Balancing events, not patients, maximizes power in randomized survival studies

Godwin Yung

LiDS
June 1, 2023
Raleigh, NC

Acknowledgements

- Ray Lin, Genentech
- Yi Liu, Nektar Therapeutics
- Kaspar Rufibach, Roche
- Marcel Wolbers, Roche

Pivotal survival trials

Common practice:

1. randomize patients 1:1
2. follow patients until d events have been observed

Pivotal survival trials

Common practice:

1. randomize patients 1:1
2. follow patients until d events have been observed

Why?

Why event-driven studies?

Why event-driven studies?

Power is driven by number of events.

Why event-driven studies?

Power is driven by number of events.

Schoenfeld's (1981) large sample approximation for the log-rank statistic:

$$Z \sim N(\theta\sqrt{p_0 p_1 d}, 1)$$

where

- θ: log hazard ratio (HR)
- p_j: proportion of patients randomized to arm j ($0 =$ control)
- d: number of events at final analysis

Why randomize patients 1:1?

Why randomize patients 1:1?

- ethics

Why randomize patients 1:1?

- ethics
- habit

Why randomize patients 1:1?

- ethics
- habit
- statistical power
  - "1:1 maximizes power for continuous endpoints"
  - "1:1 maximizes power for survival endpoints" (Schoenfeld, 1981)

How reliable is Schoenfeld's equation?

How reliable is Schoenfeld's equation?

- "Reliable for typical hazard ratios"
  - HR > 0.4 (Barthel et al., 2006)

How reliable is Schoenfeld's equation?

- "Reliable for typical hazard ratios"
  - HR > 0.4 (Barthel et al., 2006)
- "Reliable for typical hazard ratios under 1:1. Not so much under unequal randomization." (Yung et al.)

Alternative approximations for the log-rank statistic

	Randomization ratio that maximizes power (assuming fixed event size)
Schoenfeld (1981)	1:1
Freedman (1982)	$1/HR$
Rubinstein (1981)	that which results in balance of events (i.e., 1:1 event size ratio)

Research questions

I. What randomization ratio maximizes power?

II. When might unequal randomization be attractive, considering power, accrual duration, trial duration, and sample size?

I. What randomization ratio maximizes power?

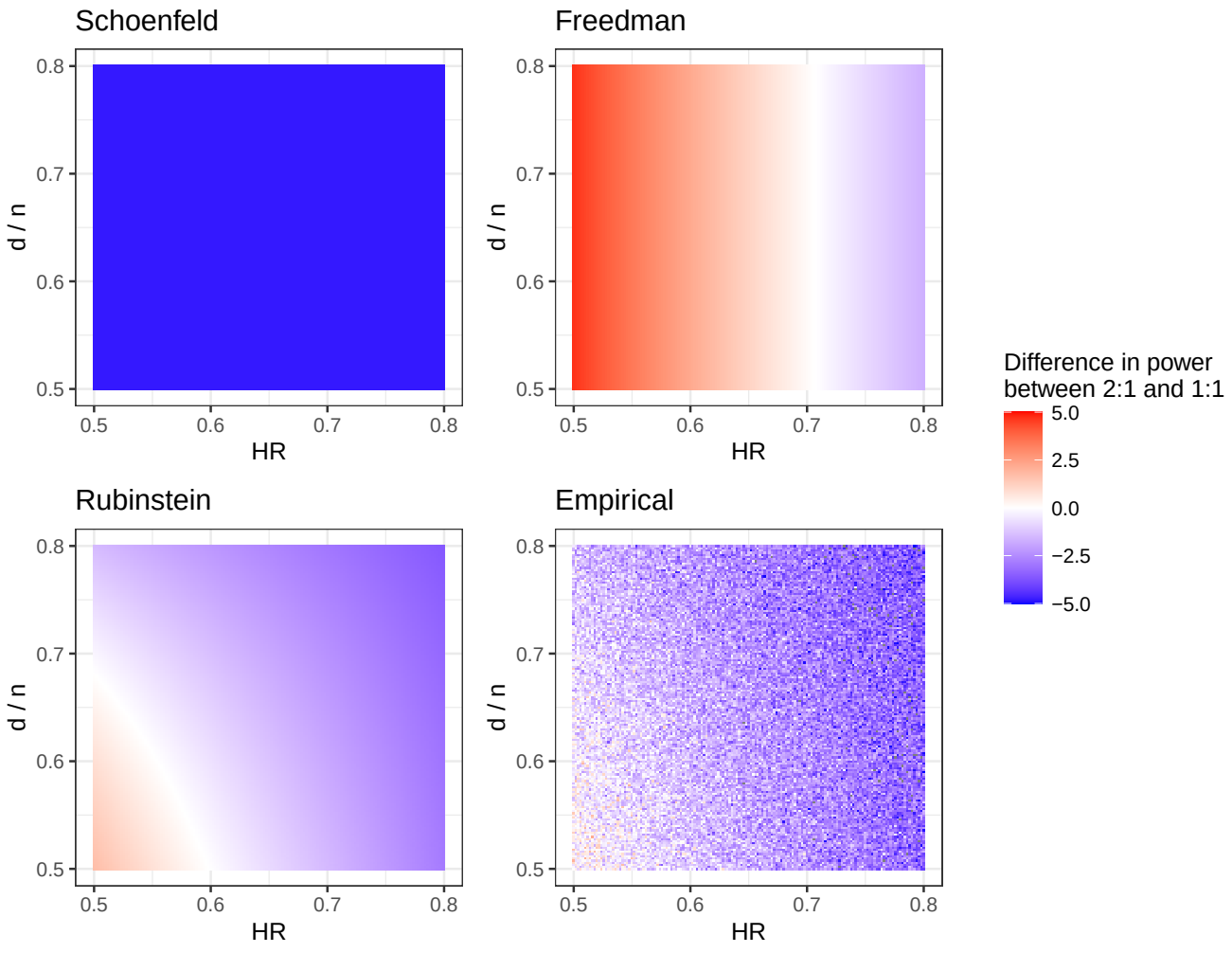
Given **hazard ratio** HR, **event-patient ratio** d/n, and **control median** CM=1y, we set ...

- **number of events** d based on Schoenfeld's equation with $p_0 = 0.5$, one-sided alpha 0.025, and 80% power
- **sample size** as $n = d * (d/n)^{-1}$
- **accrual rate** between 20-50 patients/month to ensure that trial duration is realistic

We then varied randomization ratio (**1:1, 3:2, 2:1**) and compared their power under various calculations (**Schoenfeld, Freedman, Rubinstein, empirical**).

Comparisons were made across the grid $(HR, d/n) \in [0.5, 0.8]^2$.

$$\text{Diff. in power} = \text{Power}(2{:}1, \text{Method}; \text{HR}, d/n) - \text{Power}(1{:}1, \text{Method}; \text{HR}, d/n)$$


Schoenfeld
Freedman
Rubinstein
Empirical
d / n
HR
0.5
0.6
0.7
0.8
Difference in power
between 2:1 and 1:1
5.0
2.5
0.0
−2.5
−5.0

II. When might unequal randomization be attractive?

Case-study with 6 alternative designs ("edge-cases").

II. When might unequal randomization be attractive?

Case-study with 6 alternative designs ("edge-cases").

Checkmate-017, a randomized open-label Ph3 study comparing nivolumab vs. docetaxel in patients with NSCLC.

II. When might unequal randomization be attractive?

Case-study with 6 alternative designs ("edge-cases").

Checkmate-017, a randomized open-label Ph3 study comparing nivolumab vs. docetaxel in patients with NSCLC.

Protocol:

- 264 patients, randomized 1:1
- 189 OS events
- Median survival 11.4m vs. 7.0m (HR=0.61)
- two-sided $\alpha = 0.04$, 90% power
- 12m accrual duration, 24m trial duration

Motivating questions for alternative study designs:

Motivating questions for alternative study designs:

- How does unequal randomization impact **trial duration**, assuming accrual rate does not change?

Motivating questions for alternative study designs:

- How does unequal randomization impact **trial duration**, assuming accrual rate does not change?
- How much faster does **accrual rate** need to be in order to avoid any delay in trial readout?

Motivating questions for alternative study designs:

- How does unequal randomization impact **trial duration**, assuming accrual rate does not change?
- How much faster does **accrual rate** need to be in order to avoid any delay in trial readout?
- How many more **patients** need to be recruited in order to avoid any delay in trial readout (and assuming accrual rate does not change)?

Motivating questions for alternative study designs:

- How does unequal randomization impact **trial duration**, assuming accrual rate does not change?
- How much faster does **accrual rate** need to be in order to avoid any delay in trial readout?
- How many more **patients** need to be recruited in order to avoid any delay in trial readout (and assuming accrual rate does not change)?

Given 3:2 or 2:1 randomization, 6 alternatives total.

Design	RR, Description	d	n	Accrual Dur. (months)	Trial Dur. (months)
CM-017	1:1, Protocol	189	264	12.0	23.8
Alt 1	3:2, Prolonged study	190	(264)	(12.0)	24.9
Alt 2	3:2, Accelerated accrual	190	(264)	10.0	(23.8)
Alt 3	3:2, Increased enrollment	190	274	12.5	(23.7)
Alt 4	2:1, Prolonged study	198	(264)	(12.0)	27.6
Alt 5	2:1, Accelerated accrual	198	(264)	5.1	(23.8)
Alt 6	2:1, Increased enrollment	198	294	13.4	(23.8)

Design	RR, Description	d	n	Accrual Dur. (months)	Trial Dur. (months)
CM-017	1:1, Protocol	189	264	12.0	23.8
Alt 1	3:2, Prolonged study	190	(264)	(12.0)	24.9
Alt 2	3:2, Accelerated accrual	190	(264)	10.0	(23.8)
Alt 3	3:2, Increased enrollment	190	274	12.5	(23.7)
Alt 4	2:1, Prolonged study	198	(264)	(12.0)	27.6
Alt 5	2:1, Accelerated accrual	198	(264)	5.1	(23.8)
Alt 6	2:1, Increased enrollment	198	294	13.4	(23.8)

Similar results were observed across a general set of scenarios:

- 3:2 minimal impact, easier to mitigate
- 2:1 greater impact, harder to mitigate

Conclusion

Conclusion

- Rubinstein's equation allows us to quickly and accurately calculate design parameters (power, trial duration, accrual rate, etc.).

Conclusion

- Rubinstein's equation allows us to quickly and accurately calculate design parameters (power, trial duration, accrual rate, etc.).
- Fixing event size, power is maximized under 1:1 event size ratio (which depends on randomization, hazard, and event-patient ratio).

Conclusion

- Rubinstein's equation allows us to quickly and accurately calculate design parameters (power, trial duration, accrual rate, etc.).
- Fixing event size, power is maximized under 1:1 event size ratio (which depends on randomization, hazard, and event-patient ratio).
- 3:2 is a sensible option in most cases. Some additional patients can be randomized to the active arm with minimal impact on event size and trial duration.

Conclusion

- Rubinstein's equation allows us to quickly and accurately calculate design parameters (power, trial duration, accrual rate, etc.).
- Fixing event size, power is maximized under 1:1 event size ratio (which depends on randomization, hazard, and event-patient ratio).
- 3:2 is a sensible option in most cases. Some additional patients can be randomized to the active arm with minimal impact on event size and trial duration.
- 2:1 may also be considered if event-patient ratio is small ($d/n \leq 0.5$). If event-patient ratio is large, then unequal randomization may be hard to justify given the increase in event size and trial duration.

References

- Barthel, F. et al. (2006) Evaluation of sample size and power for multi-arm survival trails allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. *Stat Med* 25:-2521-2542.
- Freedman, L. (1982) Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1:121-129.
- Hsieh, F. (1992) Comparing sample size formulae for trials with unbalanced allocation using the logrank test. *Stat Med* 11:1091-1098.
- Rubinstein, L., Gail, M., and Santner, T. (1981) Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J Chron Dis* 34:469-479.
- Schoenfeld, D. (1981) The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 68:316-319.
- Yung, G. and Liu, Y. (2019) Sample size and power for the weighted log-rank test and Kaplan-Meier based tests with allowance for nonproportional hazards. *Biometrics* 76:939-950.