Likelihood Study Design Proposal
"Just do it."

Jeffrey D. Blume, PhD
Department of Biostatistics
Vanderbilt University
SCT 2011 Vancouver

The Law of Likelihood says

- Strength of evidence for H_A over H_B is measured by the likelihood ratio: $LR = P_A(x)/P_B(x)$

- "H_A is supported over H_B by a factor of LR."
  - If LR=1, the evidence is neutral
  - If LR>1, the evidence supports H_A over H_B
  - If LR<1, the evidence supports H_B over H_A

- Weak evidence
  - for H_A over H_B: $1<LR<8$
  - for H_B over H_A: $1/8<LR<1$
- Moderate evidence
  - for H_A over H_B: $8<LR<32$
  - for H_B over H_A: $1/32<LR<1/8$
- Strong evidence
  - for H_A over H_B: $32<LR$
  - for H_B over H_A: $LR<1/32$

- *p=0.05 maps to LR=6.8 (with one look)*
- *p-values do not distinguish between weak evidence and evidence in favor of the null*

Analysis model

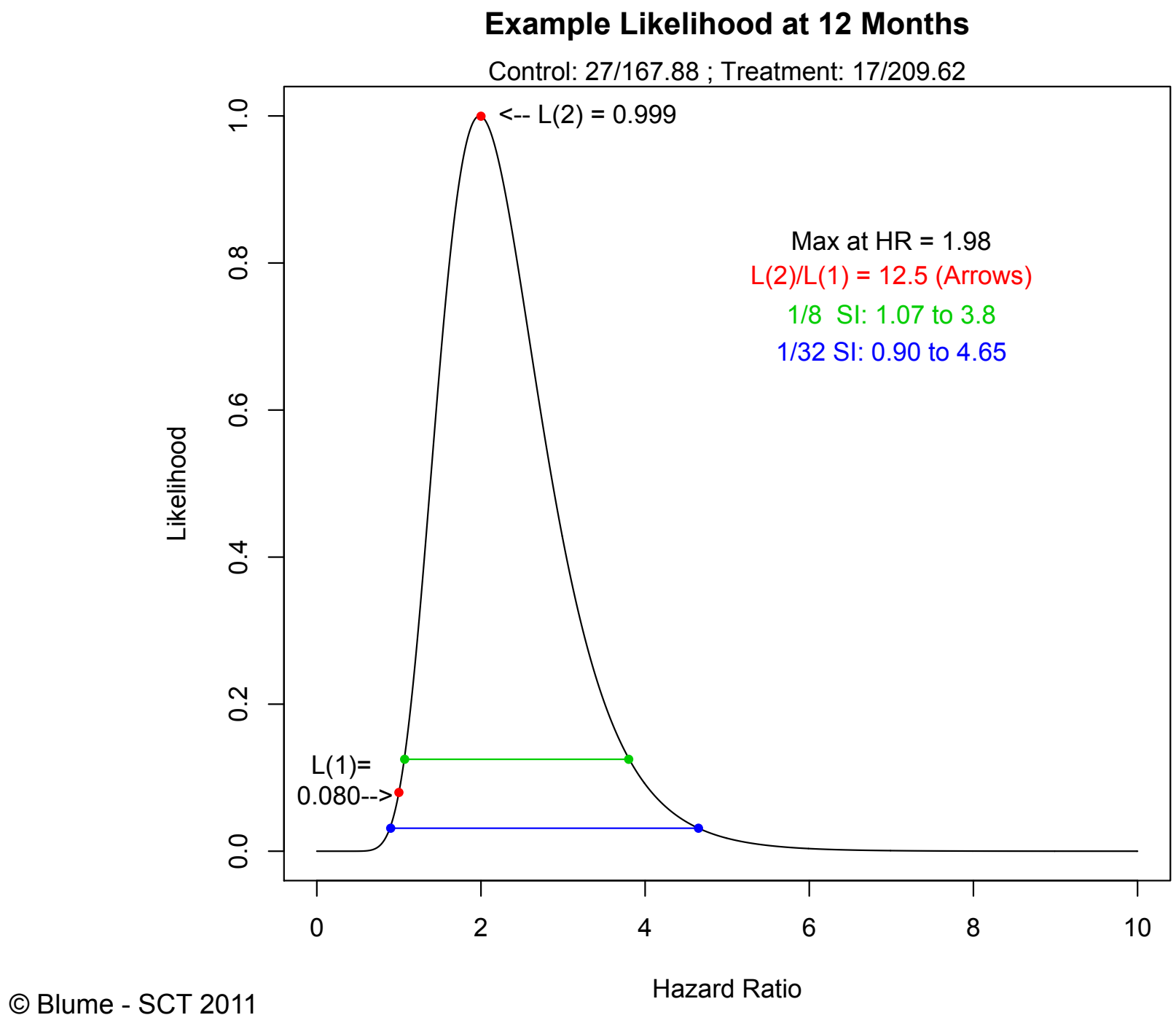
- Control: events (e_c), exposure (t_c), rate (λ_c)
- Treatment: events (e_t), exposure (t_t), rate (λ_t)

$$\left.\begin{aligned} e_c &\sim Poiss(\lambda_c t_c) \\ e_t &\sim Poiss(\lambda_t t_t) \end{aligned}\right\} \Rightarrow \quad e_t \mid e_c + e_t \sim Binom(e_c + e_t, p)$$

$$\text{where} \quad p = \frac{h}{h+g} \qquad h = \frac{\lambda_t}{\lambda_c} \qquad g = \frac{t_c}{t_t}$$

- Mapping: $(h_i, g_i) \rightarrow p_i \;\; \text{for} \;\; i = 0,1$
  - Example (under null): $h_0=1$, $g_0=1$ so $p_0=1/2$

Note: data actually generated under exponential assumption



Properties of likelihood

- Fixes the scale of evidence (e.g., LR not affected by looks)
- 'Strength of evidence' & 'probability of bad result' not confused
- Minimizes average error rate $(\alpha+\beta)/2$
  (instead of minimizing β for given α)
- Type I & II rates, FDR_0 & FDR_1 all converge to 0
  (in non-sequential case)
- Robust options available for analysis
- Maximum flexibility for conducting and reporting analyses

Metrics for evidential analysis

1. How strong is the observed evidence for H_1 over H_2?
   (Observed likelihood ratio)
2. Is the study design reliable?
   (Type I & II error rates ; expected sample size)
3. What is the chance that the observed data are misleading?
   (Posterior probabilities)

When interpreting observed data, #2 is completely irrelevant

Likelihood's evidential metrics

Evidential Quantity	Name	What it measures	Mathematical Representation
1	likelihood ratio	strength of the evidence	LR
2	probability of observing misleading evidence	propensity for study to yield misleading evidence	$mis_0=P(\ LR > k \mid H_0)$ $mis_1=P(\ LR < 1/k \mid H_1)$
3	probability that observed evidence is misleading	propensity for observed results to be misleading	$P(H_0 \mid LR > k)$ $P(H_1 \mid LR < 1/k)$

Current Paradigms

Evidential Quantity	What it measures	Hypothesis/ Significance Testing	False Discovery Rates	Bayesian Inference
1	strength of the evidence	Tail-area probability (p-value)	Tail-area probability (p-value) (?!)	Bayes Factor or Posterior Probability (?)
2	propensity for study to yield misleading evidence	Tail-area probability (Type I & II errors)	Tail-area probability (Type I & II errors)	*Operational Characteristics*
3	propensity for observed results to be misleading	Misinterpret the tail-area probability	False Discovery Rate(s)	Posterior Probability (??)

Important side note

- Likelihood's 'error rates' are driven to zero as the sample size increase (non-sequential)
- Probabilities of misleading evidence
  - Bounded and well understood
  - Example: $mis_0 = P(\ LR > k \mid H_0) \leq 1/k$
  - Bound holds in sequential case (e.g. w/ repeated looks)
  - Probabilities are typically much lower than bound
- Evidence for all alternatives is always reported

Design Summary: '*Just do it.*'

- Enroll 8 participants per month
  - 1:1 randomization to E vs. E+TT (4 / arm / month)
- Examine likelihood ratio sequentially
  - After every event or after every month
  - Continue with weak evidence: $1/8 < L(h_1)/L(h_0) < 8$
  - Stop when evidence not weak: LR < 1/8 or LR > 8
- Repeat in three strata: All, M+ only, M- Only
  - Alternative varies by strata: $h_1 = 1.5, 2, 1.4$
  - One strata may be stopped while other continues

All (M+ & M-)

Randomization : 1:1

Accrual per mo.: 8 participants
(4 E & 4 E+TT)

Stopping criteria:

k_1 = 1/8 (favor null) &

k_2 = 8 (favor alt)

Hypothesis Assumptions:

H_0: $h = 1$

H_1: $h = 1.5$

Baseline rates:

E: PFS = 12 months
(rate= 0.058)

E+TT: PFS = 18 months
(rate= 0.039)

		Planned Looks	
		Event-ly	**Monthly**
Type I Error		0.10	0.02
Power		0.90	0.86
Study Length (months)	null	23	27
	alt	21	25
Total Events	null	87	92
	alt	88	92
Total Exposure (person/mo.)	null	2226	2345
	alt	1805	1909

Early Termination (PET) in Months

	Min	**25%**	**50%**	**75%**	**Max**	**@48**
Null	7	20	26	33	48	3%
Alt	5	17	23	31	48	7%

Stratum M+

Randomization : 1:1

Accrual per mo.: 2 participants
(1 E & 1 E+TT)

Stopping criteria:

k_1 = 1/8 (favor null) &

k_2 = 8 (favor alt)

Hypothesis Assumptions:

H_0: h = 1

H_1: h = 2

Baseline rates:

E: PFS = 6 months
(rate= 0.116)

E+TT: PFS = 12 months
(rate= 0.058)

		Planned Looks	
		Event-ly	**Monthly**
Type I Error		0.09	0.01
Power		0.91	0.85
Study Length (months)	null	23	29
	alt	19	27
Total Events	null	32	32
	alt	34	35
Total Exposure (person/mo.)	null	552	538
	alt	382	412

Early Termination (PET) in Months

	Min	**25%**	**50%**	**75%**	**Max**	**@48**
Null	13	21	27	35	48	3%
Alt	7	17	23	36	48	7%

Stratum M-

Randomization : 1:1

Accrual per mo.: 6 participants
(3 E & 3 E+TT)

Stopping criteria:

k_1 = 1/8 (favor null) &

k_2 = 8 (favor alt)

Hypothesis Assumptions:

H_0: h = 1

H_1: h = 1.4

Baseline rates:

E: PFS = 15 months
(rate= 0.046)

E+TT: PFS = 21 months
(rate= 0.033)

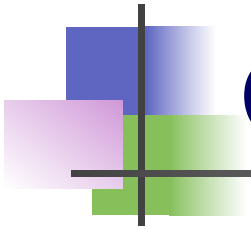
		Planned Looks	
		Event-ly	**Monthly**
Type I Error		0.10	0.01
Power		0.89	0.65*
Study Length (months)	null	35	40
	alt	32	36
Total Events	null	124	109
	alt	125	105
Total Exposure (person/mo.)	null	3754	3277
	alt	3160	2672

Early Termination (PET) in Months

	Min	**25%**	**50%**	**75%**	**Max**	**@48**
Null	9	32	40	48	48	34%
Alt	8	27	35	48	48	29% *

Recap of monthly monitoring

Strata	Evidence Favors…	Null	Alt	Stopping Month (50%) [25% to 75%]	
				Null	Alt
All	E+TT	0.02	0.86	26 [20 to 33]	23 [17 to 31]
	E	0.95	0.07		
	Neither	0.03	0.07		
M+	E+TT	0.01	0.85	27 [21 to 35]	23 [17 to 36]
	E	0.96	0.08		
	Neither	0.03	0.07		
M-	E+TT	0.01	0.65	40 [32 to 48]	35 [27 to 48]
	E	0.65	0.06		
	Neither	0.34	0.29		



Conclusions

- No formal test for marker by treatment interaction
- Accrue until desired evidence obtained or limit of resources is reached
- Design is very flexible
  - Look when you want
  - See what you see: “Pigs is pigs; Data is data” – *J Cornfield*
- Composite alternatives yield similar results
- Trials can end with weak evidence, but it is not misinterpreted as supporting the null hypothesis