PFS in oncology trials: Some common statistical issues

Kevin J Carroll

Is an increase in PFS a benefit in and of itself, or merely a surrogate for survival?

"I suggest that nobody in the field of oncology really doubts that it is good to delay the growth of cancer. That is not really the question that we need to answer.

The real question is whether you can reliably measure TTP and, if you can, what does it mean?."

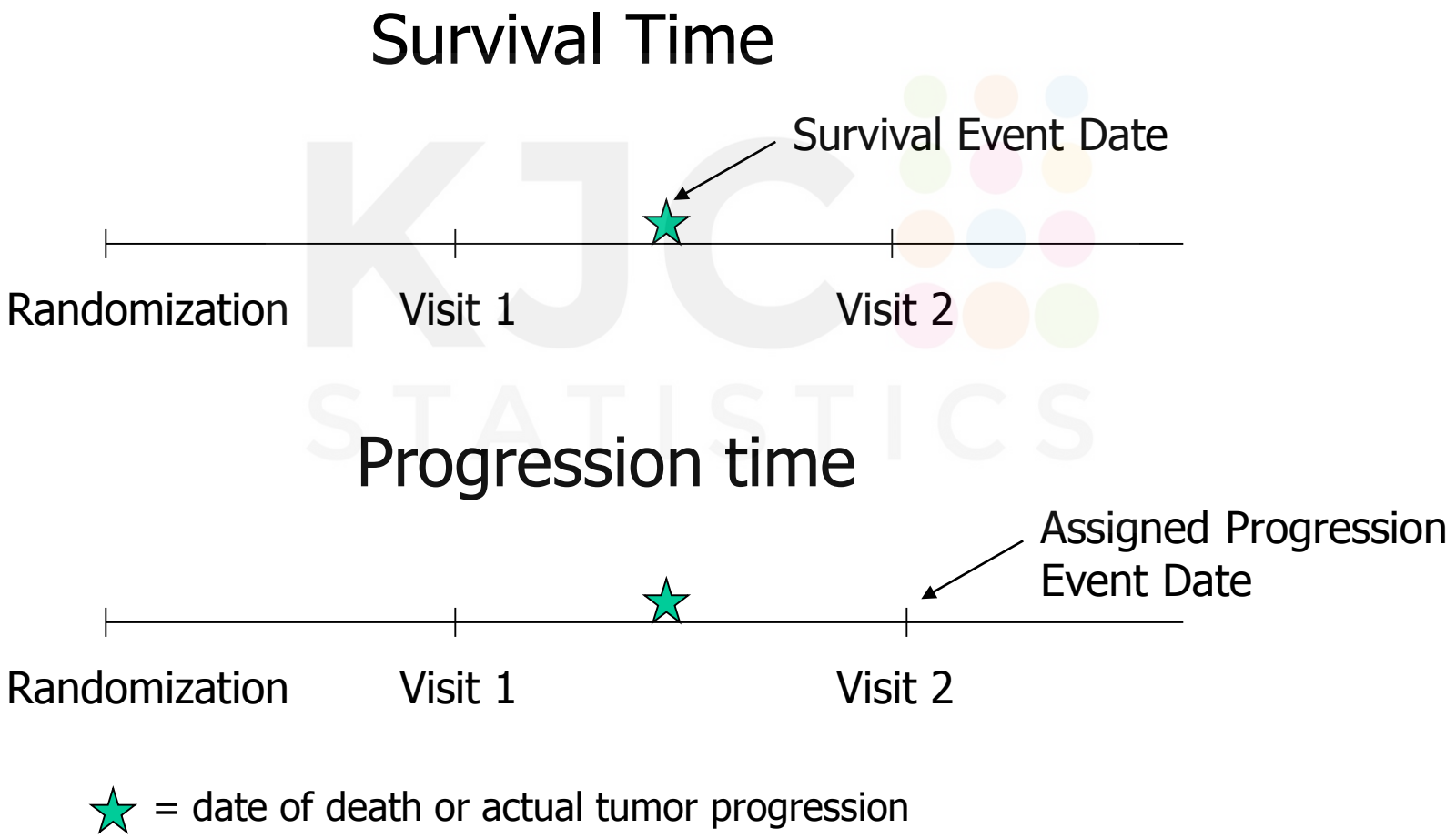
Dr Grant Williams, FDA, 2003

FDA ODAC: Endpoints in clinical cancer trials and endpoints in lung cancer clinical trials. 16th December 2003. http://www.fda.gov/ohrms/dockets/ac/03/transcripts/4009T1.DOC Transcript, page 34.

Some commonly cited PFS issues…

- Unlike survival, exact progression times are unknown, being interval censored between clinic visits.
- This can result in underestimating the treatment effect and, thus, result in reduced power.
- To avoid bias, tumor assessment frequency should be the same across study arms even when treatment cycles are of different lengths
- To be meaningful, the minimum interval between tumor assessments should be smaller than the expected treatment effect size
- Where possible, trials should be blinded – if blinding is not possible, then independent verification of tumor assessment data should be implemented
- PFS times should be censored at the time of dropout due to toxicity or the introduction of additional anti-cancer therapy

Actual progression event times are unknown in clinical trials



Type I error is increased if clinic visits are asymmetric and progression is assigned to the visit at which it was detected

Median PFS on E and C (HR=1)	Interval between visits on C (months)	Interval between visits on E (months)	Type I error (1-sided)
4	0.5	1	0.069
	1	2	0.152
6	1	1.5	0.050
	1	2	0.092
	2	3	0.090
9	1	1.5	0.040
	2	3	0.062
	3	4	0.061
12	1	2	0.050
	2	3	0.050
	3	4	0.050
	4	6	0.090

Trial with 508 events, sufficient to detect a true HR of 0.75 with 90% power, 2.5% 1 sided α

Even when clinic visits are symmetric, there is bias and a loss of power associated with assigning time of progression to the scheduled clinic visit at which it was detected

Hazard Ratio, θ	Median PFS on E (months)	Median PFS on C (months)	Interval between clinic visits, V, (months)	Log rank power[1] (%)	Relative increase in E to compensate for loss in power
0.667	6	4	0.5	87.8	1.07
			1	85.4	1.16
			2	80.0	1.34
			4	67.2	1.81
0.75	8	6	0.5	88.5	1.05
			1	86.9	1.11
			2	83.3	1.23
			4	75.0	1.51
0.80	12	9.6	0.5	89.1	1.03
			1	88.1	1.07
			2	85.9	1.14
			4	81.1	1.30

Trial originally powered at 90% (β=0.1) to detect HR size θ

Possible Implications For Trial Design and Planning

- Experimental (E) and Control (C) are to be compared in terms of PFS.
- Trial powered to detect an underlying hazard ratio, E:C, of size θ, with a 1-sided type I error rate of α and power 1-β so that a total of d events are required.
- Assuming exponential times to event with events rates of λ_E and λ_C, uniform patient recruitment over R months and a minimum follow-up period of F months, 2N patients are to be randomised on a 1:1 basis.
- Disease status assessed at every V months, say.
- For simplicity, assume that V is chosen such that $\frac{R}{V}$ and $\frac{F}{V}$ are both integer so that $\frac{F}{V}$ is the minimum and $\frac{R+F}{V}$ the maximum number of scheduled assessments per patient and a clinic visit always takes place at the end of the trial follow-up period.

Possible Implications For Trial Design and Planning

- Assigning the time of progression to the scheduled clinic visit at which it was detected introduces a diluting bias which reduces power:

$$E\left[\hat{\theta}\right] = \frac{V_C\left(1-e^{-\lambda_E V_E}\right)}{V_E\left(1-e^{-\lambda_C V_C}\right)} \neq \frac{\lambda_E}{\lambda_C}$$

- Thus, power not for θ but for $E\left[\hat{\theta}\right]$ or accept a reduction in power to

$$\phi^{-1}\left[\left(z_\alpha + z_\beta\right)\omega - z_\alpha\right] \quad \text{where} \quad \omega = \text{abs}\left(\frac{\ln\left(E\left[\hat{\theta}\right]\right)}{\ln(\theta)}\right)$$

Alternatively, size for θ and analyse data on an interval-censored basis

- If T_i denotes the observed PFS time, then $\overline{T} = \frac{\sum_{i=1}^{N} T_i}{d} = \frac{1}{\text{observed event rate}}$

  so that the asymptotically unbiased MLE of θ is given by

$$\breve{\theta} = \frac{\log\left(1 - \frac{V_E}{\overline{T}_E}\right)}{\log\left(1 - \frac{V_C}{\overline{T}_C}\right)}$$ with variance

$$\hat{V}ar\left[\log\left(\breve{\theta}\right)\right] = \frac{V_E^2}{d_E \overline{T}_E^2 \left\{\log\left(1 - \frac{V_E}{\overline{T}_E}\right)\right\}^2 \left(1 - \frac{V_E}{\overline{T}_E}\right)} + \frac{V_C^2}{d_C \overline{T}_C^2 \left\{\log\left(1 - \frac{V_C}{\overline{T}_C}\right)\right\}^2 \left(1 - \frac{V_C}{\overline{T}_C}\right)}$$

Interval censored analysis results in unbiased treatment effect estimates and thus maintains power

Hazard Ratio, θ	Median PFS on E (months)	Median PFS on C (months)	Interval between clinic visits, V, (months)	$E[\hat{\theta}]$	$\hat{\theta}$	$\breve{\theta}$	SE $\ln(\breve{\theta})$
0.667	6	4	0.5	0.677	0.679	0.672	0.1438
			1	0.686	0.681	0.669	0.1418
			2	0.705	0.69	0.664	0.1388
			4	0.74	0.718	0.668	0.1428
0.75	8	6	0.5	0.755	0.751	0.746	0.1483
			1	0.761	0.759	0.75	0.1438
			2	0.771	0.764	0.748	0.1376
			4	0.792	0.782	0.751	0.1433
0.8	12	9.6	0.5	0.803	0.804	0.802	0.1425
			1	0.806	0.808	0.803	0.144
			2	0.811	0.811	0.802	0.1377
			4	0.822	0.817	0.799	0.1474

HR estimates resulting from 1000 trial simulations with 200 patients (100 per arm) in which all patients achieve an event

Interval censored analysis is readily available for irregular and asymmetric clinic visit times[1]

- Construct survivor estimate using Turnbull (EM) algorithm.
- Use survivor estimate to calculate numbers of deaths, d_k, and numbers at risk, r_k, over time.
- Compare treatment groups via long rank test on d_k and r_k.

- Or…analyse data via accelerated failure time approach (e.g. via PROC LIFEREG).

1,…refs

Turnbull's (EM) algorithm for interval censored analyses

- Data are $(L_i, U_i]$, for i=1,…,n patients with $U_i=\infty$ meaning the i^{th} patient is right censored at L_i.
- $t_0, t_1,\ldots, t_m$ = set of time-points that includes all the points L_i and U_i, i=1,…,n.
- For each patient define $a_{ik}=1$ if $(t_{k-1}, t_k]$, k=1,…,m, is contained in the interval $(L_i, U_i]$, and 0 otherwise.
- Let $S(t_k)$ be an initial estimate of the survivor function. Update $S(t_k)$ as follows:
  - Calculate $p_k = S(t_{k-1}) - S(t_k)$, k=1,…,m
  - Estimate the number of events which occurred at t_k by

$$d_k = \sum_{i=1}^{n} \frac{\alpha_{ik} p_k}{\sum_{j=1}^{m} \alpha_{ik} p_j} \quad \text{and number at risk by} \quad r_k = \sum_{j=k}^{m} d_j$$

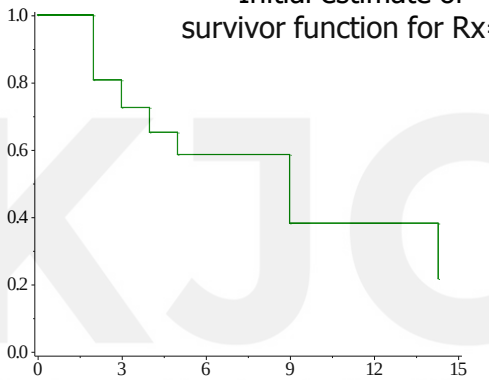
  - Use d_k and r_k to provide an updated product-limit estimate of the survivor function. Repeat until convergence.

Turnbull's (EM) algorithm for interval censored analyses

Data

Rx	Pt	PFS
0	1	(0, 2]
0	2	(0, 4]
0	3	(3, 5]
0	4	≥ 4
0	5	≥ 9
1	1	(0, 3]
1	2	(1, 5]
1	3	(2, 7]
1	4	≥ 3
1	5	(8, 11]
1	6	≥ 13

Initial estimate of survivor function for Rx=0



a_{ik} for Rx=0

Pt	0-2	2-3	3-4	4-5	5-9	>9
#1	1	0	0	0	0	0
#2	1	1	1	0	0	0
#3	0	0	1	1	0	0
#4	0	0	0	1	1	1
#5	0	0	0	0	0	1

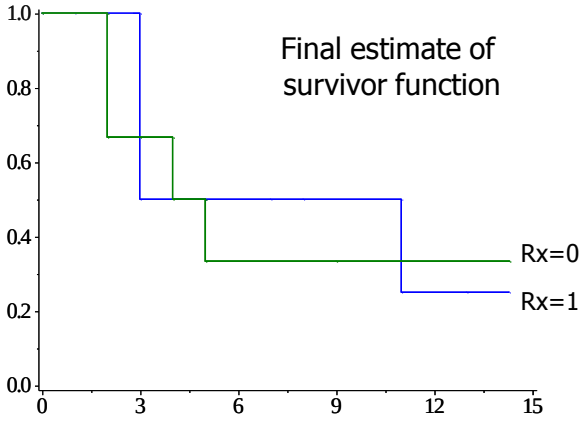
$$\frac{\alpha_{ik}p_k}{\sum_{j=1}^{m}\alpha_{ik}p_j}$$

Pt	0-2	2-3	3-4	4-5	5-9	>9
#1	1.000	0	0	0	0	0
#2	0.667	0	0.333	0	0	0
#3	0	0	0.500	0.500	0	0
#4	0	0	0	0.333	0	0.667
#5	0	0	0	0	0	1.000

number of events, number at risk
product limit survivor estimate

	d	r	s
(0, 2]	1.67	5.00	1.000
(2, 3]	0	3.33	0.667
(3, 4]	0.83	3.33	0.667
(4, 5]	0.83	2.50	0.500
(5, 9]	0	1.67	0.333
>9	1.67	1.67	0.333

Final estimate of survivor function

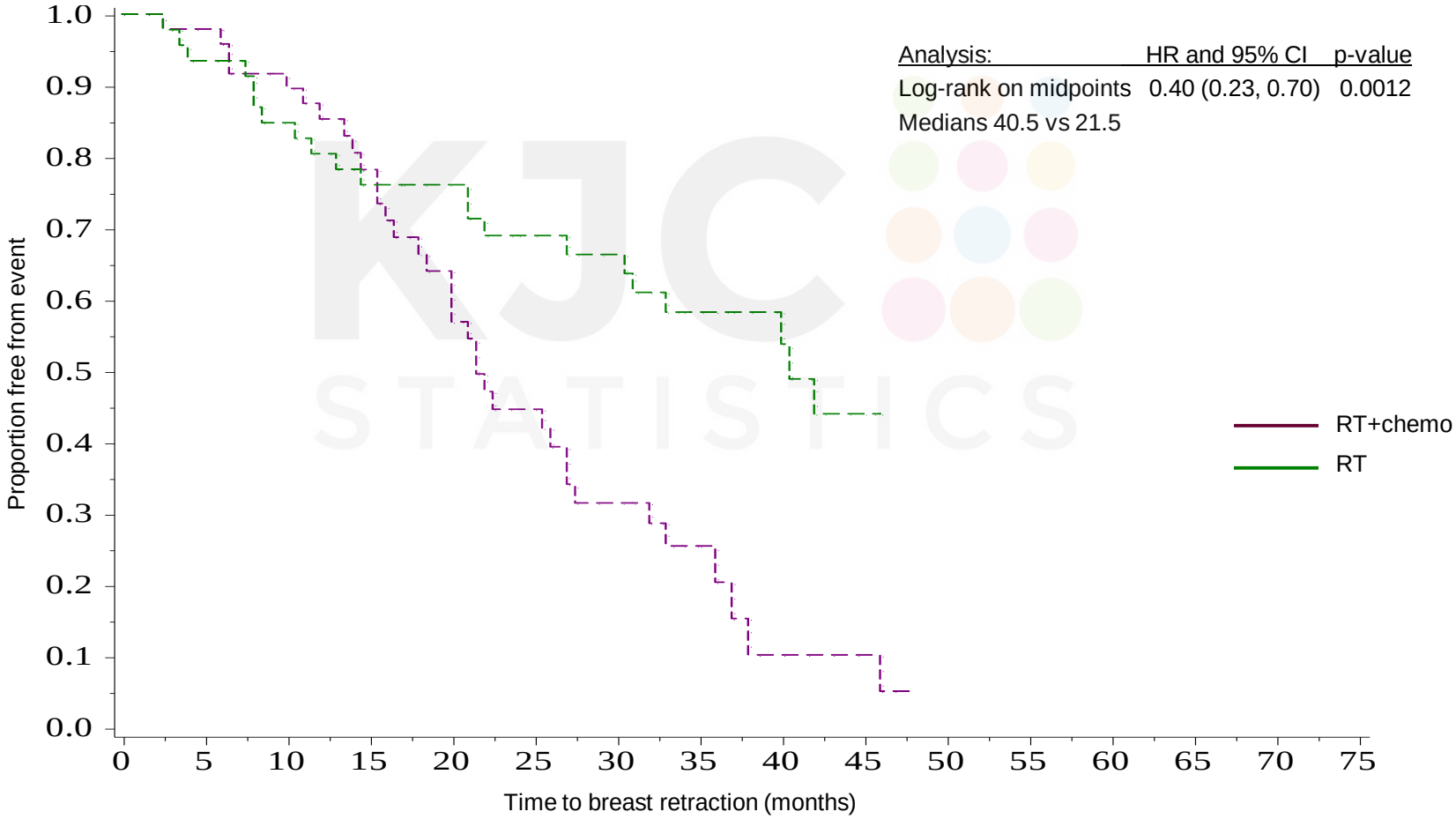


Breast cosmesis data

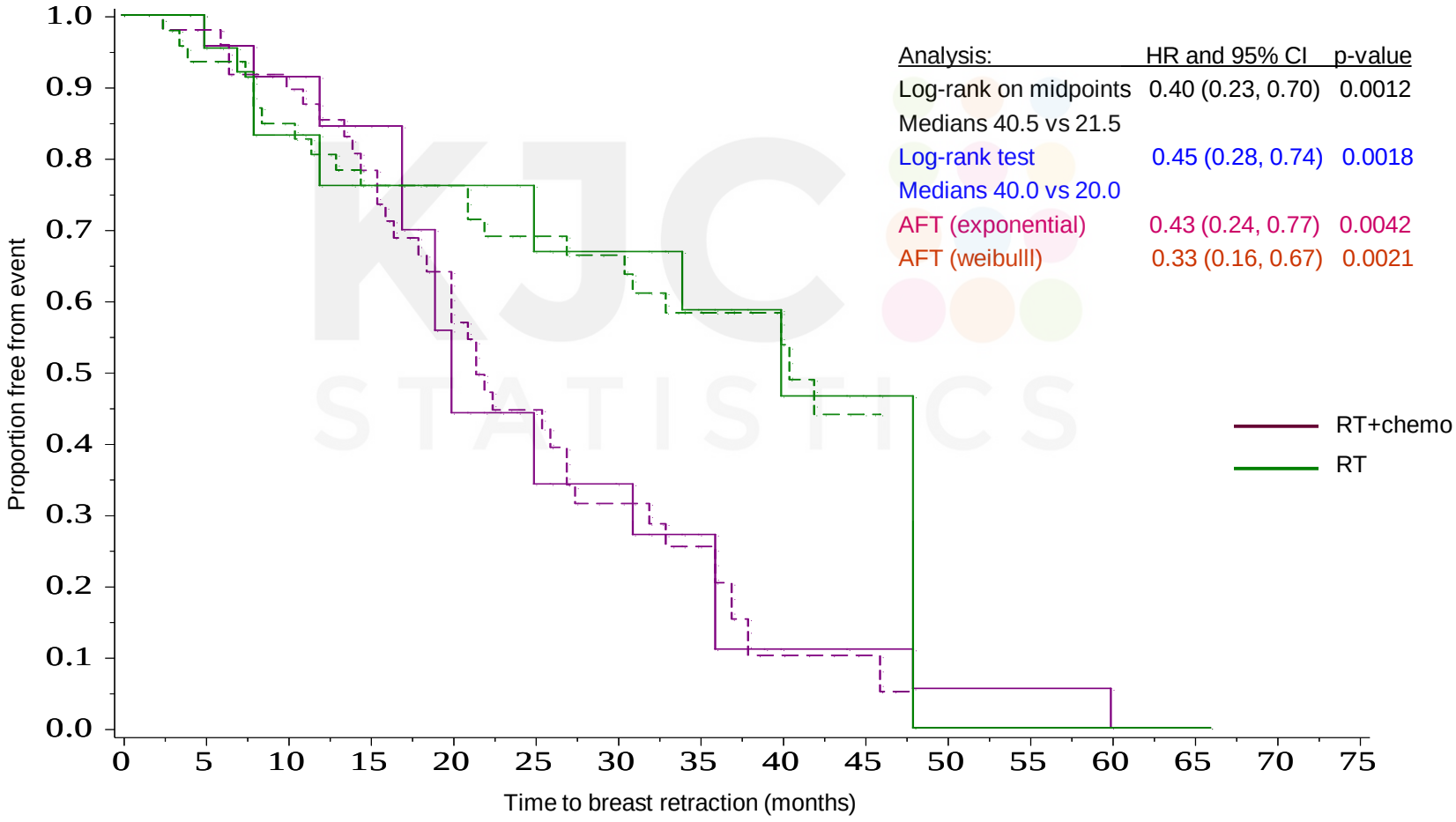
RT+chemo		RT alone	
(0 , 5]	(16 , 24]	(0 , 5]	≥ 32
(0 , 22]	(16 , 60]	(0 , 7]	≥ 33
(4 , 8]	(17 , 23]	(0 , 8]	≥ 34
(4 , 9]	(17 , 26]	(4 , 11]	≥ 36
(5 , 8]	(17 , 27]	(5 , 11]	≥ 36
(8 , 12]	(18 , 24]	(5 , 12]	(36 , 44]
(8 , 21]	(18 , 25]	(6 , 10]	(36 , 48]
(10 , 17]	(19 , 32]	(7 , 14]	≥ 37
(10 , 35]	≥ 21	(7 , 16]	≥ 37
≥ 11	(22 , 32]	(11 , 15]	≥ 37
≥ 11	≥ 23	(11 , 18]	(37 , 44]
(11 , 13]	(24 , 30]	≥ 15	≥ 38
(11 , 17]	(24 , 31]	≥ 17	≥ 40
(11 , 20]	(30 , 34]	(17 , 25]	≥ 45
(12 , 20]	(30 , 36]	(17 , 25]	≥ 46
≥ 13	≥ 31	≥ 18	≥ 46
≥ 13	≥ 32	(18 , 26]	≥ 46
≥ 13	(32 , 40]	(19 , 35]	≥ 46
(13 , 39]	≥ 34	≥ 22	≥ 46
(14 , 17]	≥ 34	≥ 24	≥ 46
(14 , 19]	≥ 35	≥ 24	≥ 46
(15 , 22]	(35 , 39]	(25 , 37]	≥ 46
(16 , 20]	(44 , 48]	(26 , 40]	
(16 , 24]	≥ 48	(27 , 34]	

Finkelstein, D.M., 1986. A proportional hazards model for interval-censored failure time data. Biometrics 42, 845–854.

Log-rank analysis on midpoints



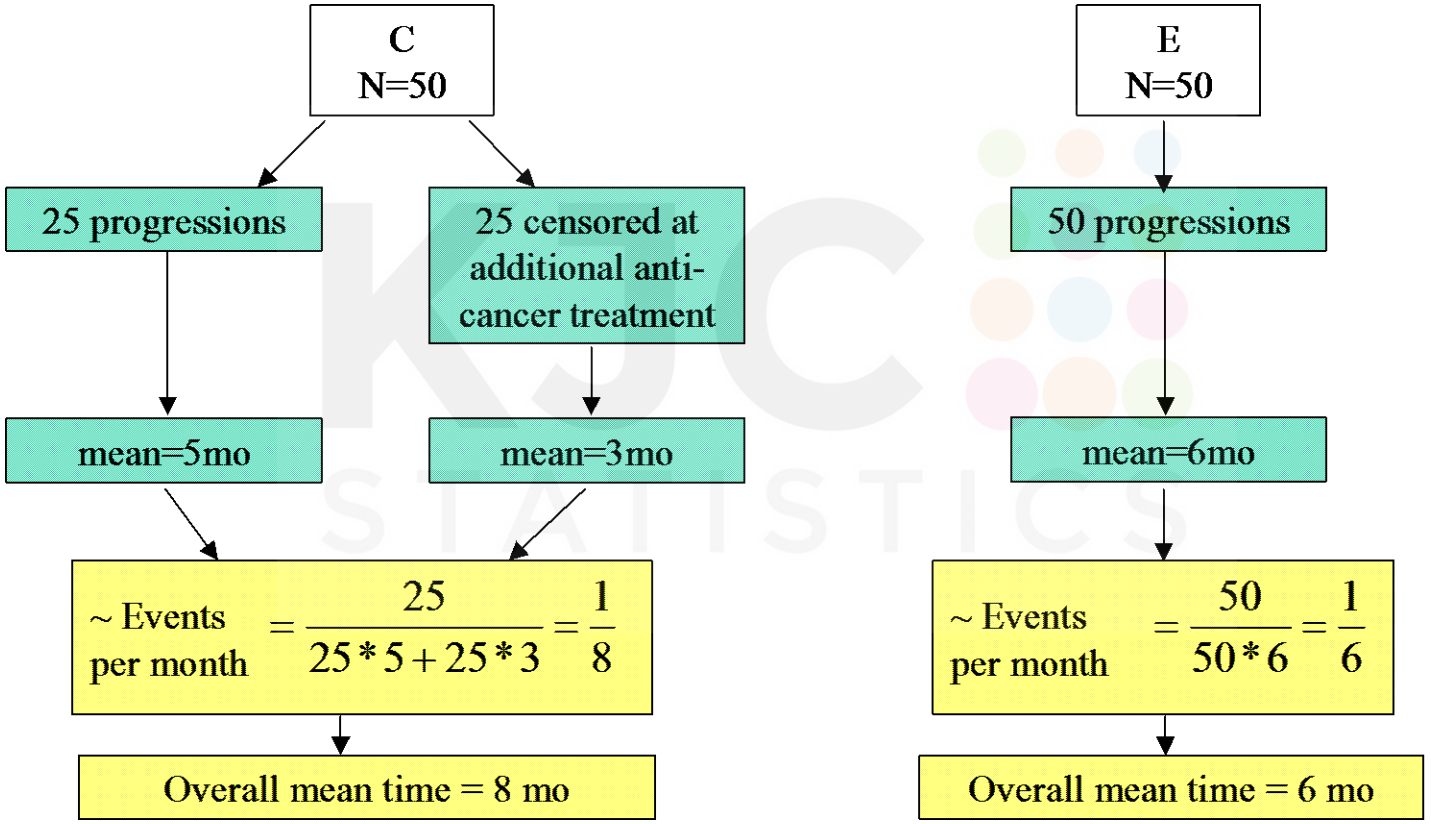
Interval censored analysis



Conservative censoring rules should be applied to progression data

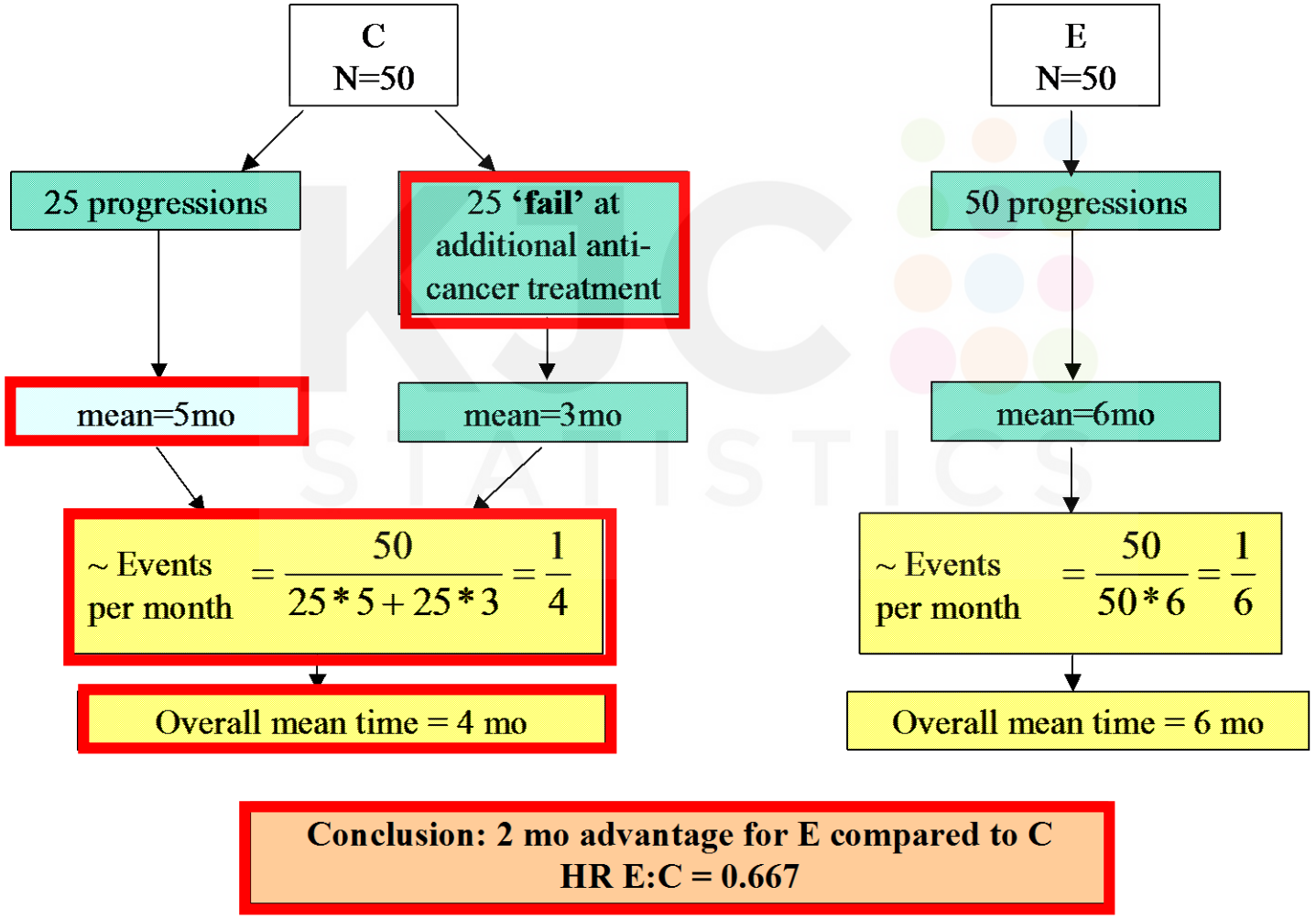
Situation	Date	Censored
No baseline or on-treatment tumor assessment	Randomization	Yes
Treatment discontinuation for PD	Last on-study assessment	No
Treatment discontinuation for other than PD or death	Last on-study assessment without PD	Yes
New anticancer treatment started	Last on-study assessment before start of new treatment	Yes
Death before first PD assessment	Death	No
Death after ≥1 assessment but before PD	Last on-study assessment without PD	Yes
Patients still on treatment	Last on-study assessment without PD	Yes

Censoring PFS on an informative event is a bad, bad idea



Conclusion: 2 mo advantage for C compared to E
HR E:C = 1.33

Censoring PFS on an informative event is a bad, bad idea



A consistent ITT philosophy is the only, common sense way to go

- Survival: No sensible person will accept a primary analysis of survival that includes only those deaths that occurred on treatment
- ITT philosophy well established for survival to ensure the true mortality benefit or disbenefit associated with the policy of treatment is captured.
- So why, in the same clinical trial, do we have one approach for survival and another approach for progression free survival?
- Makes no sense.
- Like survival, patients should be followed for tumour assessment and confirmation of progression irrespective of whether they stop taking randomised treatment for any reason or whether other anti-cancer treatments are initiated.
- Analyse and interpret PFS and survival on the same ITT basis.

Potential progression assessment and bias

- Open trials or blinded trials with drugs having distinct pharmacological effects may be subject to some degree of bias.

- Acquisition bias: Where investigators acquire more scans and/or assess patients earlier than scheduled
  - Address in trial design by requiring a radiographic scan at a fixed time point, irrespective of clinical indication.

- Assignment bias: Where the investigators read radiographic scan more positively or negatively, depending on the (assumed) randomised treatment.
  - Independent review of radiographic data may be requested

Impact of assignment bias
N=1000 per arm, 1 year event rate=30%

Treatment A			Treatment B				
unblind	pos neg	neg pos	unblind	pos neg	neg pos	HR	p
0%	10%	10%	0%	10%	10%	1.00	1.000
10%	10%	10%	10%	10%	10%	1.00	1.000
20%	10%	10%	20%	10%	10%	1.00	1.000
30%	10%	10%	30%	10%	10%	1.00	1.000
40%	10%	10%	40%	10%	10%	1.00	1.000
50%	10%	10%	50%	10%	10%	1.00	1.000

No difference in pos-neg or neg-pos
reassignment rates between treatment = no bias

Impact of assignment bias
N=1000 per arm, 1 year event rate=30%

Treatment A			Treatment B				
unblind	pos neg	neg pos	unblind	pos neg	neg pos	HR	p
0%	20%	10%	0%	5%	10%	1.00	1.000
10%	20%	10%	10%	5%	10%	0.98	0.827
20%	20%	10%	20%	5%	10%	0.97	0.662
30%	20%	10%	30%	5%	10%	0.95	0.514
40%	20%	10%	40%	5%	10%	0.93	0.385
50%	20%	10%	50%	5%	10%	0.92	0.279

Differential pos-neg reassignment
has relatively little impact

Impact of assignment bias
N=1000 per arm, 1 year event rate=30%

Treatment A			Treatment B				
unblind	pos neg	neg pos	unblind	pos neg	neg pos	HR	p
0%	10%	20%	0%	10%	5%	1.00	1.000
10%	10%	20%	10%	10%	5%	1.04	0.610
20%	10%	20%	20%	10%	5%	1.09	0.311
30%	10%	20%	30%	10%	5%	1.13	0.130
40%	10%	20%	40%	10%	5%	1.17	0.045
50%	10%	20%	50%	10%	5%	1.22	0.013

…but neg-pos reassignment matters more

Impact of assignment bias
N=1000 per arm, 1 year event rate=7.5%

Treatment A			Treatment B				
unblind	pos neg	neg pos	unblind	pos neg	neg pos	HR	p
0%	20%	10%	0%	5%	10%	1.00	1.000
10%	20%	10%	10%	5%	10%	0.99	0.927
20%	20%	10%	20%	5%	10%	0.97	0.862
30%	20%	10%	30%	5%	10%	0.97	0.801
40%	20%	10%	40%	5%	10%	0.96	0.746
50%	20%	10%	50%	5%	10%	0.95	0.695

Again, large differences in pos-neg reassignment have relatively little impact

Impact of assignment bias
N=1000 per arm, 1 year event rate=7.5%

Treatment A			Treatment B				
unblind	pos neg	neg pos	unblind	pos neg	neg pos	HR	p
0%	10%	20%	0%	10%	5%	1.00	1.000
10%	10%	20%	10%	10%	5%	1.18	0.269
20%	10%	20%	20%	10%	5%	1.36	0.036
30%	10%	20%	30%	10%	5%	1.51	0.003
40%	10%	20%	40%	10%	5%	1.66	0.000
50%	10%	20%	50%	10%	5%	1.80	0.000

… but similar sized differences in neg-pos reassignment matter more

Casodex Early Prostate Cancer programme: Retrospective re-evaluation of progression outcomes in over 1450 patients

- 8113 patient randomised double blind trial
- All 339 patients with a PFS event determined by positive bone scan or x-ray
- Plus a Random Sample of 1120 patients without a PFS event
  - 2 year bone scan or x-ray retrieved for re-evaluation.
  - N=1120 to estimate the negative-positive reclassification rate to within $\pm$2.5% with 90% confidence.
- Baseline scans retrieved for all above patients.

No evidence of investigator assignment bias in the Casodex EPC programme

		Re-evaluation outcome*	
		+	-
Casodex Investigator evaluation	**+ (n=95)**	**76.8% (n=73)**	**23.2% (n=22)**
	- (n=474)	**6.5% (n=31)**	**93.5% (n=443)**
placebo Investigator evaluation	**+ (n=180)**	**80.0% (n=144)**	**20.0% (n=36)**
	- (n=381)	**5.8% (n=22)**	**94.2% (n=359)**

- No difference in bone scan reclassification rates between treatment groups
  - Approx 6 % negative to positive reclassification
  - Approx 20% positive to negative reclassification
- Analyses based on re-evaluation outcomes continued to show a significant treatment effect.
  - Investigator assessment: RR = 0.63, p<0.0001
  - Re-evaluation: RR = 0.75, p<0.0001

*conditional on a determination being made

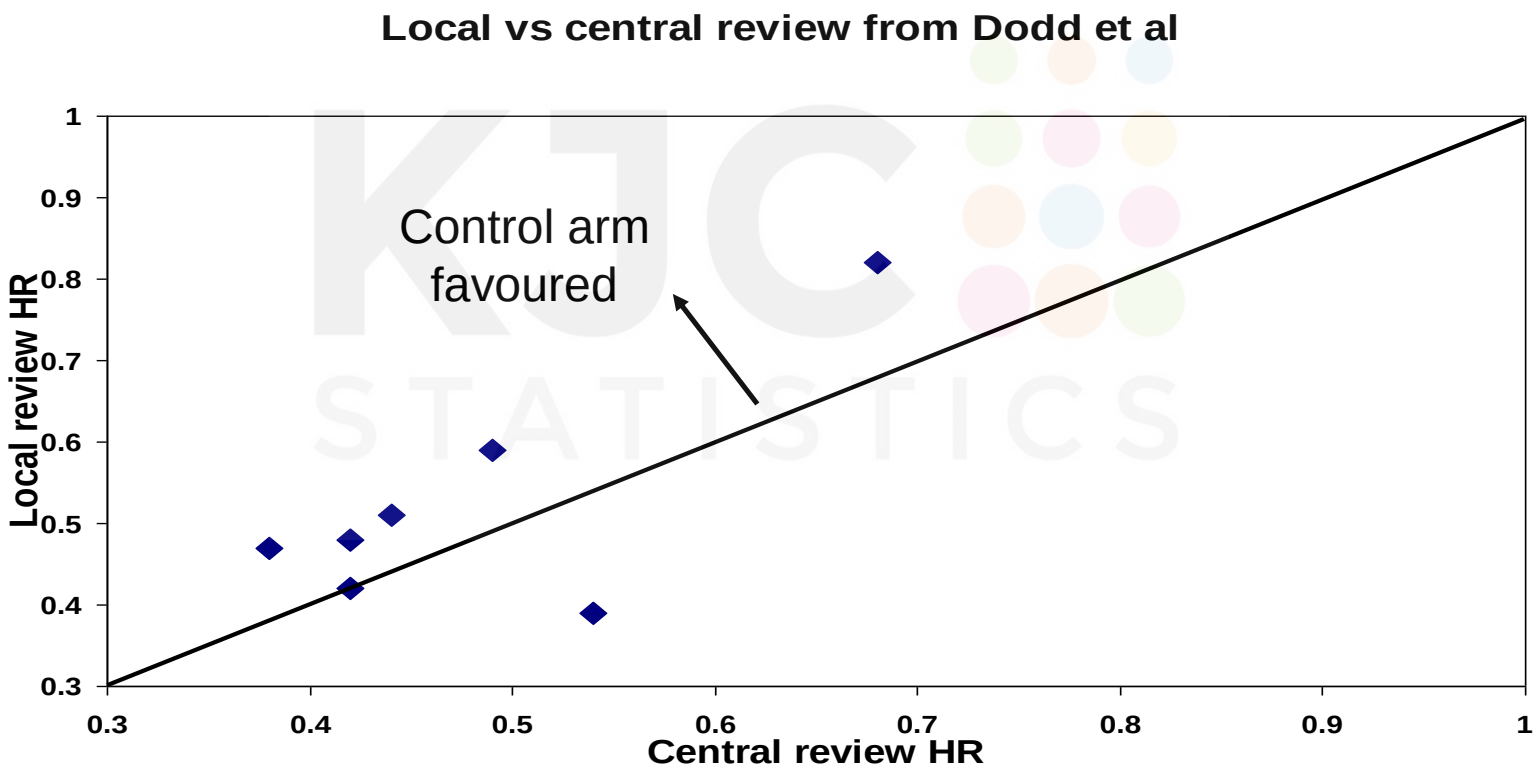
Summary of the Casodex EPC experience

- Large scale re-evaluation of radiographic outcomes in the Casodex EPC program did not change the result.
- The Casodex experience indicates that in randomised, blinded trials the investigator evaluation of the patient is a reliable basis on which to compare treatments.
- Consistent with findings by Dodd[1].
- The re-evaluation exercise took 1.5 yrs to complete at a cost of $5m.

1: Dodd L et al JCO 2008;26:3791-96

Little evidence investigators favour the experimental arm even in open-label trials

Data from 6 trials (7 treatment comparisons) of which 4 were open-label



Dodd L et al JCO 2008;26:3791-96
Sunitinib (renal), sorafenib (renal), bevacizumab (breast and renal), Panitumumab (CRC) & lapatinib (breast)

Summary

- Too much fuss around PFS from a statistical design and analysis perspective.
- Plan for and execute an interval censored analysis.
- Describe the treatment effect in terms of the HR and 95% CI – do not rely on medians.
- Do not censor on informative events like dropout due to AE or additional anti-cancer treatment.
- Execute an ITT approach and analyse PFS and survival on the same basis.
- Independent evaluation of a sample of progression data only warranted if:
  - (i) open trial or strong evidence that pharmacologic AEs may compromise the blind.
  - (ii) possibility of high unblinding rate (e.g >30%), with a low true event rate (e.g <10%) and the neg-pos reassignment rate differing between treatments.