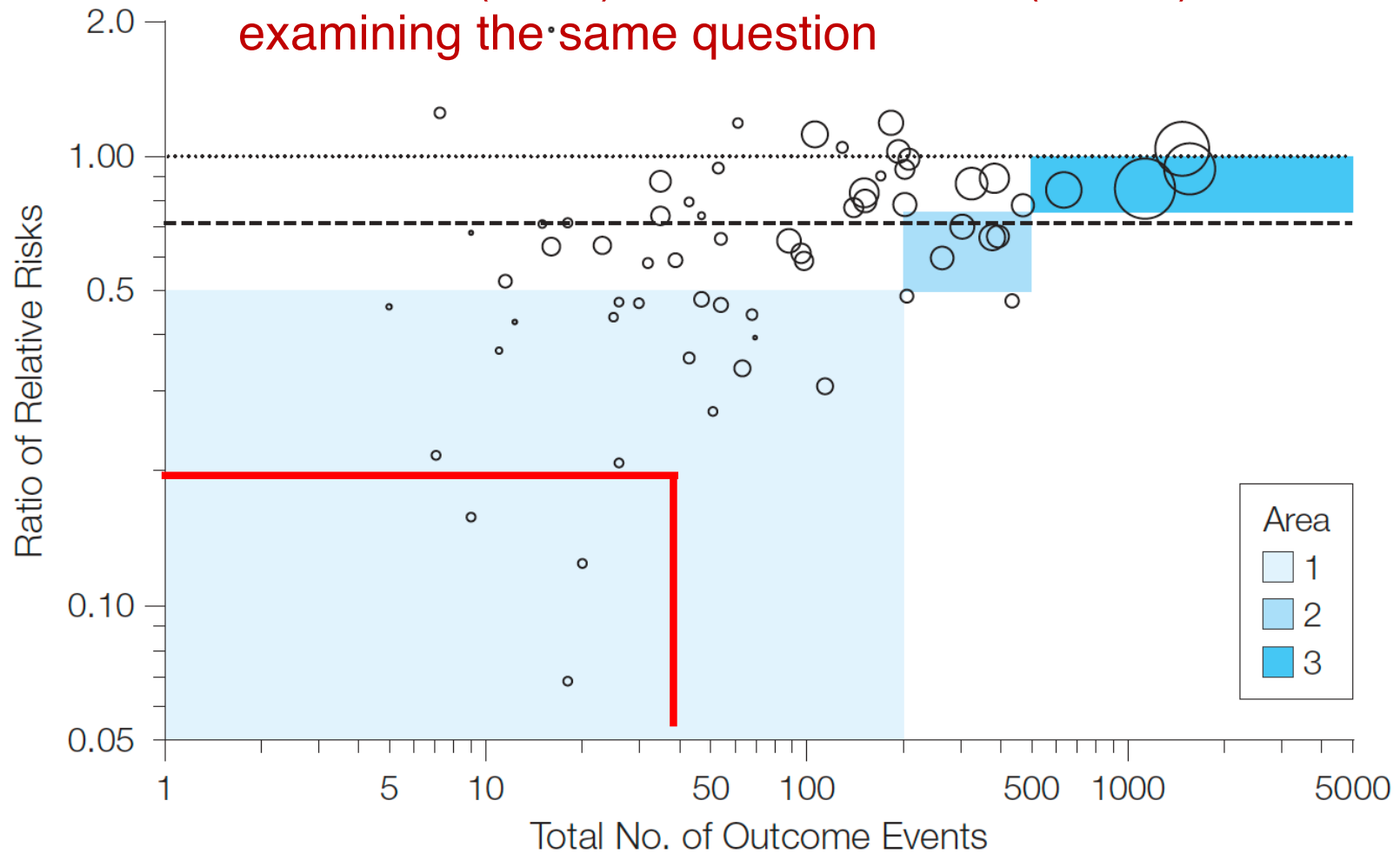
expTx = ablation
stdTx = standard treatment
LVAD = left ventricular assist device
Tx = transplant

Are there potential biases?

- If yes, is the current model adequate?
- How would you model any biases?

The bias of early stopping

Truncated (n=91) vs. non-truncated (n=424) trials examining the same question



≅ 5 fold overestimation for very small trials compared to effects in non-truncated trials

Potential biases

Inadequate concealment

Table 2.—Odds Ratios in the Unclearly and Inadequately Concealed Trials Compared With Those in Adequately Concealed Trials*

Level of Allocation Concealment	Ratio of Odds Ratios (95% Confidence Interval)	χ^2 (df)	P
Adequate	1.00 (referent)		
Unclear	0.67 (0.60-0.75)	57.9 (2)	<.001
Inadequate	0.59 (0.48-0.73)		

≅ 40% overestimation for trials with inadequate concealment

Potential bias

(Unblinding or ascertainment / performance bias)

Double-blinded		χ^2	p
Yes	1.00 (referent)	6.16 (1)	.01
No	0.83 (0.71-0.96)		

≅ 20% overestimation of treatment effect for unblinded compared to effects from blinded trials

Potential bias

(Concealment bias or bad luck with randomization)

Table 1. Characteristics of the Patients at Baseline.*

Characteristic	Ablation Group (N=97)	Medical-Therapy Group (N=97)	
Age — yr	62±12	65±10	✓
Male sex — no. (%)	85 (88)	72 (74)	✓
Body-mass index†	28±4	28±5	
NYHA functional class — no. (%)‡			
II	33 (34)	28 (29)	
III	52 (54)	54 (56)	✓
IV	12 (12)	15 (15)	
Left ventricular ejection fraction — %	29±6	25±6	✓
Diabetes mellitus — no. (%)	25 (26)	31 (32)	✓
N-terminal pro-BNP level			
No. of patients evaluated (%)	46 (47)	52 (54)	✓
Value — pg/ml	3852±3261	4461±5191	

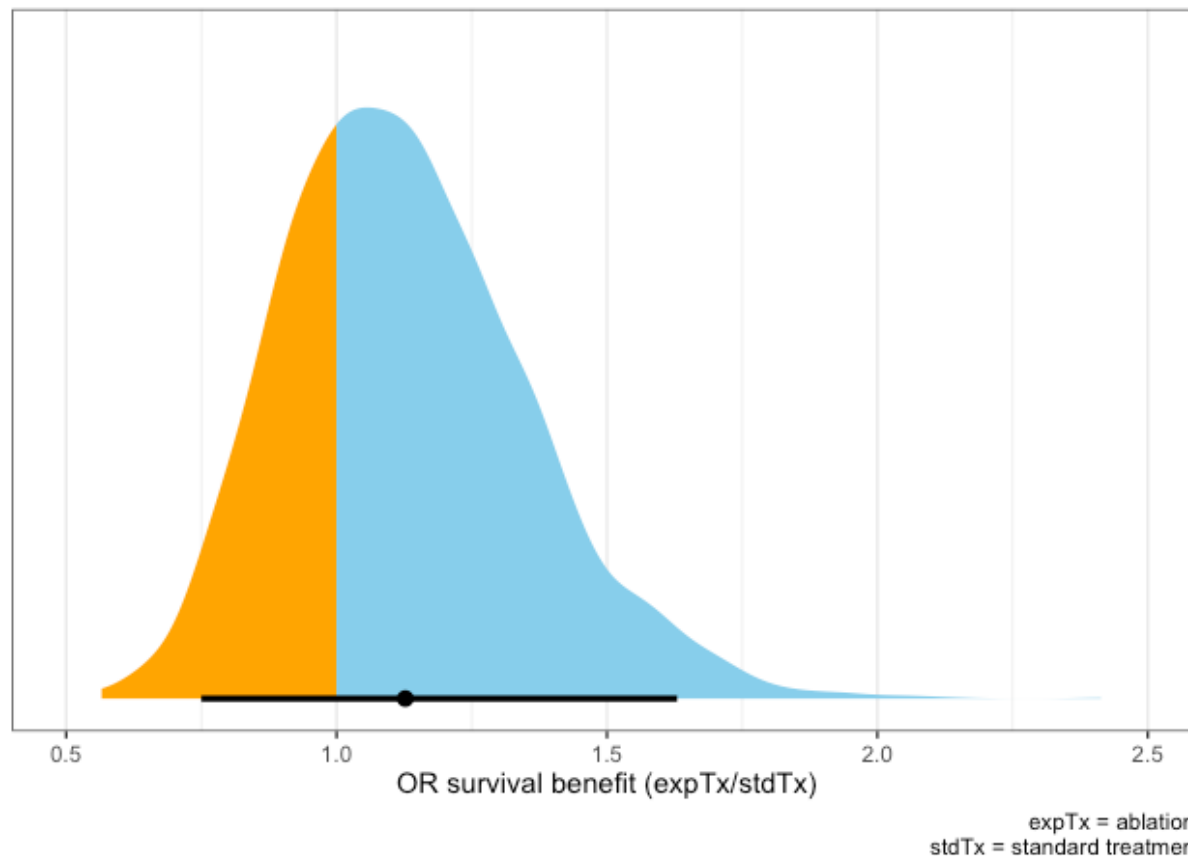
Prognosis worse with older pts, females, worse FC, lower EF, DM, higher BNP

What's the probability of 6 heads tossing a fair coin?

Prior beliefs (summation)

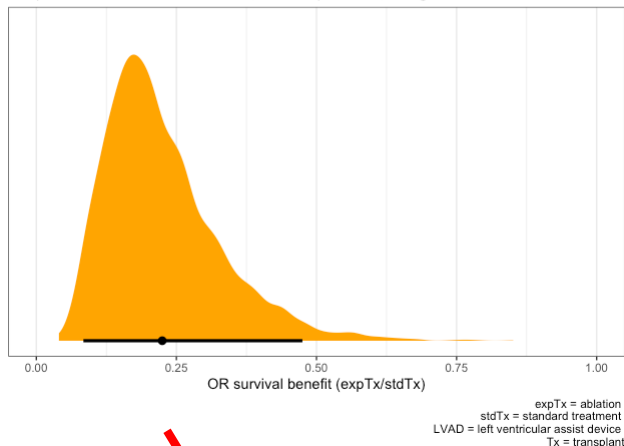
- Premature stopping (\cong 5 fold overestimation) OR 0.2 \rightarrow \cong 0.8
- Poor concealment (40% overestimation) OR 0.8 \rightarrow \cong 1.0
- Lack of blinding (\cong 10-20% overestimation) OR 1.0 \rightarrow 1.1
- Combined prior belief, based on study characteristic, not based on study data, $N(1.1, 0.2)$ (OR range 0.75 – 1.60)

CASTLE HTx prior belief of odds ratio (expTx/stdTx)
considering multiple biases (early stopping), poor concealment, unblinding

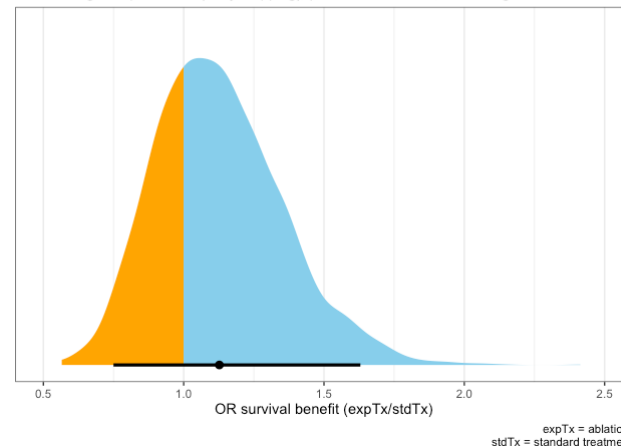


Updated CASTLE-HF Bayesian analysis

CASTLE HTx odds ratio (expTx/stdTx) with vague prior
composite outcome = all cause death, LVAD implantation, or urgent heart Tx



CASTLE HTx prior belief of odds ratio (expTx/stdTx)
considering multiple biases (early stopping, poor concealment, unblinding)

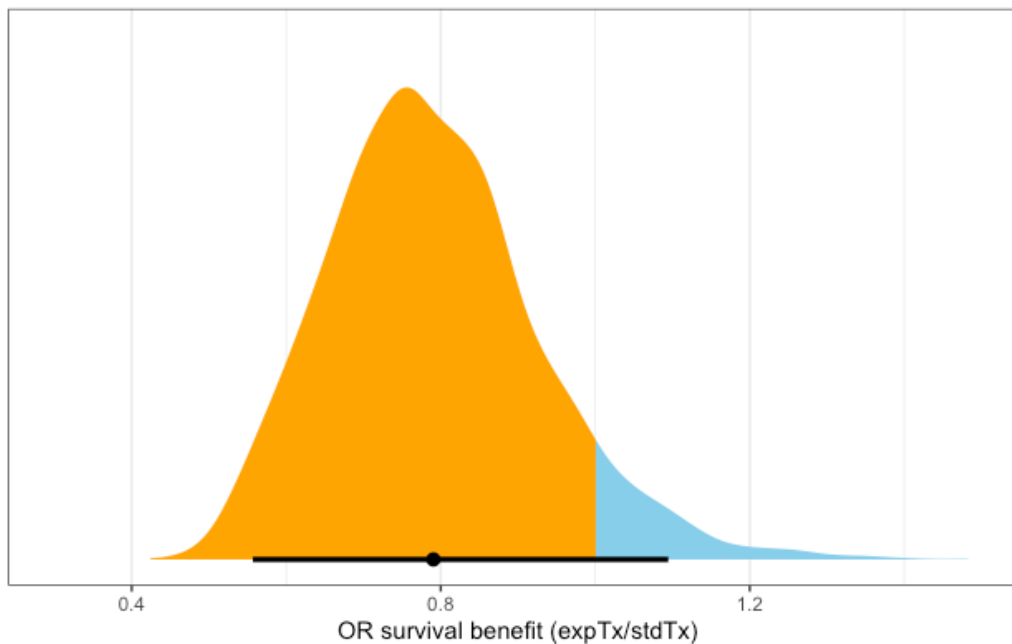


+

likelihood



CASTLE HTx odds ratio (expTx/stdTx) with informative prior
composite outcome = all cause death, LVAD implantation, or urgent heart Tx



prior



posterior

expTx = ablation
stdTx = standard treatment
LVAD = left ventricular assist device
Tx = transplant

Example # 4

A “negative” trial

The NEW ENGLAND
JOURNAL of MEDICINE

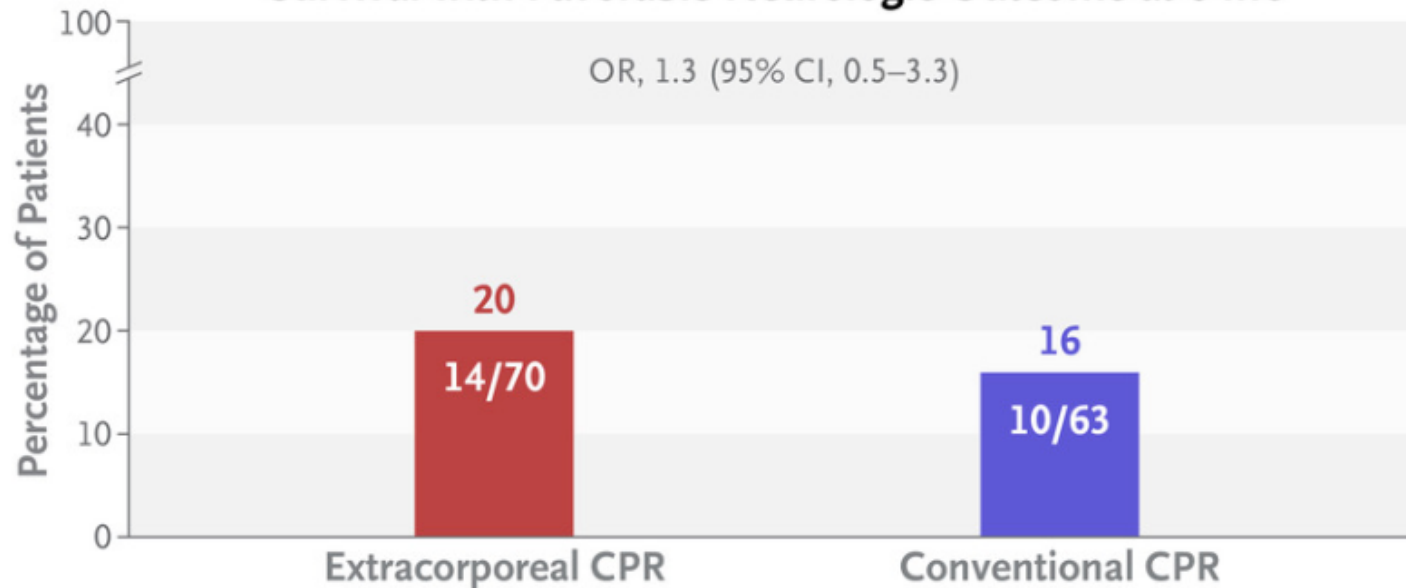
ESTABLISHED IN 1812

JANUARY 26, 2023

VOL. 388 NO. 4

Early Extracorporeal CPR for Refractory Out-of-Hospital
Cardiac Arrest

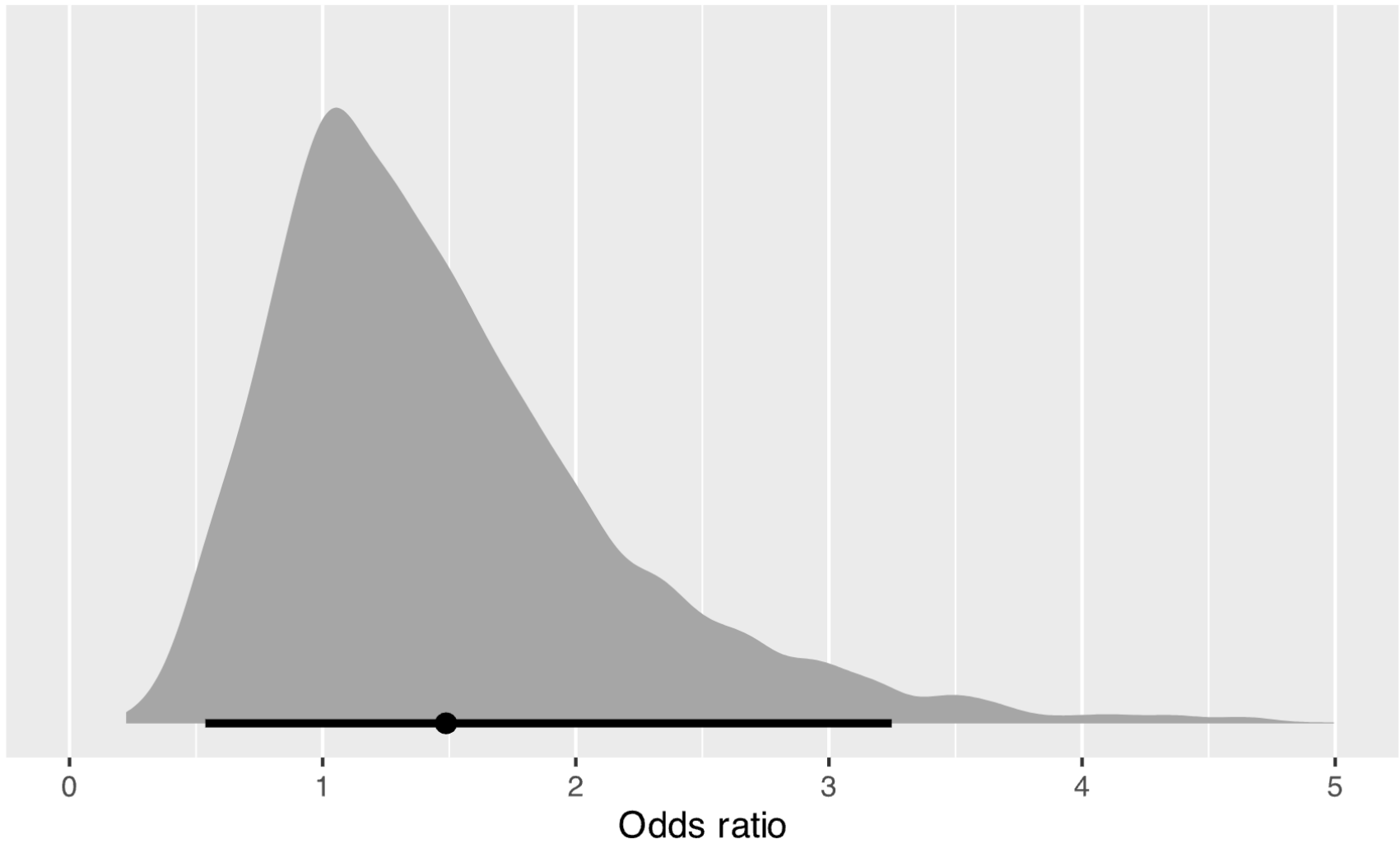
Survival with Favorable Neurologic Outcome at 6 Mo



CONCLUSIONS

In patients with refractory out-of-hospital cardiac arrest, extracorporeal CPR and conventional CPR had similar effects on survival with a favorable neurologic outcome. (Funded by the Netherlands Organization for Health Research and Develop-

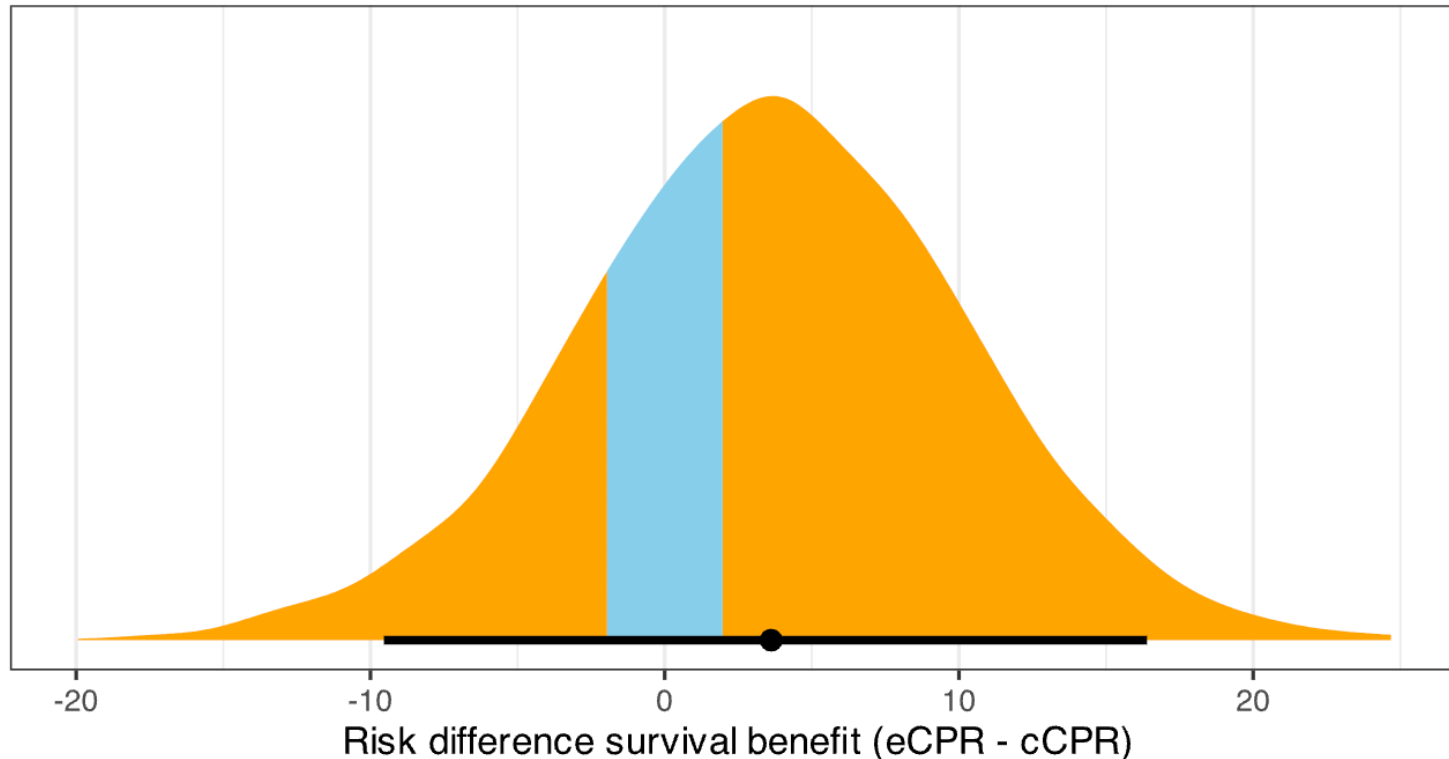
Reproducing results with Bayesian analysis (using vague prior)



Bayesian analysis (vague prior)

Risk difference of survival benefit (/100 patients)

Blue area = range of practical equivalence (ROPE) $\pm 2\%$



Remember the authors' conclusions **"eCPR and cCPR had similar effects on survival"**

What is the probability this is true? Need to define *similar*?

Assuming ± 2 lives / 100 is similar,

blue area represents this equivalence probability, 20.8%.

There remains a 60.2% that eCPR offers a clinically meaningful survival benefit (orange area to right of blue area).

Bayes has certainly deepened our appreciation of this data

Bayesian analysis (informative prior)

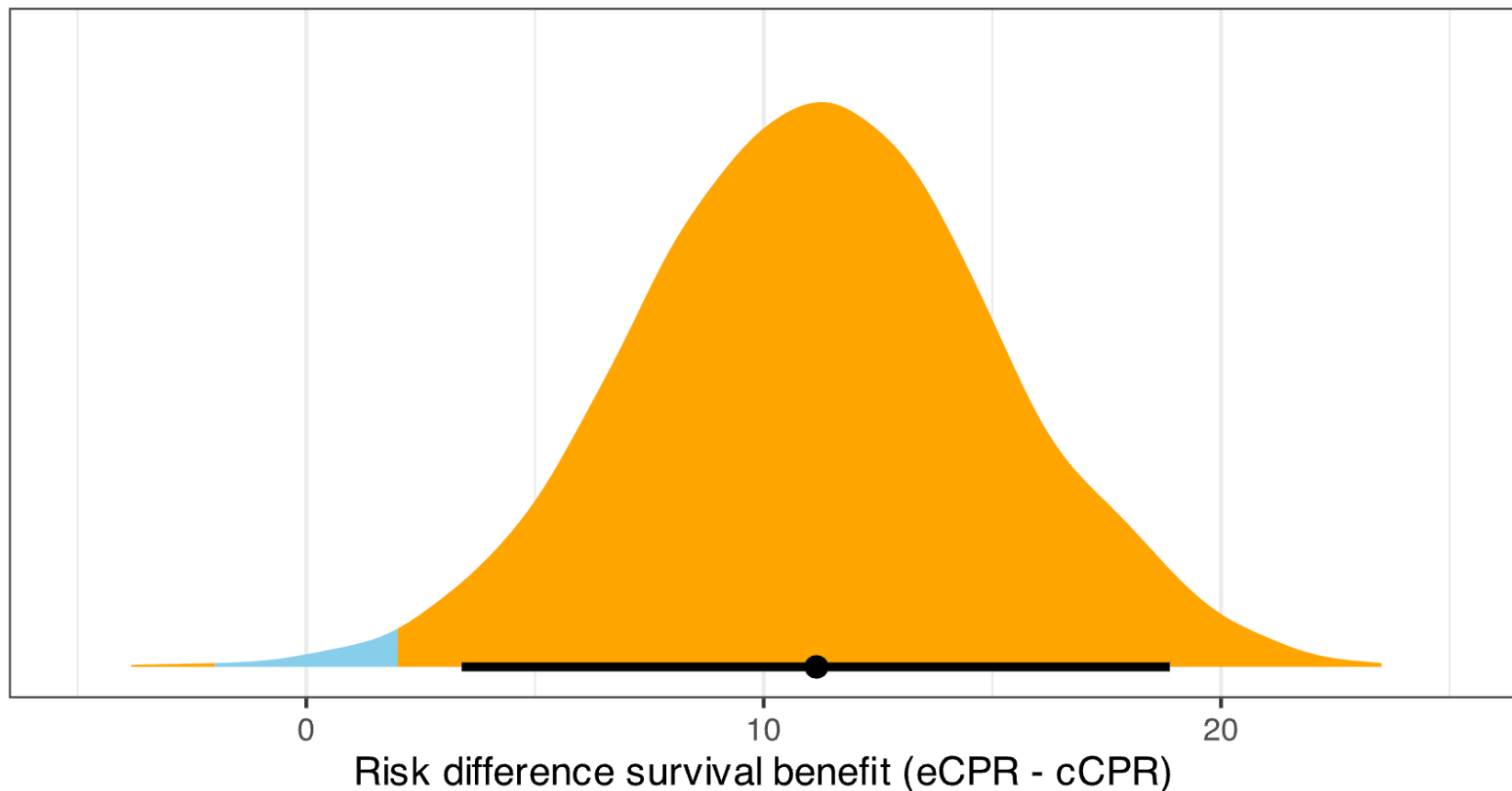
Some prior information exists from 2 previous RCTs and the Bayesian analysis can take this information into account.

PRAGUE & ARREST trials (combined) 25 successes and 122 failures in cCRP (beta(25,122)).

PRAGUE & ARREST trials (combined) 44 successes and 94 failures in eCRP (beta(44,94)).

Risk difference of survival benefit with informative prior

Blue area = range of practical equivalence (ROPE) $\pm 2\%$



Example # 3

Updating prior knowledge with new evidence

Ticagrelor Compared to Clopidogrel in aCute Coronary syndromes – TC4 a pragmatic cluster randomized controlled trial

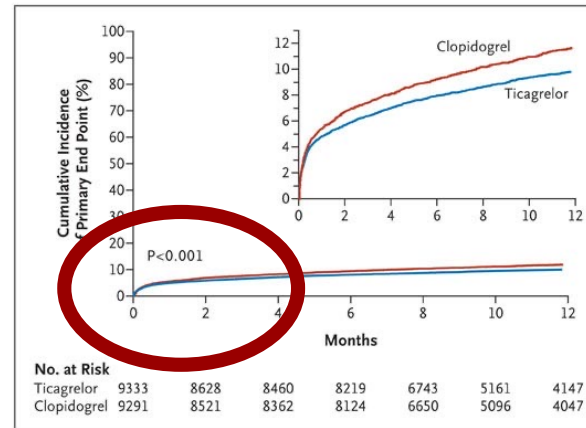
EBM publications & guidelines

PLATO

The NEW ENGLAND
JOURNAL of MEDICINE

ESTABLISHED IN 1812 SEPTEMBER 10, 2009 VOL. 361 NO. 11

Ticagrelor versus Clopidogrel in Patients with Acute
Coronary Syndromes



CCS Guidelines 2012 & 2018



Canadian Journal of Cardiology 29 (2013) 1334–1345

Society Guidelines

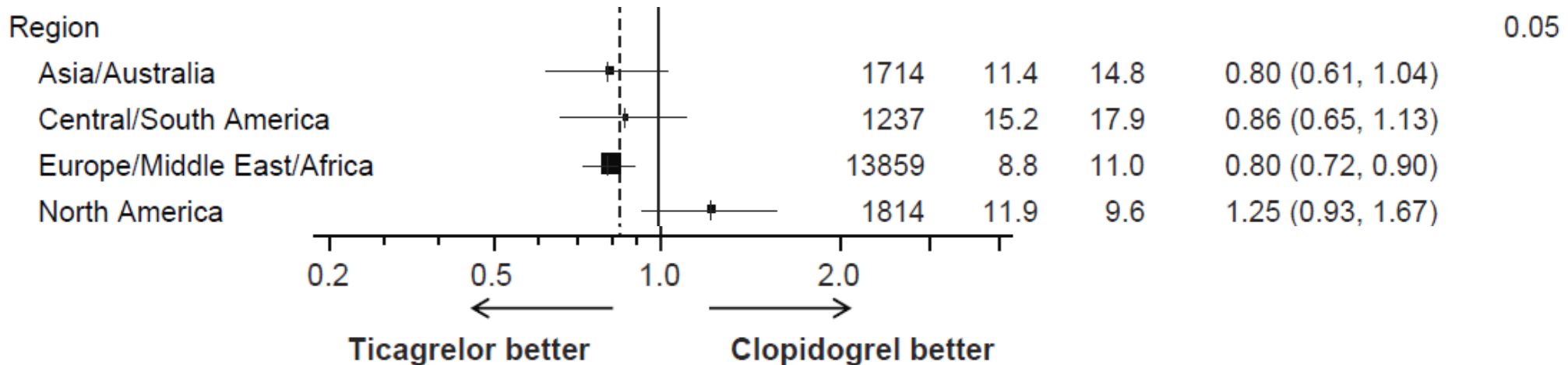
Focused 2012 Update of the Canadian Cardiovascular Society Guidelines for the Use of Antiplatelet Therapy

2. We recommend ticagrelor 90 mg twice daily over clopidogrel 75 mg daily for 12 months in addition to ASA 81 mg daily in patients with moderate to high risk NSTEACS (as defined in PLATO¹⁶: ≥ 2 or more of (1)

(Strong Recommendation, High-Quality Evidence)

Some (transient) doubters

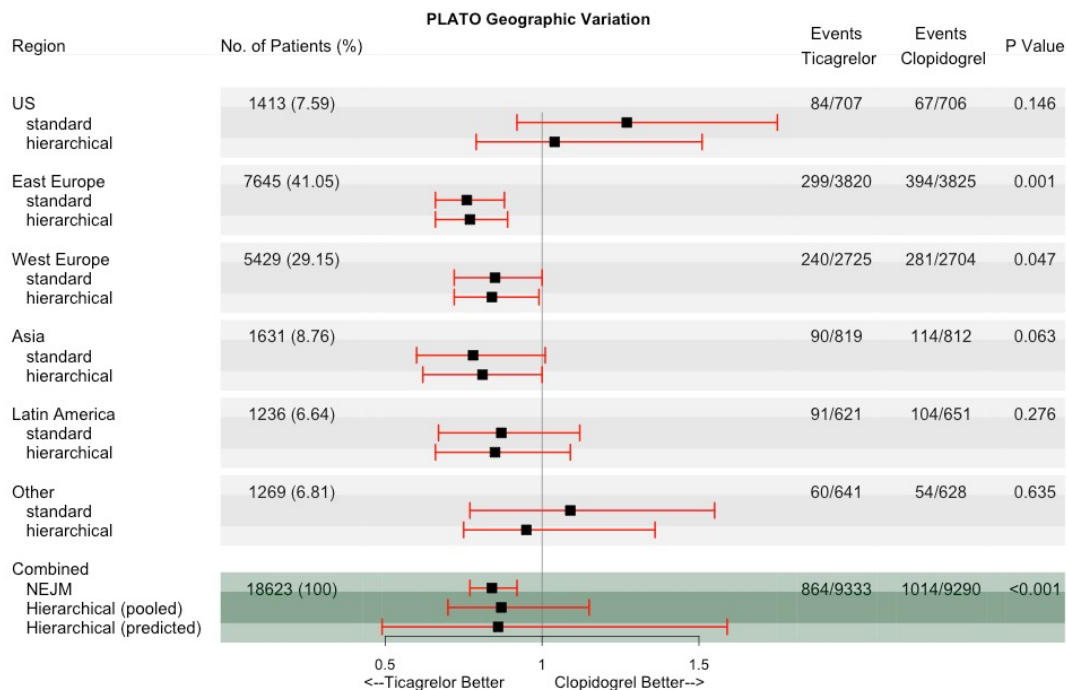
- FDA refused 1st review in 2009 , accepted 2nd in 2011
dissenting opinions (6-4)
 - “Lack of Robustness of PLATO Superiority with Failure in the US Makes a Confirmatory Study Mandatory.”
 - “Besides failure in the US, superiority was only evident in the adjudicated results.”



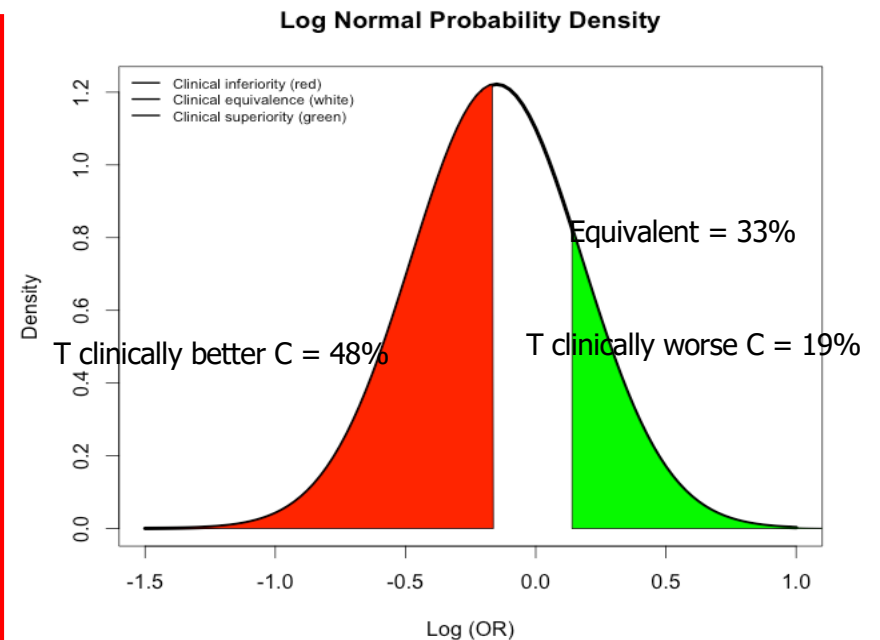
Why different conclusions – same data

- 1st review emphasis on **separate** models – “splitters”
- 2nd review emphasis on **pooled** model – “lumpers” treats all patients as identical -> inferences on “average patient”
- 3rd model option **hierarchical** model – borrowing information

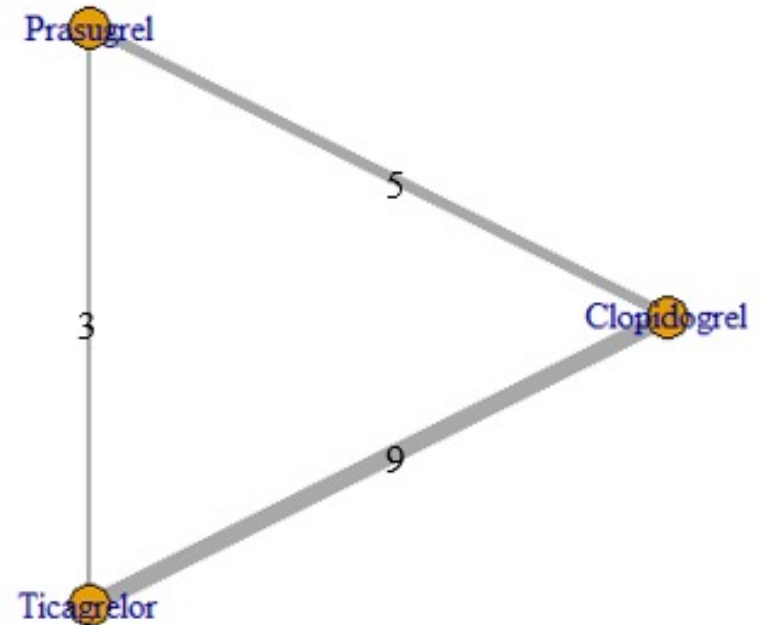
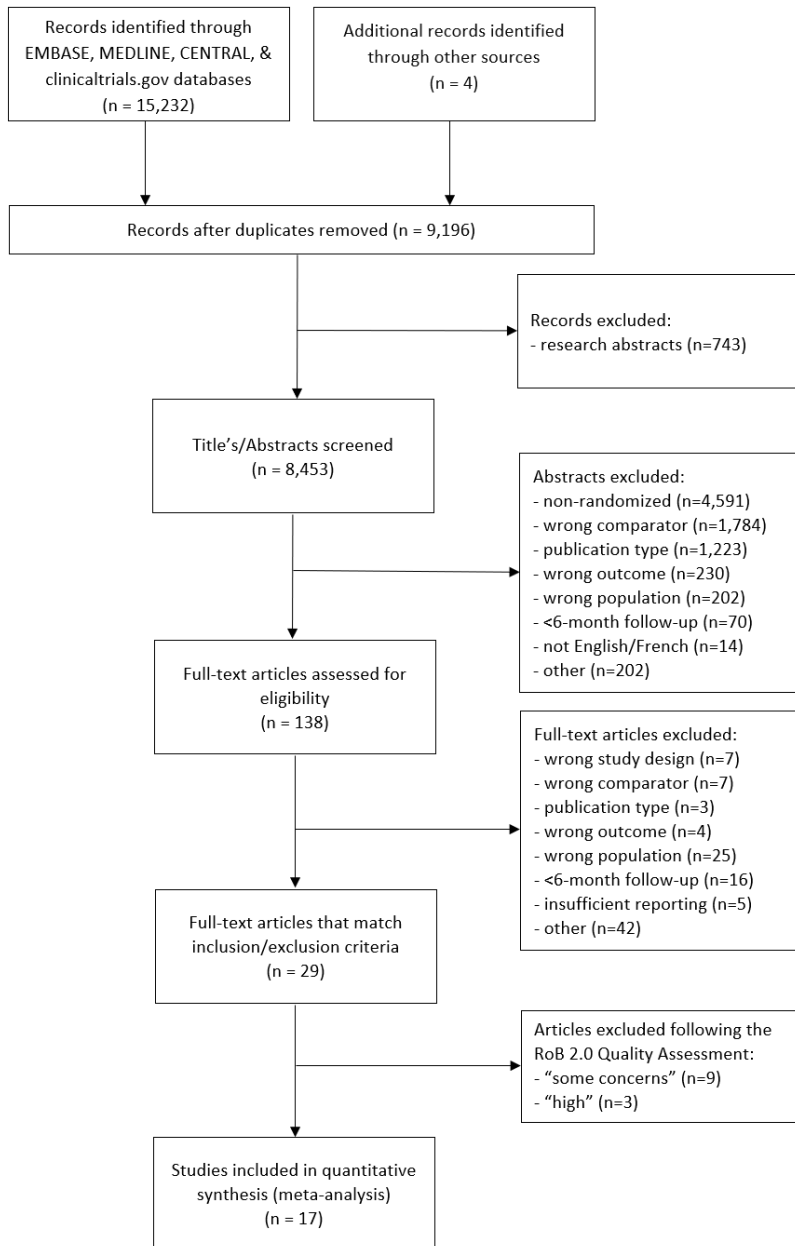
Hierarchical



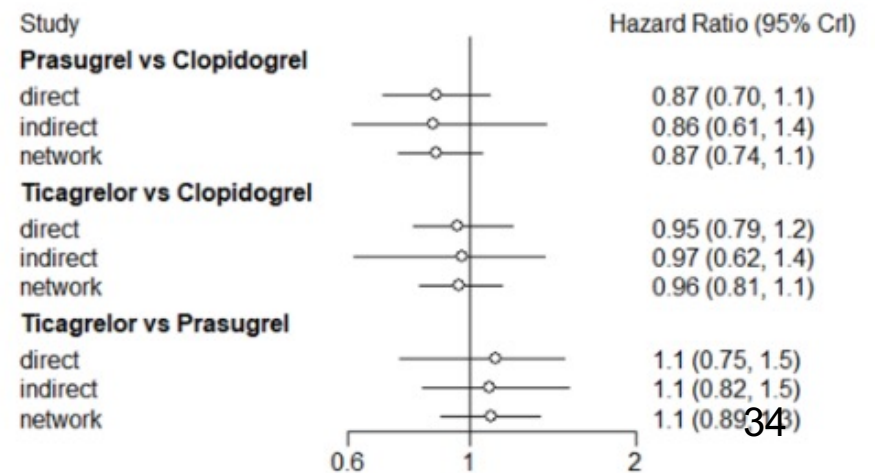
Bayesian



Bayesian network meta-analysis



MACE



The (PICO) research question

Is a DAPT regime of ticagrelor / aspirin superior to clopidogrel / aspirin in reducing cardiovascular (CV) events in patients undergoing percutaneous coronary interventions (PCI) following an acute coronary syndrome (ACS)?

- Population – ACS pts post PCI
- Intervention - ticagrelor / aspirin
- Comparator - clopidogrel / aspirin
- Outcome - death or CV hospitalizations

Doing New Research? Don't Forget the Old

Nobody should do a trial without reviewing what is known

Mike Clarke

On May 2, 1898, George Gould used his address to the founding meeting of the Association of Medical Librarians in Philadelphia to present a vision of the future of health information. 'I look forward,' he said, 'to such an organisation of the literary records of medicine that a puzzled worker in any part of the civilised world shall in an hour be able to gain a knowledge pertaining to a subject of the experience of every other man in the world' [1]. Has his vision been realised?

good quality, but some of it is not. Thus, anyone wishing to use the health literature to make well-informed decisions must both identify the relevant research from amidst this vast amount of information and then appraise it. This is an impossible task for many. Even though making access to the literature easier and cheaper will increase the ability of people to find research, it will also reveal just how much information there is out there and how daunting is the task of making sense of it.

with one or more search engines? Almost certainly, as the speed of the search increased through these four

Citation: Clarke M (2004) Doing new research? Don't forget the old. *PLoS Med* 1(2):e35.

Copyright: © 2004 access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, any medium, provided the original author and source are credited.

Mike Clarke is director, Cochrane Centre, mclarke@cochrane.org

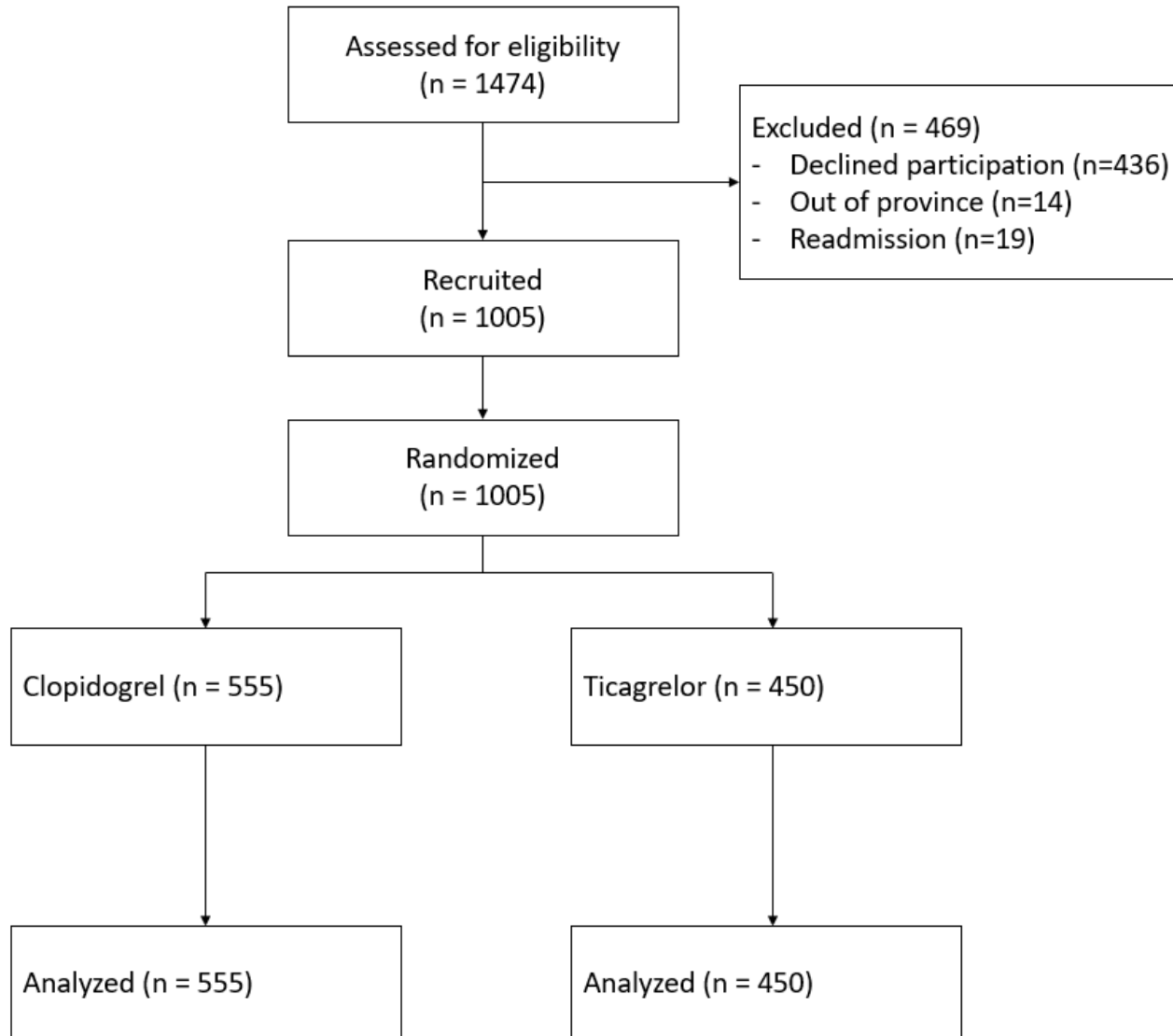
Box 1. Practical Suggestions for Researchers

- Conduct a systematic review of your research question before embarking on a new study, or identify a relevant review done by someone else.
- Design your study to take account of the relevant successes and failures of the prior studies, and of the evidence within them.
- Discuss the findings of your study in the context of an updated systematic review of relevant research.
- Publish the systematic review within, alongside, or shortly after the report of your study.
- Provide information from your study to others doing systematic reviews of similar topics.

Methods

- From Oct 2018 to Mar 2021, ACS patients with PCI
- Randomized into pragmatic, open-label, time clustered, trial
- 1^o endpoint composite of all-cause mortality, non-fatal MI, or ischemic stroke (MACE).
- 1^o safety endpoint was hemorrhagic stroke or GI bleeding requiring hospitalization.
- Outcomes were ascertained within 12 months using administrative databases
- Bayesian Cox proportional hazard models were used to evaluate all outcomes, using vague, “skeptical”, “enthusiastic”, and “summary” informative priors.

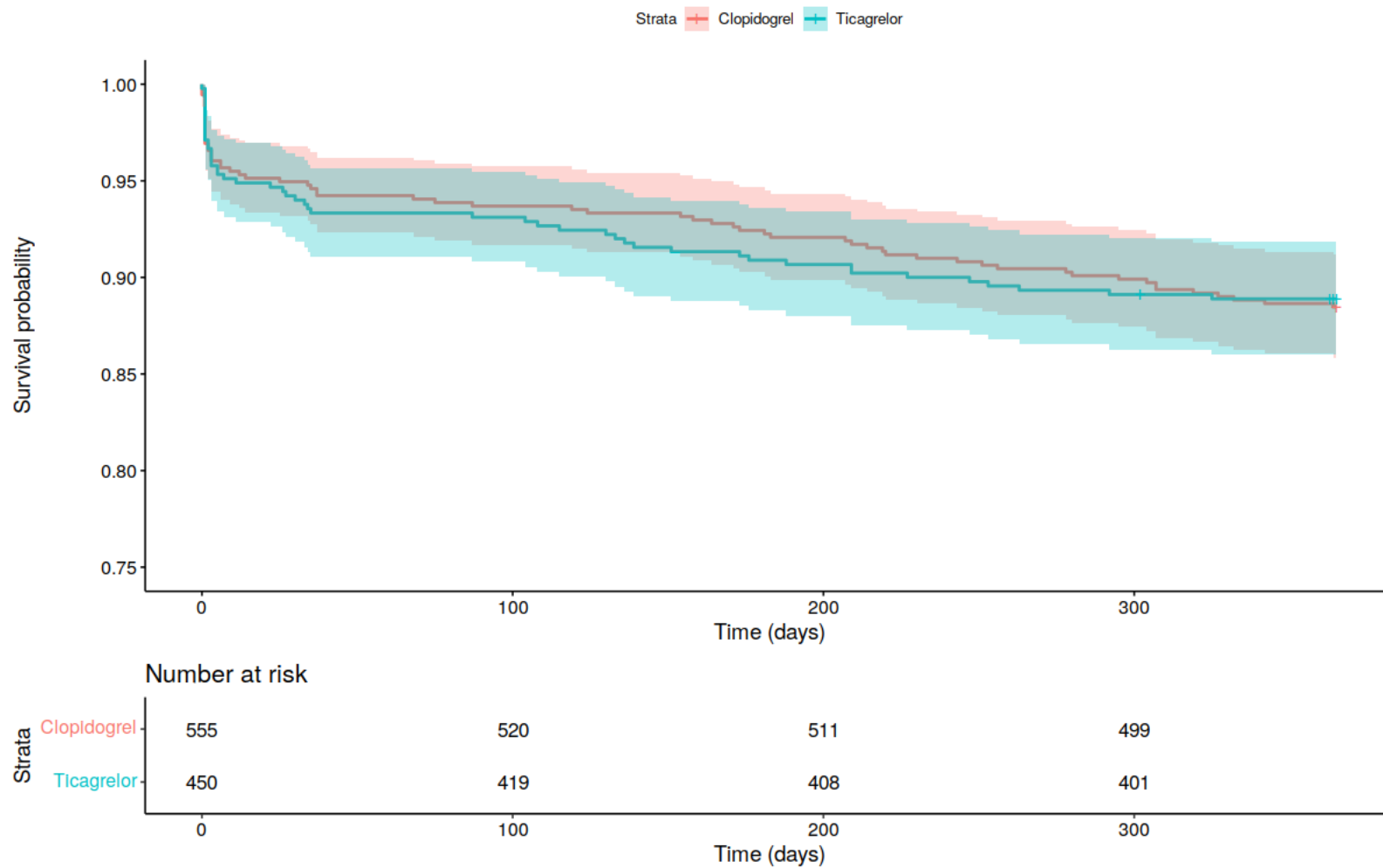
Results



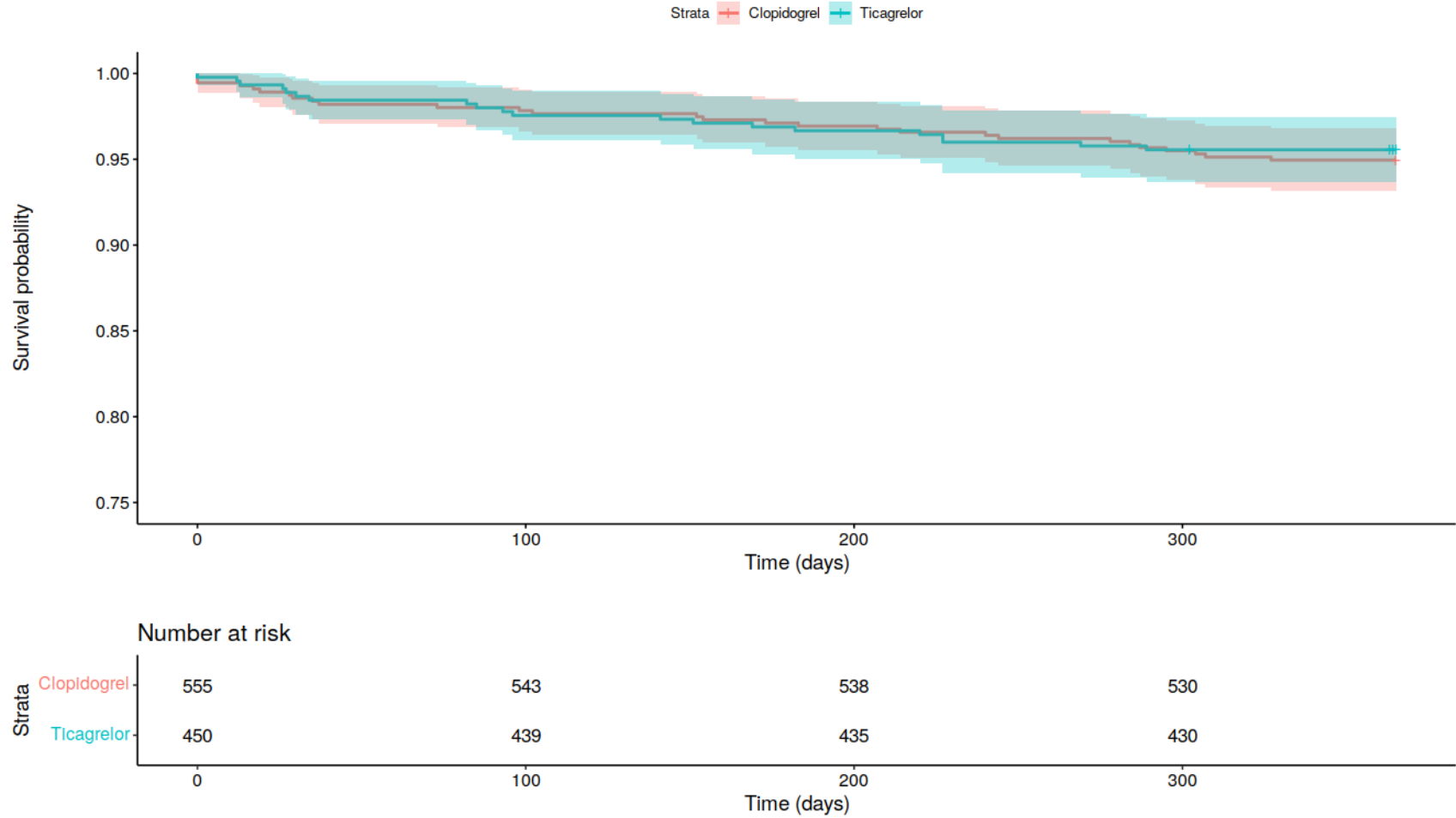
Results

	Clopidogrel	Ticagrelor
n	555	450
Age (mean (SD))	67.56 (10.92)	65.16 (11.33)
Sex (male), n (%)	420 (75.7)	338 (75.1)
Height, cm (mean (SD))	170.60 (9.47)	171.04 (9.30)
Weight, kg (mean (SD))	83.05 (21.99)	83.31 (17.78)
Smoking status, n (%)		
Current	136 (24.6)	110 (24.6)
Race, n (%)		
Caucasian	453 (81.6)	376 (83.6)
Previous DAPT, n (%)		
No	409 (74.1)	341 (76.3)
ACS diagnosis, n (%)		
STEMI	116 (20.9)	94 (20.9)
NSTEMI	210 (37.9)	207 (46.1)
Unstable Angina	89 (16.1)	69 (15.4)
Other	139(25.1)	79(17.6)
Hypertension, n (%)	387 (69.9)	300 (67.0)
SBP (mean (SD))	140.62 (22.23)	140.02 (22.62)
DBP (mean (SD))	79.72 (13.69)	80.43 (14.99)
Heart rate (mean (SD))	72.94 (15.43)	72.39 (15.11)
Dyslipidemia, n (%)	376 (68.0)	301 (67.2)
Diabetic, n (%)	185 (33.5)	139 (31.0)
Previous MI, n (%)	159 (28.6)	120 (26.9)
Previous PCI, n (%)	144 (25.9)	114 (25.4)
CHF, n (%)	32 (5.8)	15 (3.3)

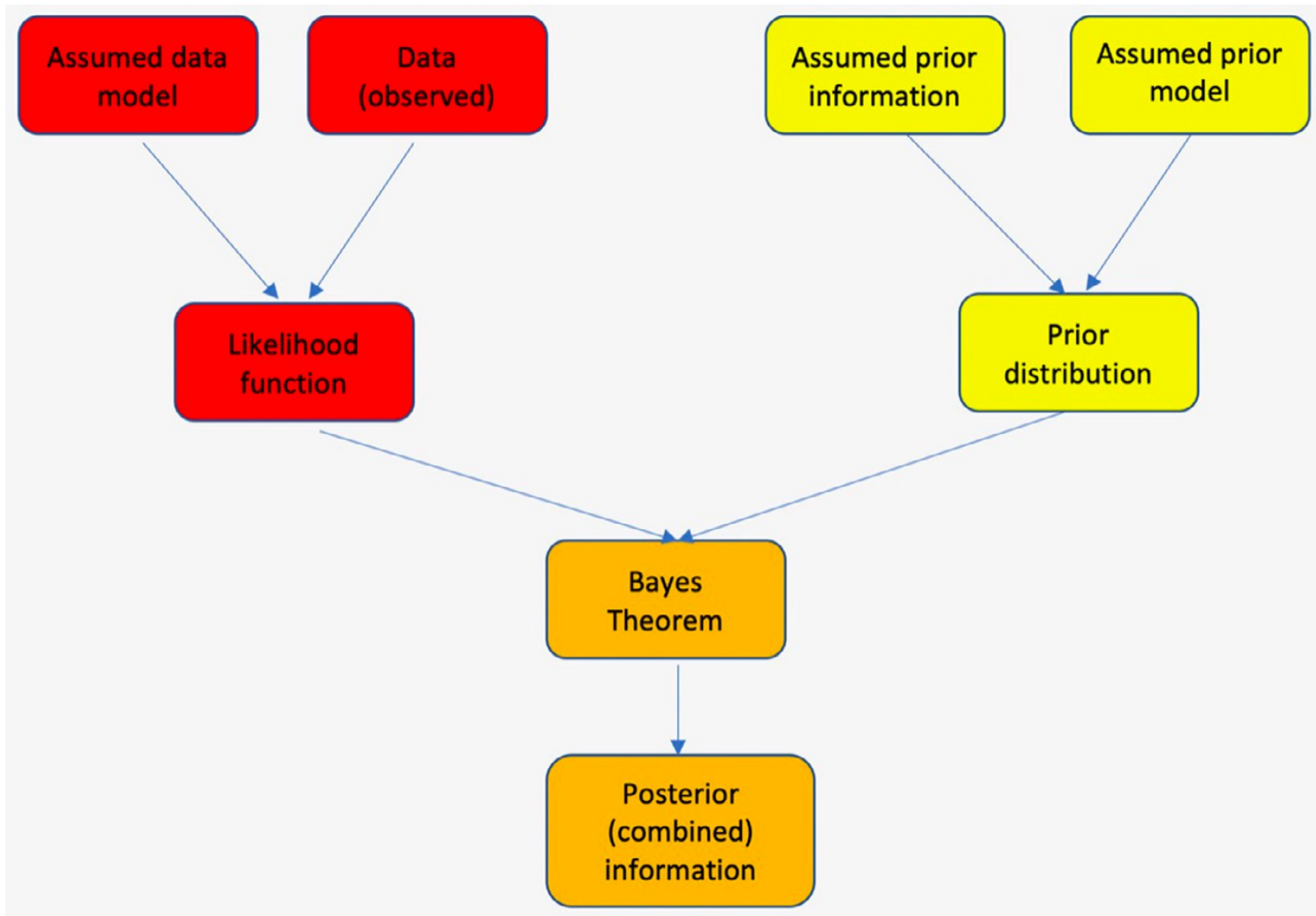
Results – Kaplan Meier Curve (MACE)



Results – Kaplan Meier Curve (Bleeding)



Bayesian paradigm



Which prior?

- **Vague** – only data from our trial
- **Enthusiastic** – data from largest most positive trial (PLATO)
- **Skeptical** – data from NA PLATO subgroup
- **Summary** - data from our Bayesian meta-analysis of all RCTs

Results (MACE)

Range of practical equivalence btw HR 0.9 – 1.1



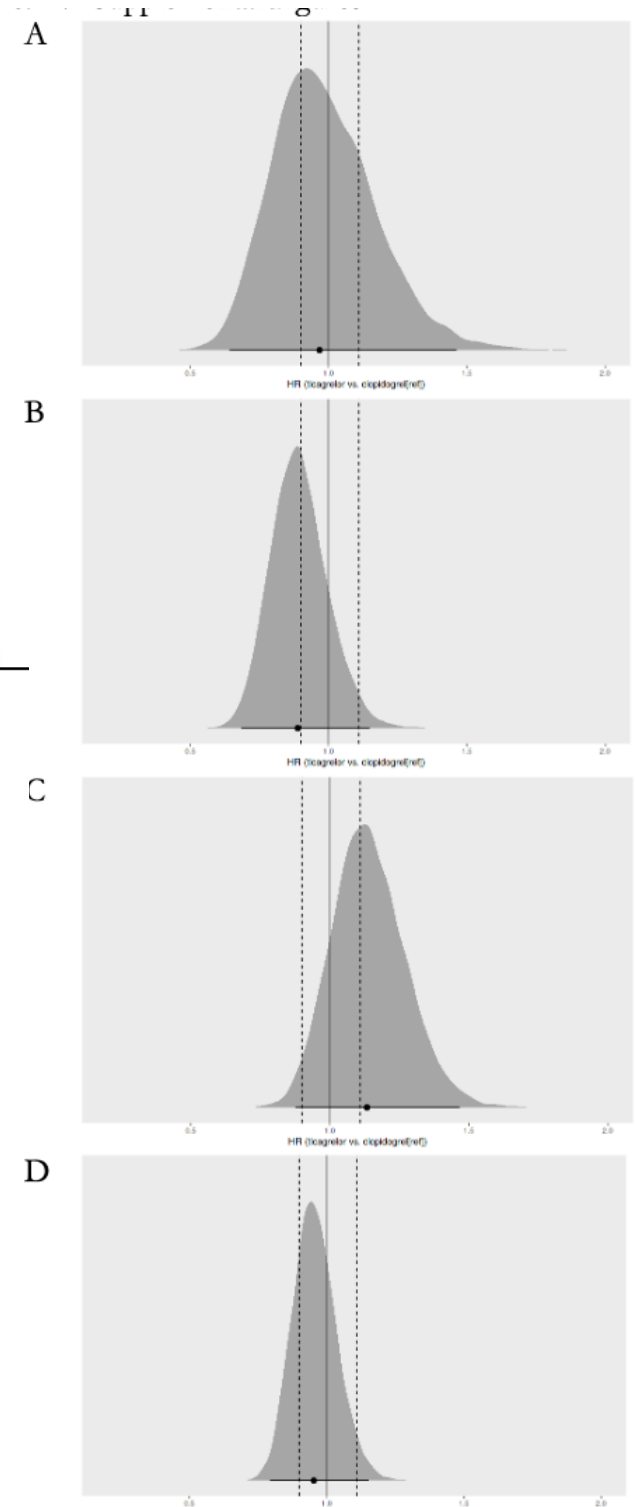
Prior	HR (95% CrI) <i>Pooled</i>	Posterior distribution		
		<u>Pr</u> $HR < 0.9$	<u>Pr</u> $HR [0.9, 1.1]$	<u>Pr</u> $HR > 1.1$
A Vague	0.97 (0.67, 1.40)	0.35	0.40	0.25
C skeptical	1.13 (0.90, 1.42)	0.02	0.38	0.60
B enthusiastic	0.89 (0.71, 1.11)	0.55	0.42	0.03
D summary	0.95 (0.81, 1.12)	0.24	0.72	0.04



Clinically meaningful benefit HR < 0.9



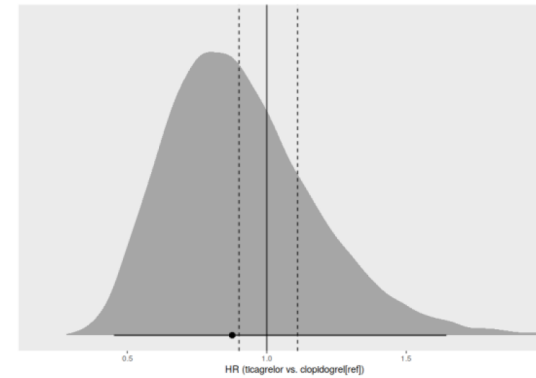
Clinically meaningful harm HR > 1.1



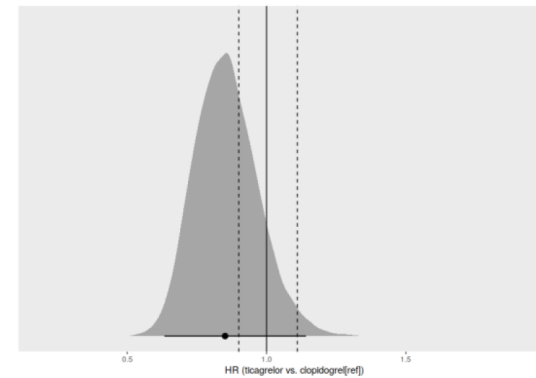
Results (Bleeding)

	Clopidogrel N=555	Ticagrelor N=450	Prior	HR (95% CrI) <i>Pooled</i>	Posterior distribution		
					Pr _{HR<0.9}	Pr _{HR[0.9, 1.1]}	Pr _{HR>1.1}
Bleeding	28 (5.0%)	20 (4.4%)	E Vague	0.88 (0.49, 1.50)	0.53	0.25	0.22
			G skeptical	1.01 (0.76, 1.34)	0.22	0.51	0.27
			F enthusiastic	0.85 (0.66, 1.10)	0.67	0.31	0.02
			H summary	1.06 (0.97, 1.16)	0.00	0.77	0.23

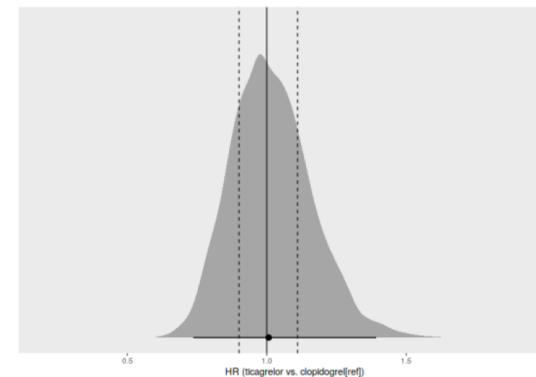
E



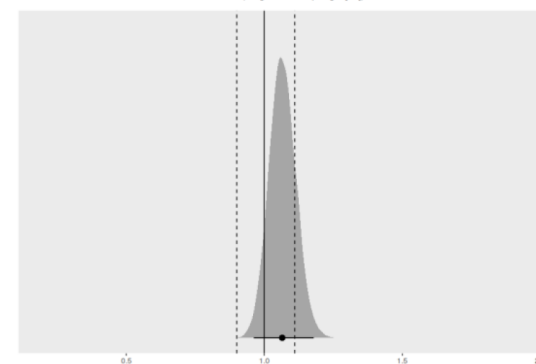
F




G



H



TC4 - Conclusions

- 1st RCT with NA pts since PLATO (2009),  NA evidence base > 50%
- Accomplished for 300K, original trial > 100MM
- Regardless of the choice of prior, there is only a low probability that ticagrelor (@\$1200/y) is clinically superior to clopidogrel (@\$168/y)
- Weak evidence ($\approx 20\%$ probability) for clinical important ($HR > 1.1$) risk of excessive bleeding with ticagrelor
- Additional annual Quebec health care cost \$25MM for a ticagrelor first policy needs re-evaluation
- This conclusion is also supported by
 - Plato hierarchical reanalysis
 - Bayesian network meta-analysis

Final thoughts of Bayesian RCT analyses

- can provide meaningful probability statements
- can avoid common NHST misinterpretations (e.g. absence of evidence is evidence of absence)
- can better account for uncertainties by considering complex models
- can allow for updating of existing knowledge with new evidence

Acknowledgements

**TC4 trial comes from
Stephen Kutcher's PhD thesis**

Financial support



MUHC- RI doctoral training grants



CIHR project grant (#PH2-388823)



FRQS (EBM Chair) salary support