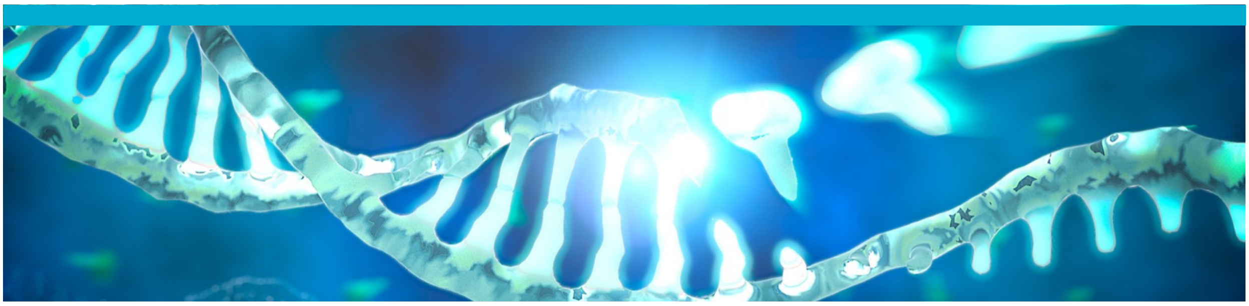
AstraZeneca
IMED Biotech Unit

Decision Making in Early Clinical Development: The framework used within AstraZeneca

Paul Frewer, Associate Director, Statistics
Early Clinical Biometrics, Early Clinical Development, IMED Biotech Unit, AstraZeneca,

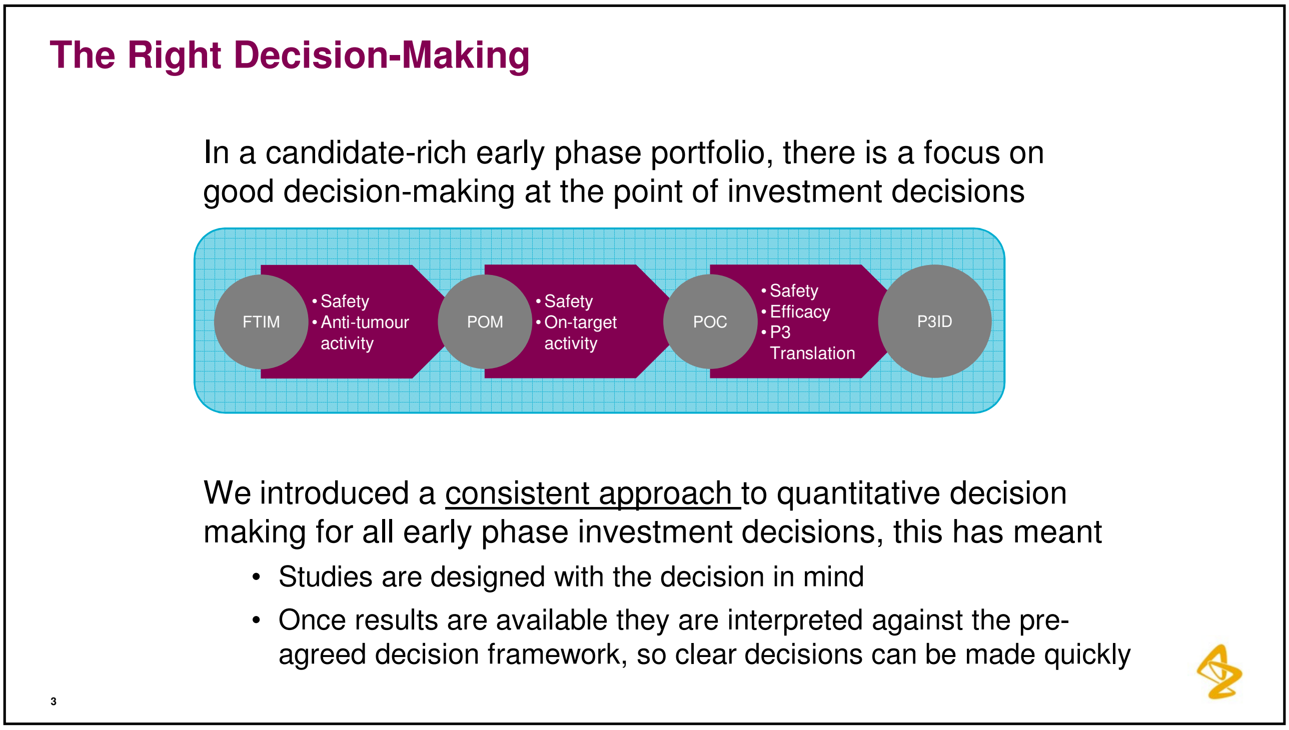
European Statistical Meeting on Decision Making in Drug Development, Paris, Dec 12th 2018



What we will cover today

- Background
- Decision Framework
- Special Considerations
  - Actions in Consider Zone, Multiple Endpoints, Accelerating Development
- What are Acceptable Operating Characteristics
- Sizing a Study based on the Decision Framework
- Interim Analyses (Futility and Administrative)
- Implementation, Software Development and Experience to date

2



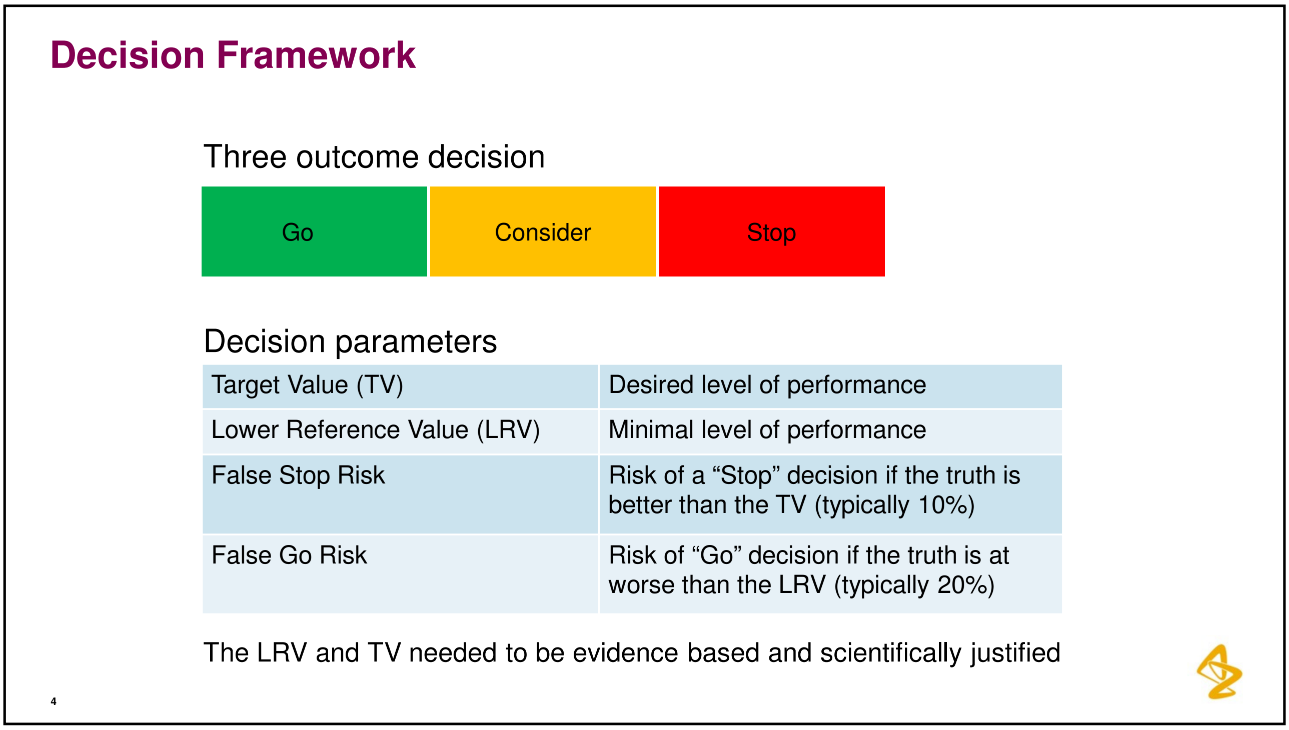
The Right Decision-Making

In a candidate-rich early phase portfolio, there is a focus on good decision-making at the point of investment decisions

FTIM
• Safety
• Anti-tumour activity
POM
• Safety
• On-target activity
POC
• Safety
• Efficacy
• P3 Translation
P3ID

We introduced a consistent approach to quantitative decision making for all early phase investment decisions, this has meant

- Studies are designed with the decision in mind
- Once results are available they are interpreted against the pre-agreed decision framework, so clear decisions can be made quickly



Decision Framework

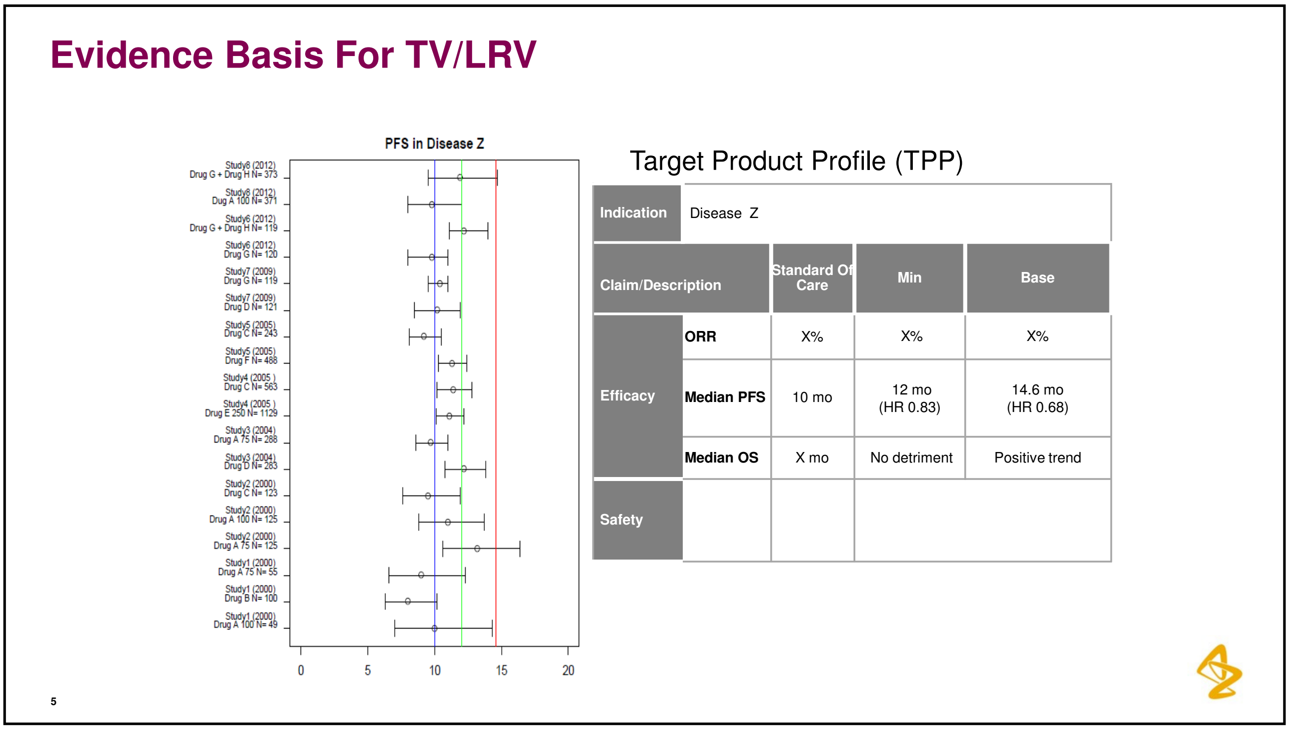
Three outcome decision

Go	Consider	Stop

Decision parameters

Target Value (TV)	Desired level of performance
Lower Reference Value (LRV)	Minimal level of performance
False Stop Risk	Risk of a “Stop” decision if the truth is better than the TV (typically 10%)
False Go Risk	Risk of “Go” decision if the truth is at worse than the LRV (typically 20%)

The LRV and TV needed to be evidence based and scientifically justified

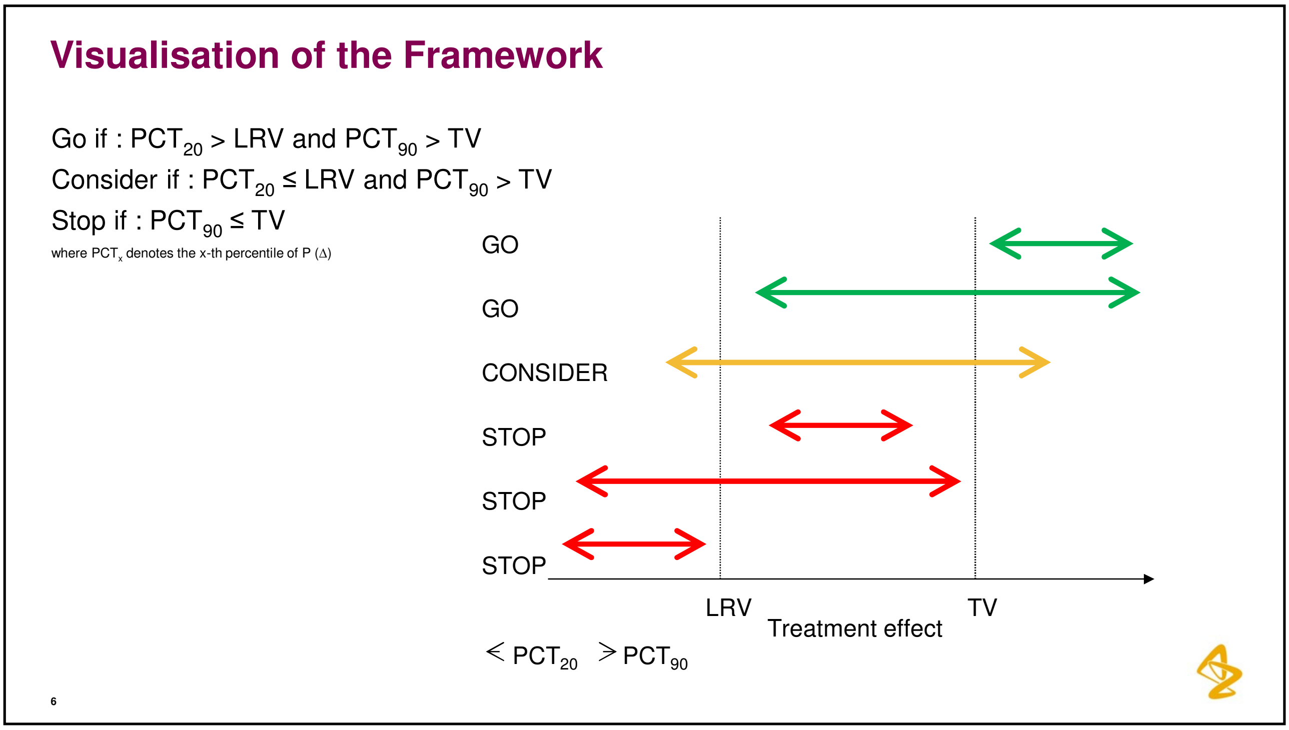


Evidence Basis For TV/LRV

Target Product Profile (TPP)

Indication	Disease Z			
Claim/Description		Standard Of Care	Min	Base
Efficacy	ORR	X%	X%	X%
	Median PFS	10 mo	12 mo (HR 0.83)	14.6 mo (HR 0.68)
	Median OS	X mo	No detriment	Positive trend
Safety				

5



Visualisation of the Framework

Go if : PCT_{20} > LRV and PCT_{90} > TV

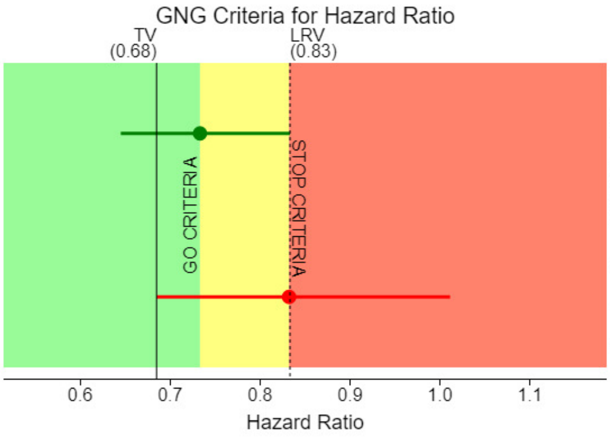
Consider if : $PCT_{20} \leq$ LRV and PCT_{90} > TV

Stop if : $PCT_{90} \leq$ TV

where PCT_x denotes the x-th percentile of P (Δ)

6

Decision Plot



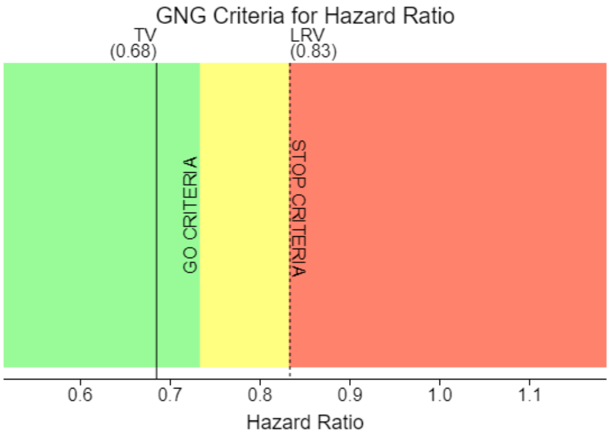
Value/Criteria	Description
Target Value	A hazard ratio of 0.68
Lower Reference Value	A hazard ratio of 0.83
Go	If there is ≥ 80% chance that the hazard ratio is ≤ 0.83* e.g. observed hazard ratio** ≤ 0.73
Stop	If there is <10% chance that the hazard ratio is ≤ 0.68* e.g. observed hazard ratio** ≥ 0.83

** Assuming 248 patients, allocation ratio = 1.0, 173 events
* Stop and Go correspond to lower-limit of 1-sided 90% CI and upper-limit of 1-sided 80% CI
The actual criteria will be driven by the stated probabilities so that if the observed data do not follow the assumptions, the GNG values will change

The sample size had been calculated to detect a Hazard Ratio=0.685 assuming 80% power and a 1-sided alpha=0.05

7

Decision Plot for Governance



Value/Criteria	Description
Target Value	A hazard ratio of 0.68
Lower Reference Value	A hazard ratio of 0.83
Go	If there is ≥ 80% chance that the hazard ratio is ≤ 0.83* e.g. observed hazard ratio** ≤ 0.73
Stop	If there is <10% chance that the hazard ratio is ≤ 0.68* e.g. observed hazard ratio** ≥ 0.83

** Assuming 248 patients, allocation ratio = 1.0, 173 events
* Stop and Go correspond to lower-limit of 1-sided 90% CI and upper-limit of 1-sided 80% CI
The actual criteria will be driven by the stated probabilities so that if the observed data do not follow the assumptions, the GNG values will change

The sample size had been calculated to detect a Hazard Ratio=0.685 assuming 80% power and a 1-sided alpha=0.05

8

Operating Characteristics

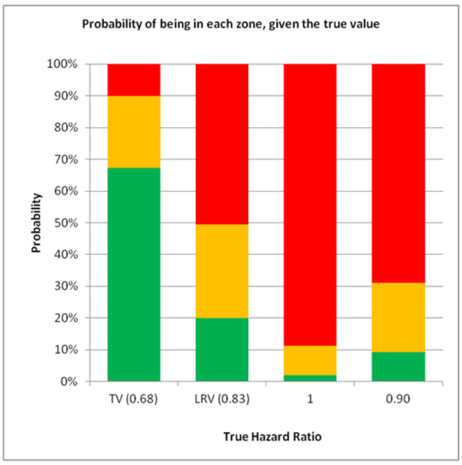
Why are the operating characteristics important?

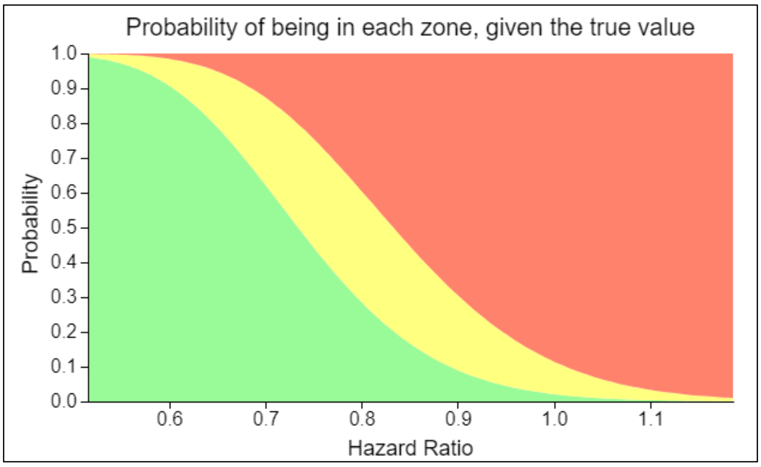
They enable evaluation of whether the framework is robust and will enable clear decisions or if the chance of being in the consider zone is too high

True effect	Probability of Making each Decision for a given True Effect		
	Go	Consider	Stop
Good (TV; HR=0.68)	67.3%	22.7%	10.0%
Reasonable (LRV; HR=0.83)	20.0%	29.7%	50.3%
Minimal Effect (1/4 TV HR=0.90)	9.3%	21.9%	68.8%
No Effect (HR=1)	2.1%	9.3%	88.6%

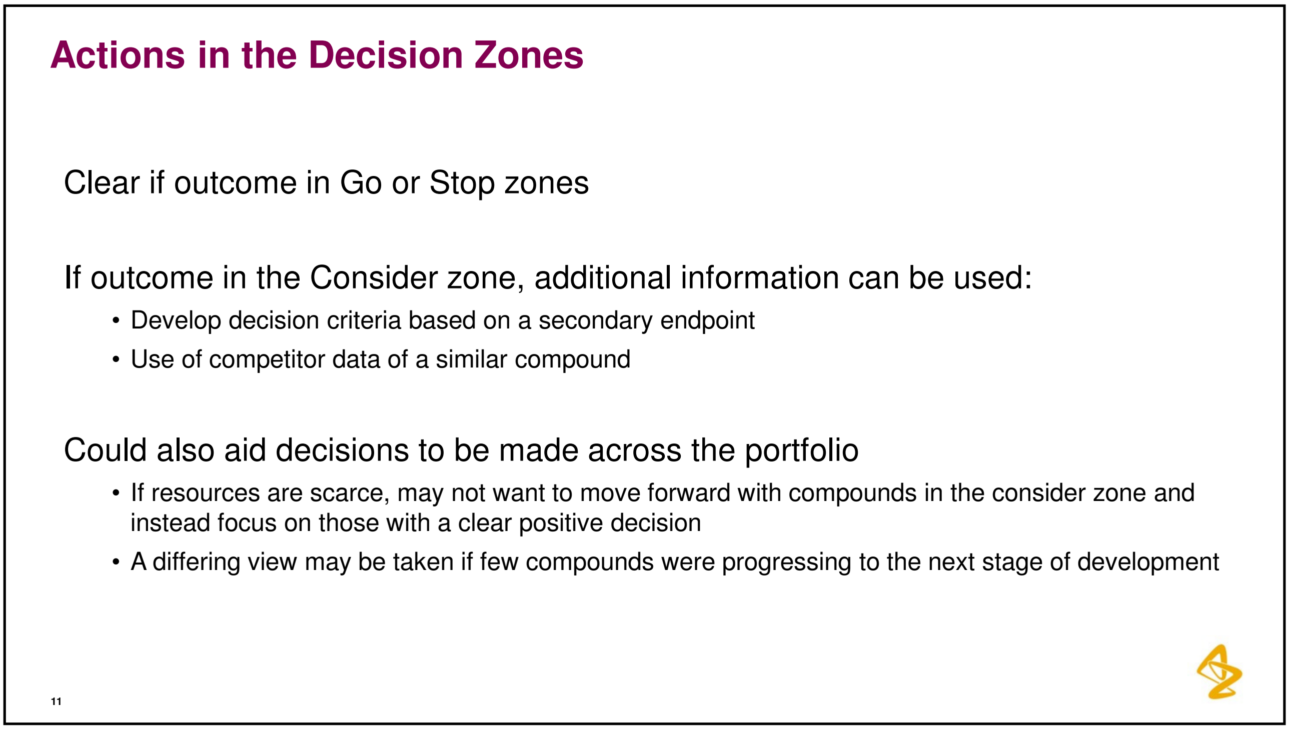
9

Graphical Displays of Operating Characteristics





10



Actions in the Decision Zones

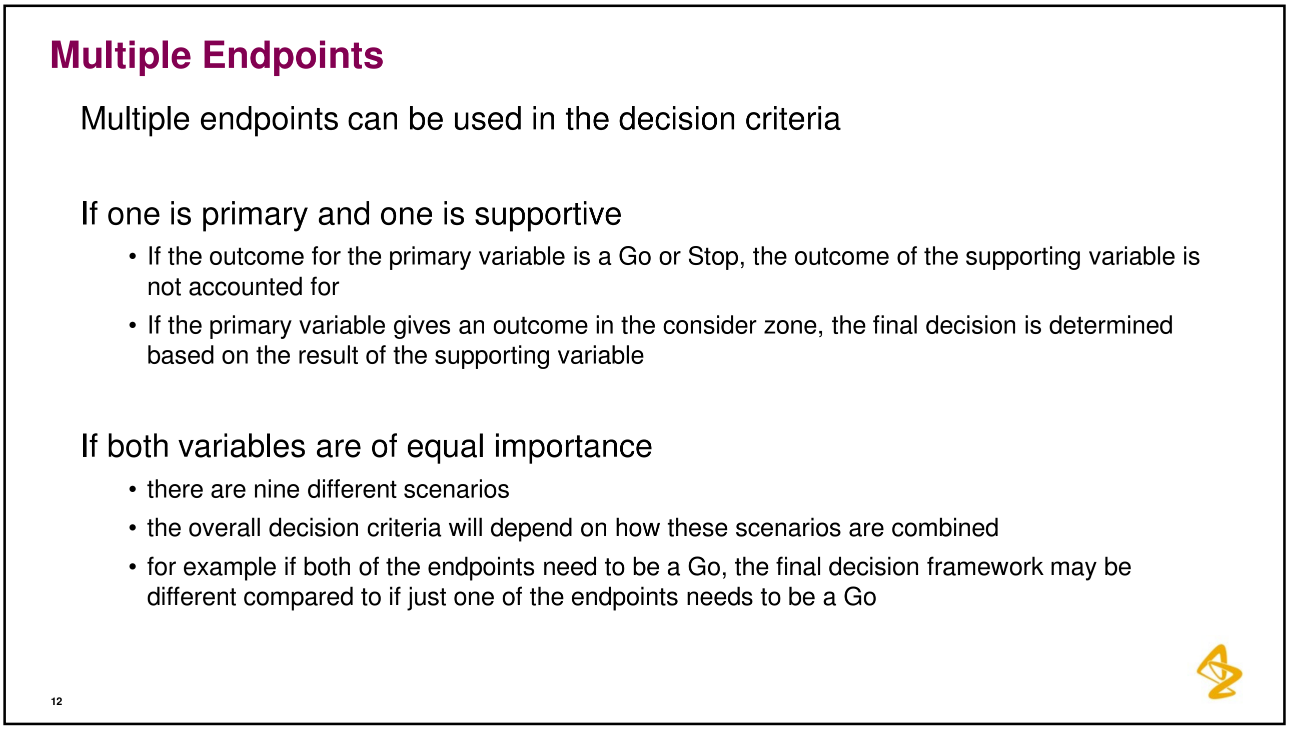
Clear if outcome in Go or Stop zones

If outcome in the Consider zone, additional information can be used:

- Develop decision criteria based on a secondary endpoint
- Use of competitor data of a similar compound

Could also aid decisions to be made across the portfolio

- If resources are scarce, may not want to move forward with compounds in the consider zone and instead focus on those with a clear positive decision
- A differing view may be taken if few compounds were progressing to the next stage of development



Multiple Endpoints

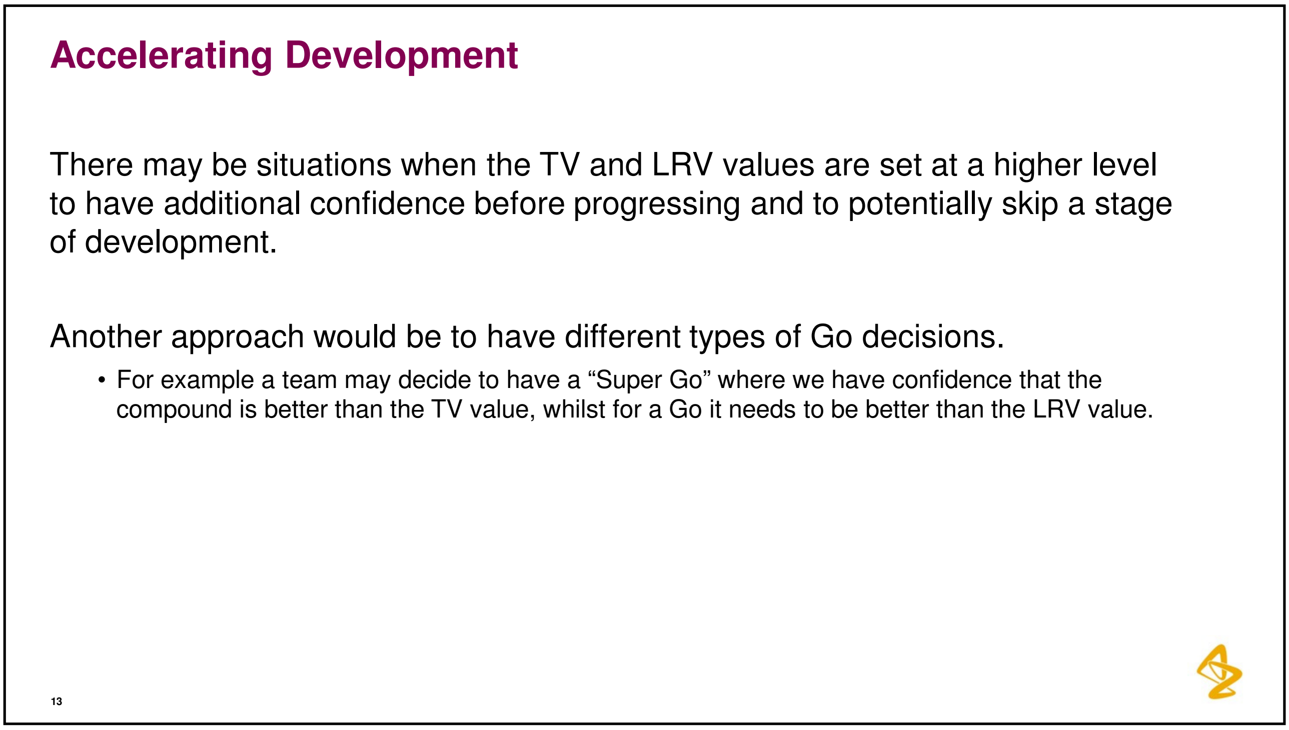
Multiple endpoints can be used in the decision criteria

If one is primary and one is supportive

- If the outcome for the primary variable is a Go or Stop, the outcome of the supporting variable is not accounted for
- If the primary variable gives an outcome in the consider zone, the final decision is determined based on the result of the supporting variable

If both variables are of equal importance

- there are nine different scenarios
- the overall decision criteria will depend on how these scenarios are combined
- for example if both of the endpoints need to be a Go, the final decision framework may be different compared to if just one of the endpoints needs to be a Go

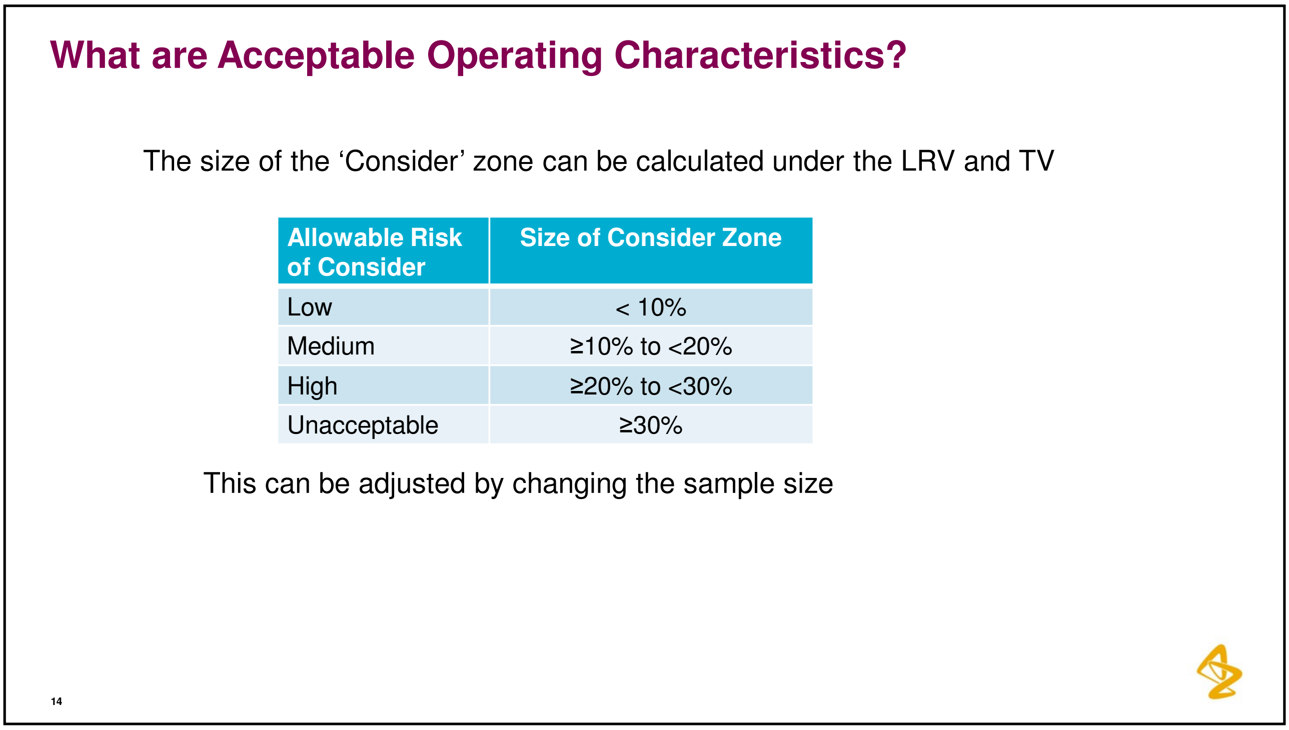


Accelerating Development

There may be situations when the TV and LRV values are set at a higher level to have additional confidence before progressing and to potentially skip a stage of development.

Another approach would be to have different types of Go decisions.

- For example a team may decide to have a “Super Go” where we have confidence that the compound is better than the TV value, whilst for a Go it needs to be better than the LRV value.



What are Acceptable Operating Characteristics?

The size of the ‘Consider’ zone can be calculated under the LRV and TV

Allowable Risk of Consider	Size of Consider Zone
Low	< 10%
Medium	≥10% to <20%
High	≥20% to <30%
Unacceptable	≥30%

This can be adjusted by changing the sample size

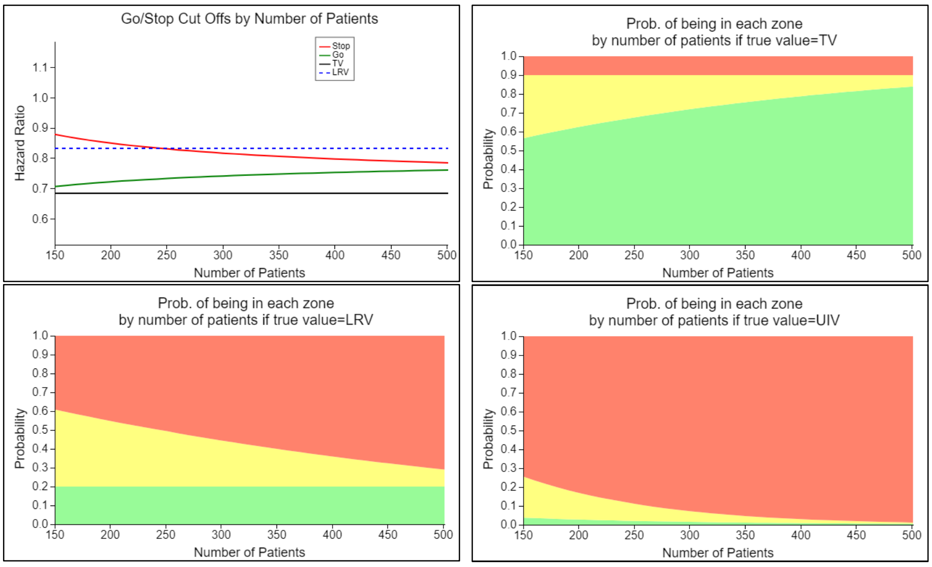
Operating Characteristics: 126 Events

If we had sized the study to detect a Hazard Ratio=0.685 assuming 90% power and a 1-sided alpha=0.2 (false go and stop risks in decision framework), 126 events (180 patients) would be required

True effect	Probability of Making each Decision for a given True Effect		
	Go	Consider	Stop
Good (TV; HR=0.68)	60.2%	29.8%	10.0%
Reasonable (LRV; HR=0.83)	20.0%	37.2%	42.8%
Minimal Effect (1/4 TV HR=0.90)	10.5%	30.4%	59.2%
No Effect (HR=1)	3.1%	16.9%	80.0%

The operating characteristics assuming 126 events would be unacceptable

Operating Characteristics by Sample Size



Assumes data maturity of 70%, e.g. 150 patients have 105 events and 500 patients have 350 events, UIV=1

Sizing a Study based on the Decision Framework

Could the sample size be an output from the decision criteria rather than calculated via a power calculation?

Yes - If we set either the P(Go|TV) or P(Stop/LRV) as an input, the required sample size to achieve this is an output from the decision framework

- For binary endpoints both of these may need to be specified

Questions may arise on how the sample size is written in the protocol

The advantage within early development is that the trial is being sized according to the decision and the risks you want to undertake

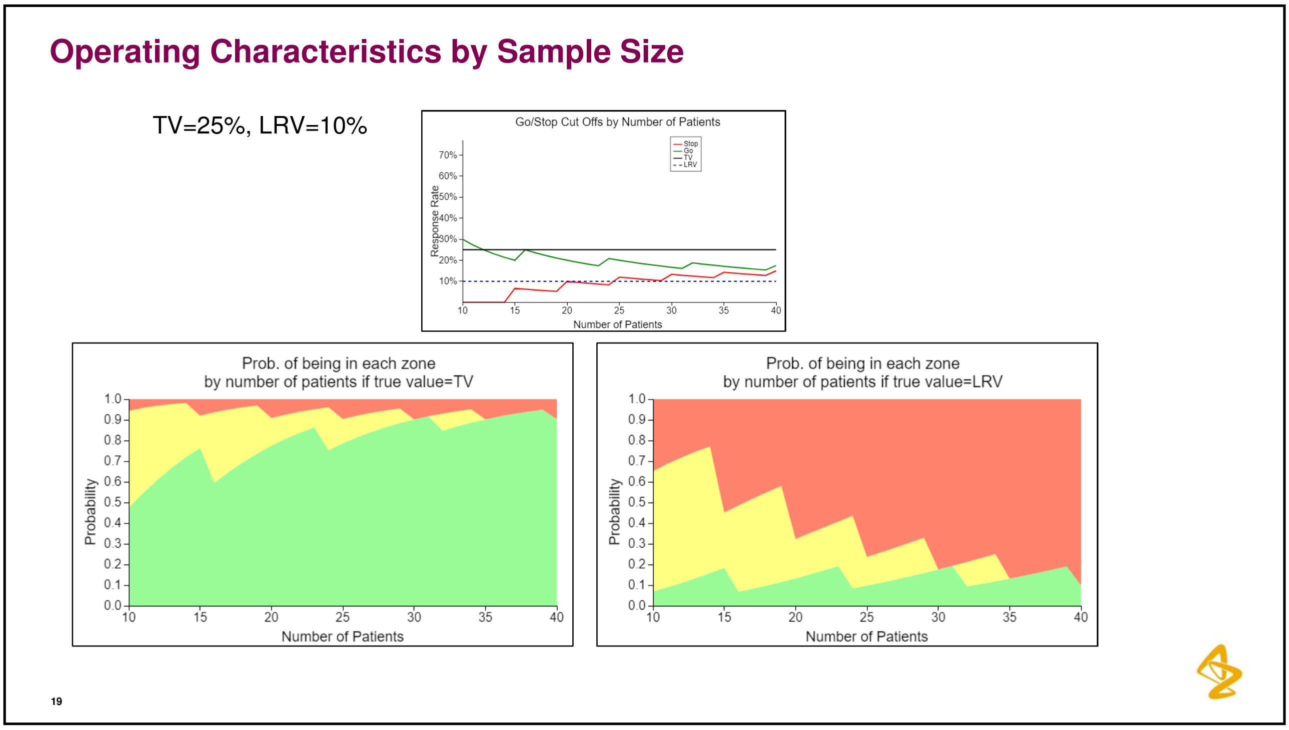
May be able to perform a smaller, shorter trial and to reach a decision earlier

Stability of Operating Characteristics in Single Arm Studies with a Binary Endpoint

Due to the nature of the binomial distribution, if an additional patient was added the operating characteristics of the decision criteria can get worse (see example on following slide)

When selecting a sample size, should we be looking at

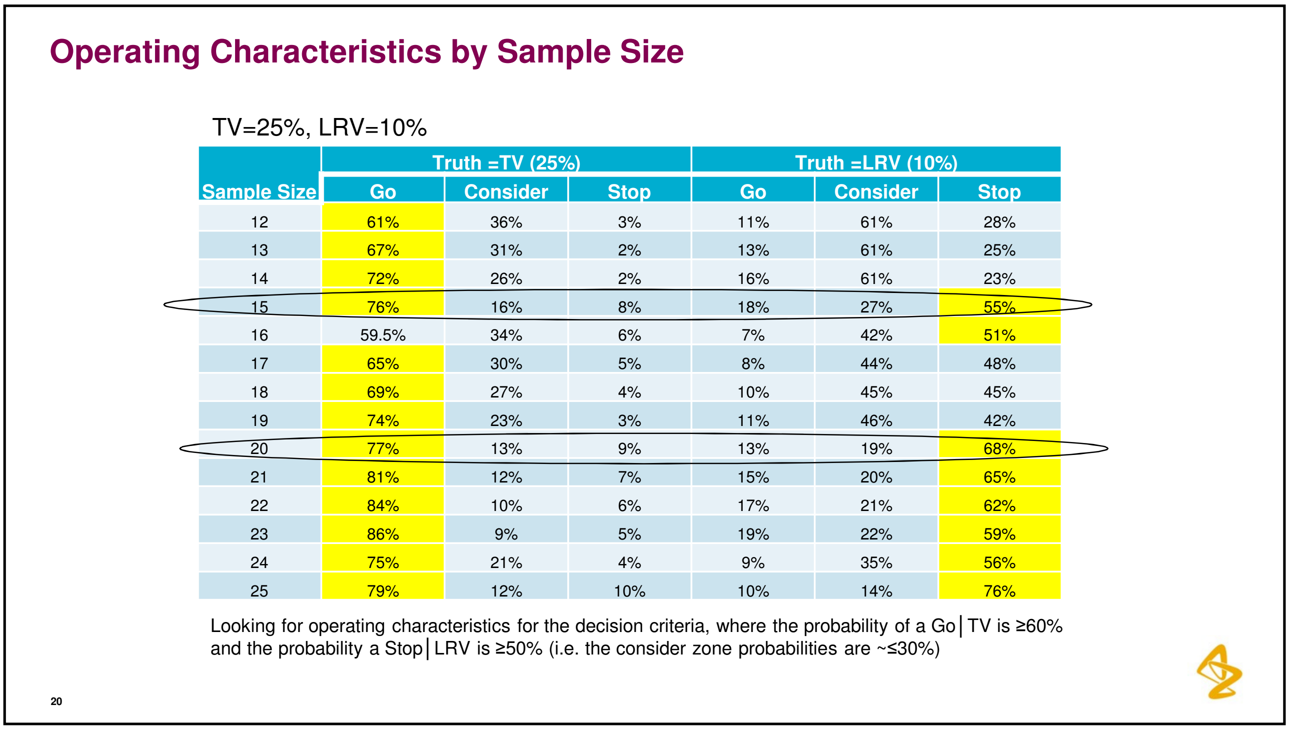
1) the first occurrence of acceptable criteria
2) the minimum number required to always have acceptable criteria



Operating Characteristics by Sample Size

TV=25%, LRV=10%

19



Operating Characteristics by Sample Size

TV=25%, LRV=10%

	Truth =TV (25%)			Truth =LRV (10%)		
Sample Size	Go	Consider	Stop	Go	Consider	Stop
12	61%	36%	3%	11%	61%	28%
13	67%	31%	2%	13%	61%	25%
14	72%	26%	2%	16%	61%	23%
15	76%	16%	8%	18%	27%	55%
16	59.5%	34%	6%	7%	42%	51%
17	65%	30%	5%	8%	44%	48%
18	69%	27%	4%	10%	45%	45%
19	74%	23%	3%	11%	46%	42%
20	77%	13%	9%	13%	19%	68%
21	81%	12%	7%	15%	20%	65%
22	84%	10%	6%	17%	21%	62%
23	86%	9%	5%	19%	22%	59%
24	75%	21%	4%	9%	35%	56%
25	79%	12%	10%	10%	14%	76%

Looking for operating characteristics for the decision criteria, where the probability of a Go│TV is ≥60% and the probability a Stop│LRV is ≥50% (i.e. the consider zone probabilities are ~≤30%)

20

Sample Size Look Up Tables

Sample size look up tables are provided (based on the minimum sample size to always have “acceptable” operating characteristics)

They all assume the standard probabilities for a False Go and a False Stop of 20% and 10% respectively

Sizes are given for a range of what are acceptable operating characteristics

1) The probabilities of a Go│TV is ≥60% and a Stop│LRV is ≥50% (i.e. Consider probabilities are ~≤30%)
2) The probabilities of a Go│TV is ≥70% and a Stop│LRV is ≥60% (i.e. Consider probabilities are ~≤20%)
3) The probabilities of a Go│TV is ≥80% and a Stop│LRV is ≥70% (i.e. Consider probabilities are ~≤10%)
4) The probabilities of a Go│TV is ≥90% and a Stop│LRV is ≥80% (i.e. No Consider zone)

Look Up Table: 15% Difference between LRV and TV

		Minimum Sample Required to ensure acceptable pre-defined operating characteristics			
		Approx size of the Consider Zone			
LRV	TV	~ 30%	~ 20%	~ 10%	None
5%	20%	18	18	25	32
10%	25%	20	25	30	35
15%	30%	21	29	33	45
20%	35%	25	32	38	48
25%	40%	24	33	42	53
30%	45%	27	34	42	57
35%	50%	26	35	44	57
40%	55%	25	38	48	55
45%	60%	29	34	45	58
50%	65%	24	33	43	58
55%	70%	24	32	41	58
60%	75%	22	31	41	52
65%	80%	22	30	38	47
70%	85%	19	24	34	44
75%	90%	16	21	27	35
80%	95%	14	15	21	27

Interim Analyses

The decision framework can also be used to set interim decision criteria. In general, interim analyses in early phase studies fall into two categories

Adaptive designs, where internal changes are made to the trial

- Futility analyses – the current trial is stopped early if it is unlikely to be successful

Non-adaptive designs, where changes are made externally to the trial

- Administrative analyses – other project activities are accelerated (or decelerated) on the basis of interim data from the current trial, but the current trial is not changed.

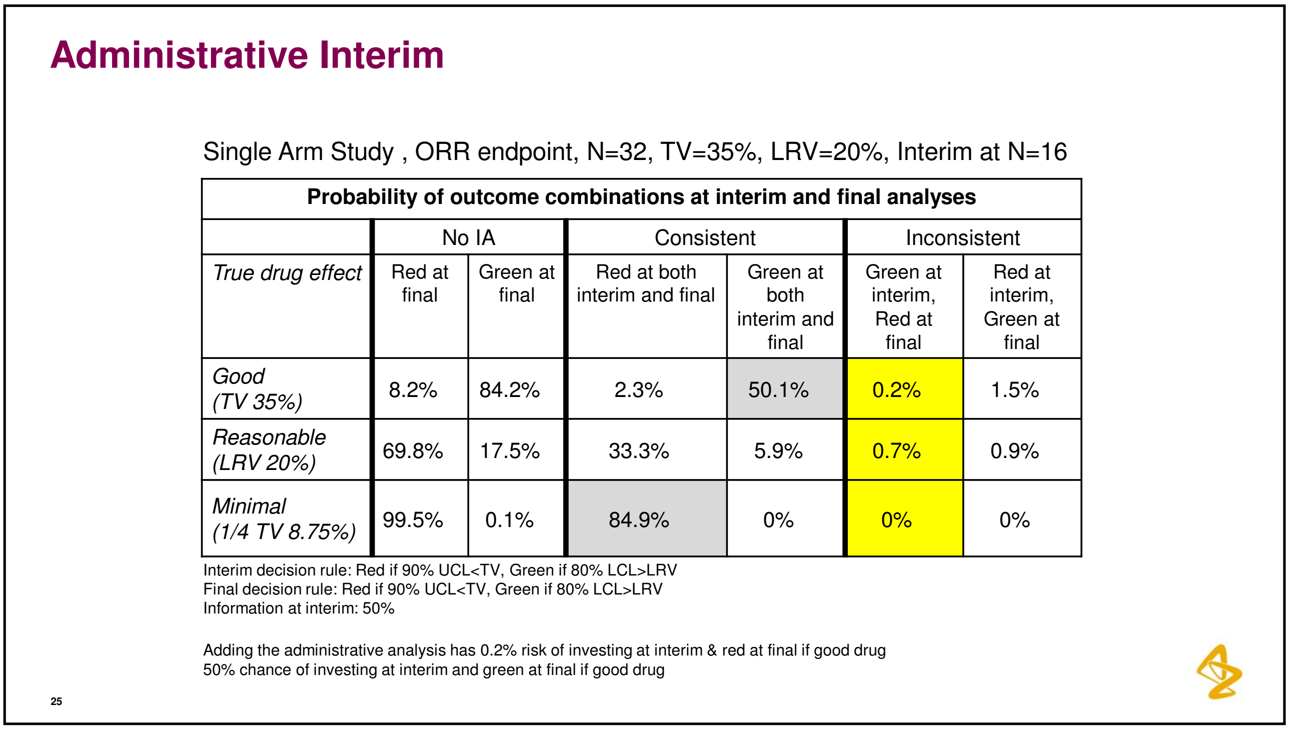
Futility Interim

An interim analysis for futility was also investigated after 87 events in the previous PFS example. The same framework for the TV, LRV and the risks was applied to the interim data and the interim decision criteria were as follows:

- Continue: HR < 0.90
- Stop: HR ≥ 0.90

Probability of stopping	True drug effect	IA stopping rule	
		No Interim	Interim (87 Events)
At any time (IA or Final analysis)	Good (TV; HR=0.68)	10.0%	15.2%
	Reasonable (LRV; HR=0.83)	50.3%	56.5%
	Minimal Effect (1/4 TV HR=0.90)	68.8%	73.3%
	No Effect (HR=1)	88.6%	90.6%
Early (At IA)	Good (TV; HR=0.68)		10.0%
	Reasonable (LRV; HR=0.83)		35.7%
	Minimal Effect (1/4 TV HR=0.90)		49.0%
	No Effect (HR=1)		68.6%

The probability of stopping an ineffective drug at the interim was high, and the overall probability of stopping a good drug was only increased by 5.2% to 15.2%



Administrative Interim

Single Arm Study , ORR endpoint, N=32, TV=35%, LRV=20%, Interim at N=16

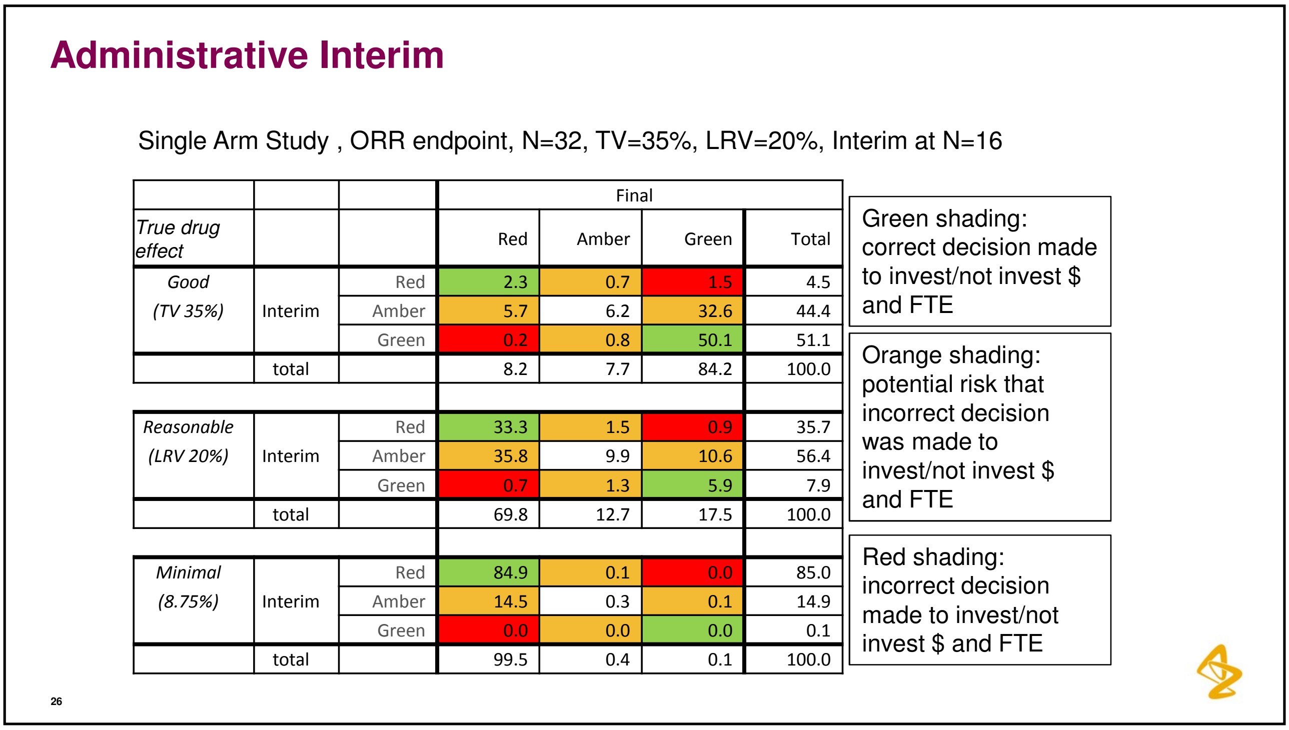
Probability of outcome combinations at interim and final analyses

	No IA		Consistent		Inconsistent	
True drug effect	Red at final	Green at final	Red at both interim and final	Green at both interim and final	Green at interim, Red at final	Red at interim, Green at final
Good (TV 35%)	8.2%	84.2%	2.3%	50.1%	0.2%	1.5%
Reasonable (LRV 20%)	69.8%	17.5%	33.3%	5.9%	0.7%	0.9%
Minimal (1/4 TV 8.75%)	99.5%	0.1%	84.9%	0%	0%	0%

Interim decision rule: Red if 90% UCL<TV, Green if 80% LCL>LRV
Final decision rule: Red if 90% UCL<TV, Green if 80% LCL>LRV
Information at interim: 50%

Adding the administrative analysis has 0.2% risk of investing at interim & red at final if good drug
50% chance of investing at interim and green at final if good drug

25



Administrative Interim

Single Arm Study , ORR endpoint, N=32, TV=35%, LRV=20%, Interim at N=16

			Final			
True drug effect			Red	Amber	Green	Total
Good (TV 35%)	Interim	Red	🟩 2.3	🟧 0.7	🟥 1.5	4.5
		Amber	🟧 5.7	6.2	🟧 32.6	44.4
		Green	🟥 0.2	🟧 0.8	🟩 50.1	51.1
	total		8.2	7.7	84.2	100.0
Reasonable (LRV 20%)	Interim	Red	🟩 33.3	🟧 1.5	🟥 0.9	35.7
		Amber	🟧 35.8	9.9	🟧 10.6	56.4
		Green	🟥 0.7	🟧 1.3	🟩 5.9	7.9
	total		69.8	12.7	17.5	100.0
Minimal (8.75%)	Interim	Red	🟩 84.9	🟧 0.1	🟥 0.0	85.0
		Amber	🟧 14.5	0.3	🟧 0.1	14.9
		Green	🟥 0.0	🟧 0.0	🟩 0.0	0.1
	total		99.5	0.4	0.1	100.0

Green shading: correct decision made to invest/not invest $ and FTE

Orange shading: potential risk that incorrect decision was made to invest/not invest $ and FTE

Red shading: incorrect decision made to invest/not invest $ and FTE

26

Timing of Interim in Single Arm Studies with a Binary Endpoint

When deciding on the timing of a futility interim in these studies, in the past generally picked a point in time (e.g. with 50% of the patients) rather than look at the range of possible timings for an interim and selected which one is “best”

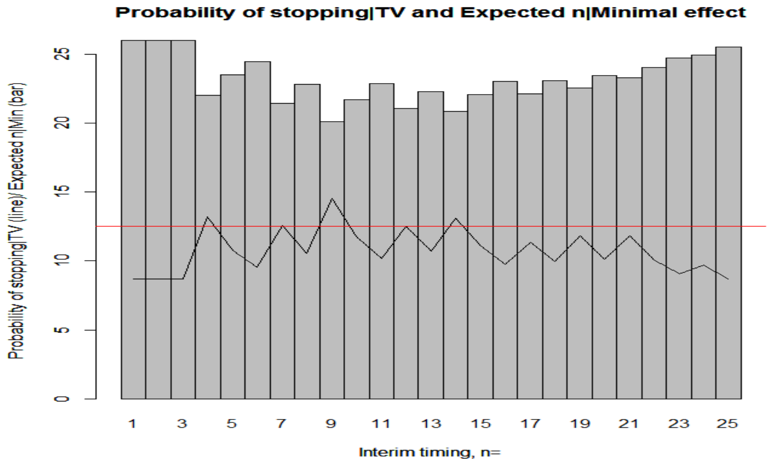
In order to decide what is “best” need to assess

1) Expected N if LRV is true
2) Probability of stopping at an interim or at the final analysis if TV is true
3) Operational considerations

Code developed which allow assessment of 1) and 2) over all possible timings for the interim to enable the interim to have the most benefit

27

Timing of Interim: N=26, TV=50%, LRV=35%



In deciding when to schedule the interim, we decided we did not want the probability of stopping if the TV (50%) was the truth to be > 12.5%

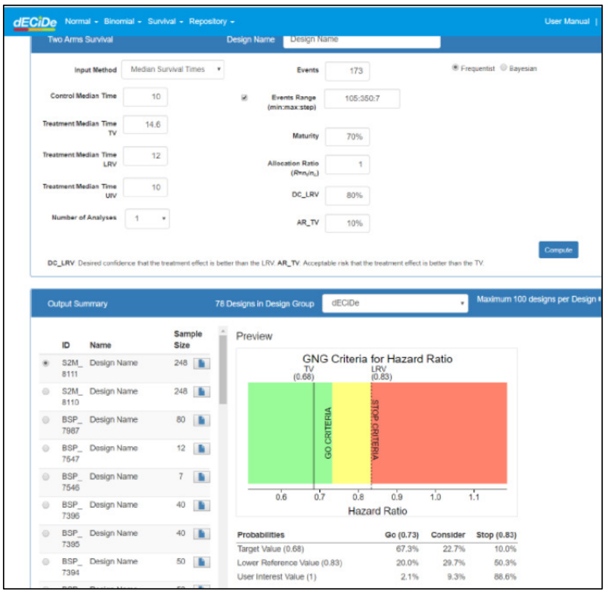
The expected N if the LRV (35%) was the truth is minimised if the interim is at N=12

28

Implementation

- Implemented in 2013
- Initially Excel, SAS and R functions developed for setting frequentist decision cut-off values and simulating operating characteristics
- Standardized presentations to governance
- Software solution developed with Cytel has been in place for 2 years
- Bayesian designs included in the software
- Published in Pharmaceutical Statistics and presented externally

Software Solution



Two Arms Survival
Design Name: Design Name

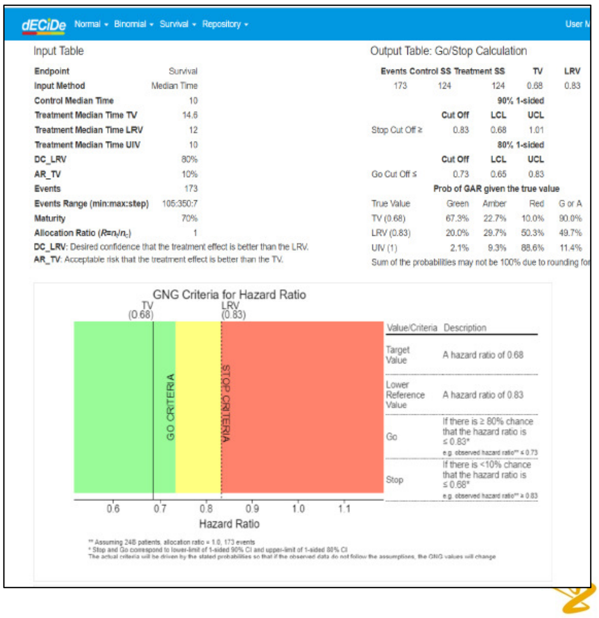
Input Method: Median Survival Times
Events: 173
● Frequentist ○ Bayesian
Control Median Time: 10
☑ Events Range (min:max:step): 105:350:7
Treatment Median Time TV: 14.6
Maturity: 70%
Treatment Median Time LRV: 12
Allocation Ratio ($R=n_t/n_c$): 1
Treatment Median Time UIV: 10
DC_LRV: 80%
Number of Analyses: 1
AR_TV: 10%

DC_LRV: Desired confidence that the treatment effect is better than the LRV. **AR_TV**: Acceptable risk that the treatment effect is better than the TV.

Output Summary
78 Designs in Design Group: dECiDe
Maximum 100 designs per Design

	ID	Name	Sample Size
●	S2M_8111	Design Name	248
○	S2M_8110	Design Name	248
○	BSP_7987	Design Name	80
○	BSP_7547	Design Name	12
○	BSP_7546	Design Name	7
○	BSP_7396	Design Name	40
○	BSP_7395	Design Name	40
○	BSP_7394	Design Name	50

Probabilities	Go (0.73)	Consider	Stop (0.83)
Target Value (0.68)	67.3%	22.7%	10.0%
Lower Reference Value (0.83)	20.0%	29.7%	50.3%
User Interest Value (1)	2.1%	9.3%	88.6%



Input Table

Endpoint	Survival
Input Method	Median Time
Control Median Time	10
Treatment Median Time TV	14.6
Treatment Median Time LRV	12
Treatment Median Time UIV	10
DC_LRV	80%
AR_TV	10%
Events	173
Events Range (min:max:step)	105:350:7
Maturity	70%
Allocation Ratio ($R=n_t/n_c$)	1

DC_LRV: Desired confidence that the treatment effect is better than the LRV.
AR_TV: Acceptable risk that the treatment effect is better than the TV.

Output Table: Go/Stop Calculation

Events	Control SS	Treatment SS	TV	LRV
173	124	124	0.68	0.83

		90% 1-sided	
	Cut Off	LCL	UCL
Stop Cut Off ≥	0.83	0.68	1.01

		80% 1-sided	
	Cut Off	LCL	UCL
Go Cut Off ≤	0.73	0.65	0.83

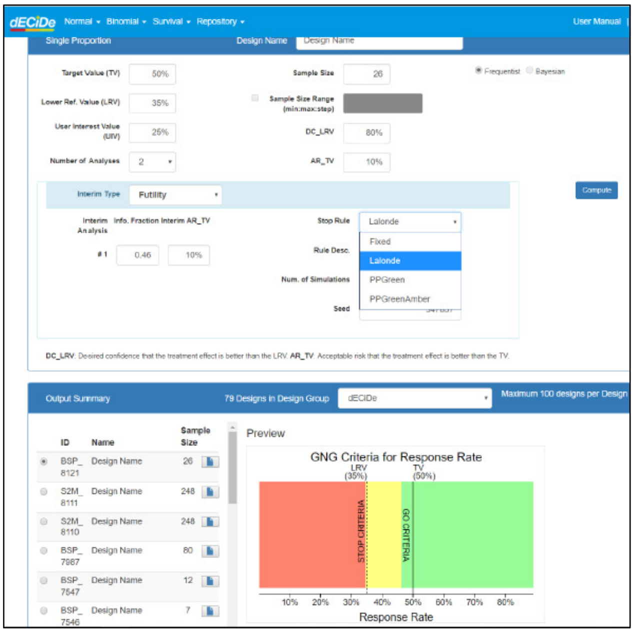
	Prob of GAR given the true value			
True Value	Green	Amber	Red	G or A
TV (0.68)	67.3%	22.7%	10.0%	90.0%
LRV (0.83)	20.0%	29.7%	50.3%	49.7%
UIV (1)	2.1%	9.3%	88.6%	11.4%

Sum of the probabilities may not be 100% due to rounding for

Value/Criteria	Description
Target Value	A hazard ratio of 0.68
Lower Reference Value	A hazard ratio of 0.83
Go	If there is ≥ 80% chance that the hazard ratio is ≤ 0.83* e.g. observed hazard ratio** ≤ 0.73
Stop	If there is <10% chance that the hazard ratio is ≤ 0.68* e.g. observed hazard ratio** ≥ 0.83

** Assuming 248 patients, allocation ratio = 1.0, 173 events
* Stop and Go correspond to lower-limit of 1-sided 90% CI and upper-limit of 1-sided 80% CI
The actual criteria will be driven by the stated probabilities so that if the observed data do not follow the assumptions, the GNG values will change

Software Solution



Single Proportion

Design Name: Design Name

Target Value (TV): 50%

Lower Ref. Value (LRV): 35%

User Interest Value (UIV): 25%

Number of Analyses: 2 ▾

Sample Size: 26

□ Sample Size Range (min:max:step)

DC_LRV: 80%

AR_TV: 10%

● Frequentist ○ Bayesian

Interim Type: Futility ▾

Interim Analysis	Info. Fraction	Interim AR_TV
# 1	0.46	10%

Stop Rule: Lalonde ▾ (Fixed / Lalonde / PPGreen / PPGreenAmber)

Rule Desc.

Num. of Simulations

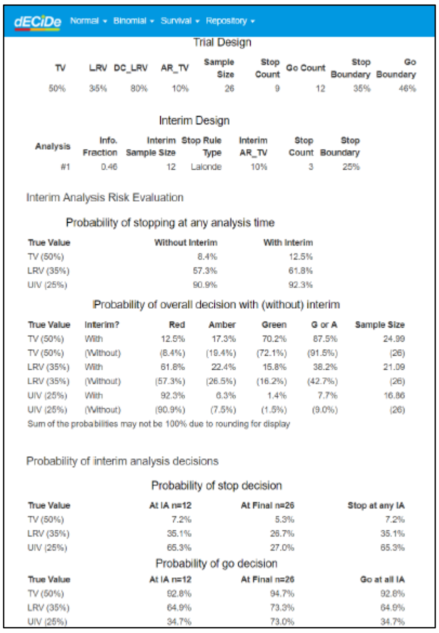
Seed

DC_LRV: Desired confidence that the treatment effect is better than the LRV. AR_TV: Acceptable risk that the treatment effect is better than the TV.

Output Summary — 79 Designs in Design Group: dECIDe ▾ — Maximum 100 designs per Design

	ID	Name	Sample Size
●	BSP_8121	Design Name	26
○	S2M_8111	Design Name	248
○	S2M_8110	Design Name	248
○	BSP_7967	Design Name	80
○	BSP_7547	Design Name	12
○	BSP_7546	Design Name	7

Preview



Trial Design

TV	LRV	DC_LRV	AR_TV	Sample Size	Stop Count	Go Count	Stop Boundary	Go Boundary
50%	35%	80%	10%	26	9	12	35%	46%

Interim Design

Analysis	Info. Fraction	Interim Sample Size	Stop Rule Type	Interim AR_TV	Stop Count	Stop Boundary
#1	0.46	12	Lalonde	10%	3	25%

Interim Analysis Risk Evaluation

Probability of stopping at any analysis time

True Value	Without Interim	With Interim
TV (50%)	8.4%	12.5%
LRV (35%)	57.3%	61.8%
UIV (25%)	90.9%	92.3%

Probability of overall decision with (without) interim

True Value	Interim?	Red	Amber	Green	G or A	Sample Size
TV (50%)	With	12.5%	17.3%	70.2%	87.5%	24.99
TV (50%)	(Without)	(8.4%)	(19.4%)	(72.1%)	(91.5%)	(26)
LRV (35%)	With	61.8%	22.4%	15.8%	38.2%	21.09
LRV (35%)	(Without)	(57.3%)	(26.5%)	(16.2%)	(42.7%)	(26)
UIV (25%)	With	92.3%	6.3%	1.4%	7.7%	16.86
UIV (25%)	(Without)	(90.9%)	(7.5%)	(1.5%)	(9.0%)	(26)

Sum of the probabilities may not be 100% due to rounding for display

Probability of interim analysis decisions

Probability of stop decision

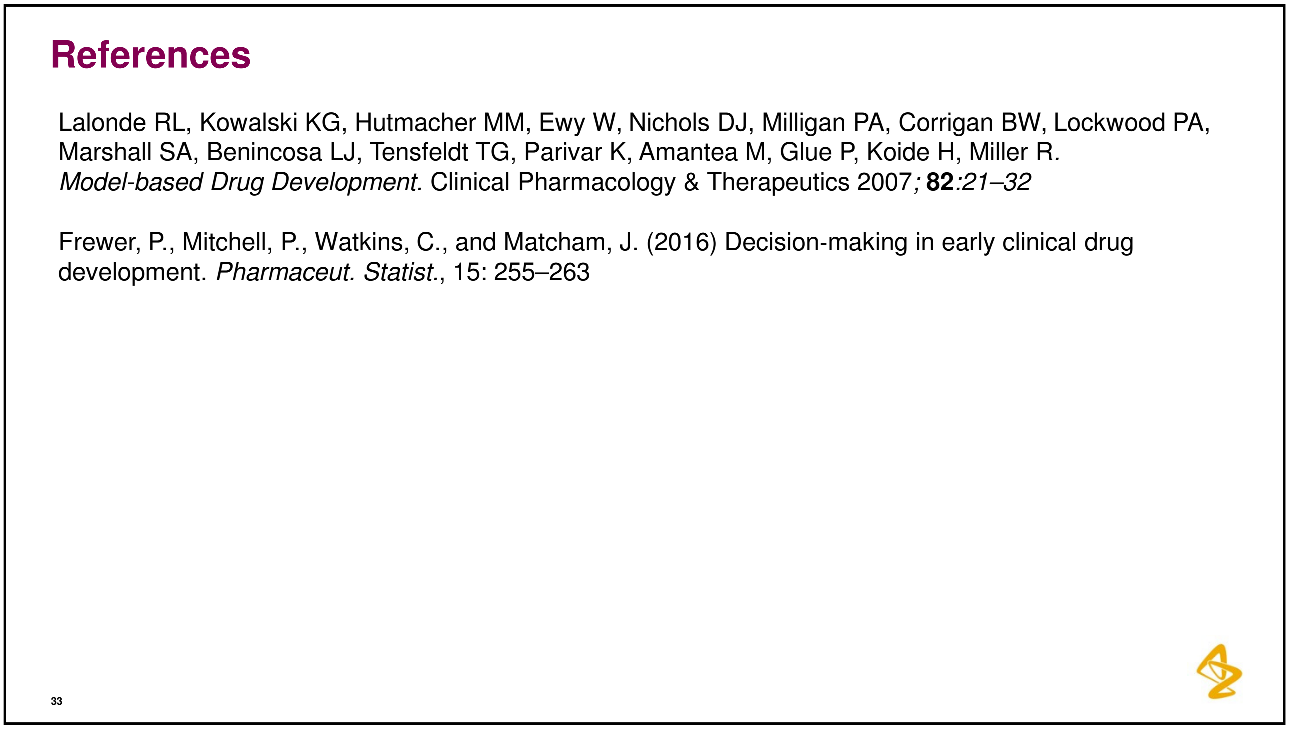
True Value	At IA n=12	At Final n=26	Stop at any IA
TV (50%)	7.2%	5.3%	7.2%
LRV (35%)	35.1%	26.7%	35.1%
UIV (25%)	65.3%	27.0%	65.3%

Probability of go decision

True Value	At IA n=12	At Final n=26	Go at all IA
TV (50%)	92.8%	94.7%	92.8%
LRV (35%)	64.9%	73.3%	64.9%
UIV (25%)	34.7%	73.0%	34.7%

Experience

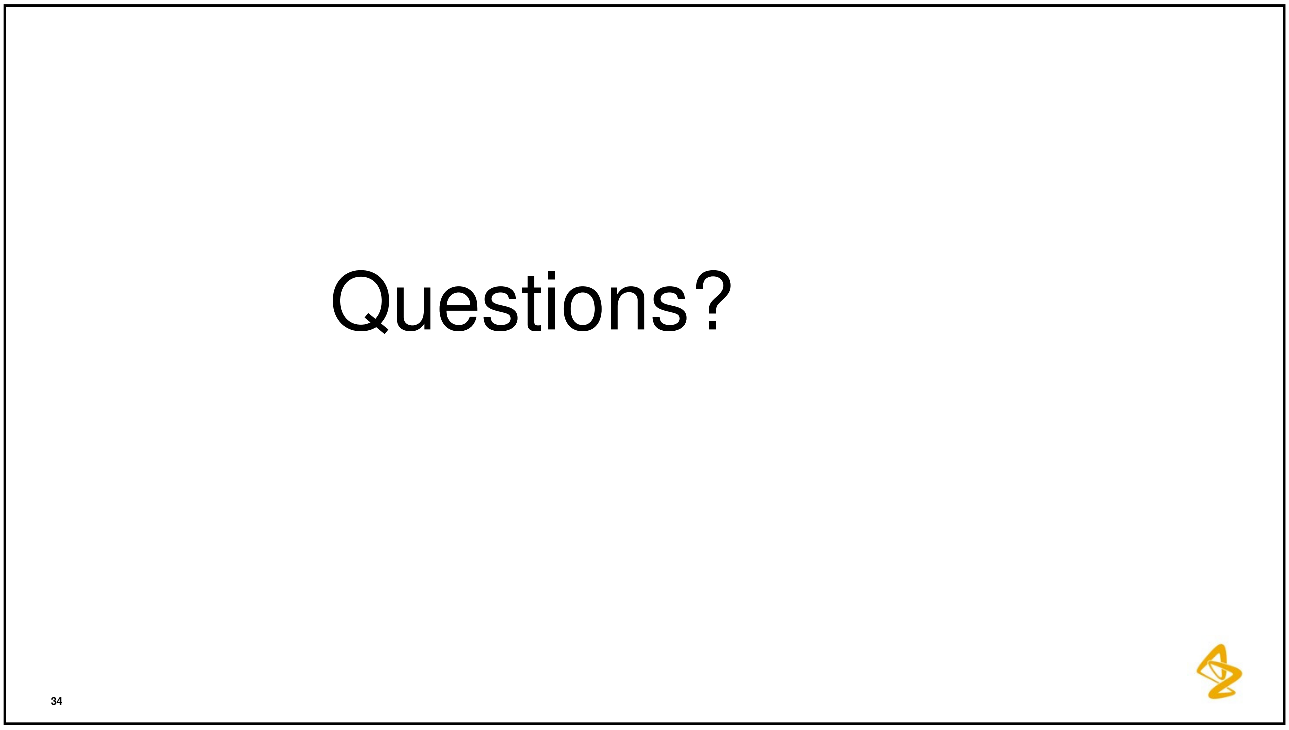
- This methodology is used throughout Early Clinical Development at AstraZeneca, teams are required to create prospective decision criteria using this approach
- Governance reviews and approves the decision criteria prospectively at the time of an investment decision
- Decision criteria are now produced routinely within the teams as part of the design of all studies
- Decisions made are based on trial data and the previously agreed decision criteria
- The role of the statistician in developing the decision criteria is key
  - evidence-base the TV and LRV
  - generate the operating characteristics of the decision
  - consult on how to improve operating characteristics and the use interim analyses to investigate decision timings.



References

Lalonde RL, Kowalski KG, Hutmacher MM, Ewy W, Nichols DJ, Milligan PA, Corrigan BW, Lockwood PA, Marshall SA, Benincosa LJ, Tensfeldt TG, Parivar K, Amantea M, Glue P, Koide H, Miller R. *Model-based Drug Development.* Clinical Pharmacology & Therapeutics 2007*;* **82***:21–32*

Frewer, P., Mitchell, P., Watkins, C., and Matcham, J. (2016) Decision-making in early clinical drug development. *Pharmaceut. Statist.*, 15: 255–263



Questions?