NEJM Statistical Guidelines for Authors

Dave Harrington

17 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

Background

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

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

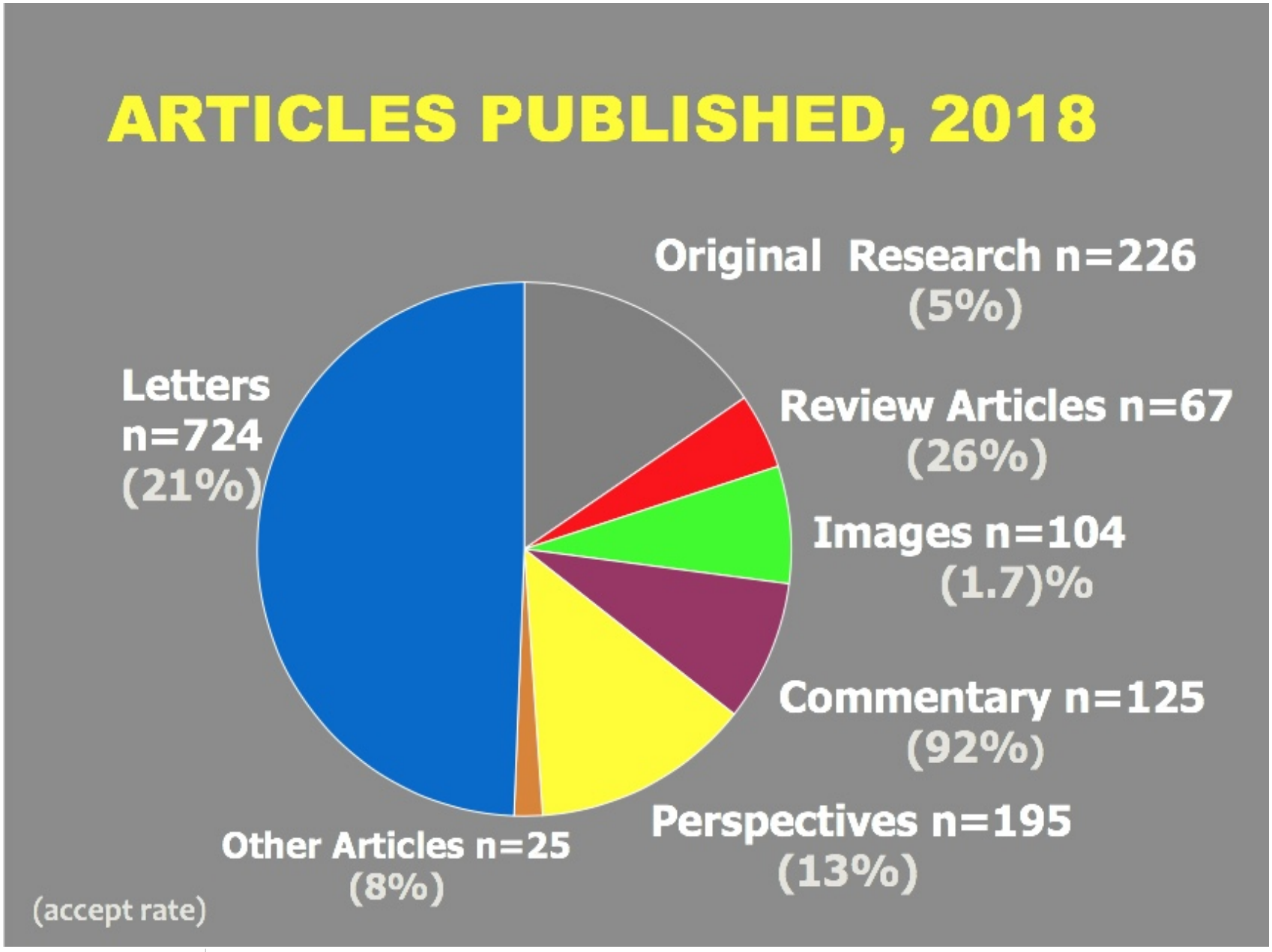
Editorial describing the guidelines published 18 July 2019

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

Guidelines cover many aspects of statistical reporting

Editorial emphasized section on p-values

Navigating NEJM



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

Why did we fuss with p-values?

Many submitted manuscripts sidestepped the issue of multiplicity

- Possibly inflated type 1 error rates when comparing many endpoints
- True in both randomized trials and observational studies
- Becoming more prevalent in complex studies with many measurements

Particularly important in trials with negative primary outcomes

VITAL trial is a useful example

VITAL: Vitamin D and Omega-3 Trial

Manson JE, et al. NEJM 2019; 380:23-32 (3 Jan 2019)

Factorial design, (Vit D vs placebo) $\times$ (Omega-3 vs placebo)

25,871 participants randomized

- 12,927 Vit D vs 12,944 placebo
- 12,933 n-3 vs 12,938 placebo

Primary endpoints: invasive cancer, composite cardiovascular outcome

Omega-3 vs placebo comparison

2 co-primary endpoints: invasive cancer, composite CV outcome

- Both negative

22 secondary/exploratory outcomes

2 traditionally significant

- Total MI, total coronary heart disease (composite)

13 subgroups analyzed for possible treatment interactions

- No 'significant' ($p < 0.05$) interactions

'Black Box' warning in the protocol

> *There was no control for multiple hypothesis testing, and no formal adjustment was made to the P values or confidence intervals. Thus, the results regarding exploratory end points and subgroups should be interpreted with caution.*

What is the potential harm in this?

- None, if the *p*-value is a simple descriptor of a calculation
- Considerably more, if a few significant comparisons are viewed as evidence for a treatment effect
  - Especially when the primary outcome is negative

P-values in medical literature

$p < 0.05$ proposed by Fisher (1926) for single comparisons in randomized experiments:

> *The value for which $P = 0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not.*

A p-value measures how much the observed data disagree with a hypothesis of no treatment effect.

It is often incorrectly interpreted as a measure of the reproducibility of a trial

- Or as the likelihood that the null hypothesis is correct (less than 5% for a statistically significant result)

Analogy with diagnostic testing

In a hypothesis test

Power of the test is likelihood of detecting a true effect

- Sensitivity

Significance level of a test is the chance of test being positive when effect is null

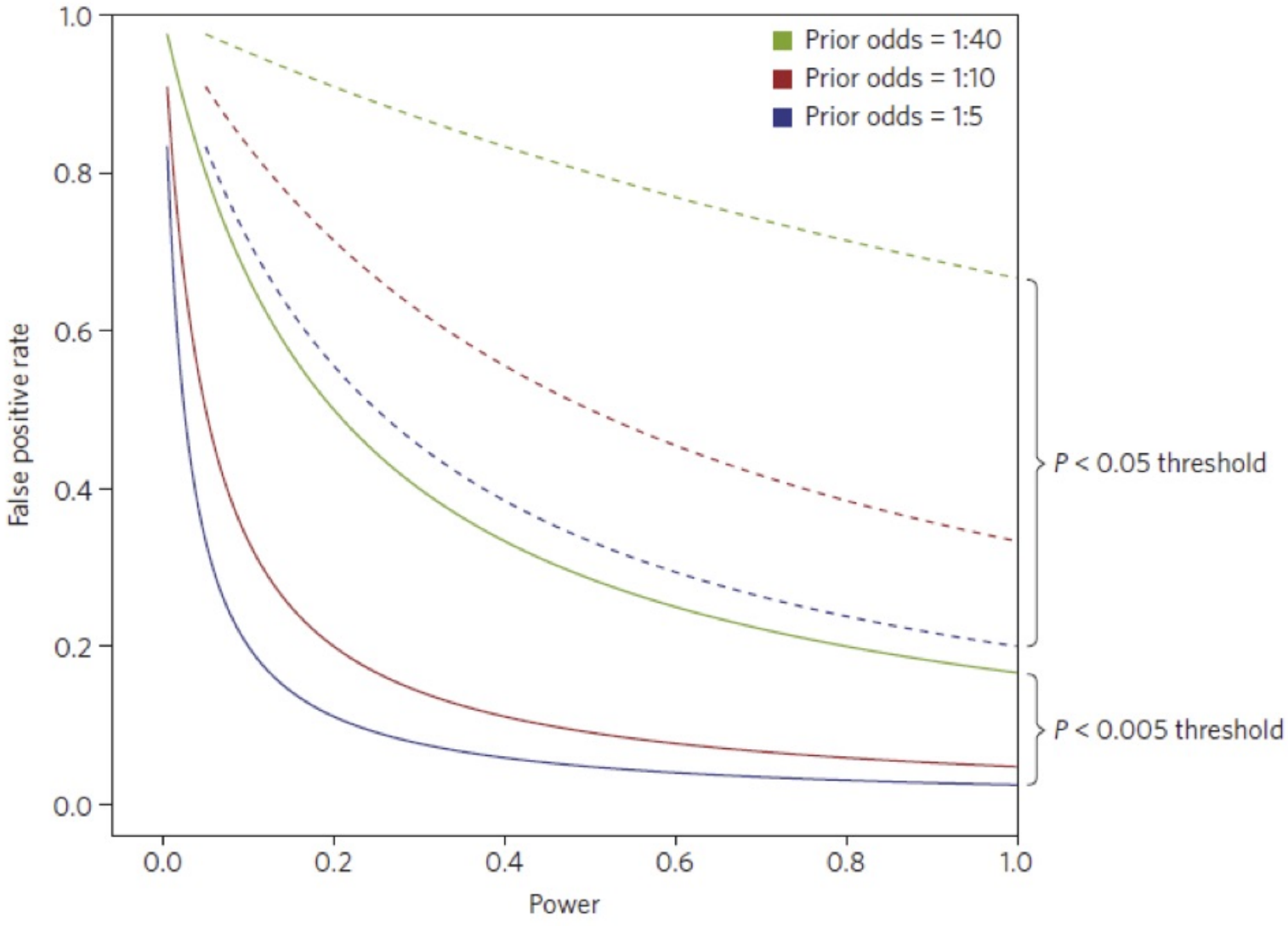
- It is the false positive rate, or 1 - Specificity

Prevalence affects false positive rate

Next slide shows a graph of probability of a false positive conclusion in favor after a statistically significant result, based on

- Power of the study
- Significance level threshold
- Prior odds of a treatment effect

False positive probability, p-value, power (Nat. Human Behavior, Jan 2018)



What if the actual threshold for a 'significant' result is larger than 0.05?

Alpha	Prior Odds	Prob Null	False Pos. Prob
0.1	1:5	0.83	0.38
0.1	1:10	0.91	0.56
0.1	1:40	0.98	0.83
0.25	1:5	0.83	0.61
0.25	1:10	0.91	0.76
0.25	1:40	0.98	0.93

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

Calculations assume power = 0.80

The effect of multiplicity

The more tests in a set of comparisons, the more likely it is that at least one will be a false positive.

Suppose each test is done at level $\alpha = 0.05$, and the endpoints are independent.

Number of Comparisons	Overall Type I Error Prob.
1	0.05
2	0.10
3	0.14
5	0.23
10	0.40
20	0.64

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

Not just a theoretical issue

Between 2000 and 2010:

- 1425 RCTs published in NEJM
- Among these trials, 222/1425 negative primary outcomes
- Among trials with negative primary outcomes, 121/222 "positive" secondary outcomes or subgroup by treatment interactions
  - 73 with a positive subgroup
  - 36 positive secondary outcome
  - 12 with 'nearly positive' outcome
- Of 121 with a presumed signal
  - 21 were replicated and showed positive primary outcome

Our recommendations in the current guidelines

If a statistically sound method for multiple tests was specified in the protocol or SAP, please follow it explicitly.

If there was no such plan

- Acknowledge the lack of a plan.
- Report secondary outcomes using only point estimates of treatment effect and 95% confidence intervals.
- Specifically state that the confidence intervals have not been adjusted for multiplicity and cannot be used to support claims about treatment effects.

Implications of the recommendations

Makes available all data on primary and secondary outcomes

- Without conclusions not likely to be reproducible

Confidence intervals are more informative than p-values

Makes demands of our readers

Requires careful editing of manuscripts

Policy applies to observational studies as well as RCTs

A purely statistical perspective oversimplifies the treatment of negative trials.

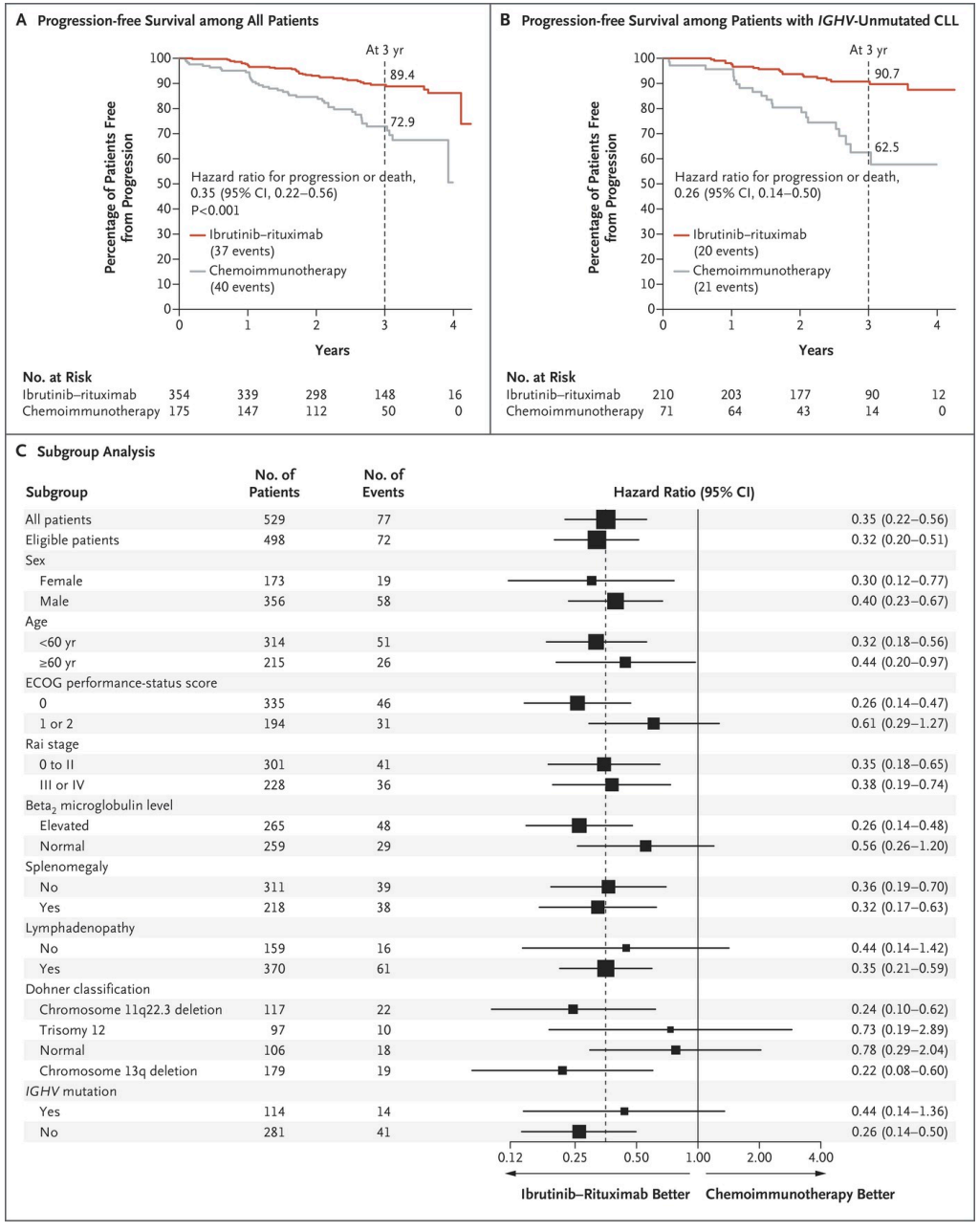
- See Pocock and Stone, NEJM 2016: "The primary outcome fails . . ."

Shanafelt, et al. NEJM 2019

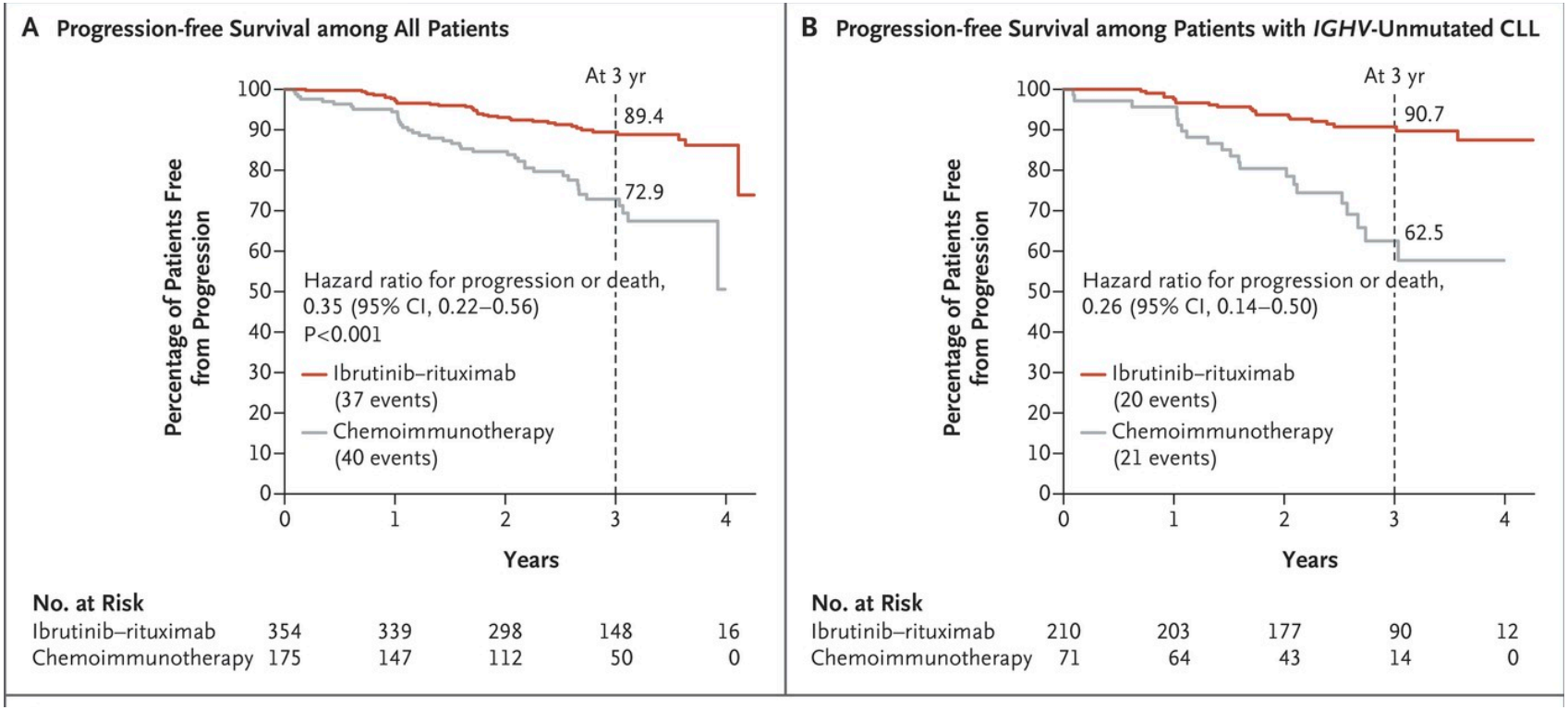
Patients 70 years of age or younger with previously untreated CLL were randomly assigned to receive ibrutinib plus rituximab or chemoimmunotherapy with fludarabine, cyclophosphamide, and rituximab.

The ibrutinib-based regimen led to prolonged progression-free and overall survival.

Figure 2 from article on next slide


A Progression-free Survival among All Patients
Percentage of Patients Free from Progression
At 3 yr
89.4
72.9
Hazard ratio for progression or death, 0.35 (95% CI, 0.22–0.56)
P<0.001
Ibrutinib–rituximab (37 events)
Chemoimmunotherapy (40 events)
Years
No. at Risk
Ibrutinib–rituximab 354 339 298 148 16
Chemoimmunotherapy 175 147 112 50 0
B Progression-free Survival among Patients with IGHV-Unmutated CLL
Percentage of Patients Free from Progression
At 3 yr
90.7
62.5
Hazard ratio for progression or death, 0.26 (95% CI, 0.14–0.50)
Ibrutinib–rituximab (20 events)
Chemoimmunotherapy (21 events)
Years
No. at Risk
Ibrutinib–rituximab 210 203 177 90 12
Chemoimmunotherapy 71 64 43 14 0
C Subgroup Analysis
Subgroup
No. of Patients
No. of Events
Hazard Ratio (95% CI)
All patients 529 77 0.35 (0.22–0.56)
Eligible patients 498 72 0.32 (0.20–0.51)
Sex
Female 173 19 0.30 (0.12–0.77)
Male 356 58 0.40 (0.23–0.67)
Age
<60 yr 314 51 0.32 (0.18–0.56)
≥60 yr 215 26 0.44 (0.20–0.97)
ECOG performance-status score
0 335 46 0.26 (0.14–0.47)
1 or 2 194 31 0.61 (0.29–1.27)
Rai stage
0 to II 301 41 0.35 (0.18–0.65)
III or IV 228 36 0.38 (0.19–0.74)
Beta2 microglobulin level
Elevated 265 48 0.26 (0.14–0.48)
Normal 259 29 0.56 (0.26–1.20)
Splenomegaly
No 311 39 0.36 (0.19–0.70)
Yes 218 38 0.32 (0.17–0.63)
Lymphadenopathy
No 159 16 0.44 (0.14–1.42)
Yes 370 61 0.35 (0.21–0.59)
Dohner classification
Chromosome 11q22.3 deletion 117 22 0.24 (0.10–0.62)
Trisomy 12 97 10 0.73 (0.19–2.89)
Normal 106 18 0.78 (0.29–2.04)
Chromosome 13q deletion 179 19 0.22 (0.08–0.60)
IGHV mutation
Yes 114 14 0.44 (0.14–1.36)
No 281 41 0.26 (0.14–0.50)
0.12 0.25 0.50 1.00 2.00 4.00
Ibrutinib–Rituximab Better
Chemoimmunotherapy Better

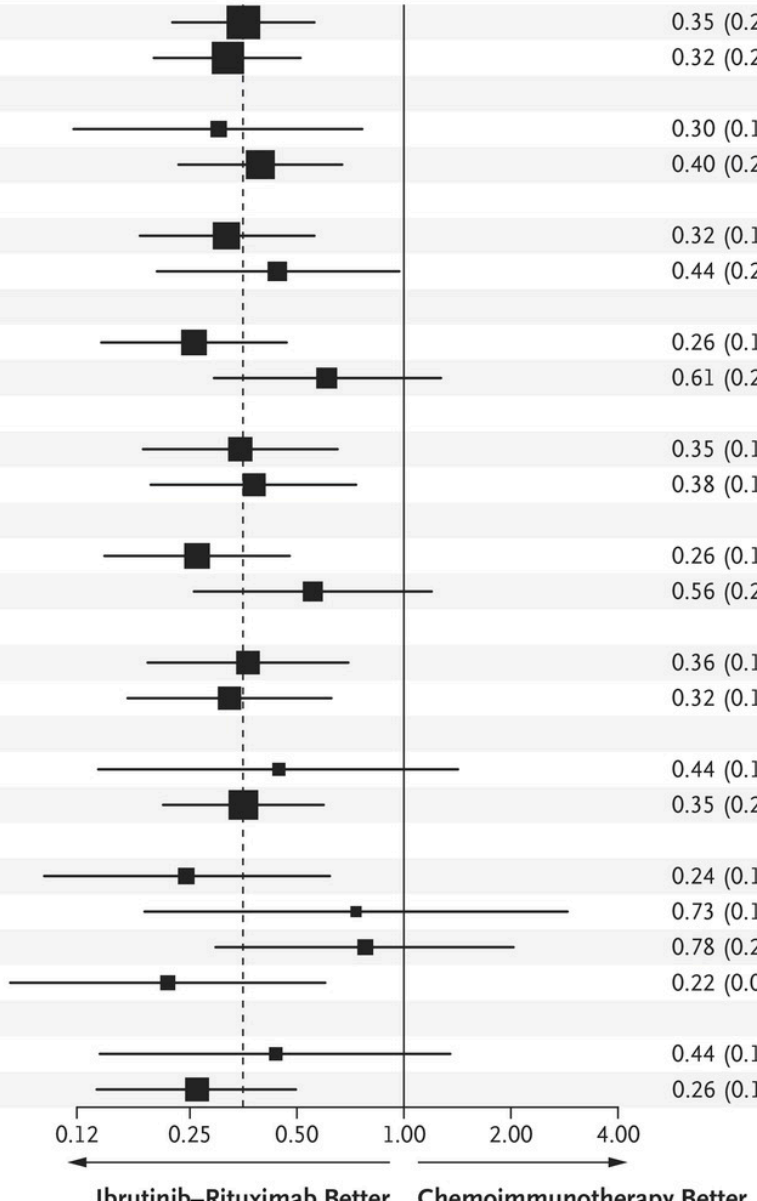
The PFS curves



Subgroups

C Subgroup Analysis

Subgroup	No. of Patients	No. of Events	Hazard Ratio (95% CI)
All patients	529	77	0.35 (0.22–0.56)
Eligible patients	498	72	0.32 (0.20–0.51)
Sex			
Female	173	19	0.30 (0.12–0.77)
Male	356	58	0.40 (0.23–0.67)
Age			
<60 yr	314	51	0.32 (0.18–0.56)
≥60 yr	215	26	0.44 (0.20–0.97)
ECOG performance-status score			
0	335	46	0.26 (0.14–0.47)
1 or 2	194	31	0.61 (0.29–1.27)
Rai stage			
0 to II	301	41	0.35 (0.18–0.65)
III or IV	228	36	0.38 (0.19–0.74)
$Beta_2$ microglobulin level			
Elevated	265	48	0.26 (0.14–0.48)
Normal	259	29	0.56 (0.26–1.20)
Splenomegaly			
No	311	39	0.36 (0.19–0.70)
Yes	218	38	0.32 (0.17–0.63)
Lymphadenopathy			
No	159	16	0.44 (0.14–1.42)
Yes	370	61	0.35 (0.21–0.59)
Dohner classification			
Chromosome 11q22.3 deletion	117	22	0.24 (0.10–0.62)
Trisomy 12	97	10	0.73 (0.19–2.89)
Normal	106	18	0.78 (0.29–2.04)
Chromosome 13q deletion	179	19	0.22 (0.08–0.60)
IGHV mutation			
Yes	114	14	0.44 (0.14–1.36)
No	281	41	0.26 (0.14–0.50)



The wider debate about p-values

p-values

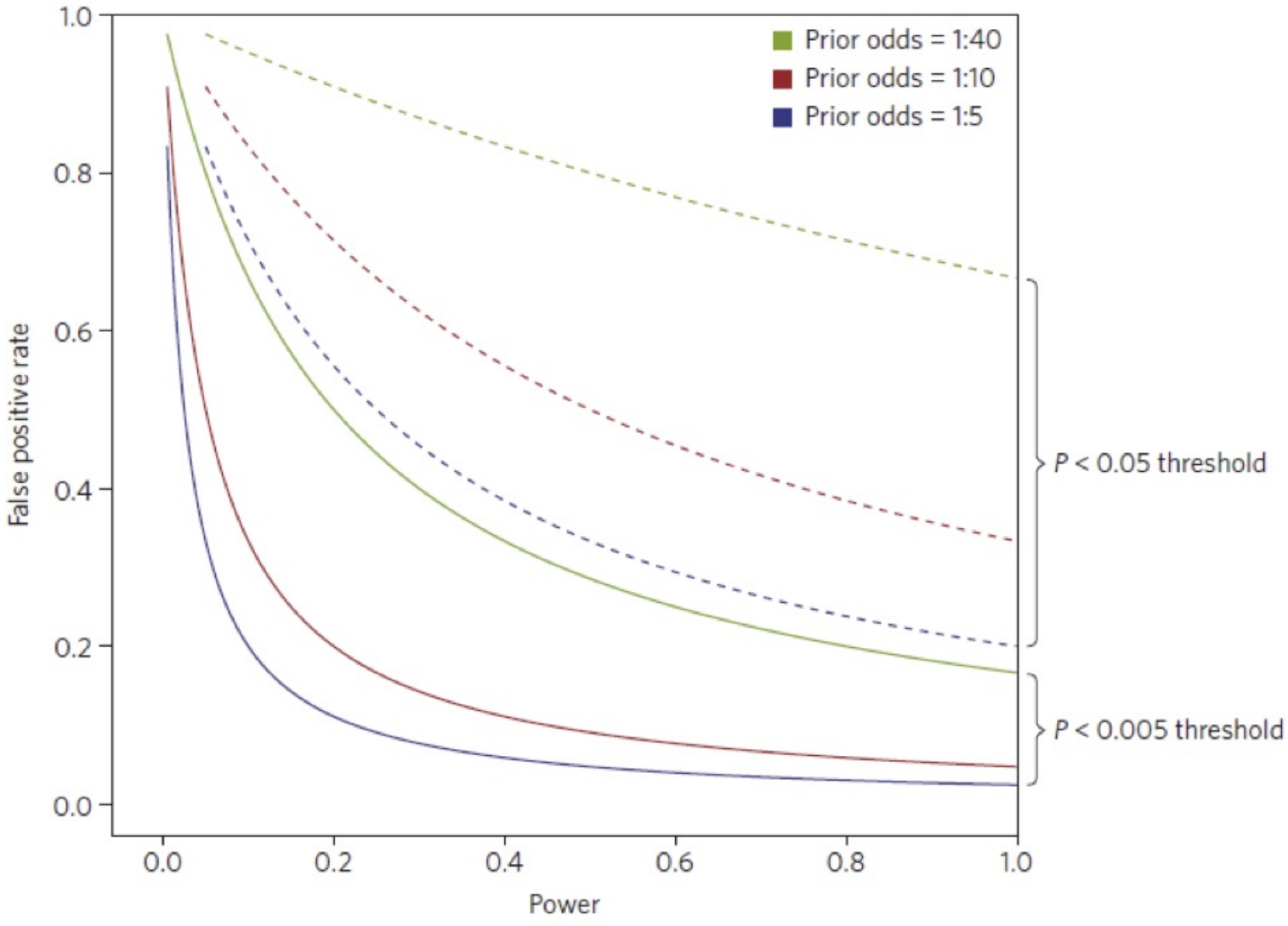
- do not indicate the size of an effect
- nor do they indicate the likelihood of an effect

Fisher never intended 'significant' to mean 'statistically significant'

- Is 'statistically significant' a meaningless term?

A true $p < 0.05$ may not reduce the false positive rate enough

False positive probability, p-value, power (Nat. Human Behavior, Jan 2018)



The wider debate...

From our editorial

> *The notion that a treatment is effective for a particular outcome if $P<0.05$ and ineffective if that threshold is not reached is a reductionist view of medicine that does not always reflect reality.*

But we need clearly articulated decision rules in evidence-based medicine

- Reproducibility and Replicability in Science (2019) http://nap.edu/25303

Also in the guidelines

Principled analyses of studies with missing data

Acceptable to use unadjusted p-values for safety outcomes

Protocol and Statistical Analysis Plan (SAP) required for clinical trials

Submit SAP for observational studies if it exists

Require model diagnostics for observational studies

Accompanying editorial gives references to methods for controlling error rates with multiple tests

Talk is available at under mskcc_2019

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

Some references

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

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

Dmitrienko A, D'Agostino RB Sr. Multiplicity considerations in clinical trials. N Engl J Med 2018;378:2115-22.

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