in partnership with

Interpreting clinical trial results

Professor Judith Bliss
Clinical Trials and Statistics Unit (ICR-CTSU)
The Institute of Cancer Research, London

Phases of clinical trials

Phase 1: **TOXICITY**

What is the maximum tolerated dose (MTD)?

safety, 3+3 vs. more complex dose escalation procedures eg continual reassessment methods (CRM), size of expansion cohorts

Phase 2: **ACTIVITY**

Does it do anyone any good?

establishing sufficient evidence of activity to justify phase III, formal stop/go criteria, single group or randomised

Phase 3 **THERAPEUTIC BENEFIT**

Is it any better than existing treatment?

efficacy comparison with standard of care, robust results with the potential to change practice, choice of endpoints, risks & benefits

Ultimate goal is to change routine clinical practice & target treatment towards those patients with the most to gain

Trial considerations: effect size

Superiority

what is the minimum clinically important improvement in efficacy with new treatment compared with standard treatment?

- e.g. treatment A is at least 6% better than treatment B

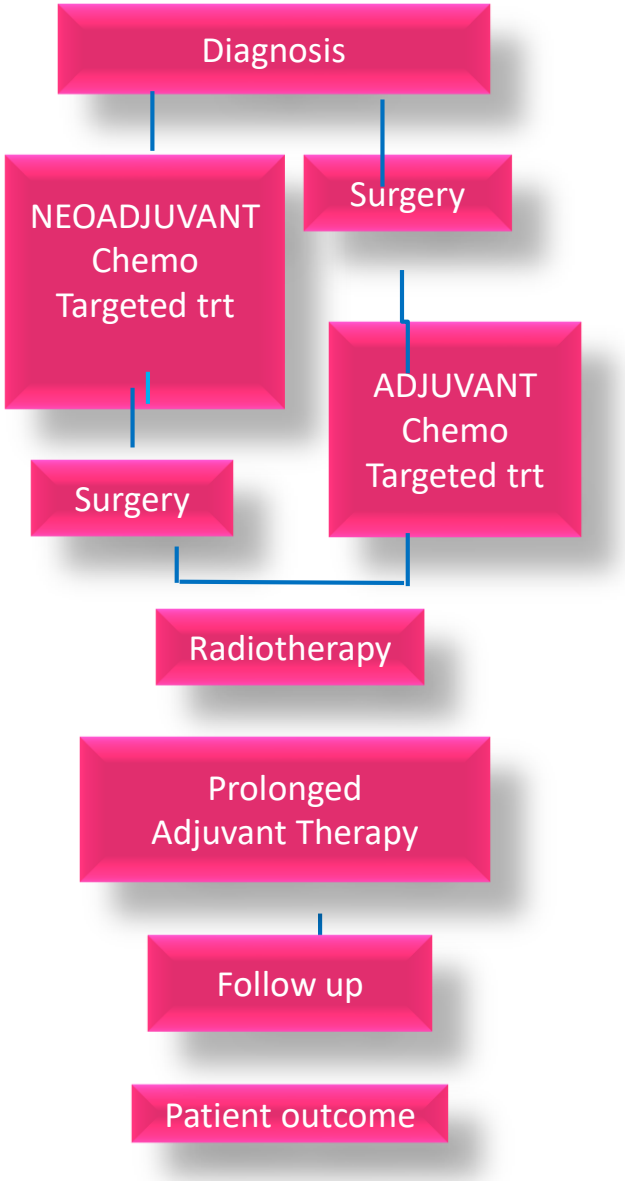
Non-inferiority

show that new treatment is not worse than standard by more than pre-specified, small amount (non-inferiority margin)

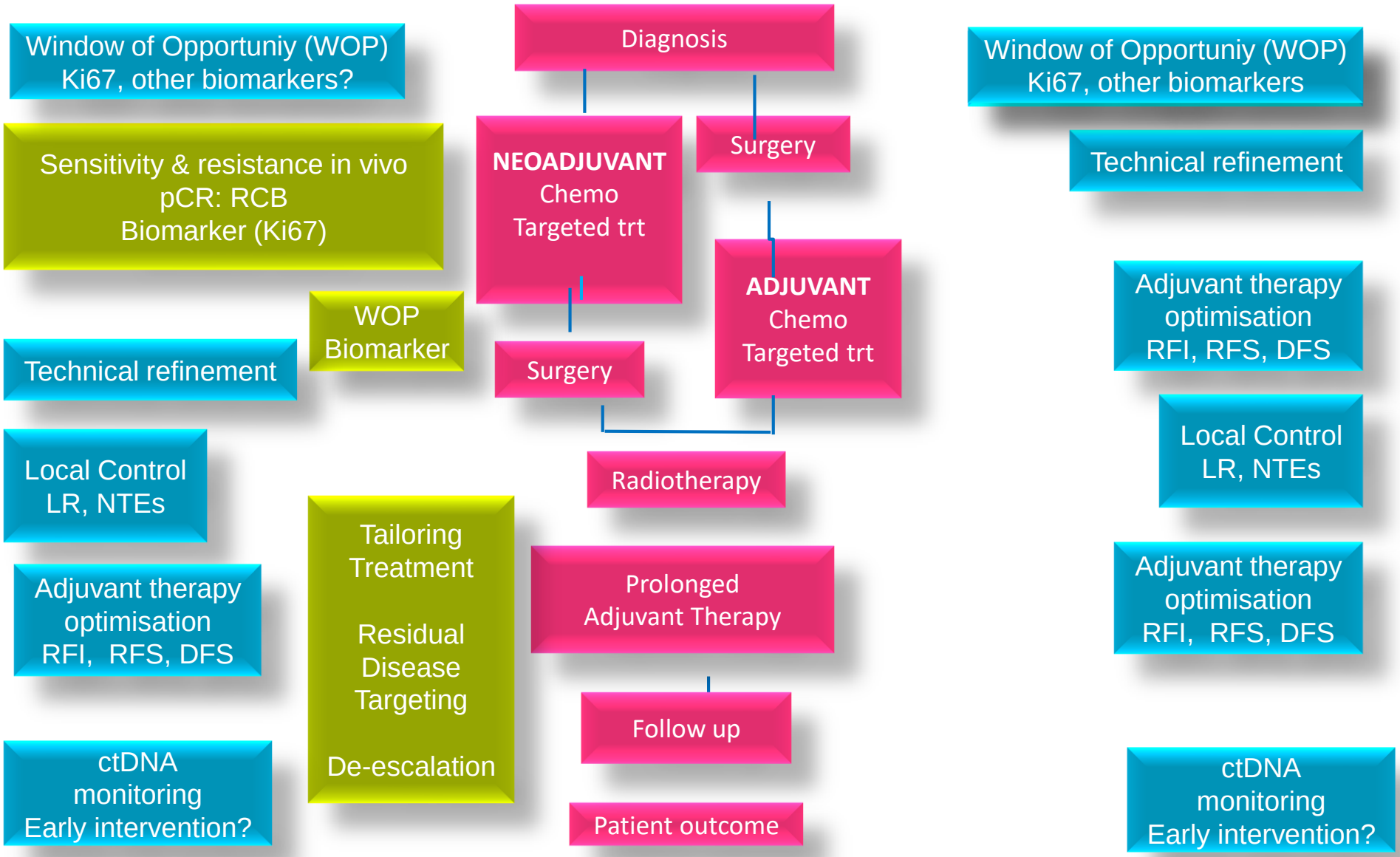
- e.g. treatment A is no more than 3% worse than treatment B

Smaller effect size → larger sample size

Early breast cancer: “patient pathway”



Early breast cancer: “experimental settings”



Types of outcome measures used - Endpoints

Disease outcomes

- Relapse-free survival (RFS) / Disease-free survival (DFS) / Relapse-free interval (RFI
- Includes as "events" when a patient has a breast cancer recurrence, develops a new cancer, or dies
- TIME TO EVENT ENDPOINT – Kaplan Meier plot (graph), Logrank test, Hazard ratio (HR)

Response to treatment

- Response rate (RR) / pCT rate / Clinical Benefit Rate (CBR)
- Measures how much a tumour/s has changed in size
- CATEGORICAL OR BINARY ENDPOINT - % responders, % change in tumour size, Odds Ratio (OR)

Patient reported outcomes

- Quality of Life (QL), treatment related symptoms, Impact on Activities of Daily Living, Well-being
- EORTC QLQ C-30, FACT-B, HADS, EQ5D
- QUESTIONAIRE BASED –CONTINUOUS SCORES AVERAGED OR % RESPONDERS

Biomarkers

- Ki67, ctDNA+, Apoptosis, PEPI score
- Often exploratory
- CONTINUOUS SCORES AVERAGED OR % RESPONDERS

Statistical considerations in clinical trials

At the concept/design stage (pre-funding application)

Trial design:

Treatment allocation method – randomisation / blinding

Stratification variables - centre / biomarkers

Protecting against other sources of bias

Endpoints – clinically informative, reliable & valid measurement?

Sample size – “study appropriately powered and minimise random errors”

Power (1-β) = probability of detecting a difference if such a difference truly exists

Significance level (α) = probability of concluding there is a difference when no difference exists

Power = 80%- 90% **α = 0.05** (usual)

Statistical considerations in clinical trials

Statistical Analysis Plan defines plans and scope for

During the running of the trial

Trial monitoring

- Data quality & completeness

Interim analyses *(for review by Independent Data Monitoring Committee)*

- Review of emerging data - safe & ethical to continue?
- Futility assessment

Analysis

Analysis of primary endpoint

- maturity of data, ITT or PP populations

Estimate of treatment effect & of precision of estimate

- 95% confidence interval

Subgroups/exploratory or hypothesis generating analyses

- Multiplicity Adjustment

Trial considerations: Null hypothesis

- **It is simpler to set out to <u>disprove</u> a hypothesis than to prove it**

  e.g. in a metastatic breast cancer trial of A vs B:
  Response rate A = 53% Response rate B = 20%

  The **null hypothesis** is that the treatments are **equally** effective in the population of all metastatic breast cancer patients (there is no true difference in response rates)

  The *alternative hypothesis* is that there *is* a true difference in response rates for A & B.
  Note: difference could be in either direction; alternative hypothesis is "*2-sided*"

Statistical fundamentals: Significance test

- After defining the **null hypothesis**, the main question is:

  If the null hypothesis were true, what are the chances of getting a difference at least as big as that observed?

  e.g. in the breast cancer trial, if there really is no true difference between the 2 drugs in terms of tumour response, what is the probability of observing a treatment difference as large (or even larger than) 53% versus 20%?

- This **probability** (the **p-value**) is determined by applying an appropriate statistical significance test

- There are different significance tests for different types of data, but the principle is the same

Statistical fundamentals: Significant or not significant?

Arbitrary cut-off of $p<0.05$ often used to indicate statistical significance, but better to present exact p-values & interpret accordingly

e.g. would you interpret p=0.04 very differently from p=0.06?

Note!!!

"Not significant" does not automatically mean that there is no actual difference (we can't *prove* the null hypothesis), but merely that we have been unable to show evidence of a difference with certainty

i.e. "No evidence of an effect" is NOT the same as "evidence of no effect" – this is subtle but important

Reasons for non-significant results include: no true difference in the population, sample size may be too small, estimates too imprecise, bias

Statistical fundamentals: Statistical versus clinical significance

Size of the p-value depends on observed difference & sample size

- If sample size is small, results may produce a p-value which is not statistically significant, even if there is actually a large true difference
- If sample size is large, small observed differences (which may be clinically irrelevant) may achieve statistical significance
- Need to think about what size differences are **clinically important** in order to interpret statistical significance results sensibly

e.g. supposing we found a mean difference in weight of 2kg between 2 groups of patients

In a small study, this difference might not be statistically significant, but in a much larger study might be highly statistically significant. So what?!

Need to use clinical judgement to decide whether 2kg is clinically important (not a statistical decision)

Statistical fundamentals: Confidence intervals & hypothesis testing (1)

Significance tests (p values) help us decide whether or not study results are compatible with a *hypothesis*

BUT they provide *no* information on the *size of the difference*

e.g. in the breast cancer trial, the 33% difference in tumour response rates was statistically significant with p<0.001

Confidence intervals help us to estimate the **size** of the difference with some measure of precision

e.g. 95% CI for the 33% difference in response rates in the breast cancer trial is:

95% Confidence Interval (20.5% to 45.5%)

i.e. effectively 95% confident that real difference between A & B tumour response is between 20.5% & 45.5%

Statistical fundamentals: Confidence intervals & hypothesis testing (2)

- There is a link between p-values & CIs
- If 95% CI for a *difference* between groups **does not include** the null hypothesis value of 0, then $p<0.05$
- If the 95% CI **includes** the null hypothesis value, $p>0.05$

In the e.g., the null hypothesis is that there is no difference between the tumour response rates in the population (i.e. the null hypothesis value = 0)

Does the 95% CI for the 33% difference in response rates include 0?

No (20.5% to 45.5%), so we can infer that $p<0.05$

Randomised Clinical Trials – superiority trials

Aim: to demonstrate that EXP is ***better*** to ST

Endpoint: e.g. Disease-free survival (recurrence, deaths)

Analysis: e.g. Hazard ratio & 95% confidence interval, p value

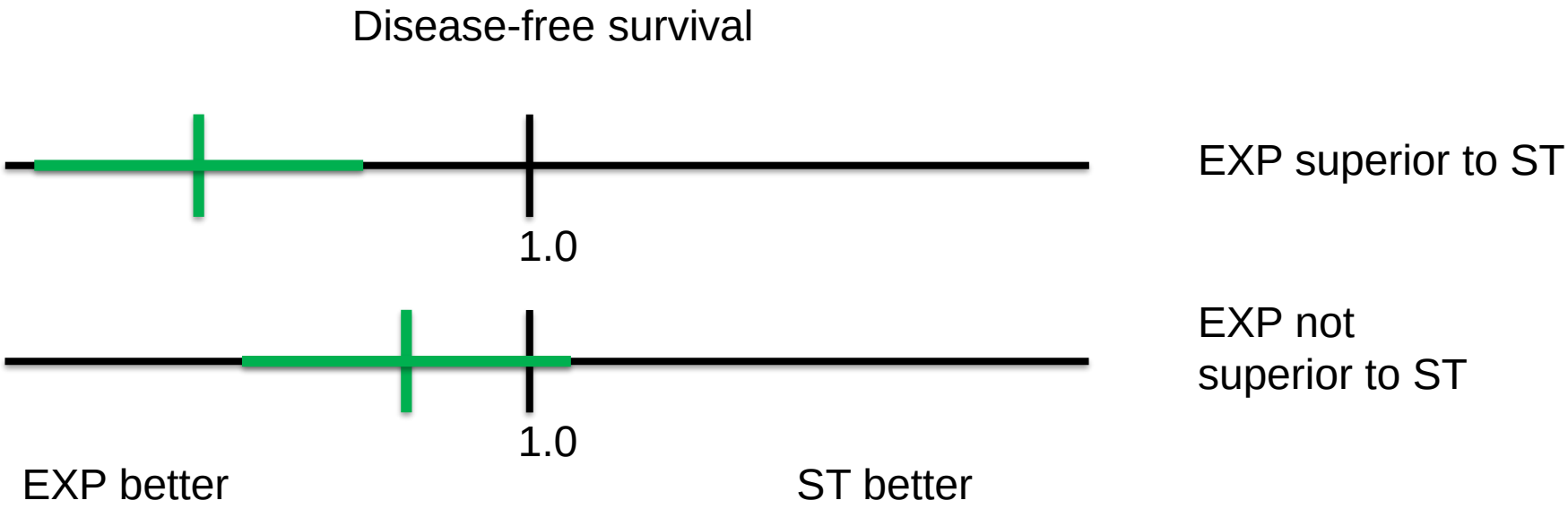
HR = 0.62 (95%CI 0.50-0.77) p<0.001 - clear-cut benefit

HR = 0.78 (95%CI 0.62-0.99), p=0.04 - marginal

Superiority/Non-Inferiority

- When the aim of a trial is to demonstrate that an experimental treatment (EXP) is superior to standard treatment (ST) this is called a **superiority trial**.
- If Δ be the difference in treatment effects e.g. EXP / ST
- H_0: Δ=1.0
- H_1: $\Delta \neq 1.0$

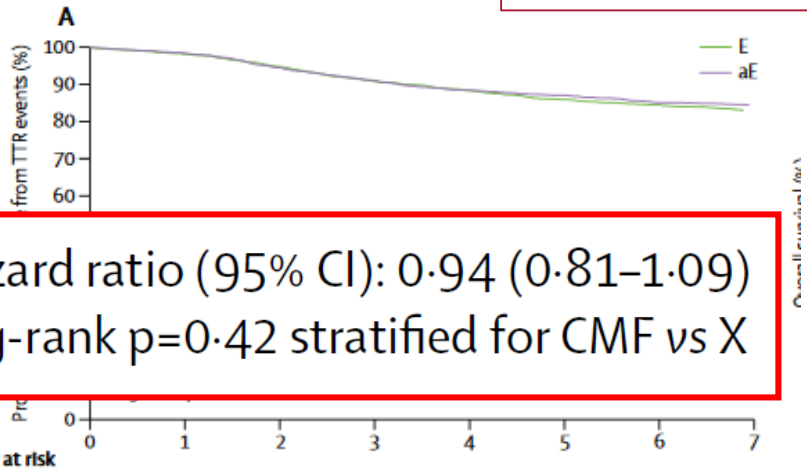
Conduct the trial, estimate Δ with 95% CI



Accelerated versus standard epirubicin followed by cyclophosphamide, methotrexate, and fluorouracil or capecitabine as adjuvant therapy for breast cancer in the randomised UK TACT2 trial (CRUK/05/19): a multicentre, phase 3, open-label, randomised, controlled trial

David Cameron, James P Morden, Peter Canney, Galina Velikova, Robert Coleman, John Bartlett, Rajiv Agrawal, Jane Banerji, Gianfilippo Bertelli, David Bloomfield, A Murray Brunt, Helena Earl, Paul Ellis, Claire Gaunt, Alexa Gillman, Nicholas Hearfield, Robert Laing, Nicholas Murray, Niki Couper, Robert C Stein, Mark Verrill, Andrew Wardley, Peter Barrett-Lee, Judith M Bliss, on behalf of the TACT2 Investigators

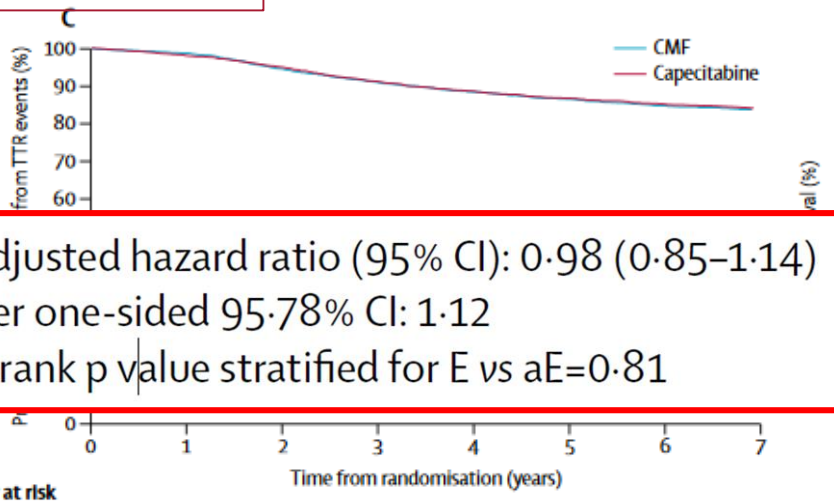
KAPLAN MEIER PLOTS



Hazard ratio (95% CI): 0·94 (0·81–1·09)
Log-rank p=0·42 stratified for CMF *vs* X

Number at risk (number of events)								
E	2221 (40)	2155 (75)	2073 (86)	1975 (52)	1914 (55)	1823 (30)	1645 (26)	1098 (13
aE	2170 (35)	2115 (82)	2023 (74)	1941 (59)	1865 (26)	1810 (40)	1597 (40)	1170 (20

Numbers censored								
E	0	26	33	45	54	90	238	759
aE	0	20	30	38	55	84	257	744



Unadjusted hazard ratio (95% CI): 0·98 (0·85–1·14)
Upper one-sided 95·78% CI: 1·12
Log-rank p value stratified for E *vs* aE=0·81

Number at risk (number of events)								
CMF	2178 (27)	2138 (91)	2035 (77)	1950 (55)	1882 (40)	1813 (38)	1618 (17)	1117 (17*)
X	2180 (43)	2117 (65)	2047 (82)	1953 (56)	1884 (40)	1809 (32)	1614 (20)	1084 (16*)

Numbers censored								
CMF	0	13	25	33	46	75	232	716
X	0	20	25	37	50	85	248	758

Lancet Oncol 2017; 18: 929–45

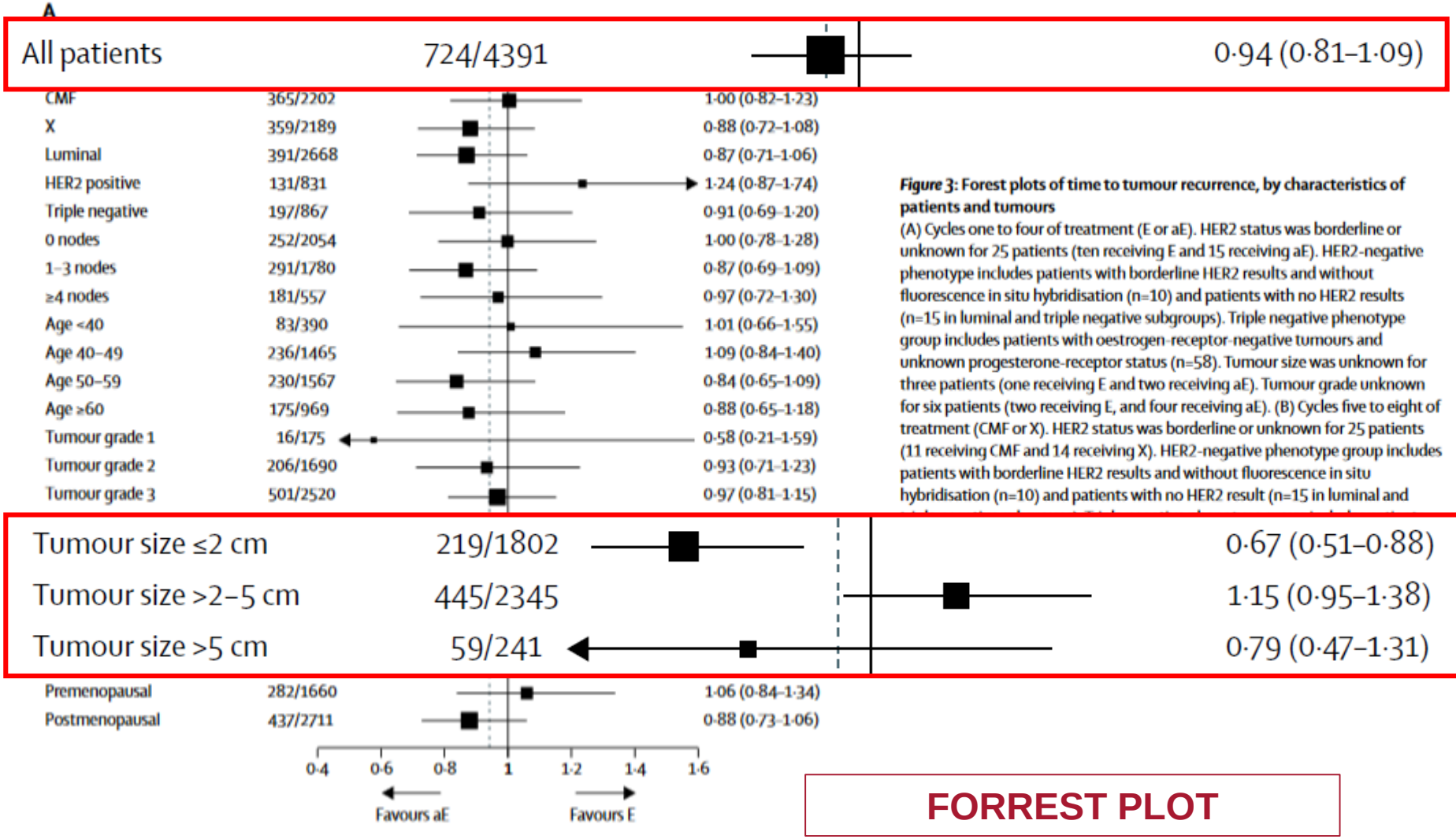


***Figure 3*: Forest plots of time to tumour recurrence, by characteristics of patients and tumours**

(A) Cycles one to four of treatment (E or aE). HER2 status was borderline or unknown for 25 patients (ten receiving E and 15 receiving aE). HER2-negative phenotype includes patients with borderline HER2 results and without fluorescence in situ hybridisation (n=10) and patients with no HER2 results (n=15 in luminal and triple negative subgroups). Triple negative phenotype group includes patients with oestrogen-receptor-negative tumours and unknown progesterone-receptor status (n=58). Tumour size was unknown for three patients (one receiving E and two receiving aE). Tumour grade unknown for six patients (two receiving E, and four receiving aE). (B) Cycles five to eight of treatment (CMF or X). HER2 status was borderline or unknown for 25 patients (11 receiving CMF and 14 receiving X). HER2-negative phenotype group includes patients with borderline HER2 results and without fluorescence in situ hybridisation (n=10) and patients with no HER2 result (n=15 in luminal and

FORREST PLOT

Randomised Clinical Trials – Non-inferiority trials

Aim: to demonstrate that EXP is ***no worse*** than to ST

Randomised Clinical Trials – Non-inferiority trials

Aim: to demonstrate that EXP is ***not substantially worse (no clinically meaningful loss of effect)*** than to ST

Endpoint: e.g. Disease-free survival (recurrence, deaths)

Analysis: e.g. Hazard ratio & 95% confidence interval, p value

Pre- define: threshold of non-inferiority based on difference in event rates

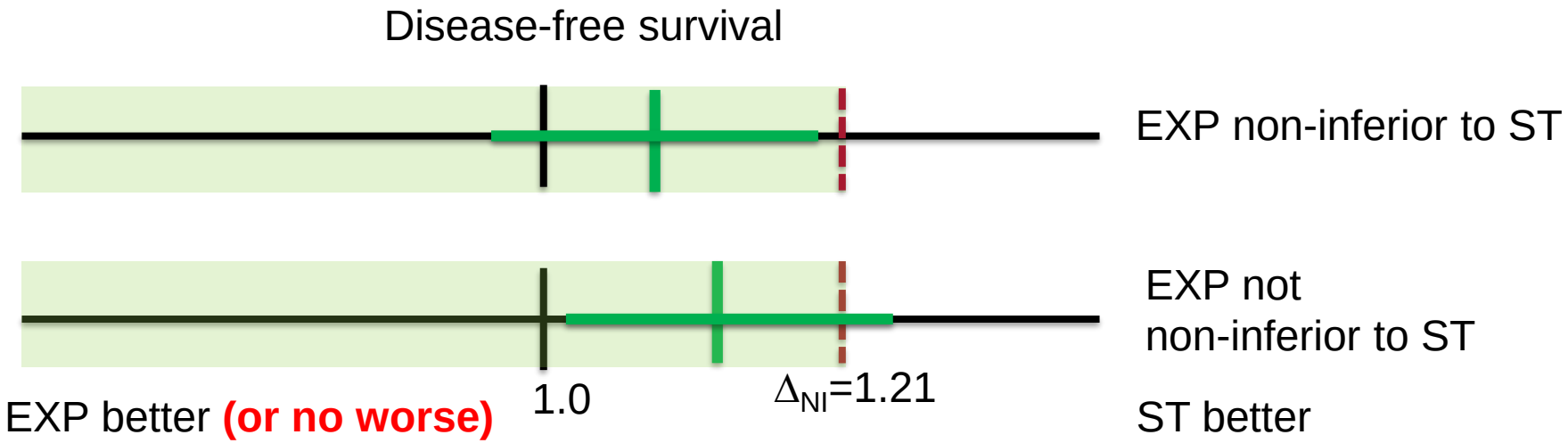
- Absolute - eg ≤2% EXP 94% vs 96% DFS or EXP 74% vs ST 76%
- Relative - eg HR ≤ 1.15 (15% increase in risk) or HR ≤ 1.30 (30% increase)

Superiority/**Non-Inferiority**

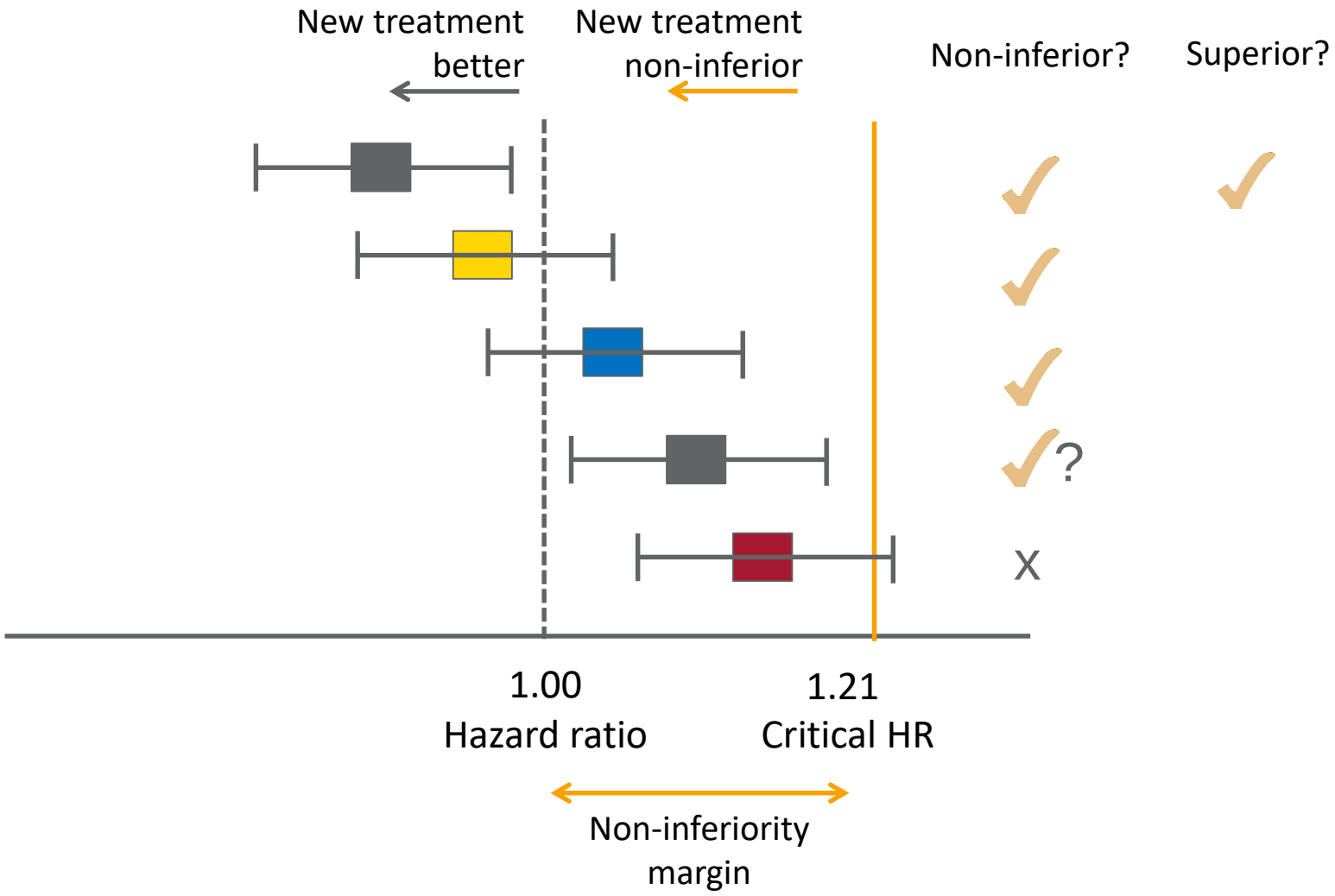
- Interested in demonstrating that a experimental treatment is not substantially worse than a current treatment. e.g when comparing shorter vs longer treatment
- Agree a threshold *before the start of the study* for "not substantially worse", Δ_{NI}
- H_0: $\Delta \geq \Delta_{NI}$.
- H_1: $\Delta < \Delta_{NI}$

e.g. $\Delta_{NI}=1.21$

Conduct the trial, estimate Δ with 95% CI

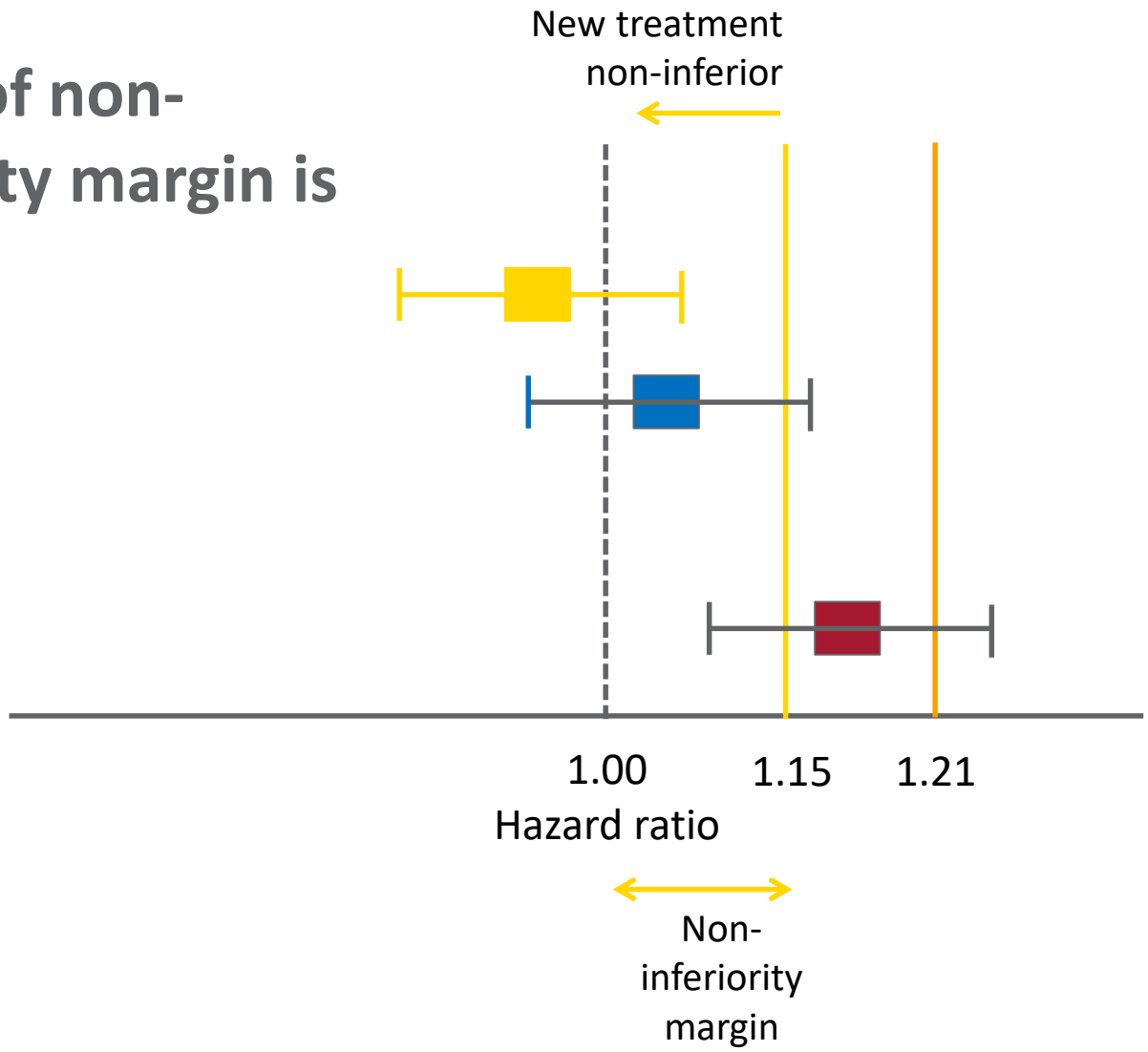


Testing for non-inferiority



Non-inferiority margin

Choice of non-inferiority margin is key



Interventional Cohort design

Aim: to demonstrate that EXP is ***no worse*** than a fixed outcome threshold

Endpoint: e.g Disease-free survival (recurrence, deaths)

Analysis: e.g DFS at (say) 5 years, 95% confidence interval, p value

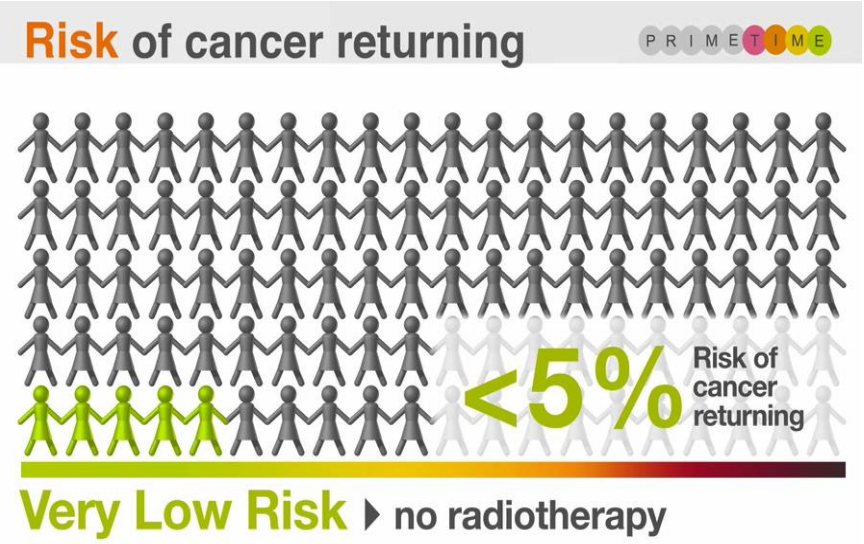
Pre- define: threshold DFS event-free

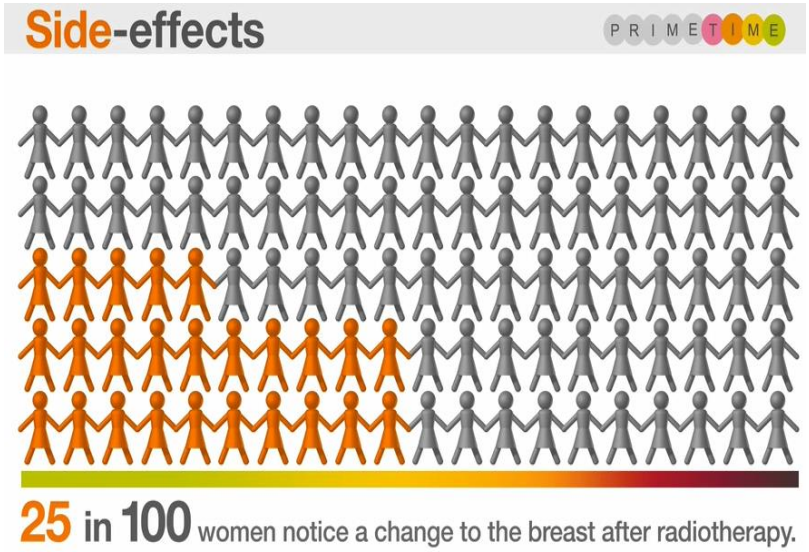
- ***92%*** 96% (95%CI **93**-98) p=0.02 94% (95%CI **91**-97) p=0.10
- ***72%*** 79% (95%CI **74**- 84) p=0.03 75% (95%CI **70**-80) p=0.15

De-escalation trials – risks vs benefits

What are the risks vs benefits of treatment?

- *How common is the risk? How common is the benefit?*
- *Are we talking about absolute or relative risks?*





De-escalation trials – considerations

What are the risks vs benefits of treatment?

What size of benefit are we prepared to "lose"?

How was the study analysed?

Is the endpoint sensitive to the important outcomes?

Is the threshold for establishing non-inferior outcome robust & well defined?

APT Trial 7 year follow up

In 410 patents, with a median follow-up of 6.5 yrs, there were 23 DFS events observed:

4 (1.0%) distant recurrences,
5 local/regional recurrences (1.2%),

6 new contralateral BC (1.5%),
8 deaths without documented recurrence (2.0%).

At 7-years
DFS was 93.3% (95% CI 90.4-96.2);
HR+ pts 94.6% (95% CI 91.8-97.5) or HR- pts 90.7% (95% CI 84.6-97.2).
RFI was 97.5% (95% CI 95.9-99.1);
BCSS is 98.6% (95% CI 97.0-100);
OS was 95.0% (95% CI 92.4-97.7).

Tolaney, ASCO 2017

Statistics – the fundamentals

Statistics is …about **understanding** data

It is NOT just about hypothesis testing and p-values - a **statistically** significant result may not be **clinically** important or vice versa

Confidence Intervals (95%) give information on the **precision** and clinical significance of an observed effect

Subgroup analyses

– open to abuse and mis-interpretation “the more you look the more you find” – adjustment for **multiple testing**, **biological plausibility,**

– quantitative vs qualitative treatment interactions - if overall trial **no effect**, identification of **sensitive subgroup** implies subgroup where treatment confers **harm**

Further reading

Books:

Clinical Trials. A Practical Approach. Stuart J Pocock. Wiley 1983

Cancer Clinical Trials. Methods and Practice. Edited by Marc Buyse, Maurice Staquet, Richard Sylvester. Oxford Medical Publications. 1984

Internet:

www-users.york.ac.uk/~mb55/pubs/pbstnote.htm BMJ Statistics Note series (Doug Altman & Martin Bland) OR on BMJ website (Research methods & reporting section)

www.ct-toolkit.ac.uk/ MRC DoH Clinical Trials Toolkit

http://csg.ncri.org.uk/portfolio/portfolio-maps/ cancer clinical studies within the NIHR portfolio

www.clinicaltrials.gov US NIH service – general information

And finally…….. (Power, P-values, publication bias, statistical evidence)

https://www.youtube.com/watch?v=kMYxd6QeAss