Bayesian Adaptive Designs for Clinical Trials

Jason Connor
ConfluenceStat

Jason@ConfluenceStat.com
412-860-3113

Phase 1

Example 1: Early Phase Trials Identify Maximum Tolerated Dose (MTD)

With Kristine Broglio
& Anna McGlothin

Phase I Trials

- Key goal is to assess safety and identify the appropriate dose of a new treatment

- Typically very small sample sizes

- Most therapeutic areas:
  - healthy volunteers
  - single ascending dose (SAD) and multiple ascending dose (MAD) studies

  - Oncology (therapies often very toxic):
  - patients
  - dose escalation studies: no randomization
  - heterogeneous population (often "all-comer" trials)

Dose Escalation Studies

Typically small, uncontrolled studies.

GOAL: Determine the maximum tolerated dose (MTD), and/or a recommended Phase II dose.

Approaches:

- Algorithm-based designs

  3+3 (or the more general A+B)

  MTD is identified as the dose with fewer than some proportion of dose limiting toxicities (e.g. <2/6).

- Model-based designs

  MTD is estimated as a quantile of the dose-toxicity curve.

Key Design Elements

- Define starting dose and dose spacing
- Cohort size: how many patients are treated at a time
- Identify which adverse events will be defined as "dose-limiting toxicity" (DLT)
- Define acceptable/target toxicity level (TTL)
  - typically between 20% and 33%
  - corresponds to maximum tolerated dose (MTD)
- Define dose escalation rules

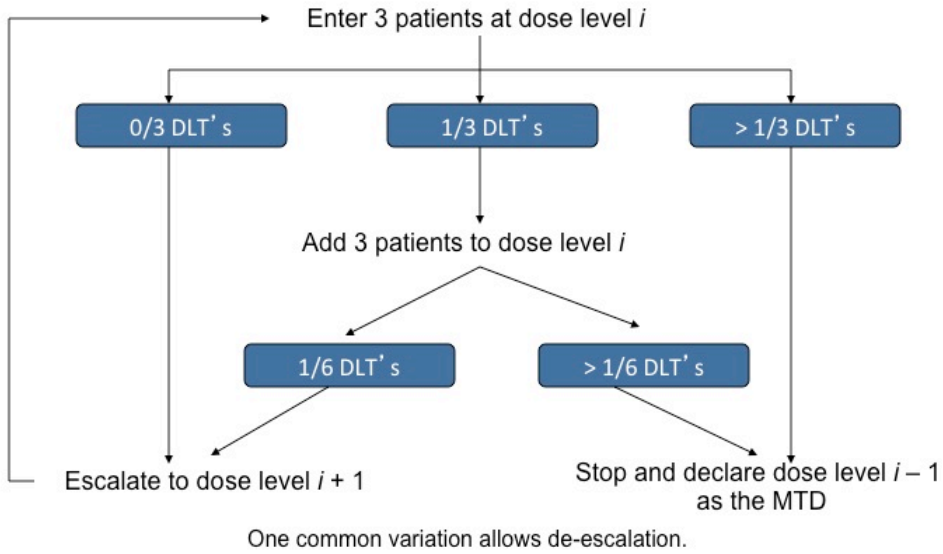
Two Strategies

1. Algorithm-based designs
   - 3+3 (or more generally A+B)
   - Accelerated titration
   - MTD is defined as the dose with fewer than some pre-specified observed proportion of dose limiting toxicities (e.g. < 2/6)

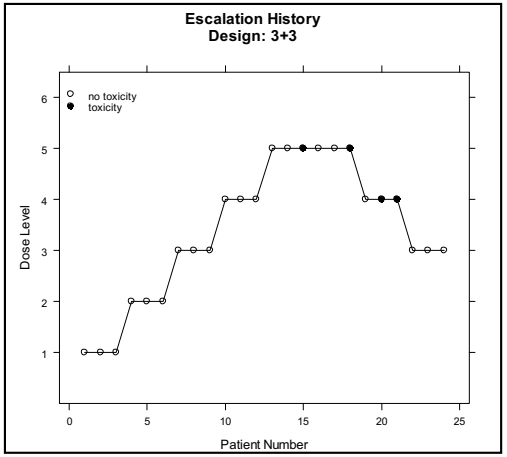
1. Model-based designs
   - MTD is estimated as a quantile of the dose-toxicity curve

"Traditional" 3+3 Design

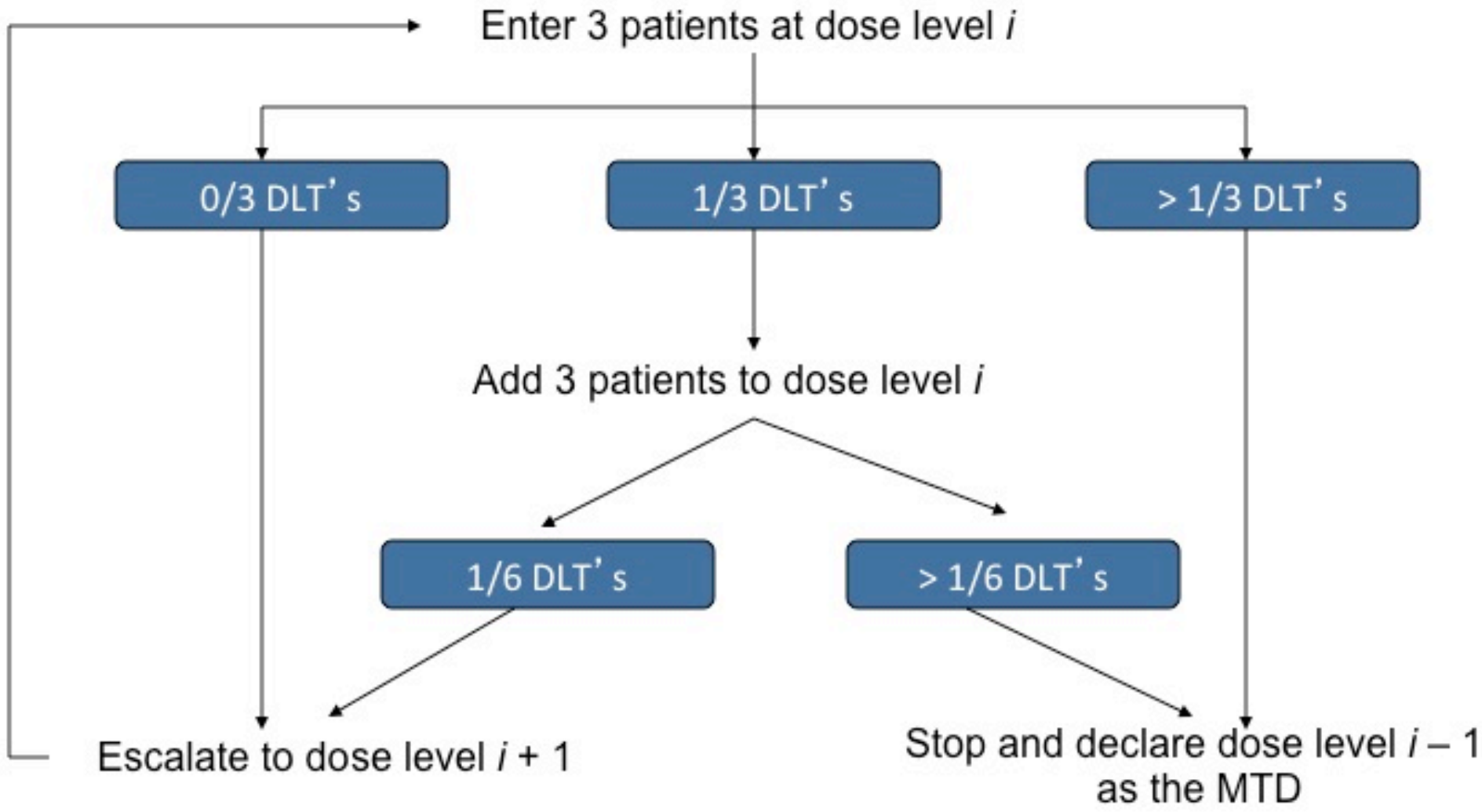
- Well accepted, easy to understand and implement
- Weaknesses are well documented:
  - The probability of stopping at an incorrect dose level is higher than generally believed (Reiner et al. 1999)
  - No statistical estimation of MTD
  - Underlying Pr(toxicity) associated with MTD is unclear
  - Behavior of the design depends on the number of cohorts before the MTD



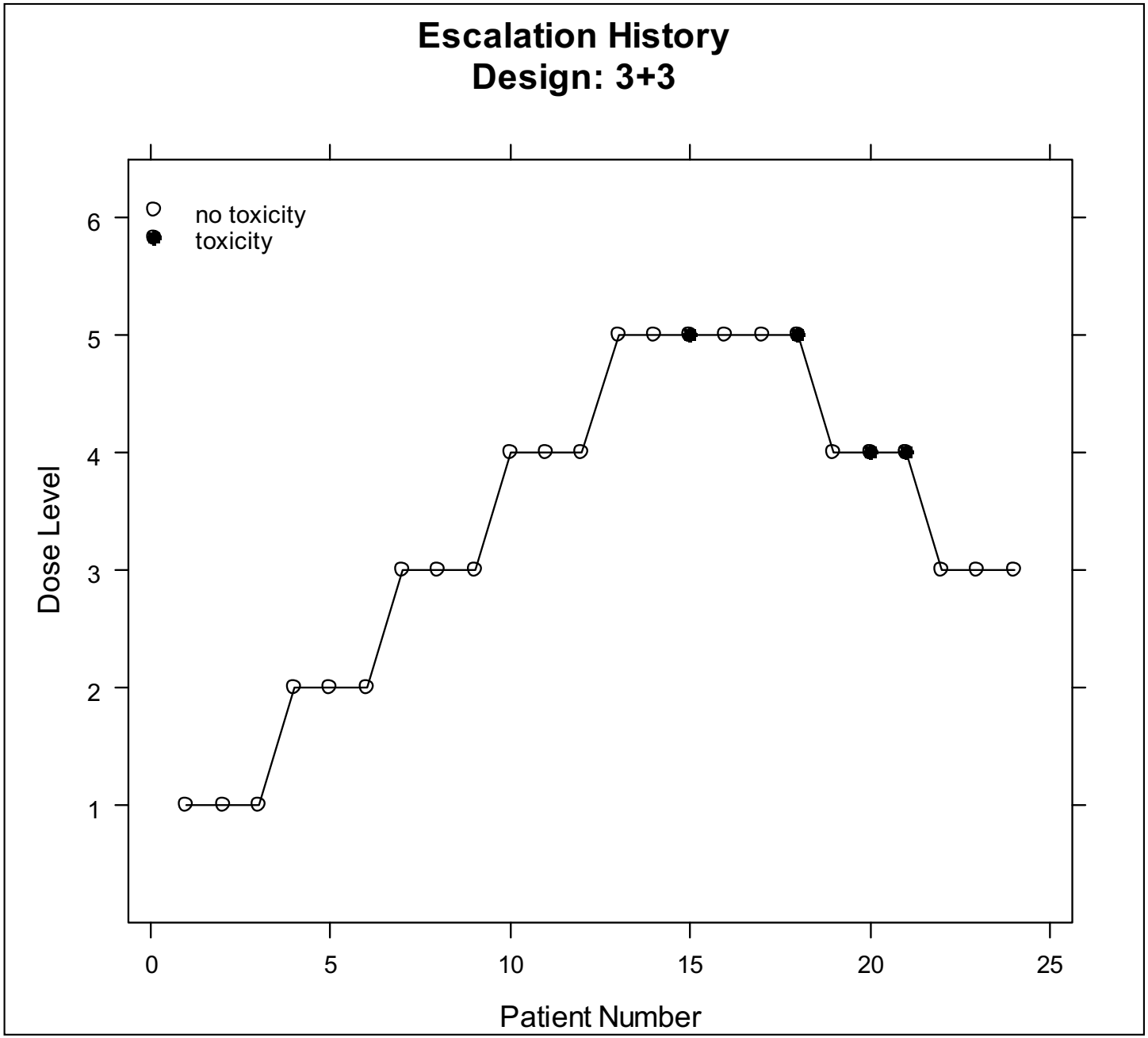
One common variation allows de-escalation.



One common variation allows de-escalation



One common variation allows de-escalation.


Escalation History
Design: 3+3
no toxicity
toxicity
Dose Level
1
2
3
4
5
6
0
5
10
15
20
25
Patient Number

Model-based Designs

Model-based designs use a statistical model to describe the relationship between dose and outcome:

Continual Reassessment Method (CRM)

- O'Quigley, Pepe, Fisher (1990)
- Faries (1994)
- Goodman, Zahurak, Piantadosi (1995)

Escalation with Overdose Control (EWOC)

- Babb, Rogatko, Zacks (1998)

Joint Toxicity/Efficacy

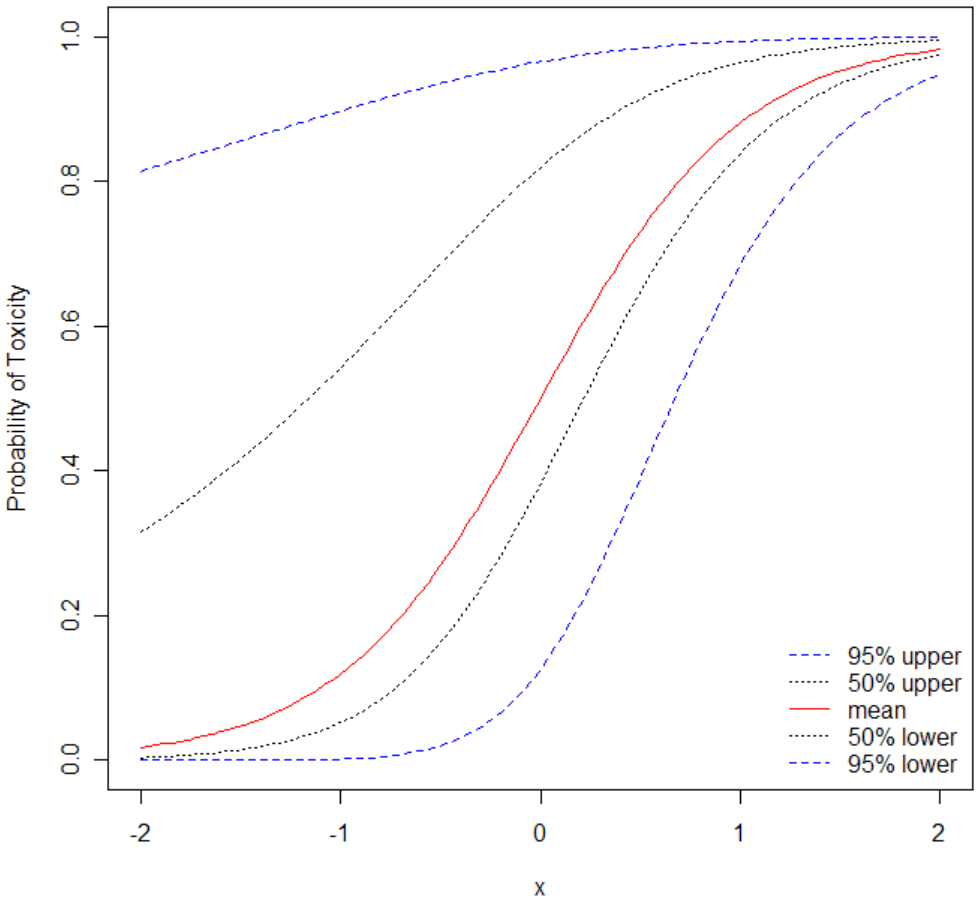
- Braun (2002)
- Thall and Cook (2004)

CRM with Cool Features

- Broglio, Berry (2015)
- Quintana, Li et al (2016)

Continual Reassessment Method

1. Select a mathematical model to describe the relationship between dose and Pr(DLT).
2. Describe uncertainty about the model by a prior distribution.
3. After each cohort, update the model, and estimate the probability of toxicity for each dose.
4. Assign patients to dose levels according to the model-based estimates of DLT and with regard to other rules that may govern escalation (e.g. no skipped doses)
5. Stop when a pre-specified rule is met (e.g. maximum number of subjects on the MTD)



Lots of customization opportunities!

Example Phase I CRM

- 8 doses: 1, 2, 3, 5, 8, 13, 21, 34 mg or generically: 1, 2, 3, 4, 5, 6, 7, 8
- DLT (dose-limiting toxicity) is defined within the first cycle
- Goal is to find the *MTD*: *maximum dose with rate of DLT's* $\leq$ *0.30*

Example Phase I CRM

- Bayesian model for DLT-dose relationship

$$\pi_d = \Pr\left(DLT \mid \text{dose d}\right)$$

- Model Log-odds

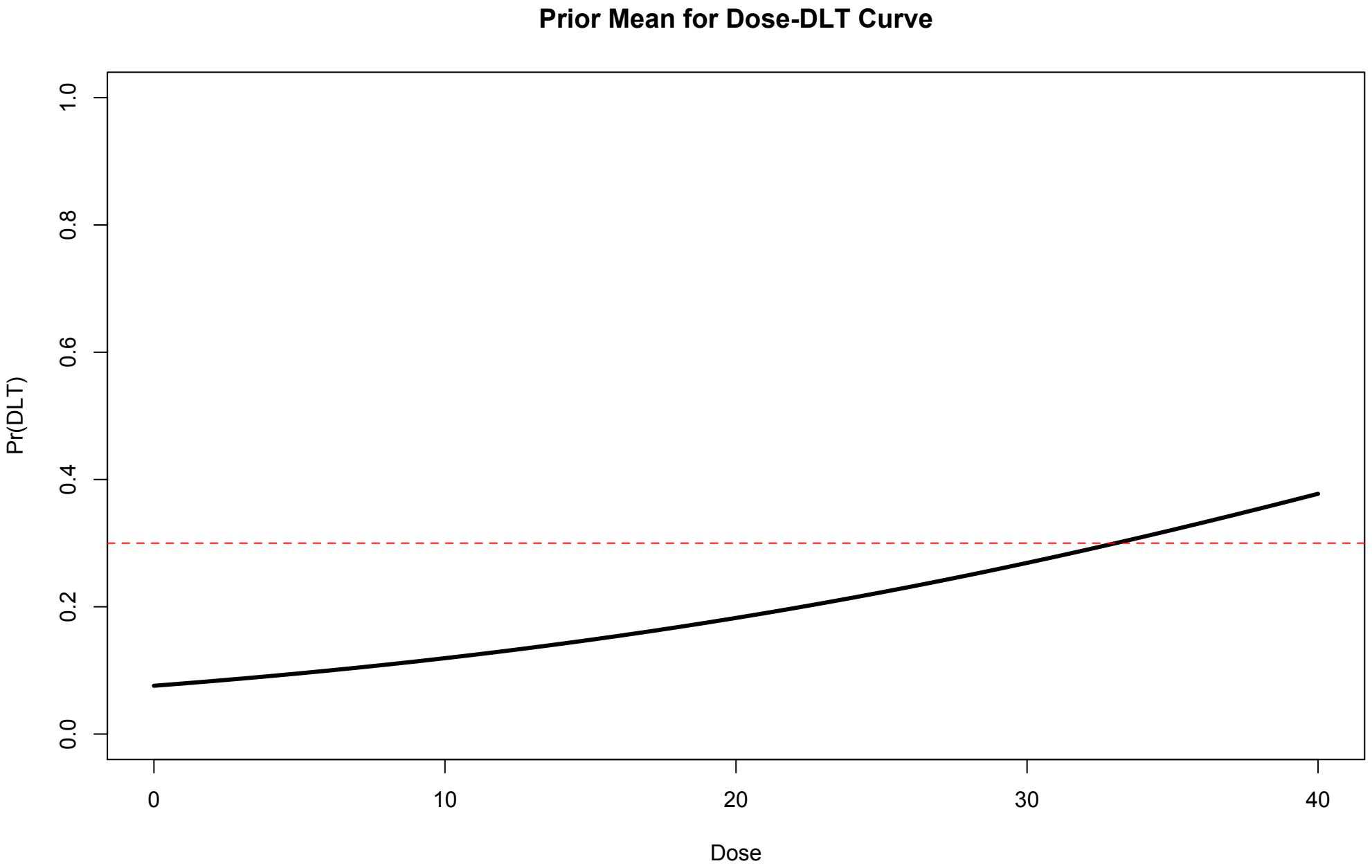
$$\theta_d = \log\left(\frac{\pi_d}{1-\pi_d}\right)$$

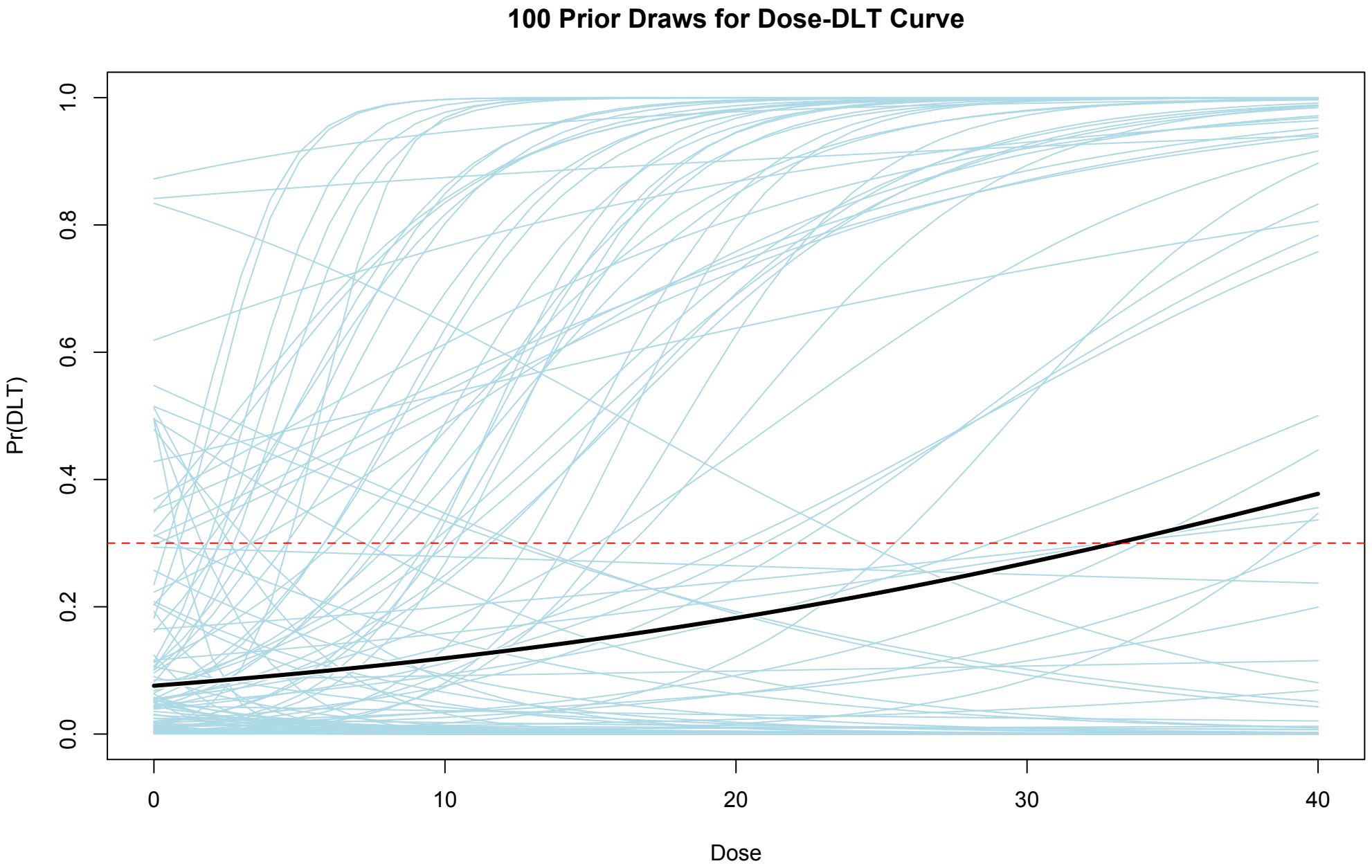
- Model:

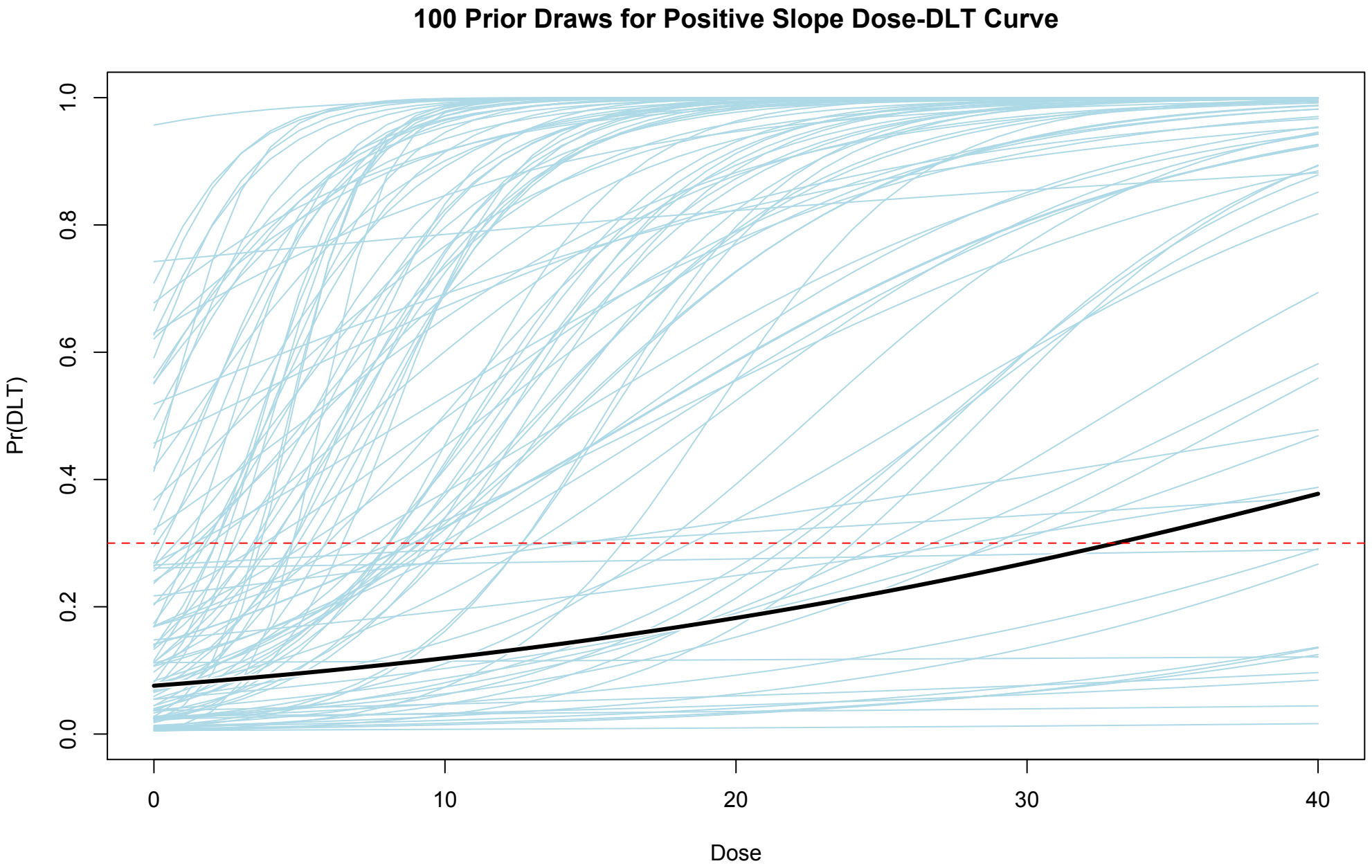
$$\theta_d = \alpha + \beta d$$

$$\alpha \sim N\left(-2.5, 2^2\right)$$

$$\beta \sim N\left(0.05, 0.35^2\right)$$


Prior Mean for Dose-DLT Curve
Pr(DLT)
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Dose


100 Prior Draws for Dose-DLT Curve
Pr(DLT)
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Dose


100 Prior Draws for Positive Slope Dose-DLT Curve
Pr(DLT)
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Dose

Code

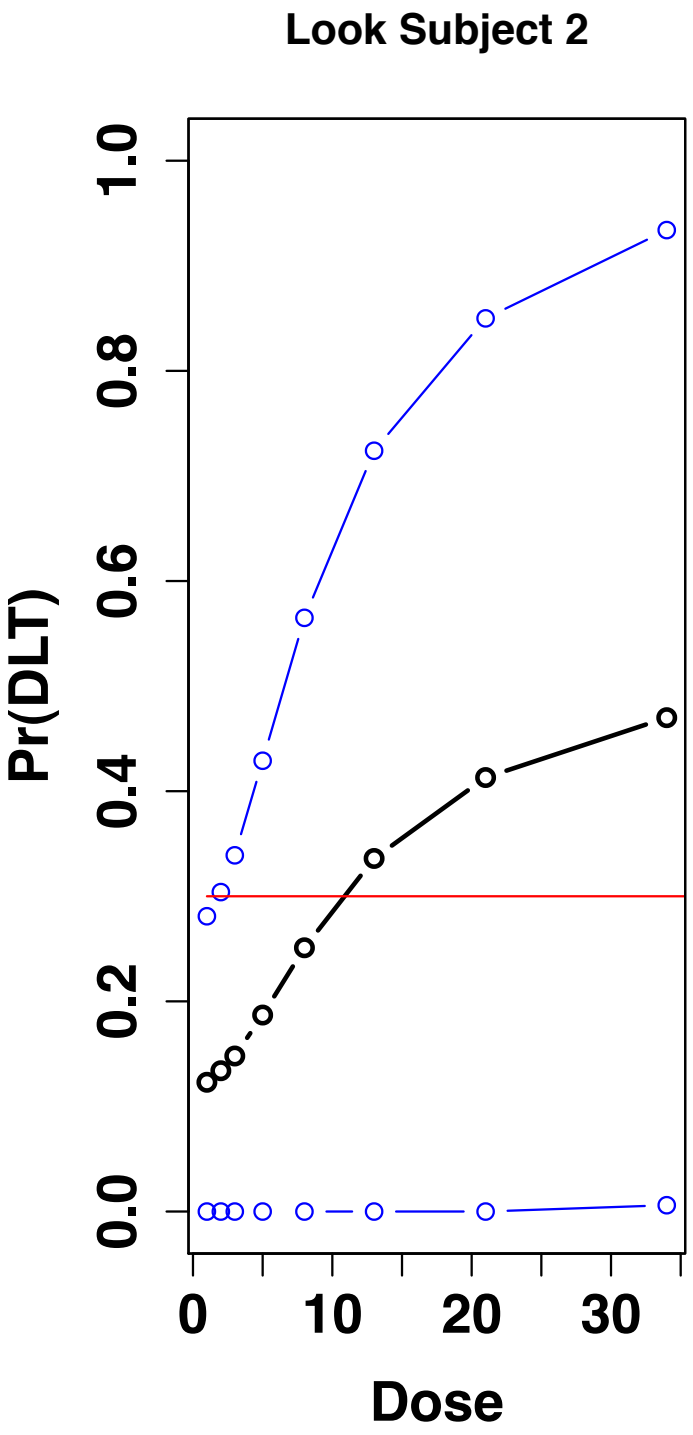
```
plot(0,0,xlim=c(0,40), ylim=c(0,1), xlab='Dose', ylab='Pr(DLT)', type='n',
main='Prior Mean for Dose-DLT Curve')
lines(d, inv.logit.fun(-2.5 + d*0.05), lwd=3)
abline(h=0.3, col=2, lty=2)


alpha <- rnorm(100, -2.5, 2)
beta <- rnorm(100, 0.05, 0.35)
plot(0,0,xlim=c(0,40), ylim=c(0,1), xlab='Dose', ylab='Pr(DLT)', type='n',
main='100 Prior Draws for Dose-DLT Curve' )
for(s in 1:100){
lines(d, inv.logit.fun(alpha[s]+beta[s]*d), col='lightblue')
}
lines(d, inv.logit.fun(-2.5 + d*0.05), lwd=5,  col='white')
lines(d, inv.logit.fun(-2.5 + d*0.05), lwd=3)
abline(h=0.3, col=2, lty=2)


alpha <- rnorm(100, -2.5, 2)
beta <- rnorm(10000, 0.05, 0.35)
beta <- beta[beta>0]
beta <- beta[1:100]

plot(0,0,xlim=c(0,40), ylim=c(0,1), xlab='Dose', ylab='Pr(DLT)', type='n',
main='100 Prior Draws for Positive Slope Dose-DLT Curve' )
for(s in 1:100){
lines(d, inv.logit.fun(alpha[s]+beta[s]*d), col='lightblue')
}
lines(d, inv.logit.fun(-2.5 + d*0.05), lwd=5,  col='white')
lines(d, inv.logit.fun(-2.5 + d*0.05), lwd=3)
abline(h=0.3, col=2, lty=2)
```

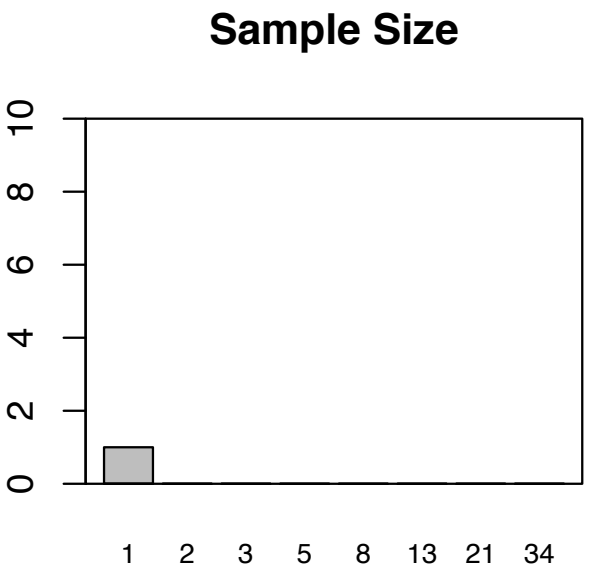
Model Update

- Based on observed DLTs or No DLTs at each dose the DLT-dose curve is updated (posterior distribution)
- Posterior Quantities:
  - Posterior for π (mean and st. dev)
  - Pr(d=SAFE): $Pr(\pi_d \leq 0.30)$
  - Pr(d=MTD): Probability dose is the MTD

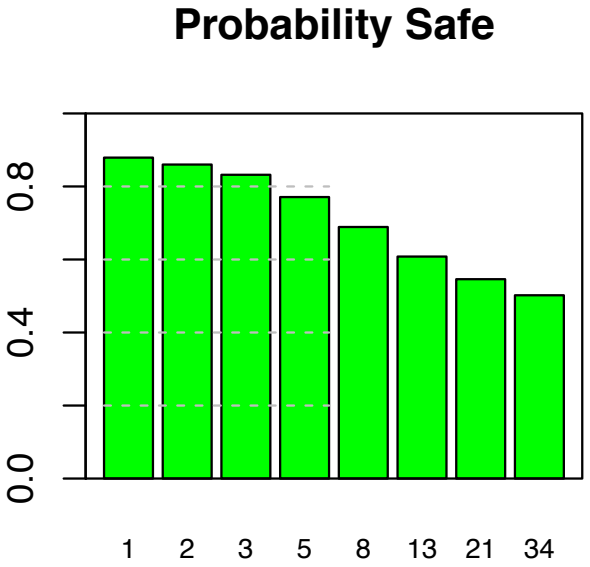
Design

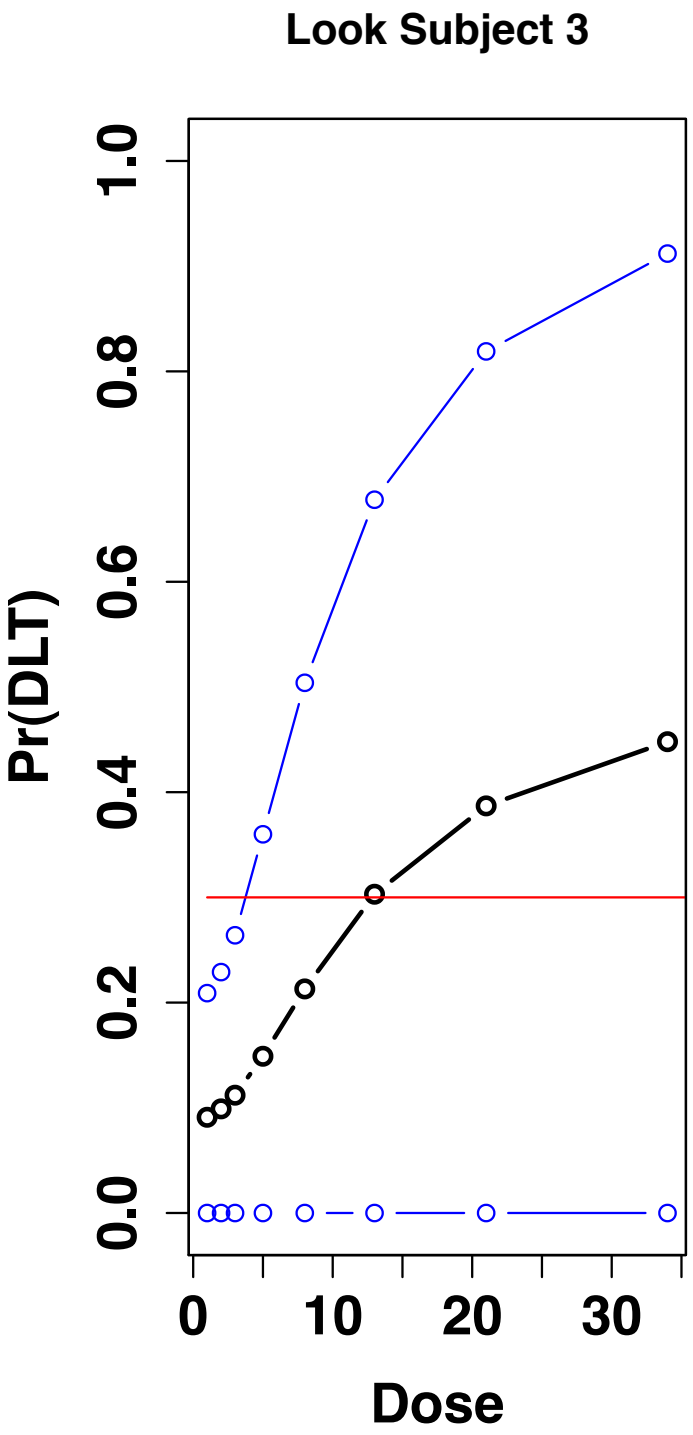
- Allocate next subject to the highest safe dose
  - "Safe" means Pr(DLT<0.30)>0.50
- Following constraints:
  - First patient enrolled to dose 1
  - Must have at least 2 with "observations" at *d* before escalate to *d+1*
  - First 6 subjects "complete information" before accrue next
- Open enrollment after 6 (controlled)
- Stop if n=10 at "current dose"
- Stop for futility if Pr(d=Safe)<0.25 for all doses (otherwise *d*=1) is used

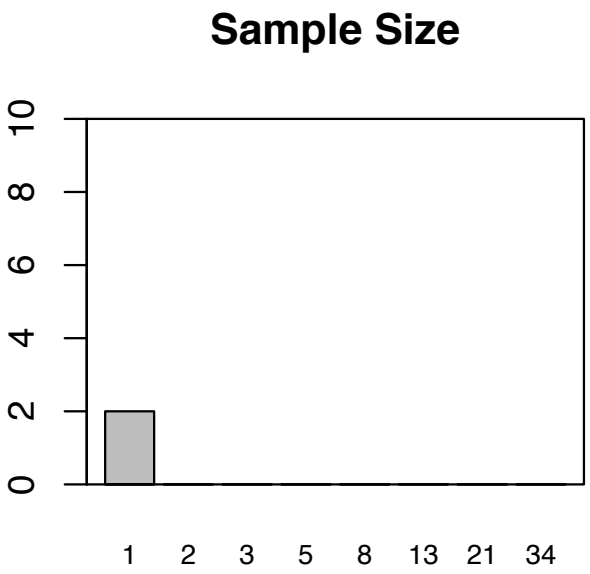




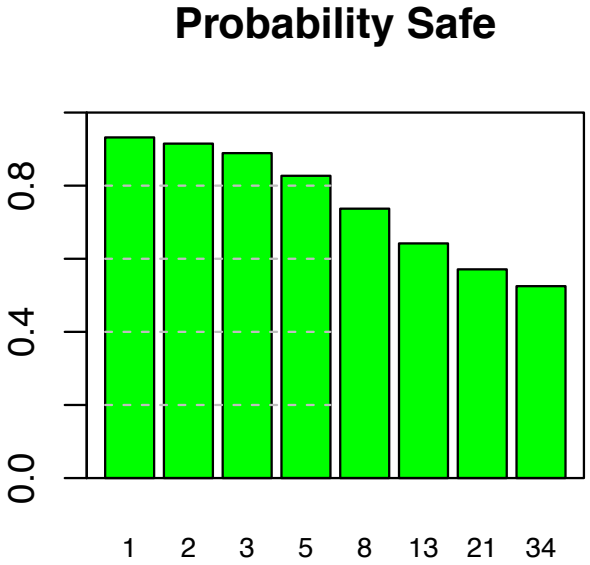
Allocate subject 2 to d=1

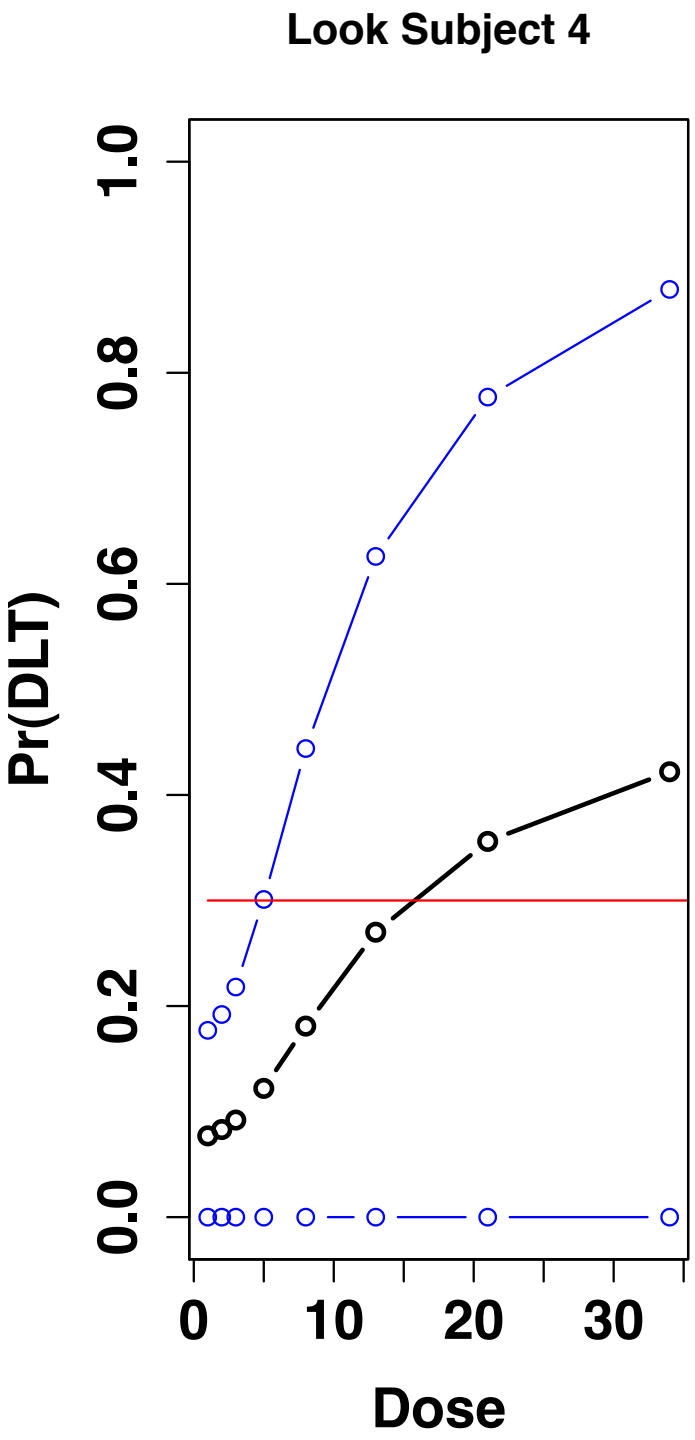


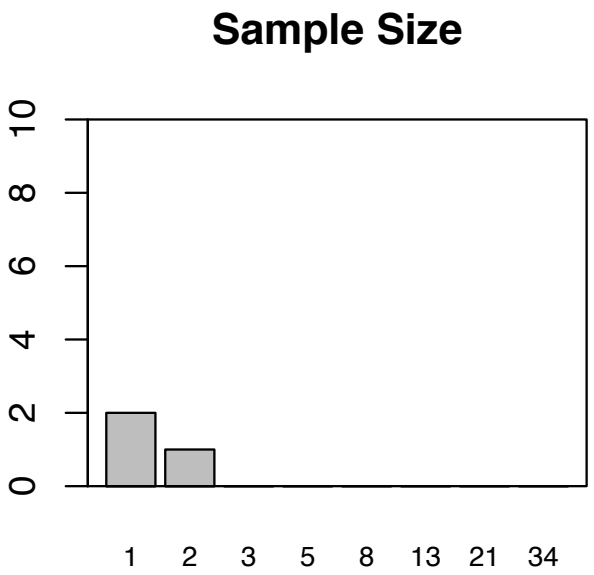


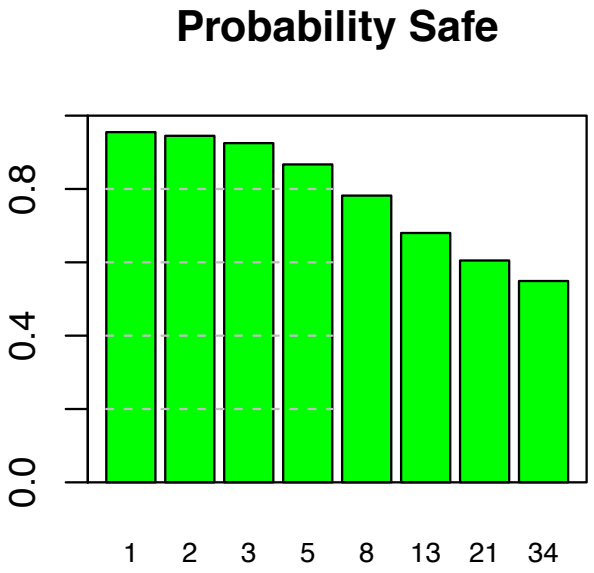


Allocate subject 3 to d=2

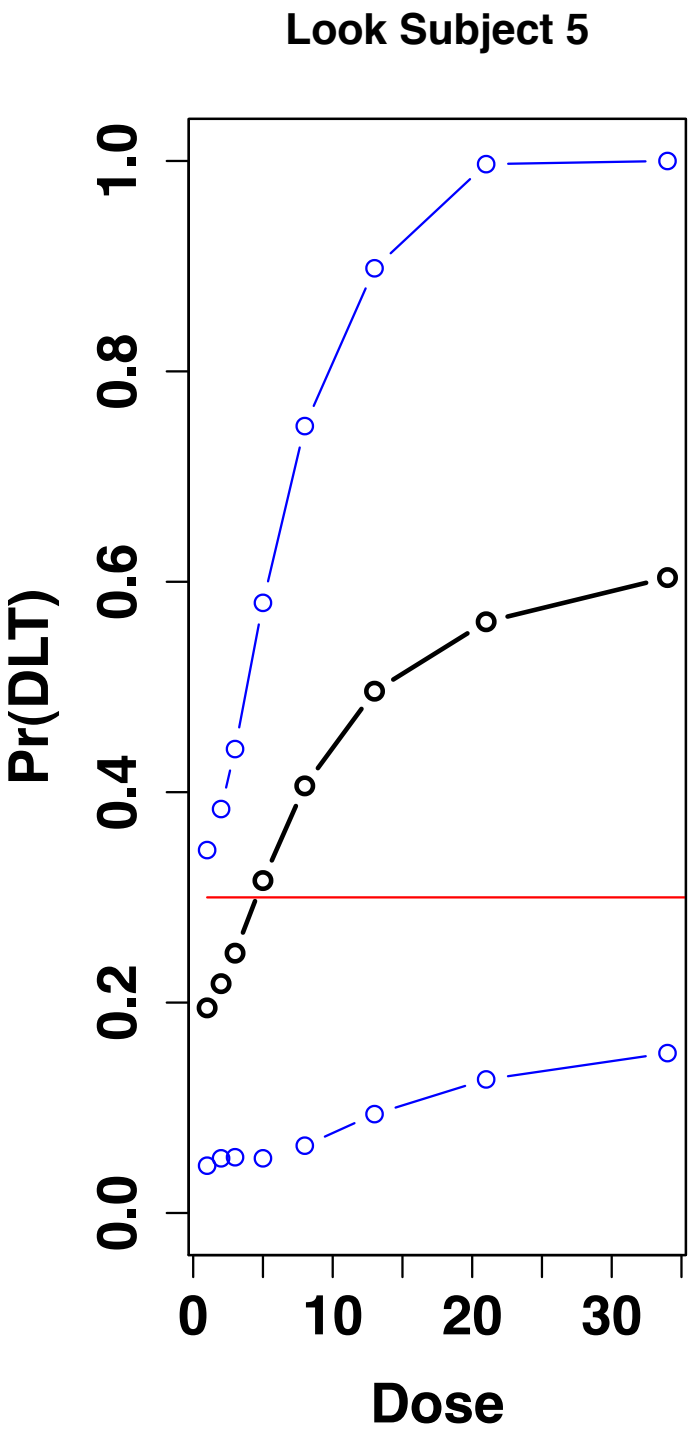


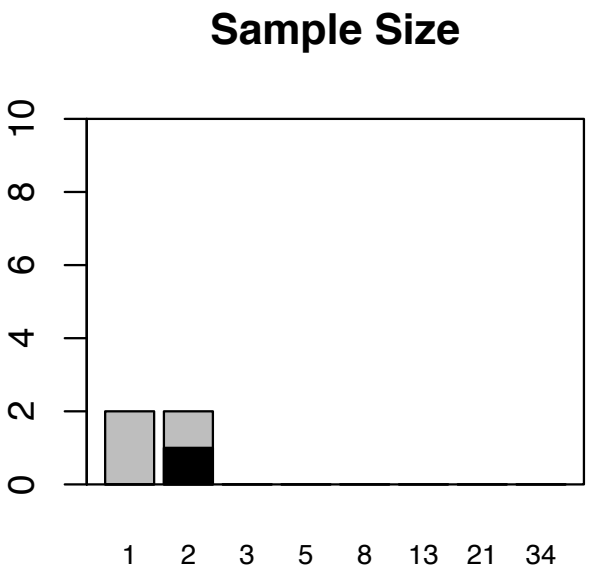




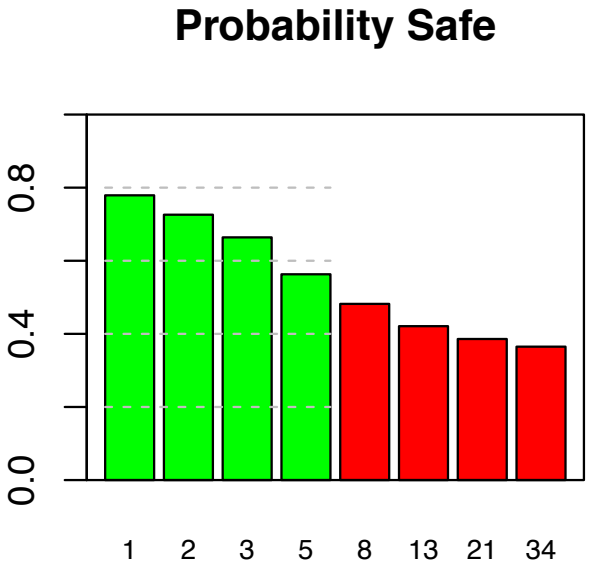


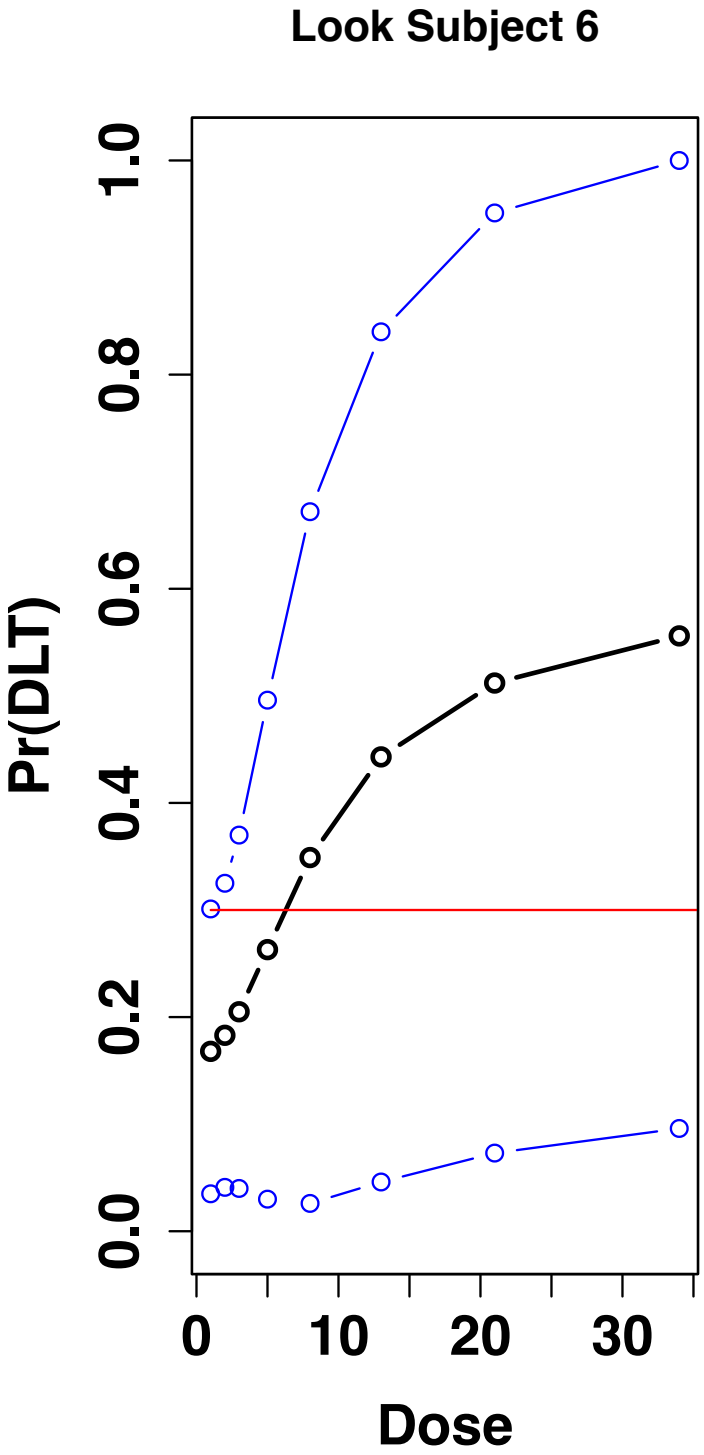
Allocate subject 4 to d=2

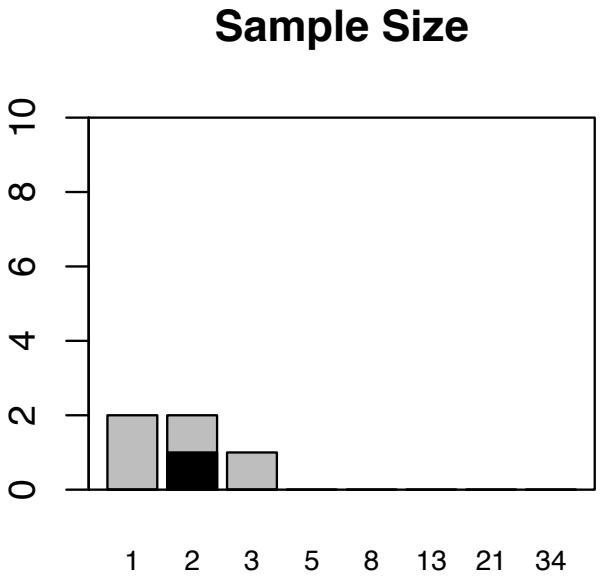




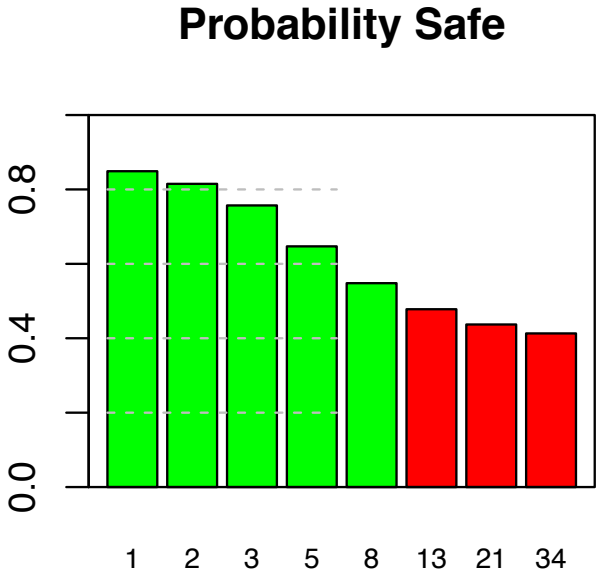
Allocate subject 5 to d=3

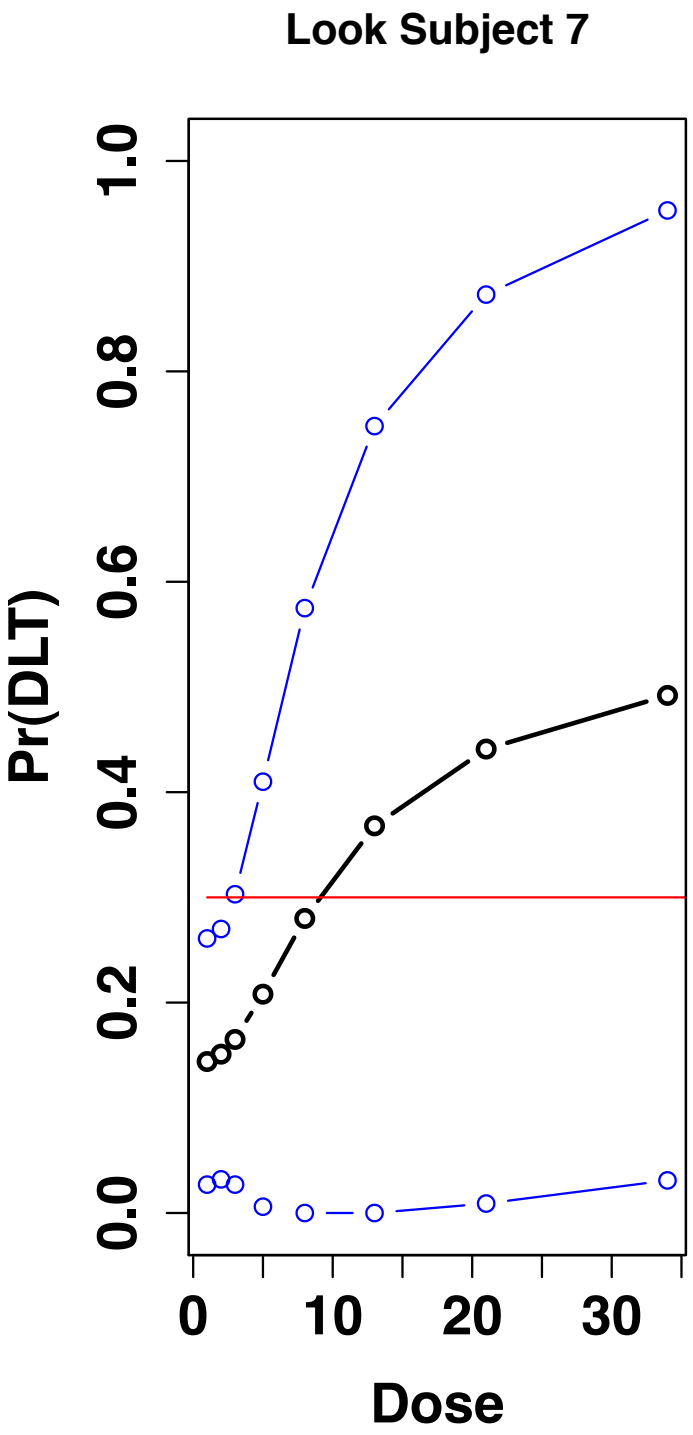


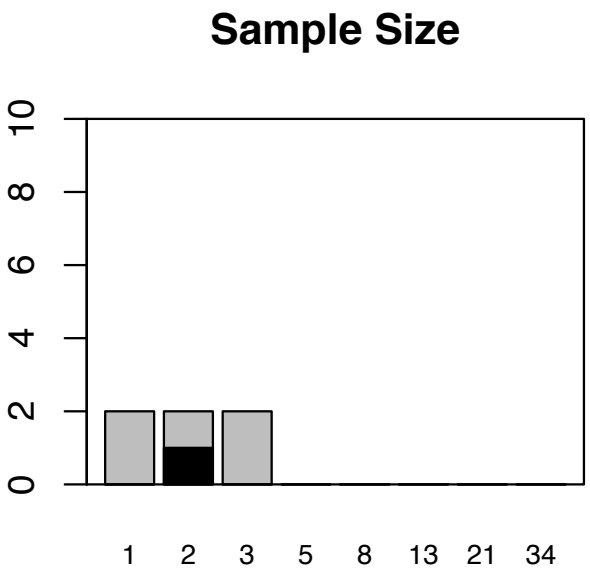




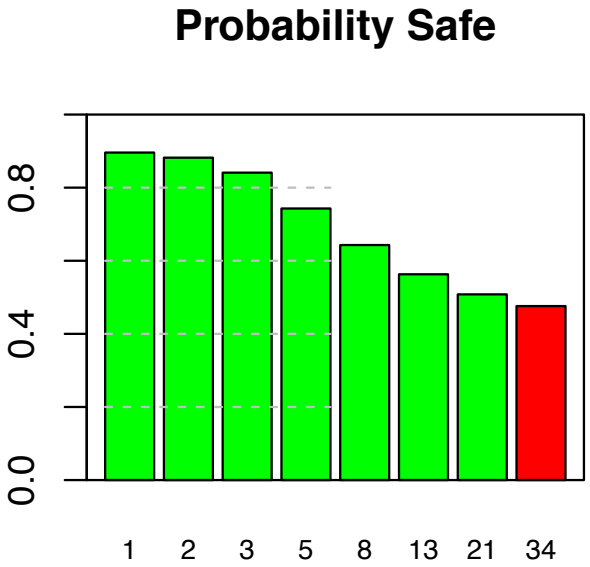
Allocate subject 6 to d=3

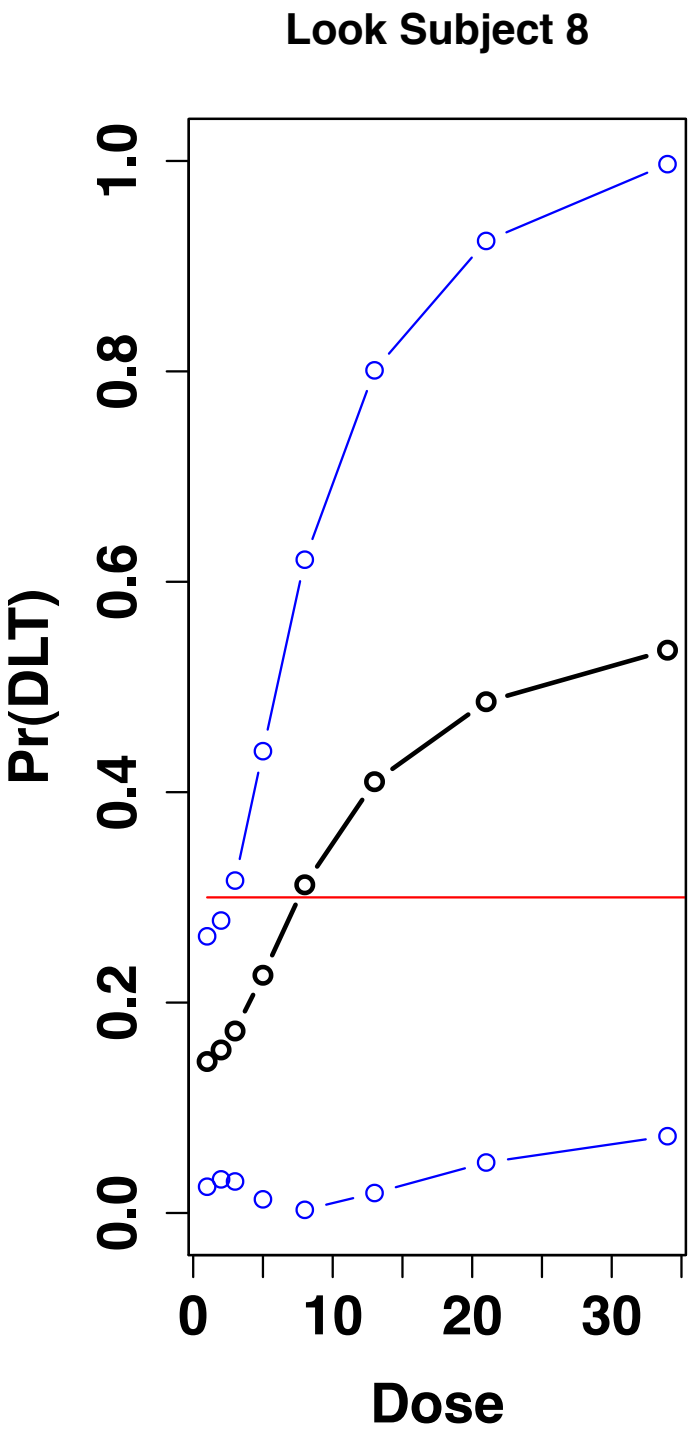


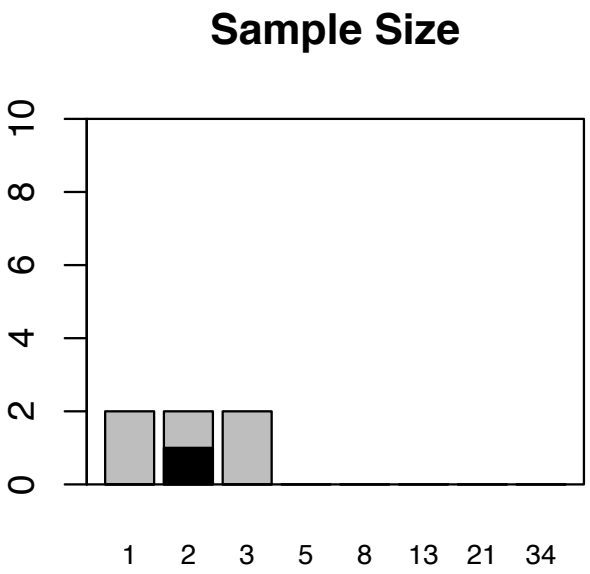




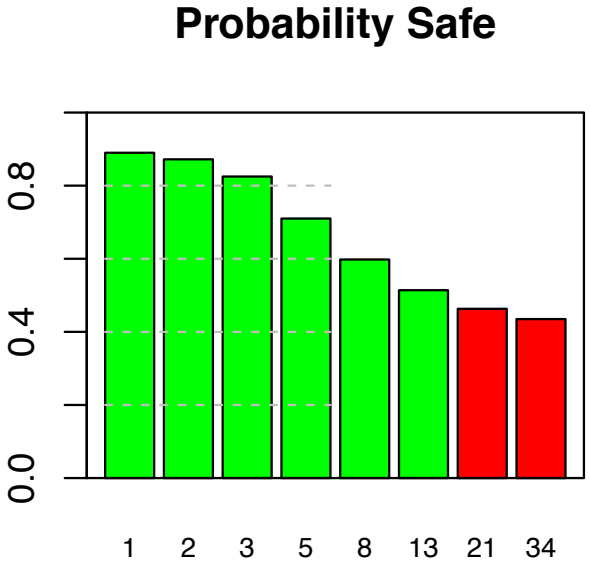
Allocate subject 7 to d=5

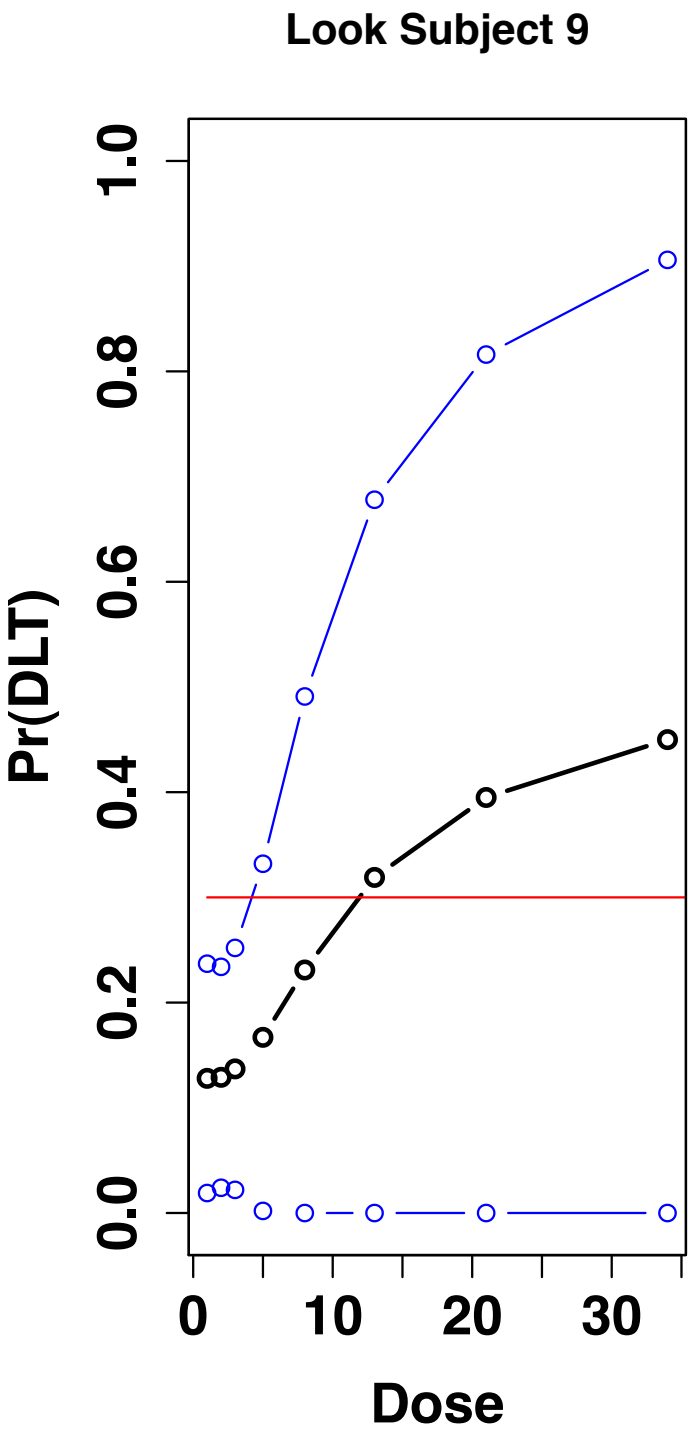


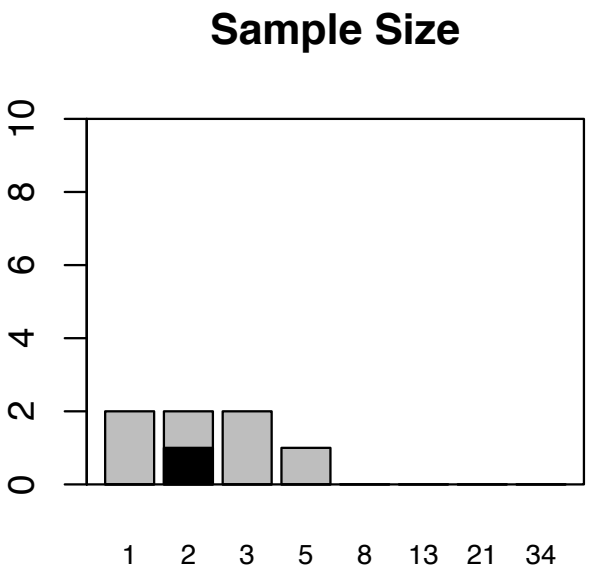




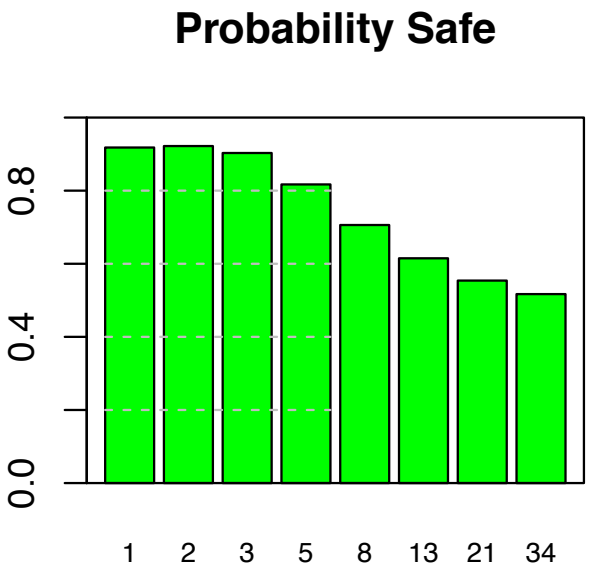
Allocate subject 8 to d=5

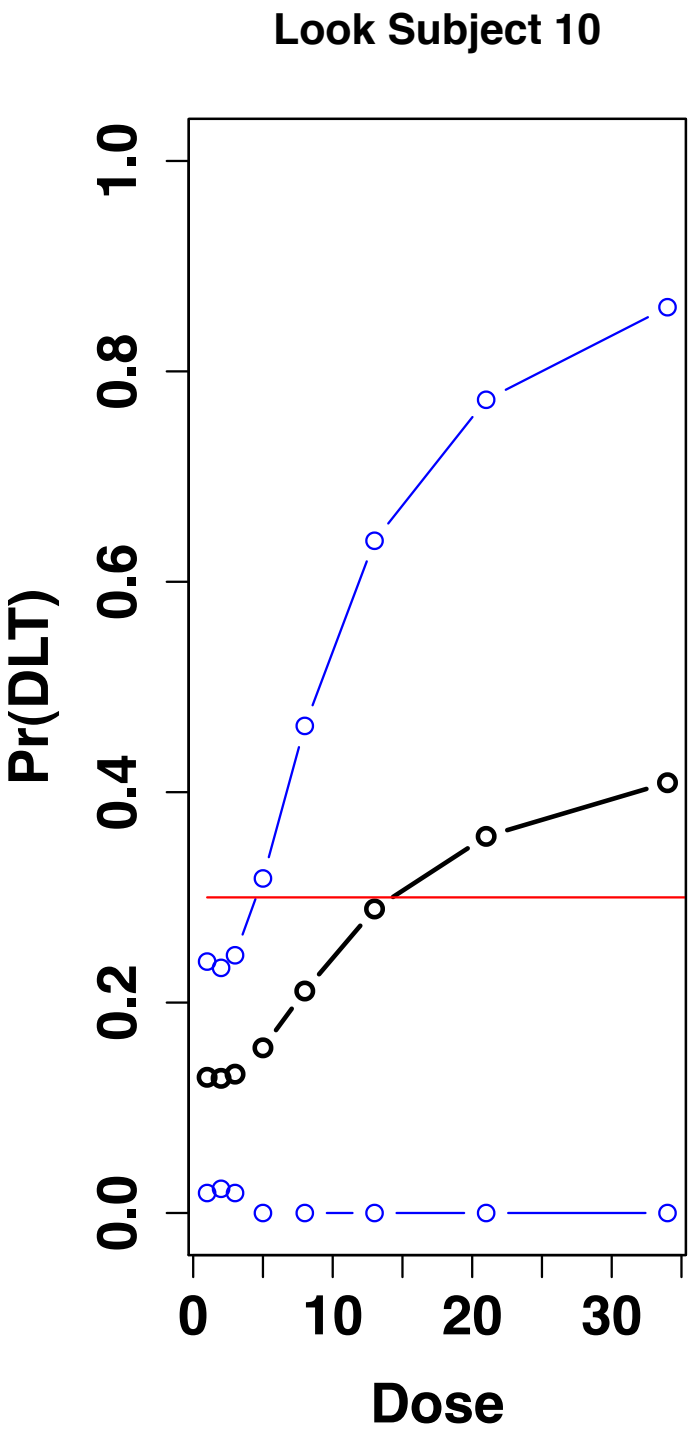


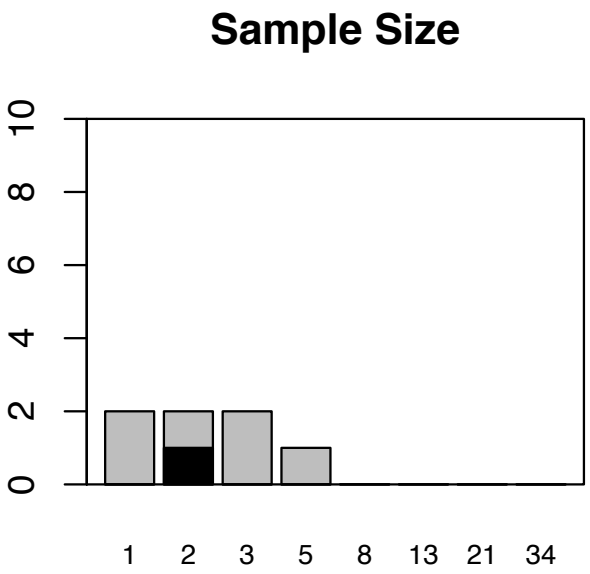


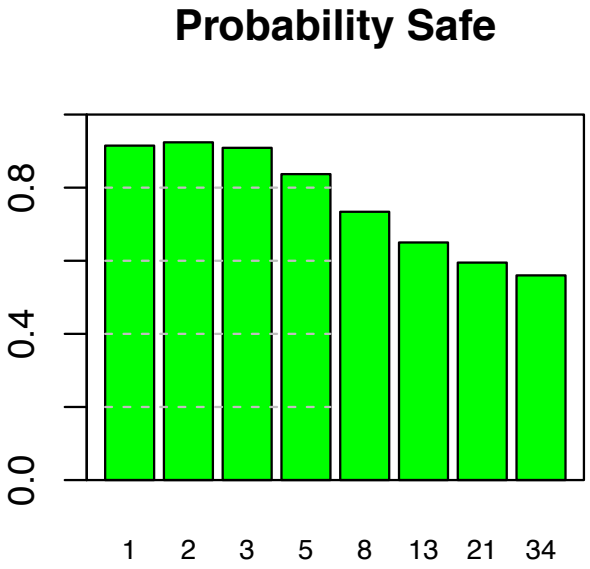


Allocate subject 9 to d=5

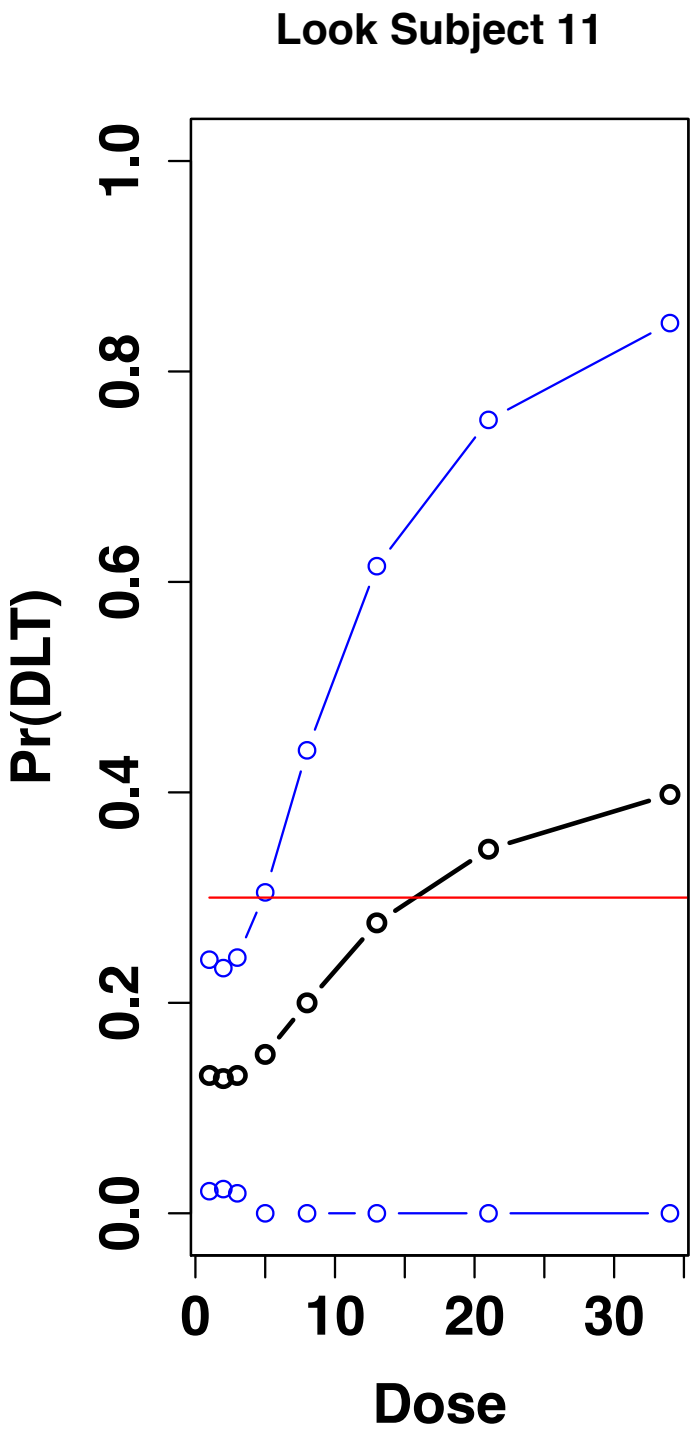


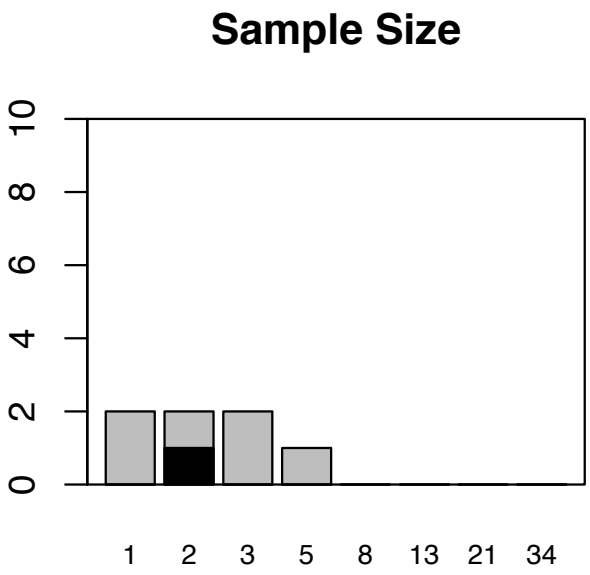




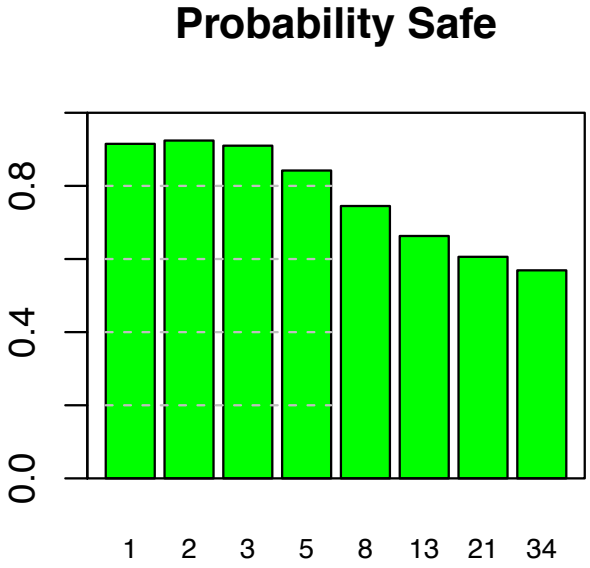


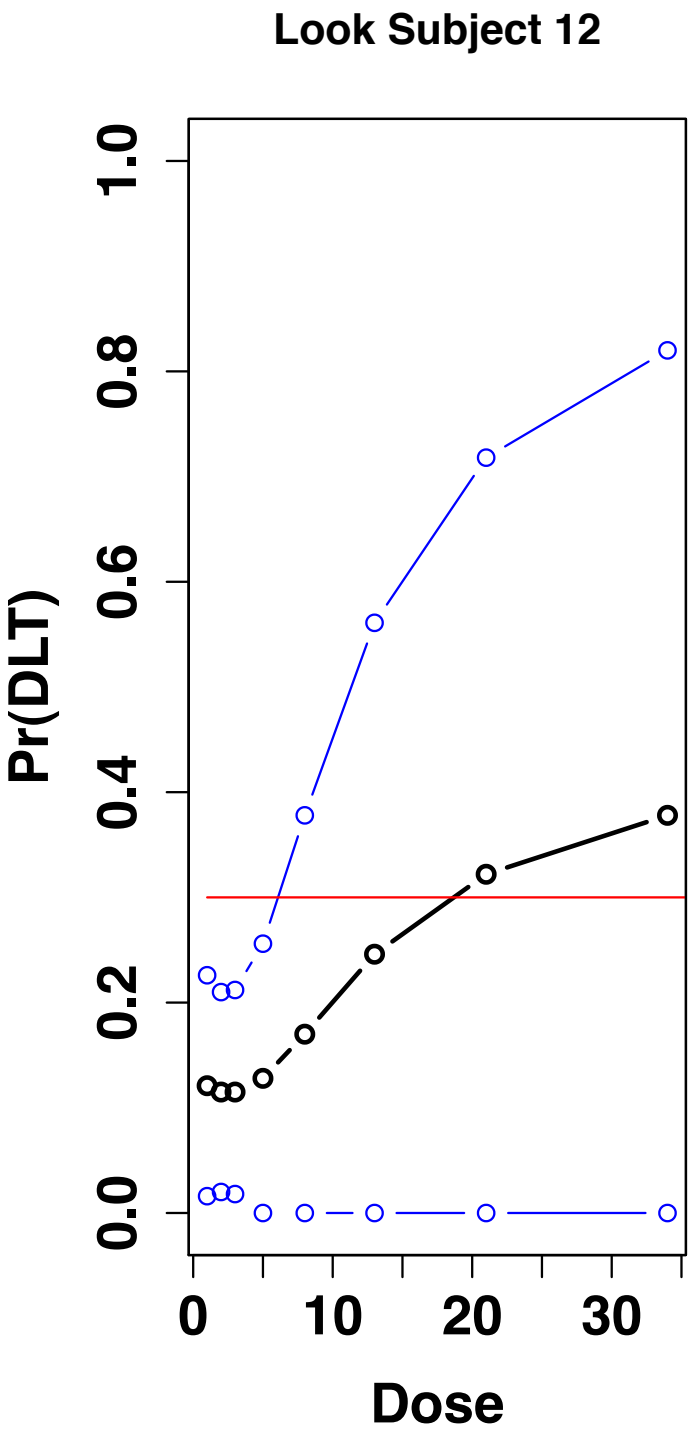
Allocate subject 10 to d=5

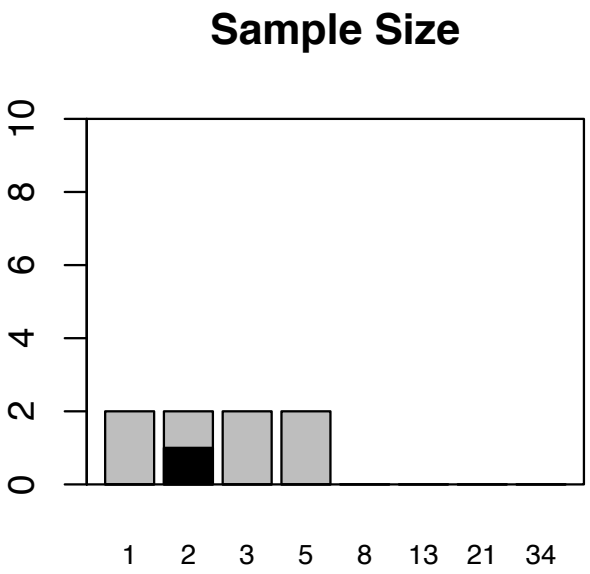


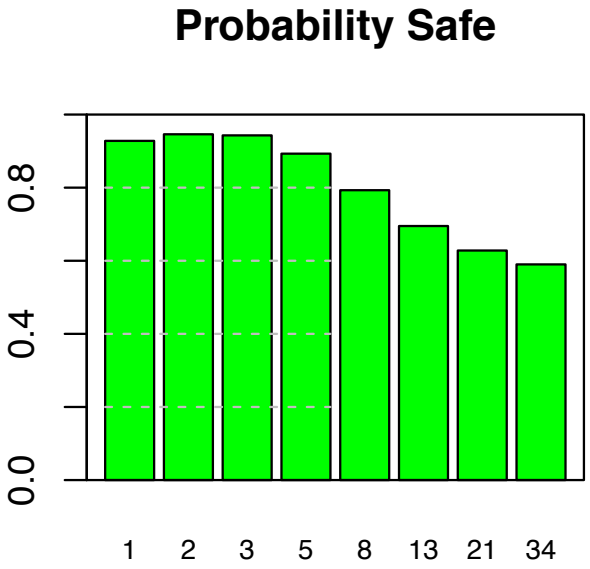


Allocate subject 11 to d=5

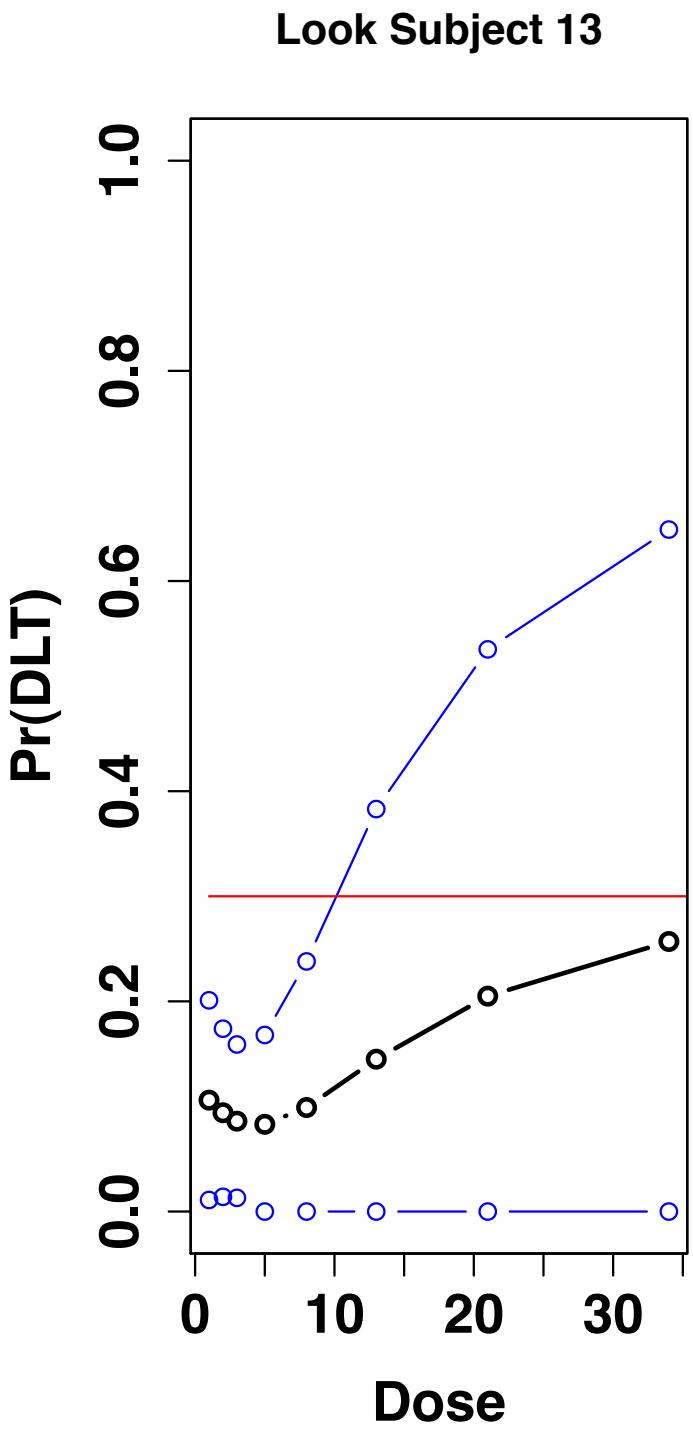


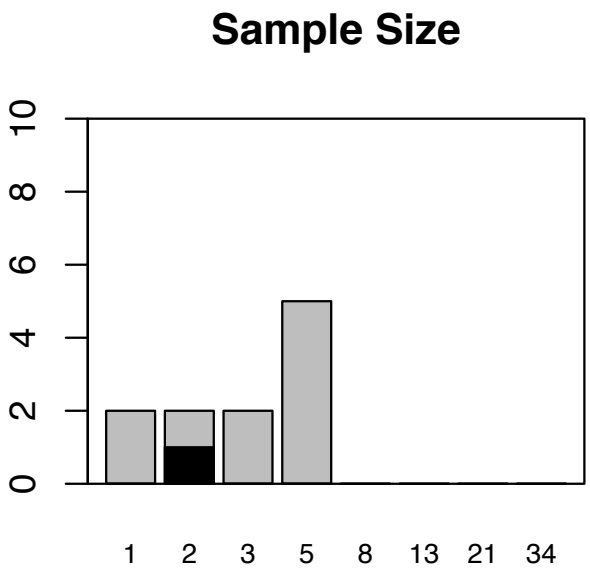


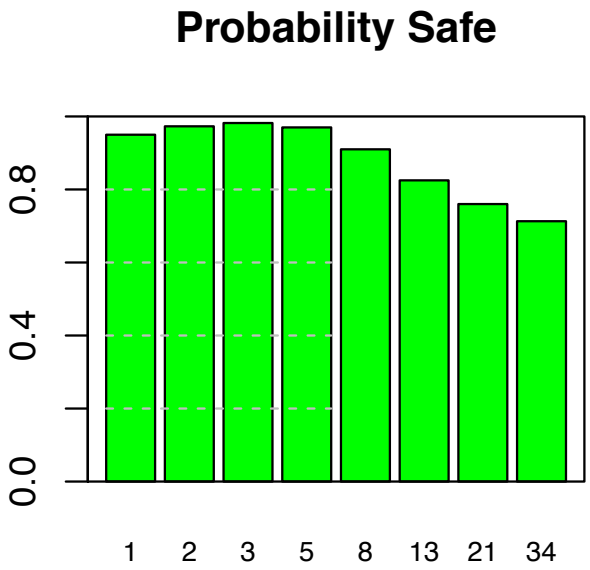




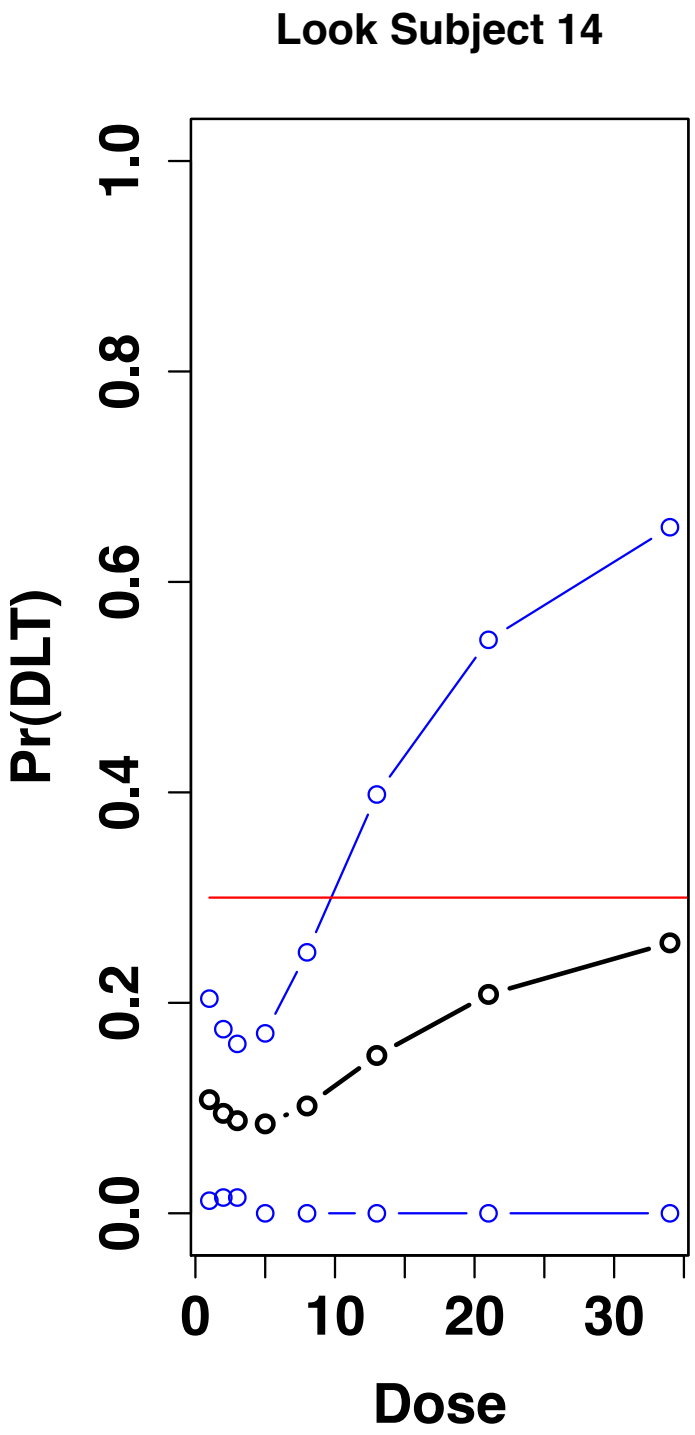
Allocate subject 12 to d=8

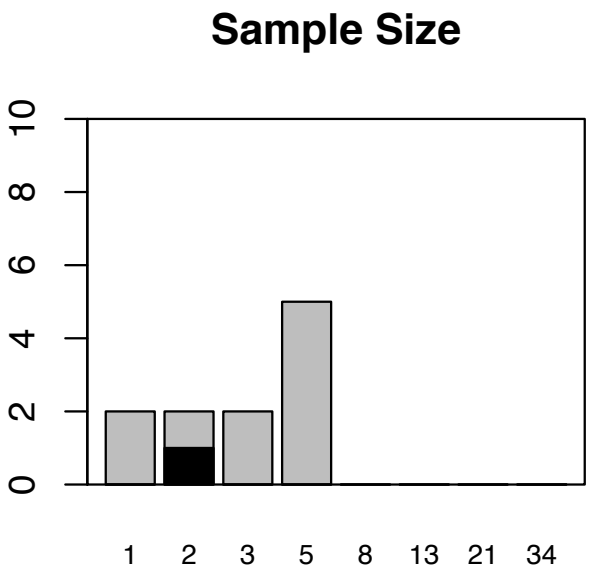


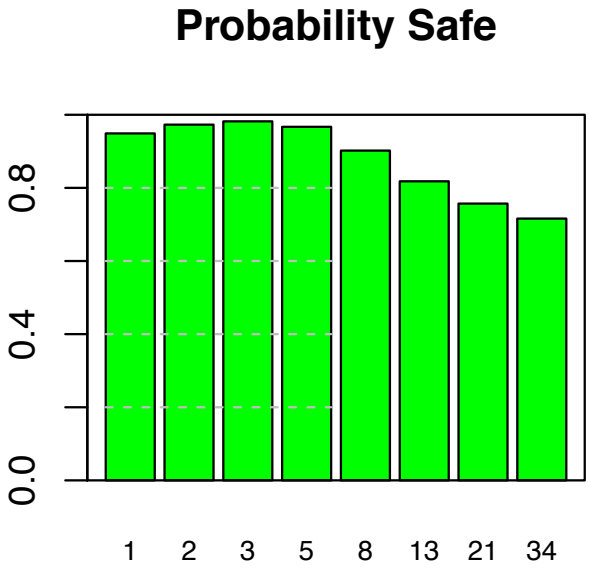




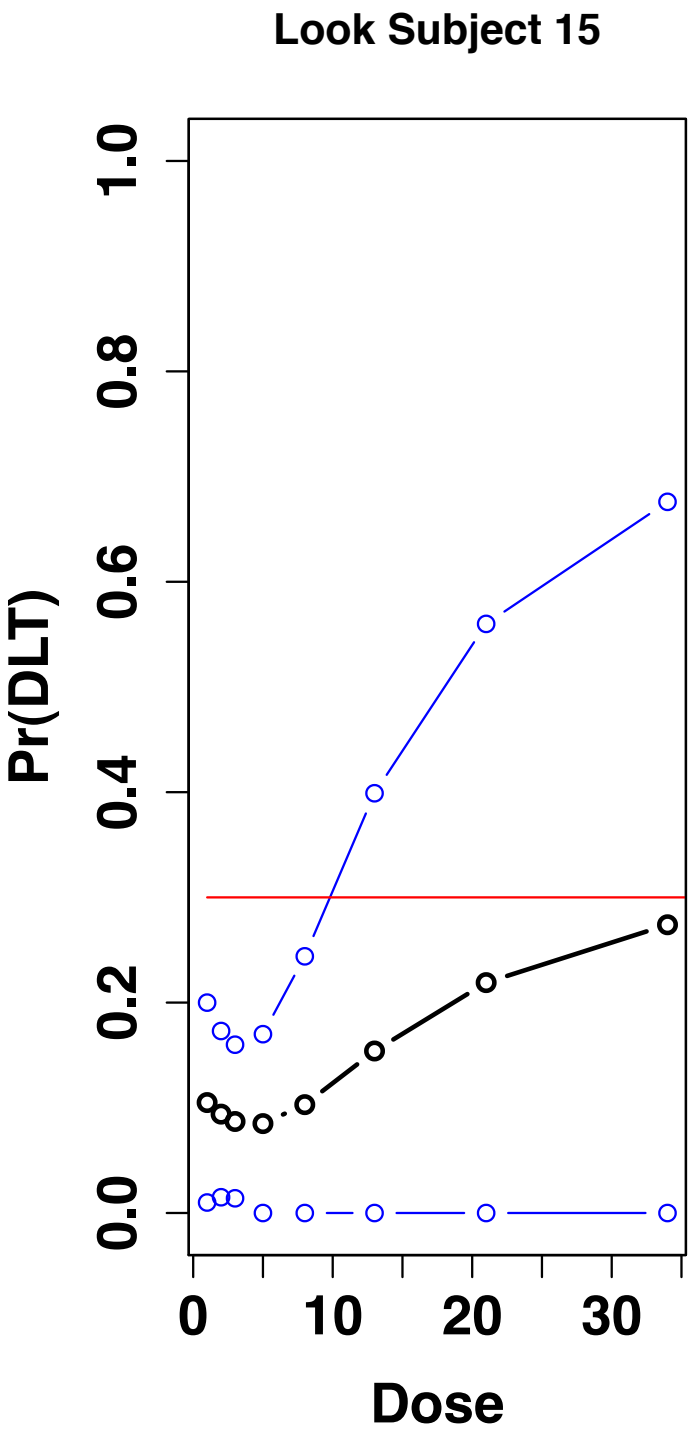
Allocate subject 13 to d=8

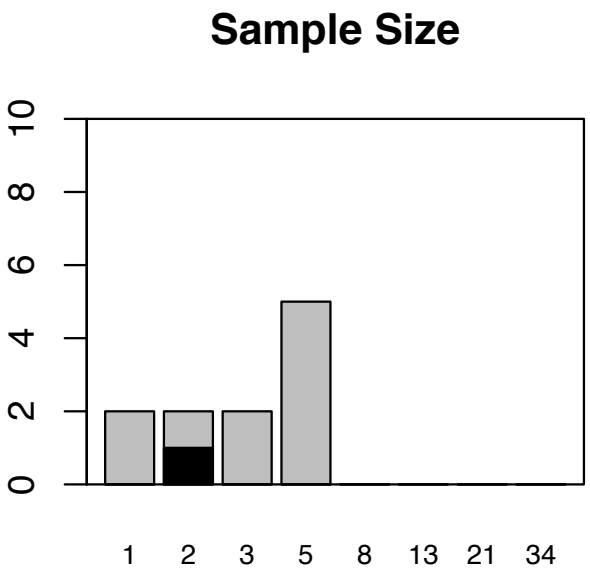




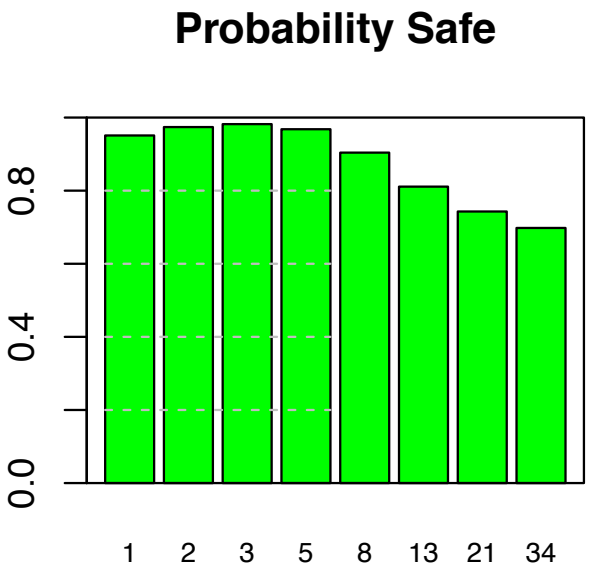


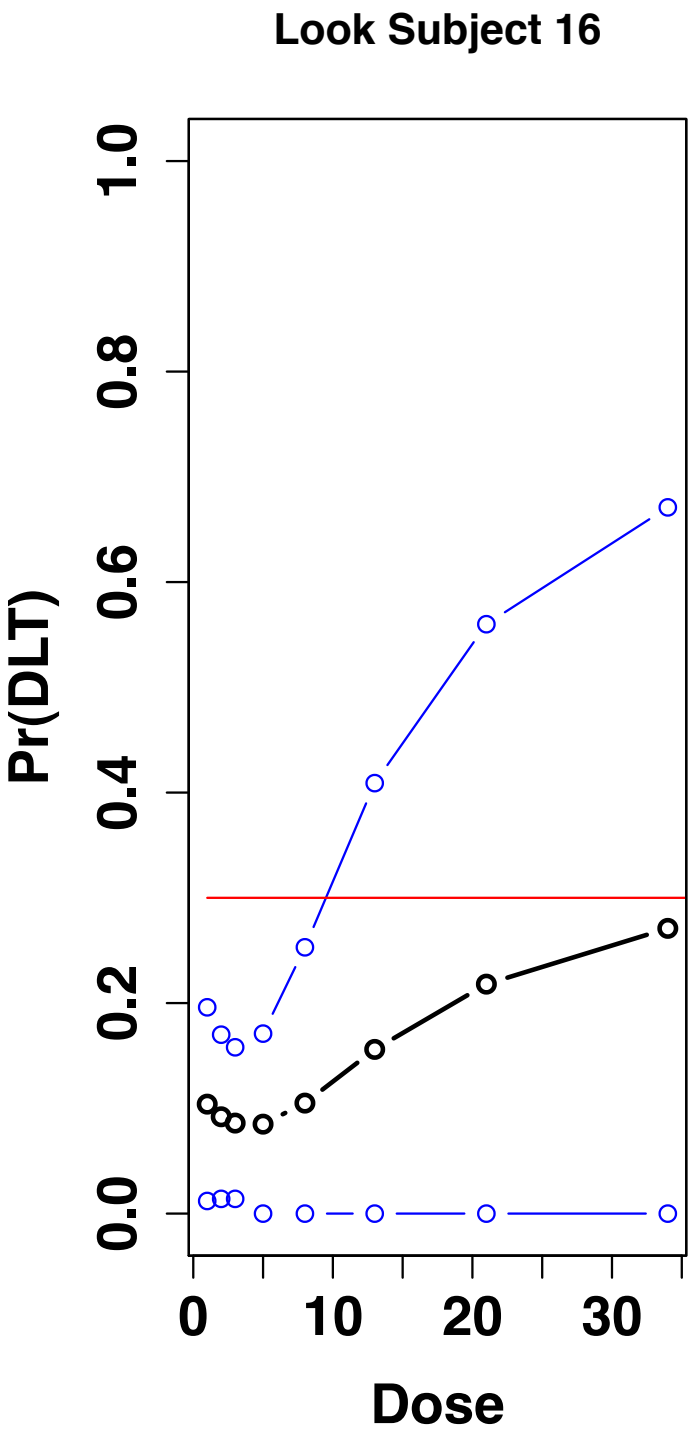
Allocate subject 14 to d=8

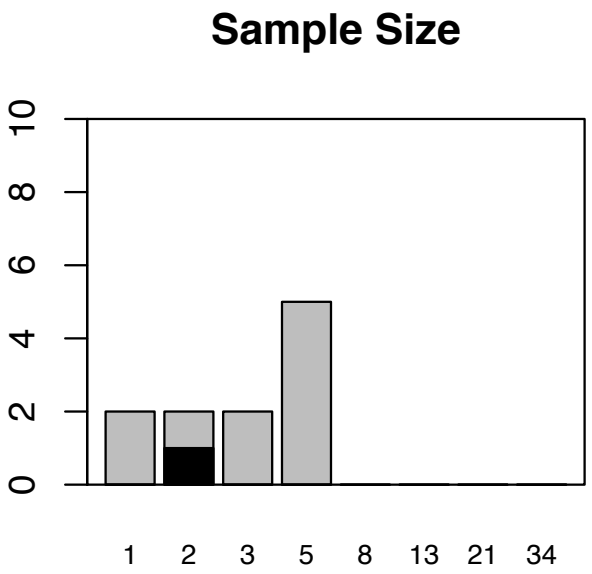


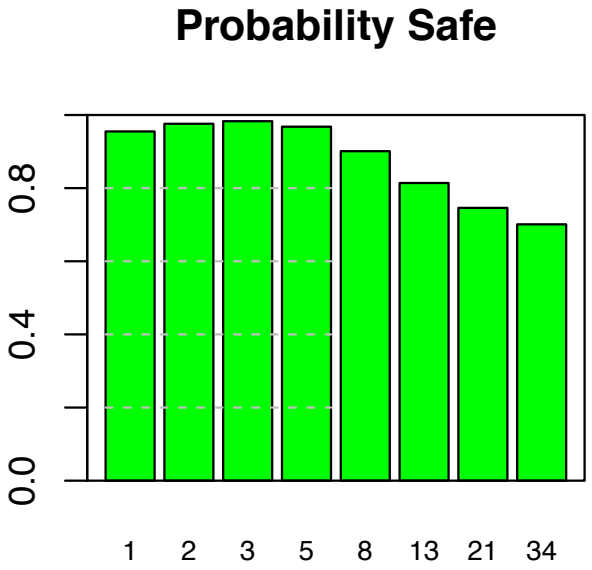


Allocate subject 15 to d=8

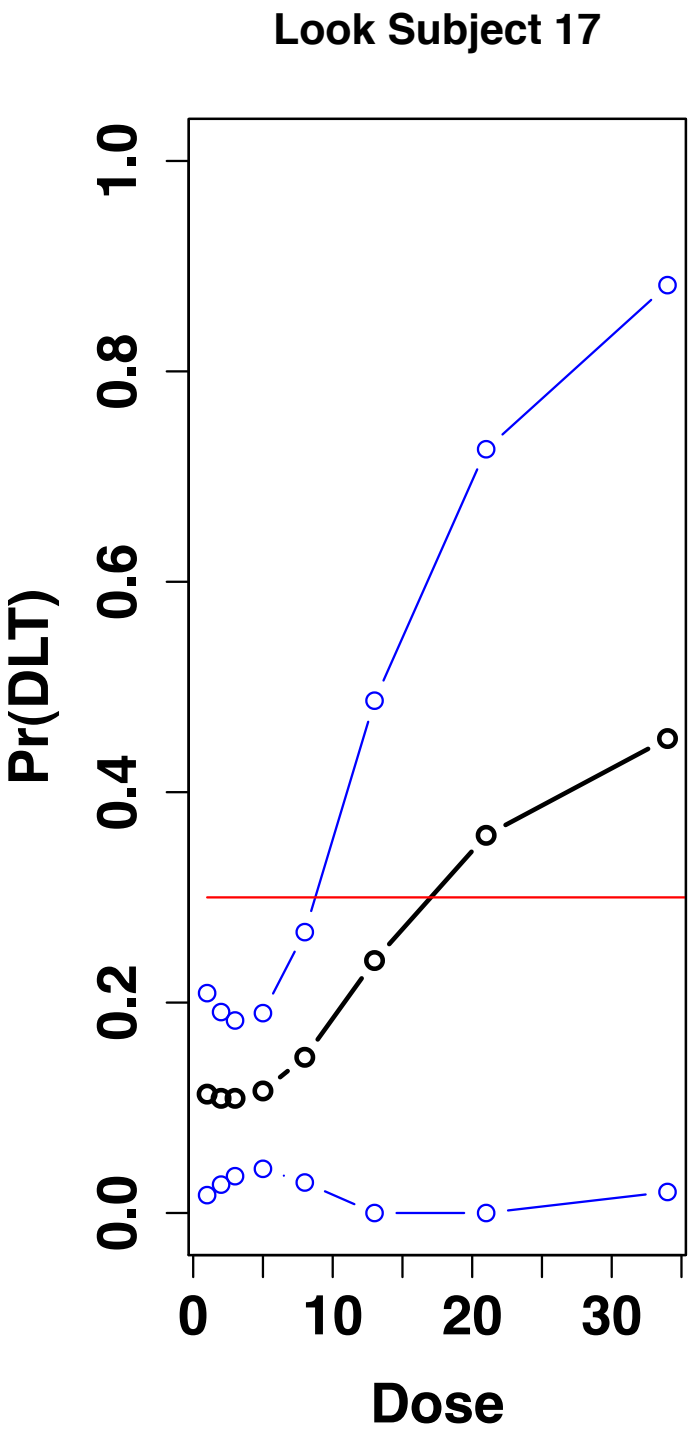


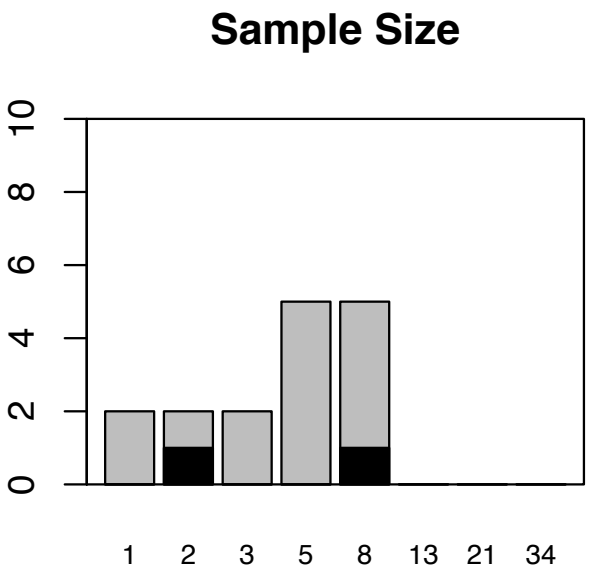


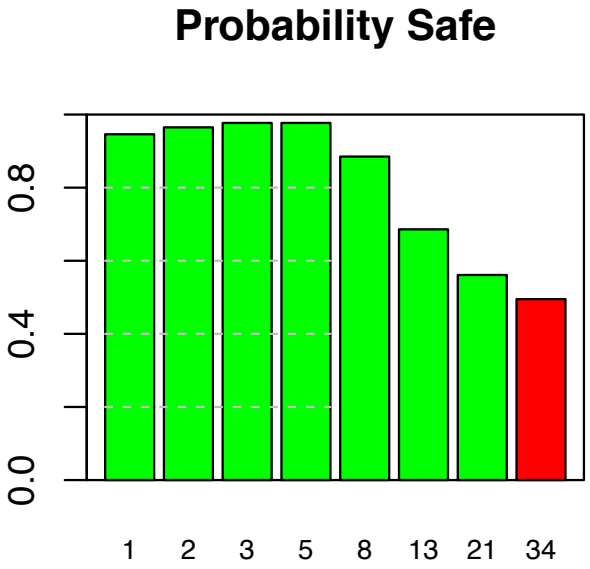




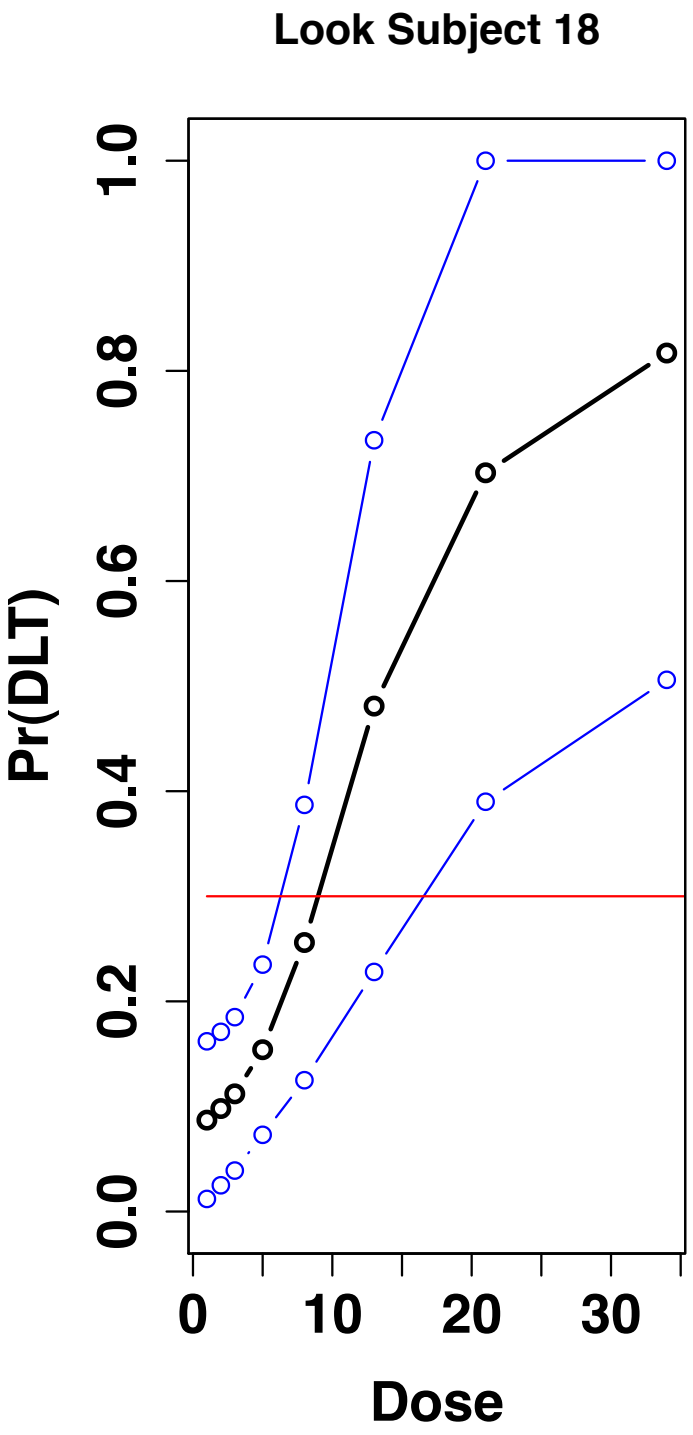
Allocate subject 16 to d=8

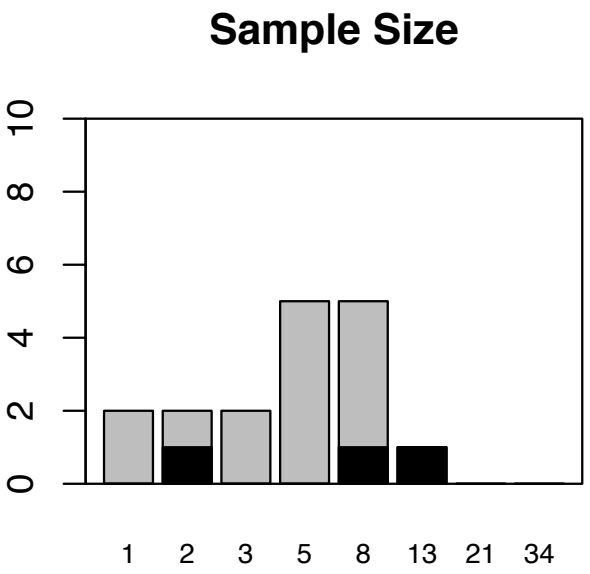


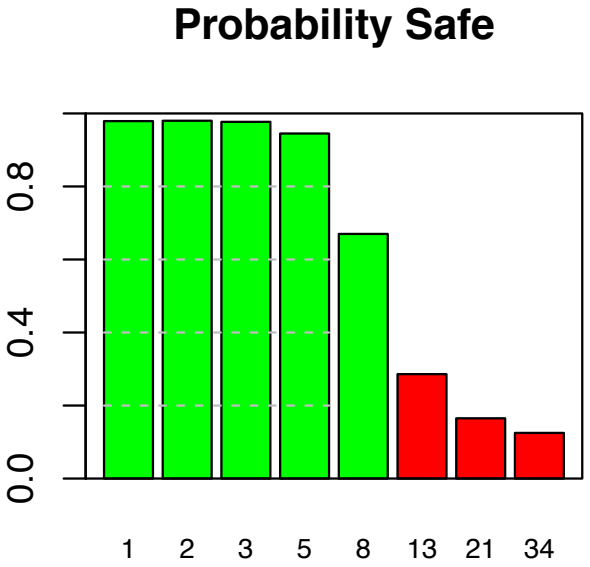




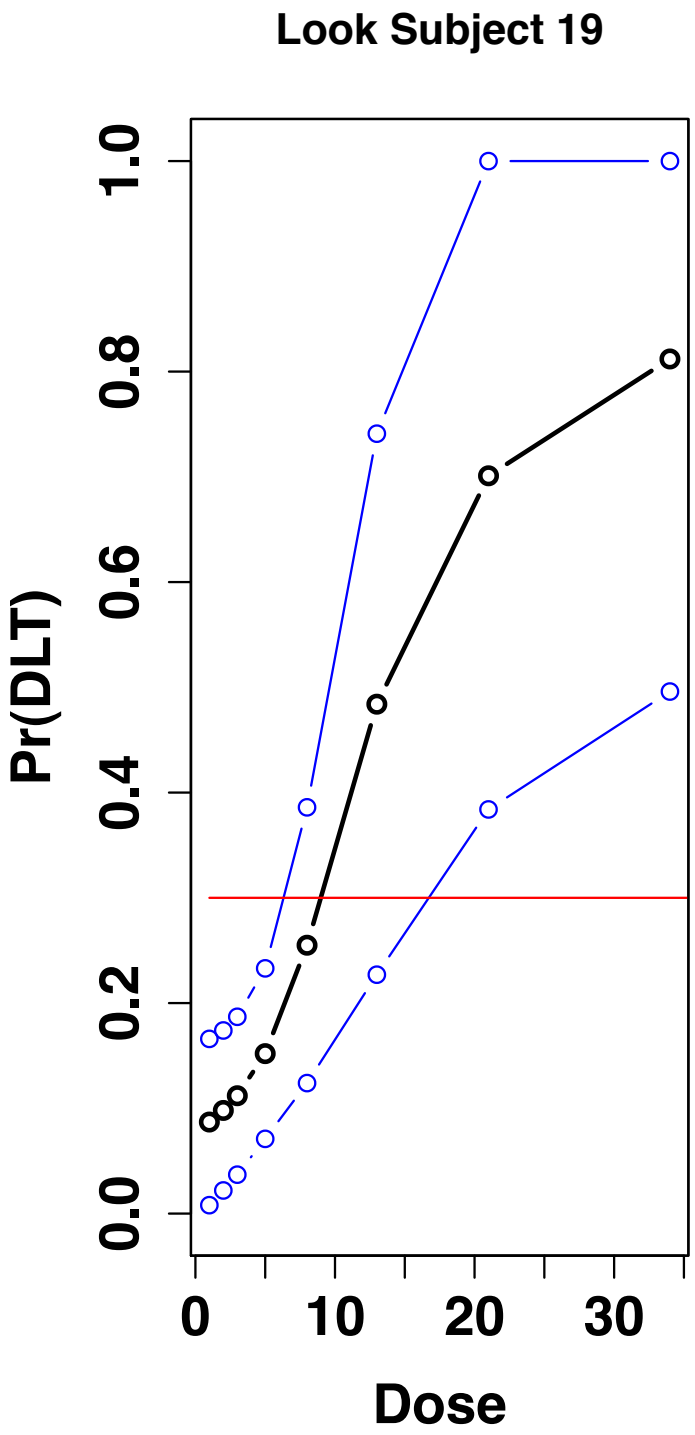
Allocate subject 17 to d=13

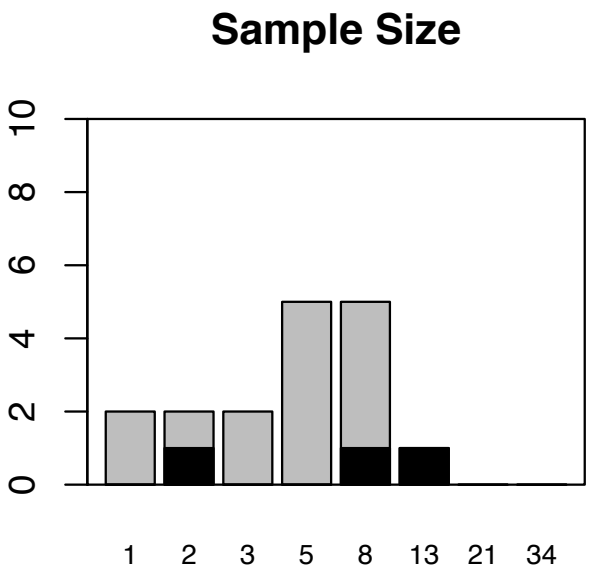


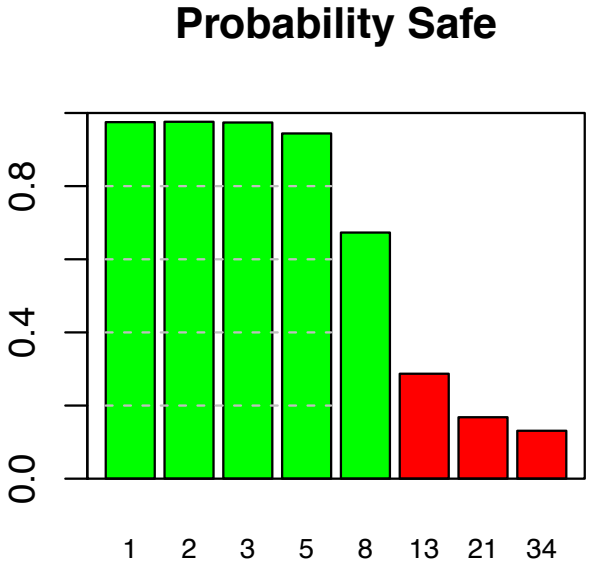




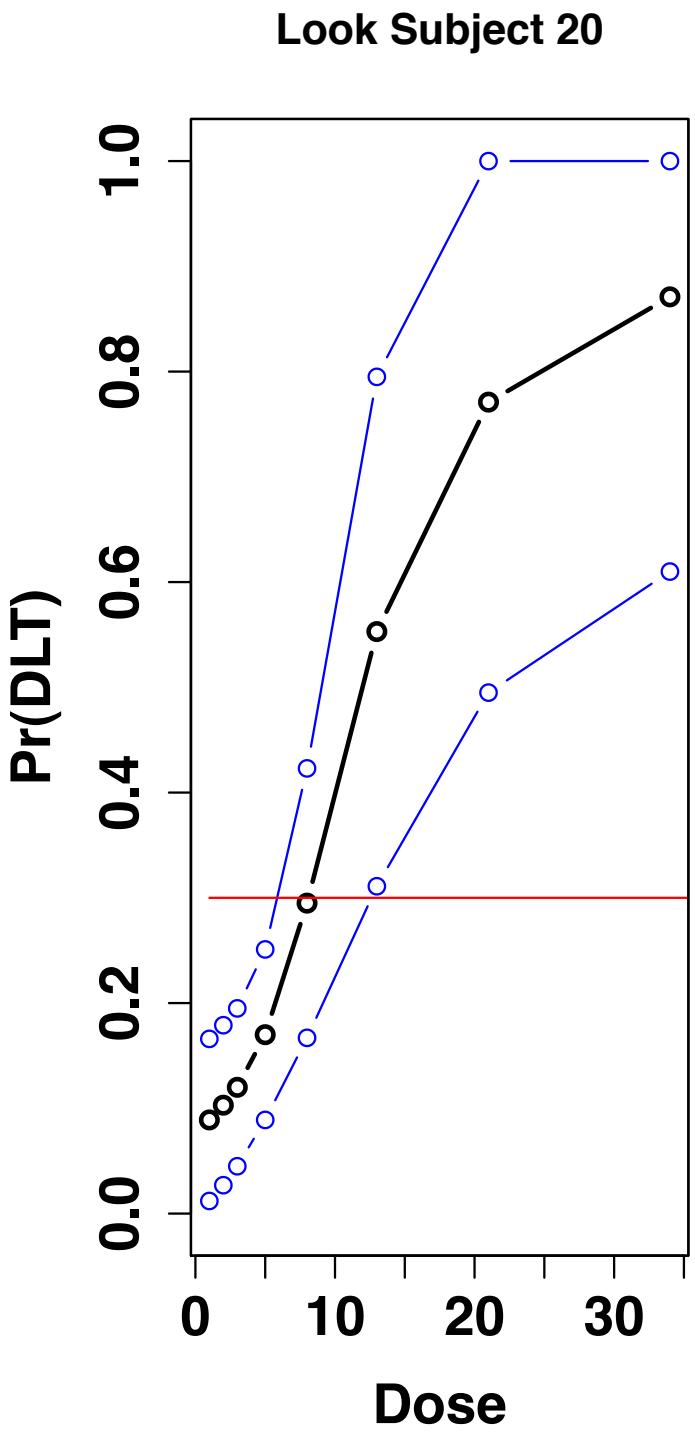
Allocate subject 18 to d=8

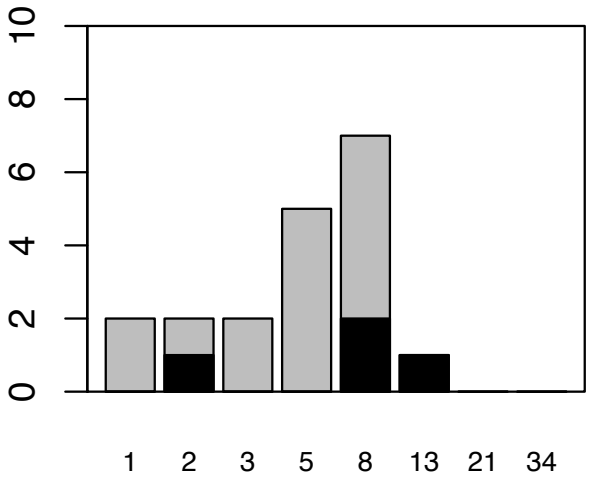


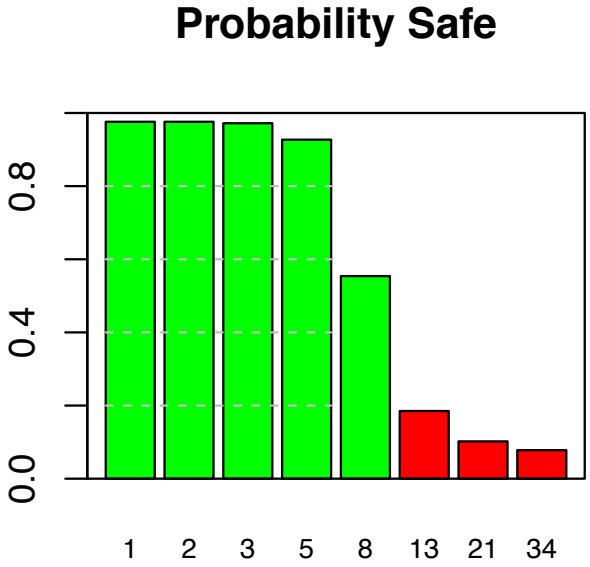




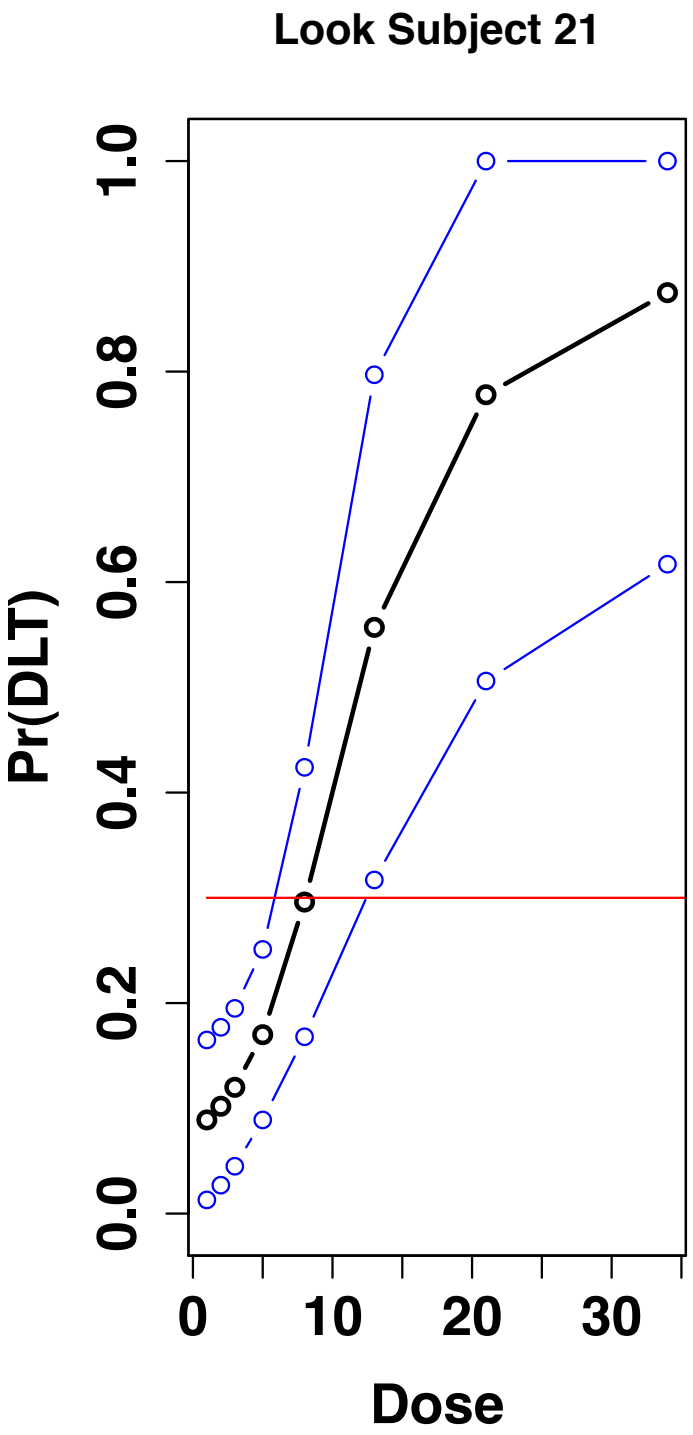
Allocate subject 19 to d=8

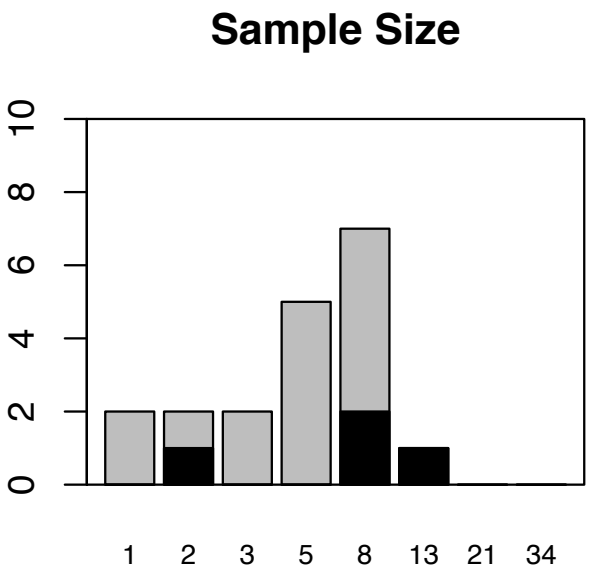


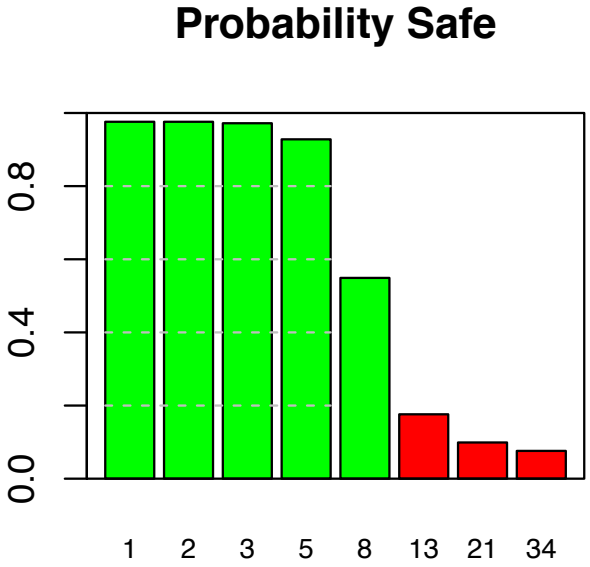




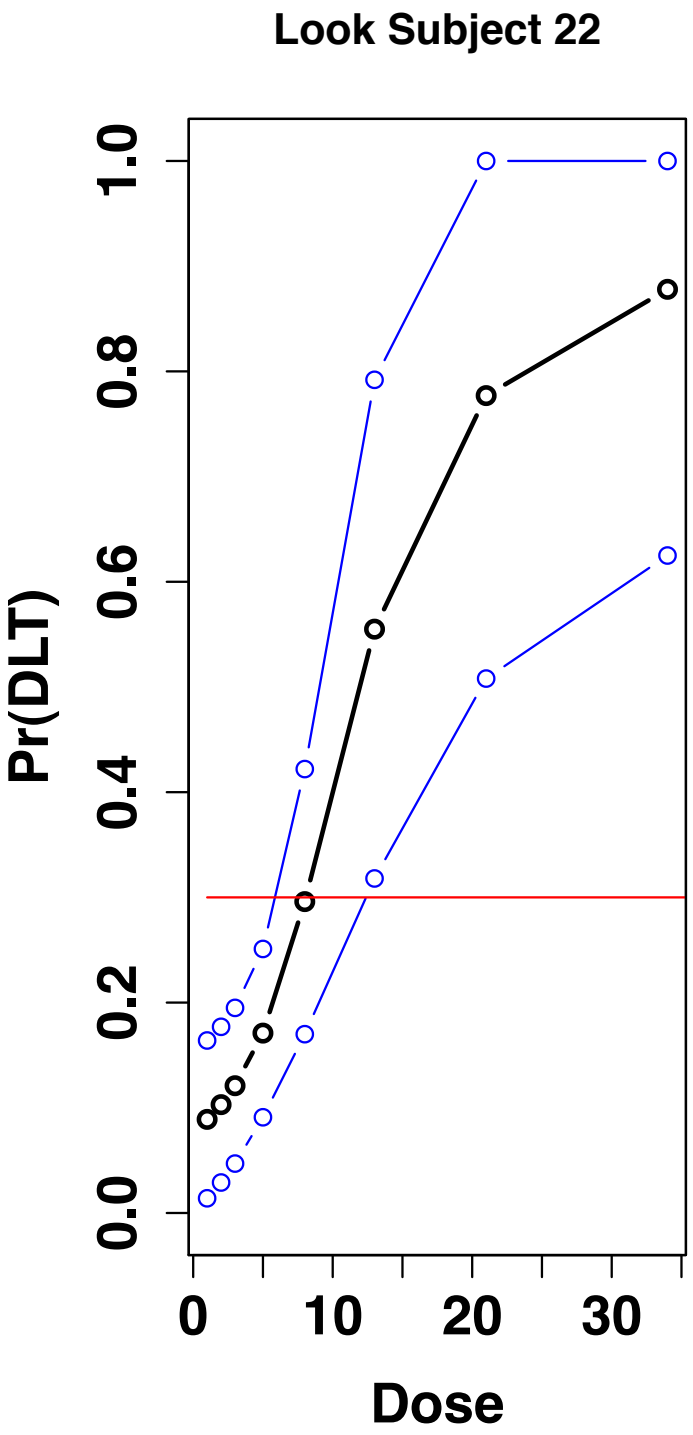
Allocate subject 20 to d=8

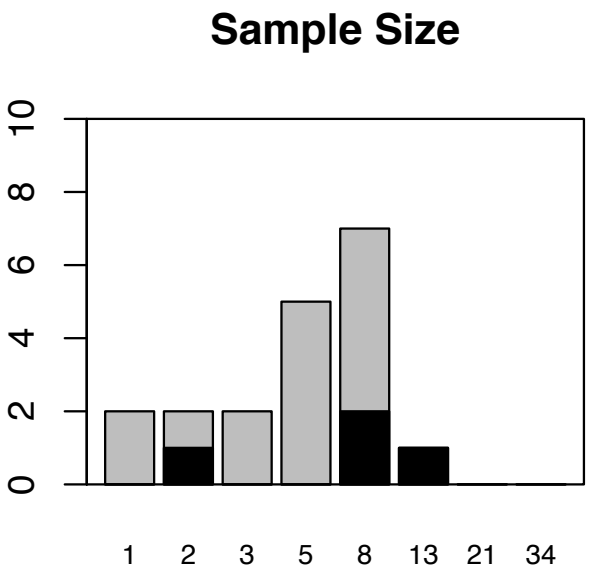


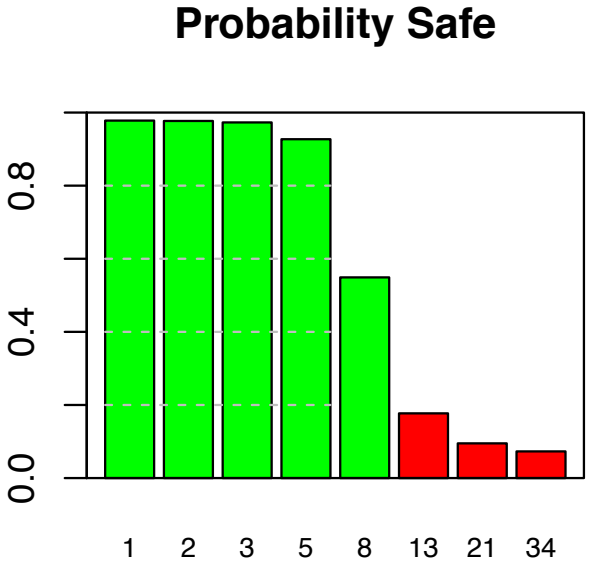




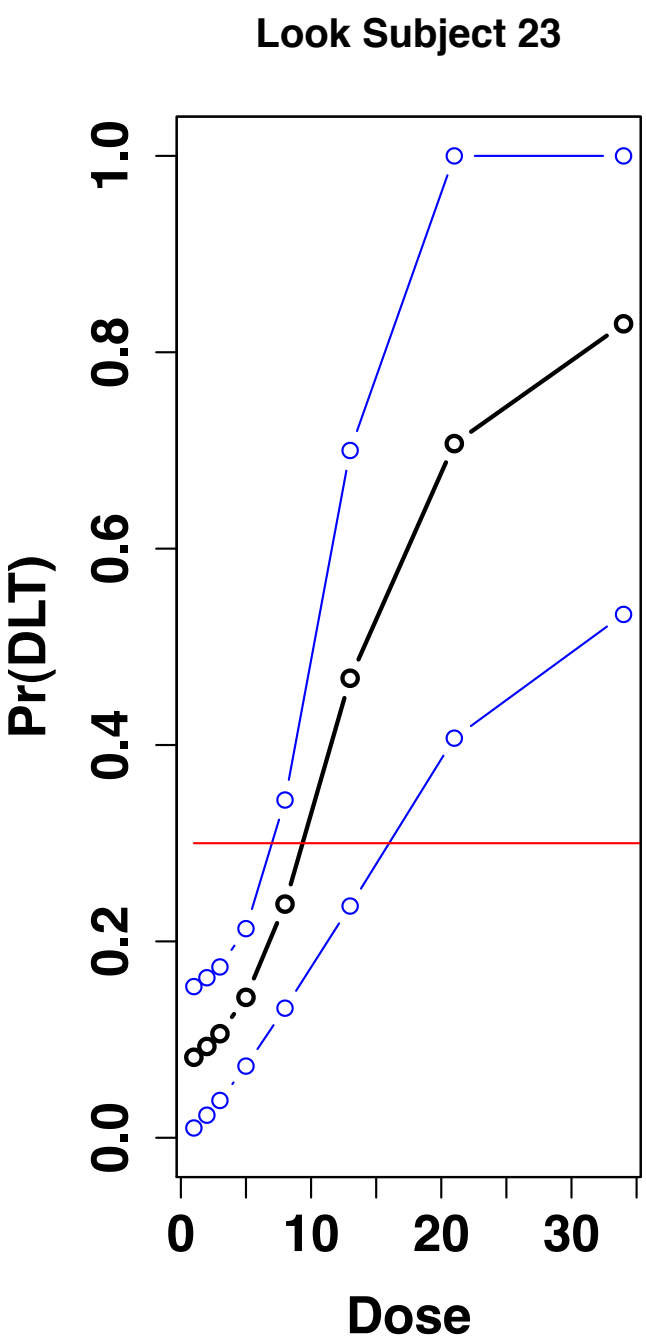
Allocate subject 21 to d=8

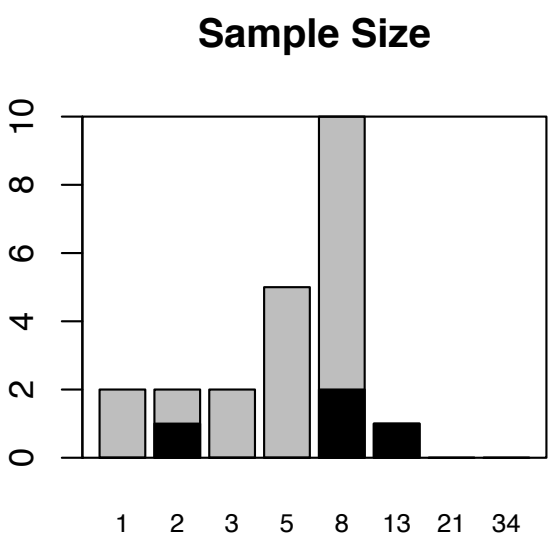


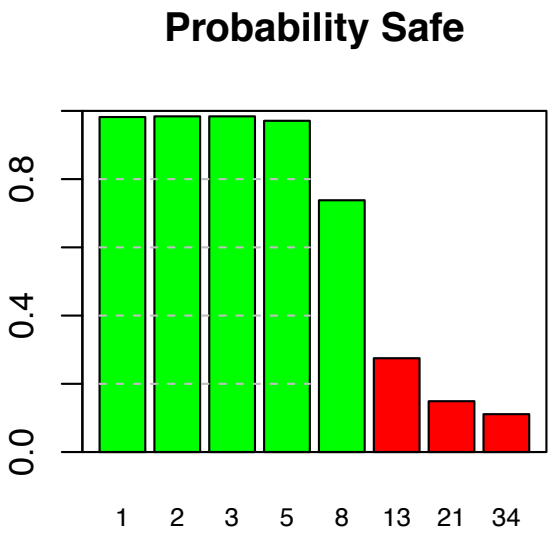




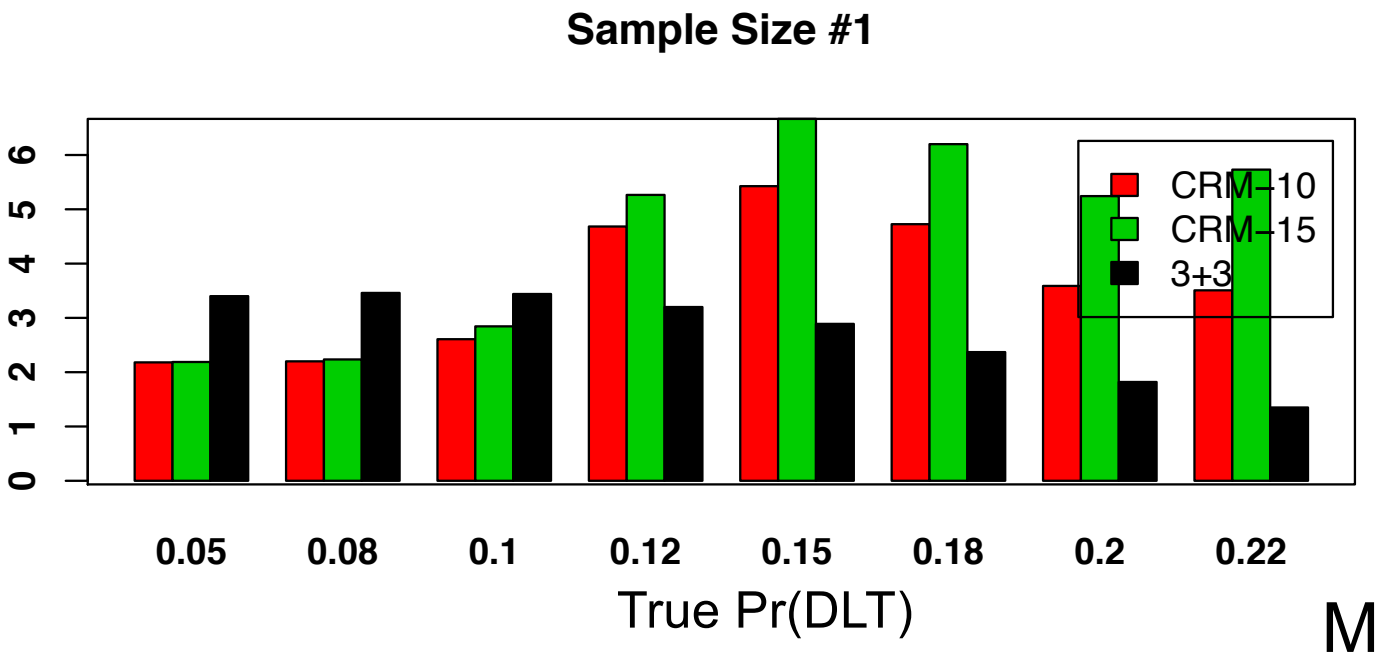
Allocate subject 22 to d=8







Stop Trial for n=10 at 8mg/kg dose

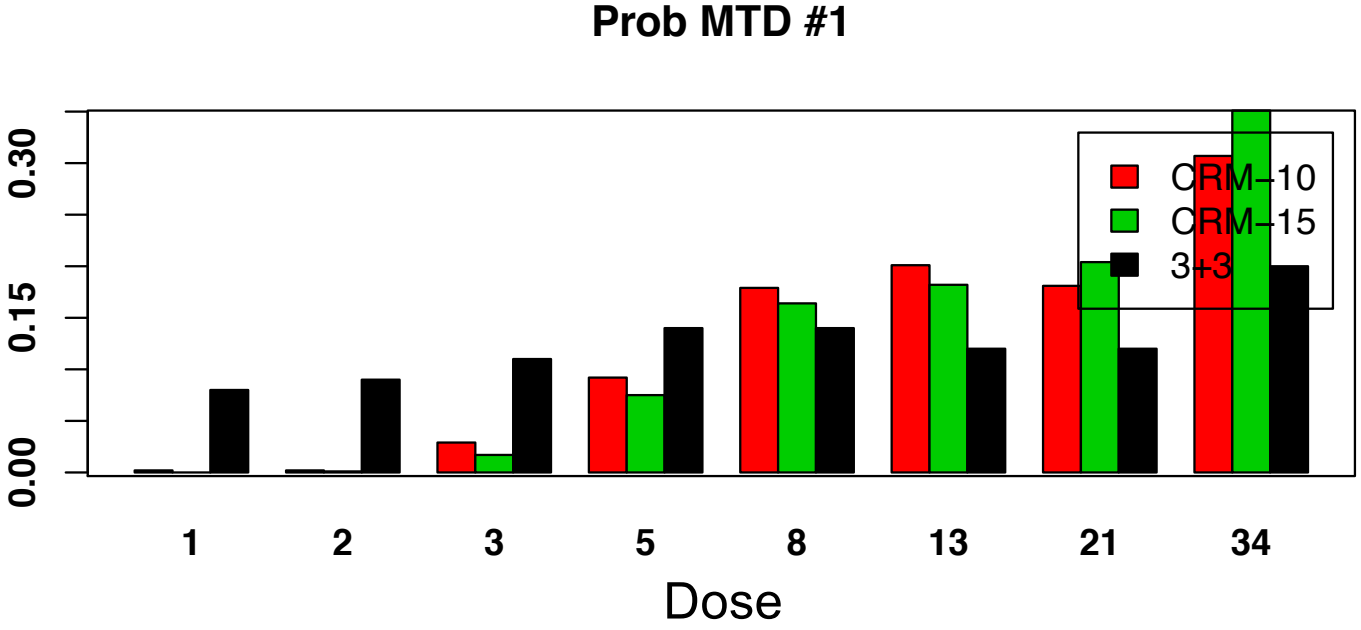


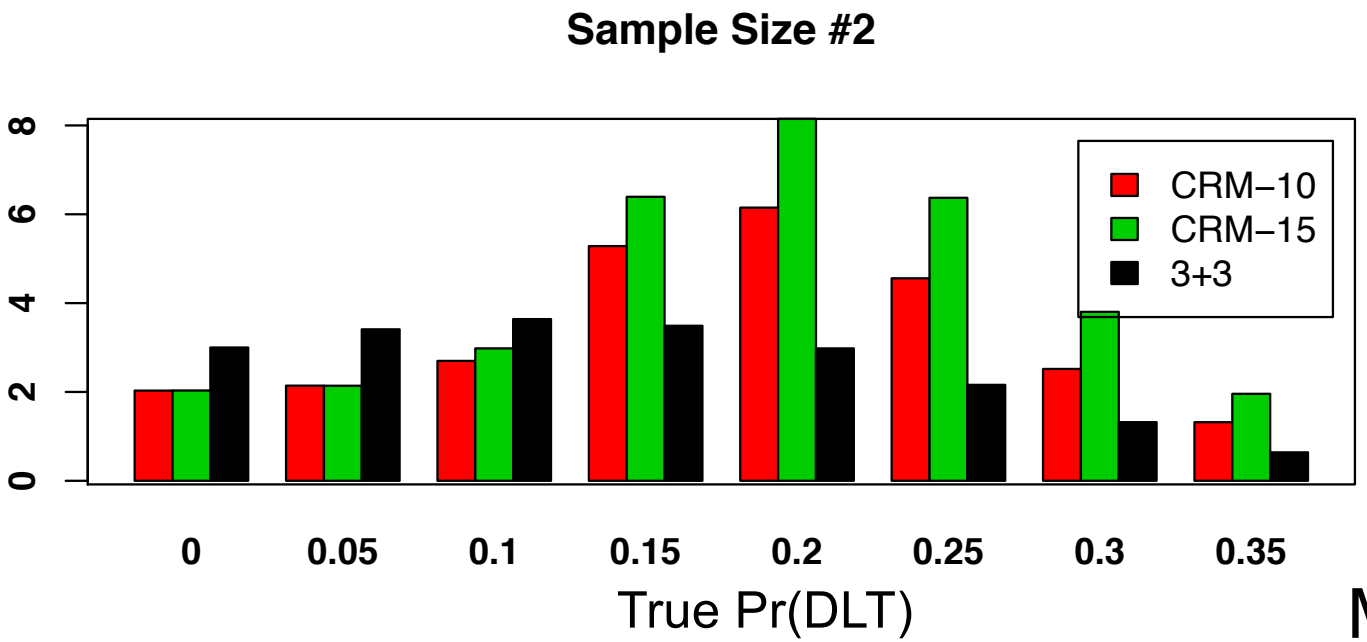
Mean Sample Sizes

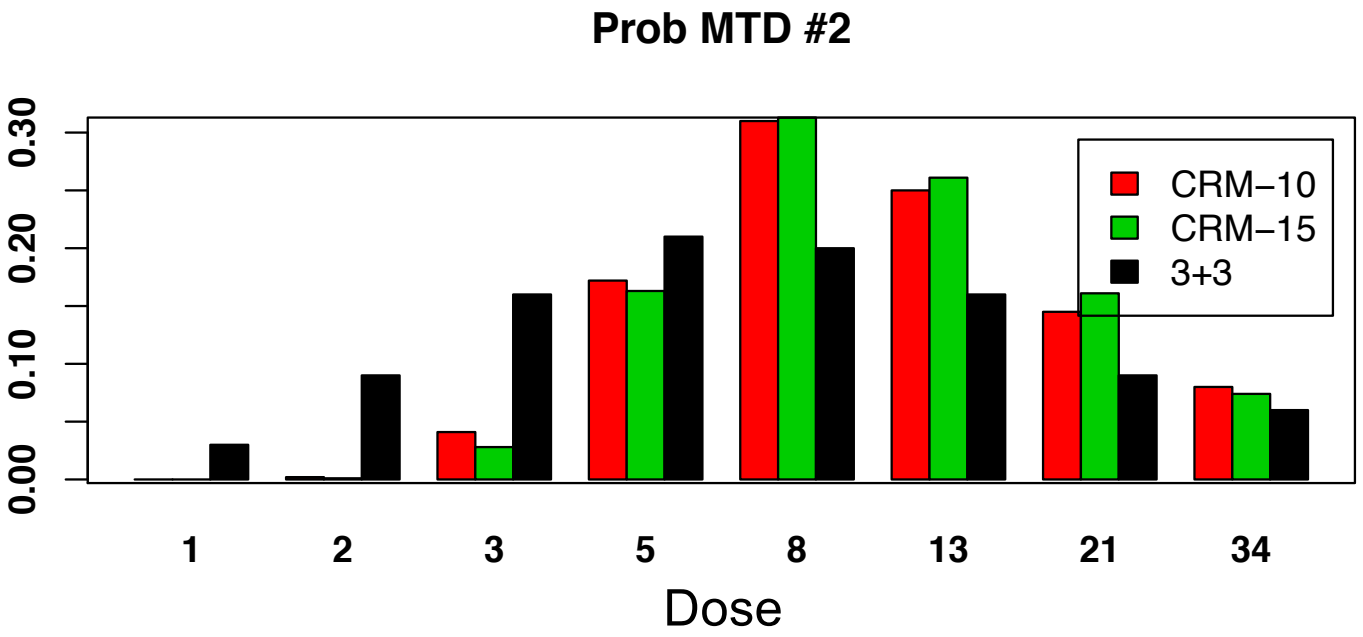
CRM-10: 28.9

CRM-15: 36.4

3+3: 21.9





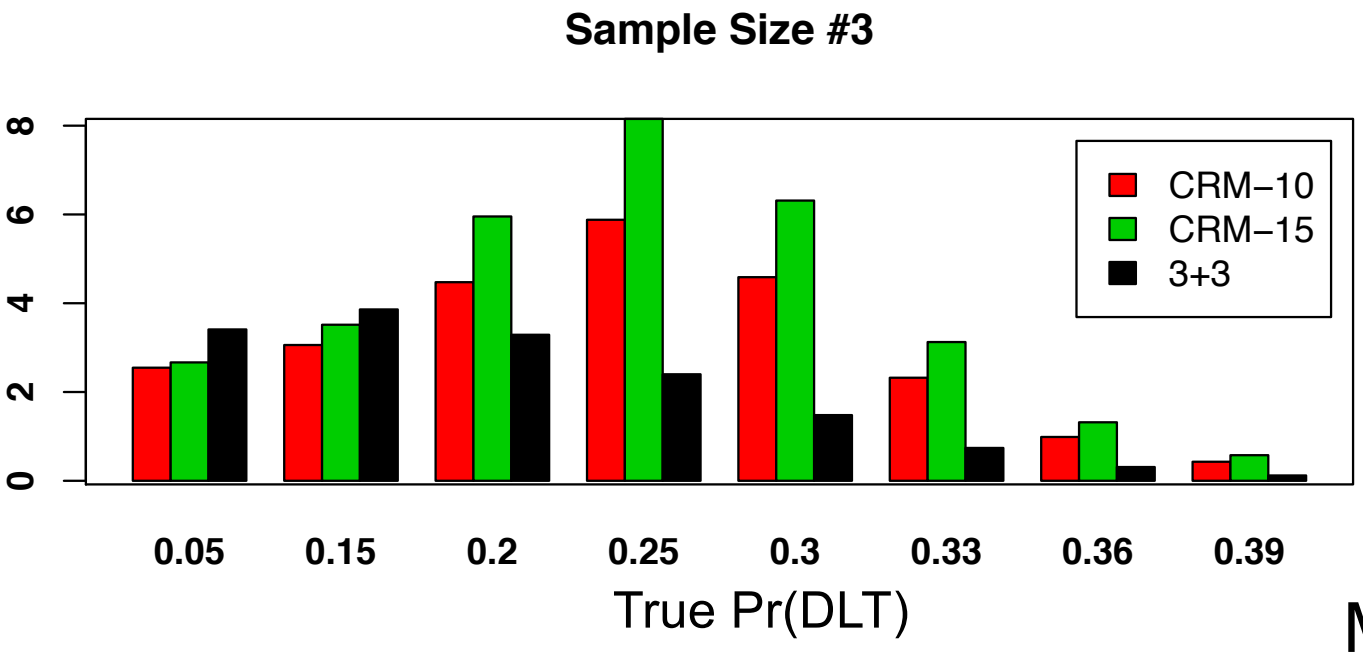


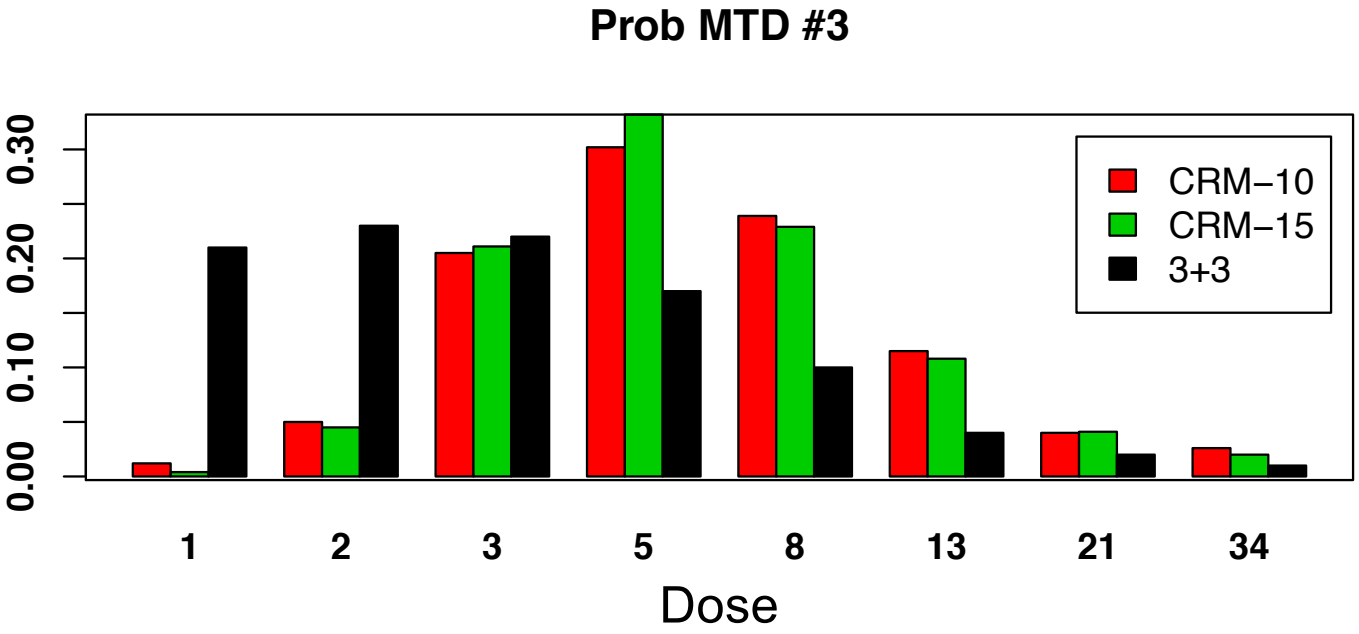
Mean Sample Sizes

CRM-10: 26.7

CRM-15: 33.8

3+3: 20.6



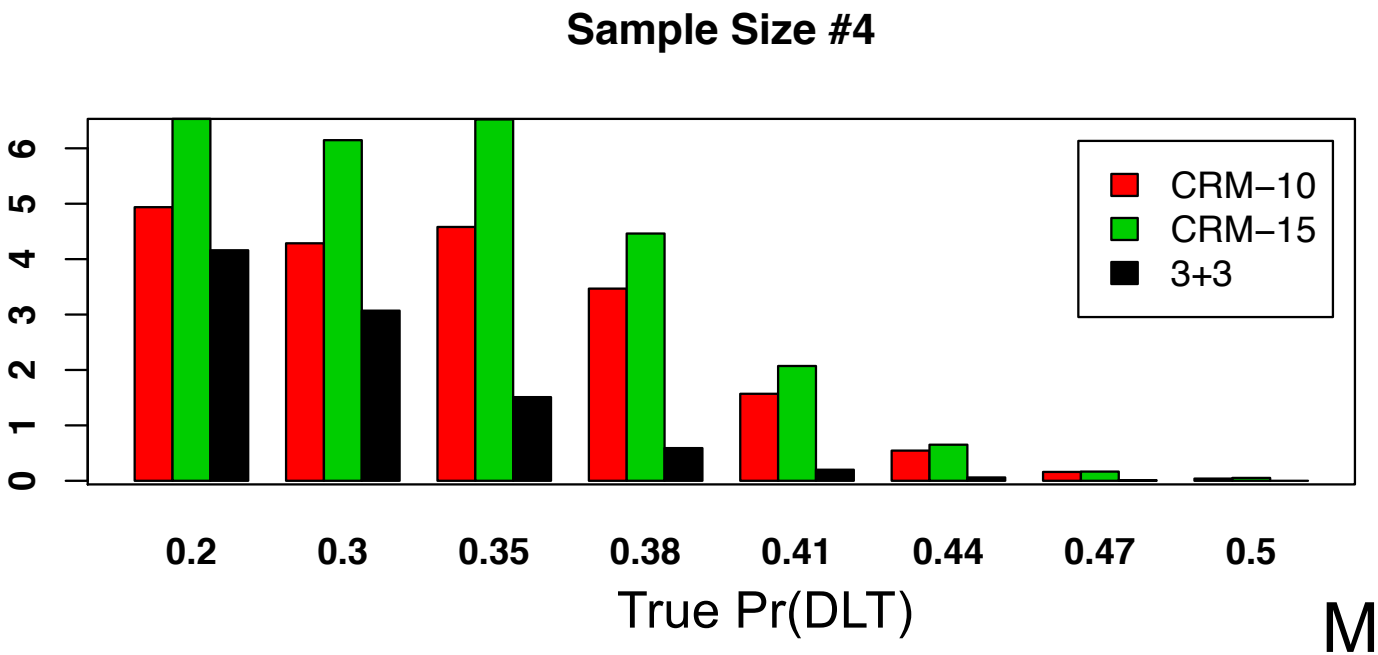


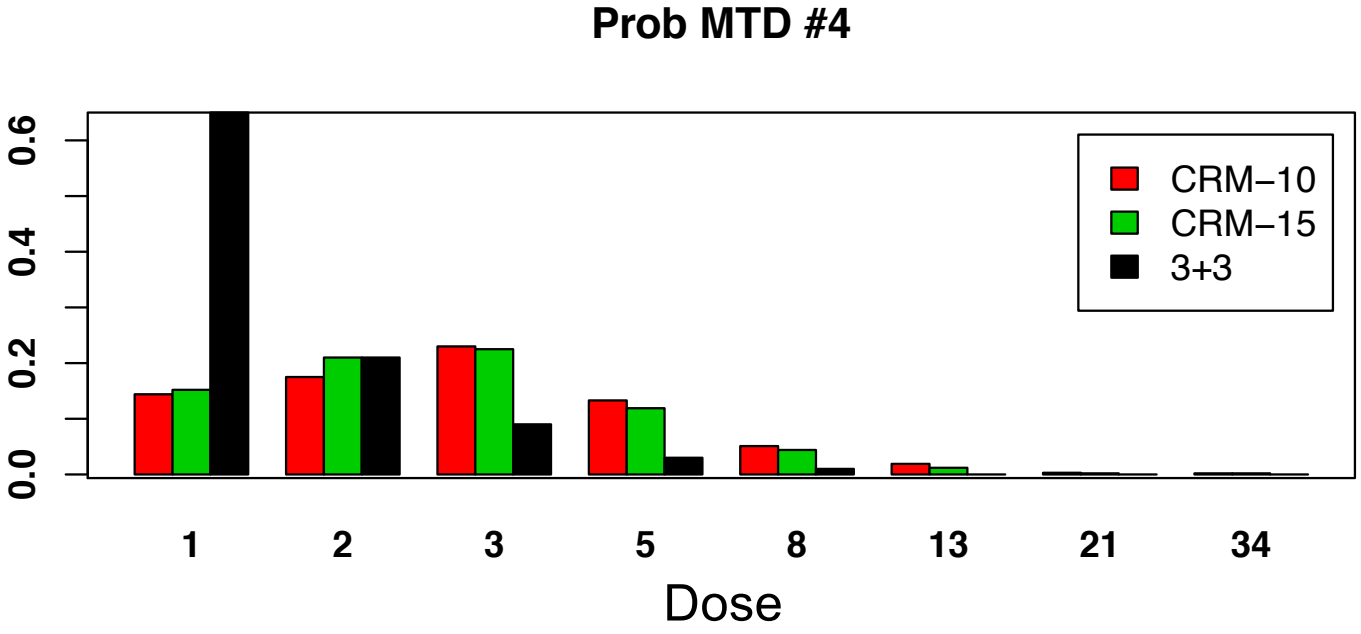
Mean Sample Sizes

CRM-10: 24.3

CRM-15: 31.6

3+3: 15.6



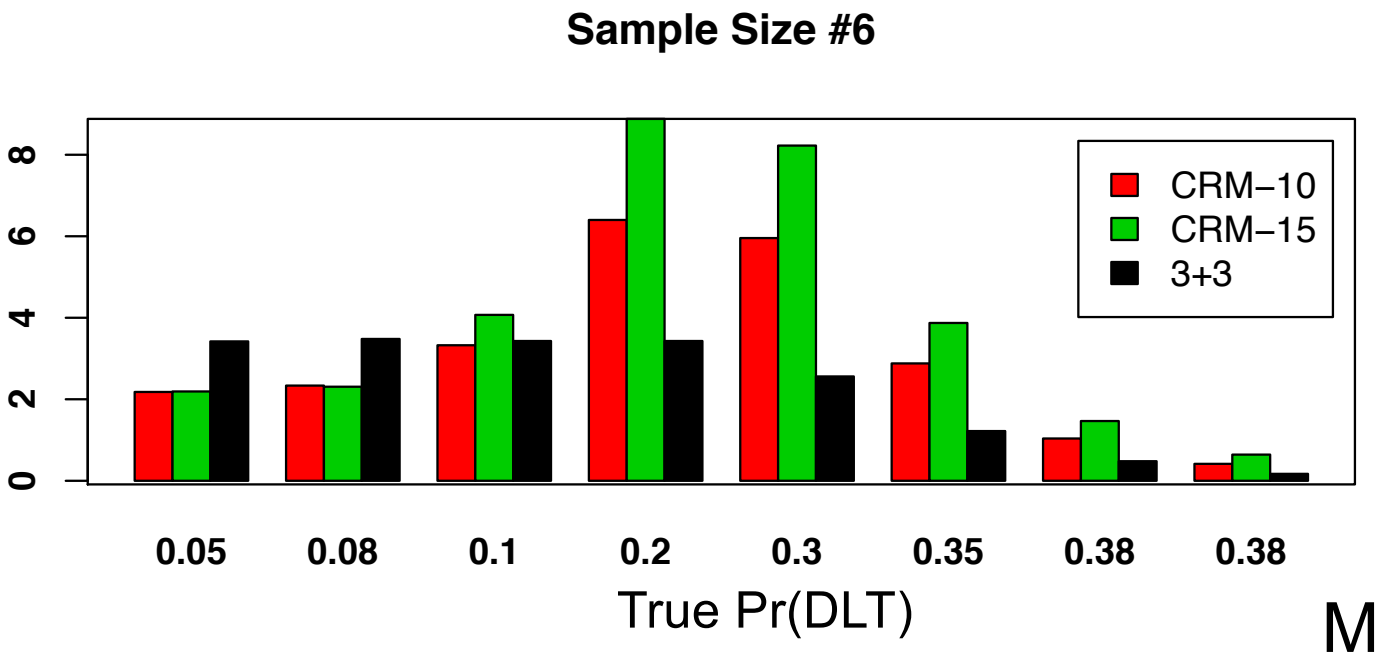


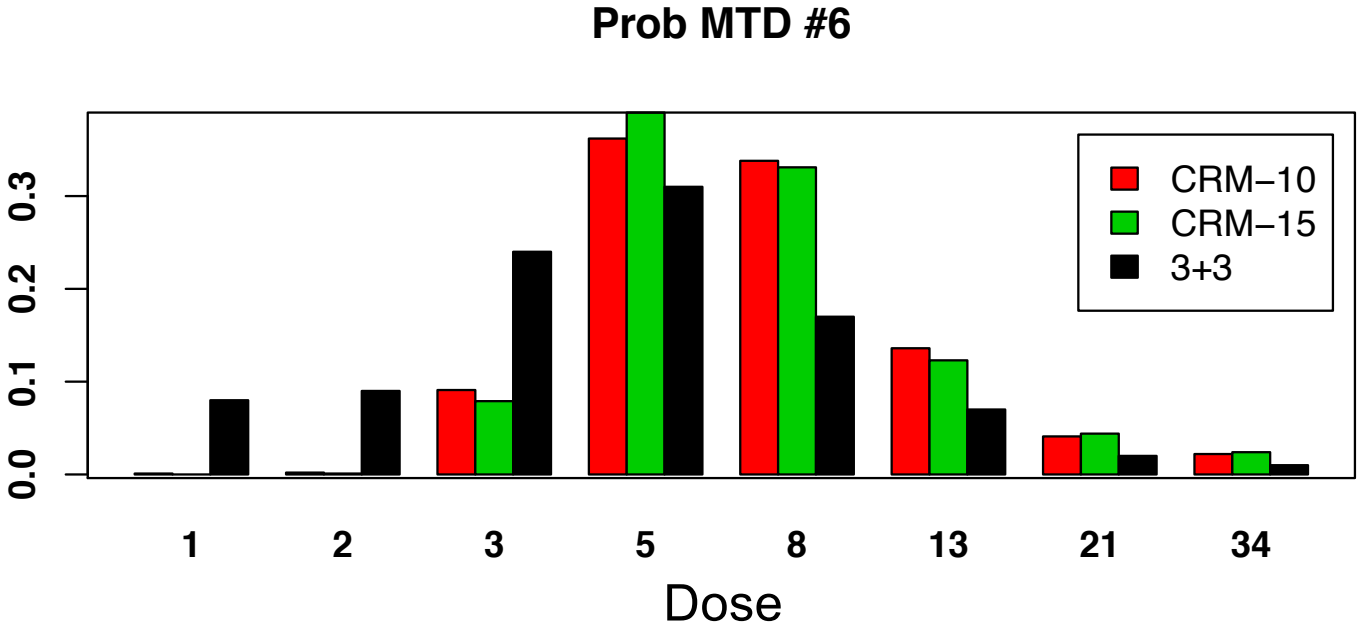
Mean Sample Sizes

CRM-10: 19.6

CRM-15: 26.6

3+3: 9.6



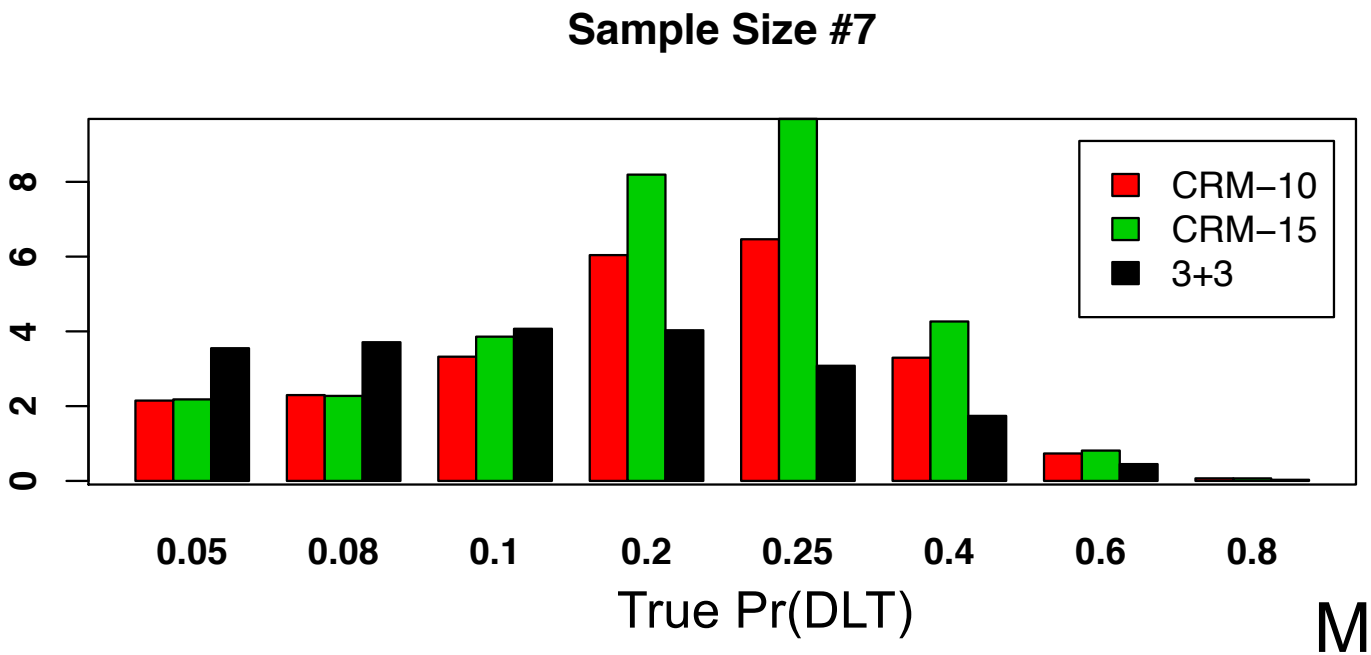


Mean Sample Sizes

CRM-10: 24.5

CRM-15: 31.7

3+3: 18.2

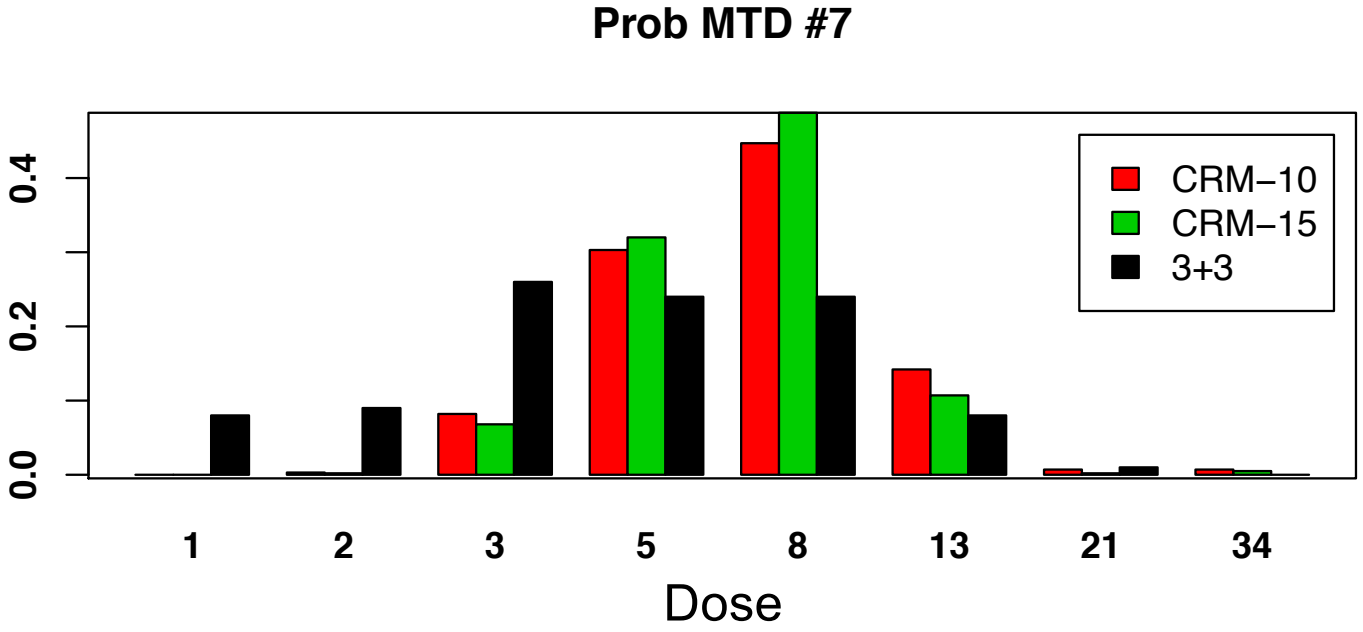


Mean Sample Sizes

CRM-10: 24.4

CRM-15: 31.3

3+3: 20.7



3+3 Pros and Cons

- Well accepted, easy to understand and implement
- Treats a higher proportion of patients at low, possibly ineffective, doses
- The probability of stopping at an incorrect dose level is higher than generally believed (Reiner, Paoletti, O'Quigley 1999).
- Tends to choose a dose below the true MTD
- In dose escalation uses information from only the most recent cohort and ignores data from previous cohorts
- There is no statistical estimation of the MTD

CRM Pros and Cons

- Requires more planning and infrastructure to execute
- Execution can be streamlined (same day); no delay between cohorts
- Could require more patients and take longer
- CRM allows complete flexibility: can target any DLT rate, flexible stopping rules, etc
- Tends to treat patients at doses close to the MTD
- In dose escalation incorporates available data from all cohorts borrowing information across all dose levels
- Provides a statistical estimate of the MTD, borrowing information from neighboring dose levels, and allows for uncertainty around this estimate

Previously Performed Extensions to the CRM

- Customization of rules for enrollment, dose escalation, stopping early for having sufficiently characterized the MTD
- Ordinal (rather than dichotomous) outcomes
- Two dimensional dose finding for combinations
- Randomized dose-escalation between two formulations – compare rate of DLTs between control and new formulation
- Dose escalate simultaneously in multiple populations and/or schedules – share information between groups

Previously Performed Extensions to the CRM

- Seamless phase I/II: escalate to the MTD and continue to a phase II portion
  - Bring select doses forward for randomized phase II
  - Bring all doses forward, continue to refine the MTD while determining optimal phase II dose based on both toxicity and efficacy
- "Backfill" of lower dose levels: enroll additional patients at lower doses during escalation to establish dose-toxicity and dose-response curves

Example #2: Combo CRM

With Scott Berry

Goal

- Experimental agent and existing agent
- Find the MTD (maximum tolerated dose) for the single agent therapy and the combination therapy
- MTD is the largest dose that achieves a 33% or less rate of DLT (dose limiting toxicity)
- Certainly can be different doses

Modeling

- 42 doses of experimental
- Probability of DLT for single agent:

$$\log\left(\frac{\pi_S}{1-\pi_S}\right) = \alpha_1 + \beta d$$

- Probability of DLT for combination:

$$\log\left(\frac{\pi_C}{1-\pi_C}\right) = \alpha_1 + \alpha_2 + \beta d$$

Prior Distributions

$$\alpha_1 \sim N\left(-3, 0.5^2\right)$$

$$\alpha_2 \sim N\left(0.20, 0.15^2\right)$$

$$\beta \sim N\left(0.10, 0.05^2\right)$$

Design--Stage 1

- 3 Subjects assigned to dose 1
- Find *Estimated MTD* (Bayesian calculation from posterior)
- Assign 3 to estimated MTD, unless more than double highest dose given (2x max increase)
- 12 subjects (4 cohorts) in Stage 1

Design--Stage 2

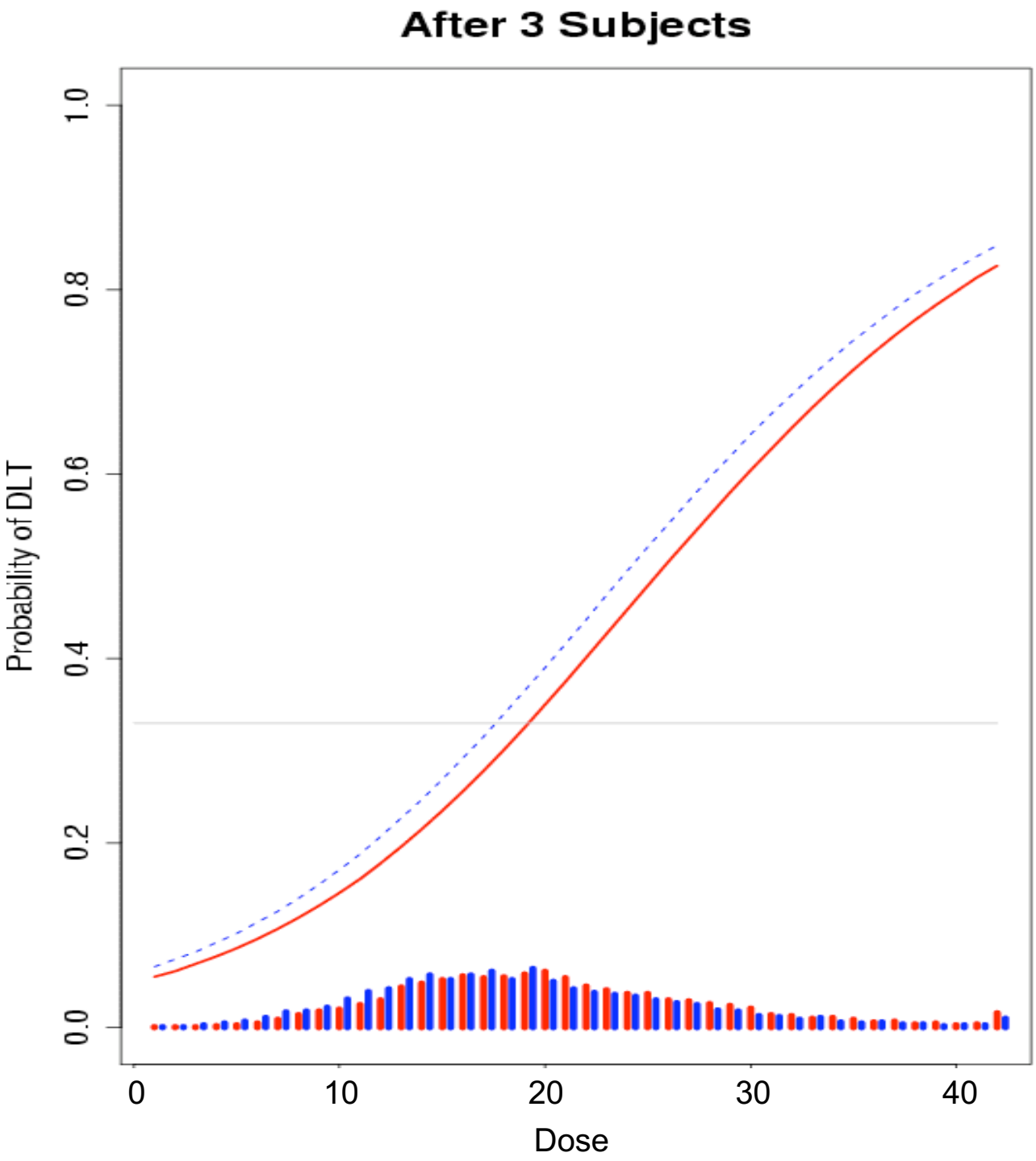
- Start combo at 50% of single agent MTD
- Rotate assignment to estimated MTD for combo and single agent
- No more than a doubling of the highest given value
- Update posterior each pair of subjects

Stopping Rules

- Futility: Drop single or combo if
  - $\Pr(\pi > 0.33 \mid d=1) > 0.90$
- Success: Stop single or combo if
  - 5 successive doses sum P(d=MTD)>0.90
  - With 42 doses no one dose will ever dominate
- Stop at the cap of 60
- If single or combo stops continue the other

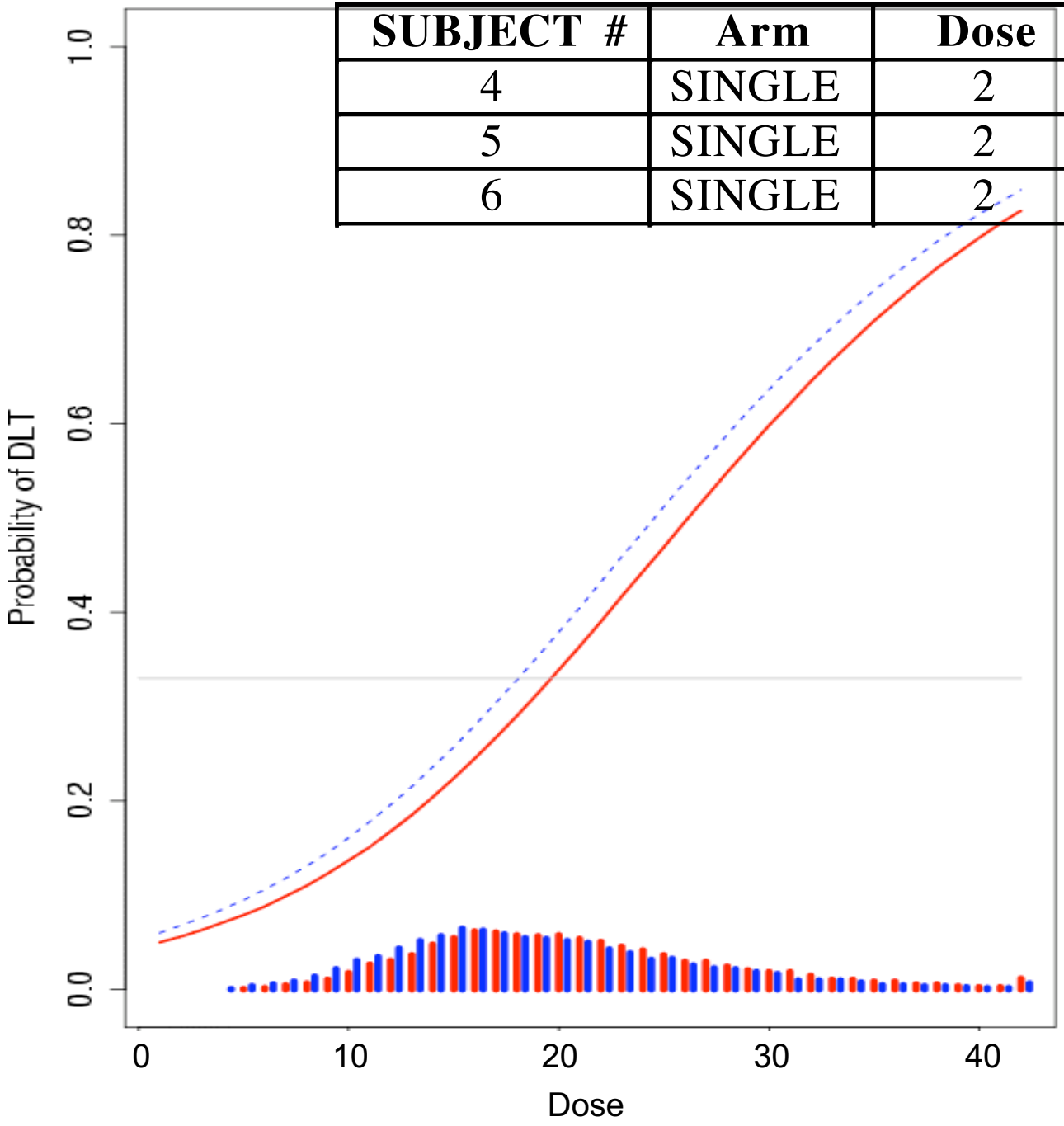
Sample Trial

SUBJECT #	Arm	Dose	DLT?
1	SINGLE	1	NO
2	SINGLE	1	NO
3	SINGLE	1	NO


After 3 Subjects
Probability of DLT
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Dose

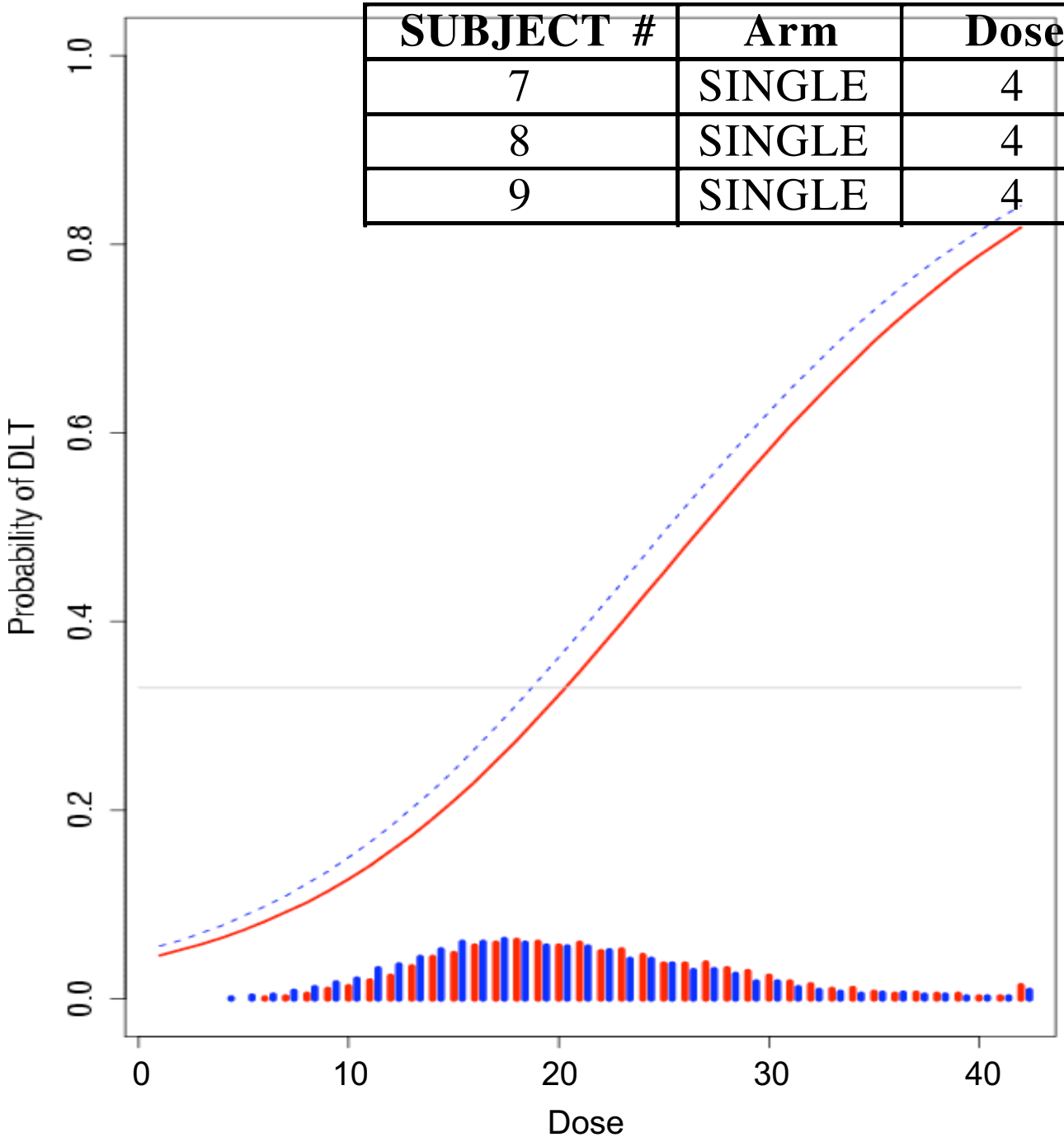
After 6 Subjects

SUBJECT #	Arm	Dose	DLT?
4	SINGLE	2	NO
5	SINGLE	2	NO
6	SINGLE	2	NO



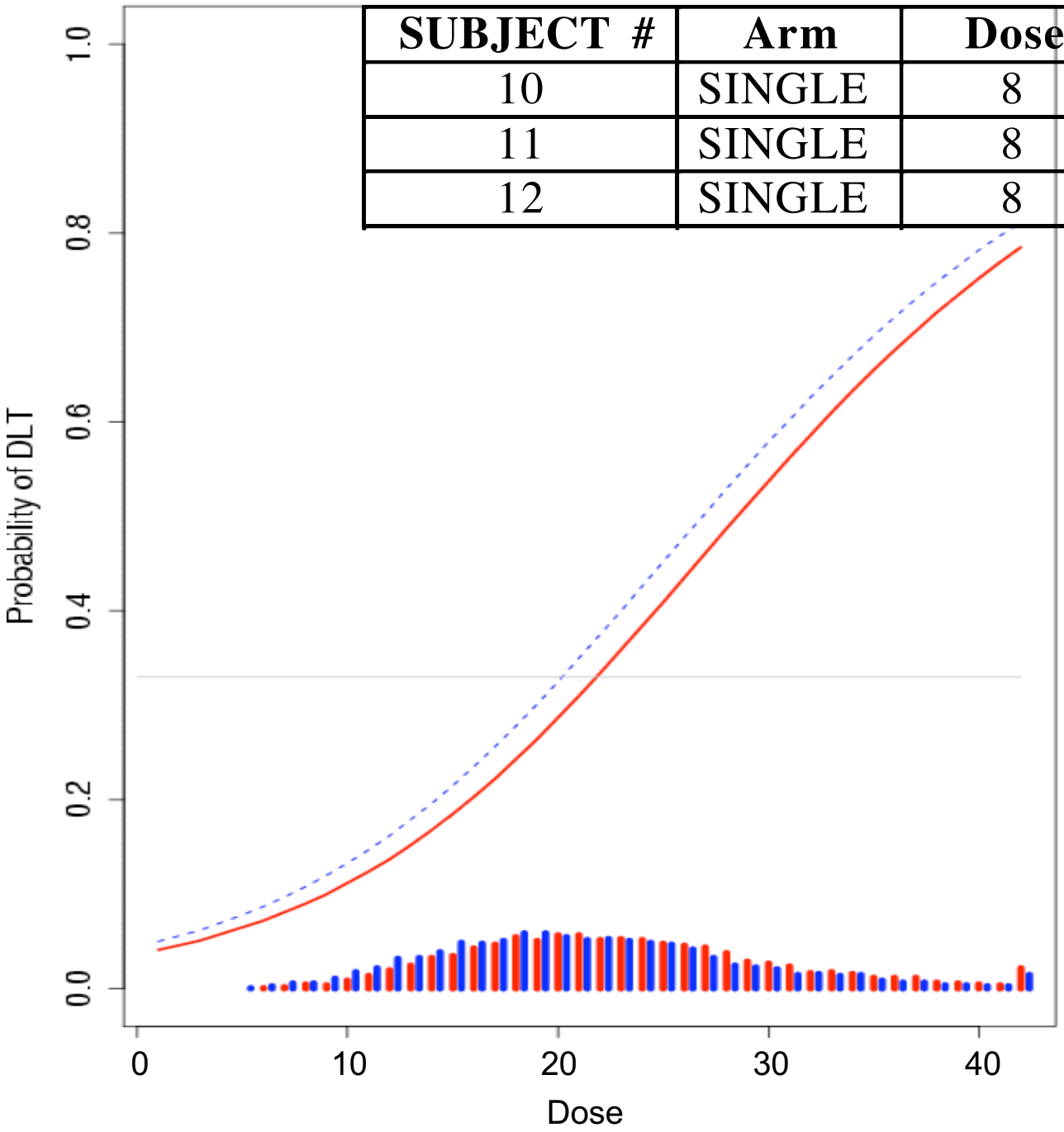
After 9 Subjects

SUBJECT #	Arm	Dose	DLT?
7	SINGLE	4	NO
8	SINGLE	4	NO
9	SINGLE	4	NO



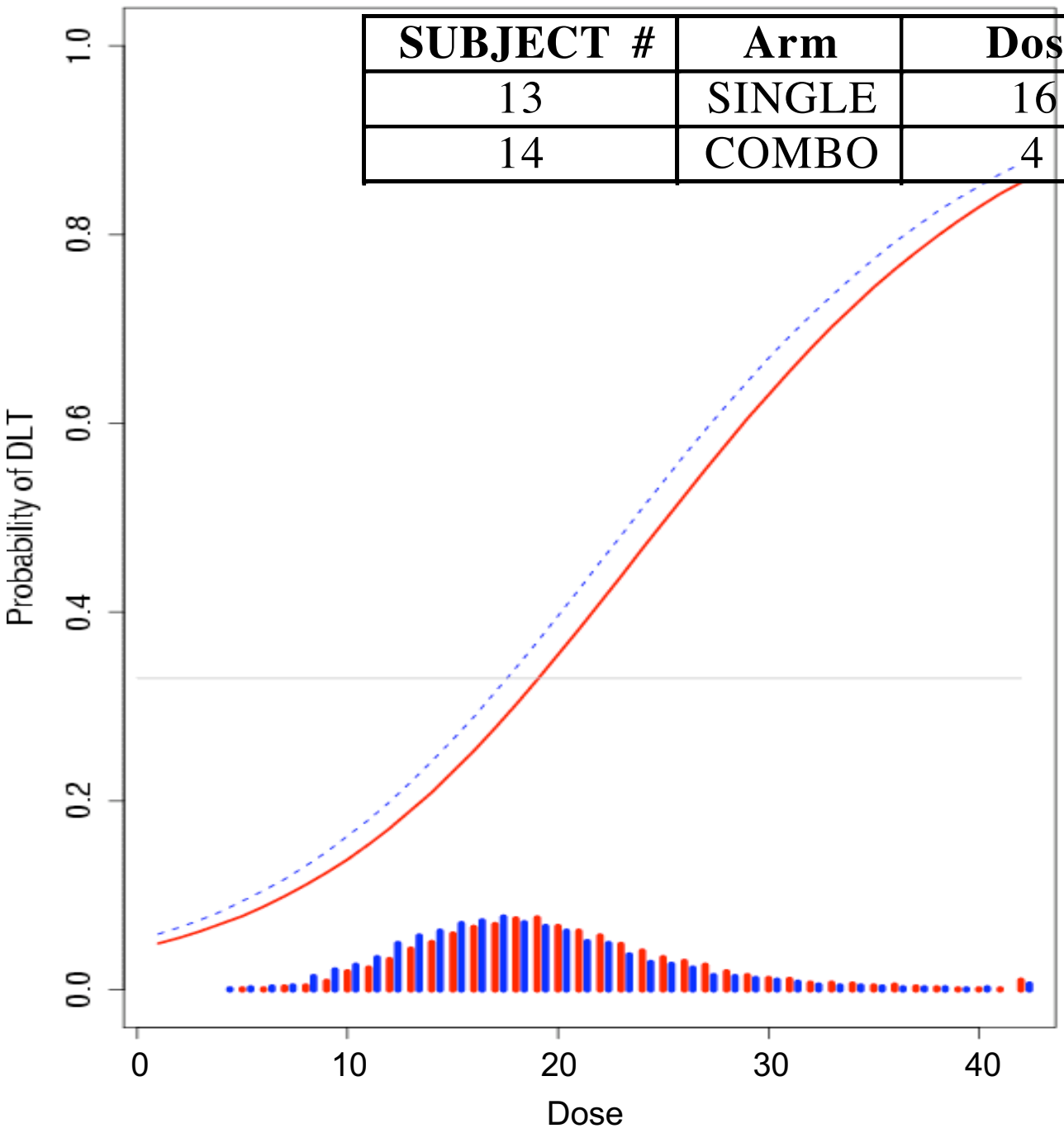
After 12 Subjects

SUBJECT #	Arm	Dose	DLT?
10	SINGLE	8	NO
11	SINGLE	8	NO
12	SINGLE	8	NO



After 14 Subjects

SUBJECT #	Arm	Dose	DLT?
13	SINGLE	16	YES
14	COMBO	4	NO



SUBJECT #	Arm	Dose	DLT?
15	SINGLE	19	YES
16	COMBO	8	NO
17	SINGLE	16	YES
18	COMBO	15	NO
19	SINGLE	15	NO
20	COMBO	14	YES

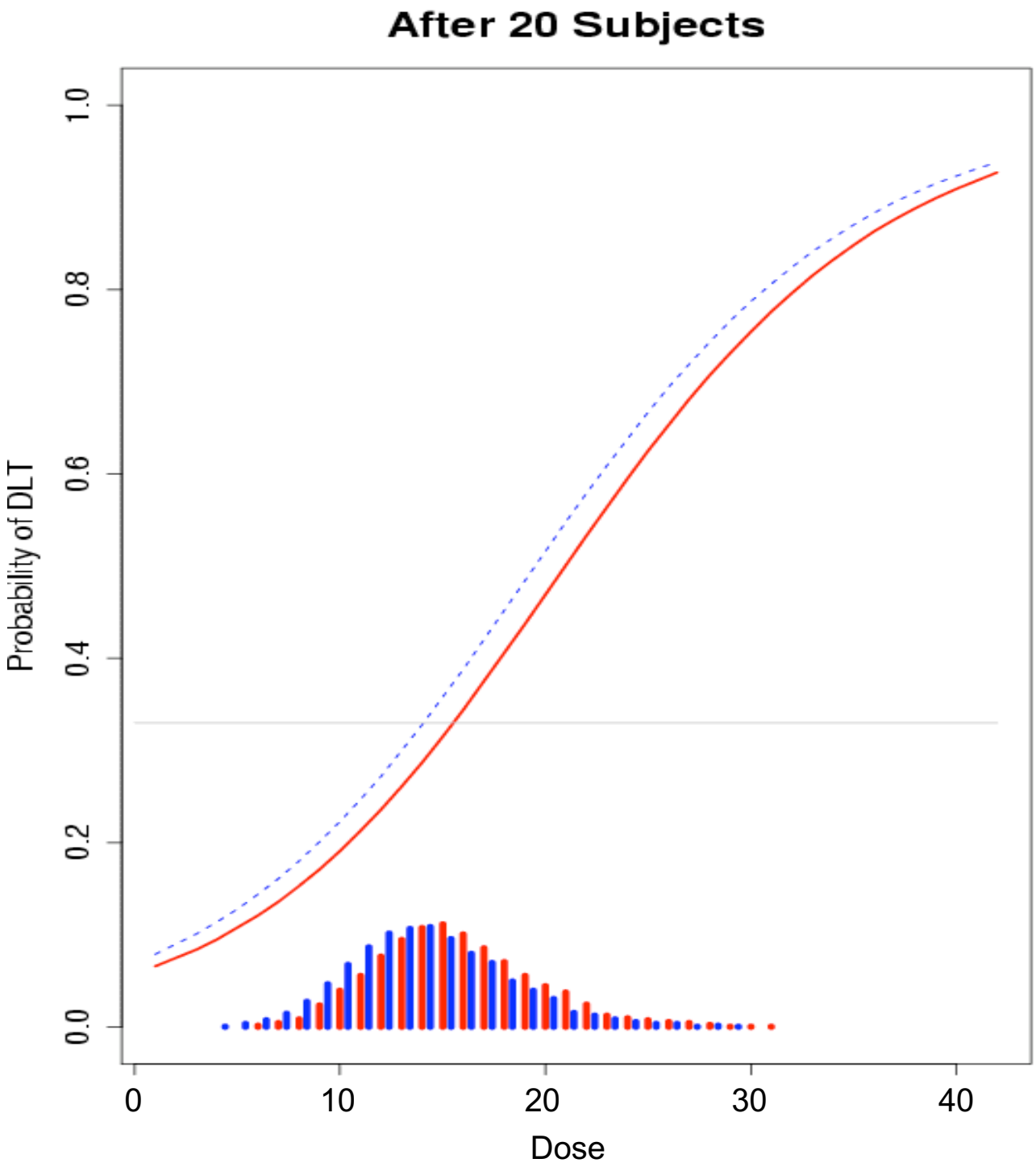
The probability that the MTD for SINGLE is dose 13 or 14 or 15 or 16 or 17 is

$0.095 + 0.108 + 0.112 + 0.101 + 0.086 = 0.502.$

This is the highest set of five adjacent doses.

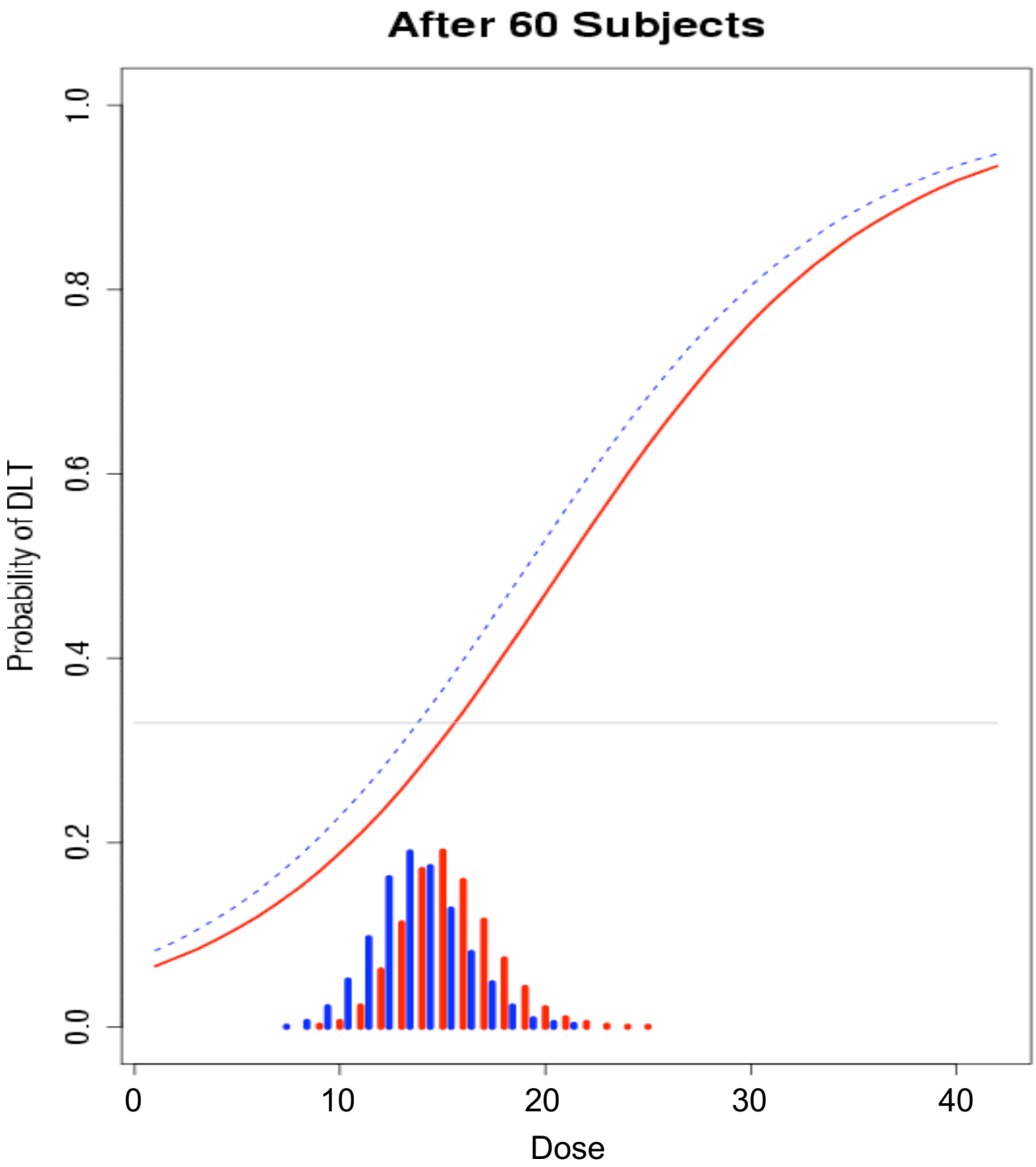
$0.502 < 0.90$: trial does not stop for successfully knowing the MTD.

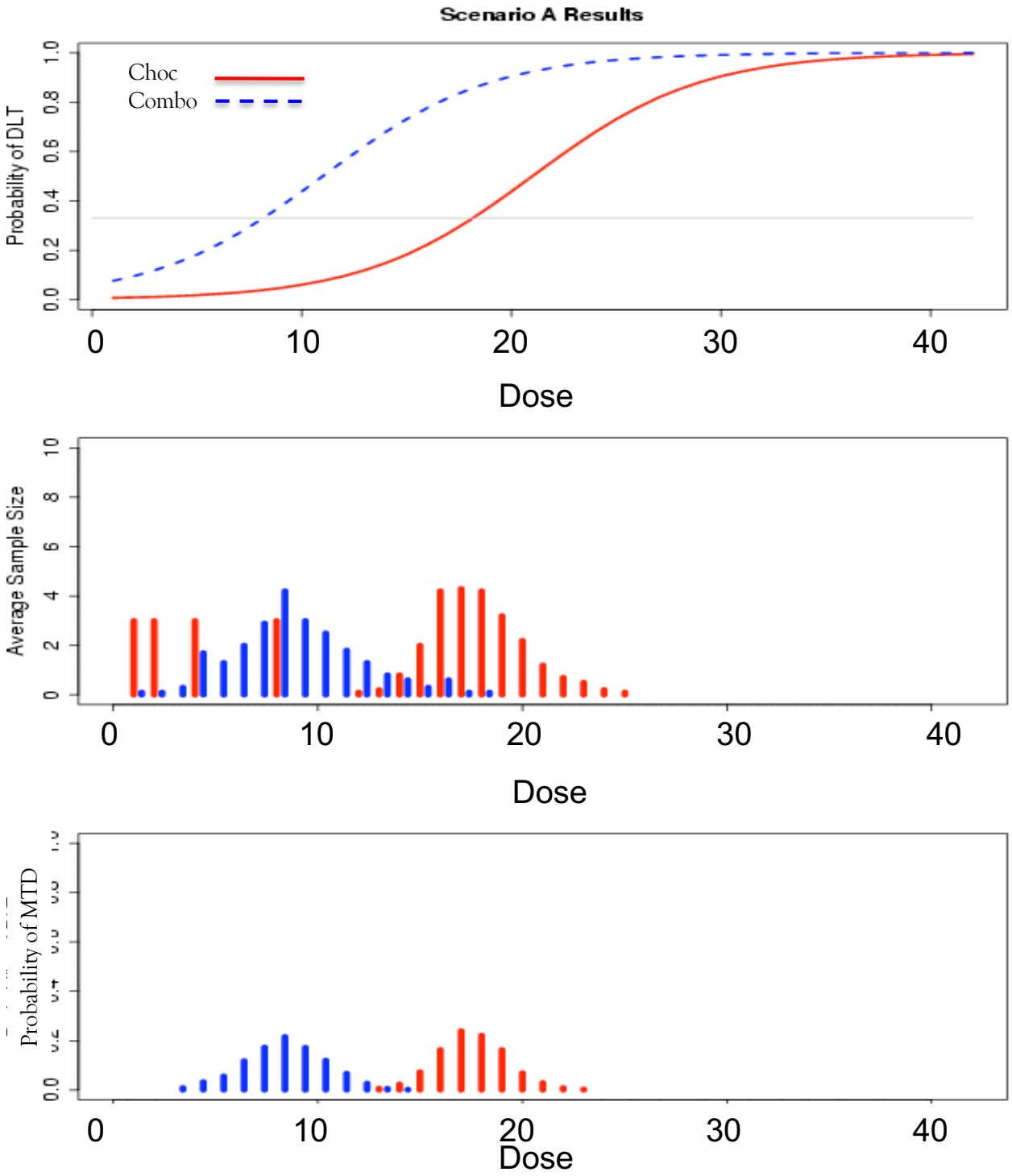
Likewise the COMBO arm does not stop for success.

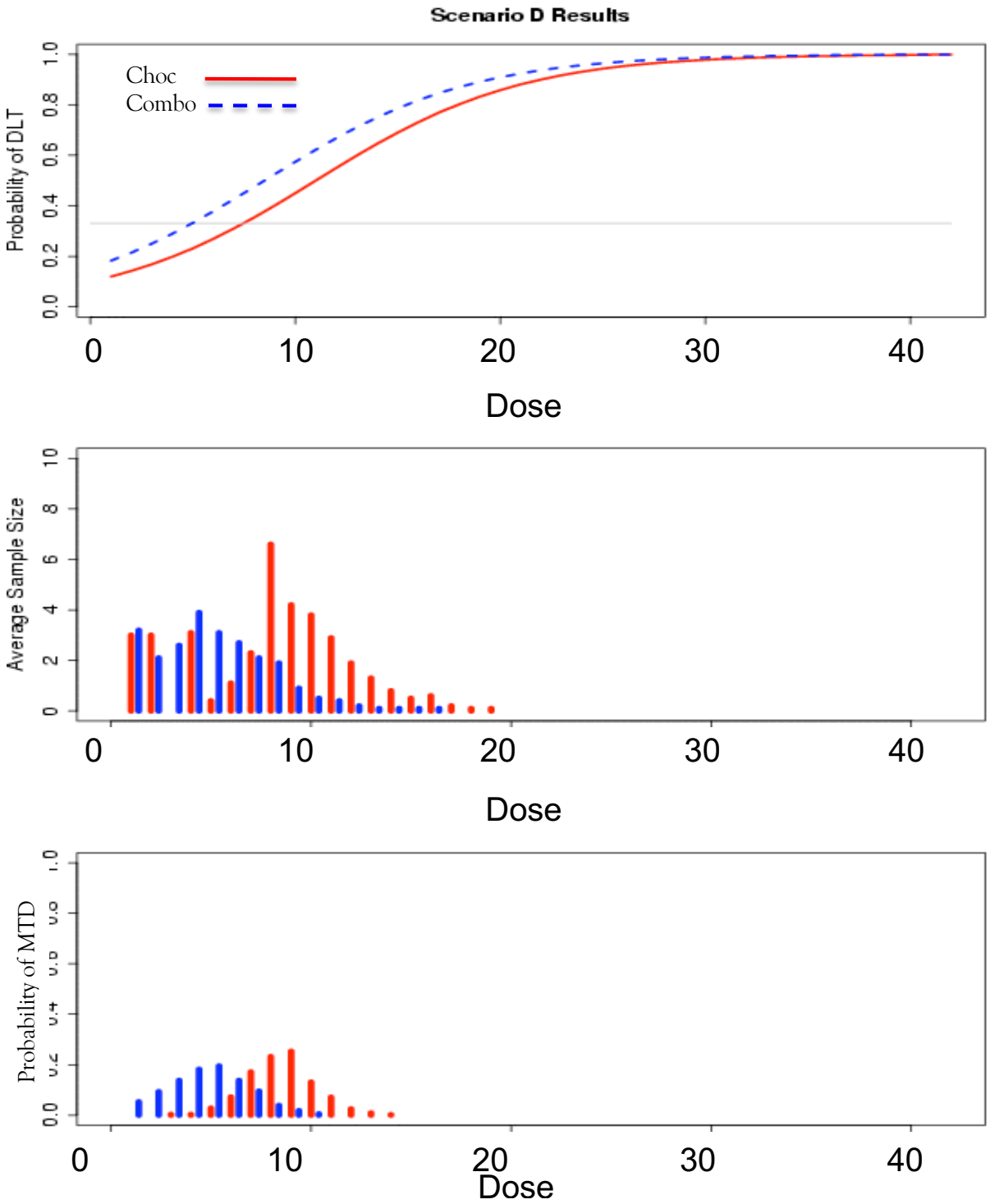

After 20 Subjects
Probability of DLT
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Dose

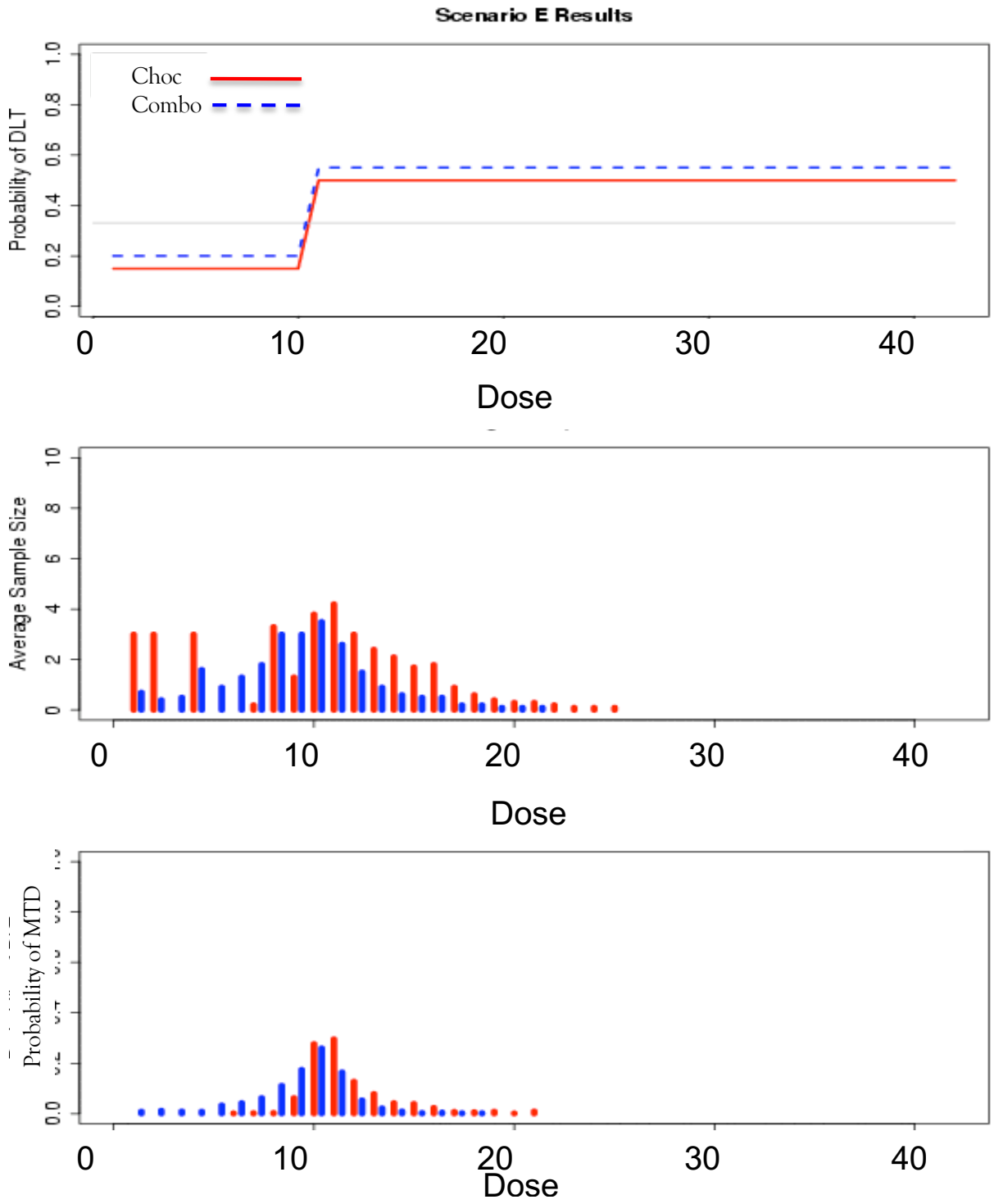
SUBJECT #	Arm	Dose	DLT?
21	SINGLE	15	NO
22	COMBO	14	NO
23	SINGLE	16	NO
24	COMBO	15	YES
25	SINGLE	16	NO
26	COMBO	14	YES
27	SINGLE	15	NO
28	COMBO	14	NO
29	SINGLE	16	YES
30	COMBO	14	NO
31	SINGLE	15	NO
32	COMBO	14	YES
33	SINGLE	15	NO
34	COMBO	13	YES
35	SINGLE	15	NO
36	COMBO	13	NO
37	SINGLE	15	NO
38	COMBO	14	NO
39	SINGLE	16	NO
40	COMBO	14	YES

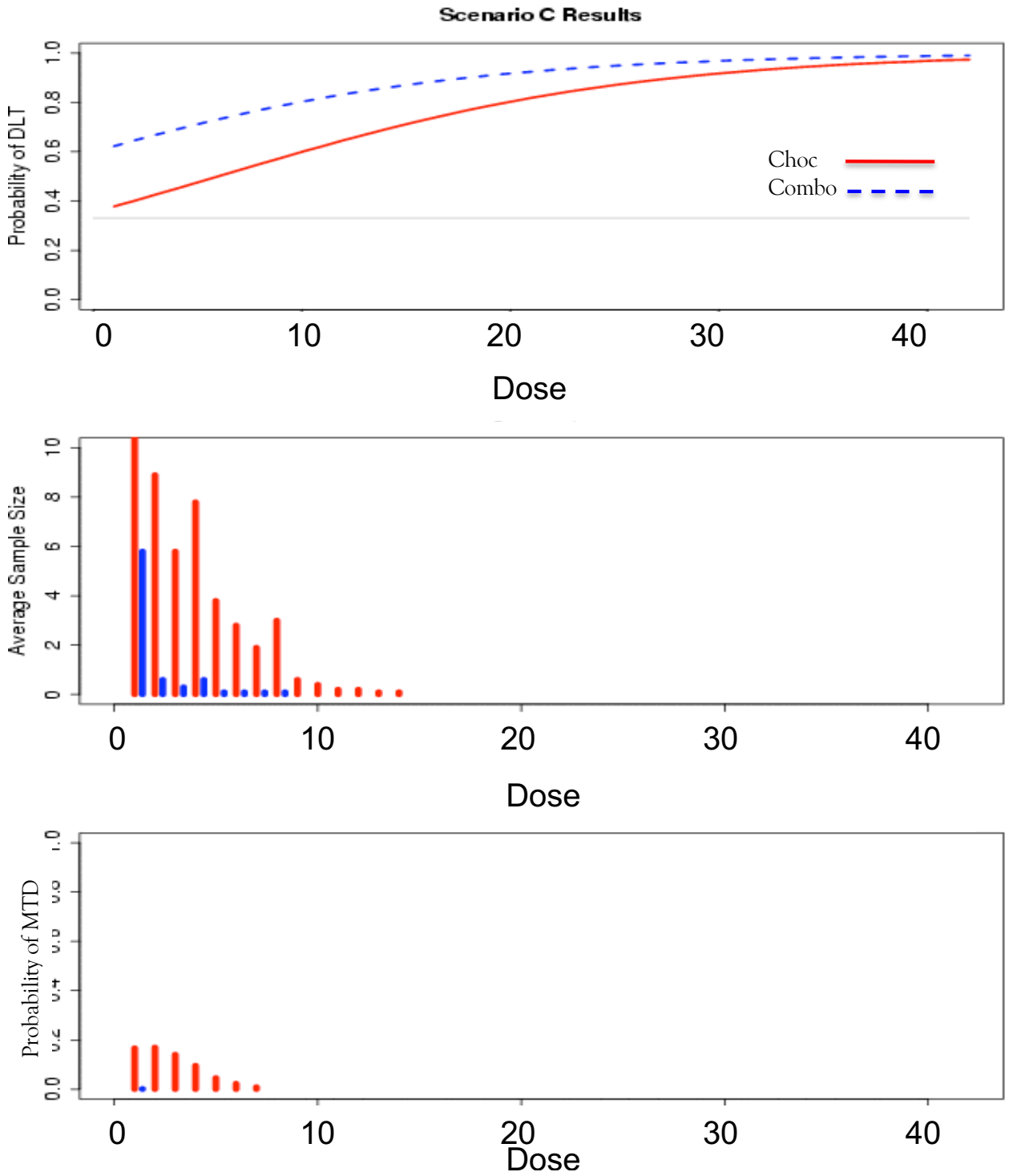
SUBJECT #	Arm	Dose	DLT?
41	SINGLE	16	NO
42	COMBO	14	YES
43	SINGLE	15	NO
44	COMBO	14	YES
45	SINGLE	15	NO
46	COMBO	13	NO
47	SINGLE	15	NO
48	COMBO	14	YES
49	SINGLE	15	NO
50	COMBO	14	YES
51	SINGLE	15	YES
52	COMBO	13	NO
53	SINGLE	15	NO
54	COMBO	13	NO
55	SINGLE	15	NO
56	COMBO	13	YES
57	SINGLE	15	NO
58	COMBO	13	NO
59	SINGLE	15	NO
60	COMBO	13	YES


After 60 Subjects
Probability of DLT
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Dose


Scenario A Results
Choc
Combo
Probability of DLT
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Dose
Average Sample Size
0
2
4
6
8
10
0
10
20
30
40
Dose
Probability of MTD
0.0
0
10
20
30
40
Dose


Scenario D Results
Choc
Combo
Probability of DLT
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Dose
Average Sample Size
0
2
4
6
8
10
0
10
20
30
40
Dose
Probability of MTD
0.0
0
10
20
30
40
Dose


Scenario E Results
Choc
Combo
Probability of DLT
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Dose
Average Sample Size
0
2
4
6
8
10
0
10
20
30
40
Dose
Probability of MTD
0.0
0
10
20
30
40
Dose


Scenario C Results
Probability of DLT
0.0
0.2
0.4
0.6
0.8
1.0
Choc
Combo
0
10
20
30
40
Dose
Average Sample Size
0
2
4
6
8
10
0
10
20
30
40
Dose
Probability of MTD
0.0
0
10
20
30
40
Dose

Example #5:

Indication Finder or Basket Trials

With Kert Viele

Phase II Trial

- Tumor Type #1, Historic response rate of 15%
- What conclusion?

Tumor Type	N	Response	%
#1	40	8	20

- $\pi \sim \text{Beta}(1,1)$; $\Pr(\pi > 15\%) = 0.848$

Phase II Trial: Additional Tumor Types

- Tumor Type #1, Historic response rate of 15%
- What conclusion?

Tumor Type	N	Response	%
#1	**40**	**8**	**20**
#2	40	3	7.5
#3	40	1	2.5
#4	40	4	10

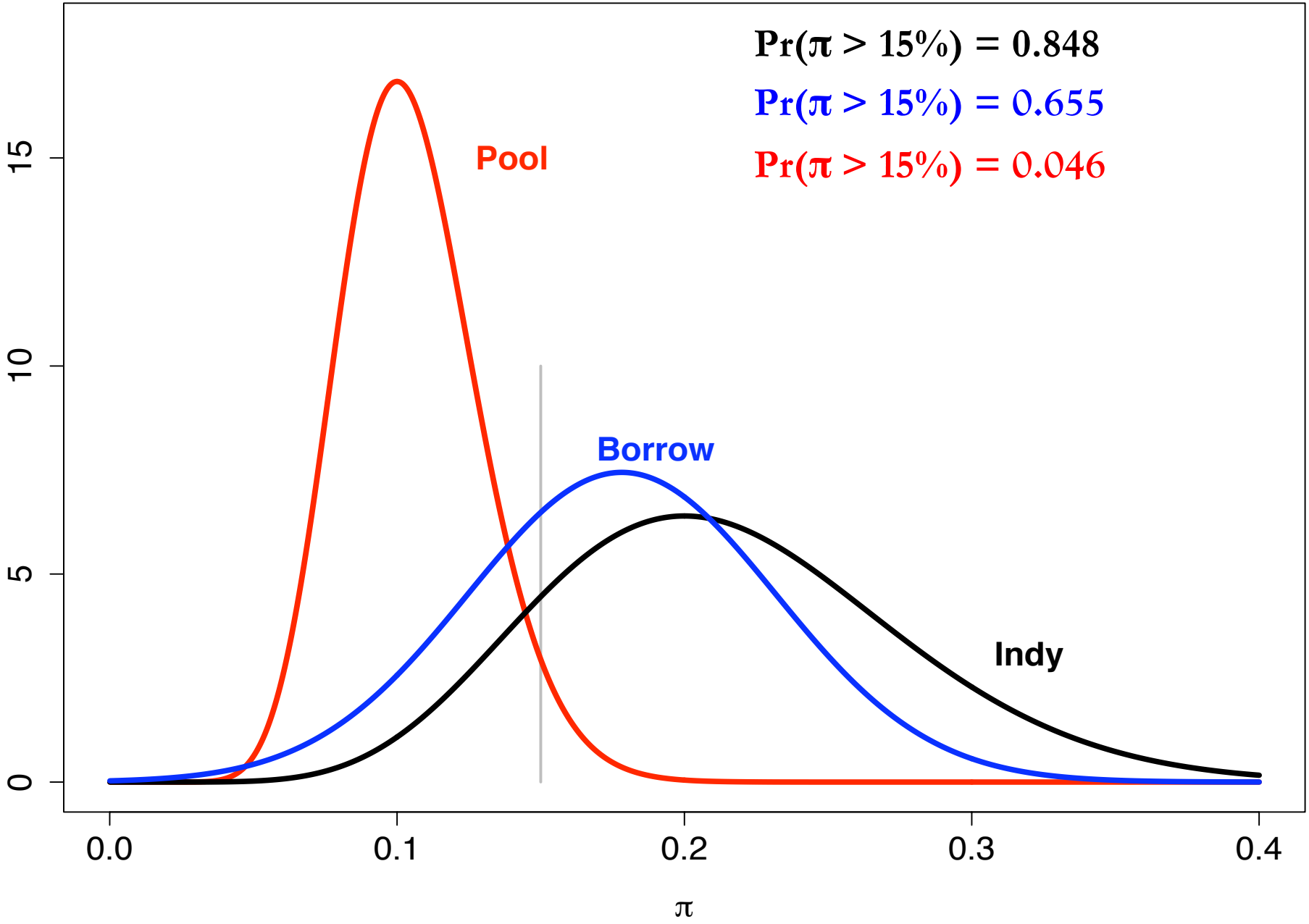
- $\pi \sim Beta(1,1)$; $Pr(\pi > 15\%) = 0.848$
- Pooled: $Pr(\pi > 15\%) = 0.046$

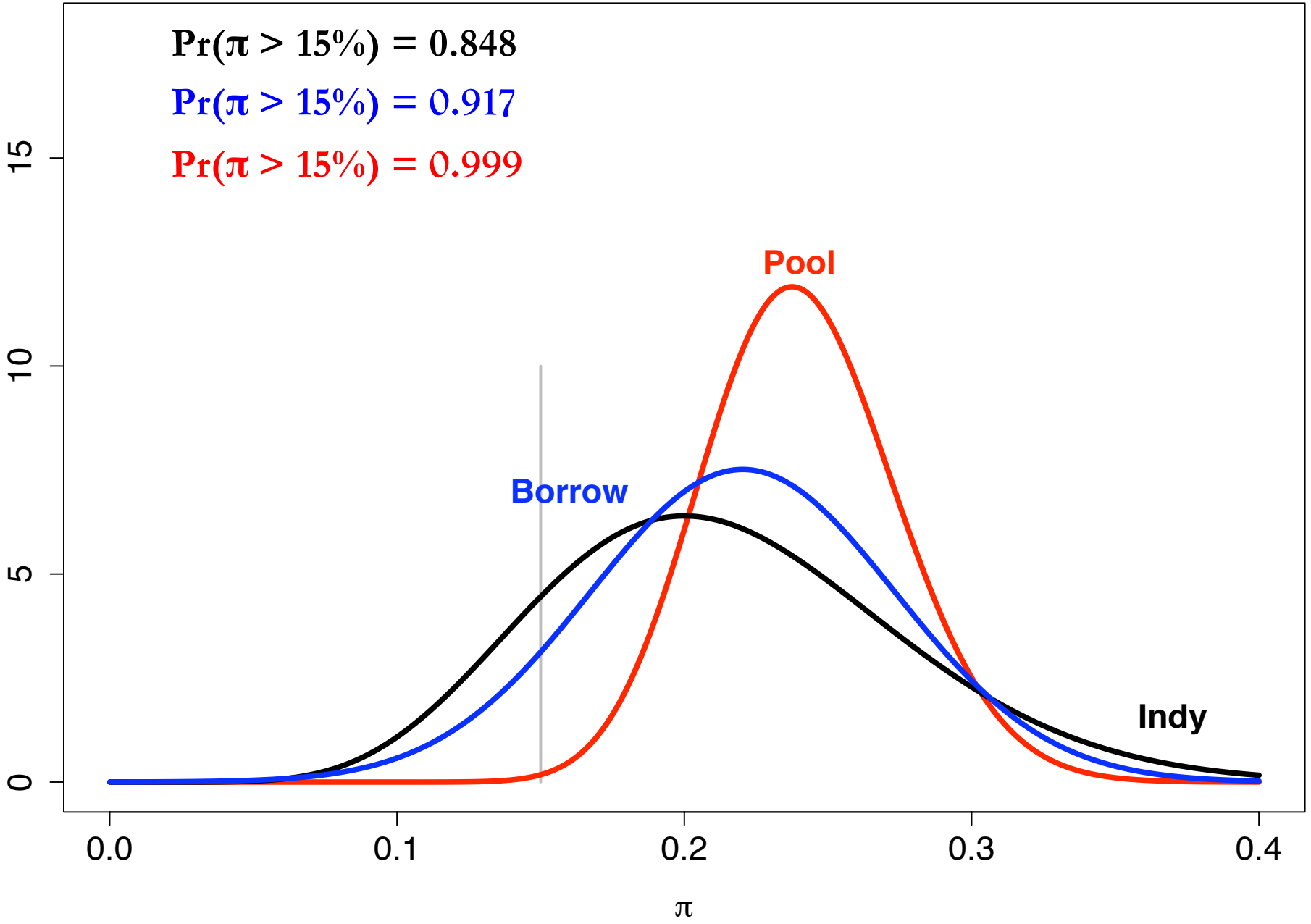
Phase II Trial: Additional Tumor Types

- Tumor Type #1, Historic response rate of 15%
- What conclusion?

Tumor Type	N	Response	%
#1	**40**	**8**	**20**
#2	40	10	25
#3	40	8	20
#4	40	12	30

- $\pi \sim Beta(1,1)$; $Pr(\pi > 15\%) = 0.848$
- Pooled: $Pr(\pi > 15\%) = 0.9986$


Pr(π > 15%) = 0.848
Pr(π > 15%) = 0.655
Pr(π > 15%) = 0.046
Pool
Borrow
Indy
0
5
10
15
0.0
0.1
0.2
0.3
0.4
π


Pr(π > 15%) = 0.848
Pr(π > 15%) = 0.917
Pr(π > 15%) = 0.999
Pool
Borrow
Indy
0
5
10
15
0.0
0.1
0.2
0.3
0.4
π

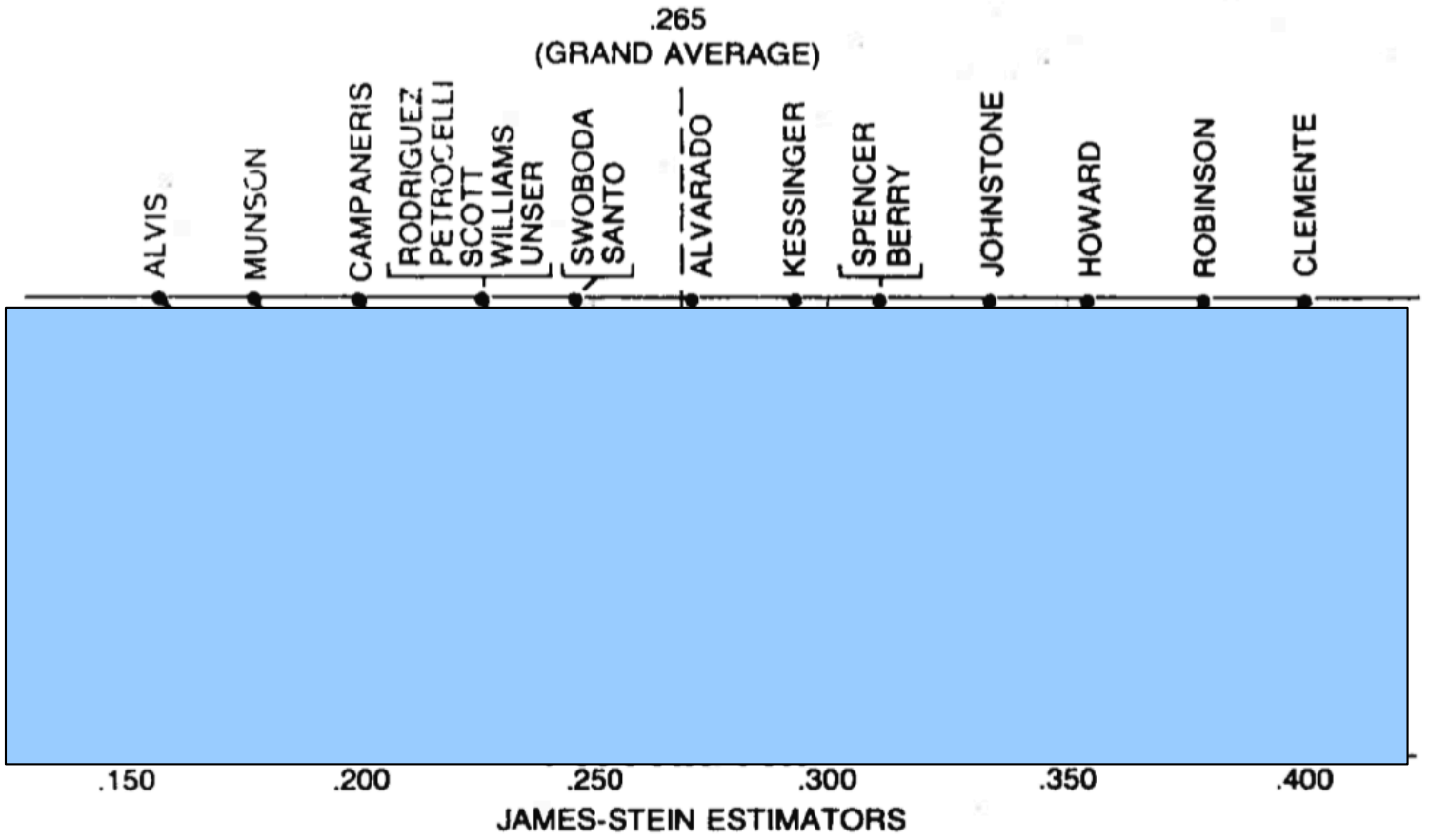
Hierarchical Modeling

- Shrinkage – the estimation of one "unit" borrows information from the other units
- Better estimation of individual units
  - Better than pooling
  - Better than ignoring
  - Think James-Stein estimator
- Realizing this is a better way to synthesize information, we use this prospectively in the design

More intuition

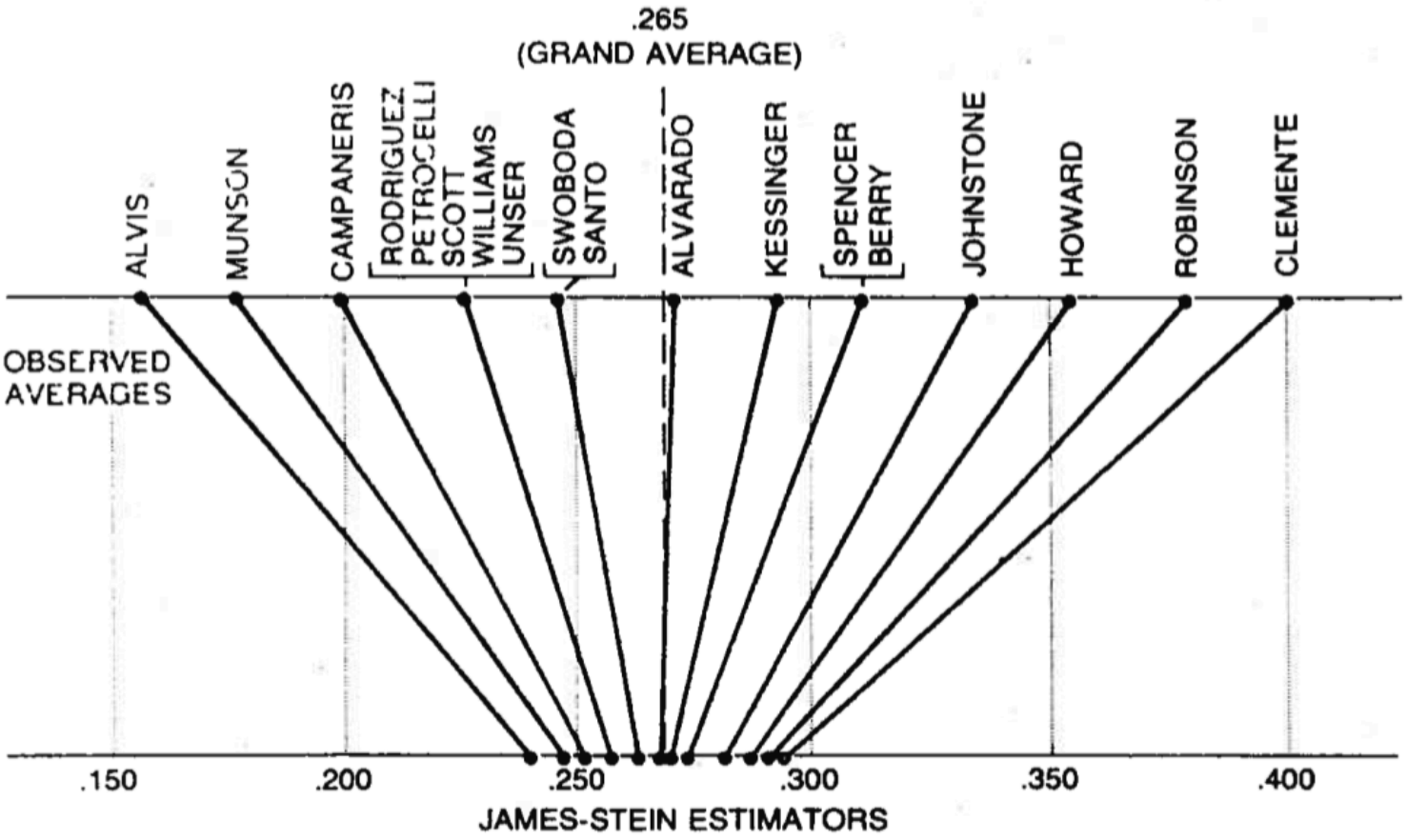
- Think of baseball batting averages in May
  - someone is almost always hitting near 0.400. This never lasts…no one is that good…but with multiple good players, someone is typically good AND lucky.
  - some otherwise very good players often have slumps.
- Good estimates separate the luck (random variation) from the skill
  - the highest averages are biased upward, lower them
  - the lowest averages are biased downward, raise them
- Pulling the observed values together makes better estimates.
  - don't pool! bring the highest down a little bit, the lowest up a little bit. But the estimates are still different.
- Efron & Morris, JASA, 1975

James-Stein Estimators



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

James-Stein Estimators



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

More intuition

1. 1970 Batting Averages for 18 Major League Players and Transformed Values X_i, θ_i

i	Player	Y_i = batting average for first 45 at bats	p_i = batting average for remainder of season	At bats for remainder of season	X_i	θ_i
		(1)	(2)	(3)	(4)	(5)
1	Clemente (Pitts, NL)	.400	.346	367	−1.35	−2.10
2	F. Robinson (Balt, AL)	.378	.298	426	−1.66	−2.79
3	F. Howard (Wash, AL)	.356	.276	521	−1.97	−3.11
4	Johnstone (Cal, AL)	.333	.222	275	−2.28	−3.96
5	Berry (Chi, AL)	.311	.273	418	−2.60	−3.17
6	Spencer (Cal, AL)	.311	.270	466	−2.60	−3.20
7	Kessinger (Chi, NL)	.289	.263	586	−2.92	−3.32
8	L. Alvarado (Bos, AL)	.267	.210	138	−3.26	−4.15
9	Santo (Chi, NL)	.244	.269	510	−3.60	−3.23
10	Swoboda (NY, NL)	.244	.230	200	−3.60	−3.83
11	Unser (Wash, AL)	.222	.264	277	−3.95	−3.30
12	Williams (Chi, AL)	.222	.256	270	−3.95	−3.43
13	Scott (Bos, AL)	.222	.303	435	−3.95	−2.71
14	Petrocelli (Bos, AL)	.222	.264	538	−3.95	−3.30
15	E. Rodriguez (KC, AL)	.222	.226	186	−3.95	−3.89
16	Campaneris (Oak, AL)	.200	.285	558	−4.32	−2.98
17	Munson (NY, AL)	.178	.316	408	−4.70	−2.53
18	Alvis (Mil, NL)	.156	.200	70	−5.10	−4.32

More intuition

- Separate trials assume independence
- Suppose you were waiting on data from the 5th group, but knew the first 4 groups succeeded
  - would you be more optimistic than when you started?
- Similarly, would repeated failures in the first 4 make you pessimistic about the 5th?
- Sharing (borrowing) of information formalizes this intuition.

Indication Finder

- Explore multiple histologies in one trial
  - Different cancer types
  - Different tumor types
- Combine the information across the histologies to make inferences about the drug effect in each histology
- Borrow strength across the histologies with hierarchical modeling

Potential Adaptations

- Sample size
  - stop accrual if histology/tumor type clearly does or does not meet goal
- Inclusion criteria
  - steer accrual to optimal subpopulations
  - steer accrual to subpopulations with remaining uncertainty

Questions for Collaborators

- How many subtypes
  - Need to prospectively define each subtype
- Goal response rate
  - May differ across subtypes
- Success & Futility stopping rules
  - Trade off more aggressive early stopping / smaller trial size vs. increase likelihood of erroneous inferences
- Maximum trial size
- Maximum subtype size
- Minimum subtype size before allow early stopping
- Prior on effect size

Model Details

$$p_t = \Pr(\text{Tumor Response} \mid \text{Tumor Type } t)$$

$$\theta_t = \log\left(\frac{p_t}{1 - p_t}\right)$$

$$\theta_t \sim N(\mu, \sigma^2)$$

$$\mu \sim N(0, 10^2)$$

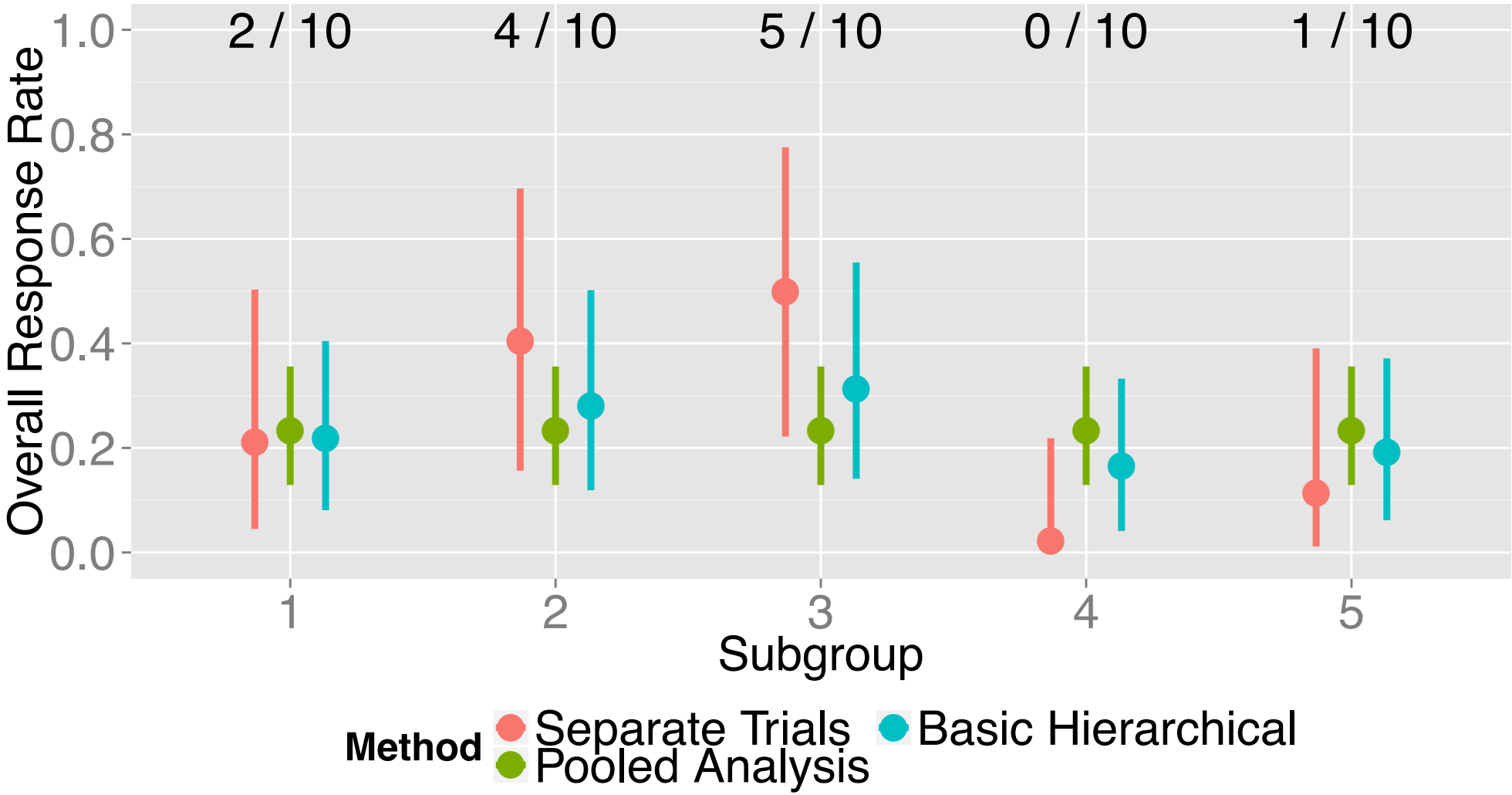
$$\sigma^2 \sim \Gamma^{-1}(1.0, 0.1)$$

At each interim analysis calculate

$$B_{t,Success} = \Pr(p_t > 0.20)$$

$$B_{t,Futility} = \Pr(p_t < 0.20)$$

Example Analysis



Indication Finder

- Statistical goals
  - Estimate response rate in each tumor subtype
  - Stop enrolling a subtype if it is unresponsive
    - Saves resources by not enrolling where futile
  - Stop enrolling a subtype if we determine is meets goal
    - Quickly move to next phase for this subpopulation
    - Saves resources for this trial for tumor types that need further study
- Statistical properties
  - If response rates similar across tumor types estimate response rates as if one large population
  - If response rates different across tumor types estimate response rates separately for each group
  - If somewhere in between, share data accordingly

Adaptive Decisions

- Enroll patients, track trial as data accumulates
- At each analysis calculate
  - Posterior distribution for response rate in tumor type t, p_t
  - Pr(Response Rate for type t > Goal rate) = Pr(p_t > Goal rate)
- If Pr(p_t > Goal) > 0.95 & 10 patients of type t
  - Stop enrolling tumor type t; move to next phase
- If Pr(p_t > Goal) < 0.10 & 10 patients of type t
  - Stop enrolling tumor type t for futility
- If 20 patients of type t, stop enrolling patients of tumor type t
- Stop when total reaches maximum sample size

Design Parameters

- Number of subtypes: 4
- Frequency of interim analysis: every 8 patients
- Goal response rate: 20%
- Success Stopping rule: $\Pr(p_t > \text{Goal}) > 0.95$
- Futility stopping rule: $\Pr(p_t > \text{Goal}) < 0.10$
- Total maximum sample size: 40
- Maximum size per subtype: 20
- Minimum size before stopping: 10
- Prior on effect: N(0, 10), vague prior
- Prior on heterogeneity: Gamma(1.0, 0.1)

Operating Characteristics, Positive Response

Scenario 5:

Hist	True	Goal	Cap	Fut1	Suc	Prev
Type1	0.50	0.20	20	0.10	0.95	0.35
Type2	0.50	0.20	20	0.10	0.95	0.25
Type3	0.50	0.20	20	0.10	0.95	0.15
Type4	0.50	0.20	20	0.10	0.95	0.25

40 max

Parameter	Fut	N Cap	Pos	SS	SD	Avg SD	Avg Pi
Type1	0.000	0.086	0.914	11.2	1.6	0.1090	0.5046
Type2	0.000	0.252	0.748	10.6	1.6	0.1110	0.4999
Type3	0.000	0.835	0.165	7.7	1.9	0.1211	0.4953
Type4	0.000	0.257	0.743	10.5	1.6	0.1110	0.5003
Total				40.0	0.0		

60 max

Parameter	Fut	N Cap	Pos	SS	SD	Avg SD	Avg Pi
Type1	0.000	0.030	0.970	11.5	2.2	0.1049	0.5059
Type2	0.001	0.034	0.965	11.7	2.2	0.1042	0.5031
Type3	0.000	0.040	0.960	12.3	2.2	0.1030	0.5037
Type4	0.001	0.032	0.967	11.7	2.2	0.1045	0.5053
Total				47.2	4.1		

Operating Characteristics, One Responder

Scenario 9:

Hist	True	Goal	Cap	Fut1	Suc	Prev
Type1	0.10	0.20	20	0.10	0.95	0.35
Type2	0.10	0.20	20	0.10	0.95	0.25
Type3	0.40	0.20	20	0.10	0.95	0.15
Type4	0.10	0.20	20	0.10	0.95	0.25

40 max

Parameter	Fut	N Cap	Pos	SS	SD	Avg SD	Avg Pi
Type1	0.560	0.436	0.004	12.6	2.6	0.0700	0.1050
Type2	0.362	0.637	0.001	10.4	2.1	0.0774	0.1075
Type3	0.004	0.950	0.046	6.7	2.3	0.1524	0.3534
Type4	0.364	0.636	0.000	10.4	2.3	0.0773	0.1102
Total				40.0	0.0		

60 max

Parameter	Fut	N Cap	Pos	SS	SD	Avg SD	Avg Pi
Type1	0.665	0.331	0.004	14.6	4.0	0.0631	0.0954
Type2	0.634	0.359	0.007	14.2	3.7	0.0643	0.0973
Type3	0.018	0.496	0.486	13.6	3.8	0.1210	0.3906
Type4	0.637	0.363	0.000	14.2	3.7	0.0650	0.0977
Total				56.7	4.1		

Operating Characteristics, Poor Response

Scenario 8:

Hist	True	Goal	Cap	Fut1	Suc	Prev
Type1	0.20	0.20	20	0.10	0.95	0.35
Type2	0.20	0.20	20	0.10	0.95	0.25
Type3	0.10	0.20	20	0.10	0.95	0.15
Type4	0.10	0.20	20	0.10	0.95	0.25

40 max

Parameter	Fut	N Cap	Pos	SS	SD	Avg SD	Avg Pi
Type1	0.230	0.714	0.056	13.5	2.8	0.0929	0.1943
Type2	0.139	0.837	0.024	10.2	2.4	0.1017	0.1881
Type3	0.040	0.960	0.000	6.3	2.2	0.0954	0.1199
Type4	0.320	0.680	0.000	10.0	2.3	0.0789	0.1117
Total				40.0	0.0		

60 max

Parameter	Fut	N Cap	Pos	SS	SD	Avg SD	Avg Pi
Type1	0.299	0.634	0.067	17.0	3.9	0.0821	0.1890
Type2	0.272	0.668	0.060	15.9	3.7	0.0839	0.1876
Type3	0.473	0.527	0.000	11.6	3.2	0.0719	0.1032
Type4	0.647	0.350	0.003	13.6	3.4	0.0650	0.0948
Total				58.1	3.0		

Benefits to Borrowing

- Suppose you are in the alternative scenario.
  - Multiplicities work both ways. Separate trials have limited power, generally you miss multiple effective groups.
  - A general trend of effective groups provides a backdrop for accepting "borderline" groups.
  - Suppose separate analyses would results in pvalues of 0.01, 0.02, 0.01, 0.03, and 0.06.
    - the last isn't significant, but is borderline.
    - the general good trend allows the hierarchical model to conclude the 0.06 is a random low, and declare success in that group.
    - Compare to separate pvalues of 0.30, 0.50, 0.15, 0.70, and 0.06 where the 5^{th} group looks more like a random high.
- More power in the alternative scenario.

Disadvantages

- Outlying subgroups or clusters can be difficult
  - If the drug works in only one subgroup, then the power for that effect can be decreased relative to separate trials
  - If the drug is effective in all but one subgroup, then Type I error for that subgroup can increase relative to separate trials
  - Intuitively, if the value really is different, "shrinkage" can pull it in the wrong direction.

Other materials

- Berry et. al. Bayesian hierarchical modeling of patients subpopulations : efficient designs of phase II oncology clinical trials. Clinical Trials 2013, 720-734. http://www.ncbi.nlm.nih.gov/pubmed/23983156
- Kert Viele's YouTube talk https://www.youtube.com/watch?v=H7C11PvybOk