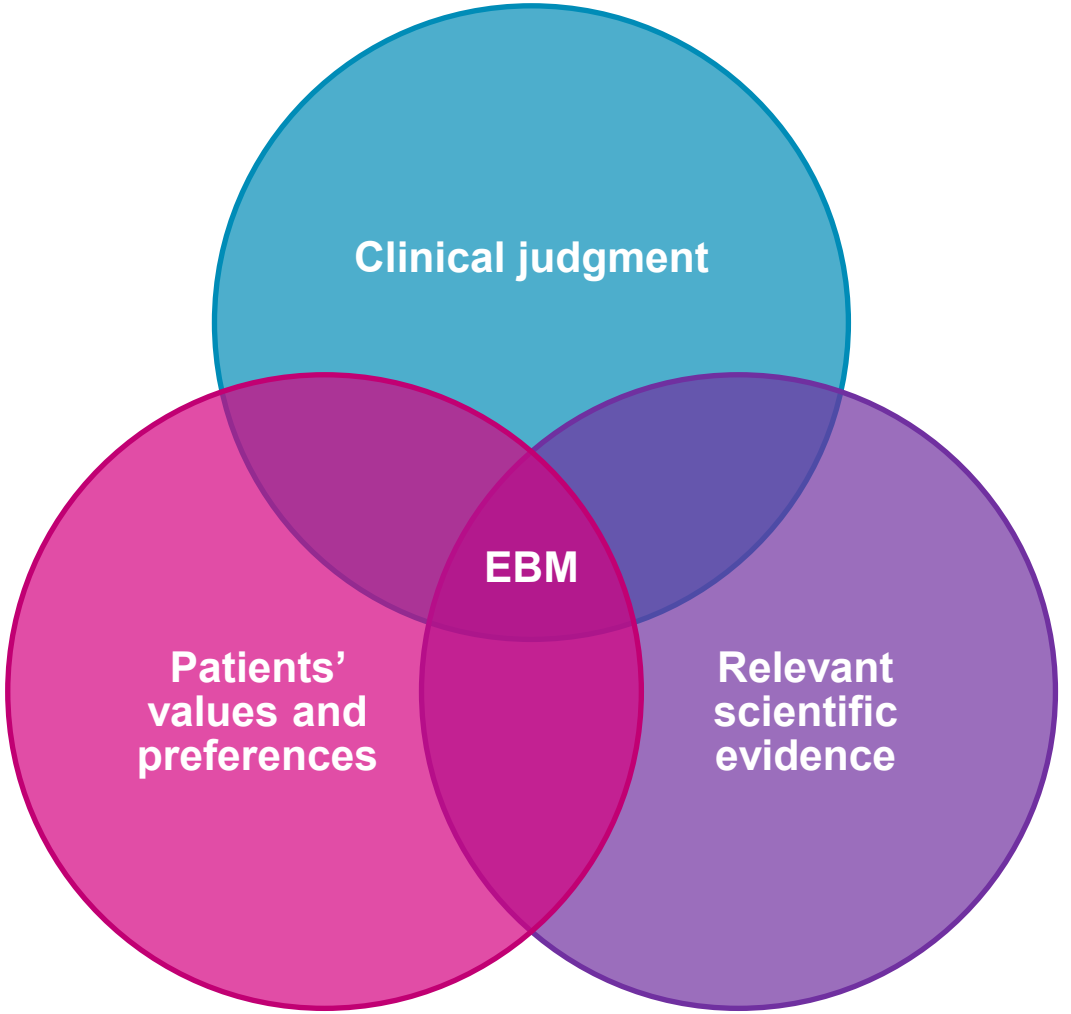
Design of Clinical Trials

Evidence-Based Medicine (EBM)

What is EBM?

Not all evidence is created equally, and a hierarchy guides clinical decision making.

Evidence alone is not enough; there needs to be a balance between the risks and benefits in the context of patient values and preferences.



Adapted from Makam AN & Nguyen OK. Circulation. 2017;135(2):180–195.

Types of Studies in Clinical Research

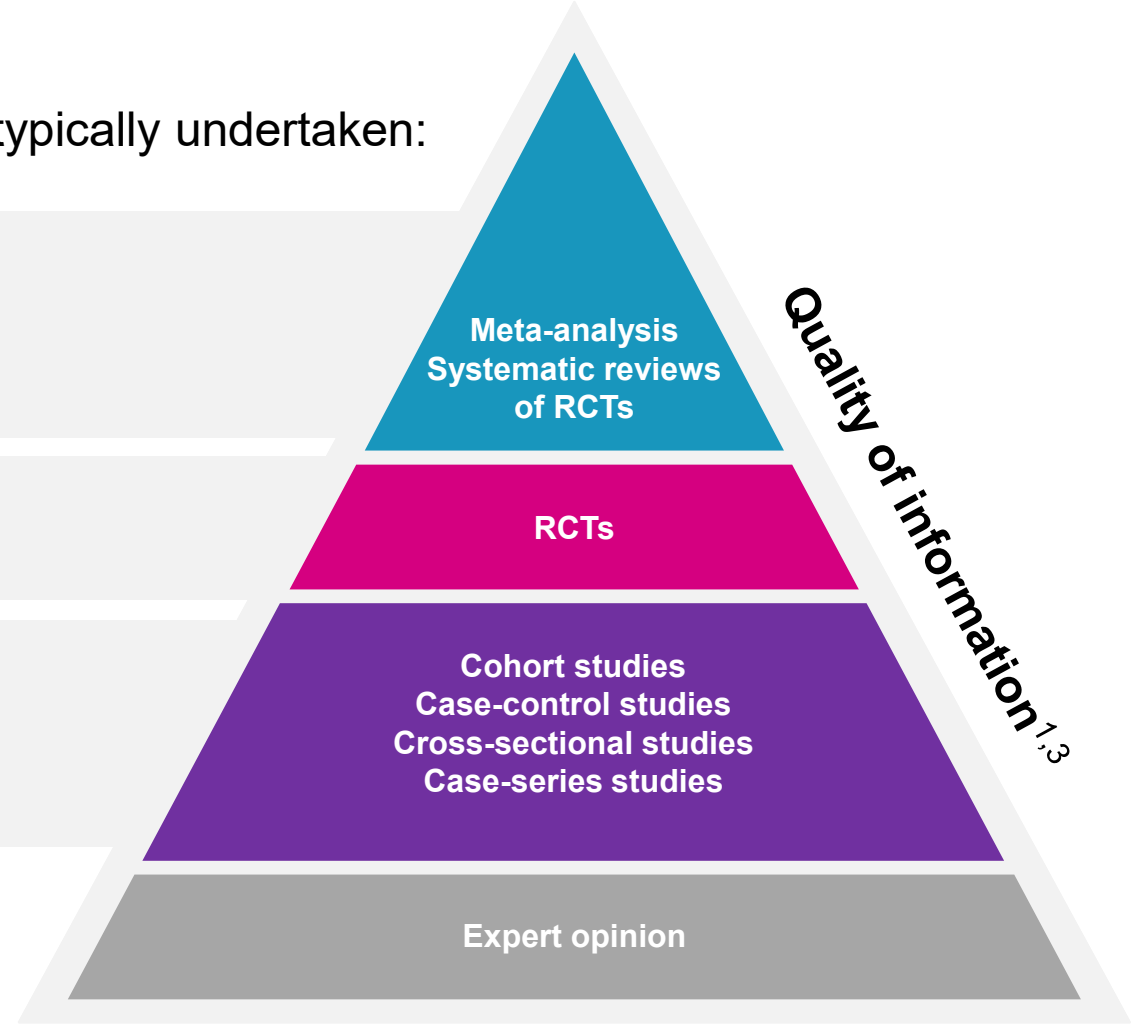
Hierarchy of evidence

In the field of clinical research, three types of study are typically undertaken:

Critical appraisals[1] *For more information, click here*

Experimental[2] *For more information, click here*

Non-experimental (observational)*[2]



This hierarchy is solely a guide; meta-analyses and systematic reviews often only provide the highest level of evidence when conducted on RCTs.[4] *Please refer to slide notes for additional information.

1. Evans DJ. Clin Nurs. 2003;12(1):77–84; **2.** Noordzij M et al. Nephron Clin Pract. 2009;113(3):218–221; **3.** Oxford Centre for Evidence-Based Medicine: Levels of Evidence (March 2009). Available at: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009 [Accessed Oct 21] **4.** Hassan Murad M et al. Evid Based Med. 2016;21(4):125–127.

Plan

- Overview of different phases of trial
- Randomisation and alternatives
- Blinding
- Types of radnomised trial
- Endpoints
- Sample size
- Some miscellaneous stuff

Phases of trial

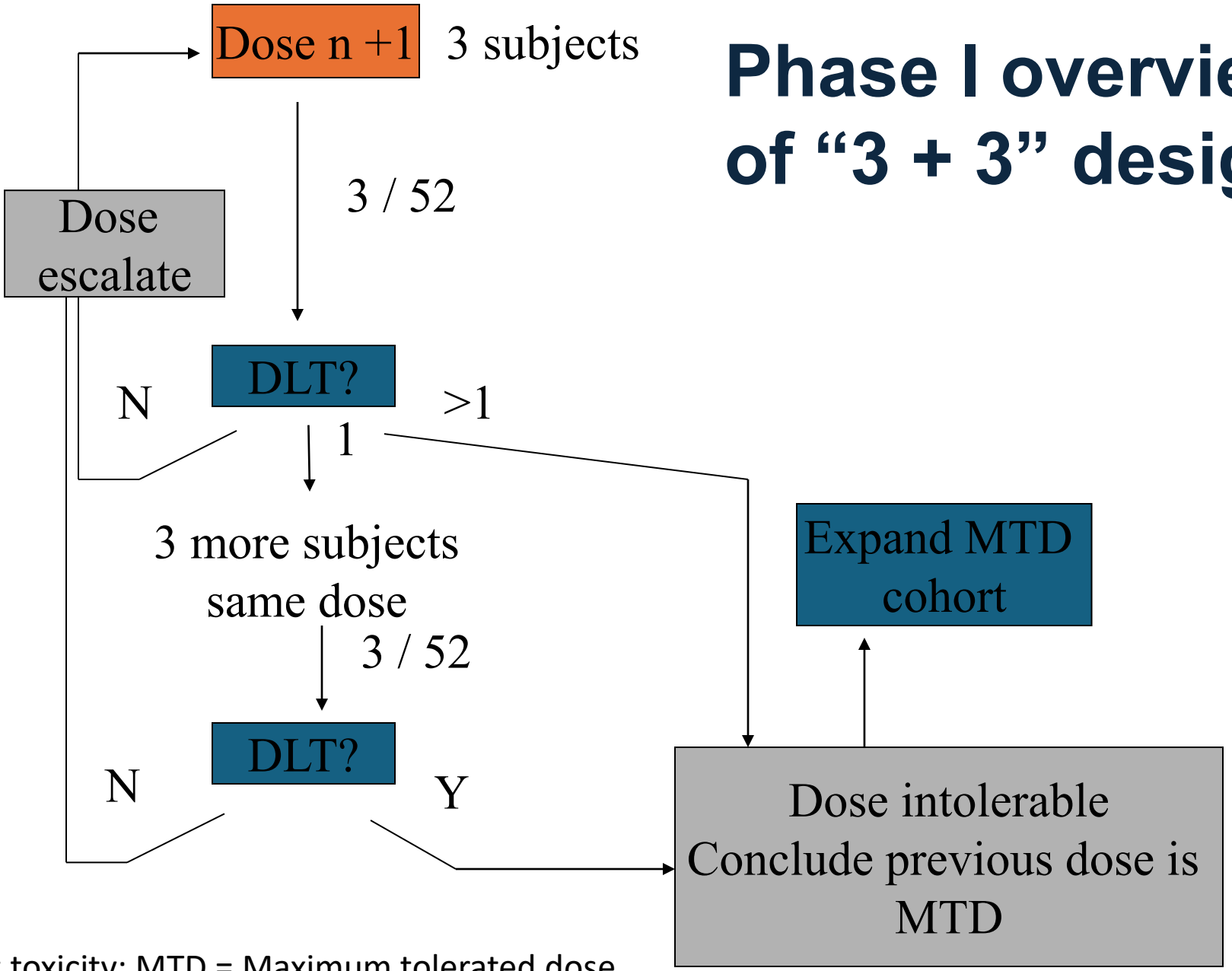
I - doses

II - diseases

III - confirmation

(IV - real-world)

Phase I overview of “3 + 3” design

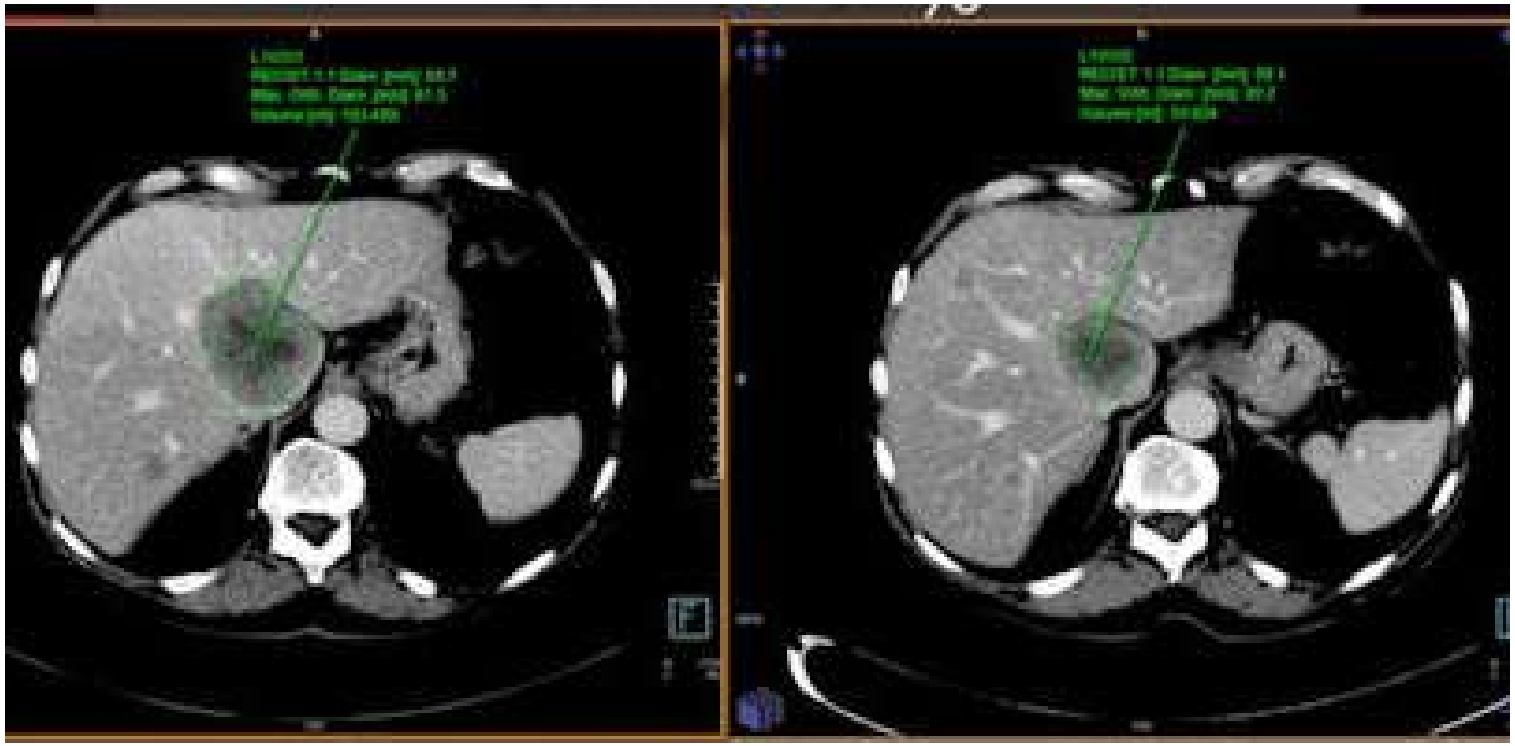


DLT = dose limiting toxicity; MTD = Maximum tolerated dose

Phase II

Activity screening

- Using dose defined by phase I
- Limited tumour types
- “Measurable” tumours



Result = proportion of patients who have signficant tumour shrinkage

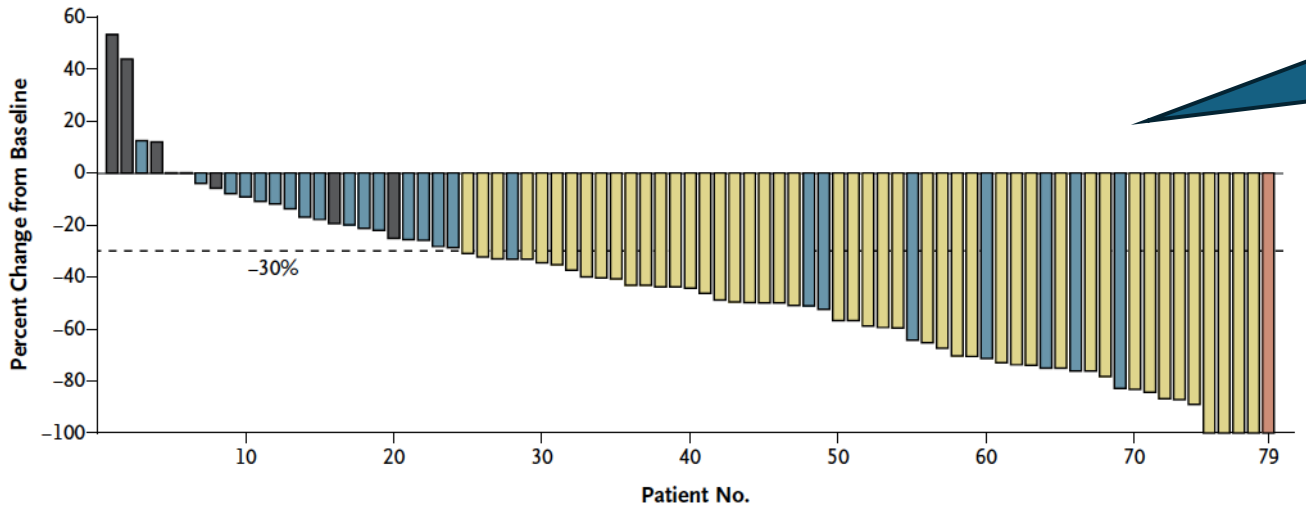
The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 OCTOBER 28, 2010 VOL. 363 NO. 18

Anaplastic Lymphoma Kinase Inhibition in Non–Small-Cell Lung Cancer

Eunice L. Kwak, M.D., Ph.D., Yung-Jue Bang, M.D., Ph.D., D. Ross Camidge, M.D., Ph.D., Alice T. Shaw, M.D., Ph.D., Benjamin Solomon, M.B., B.S., Ph.D., Robert G. Maki, M.D., Ph.D.,

A **Percent Change in Tumor Burden**



Is this a controlled trial?

2 – 7% of NSCLC has this mutation

Phase III trials

Alternatives to randomized phase III in evidence generation

- Historical controls

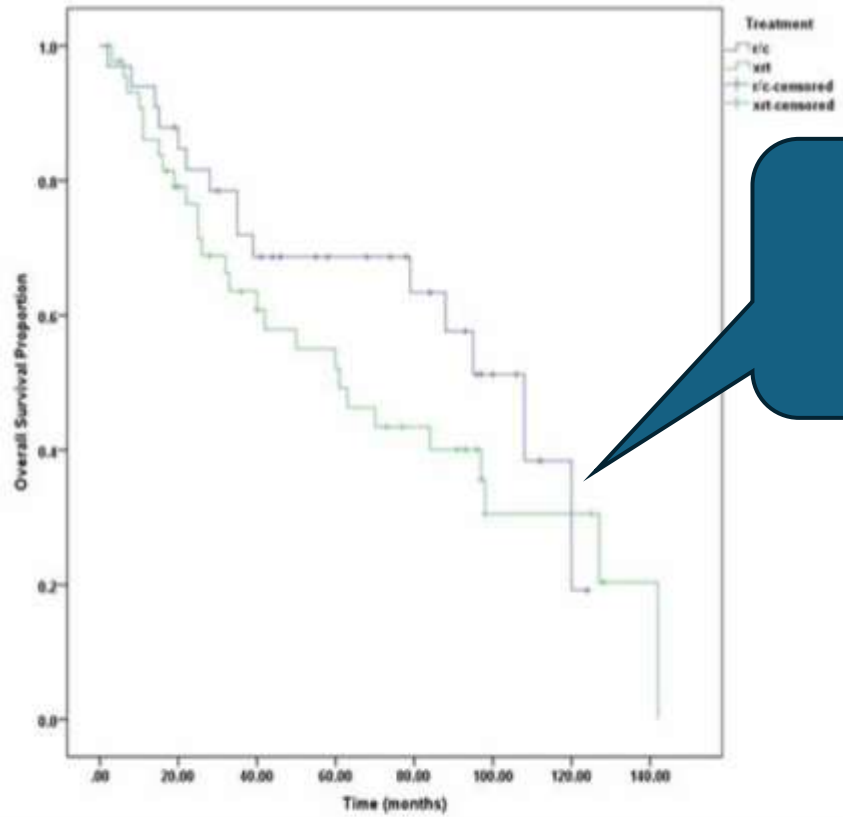


FIGURE 3: Clinical stage II radical radiotherapy (xrt) vs. radical cystectomy (r/c) overall survival

What's the main problem?

Fiaz S, Ali A, Adnan S, et al. (August 26, 2020) Comparison of Outcomes Between Radical Radiotherapy and Radical Cystectomy in Muscle Invasive Bladder Cancer in a Cancer Specialized Unit of a Developing Country. Cureus 12(8): e10057. DOI 10.7759/cureus.10057

Radical cystectomy versus trimodality therapy for muscle-invasive bladder cancer: a multi-institutional propensity score matched and weighted analysis

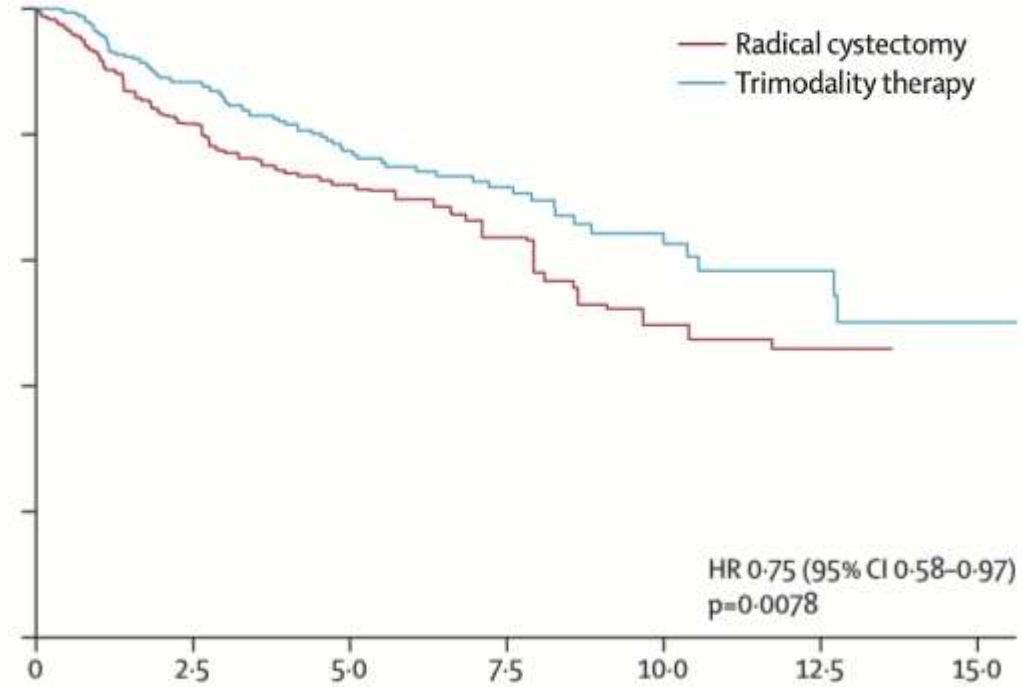
Alexandre R Zlotta, Leslie K Ballas, Andrzej Niemierko†, Katherine Lajkosz†, Cynthia Kuk, Gus Miranda, Michael Drumm, Andrea Mari, Ethan Thio, Neil E Fleshner, Girish S Kulkarni, Michael A S Jewett, Robert G Bristow, Charles Catton, Alejandro Berlin, Srikala S Sridhar, Anne Schuckman, Adam S Feldman, Matthew Wszolek, Douglas M Dahl, Richard J Lee, Philip J Saylor, M Dror Michaelson, David T Miyamoto, Anthony Zietman, William Shipley, Peter Chung, Siamak Daneshmand, Jason A Efstathiou**

Summary

Background Previous randomised controlled trials comparing bladder preservation with radical cystectomy for muscle-invasive bladder cancer closed due to insufficient accrual. Given that no further trials are foreseen, we aimed to use propensity scores to compare trimodality therapy (maximal transurethral resection of bladder tumour followed by concurrent chemoradiation) with radical cystectomy.

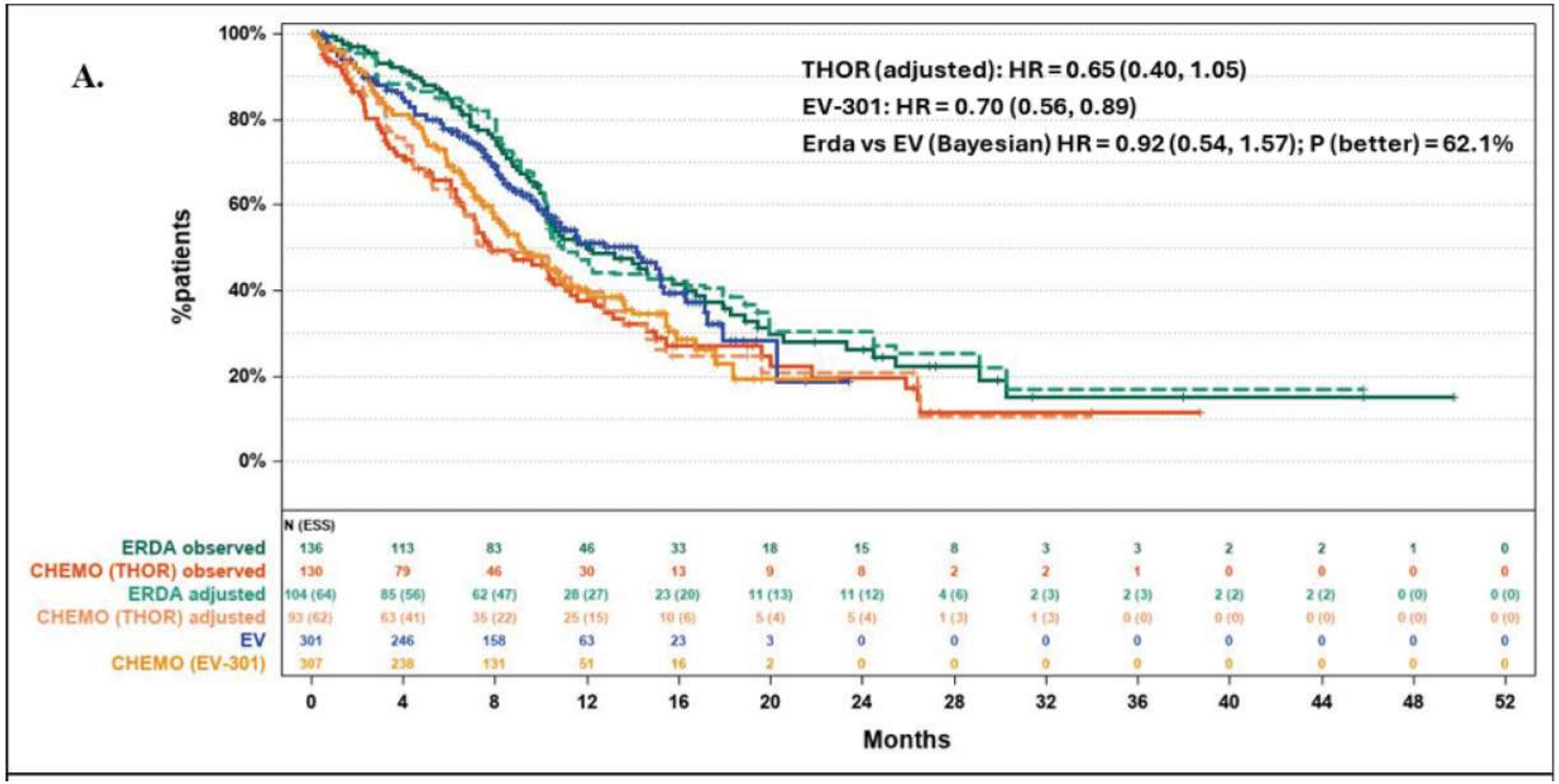
Lancet Oncol 2023; 24: 669–81
Published Online
May 12, 2023
https://doi.org/10.1016/S1470-2045(23)00170-5

Overall survival



Number at risk (number censored)							
Radical cystectomy	837 (0)	595 (124)	403 (325)	172 (408)	62 (451)	16 (467)	0 (467)
Trimodality therapy	282 (0)	222 (59)	138 (113)	76 (140)	43 (161)	19 (176)	2 (177)

Matched adjusted intertrial comparison, EV-301 and Thor trials



Sanden et al. *JHEOR*. **2024:11(2):49-57.**

Alternatives to randomized phase III in evidence generation

- Historical controls
- Other novel methods:
  - Trial within cohort study (TWICS)
  - Cluster randomisation

Alternatives to randomized phase III in evidence generation

- Historical controls
- Other novel methods:
  - Trial within cohort study (TWICS)
  - Cluster randomisation

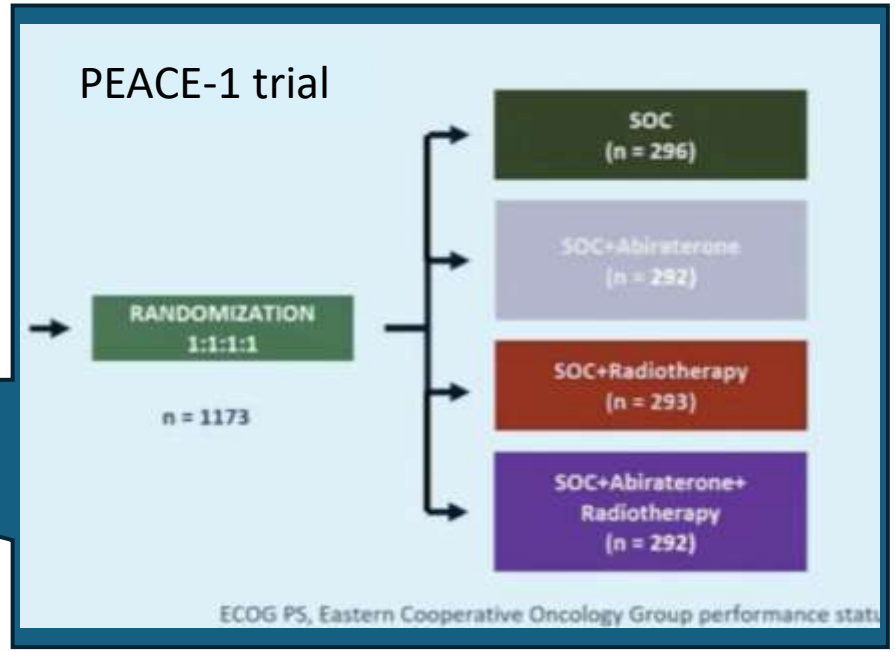
Randomised controlled phase III trials

Why do we randomize?

Randomised phase III trials

Types

- Parallel group
- Cross-over
- Factorial



Objectives

- Superiority
- Non-inferiority

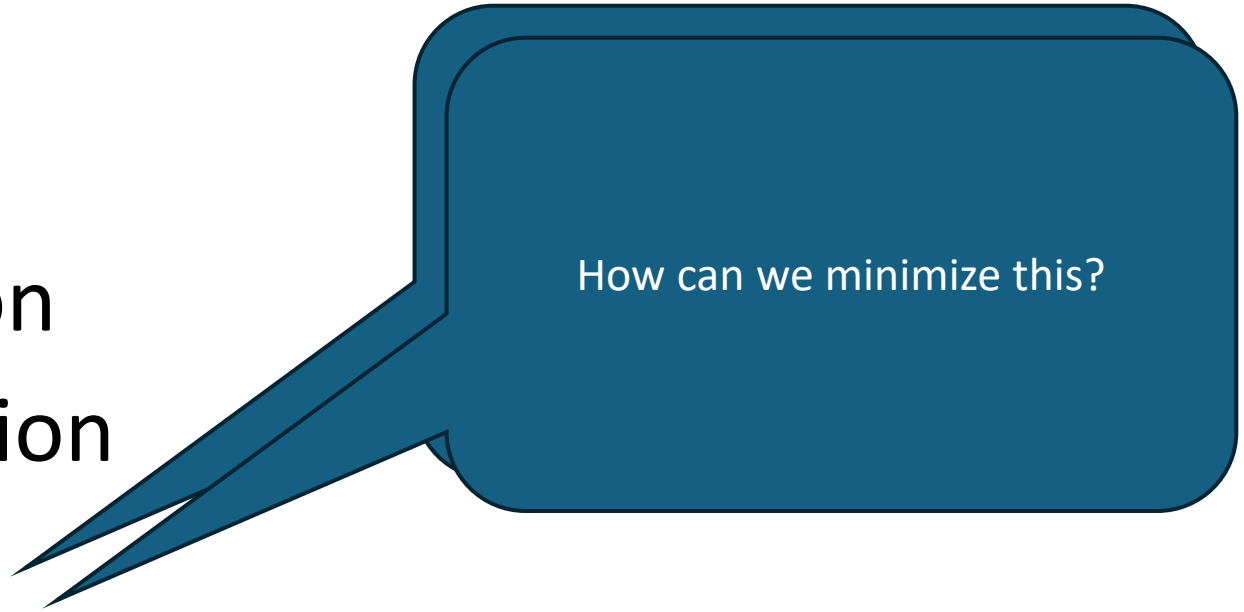
Fizazi et al. The Lancet. 2022. Volume 399, Issue 10336P1695-1707

PICO

- *P* opulation
- *I* ntervention
- *C* ontrol
- *O* utcome

PICO

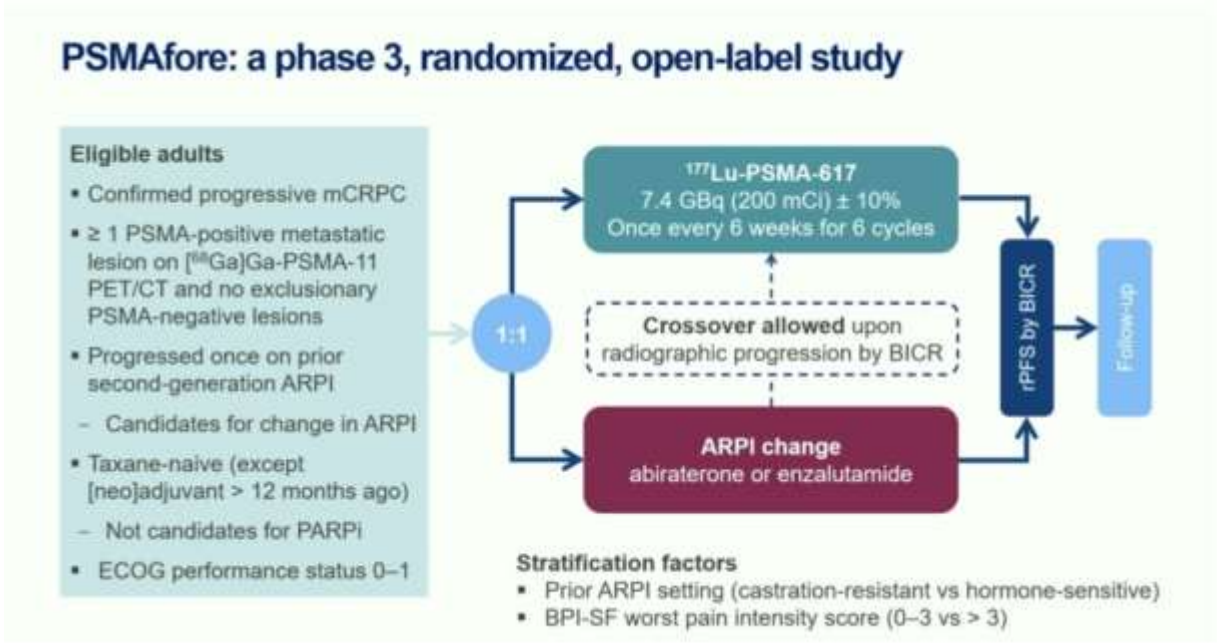
- *P* opulation
- *I* ntervention
- ***C* ontrol**
- *O* utcome

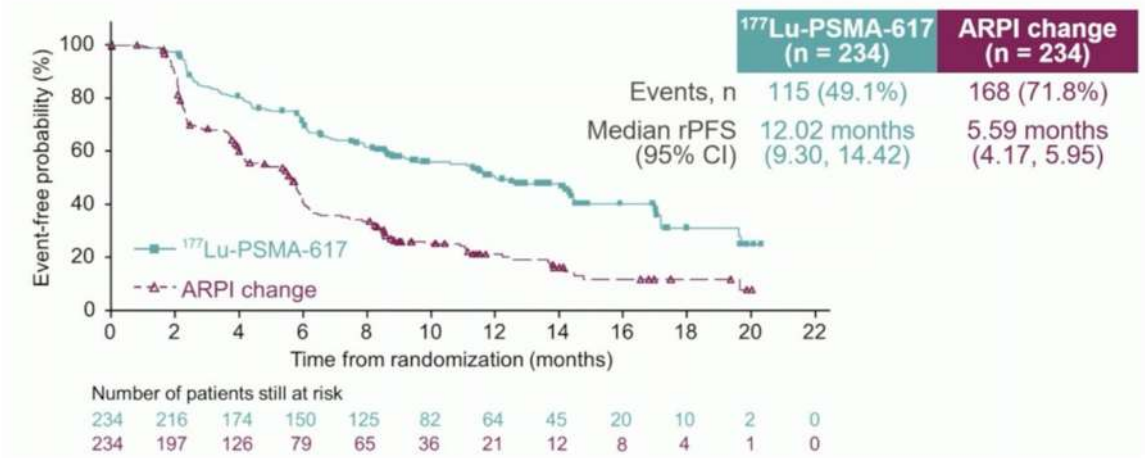


What endpoints do we use to measure effectiveness of a cancer treatment?

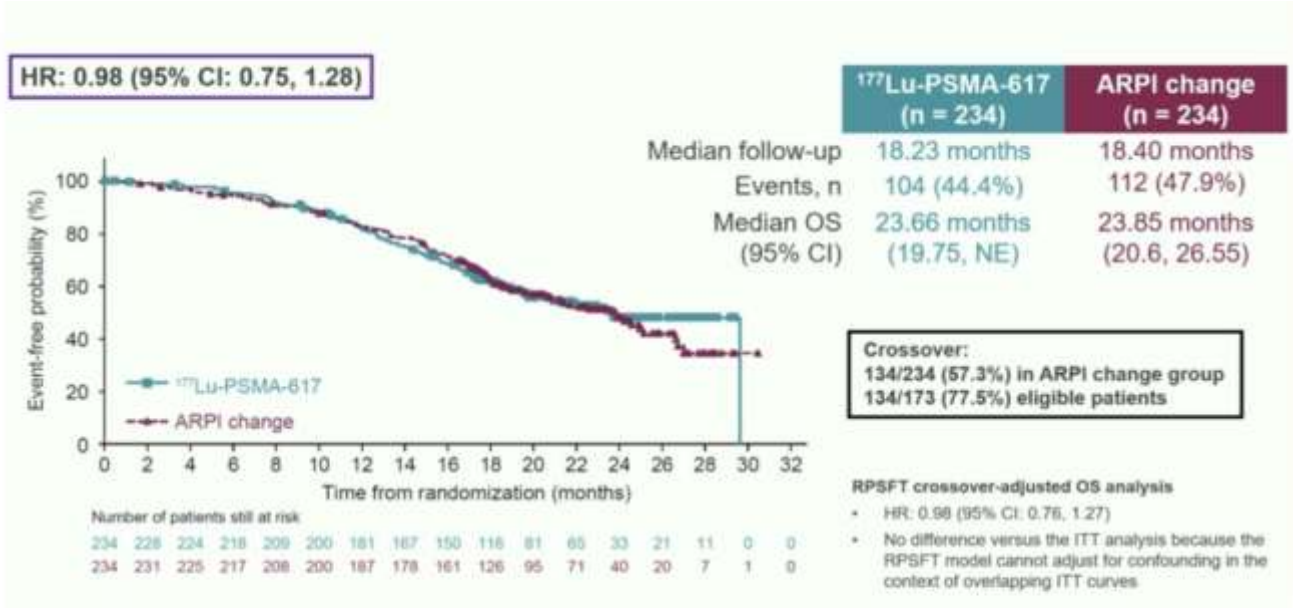
What endpoints do we use to measure effectiveness of a cancer treatment?

- Overall survival
- Cause-specific survival
- Progression free survival (disease free survival)
- Time to treatment failure
- Response rate
- Quality of life/ symptomatic change
- Situation specific endpoints
  - Metastasis free survival
  - Local failure free survival
  - Pain
- Composite endpoints
  - Time to first SRE



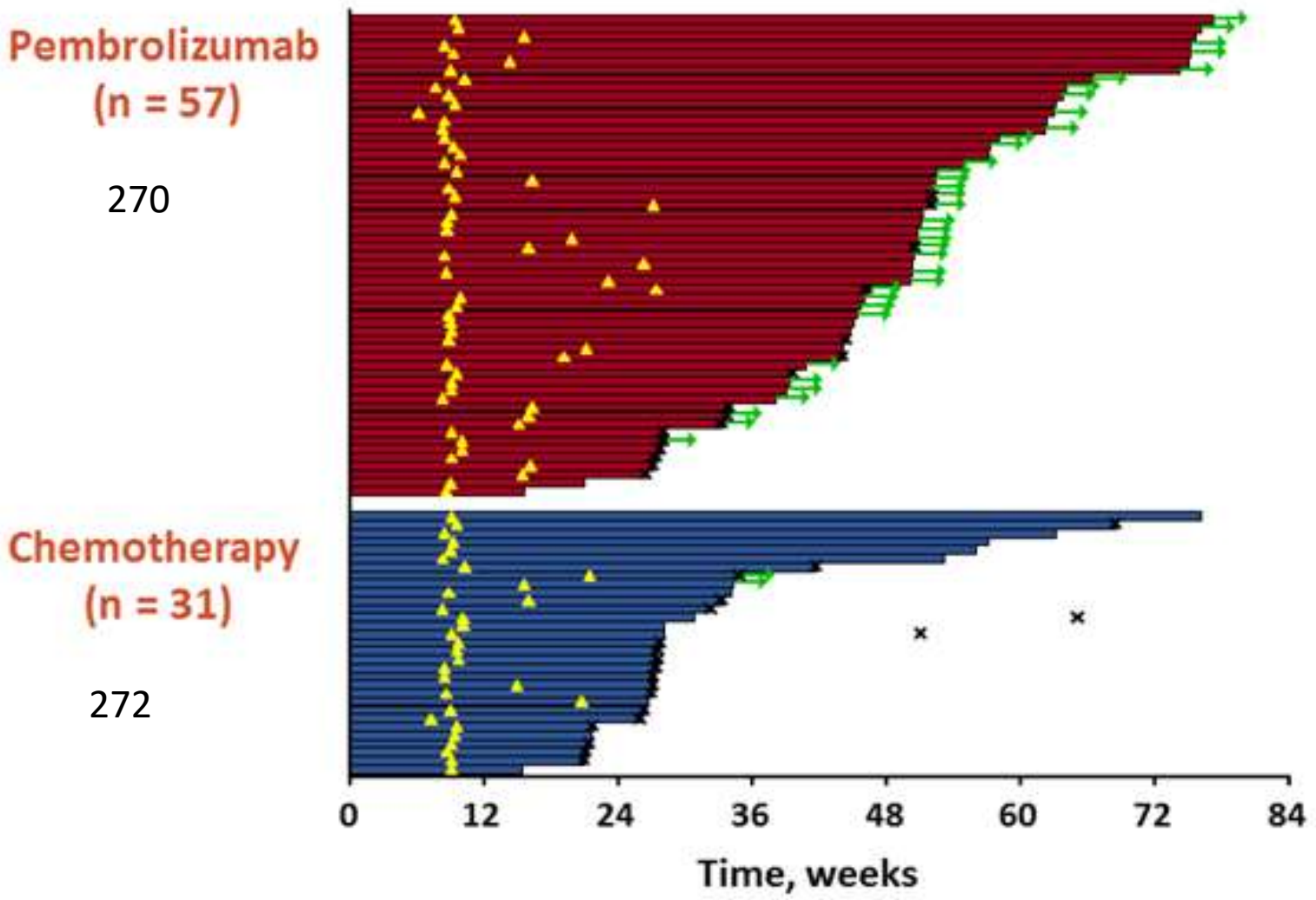


Sartor et al. The Society of Nuclear Medicine & Molecular Imaging (SNMMI) 2024 Annual Meeting held in Toronto, June 2024



With regards to adverse events, grade 3-4 events were less frequent in the Lu-PSMA arm (34% versus 43%). Similarly, serious adverse events were also less common with ^{177}Lu-PSMA-617 treatment (20% versus 28%). Adverse events leading to dose adjustment occurred less commonly with ^{177}Lu-PSMA-617 treatment (3.5% versus 15%).

Duration of response



Bellmunt et al. NEJM 2017

Survival Analysis

What is survival analysis?

The common outcome for assessment in oncology trials is time-to-event, often termed **survival time**.*[1]

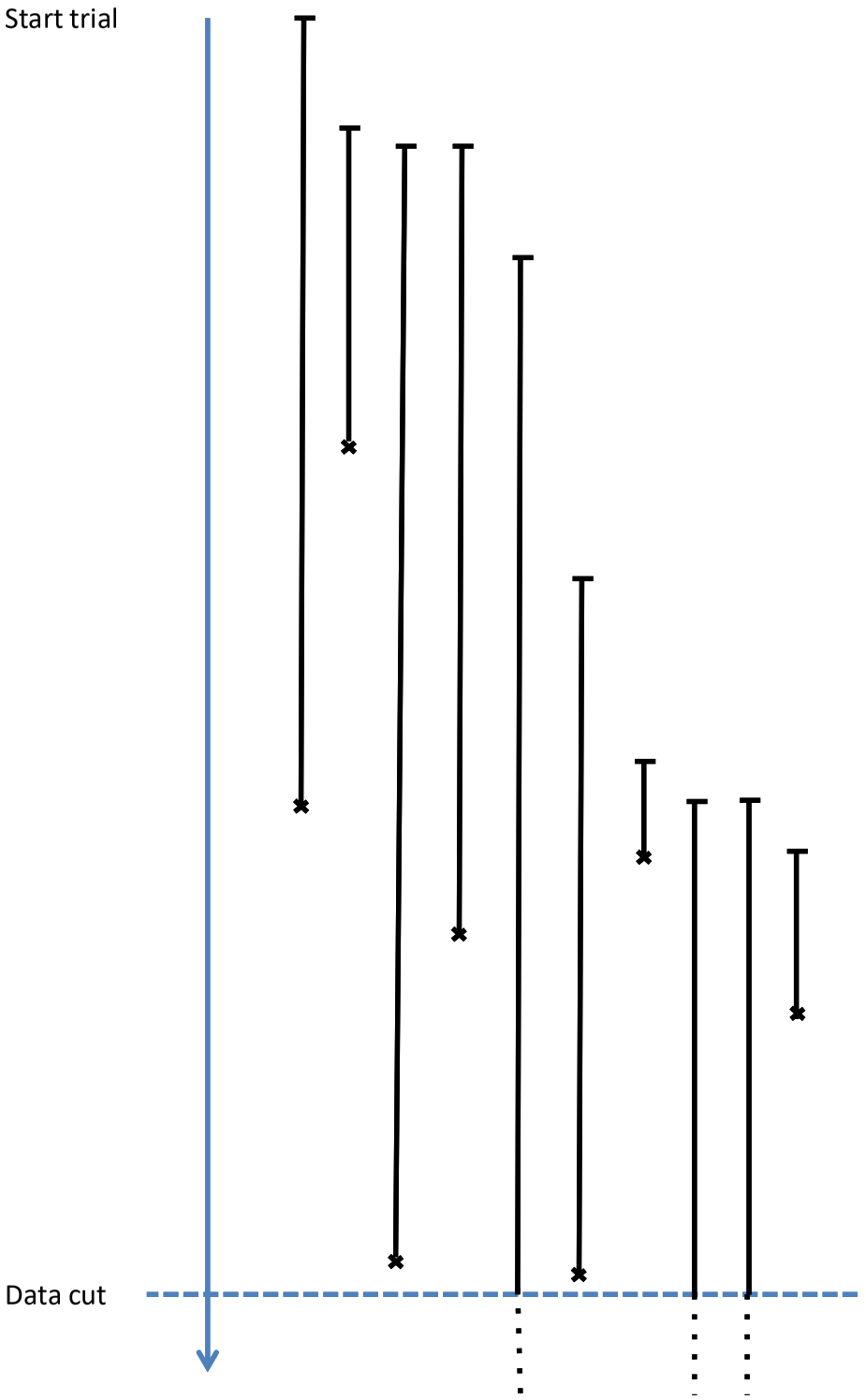
Why are survival data different?[1]

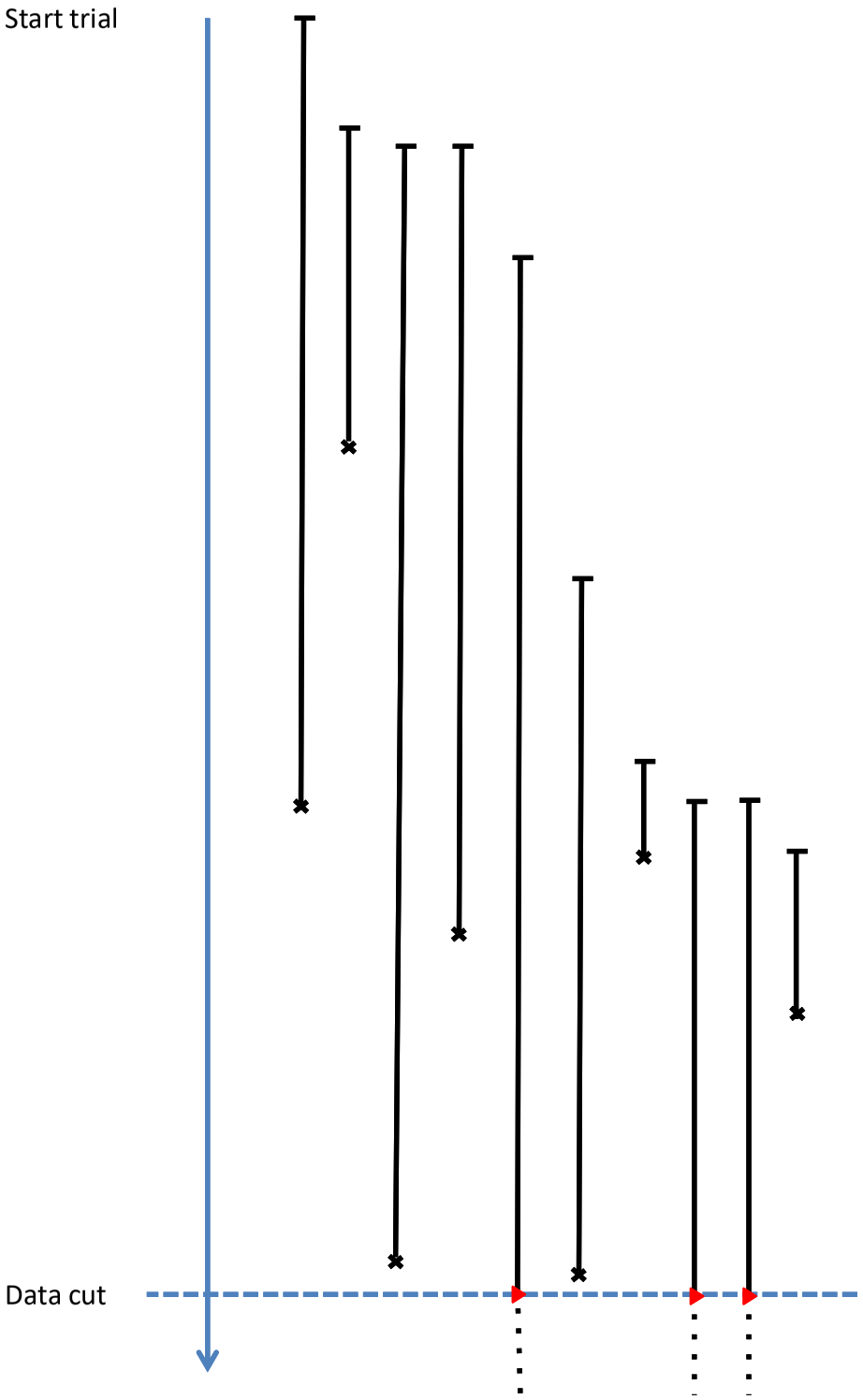
- By the end of the follow-up period, not all individuals may have experienced the event of interest, and some individuals may have been lost to follow-up. Hence, the true time-to-event is unknown (censoring)
- Survival data are typically skewed and rarely normally distributed (median [instead of mean] survival are therefore used)[2]

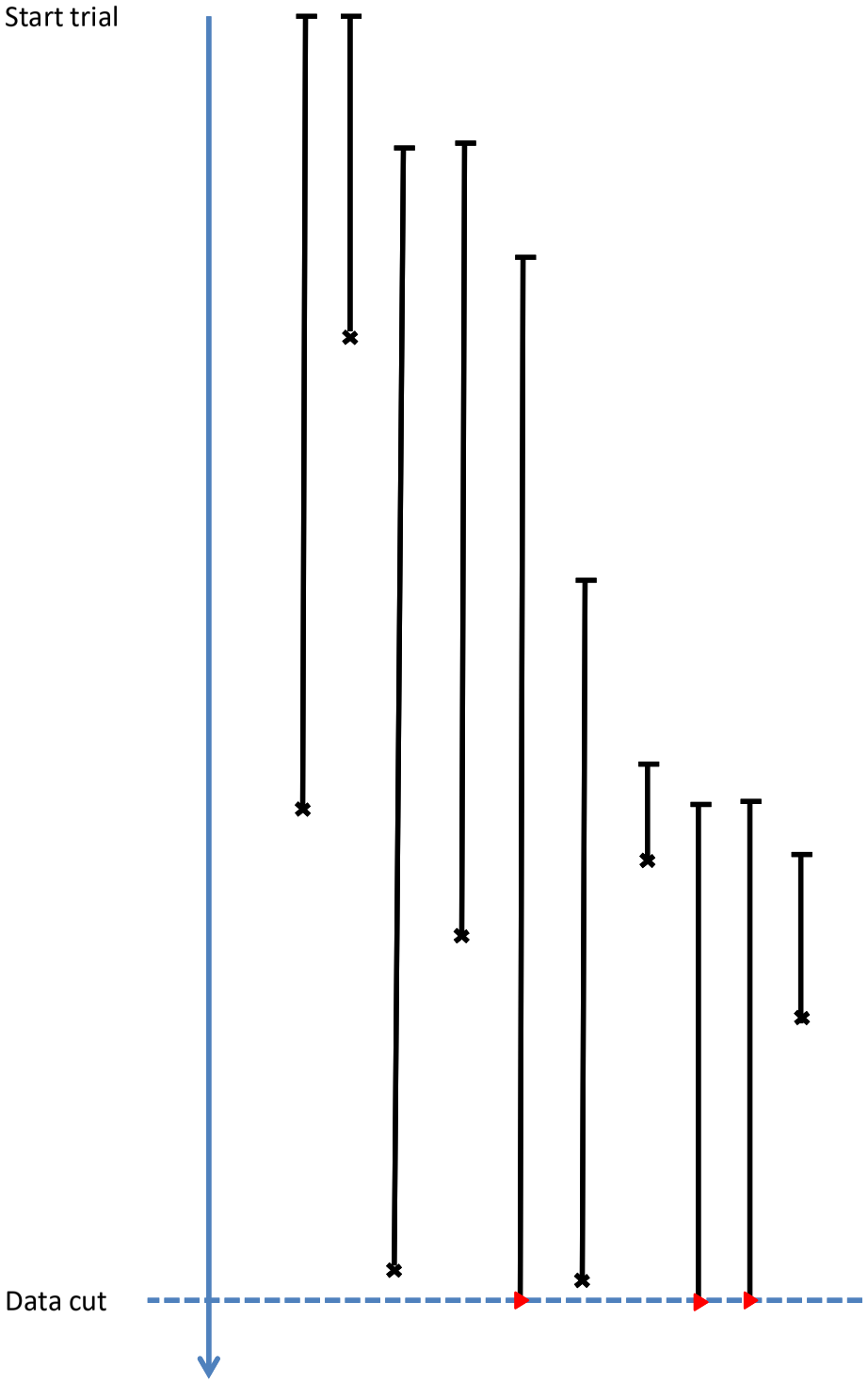
These features of survival data necessitate **survival analysis** methods, such as Kaplan-Meier curves, log-rank tests and/or the Cox regression model.[1]

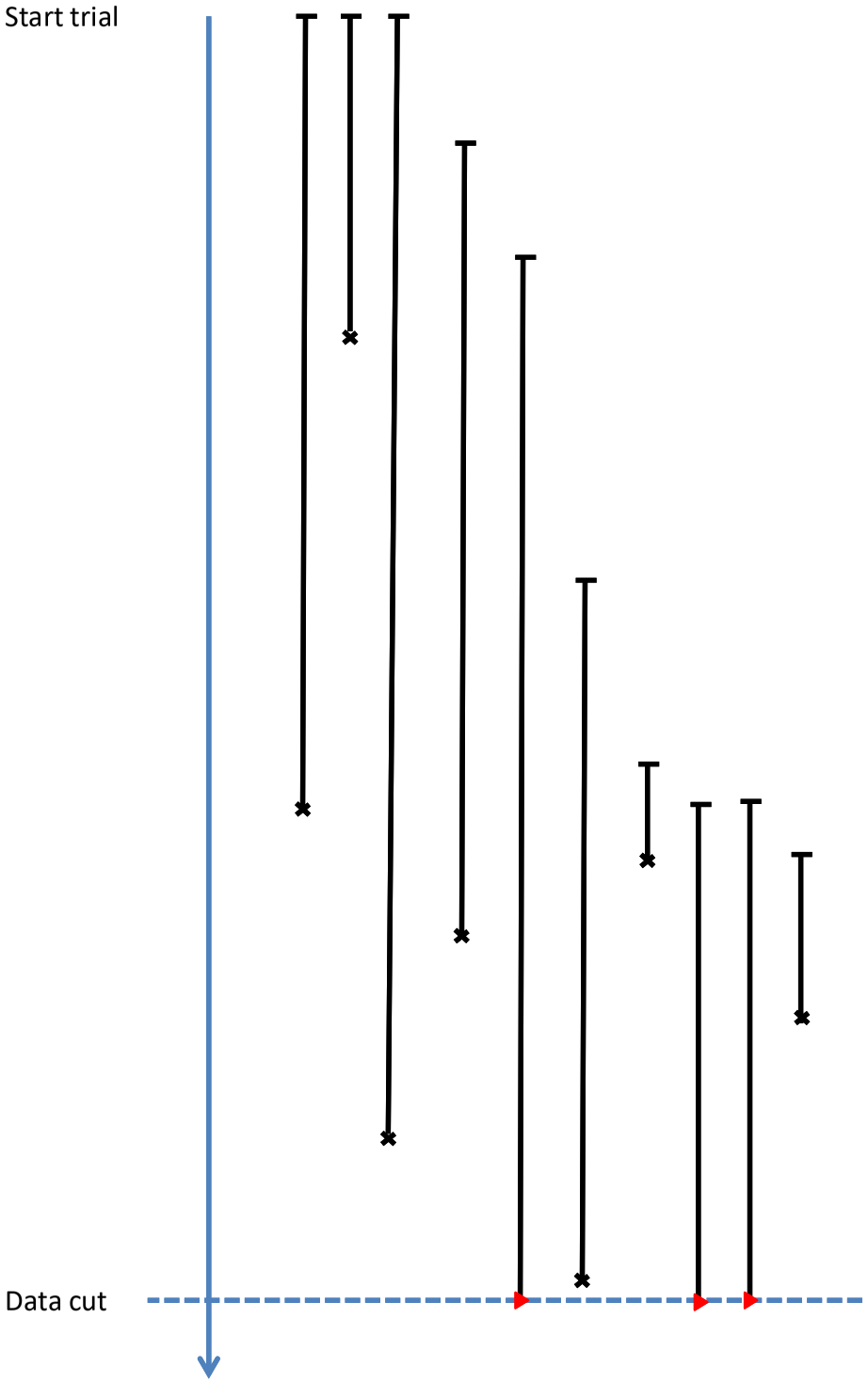
*In addition to time survived from diagnosis to death, it may also be applied to the time survived from complete remission to relapse or progression.[1]

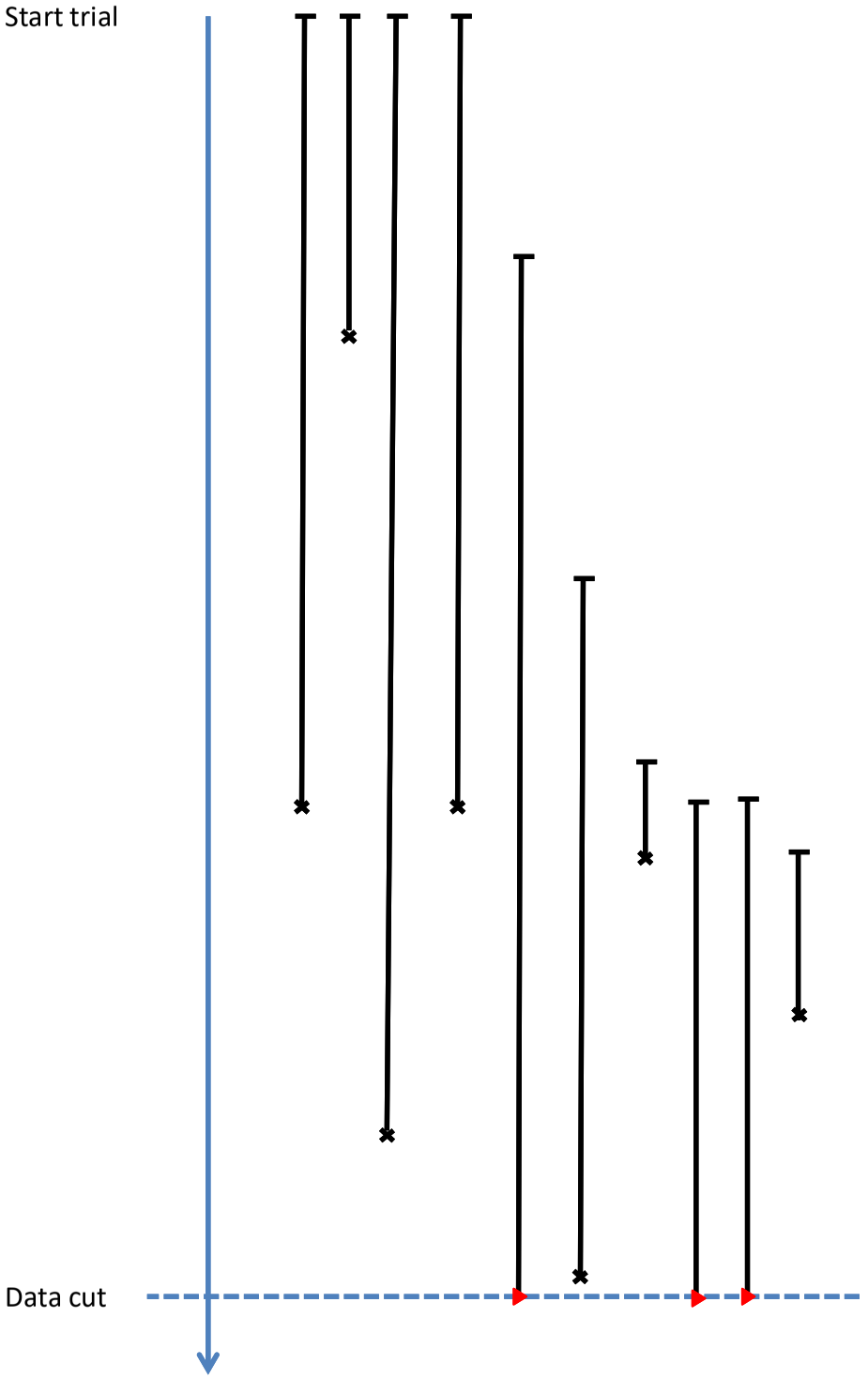
1. Clark TG et al. Br J Cancer. 2003;89(2):232–238; **2.** Jager KJ et al. Kidney Int. 2008;74(5):560–565.

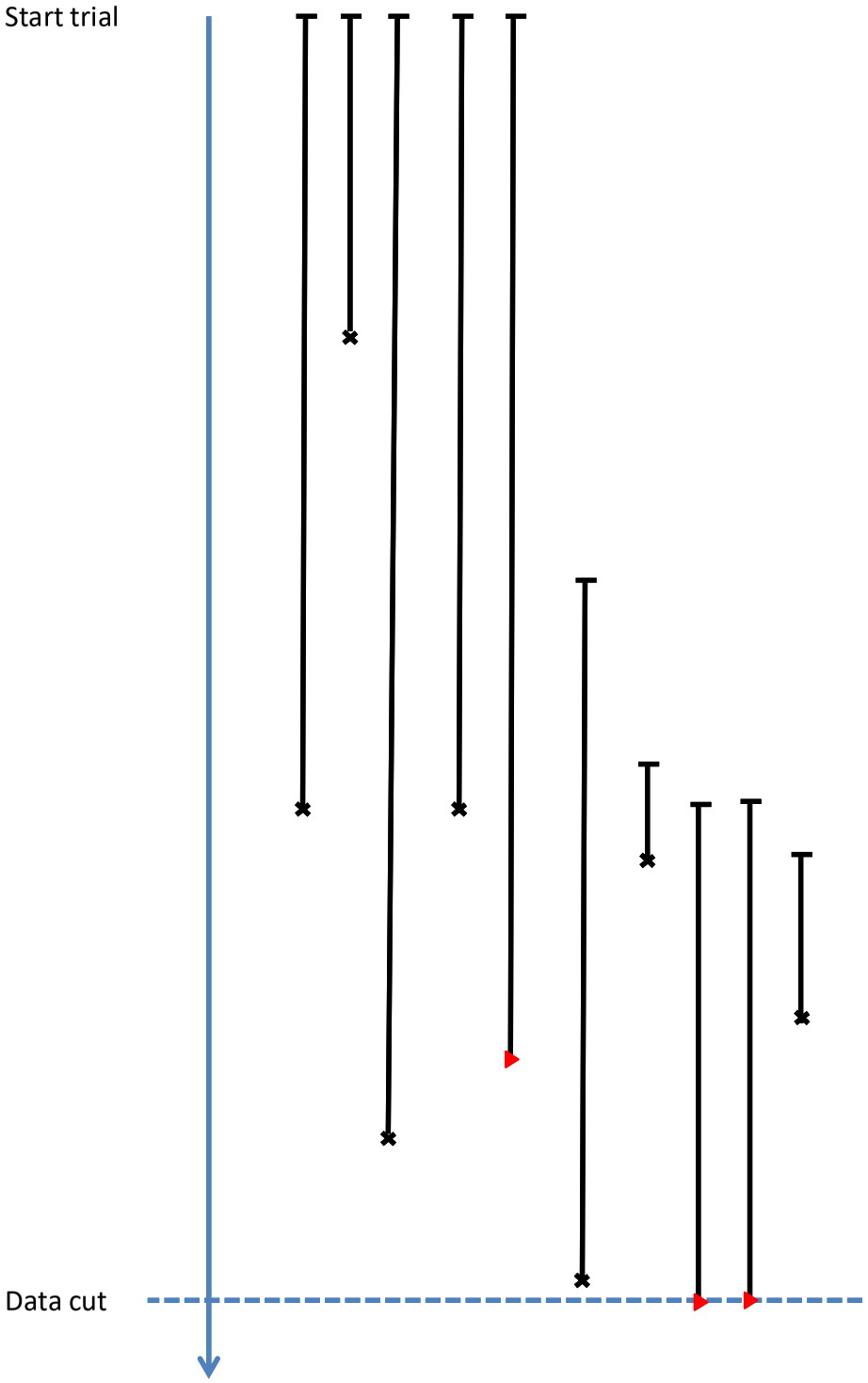

Start trial
Data cut

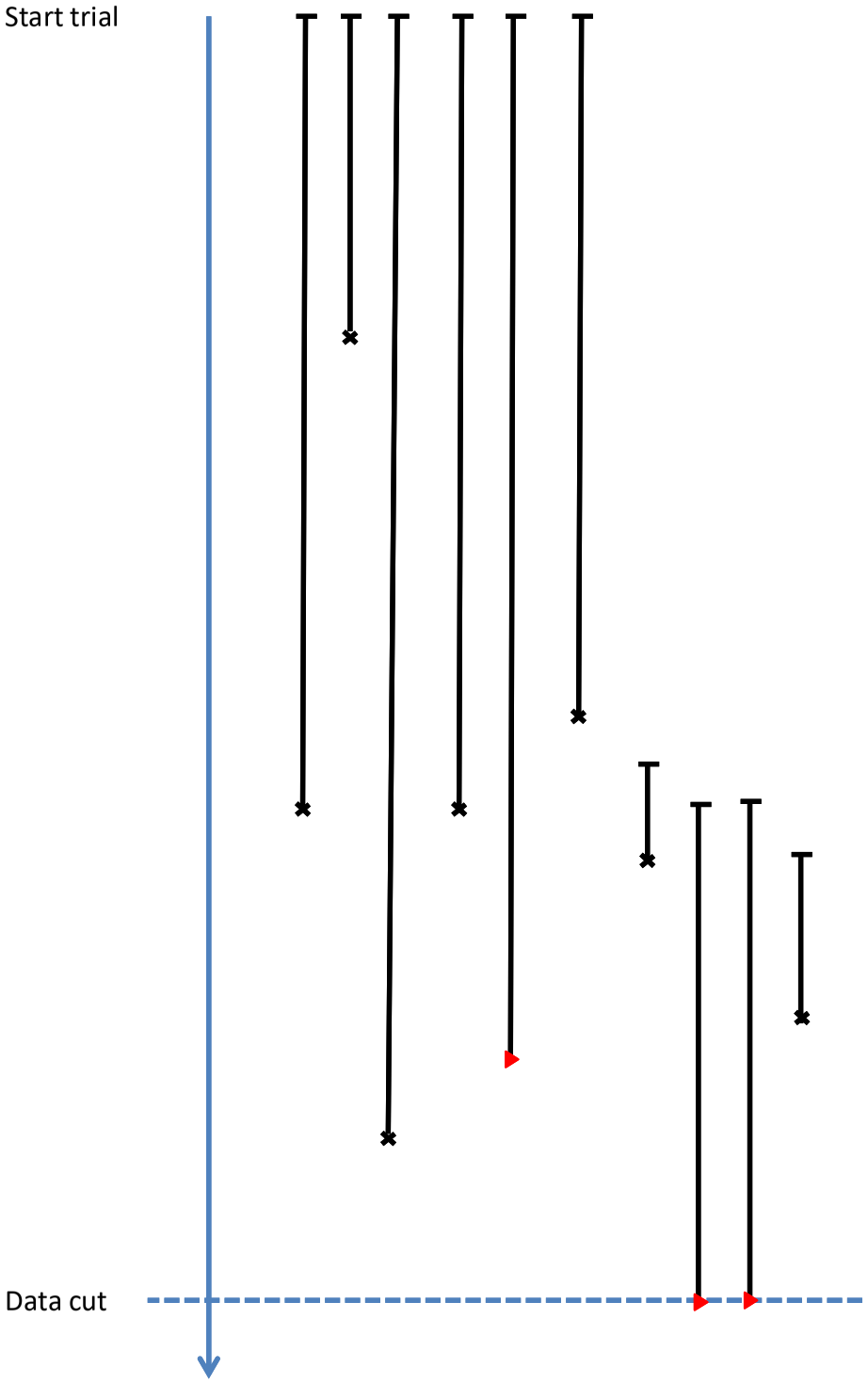

Start trial
Data cut

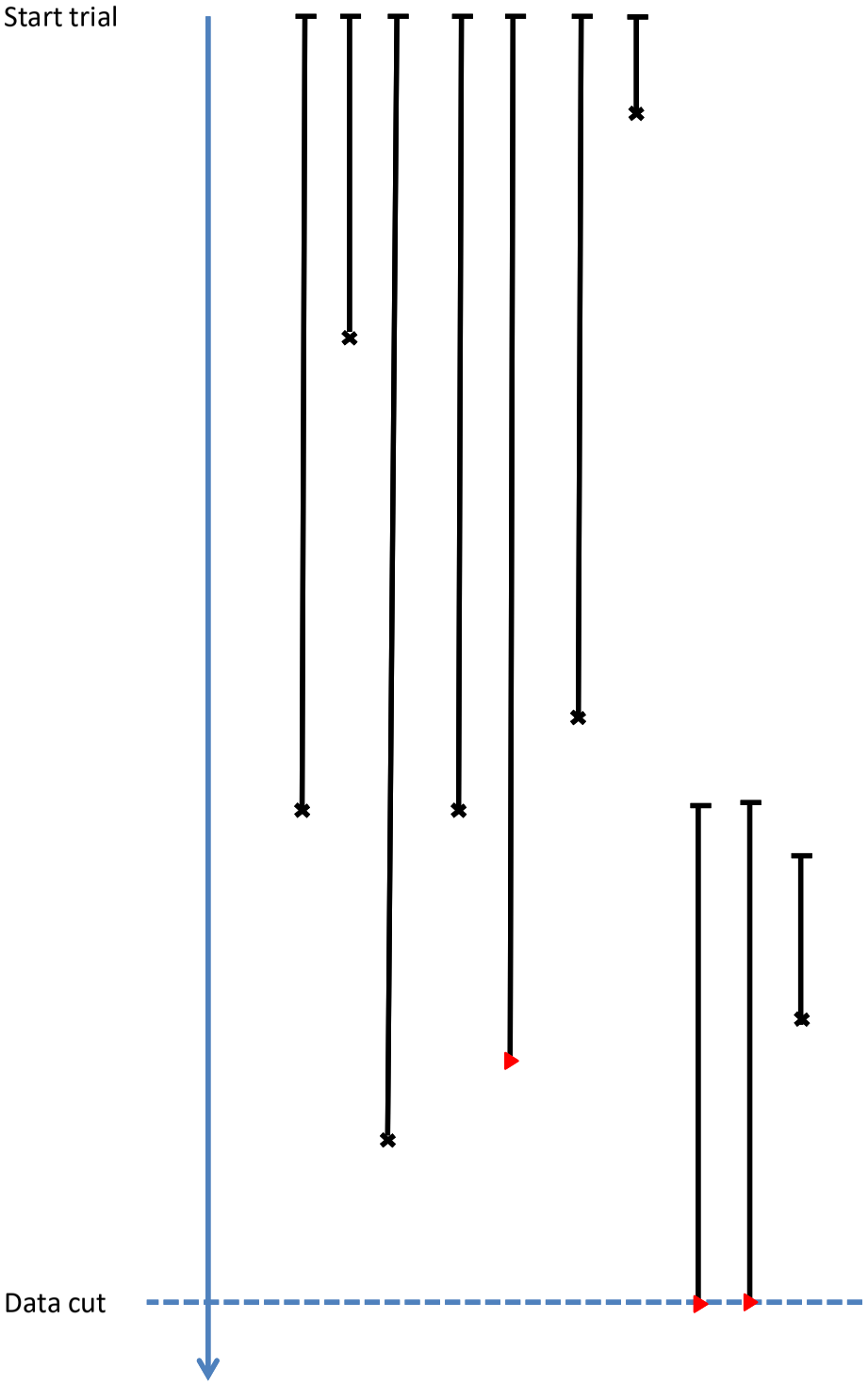

Start trial
Data cut

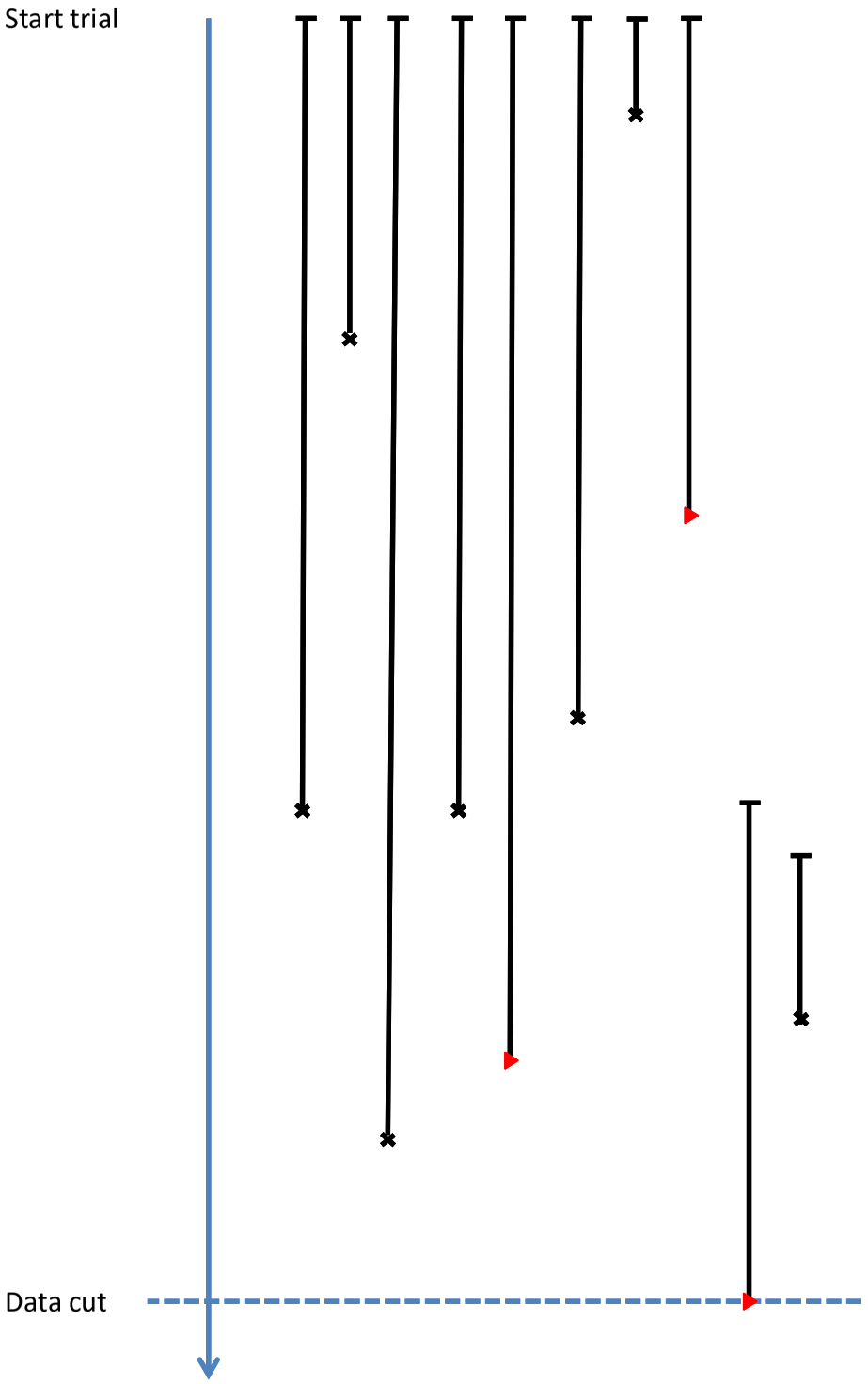

Start trial
Data cut

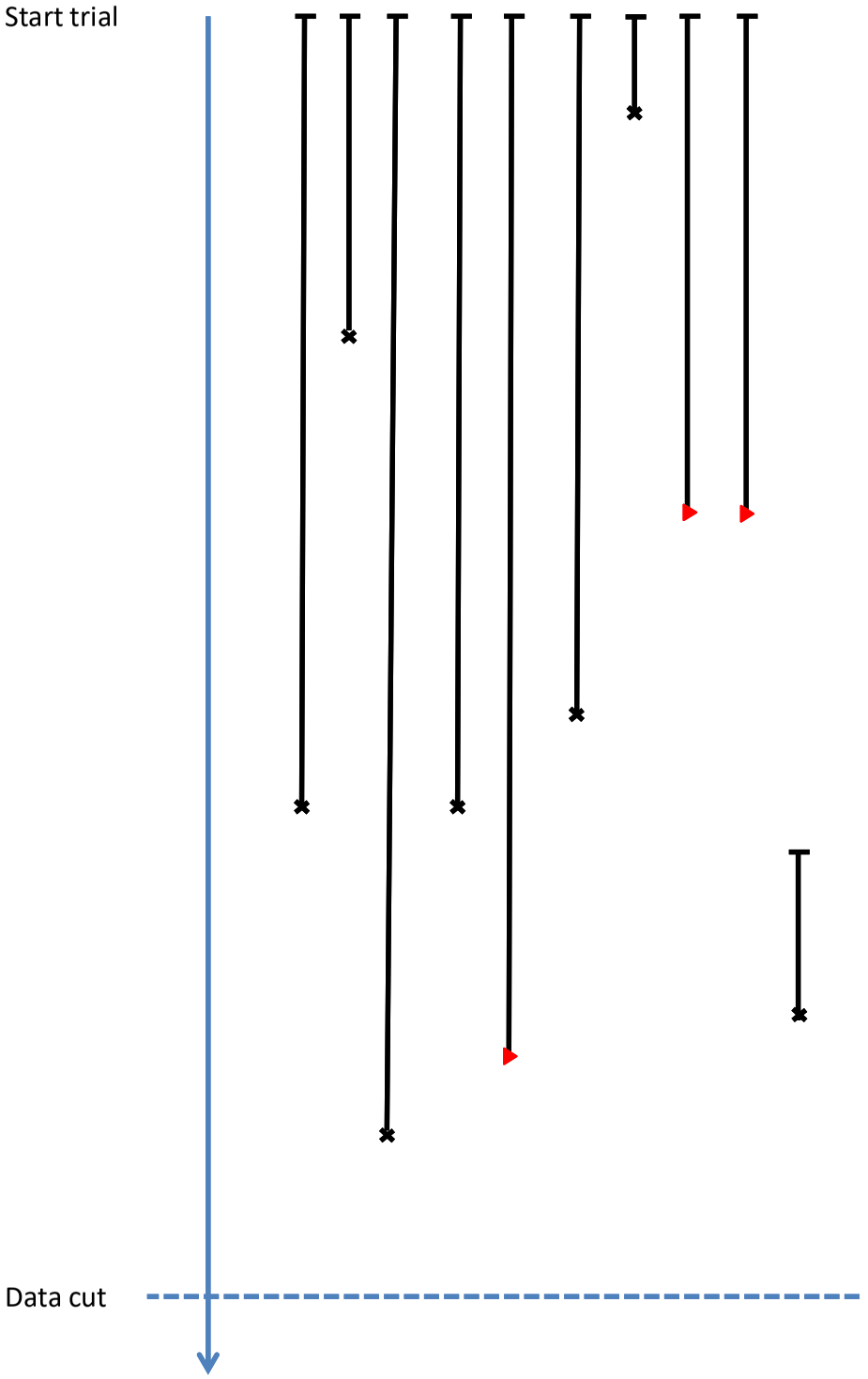

Start trial
Data cut

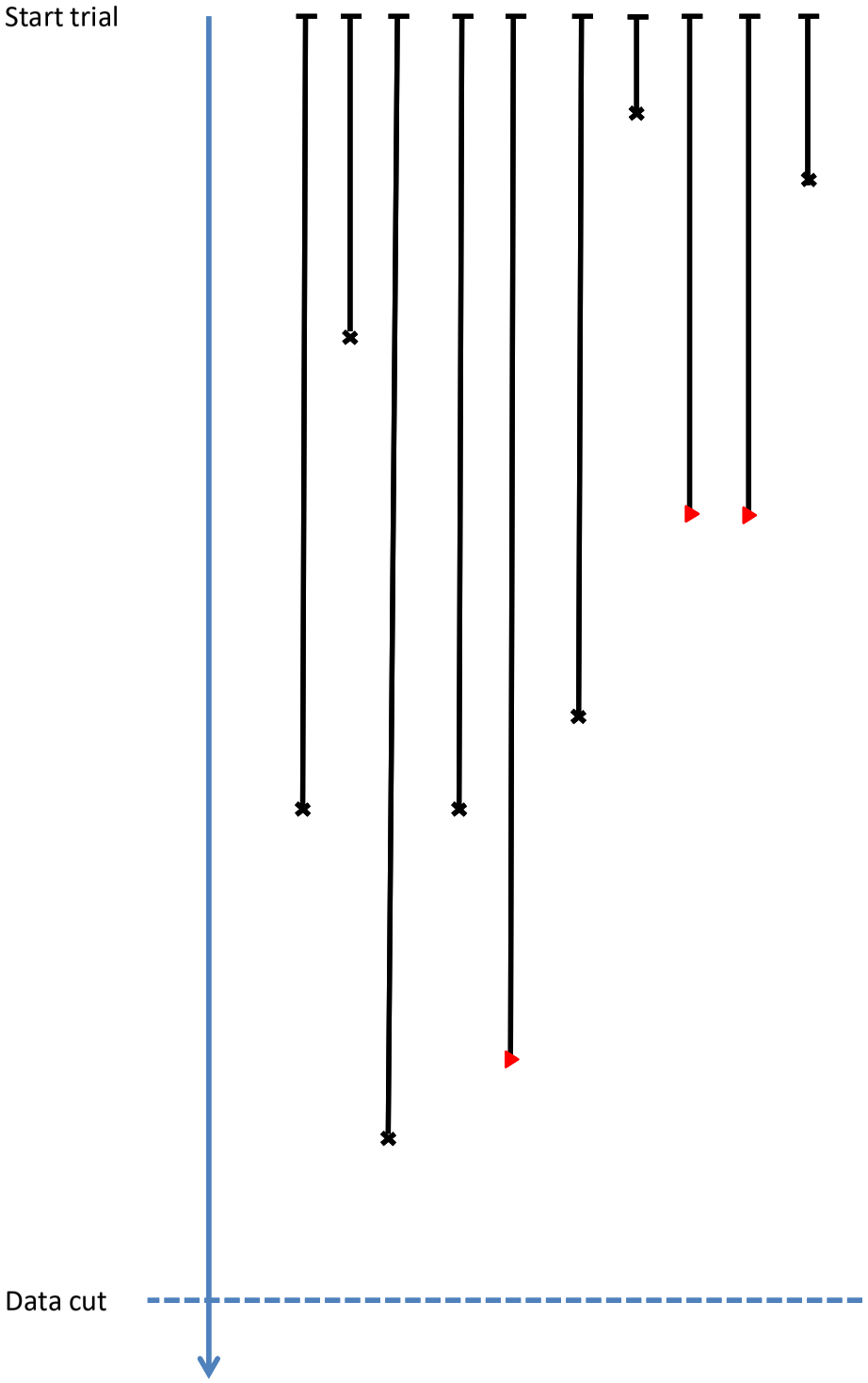

Start trial
Data cut

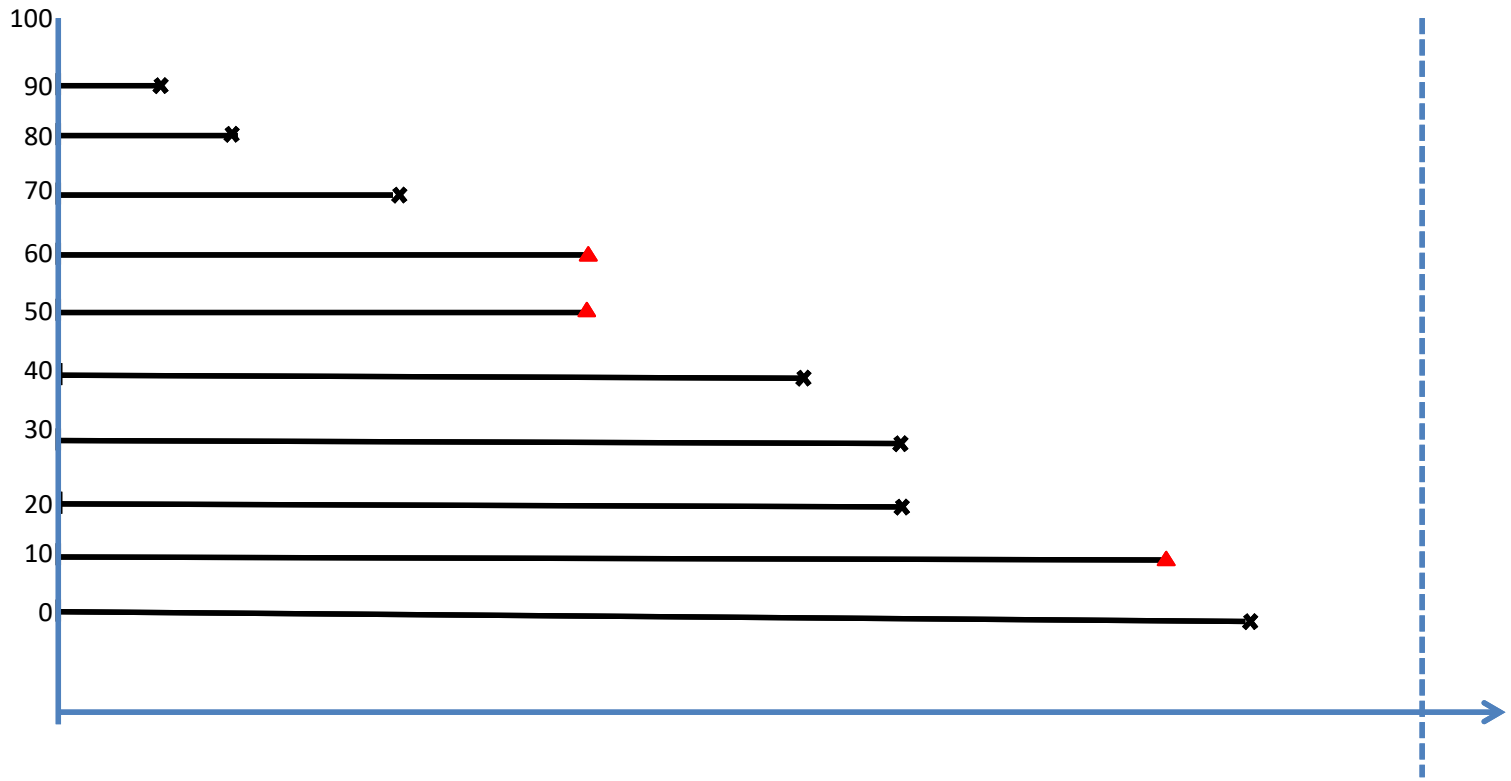

Start trial
Data cut


Start trial
Data cut

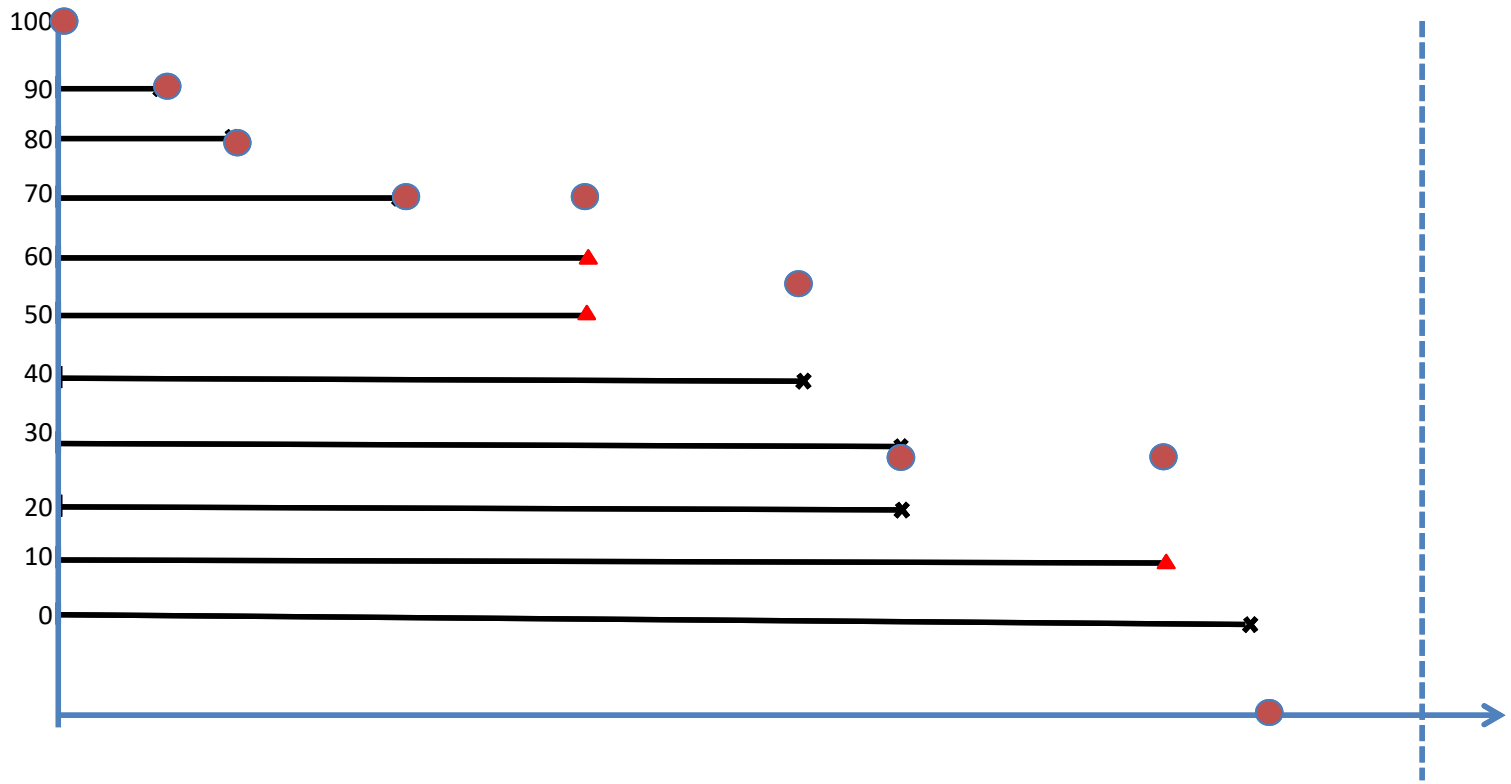

Start trial
Data cut


Start trial
Data cut

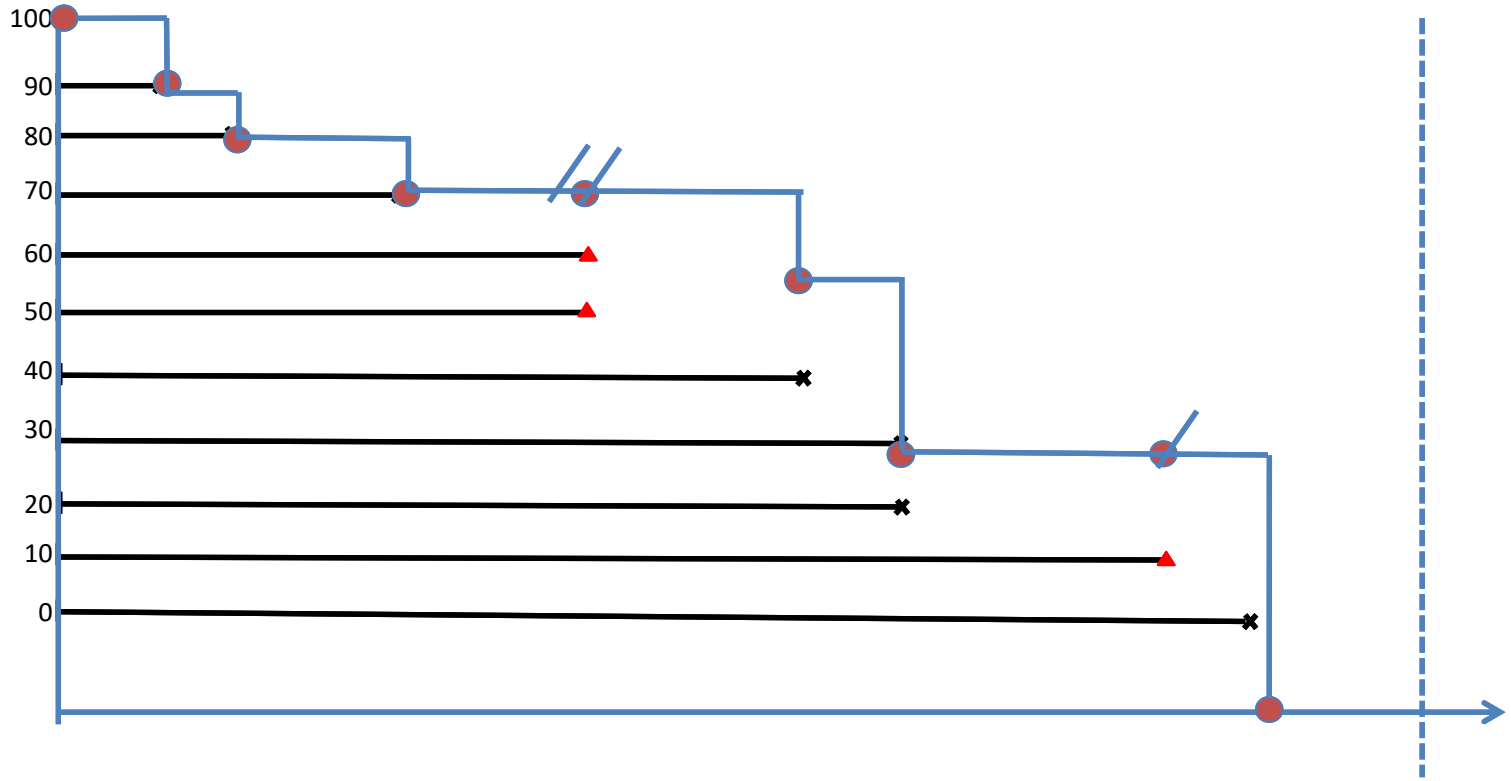

Start trial
Data cut



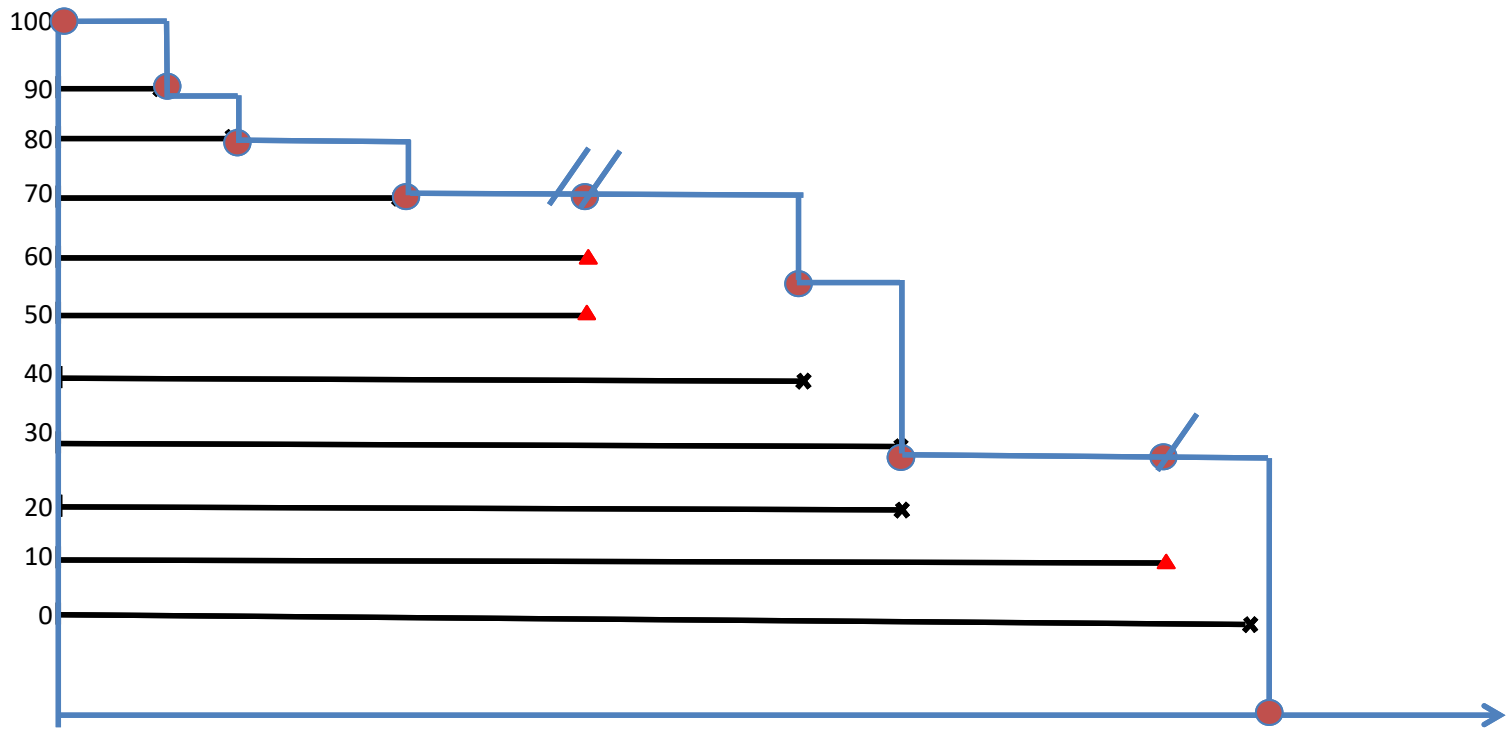
Number at risk	10	9	8	7	5	4	2	1	0
Proportion surviving	1	0.9 (9/10)	0.89 (8/9)	0.88 (7/8)	1.0	0.8 (4/5)	0.5 (2/4)	1.0	0 (0/1)
Cumulative survival	1	0.9	0.8	0.7	0.7	0.56	0.28	0.28	0



Number at risk	10	9	8	7	5	4	2	1	0
Proportion surviving	1	0.9 9/10	0.89 8/9	0.88 7/8	1.0	0.8 4/5	0.5 2/4	1.0	0 0/1
Cumulative survival	1	0.9	0.8	0.7	0.7	0.56	0.28	0.28	0



Number at risk	10	9	8	7	5	4	2	1	0
Proportion surviving	1	0.9 9/10	0.89 8/9	0.88 7/8	1.0	0.8 4/5	0.5 2/4	1.0	0 0/1
Cumulative survival	1	0.9	0.8	0.7	0.7	0.56	0.28	0.28	0

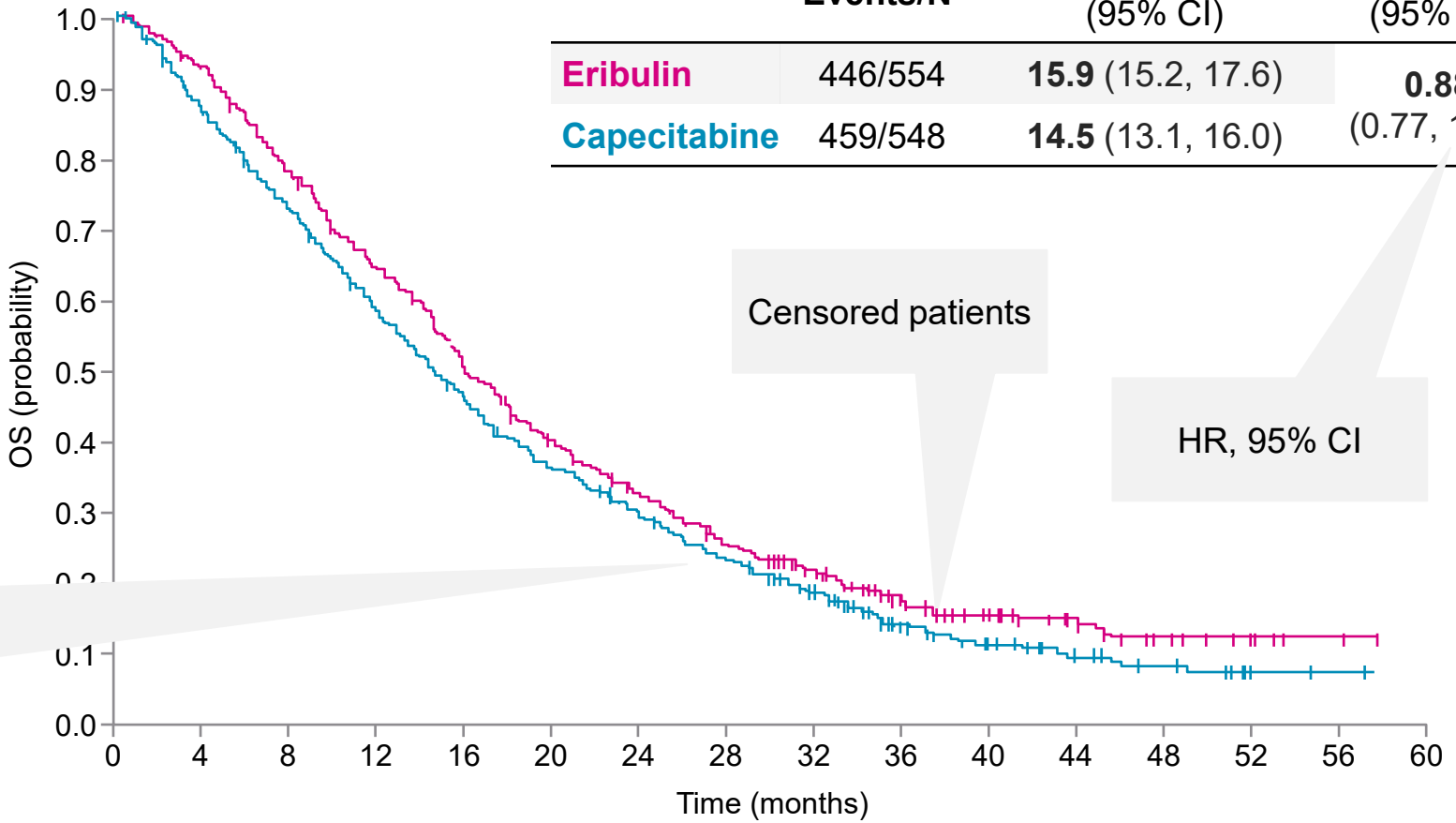


Number at risk	10	9	8	7	5	4	2	1	0
Proportion surviving	1	0.9 (9/10)	0.89 (8/9)	0.88 (7/8)	1.0	0.8 (4/5)	0.5 (2/4)	1.0	0 (0/1)
Cumulative survival	1	0.9	0.8	0.7	0.7	0.56	0.28	0.28	0

Survival Analysis

Kaplan-Meier curves

	Events/N	Median, months (95% CI)	HR (95% CI)	p value
Eribulin	446/554	**15.9** (15.2, 17.6)	**0.88** (0.77, 1.00)	0.056
Capecitabine	459/548	**14.5** (13.1, 16.0)		



Censored patients

HR, 95% CI

Statistical uncertainty

Vertical steps show when the events occur

The height of the step correlates with the number of patients still in follow-up

Number of patients at risk

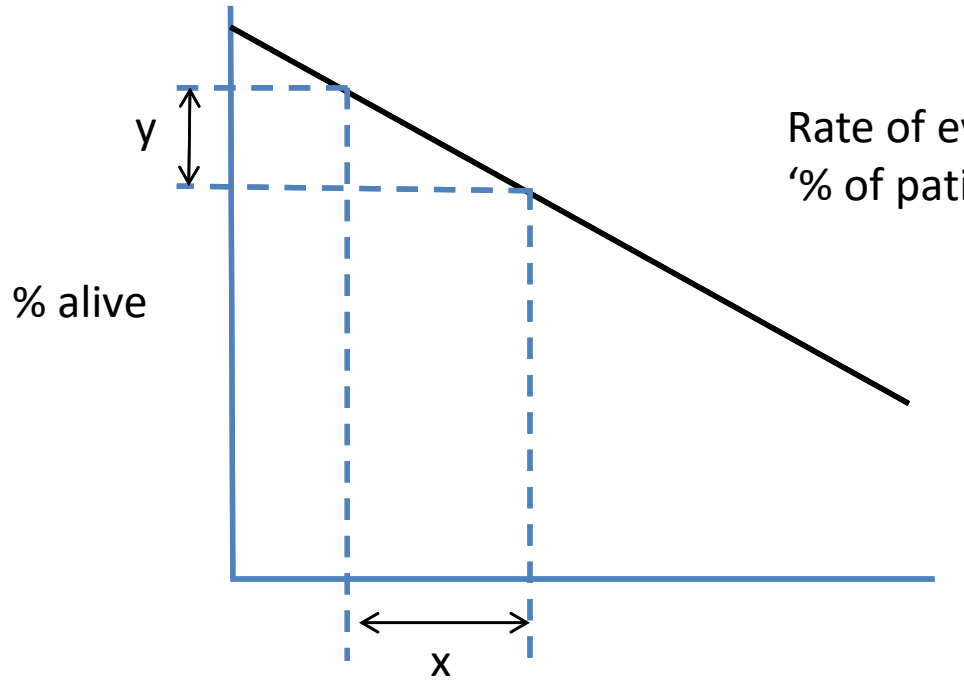
Eribulin	554	505	423	349	268	214	173	133	99	52	32	22	13	7	2	0
Capecitabine	548	466	391	308	242	191	155	122	81	42	27	17	12	2	1	0

Kaufman PA et al. J Clin Oncol. 2015;33(6):594–601.

What is a hazard ratio?

Hazard

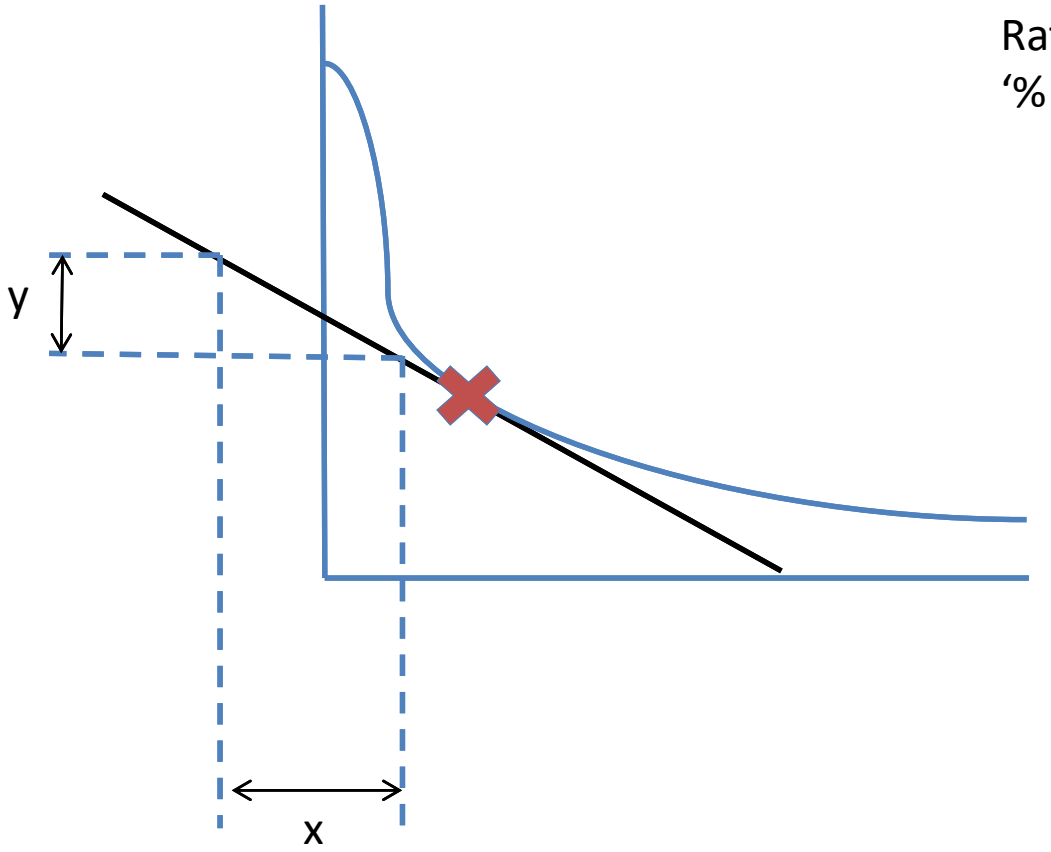
- A hazard rate is the rate at which an event occurs



Rate of events = y / x
'% of patients who die per month'

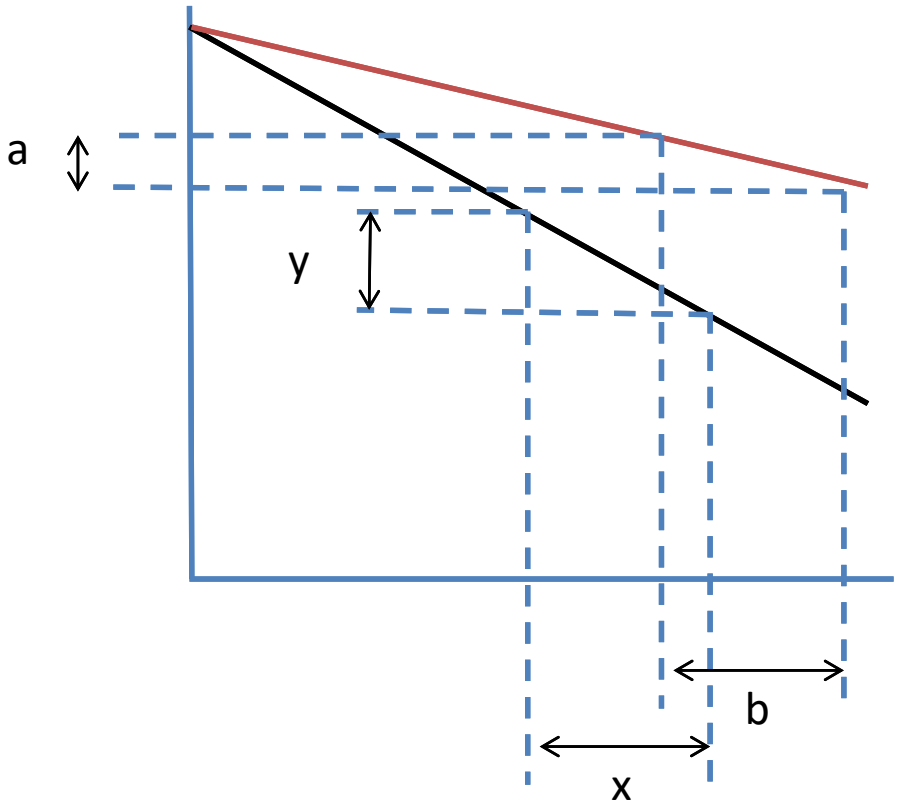
Hazard

- A hazard rate is the rate at which an event occurs



Rate of events = y / x
‘% of patients who die per month’

Hazard ratio



HR= (a/b):(y/x)

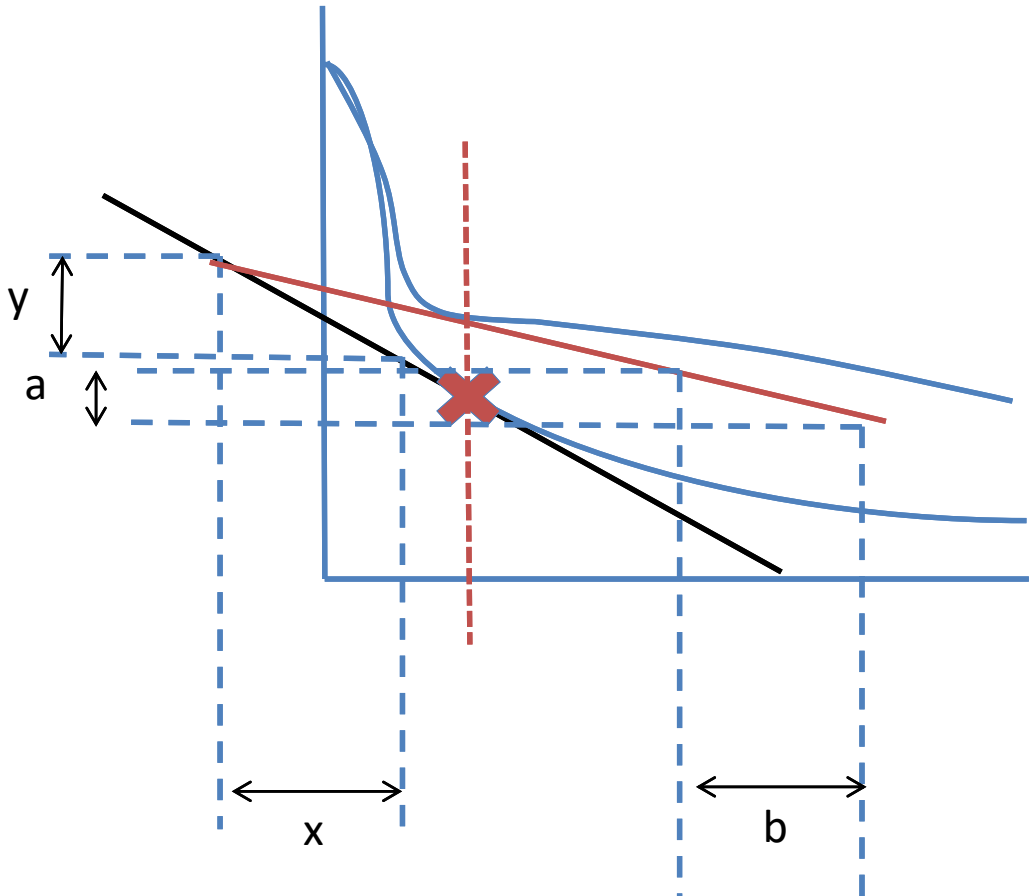
Eg. 10%/6months : 20%/6months

= 0.5

Or ‘patients are dying half as often as in control arm’

Or ‘the risk of death is reduced by 50%’

Hazard ratio



HR= (a/b):(y/x)

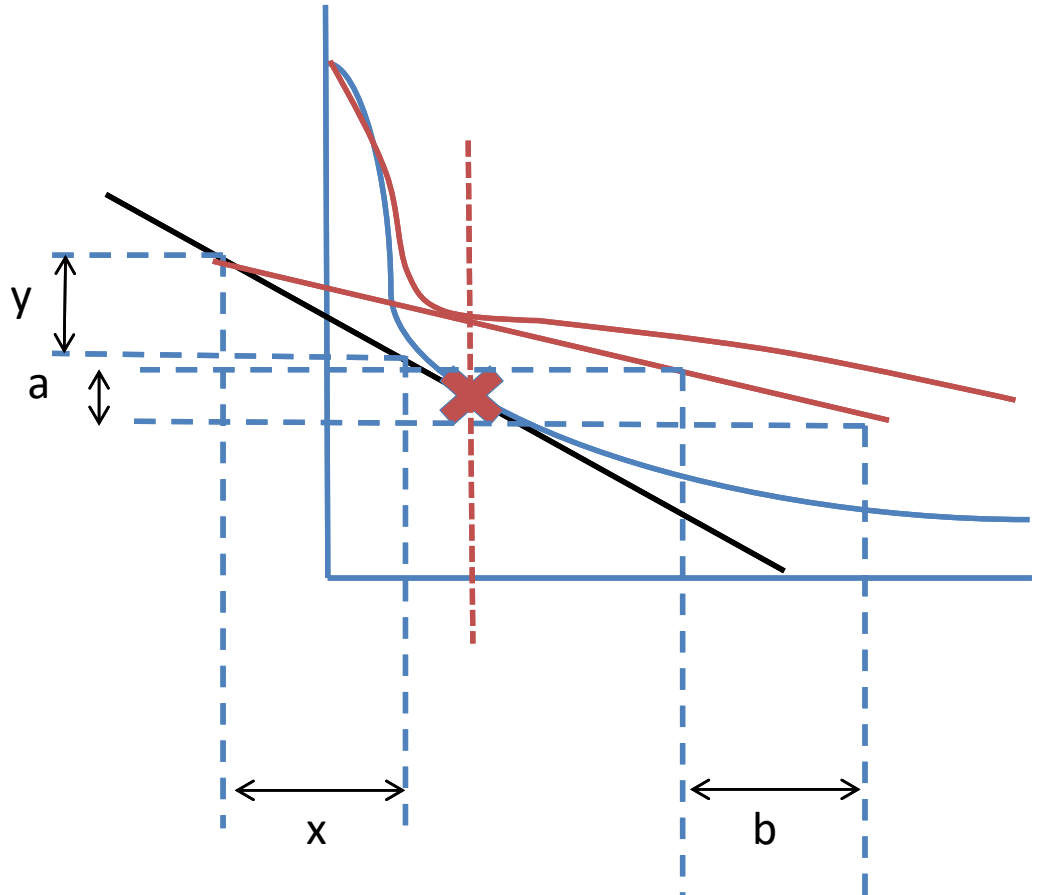
Eg. 10%/6months : 20%/6months

= 0.5

Or ‘patients are dying half as often as in control arm

Proportionate hazards

- Ratio between y/x and a/b does not change over life of curves



SAMPLE SIZE

Questions

- What effect does increase in alpha (ie. Increased risk of type 1 error) have on sample size?
- What effect does increase in beta (increased risk of type II error) have on sample size?
- If we increase the minimum effect of interest, (delta), what effect does that have?

Types of incorrect conclusions (type I/II error)

Two types of incorrect conclusion can occur, and these are classed as random (statistical) errors:[1,2]

Type I

Incorrectly concluding that there is a difference, where none exists. Rejecting the null hypothesis that is true (we observe a **false positive**).

Type II

Incorrectly concluding that there is no difference, where difference exists. Failing to reject the null hypothesis (we observe a **false negative**).

Type I and Type II error rates:

- **α** is the probability of making a type I error and is usually set to **0.05** (5%); the value of α should be fixed in advance, and is part of the study design*[3,4]

- **β** is the probability of making a type II error and is often set at **0.20** (**1- β** is termed the **statistical power** and values of 0.80 [80%] are desirable)[4]

*Please refer to slide notes for additional information. **1.** Rothman KJ. Eur J Epidemiol. 2010;25(4):223–224; **2.** Akobeng AK. J Pediatr Gastroenterol Nutr. 2008;47(3):277–282;
3. Greenland S et al. Eur J Epidemiol. 2016;31(4):337–350; **4.** Hulley SB. et al. Designing clinical research, 3rd ed. Philadelphia (PA): Lippincott Williams and Wilkins, 2007, 56–63.

Type I/II error in clinical trials

Type I and II errors differ between randomised phase II and III trials:

Phase II

Type I error (α):

- A higher α is usually acceptable (10–20%) to allow for relatively low patient numbers while still obtaining enough data to inform the decision to proceed with a phase III trial
- The consequence of a type I error is the treatment proceeding to a negative phase III trial

Type II error (β):

- Typically low to minimise obtaining a false negative

Phase III

Type I error (α):

- Typically low (compared to the α accepted in phase II trials) to minimise obtaining a false positive
- The consequence of a type I error is an ineffective treatment being deemed effective

Type II error (β):

- Typically higher than in phase II trials to increase power (the probability to detect a treatment effect)

Cannistra SA. J Clin Oncol. 2009;27(19):3073–3076.

Questions

- What effect does increase in alpha (ie. Increased risk of type 1 error) have on sample size?
- What effect does increase in beta (increased risk of type II error) have on sample size?
- If we increase the minimum effect of interest, (delta), what effect does that have?

Interim analyses

Alpha (α) risk inflation (type I error rate)

Alpha risk: When a set of hypotheses are tested, there is a risk of incorrectly concluding that there is a difference, where none exists e.g. a risk of making a type I error (**false positive**).[1]

Alpha risk increases when hypotheses are tested simultaneously within the same study e.g.[1]

- Analyses of multiple outcomes
- Multiple analyses of the same outcome at different times
- Sub-group analyses

To avoid a risk situation, authors should use statistical methods that take alpha risk inflation into consideration and, therefore, multiple comparisons.*[2]

*Please refer to slide notes for additional information.

1. Li G et al. Int J Epidemiol. 2017;46(2):746–755; **2.** Sham PC & Purcell SM. Nat Rev Genet. 2014;15(5):335–346.

INTENTION TO TREAT ANALYSES

Who counts?

N screened

And why does it matter?

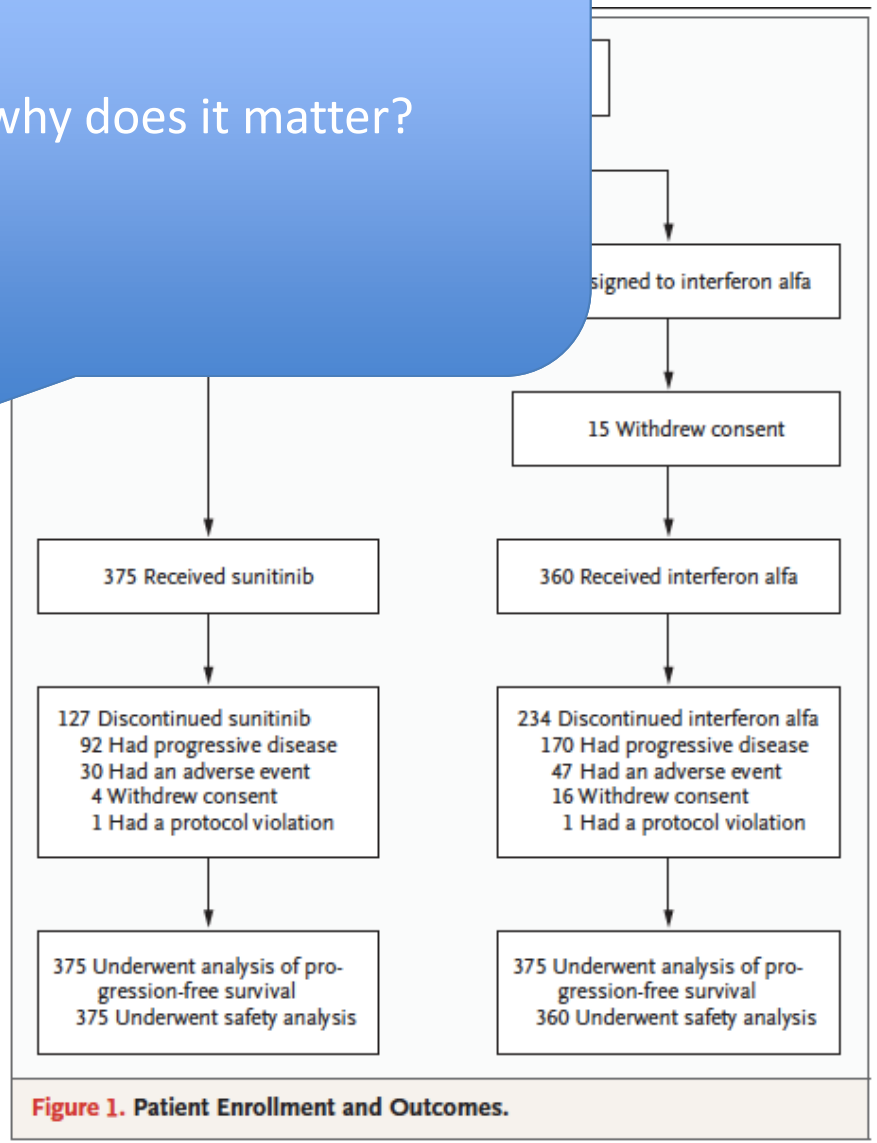


Figure 1. Patient Enrollment and Outcomes.

Motzer et al. NEJM 2007

robert.jones@glasgow.ac.uk

Back up slides

Non-Inferiority Studies (1/8)

Rationale for non-inferiority (NI) studies vs superiority studies

A new drug might offer greater efficacy (**superiority**) or it might promise easier administration, greater safety, convenience or less expense but with similar efficacy (**NI**).[1,2]

Superiority trial:

Aims to demonstrate that a new treatment is better than an active control or placebo.[1]

vs

NI trial:

Aims to demonstrate that a new treatment has an equivalent efficacy to the active control. The design is commonly used when it is **not ethical** to include a placebo or no-treatment control.[1,2]

NI study goals:[1,2]

Demonstrate that the new drug is **not unacceptably worse** than the active control by a specified amount (the **NI margin**) with a given degree of confidence.

1. Head SJ et al. Eur Heart J. 2012;33(11):1318–1324; **2.** Mauri L et al. N Engl J Med. 2017;377(14):1357–1367.

Non-Inferiority Studies (2/8)

The null hypothesis based on FDA guidance

The null (H_0) and alternative hypotheses (H_a) in a placebo-controlled trial…

- H_0 states that the response to the new drug is **less than or equal to** the response to the placebo (there is **no difference** between comparing groups)
- H_a states that the response to the new drug is **greater than** to the placebo (**there is a difference** between comparing groups)

…correspond to a null hypothesis of inferiority and an alternative hypothesis of NI:

Hypothesis	Statistical test results	Implication
H_0	Active control – new drug ≥NI margin	New drug is inferior to active control by ≥NI margin
H_a	Active control – new drug <NI margin	New drug is inferior to active control by <NI margin

A statistical test is performed by comparing the **upper-bound** of the two-sided CI for (active control – new drug) with the NI margin (specified in advance). If the **upper-bound** of the CI is **<NI margin**, NI of the new drug relative to the active control is established.

FDA guidance for industry: non-inferiority clinical trials to establish efficacy. Available at: https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf [Accessed: Oct 2021].

Non-Inferiority Studies (3/8)

Selection of the NI margin

The choice of NI margin is estimated based on historical data and/or clinical judgment, and is not measured in the trial.

Option 1:

Set the margin equal to **entire known effect** of the active control relative to placebo (largest possible margin).

Option 2 (desirable):

Set the margin equal to a **clinically relevant portion** of the entire known effect, reflecting the largest loss of effect that would be clinically acceptable.

Margin size	Consequence
Too small	• **Upper-bound** of the two-sided 95% CI for (active control – new drug) must be lower • Larger sample size needed to establish NI
Too large	• A false conclusion of NI of the new drug vs the active control

FDA guidance for industry: non-inferiority clinical trials to establish efficacy. Available at: https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf [Accessed: Oct 2021].

Non-Inferiority Studies: REFLECT (4/8)

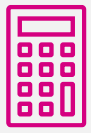
Selection of the NI margin: Example*

An open-label, phase III, NI trial of patients with previously untreated, unresectable hepatocellular carcinoma (HCC) randomised (1:1) to receive:[1]

- 8 mg/day **lenvatinib** (body weight [BW] <60 kg) or 12 mg/day (BW ≥60 kg) OR 400 mg twice daily **sorafenib**

The primary endpoint of OS was first tested for NI, then for superiority.[1] The NI margin was based on historical data and clinical judgement:

- Data from two previous phase III sorafenib trials yielded a pooled OS HR (0.69) and 95% CI (0.57–0.83) for sorafenib vs placebo[2,3]



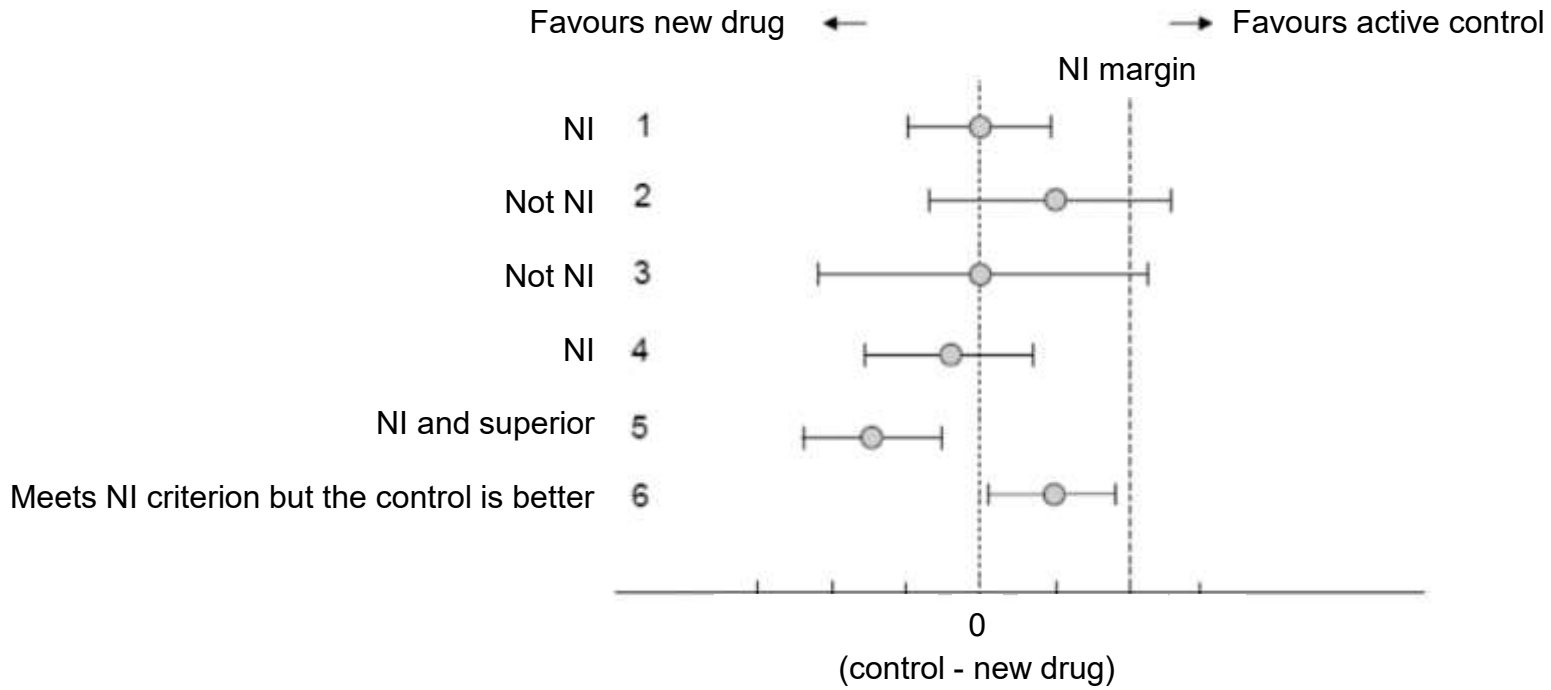
The **lower-bound** of the two-sided 95% CI of log HR was used to initially calculate the **entire known effect** (largest possible NI margin). The NI margin was then further specified and the NI margin corresponding to **60%** retention of sorafenib effect vs placebo was calculated to be **1.08**.[1,2]

NCT01761266. *Please refer to slide notes for additional information; **1.** Kudo M et al. Lancet. 2018;391(10126):1163–1173; **2.** Study 304 CSR, Section 9.7.1.6, Pages 86–87; **3.** Llovet JM et al. N Engl J Med. 2008;359(4):378–390.

Non-Inferiority Studies (5/8)

Determining NI based on the NI margin

Example results showing differences between the active control and new drug (point estimate and 95% CI):*



This example uses the largest possible NI margin value (option 1). A finding of NI means that the new drug has an effect >0 but the effect of an unacceptable loss of the active control cannot be ruled out.

*Please refer to slide notes for additional information.

FDA guidance for industry: non-inferiority clinical trials to establish efficacy. Available at: https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf [Accessed: Oct 2021].

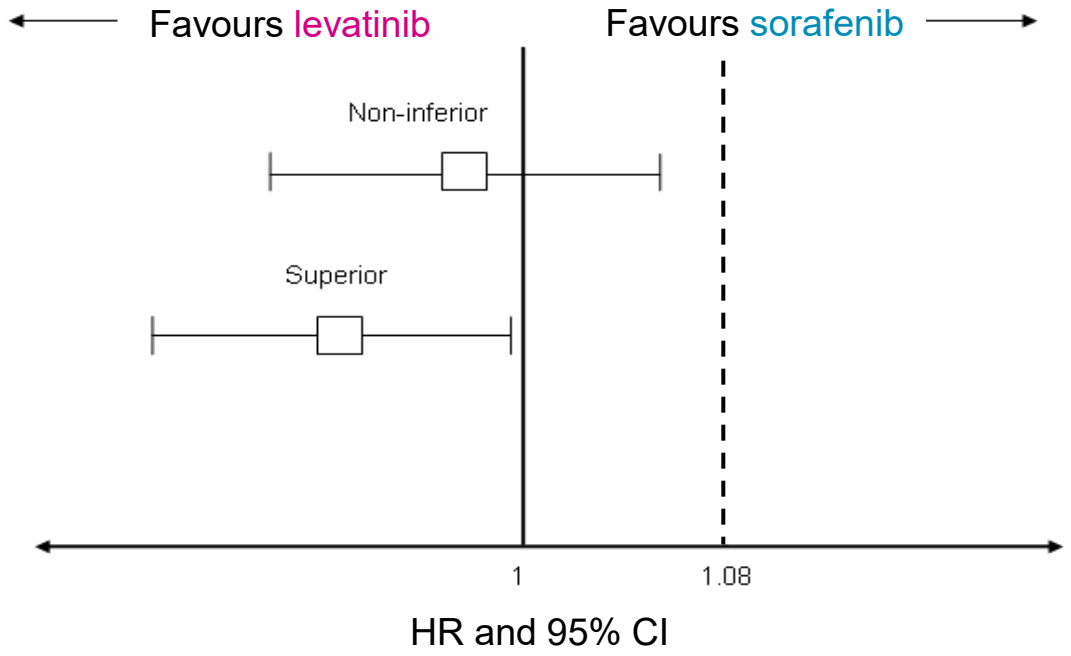
Non-Inferiority Studies: REFLECT (6/8)

Determining NI: Example

The trial would be successful if the **upper-bound** of the 95% CI for the HR (lenvatinib/sorafenib) <NI margin (1.08):

- NI of lenvatinib vs sorafenib would be inferred (60% preservation of sorafenib effect vs placebo)
- Superiority of lenvatinib vs placebo would be (indirectly) demonstrated

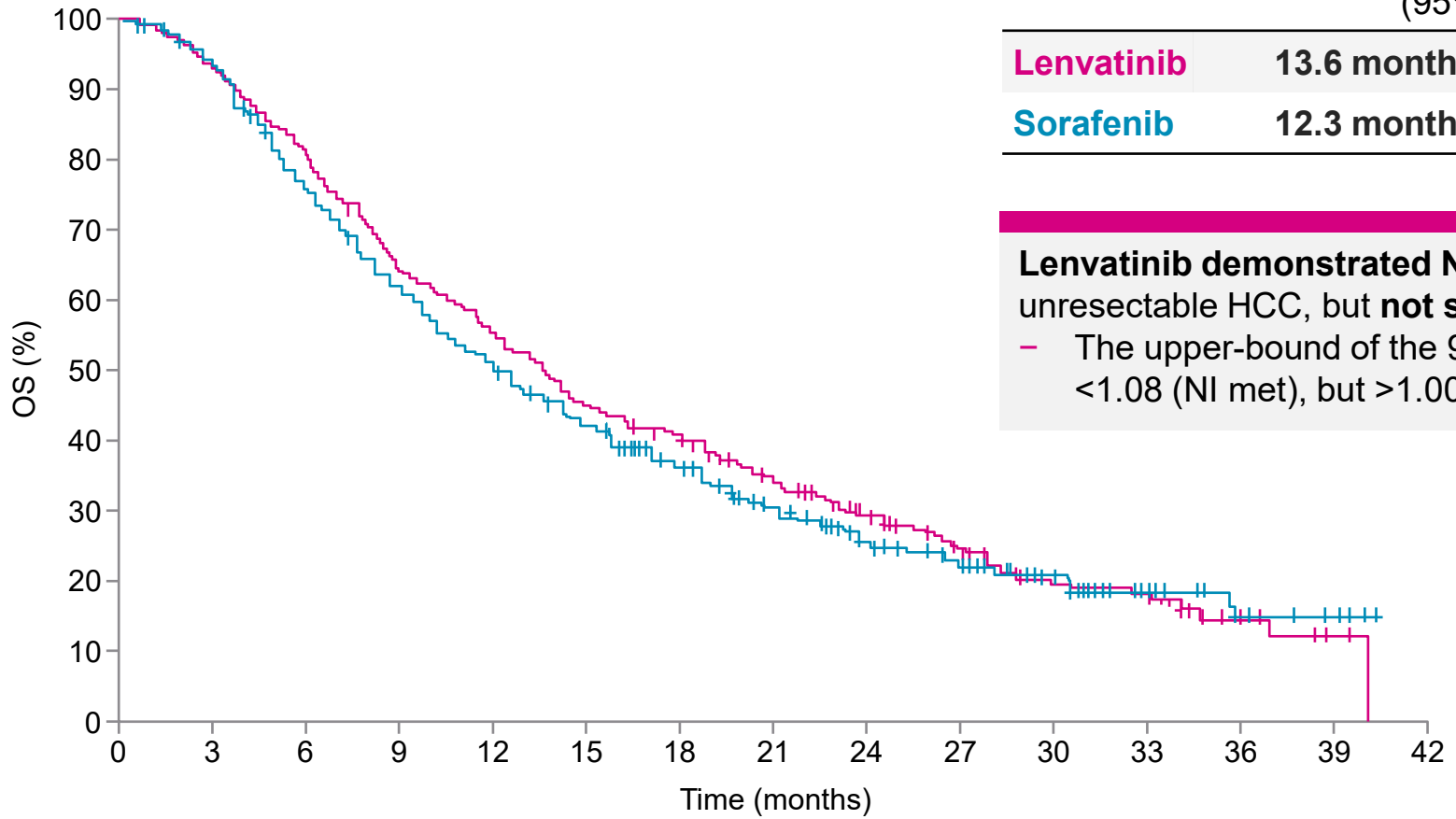
Additionally, if the 95% CI for the HR (lenvatinib/sorafenib) **<1.00**, then **superiority of lenvatinib vs sorafenib** can be claimed.



NCT01761266.
Study 304 CSR, Section 9.7.1.6, Pages 86–87.

Non-Inferiority Studies: REFLECT (7/8)

REFLECT results



	Median overall survival (95% CI)	HR (95% CI)
Lenvatinib	**13.6 months** (12.1, 14.9)	**0.92** (0.79, **1.06**)
Sorafenib	**12.3 months** (10.4, 13.9)	

Lenvatinib demonstrated NI vs sorafenib in OS in untreated unresectable HCC, but **not superiority:**

- The upper-bound of the 95% CI for the HR was 1.06 e.g., <1.08 (NI met), but >1.00 (not superior)

Number of patients at risk

Lenvatinib	478	436	374	297	253	207	178	140	102	67	40	21	8	2	0
Sorafenib	476	440	348	282	230	192	156	116	83	57	33	16	8	4	0

NCT01761266.
Adapted from Kudo M et al. Lancet. 2018;391(10126):1163–1173.

Statistical Considerations: Subgroups (2/4)

Testing for interaction

The interaction between the treatment and the subgroup baseline/demographic factor can be interpreted as **effect-measure modification**, also referred to as **effect heterogeneity**.[1]

The use in determining whether there is heterogeneity is to identify the subgroups in which treatment is most/least effective.[1] Subgroup analysis should focus on differences from the overall treatment effect via tests of heterogeneity or **interaction**.[2]

Two misinterpretations to avoid:[2]

1. Attributing an effect to a subgroup when there is no overall effect and no evidence for heterogeneity
2. Claiming lack of effect in a subgroup when the overall effect is significant

1. Brankovic M et al. Eur J Clin Invest. 2019;49(8):e13145; **2.** Cuzick J. Lancet 2005;365(9467):1308.

Statistical Considerations: Subgroups (3/4)

Testing for interaction: STAMPEDE*

STAMPEDE, a phase III RCT in which patients with newly diagnosed metastatic prostate cancer were randomised (1:1) to standard of care (control group) or standard of care and radiotherapy (radiotherapy group), provides an example of interaction testing.

The primary outcome was OS. Two prespecified subgroup analyses tested the effects of prostate radiotherapy by baseline metastatic burden (**low** vs **high**) and radiotherapy schedule (**daily** vs **weekly**).

Radiotherapy did not improve OS for all patients. However, it did improve OS in the subgroup of patients with **low metastatic burden** (HR 0.68, 95% CI 0.52–0.90; p=0.007; 3-year survival 73% [control] vs 81% [radiotherapy]).

Using an interaction test, there was some evidence of heterogeneity of treatment effect by metastatic burden (interaction p=0.0098). This result suggested a low likelihood that the apparent subgroup effect could be due to chance.

*NCT00268476.
Parker C et al. Lancet 2018;392(10162):2353–2366.