

EBCTCG Data Collection and Processing

EBCTCG process

- 
- Identify and prioritise key clinical question
 - Identify **all** relevant trials
 - Form a writing committee
 - Develop and agree a statistical analysis plan
 - Request data
 - Process and clean data
 - Perform analyses
 - Agree results with trialists
 - Present results at scientific meetings
 - Write paper, in conjunction with writing committee

EBCTCG process



- Identify all relevant trials



- Form a writing committee



- Develop and agree a statistical analysis plan



- Request data



- Process and clean data



- Perform analyses

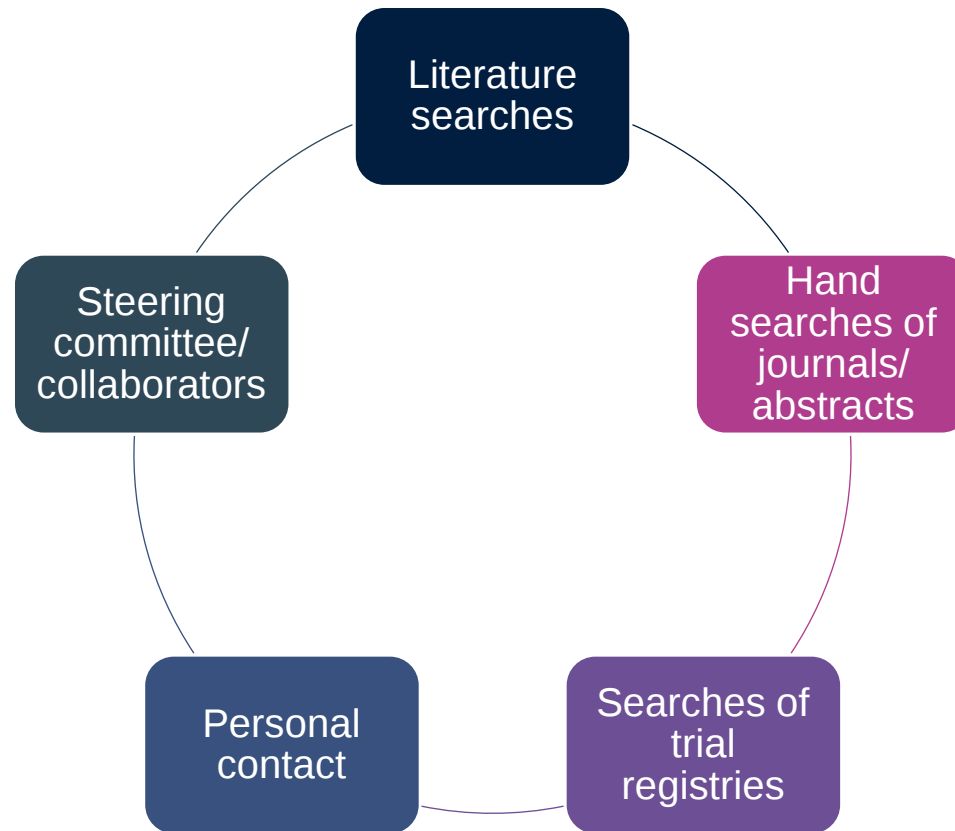


- Agree results with trialists



- Write paper, in conjunction with writing committee

Identifying trials



Requesting data – 1985 style

[illegible]

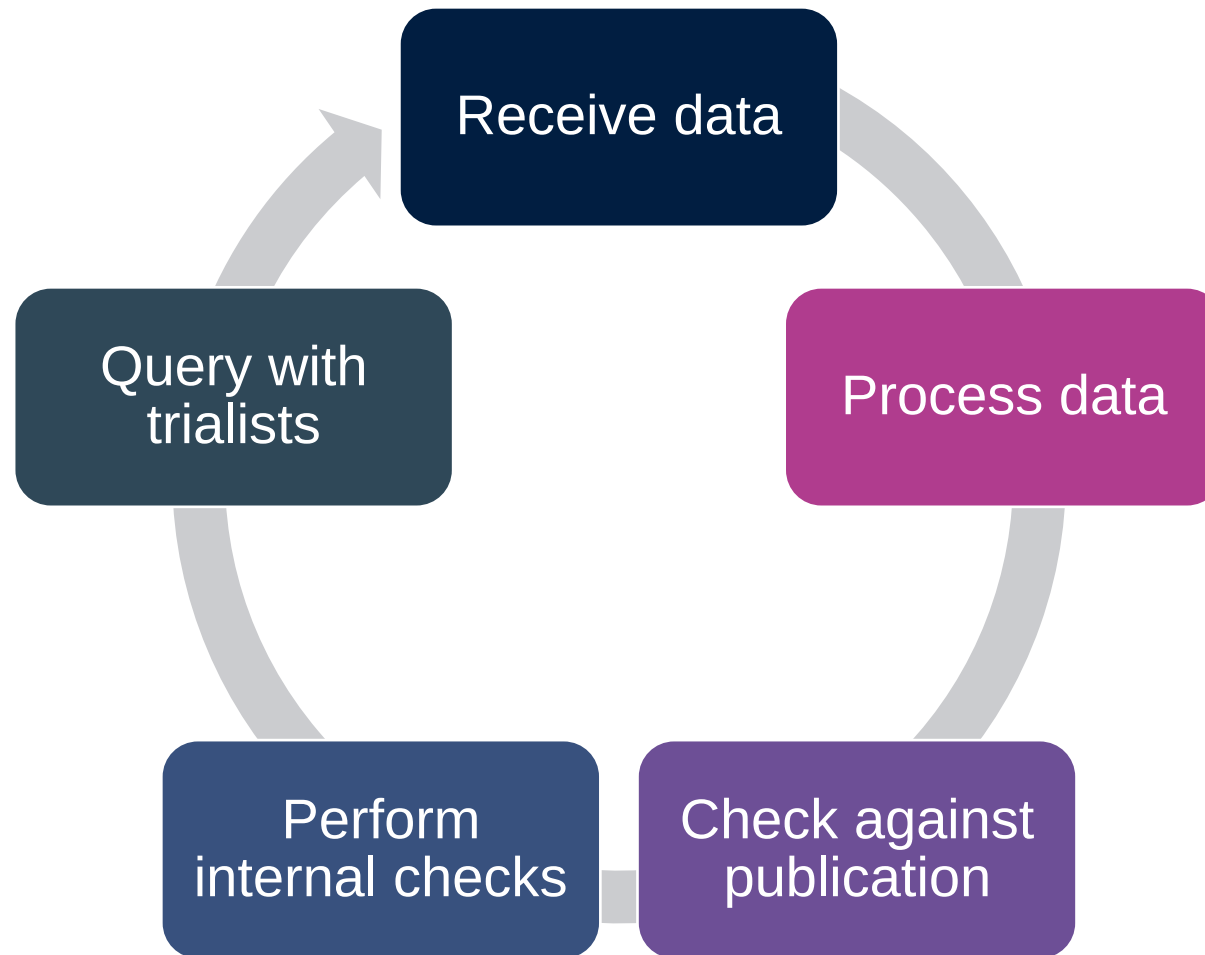
Requesting data – 2025 style

- Data now provided electronically
 - Mainly in Excel spreadsheets or SAS datasets
 - Sent either by secure email or file transfer processes
 - All datafiles stored securely at EBCTCG before conversion to main database format
- Occasionally get data from a portal – means we do not get IPD but the results of analyses on IPD – see Wednesday

Requesting data – 2025 style

- Our capacity to process data in any format reduces burden on trialists and makes it much easier to get all the available evidence
- Some data (e.g. adverse events) may be recorded incidentally or in formats specific to a trial
- Important to interpret raw data correctly

Processing and cleaning incoming data



Publication checking

- Check our figures match those reported
- May reveal differences in baseline/endpoint definitions
- Does the published data have some information we do not?
- Can information from the publication be used to fill in gaps?

Logical checks

- Many logical checks run automatically
 - Baseline characteristics balanced by arm?
 - Events in the right order?
 - Values within plausible (possible) ranges?
 - Breast cancer deaths without recurrence?

Data warnings:

Very old yet still living patients

PID	Age at Last Follow-up	Status
447	100	1
921	102	1
1288	101	1
1510	101	1

Patients with very large tumours

PID	Size
844	210

Compliant patients with NonComp date

Patient identifier	Non-compliance before any recurrence	Date of first Non-compliance/completion of protocol therapy
451	No	2004-07-14

Patients with distant recurrence but no date

Patient identifier	Site of first distant recurrence	Date of first distant recurrence
968	Only visceral (req. 7)	NA
1872	Only visceral (req. 7)	NA

Patients with contralateral recurrence but no date

Patient identifier	Invasive contralateral breast cancer	Date of first contralateral recurrence
78	Yes	NA
737	Yes	NA
748	Yes	NA
1316	Yes	NA

High rate of contralateral recurrences

Contralateral	n	Proportion
1	1368	90%
2	154	10%

Patients with date before randomisation

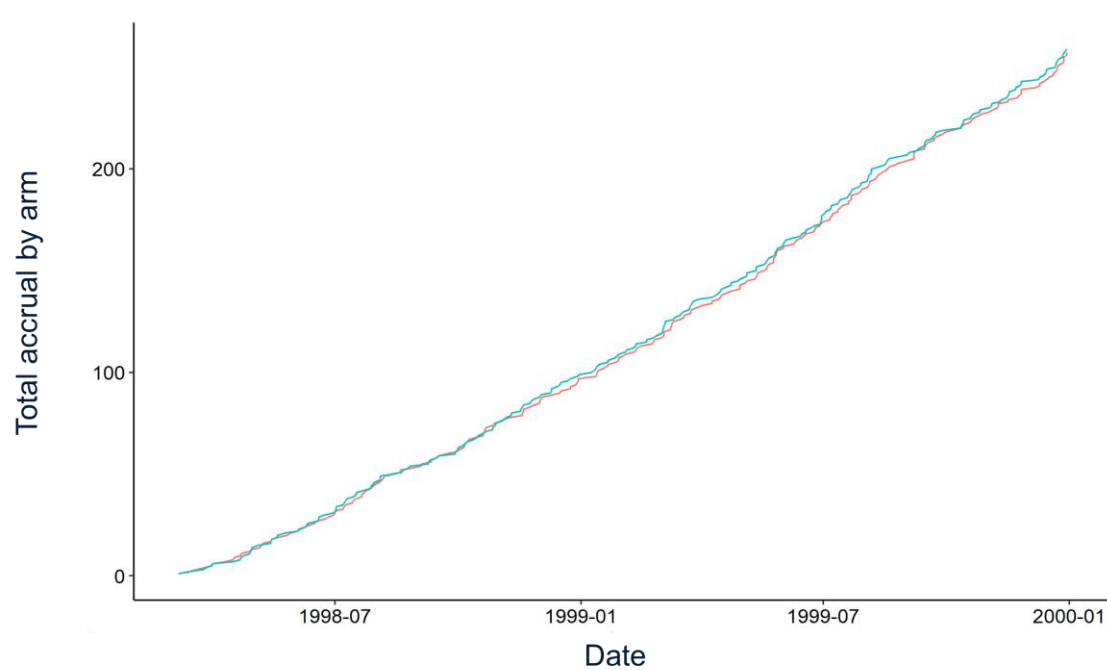
PID	RandomDate	NonCompDate
1053	1995-11-24	1995-11-21
1100	1995-12-21	1995-11-15
1342	1996-03-15	1996-03-04
1536	1996-05-23	1996-05-21

Patients with date after date last seen

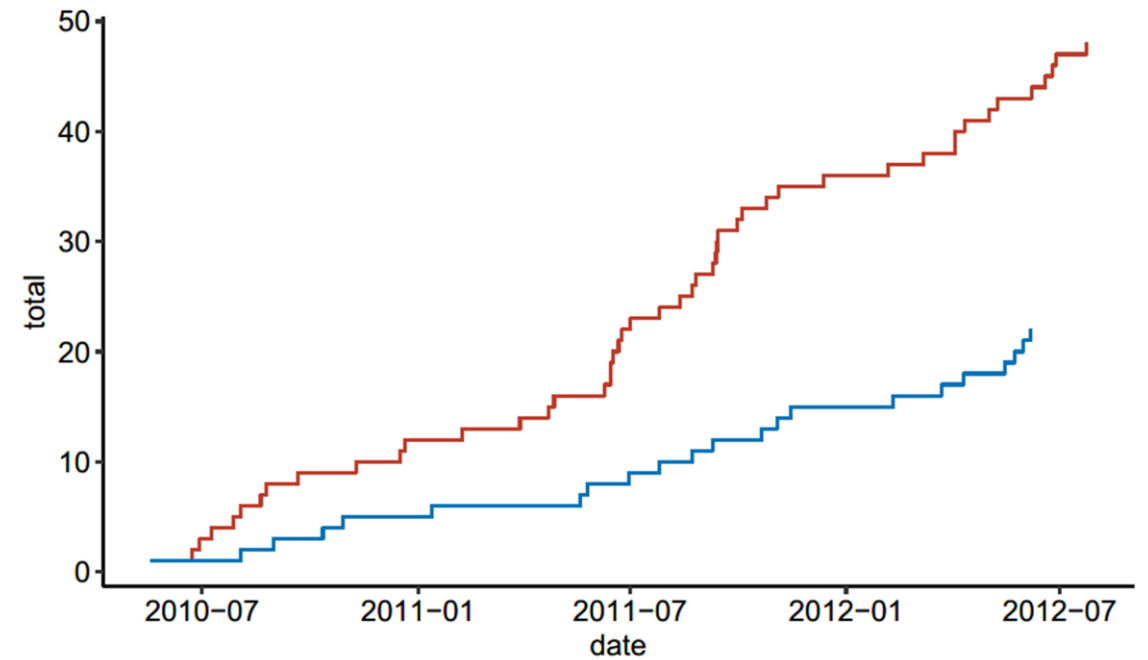
PID	NonCompDate	DistDate	LastSeeDate	DeathDate	HeartDate
253	2004-09-15	NA	2004-01-12	2004-01-12	NA
326	1996-05-15	2006-07-17	2004-09-12	2004-09-12	NA
887	NA	2009-04-23	2009-06-07	2009-06-07	2009-06-15

Recruitment checks

Randomised 1:1



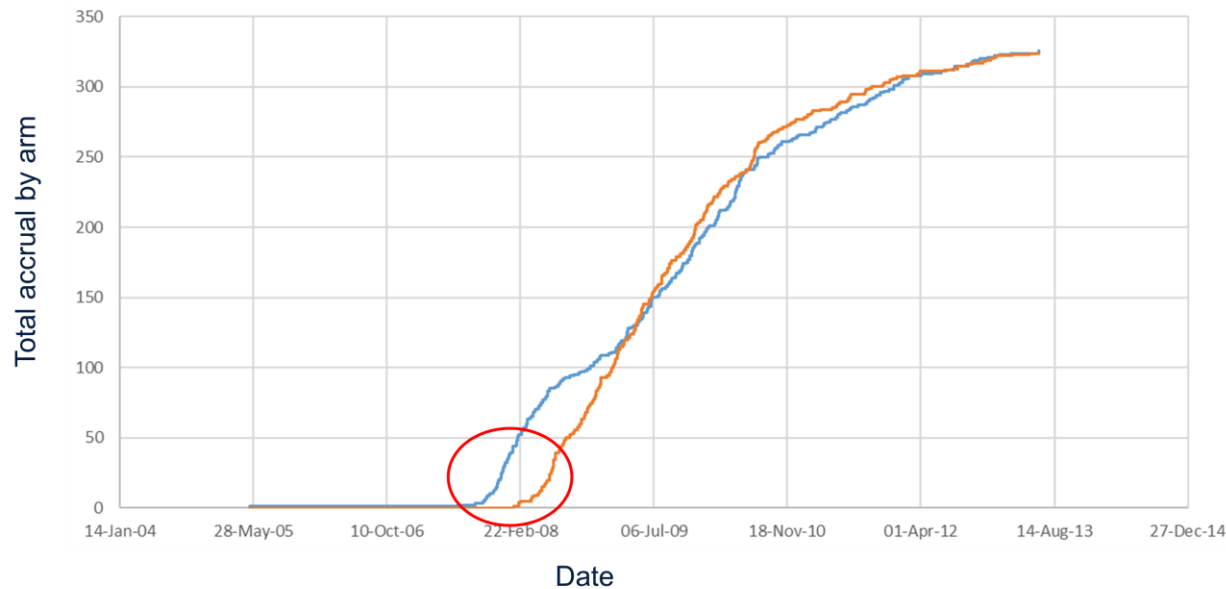
Randomised 2:1



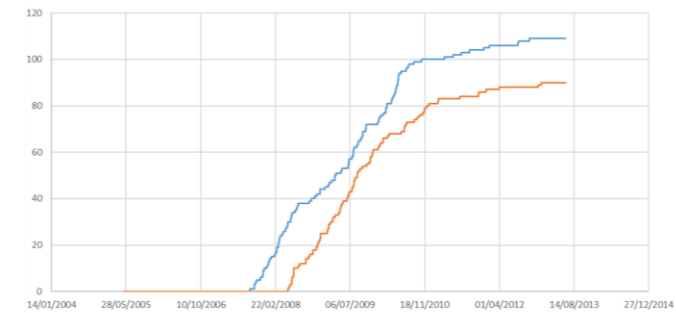
Recruitment checks

Totals *look* balanced - **but:**

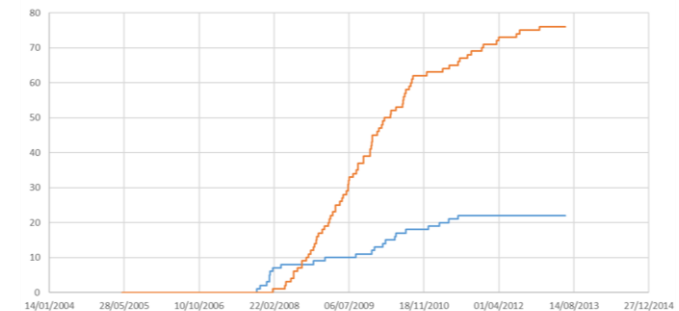
- One arm opens early
- Discrepancies by age



Younger

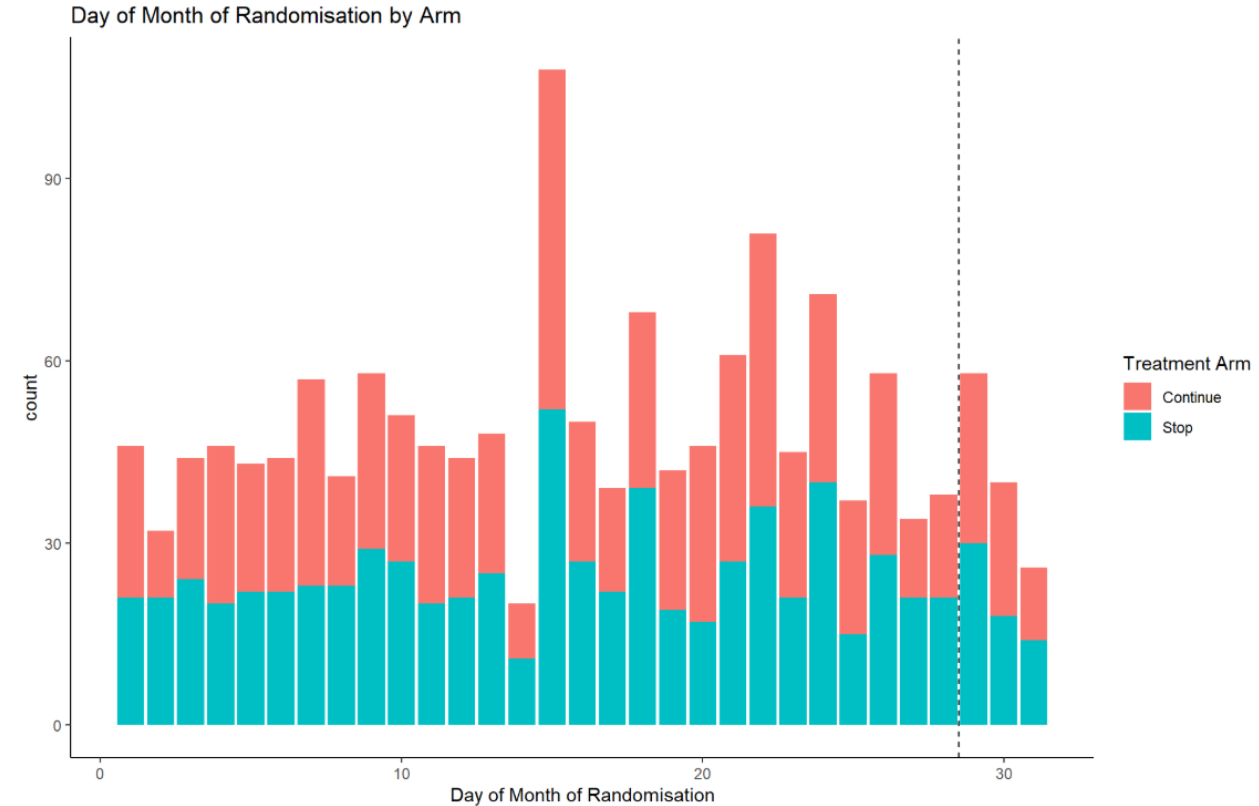
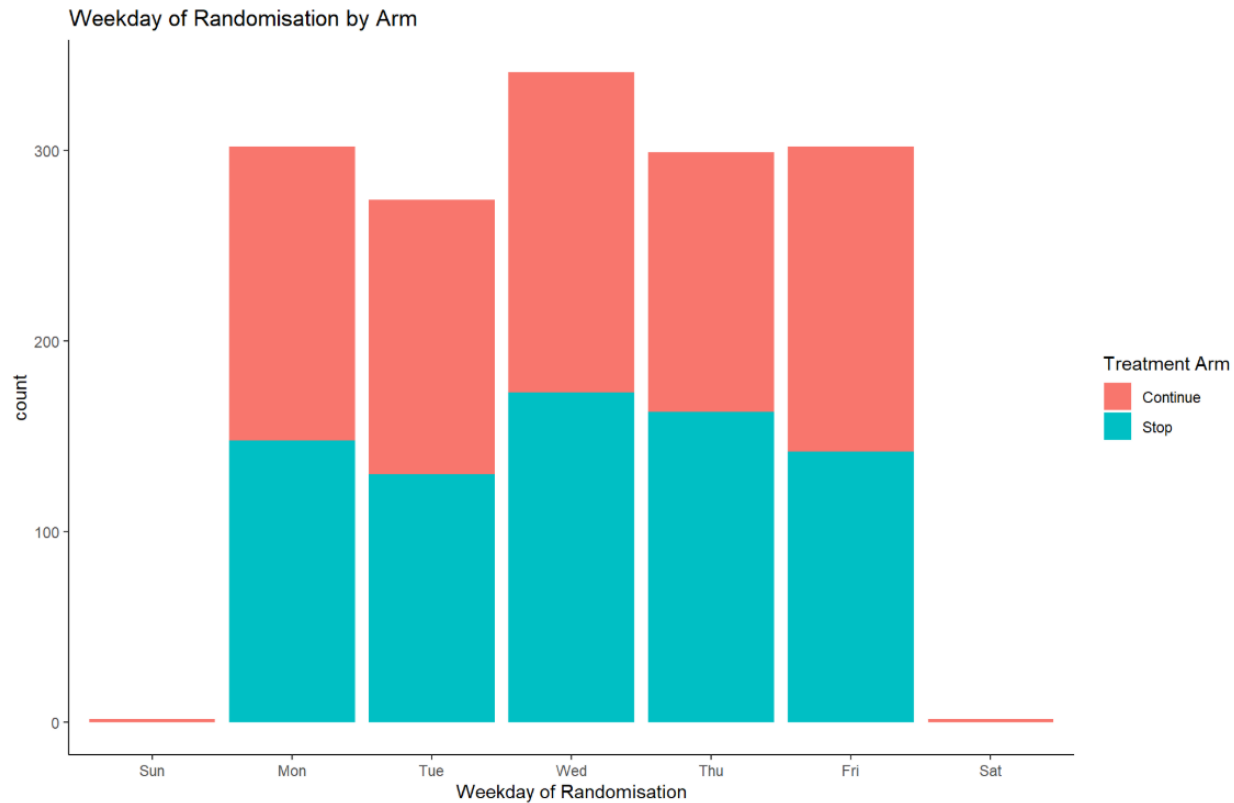


Older



Recruitment checks

Day of week & day of month randomised



Strange data patterns

Age	No. of patients
-----	-----------------

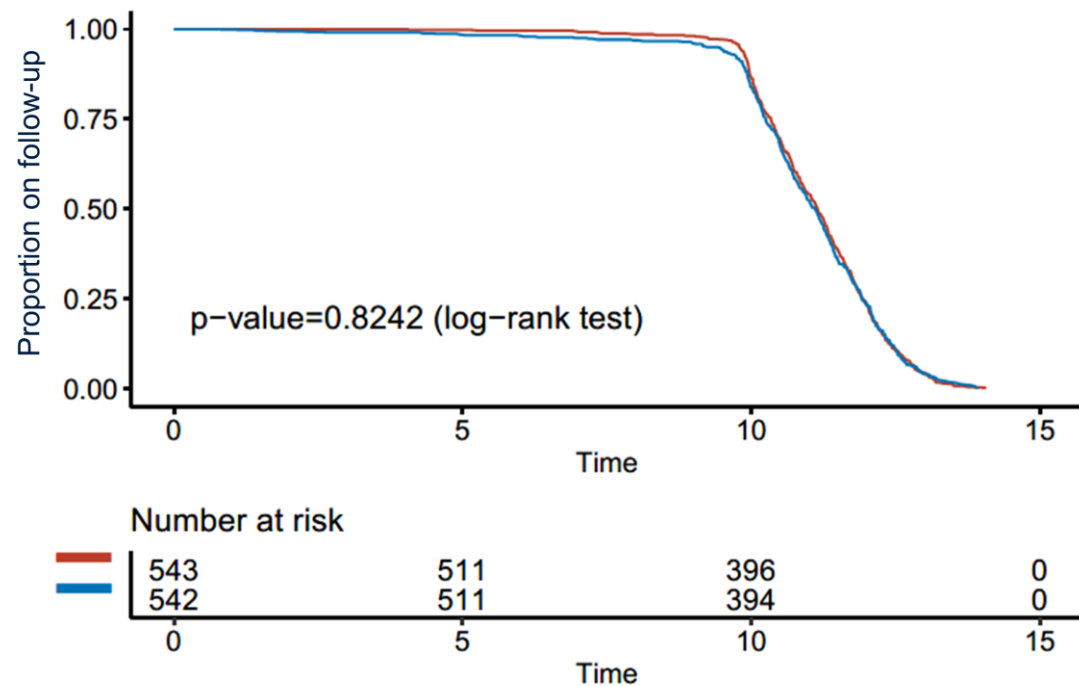
35	6
36	35
37	25
38	18
39	59
40	5
41	2
42	12
43	9
44	54

39 is a very good age. I know many women in London who of their own choice have remained 39 for several years.

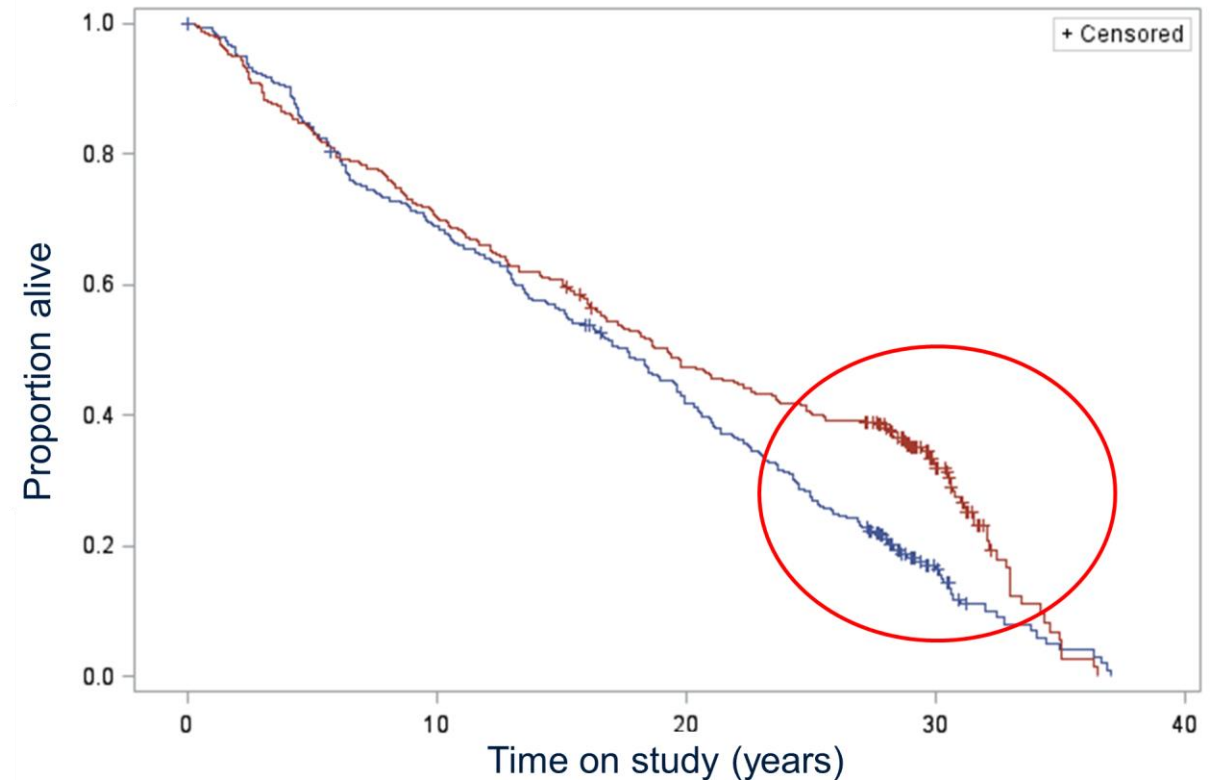
Lady Bracknell in Oscar Wilde's play
"The Importance of Being Earnest"

Checking follow-up

Pattern of follow-up identical between arms

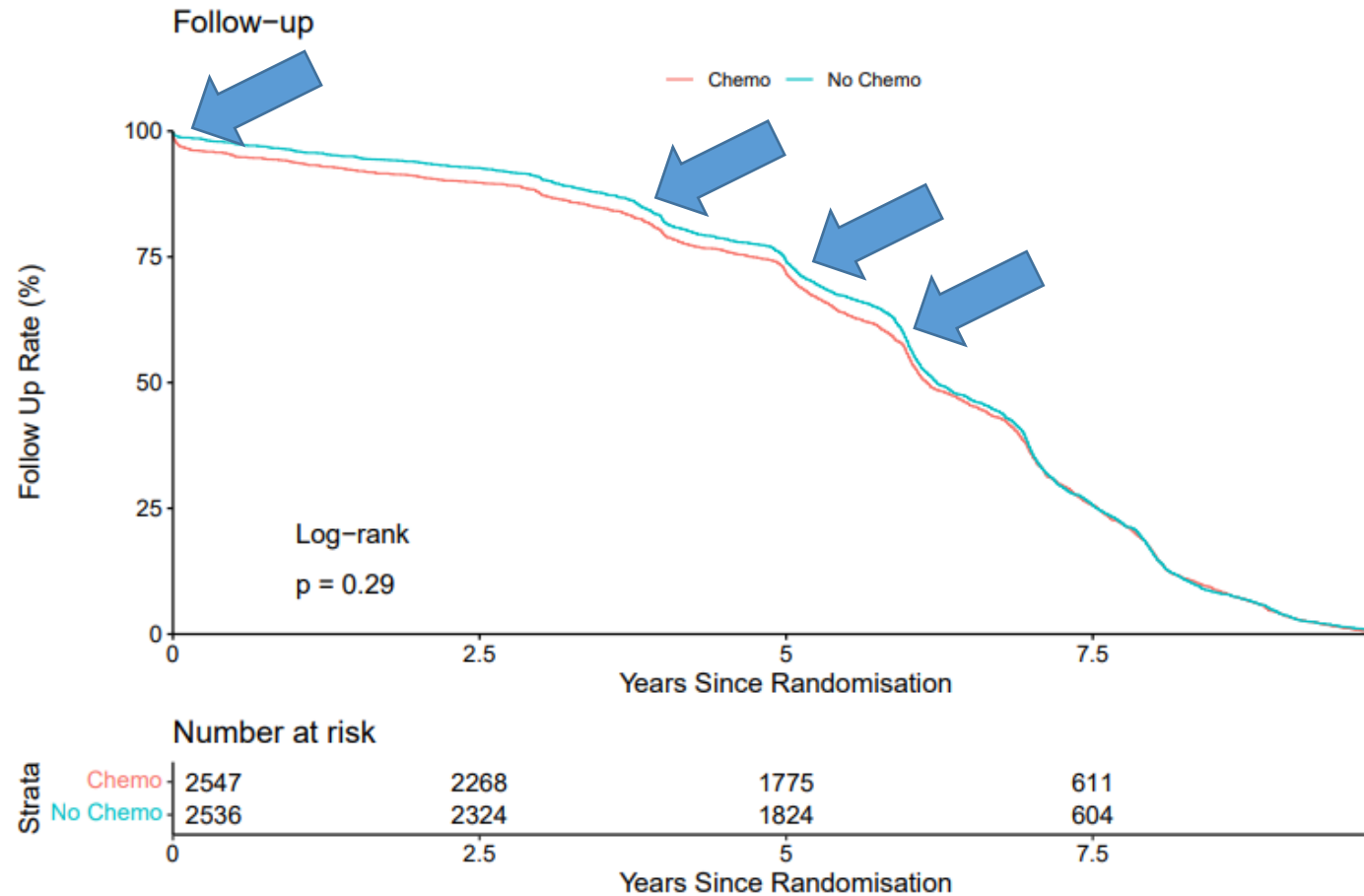


Flagged data - deaths updated but not date last seen if alive



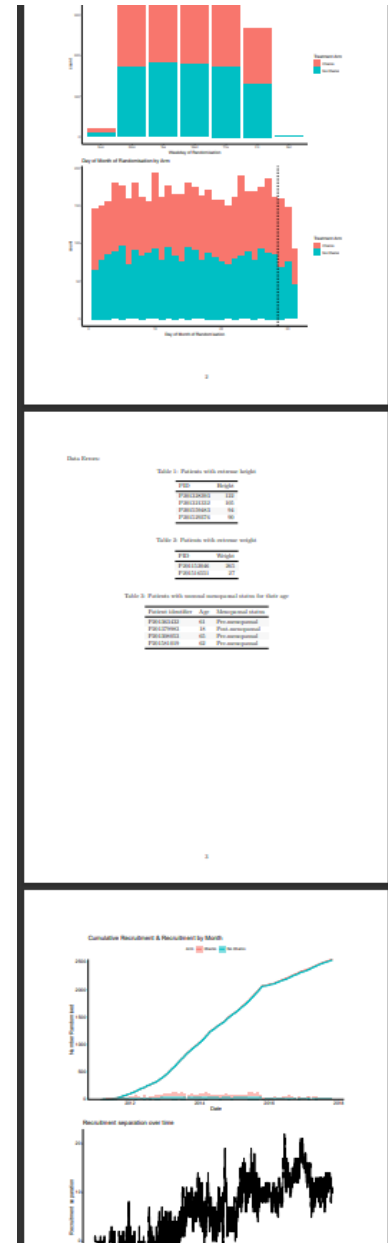
Checking follow-up

Early dropouts in one arm



Internal discussion

- Meet to discuss any difficulties which arose during processing
- Discuss completeness and whether we can resolve any ambiguity internally
- Further revision prior to sending queries
- Consider how decisions made during processing will affect analysis



Querying with trialists

- Additional data or follow-up
- Clarification
- Get to the bottom of possible mistakes or discrepancies
- Prioritising important questions