NEJM Statistical Guidelines for Authors

Under the Hood

Dave Harrington

18 September 2019

My coordinates

- Department of Biostatistics, Harvard T.H. Chan School of Public Health
- Department of Data Sciences, Dana-Farber Cancer Institute
- Statistical Consultant, New England Journal of Medicine
- davidharrington@g.harvard.edu

Guidelines were joint effort

Statistical Consultants

- Ralph D'Agostino, BU
- Constantine Gatsonis, Brown
- D Harrington, HSPH, DFCI
- Joe Hogan, Brown
- David Hunter, Oxford
- Sharon-Lise Normand, HMS

Jeff Drazen, former Editor-in-Chief Mary Beth Hamel, Executive Deputy Editor

All Deputy Editors read and approved guidelines

Statistical Review at NEJM

Six paid statistical consultants (reviewers)

We meet weekly with the editors

Review all research articles that editors decide to move through the system

- We see approximately 5% of the 5,000 - 6000 submitted articles

Statistical reviewers

- Help set statistical 'policy'
- Work toward consistency in reviews

Background

Revised author guidelines on statistical reporting posted on NEJM website July 1, 2019

https://www.nejm.org/author-center/new-manuscripts

Editorial describing the guidelines published 18 July 2019

- N Engl J Med 2019; 381:285-286

Guidelines cover many aspects of statistical reporting

Editorial emphasized section on p-values

Our goals

Emphasize aspects of good practice in analysis and reporting

- All points discussed have been in the statistical literature for decades

Add consistency to

- Statistical reviewing at NEJM
- Standards of analysis in reporting of results

Move toward standards of reporting in observational studies

- Elements of STROBE (Strengthening of Reporting of Observational Studies in Epidemiology)
- https://strobe-statement.org/

Goals . . .

Show the data

- Point estimates and confidence intervals
- Reduce irrelevant p-values

Acknowledge that we cannot influence design

Every good paper should have a path to publication, even with

- No planned adjustment for multiplicity
- No specified method for missing data

Goals . . .

Minimize misleading or misunderstood statistical jargon

- Nominal p-value
- Nominally significant
- 'Controlling' for confounders in regression models

Our constraints

For busy clinicians, conclusions in a paper should be clear, direct, and supportable.

Statistical background of NEJM readers

Word limits in NEJM make it difficult to discuss nuance.

- 250 words for abstracts
- 2700 words for article

Normal review time and weekly publication schedule make transition to major changes difficult.

Content of the guidelines

For all studies

- Describe sample size and power calculations
- p-values should be two-sided.

For RCTs:

- Provide protocol and statistical analysis plan (SAP).
- Analysis should match design
- Specify and follow multiplicity adjustments in SAP
- No plan for multiplicity adjustment: report point estimates with confidence intervals, no p-values

The guidelines . . .

RCTs:

- Forest plots for subgroups, but no p-values
- OK to use unadjusted p-values for safety comparisons
- Report event rates as well as hazard ratios
- Provide CONSORT diagram

The guidelines

Observational studies:

- Provide a SAP if you have one
- No SAP - provide a clear analysis plan in methods section or supplement
- No plan for multiplicity adjustment: report point estimates with confidence intervals, no *p*-values
- Show distributions of the measured confounders
- Model diagnostics and sensitivity analyses in the supplement.
- If possible, retest findings in a validation dataset.

What about those unadjusted confidence intervals?

Is a confidence interval always a test in disguise?

Interpretation of confidence intervals by readers

Typical insert I recommend to authors for methods:

> *Because the statistical analysis plan did not include a provision for correcting for multiplicity, secondary and other outcomes are reported as point estimates and 95% confidence intervals. The widths of the confidence intervals have not been adjusted for multiplicity, so the intervals should not be used to infer definitive treatment effects for secondary outcomes.*

Post hoc adjustments for multiplicity

Personal communication from Rebecca Betensky

For each of 10,000 replicates in a simulation

- Simulate collection of correlated p-values
- Choose a method of FWER or FDR control that is most favorable to the analyst
  - Minimize the minimum p-value or
  - Maximize the number of adjusted p-values < 0.05
- Average FWER or FDR across replicates

Choice of methods from

- FWER: Bonferroni, Holm, Hochberg, Hommel, Šidák single step, Šidák step down.
- FDR: Benjamini-Hochberg, Benjamini-Yekutieli, two-stage Benjamini-Hochberg.

Results

number of tests	correlation	maximize number of adjusted p<0.05 procedures used			minimize minimum adjusted p procedures used		
		FWER	FDR	ALL	FWER	FDR	ALL
20	0	0.052	0.052	0.053	0.003	0.050	0.053
	0.5	0.032	0.037	0.037	0.001	0.035	0.037
	0.9	0.019	0.026	0.026	0.000	0.026	0.026
50	0	0.046	0.046	0.047	0.009	0.039	0.047
	0.5	0.029	0.034	0.034	0.004	0.031	0.034
	0.9	0.015	0.025	0.025	0.000	0.025	0.025
100	0	0.051	0.052	0.053	0.028	0.025	0.053
	0.5	0.027	0.034	0.034	0.008	0.026	0.034
	0.9	0.010	0.022	0.022	0.000	0.022	0.022

Perils of post hoc adjustment

Does not distinguish between confirmation and discovery

Are all the primary secondary and subgroup analyses part of the family of tests examining a treatment effect?

- If so, reduces power for primary outcome

Unadjusted p-values in safety outcomes

From Metra, et al. N Engl J Med 2019; 381:716-726 DOI: 10.1056/NEJMoa1801291

Table 3. Most Frequent Adverse Events up to and Including Day 5.*

Event	Serelaxin Group (N=3257)	Placebo Group (N=3248)	P Value†
	no. (%)		
Hypokalemia	263 (8.1)	242 (7.5)	0.35
Cardiac failure	192 (5.9)	215 (6.6)	0.23
Muscle spasms	80 (2.5)	49 (1.5)	0.006
Hypotension	77 (2.4)	65 (2.0)	0.32
Headache	74 (2.3)	92 (2.8)	0.15
Constipation	70 (2.1)	57 (1.8)	0.25
Urinary tract infection	63 (1.9)	68 (2.1)	0.65

* Shown are events that occurred in at least 2% of patients in either group.
† P values were calculated with the use of the Cochran–Mantel–Haenszel test.

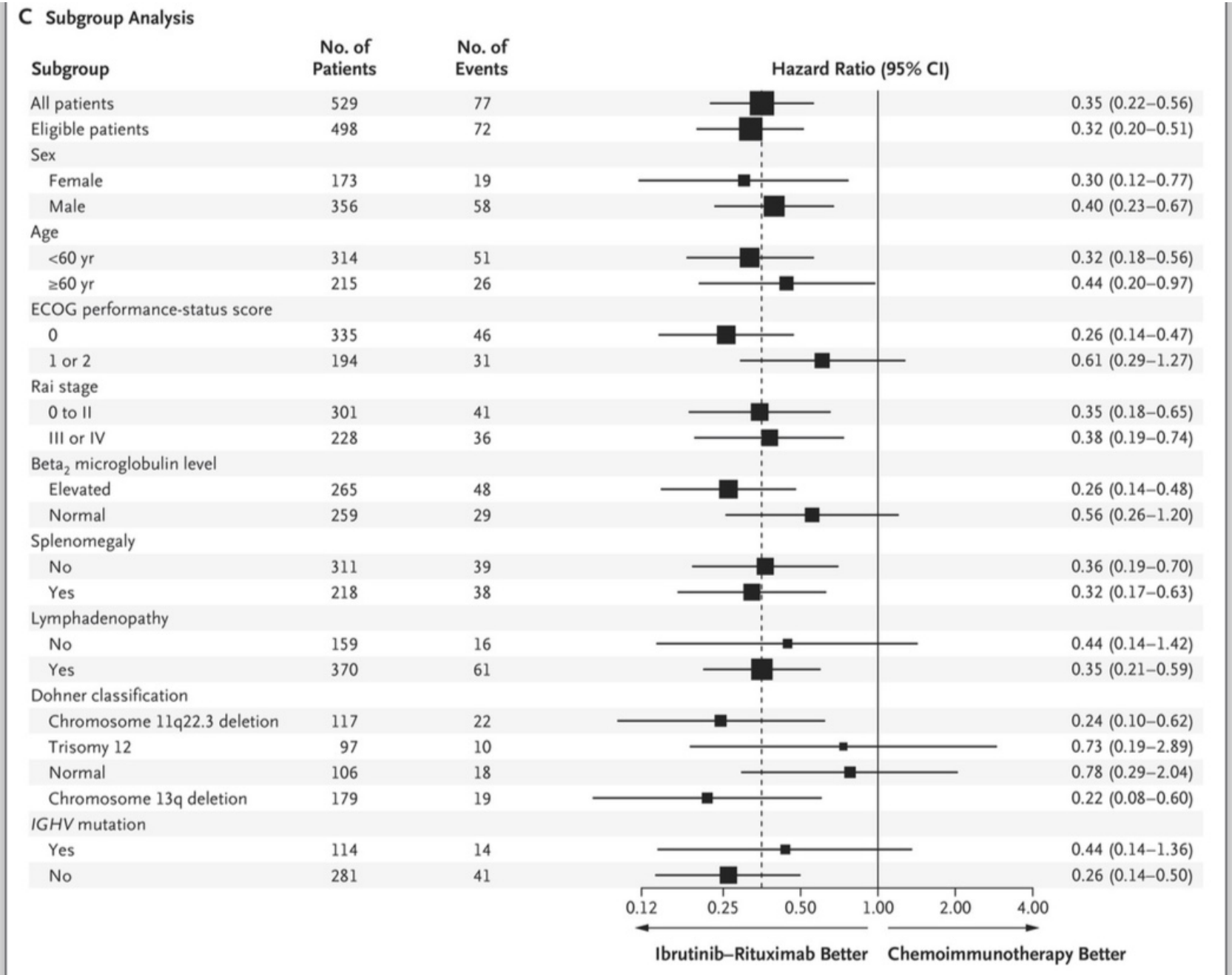
p VALUES FOR INTERACTIONS

Did we do more harm than good in eliminating p-values in forest plots of subgroups?

Deputy editors, authors and some statisticians reluctant to eliminate these

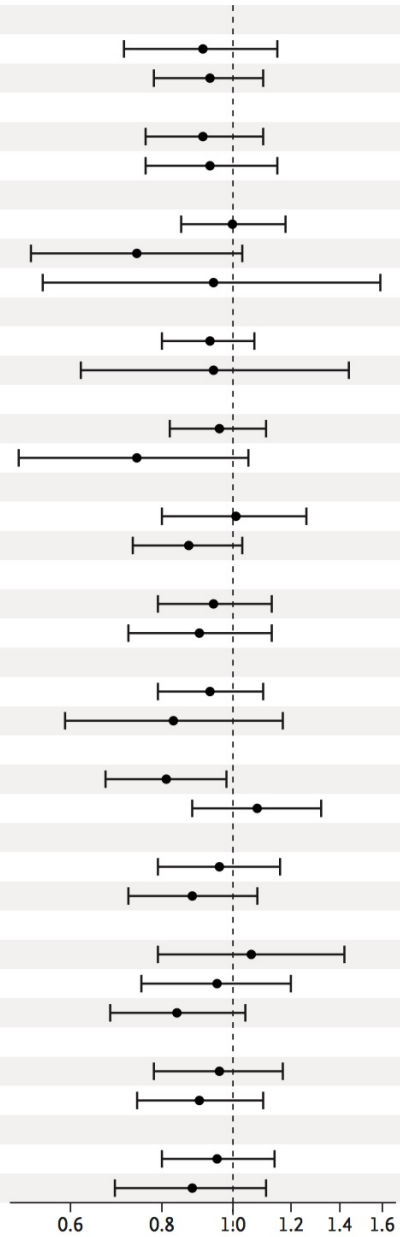
- Authors and editors found lists of non-significant p-values reassuring.
- Easy to recommend that significant p-values be ignored.
- No easy general recommendation for point estimates and confidence intervals for inteaction effects.

New form of forest plots



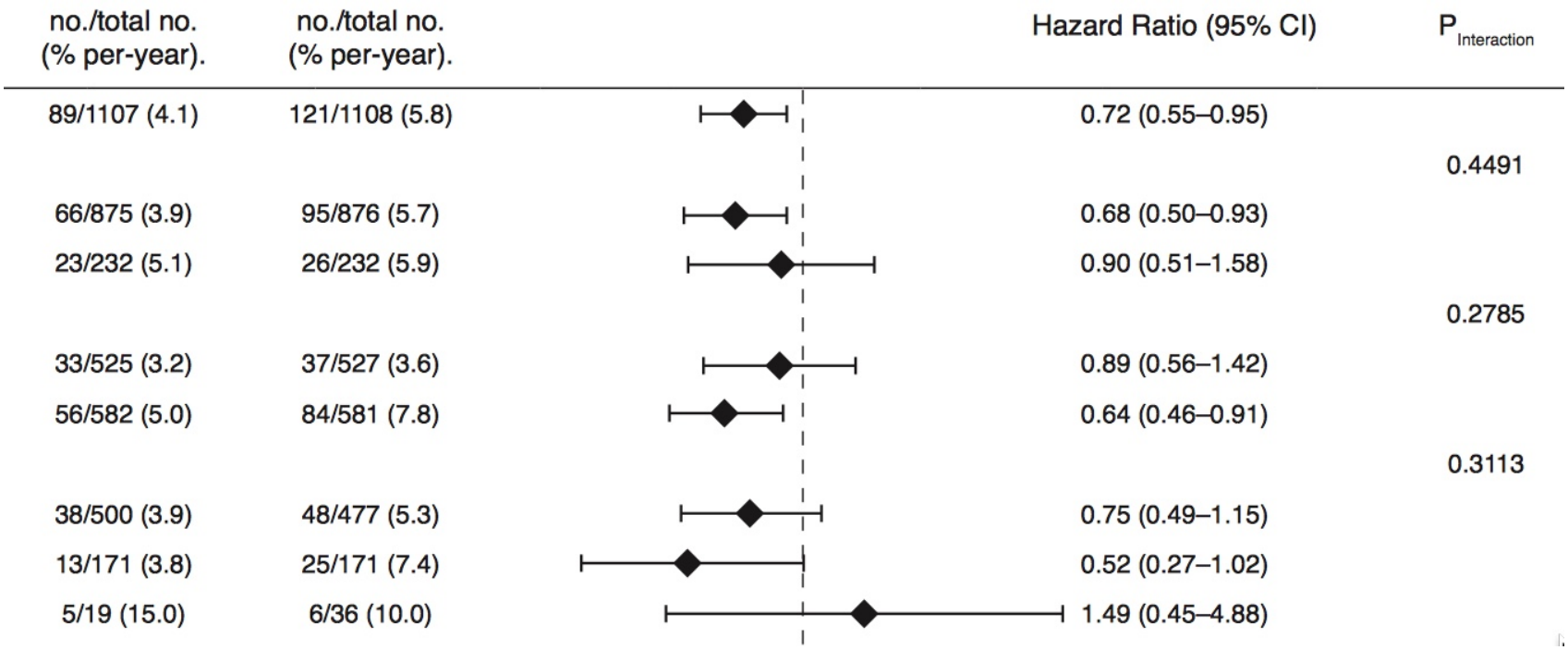
Before change

Subgroup	No. of Participants	n–3 Fatty Acids (no. of participants with event)	Placebo (no. of participants with event)	Hazard Ratio (95% CI)	P Value for Interaction
Age	25,871				0.84
<Median of 66.7 yr	12,859	129	142	0.91 (0.71–1.15)	
≥Median of 66.7 yr	13,012	257	277	0.93 (0.78–1.10)	
Sex	25,871				0.88
Male	12,786	213	233	0.91 (0.76–1.10)	
Female	13,085	173	186	0.93 (0.76–1.15)	
Race	25,304				0.26
Non-Hispanic white	18,046	292	289	1.00 (0.85–1.18)	
Black	5,106	62	83	0.74 (0.53–1.03)	
Other	2,152	26	30	0.94 (0.55–1.59)	
Current smoker	25,485				0.77
No	23,649	340	365	0.93 (0.80–1.07)	
Yes	1,836	41	44	0.94 (0.62–1.44)	
Diabetes	25,860				0.19
No	23,132	334	349	0.96 (0.82–1.11)	
Yes	2,728	52	70	0.74 (0.51–1.05)	
Hypertension	25,698				0.32
No	12,907	151	147	1.01 (0.80–1.26)	
Yes	12,791	231	270	0.87 (0.73–1.03)	
Current cholesterol medication	25,428				0.77
No	15,904	236	252	0.94 (0.79–1.13)	
Yes	9,524	140	154	0.90 (0.72–1.13)	
Parental history of myocardial infarction	22,915				0.56
No	19,262	268	288	0.93 (0.79–1.10)	
Yes	3,653	61	71	0.83 (0.59–1.17)	
Fish consumption	25,435				0.045
<Median of 1.5 servings/wk	13,514	189	232	0.81 (0.67–0.98)	
≥Median of 1.5 servings/wk	11,921	189	176	1.08 (0.88–1.32)	
Randomization in vitamin D portion of trial	25,871				0.56
Placebo group	12,944	200	209	0.96 (0.79–1.16)	
Vitamin D group	12,927	186	210	0.88 (0.72–1.08)	
No. of cardiovascular risk factors	25,871				0.19
0	7,802	92	85	1.06 (0.79–1.42)	
1	8,948	133	141	0.95 (0.75–1.20)	
≥2	9,121	161	193	0.84 (0.68–1.04)	
Baseline aspirin use	25,497				0.68
No	13,927	192	199	0.96 (0.78–1.17)	
Yes	11,570	187	209	0.90 (0.74–1.10)	
Baseline statin use	25,447				0.57
No	16,557	247	260	0.95 (0.80–1.14)	
Yes	8,890	130	147	0.88 (0.69–1.11)	

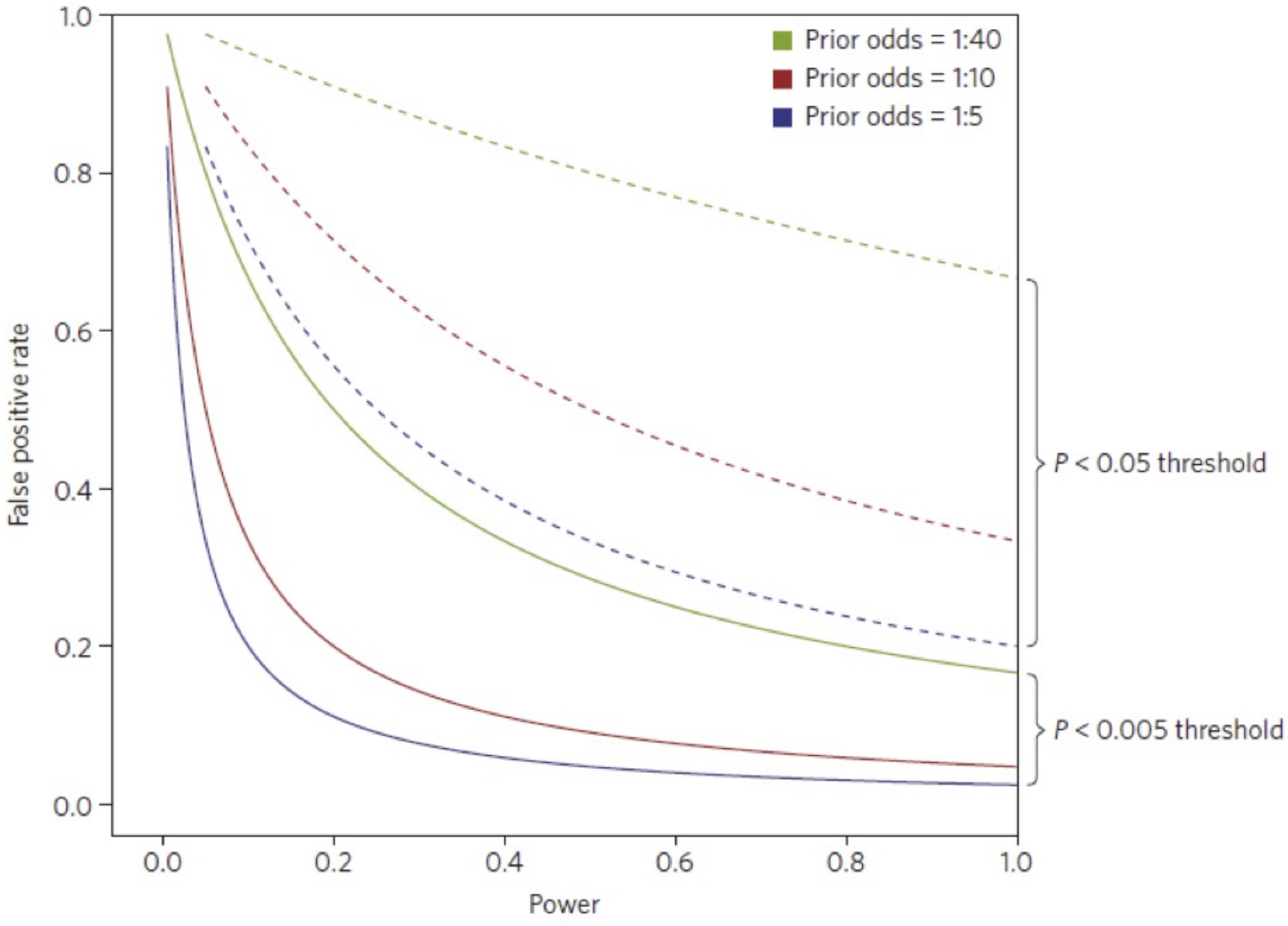


DANA-FAR CANCER INST

Perhaps p-values are helpful



Calculating posterior odds (Nat. Human Behavior, Jan 2018)



Effect of increasing α

Prior odds = 1:10, power = 0.80

Alpha	False Pos. Prob
0.10	0.56
0.15	0.65
0.20	0.71
0.25	0.76
0.40	0.83
0.60	0.88

False Pos. Prob = probability of incorrectly claiming alternative is true, given data

Assumptions behind the calculations

Decision to use an intervention based on single trial, single p-value

Prior odds is knowable, at least approximately

There is a simple alternative for which power is relevant

For practicing clinician, is reproducibility the right metric?

Have we made unsupported claims easier?

Common to encounter phrases in drafts such as

> *A consistent pattern for improved survival . . . was noted across multiple subgroups*

or

> *An exploratory . . . analysis . . . of patients who resumed [intervention] was superior to that of patients who did not*

PDF of slides available at under mskcc_2019 at

https://github.com/dave-harrington/talks

Reading

Benjamin DJ, et al. Redefine statistical significance. Nat Hum Behavior 2018;2:6-10. doi: 10.1038/s41562-017-0189-z

Ioannidis JPA. Retiring statistical significance would give bias a free pass. Nature. 2019;567 (7749):461. doi:10.1038/d41586-019-00969-2

National Academies of Sciences, Engineering, and Medicine. Reproducibility and replicability in science. Washington, DC: National Academies Press, 2019. http://nap.edu/25303

Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond $p<0.05$. Am Stat. 2019;73:1-19. doi:10.1080/00031305.2019.1583913

Dmitrienko A, Bretz F, Westfall PH, et al. Multiple testing methodology. In: Dmitrienko A, Tamhane AC, Bretz F, eds. Multiple testing problems in pharmaceutical statistics. New York: Chapman and Hall/CRC Press, 2009:35-98.