Stopping clinical trials early for futility: retrospective analysis of several randomised clinical studies

M Jitlal1, I Khan1, SM Lee2 and A Hackshaw*1

1Cancer Research UK & UCL Cancer Trials Centre, Cancer Institute, University College London, 90 Tottenham Court Road, London W1T 4TJ, UK; 2Department of Oncology, University College London Hospital, 250 Euston Road, London NW1 2PG, UK

BACKGROUND: Many clinical trials show no overall benefit. We examined futility analyses applied to trials with different effect sizes.

METHODS: Ten randomised cancer trials were retrospectively analysed; target sample size reached in all. The hazard ratio indicated no overall benefit (n = 5), or moderate (n = 4) or large (n = 1) treatment effects. Futility analyses were applied after 25, 50 and 75% of events were observed, or patients were recruited. Outcomes were conditional power (CP), and time and cost savings.

RESULTS: Futility analyses could stop some trials with no benefit, but not all. After observing 50% of the target number of events, 3 out of 5 trials with no benefit could be stopped early (low CP ≤ 15%). Trial duration for two studies could be reduced by 4–24 months, saving £44 000–231 000, but the third had already stopped recruiting, hence no savings were made. However, of concern was that 2 of the 4 trials with moderate treatment effects could be stopped early at some point, although they eventually showed worthwhile benefits.

CONCLUSIONS: Careful application of futility can lead to future patients in a trial not being given an ineffective treatment, and should therefore be used more often. A secondary consideration is that it could shorten trial duration and reduce costs. However, studies with modest treatment effects could be inappropriately stopped early. Unless there is very good evidence for futility, it is often best to continue to the planned end.

British Journal of Cancer (2012) 107, 910–917. doi:10.1038/bjc.2012.344 www.bjcancer.com
Published online 9 August 2012 © 2012 Cancer Research UK

Keywords: clinical trials; futility analysis; conditional power; early stopping

Randomised phase III trials are usually based on several hundred or thousand subjects. New interventions are often found to be ineffective, or the observed effect is lower than expected and clinically unimportant, despite preliminary evidence that it could be beneficial. If an intervention is ineffective, it is worth considering whether the trial could have been stopped earlier after examining interim data, thus avoiding the recruitment of additional subjects and giving them an ineffective therapy, particularly if there are side effects. Also, the trial treatment could be stopped among those who are already taking it. Stopping for futility has other potential advantages, including savings in staff and financial resources.

Stopping trials early for futility has been discussed as far back as the 1980s (Halperin et al, 1982; Lan et al, 1982; Lan and Wittes, 1988), and work in this area is ongoing (Whitehead and Matsushita, 2003; Pocock, 2006; Lachin, 2009). There appears to be an increasing number of trials that incorporate futility, either in the protocol or the Independent Data Monitoring Committee (IDMC) requests such analyses during the trial. The Food & Drug Administration, for example, gives some guidance on this (FDA, 2006).

Futility methods involve using earlier results from patients recruited up to a specific point, to make assumptions about future data, so there will be limitations to this. For example with time-to-event outcomes, the assumption of proportional hazards could be violated during the trial. The two main methods to assess futility are group sequential methods and conditional power (CP) (Whitehead and Matsushita, 2003; Snappin et al, 2006; Hughes et al, 2009). There are various approaches to futility analysis based on CP (Halperin et al 1982, Lan et al 1982; Lan and DeMets, 1983; Lan and Wittes, 1988). Other approaches include a Bayesian method to estimate an ‘average’ CP, called predictive power (Spiegelhalter et al, 1986; DeMets, 2006) and the use of a phase II surrogate end point in a phase III trial (Herson et al, 2011). The Cancer Research UK and UCL Cancer Trials Centre has conducted clinical trials for many years, of which several have not shown a worthwhile benefit. We retrospectively examined futility analyses in these trials.

MATERIALS AND METHODS

Ten randomised phase III trials of superiority were included, in which the target sample size was reached in all (none had stopped early); Table 1. These trials were all of those on which the authors had worked that showed either no effect (n = 5), a moderate effect (n = 4) and one additional study was chosen with a large benefit, for comparison. We aimed to see whether examining futility would stop the five ‘negative’ trials early (and if so, what the savings could be), but not any of the others. In UKHAN, no effect was shown in patients with prior surgery, whereas those without surgery did benefit, so we regarded them as two separate trials (UKHAN_1 and UKHAN_2, respectively). In ZIPP, results for two endpoints were used to show how different results could arise.
The CP is the chance of getting a statistically significant result at the end of the trial given the data so far. At each analysis, the distribution of future data is assumed to be consistent with the target hazard ratio (HR) (Snapinn et al., 2006). The CP calculation incorporates the observed and target number of events, and the observed and target HR; see Appendix 1 for the statistical methods, which are described in full elsewhere (Lan and Wittes, 1988; Proschan et al., 2006). For time-to-event outcomes, CP is usually based on the expected total number of events. However, one should be cautious in choosing the denominator of the information fraction and it is recommended not to dramatically change it (Proschan et al., 2006). Royston et al. (2003) describe a stopping rule approach based on the expected total number of events in the control group, for multiple experimental arms when each is compared with the control. Generally, the CP should get closer to 100% over time as the observed HR approaches or exceeds the planned HR (i.e., when there is a real treatment effect), and it gets closer to 0% for studies of ineffective treatments. The CP should be low to provide sufficient supporting evidence to stop early, though there is no standard threshold in practice (Snapinn et al., 2006). Here, we suggest CP ≤15%. Stata v10 (College Station, TX, USA) was used to calculate CP (Appendix 2). We also used the method in which CP is based on the observed treatment effect, rather than the target effect, at the interim analysis (Snapinn et al., 2006). Although this method can be used with the other one by the Data Monitoring Committee to examine the interim results using various assumptions, it is not often used in practice. For time-to-event outcomes, caution should be used when interpreting CP if the proportional hazards assumption is violated. This is not so much a concern for calculating the CP, but rather a limitation of the statistic to measure treatment benefit in the presence of non-proportional hazards.

Three interim analyses were specified, after 25, 50 and 75% of events had occurred, or patients had been recruited. Many researchers trigger the interim analyses on events, but using patients recruited is also used. Outcomes were: (i) CP, (ii) the number of patients left to recruit the target sample size and (iii) cost savings if a trial were stopped early.

If analyses are triggered on a specified percentage of recruited patients, we allowed some follow-up so that events could occur in the last patients accrued: 3 months for advanced disease (lung and biliary tract cancer), and 6 months for the others. Further patients would be recruited during this time, but with minimal contribution to the analysis. In addition to this, and also for interim analyses triggered on a specified percentage of events, we allowed two extra months, during which the IDMC would meet, discuss the results, and then make decisions with the trial investigators. Both allowances are expected in practice.

To provide some estimate of uncertainty when interpreting a single observed CP from a trial, we also simulated 1000 bootstrap samples for each trial when 50% patients or 50% events had occurred. Sampling was with replacement. For each trial, patients were randomly selected from the trial, such that they could contribute none or at least once to each of the 1000 bootstrap samples. The patients were sorted by the date of randomisation and the date of events where they occurred. Bootstrapping was
conducted based on the order in which patients were entered into the trial, therefore replicating the interim analyses scenario as they would have occurred prospectively. Each simulation was stratified by treatment arm so that the number of patients in each arm was the same as that observed. For each of the 1000 bootstrap samples we calculated the HR and corresponding CP, in order to assess the proportion of samples that would indicate stopping the trial early (i.e., where CP ≤ 15%). Cost savings were examined in the five trials with no overall benefit. The same unit costs were specified for all studies for comparability, without considering inflation and increased expenses over time. The costs were applied to the number of months left to complete the target recruitment at each interim analysis. Investigational drugs were always provided free of charge by the manufacturer or health service provider, as were the costs associated with extra follow-up clinic visits and assessments. Because only the direct costs of conducting the trial were considered, any estimates of savings are conservative.

RESULTS
The 10 trials are summarised in Table 1, of which 6 were relatively large (>500 patients). The observed HRs at the end of the study in the 5 trials with no overall treatment benefit were either just below or above 1.0, though one (TOPICAL) showed a clear benefit among patients who had first cycle erlotinib rash. Among the four trials with moderate effects, the HRs were no lower than 0.78. The proportional hazards assumption was met in all trials.

Interim analyses triggered after a specified percentage of events are observed
None of the five trials with no overall benefit would have been stopped early after observing 25% of events (Table 2). After 50% of events had occurred, Study 12 and UKHAN_1 could have been stopped early after observing 25% of events (Table 2). After 50% of events were observed Table 2 shows the results at each of the three specified time points. None of the five trials with no benefit would have been stopped early for futility at any point, though at 75% of events the CP (17%) was close to our specified cutoff. As expected, the trial with the large treatment effect (ABC02) would not have been stopped early for futility at any point.

Table 2: Interim analyses based on a fixed percentage of target events (assumes future data is consistent with the target HR)

| Trial          | HR (95% CI); P-value | Conditional power % | No. (%) of patients recruited | No. (%) of patients left | HR (95% CI); P-value | Conditional power % | No. (%) of patients recruited | No. (%) of patients left |
|---------------|----------------------|---------------------|-------------------------------|--------------------------|----------------------|---------------------|-------------------------------|--------------------------|
| No evidence of an overall benefit |                      |                     |                               |                          |                      |                     |                               |                          |
| Study 8       | 0.74 (0.45-1.21); 0.23 | 72                  | 125                           | 191 (60)                 | 0.88 (0.62-1.25); 0.47 | 39                  | 187                           | 129 (41)                 |
| Study 12      | 1.13 (0.82-1.56); 0.44 | 48                  | 447                           | 273 (38)                 | 1.17 (0.94-1.47); 0.17 | 2                   | 637                           | 83 (12)                  |
| Study 14      | 0.94 (0.69-1.30); 0.72 | 73                  | 511                           | 209 (29)                 | 1.05 (0.84-1.31); 0.69 | 15                  | 722                           | 0 (0)                    |
| TOPICAL       | 0.94 (0.67-1.31); 0.70 | 81                  | 266                           | 398 (60)                 | 0.87 (0.68-1.10); 0.24 | 78                  | 425                           | 239 (36)                 |
| UKHAN_1       | 1.11 (0.61-2.02); 0.73 | 19                  | 140                           | 113 (45)                 | 1.10 (0.72-1.67); 0.67 | 3                   | 217                           | 34 (16)                  |
| Moderate treatment effect | |                     |                               |                          |                      |                     |                               |                          |
| ACT I         | 0.98 (0.59-1.63); 0.93 | 49                  | 323                           | 254 (44)                 | 1.16 (0.80-1.66); 0.04 | 4                   | 526                           | 51 (9)                   |
| Over 50s      | 0.96 (0.74-1.25); 0.06 | 83                  | 2887                          | 125 (4)                  | 0.83 (0.69-1.00); 0.04 | 95                  | 3443                          | 0 (0)                    |
| UKHAN_2       | 0.95 (0.66-1.66); 0.85 | 36                  | 194                           | 205 (51)                 | 0.85 (0.61-1.18); 0.33 | 45                  | 331                           | 68 (17)                  |
| ZIPP: EFS     | 0.76 (0.57-1.01); 0.026 | 90                  | 1436                          | 291 (73)                 | 0.85 (0.70-1.04); 0.12 | 79                  | 2159                          | 541 (20)                 |
| ZIPP: OS      | 0.98 (0.74-1.31); 0.92 | 60                  | 2358                          | 342 (13)                 | 0.86 (0.70-1.05); 0.13 | 76                  | 2691                          | 9 (0.3)                  |
| Large treatment effect | |                     |                               |                          |                      |                     |                               |                          |
| ABC02         | 0.59 (0.37-0.94); 0.002 | 93                  | 109                           | 291 (73)                 | 0.68 (0.50-0.94); 0.018 | 95                  | 268                           | 132 (33)                 |

Abbreviations: CI = confidence interval; EFS = event-free survival; HR = hazard ratio; OS = overall survival. *No. of patients recruited* plus *no. of patients left* equals at least the target sample size. *Includes patients recruited while the DMEC meeting would be organised.
futility assuming future data would be consistent with the observed HR

Table A1 shows the CP when the futility analyses assume that data from future patients follow the same distribution as that observed so far (rather than the original target HR). All of the other results (observed HR, number of patients recruited, and number of patients and events left to accrue) are the same as in Tables 2 and 4.

After 50% recruitment, four trials could have been stopped early after 75% of patients had been recruited, where the CP was 0.2%, 3%, 10% and 8%, in Study 8, Study 12, TOPICAL and UKHAN_1, respectively. At this point, there remains 22, 9, 10 and 17% of patients to be recruited to complete the original target for these trials.

Among trials with a modest treatment benefit, there are two instances when recruitment could have terminated early: UKHAN_2 (CP = 11%; 50% of patients), and ACT I (CP = 7%; 75% of patients). Stopping these two trials early would be particularly concerning because the interim data for OS would not indicate any benefit at all (HRs 1.12 and 1.29 for ACT I and UKHAN_2, respectively) – very different from the final estimates (0.86 and 0.81). After 13 years followup of ACT I HR = 0.86, 95% CI 0.70–1.04 (Northover et al, 2010) and there was a clear benefit on event-free survival for UKHAN (HR = 0.72, P = 0.004; Tobias et al, 2010). For ACT I, at 50% events, about 22% of bootstrapped CPs were >15% (Table 3), which shows some uncertainty in any decision to stop early.

### DISCUSSION

To the best of our knowledge, this is the first application of futility analysis to several real phase III oncology trials. Early stopping of those with an ineffective intervention has obvious appeal – primarily not exposing further patients to it, when there is no benefit but there could be side effects. However, we show that the decision to stop recruitment early is not straightforward (unless based on safety concerns and there is clearly more harm in one group than the other). There are trials with no overall benefit that might not be stopped early, but worse still there are studies with modest effects that could. Similar conclusions have been found elsewhere (Barthel et al, 2009). Conducting clinical trials is expensive and takes several years, so a secondary consideration is the potential significant savings in accrual time and financial costs, which could be of interest to funding organisations, but should be outweighed by the ethical issues. All of these considerations should be balanced against maximising the sample size to get a more reliable estimate of the treatment effect; examination of secondary end points (DeMets, 2006) and important, pre-specified subgroup analyses; and not missing an intervention with a moderate benefit, which is still clinically worthwhile.

Occasionally, by the time there is sufficient evidence for futility, recruitment is not far from the target, so it is sometimes best to continue to the end, because the savings in time and costs are minimal (e.g., Study 12); but only if there is no unacceptable harm to patients. A further consideration is whether patients are still on treatment. A trial in which all have finished the trial treatments, but subjects are in follow-up, could still continue if there are no concerns over the schedule of clinic assessments. Continuing follow-up in a trial that has been stopped early has the advantages of minimising bias and obtaining more data on adverse events.

The worse situation is for trials where there appears to be no benefit at an interim analysis, but they do in fact have a moderate effect. It would be unsatisfactory to stop such trials early because

### Time and cost savings for the trials that showed no evidence of an overall benefit

When interim analyses are based on percentage of events, the number of months left is, as expected, lower than when based on percentages of patients recruited, but there could still be cost savings (Table 5). For example, Study 12 could be terminated early after observing 50% of events (CP ≥ 2%), but there are only 4 months to complete recruitment and the savings associated with early stopping is £44 000. Overall, after seeing 50% of events, three trials could be stopped early, avoiding 4–24 more months of accrual and saving £44 000–231 000 in two of these (Study 12 and UKHAN_1); in Study 14 no savings are made because recruitment had already finished.

With the futility analysis at 75% of events, only one trial with low CP is still recruiting (Study 8), but the savings would be 15 fewer months of recruitment and £144 000 lower costs.

Table 5 also shows the estimated time and cost savings when the analyses are based on recruited patients. The trials could only be stopped after 75% of patients had been recruited, with 4–28 fewer accrual months and £44 000–270 000 lower costs. For example, in Study 12 only 66 more patients are needed to reach the target sample size, which actually took only 4 months. Had this study been stopped early, the savings would be £44 000. However, the number of months left to complete accrual was 19 for Study 8, 6 for TOPICAL and 28 for UKHAN_1. Even after recruiting 75% of patients there could be significant cost savings by stopping early: £183 000, £58 000 and £270 000, respectively. The observed monthly accrual rates are an important factor when considering whether to stop early or not, which was high in Study 12.
Clinical Studies

Events observed

| Trial          | No evidence of an overall benefit | Study 8 | 0.61 (0.33–1.15); P = 0.13 | 80 (41) | 90 | 226 (72); P = 0.06 (0.66–1.39); 0.81 | 29 (111) | 166 | 150 (47); 0.05 (0.78–1.91); 0.01 | 21 (179) | 246 | 70 (22) |
|               | Study 12 | 0.66–1.97; P = 0.04 | 77 (53) | 330 | 390 (54); P = 0.081 | 55 (156) | 453 | 267 (37); 0.09 (0.55–1.97); 0.026 | 3 (315) | 654 | 66 (9) |
|               | Study 14 | 0.60–1.44; P = 0.03 | 81 (81) | 354 | 366 (51); P = 0.071 (0.71–1.32); 0.09 | 68 (161) | 543 | 177 (25); 0.05 (0.83–1.33); 0.026 | 21 (280) | 714 | 6 (1) |
|               | TOPICAL | 0.67–1.35; P = 0.73 | 81 (129) | 249 | 415 (63); P = 0.088 (0.69–1.12); 0.09 | 75 (277) | 423 | 241 (36); 0.09 (0.77–1.12); 0.043 | 10 (434) | 600 | 64 (10) |
|               | UKHAN_1 | 0.34–1.87; P = 0.78 | 42 (22) | 93 | 160 (63); P = 0.106 (0.61–1.85); 0.83 | 18 (50) | 157 | 96 (38); 0.101 (0.65–1.55); 0.098 | 8 (82) | 210 | 43 (17) |
|               | UKHAN_2 | 0.34–1.87; P = 0.60 | 42 (22) | 93 | 160 (63); P = 0.106 (0.61–1.85); 0.83 | 18 (50) | 157 | 96 (38); 0.101 (0.65–1.55); 0.098 | 8 (82) | 210 | 43 (17) |
|               | Moderate treatment effect | ACT 1 | 1.15 (0.51–2.61); P = 0.074 | 59 (23) | 192 | 385 (67); P = 0.093 (0.58–1.50); 0.78 | 50 (68) | 371 | 206 (36); 1.12 (0.78–1.62); 0.54 | 7 (114) | 510 | 67 (12) |
|               |          | Over 50s | 0.96 (0.46–2.02); P = 0.92 | 93 (28) | 995 | 2017 (67); P = 0.092 (0.63–1.35); 0.68 | 92 (107) | 1970 | 1042 (35); 1.03 (0.77–1.39); 0.84 | 81 (173) | 2600 | 412 (14) |
|               |          | UKHAN_1 | 1.05 (0.55–2.02); P = 0.87 | 52 (37) | 123 | 276 (69); P = 1.29 (0.85–1.94); 0.23 | 11 (91) | 241 | 158 (40); 0.86 (0.62–1.20); 0.37 | 42 (148) | 333 | 66 (17) |
|               |          | ZIPP: EFS | 0.72 (0.46–1.13); P = 0.15 | 87 (79) | 926 | 1774 (66); P = 0.80 (0.62–1.05); 0.11 | 87 (223) | 1582 | 1118 (41); 0.81 (0.66–0.98); 0.03 | 91 (405) | 2264 | 436 (16) |
|               |          | ZIPP: OS | 1.14 (0.50–2.59); P = 0.75 | 78 (23) | 926 | 1774 (66); P = 1.16 (0.74–1.81); 0.51 | 66 (78) | 1582 | 1118 (41); 1.02 (0.75–1.38); 0.91 | 59 (168) | 2264 | 436 (16) |
|               | Large treatment effect | ABC02 | 0.62 (0.40–0.97); P = 0.036 | 92 (81) | 124 | 276 (69); P = 0.65 (0.46–0.90); 0.011 | 96 (141) | 247 | 153 (38); 0.69 (0.53–0.90); 0.006 | 99 (222) | 356 | 44 (11) |

Abbreviations: CI = confidence interval; EFS = event-free survival; HR = hazard ratio; OS = overall survival. *No. of patients recruited* plus *no. of patients left* equals the target sample size.*Includes patients recruited while the first 25, 50 or 75% are being followed up, and also the time for the DMEC to meet.

Table 5 Potential savings (time and costs) associated with the interim analyses shown in Tables 2 and 4 for the five trials in which there was no overall treatment effect

| Trial          | 25% Target | 50% Target | 75% Target |
|               | CP (%) | Information fraction | No. of months left to complete recruitment | Costs saved (£’000) | CP (%) | Information fraction | No. of months left to complete recruitment | Costs saved (£’000) | CP (%) | Information fraction | No. of months left to complete recruitment | Costs saved (£’000) |
| Study 8        | 48         | 0.25         | 12       | 133      | 2           | 0.50      | 4          | 44       | 0.02               | 0.75           | 15       | 144 |
| Study 12       | 77         | 0.09         | 18       | 200      | 5.5         | 0.26      | 12        | 133      | 0.001              | 0.75           | 0       | 0 |
| Study 14       | 81         | 0.13         | 11       | 122      | 6.8         | 0.26      | 6         | 67       | 0.001              | 0.75           | 0       | 0 |
| TOPICAL        | 81         | 0.23         | 24       | 231      | 7.5         | 0.49      | 15        | 144      | 0.10               | 0.79           | 6       | 58 |
| UKHAN_1        | 42         | 0.13         | 65       | 626      | 18          | 0.29      | 47        | 452      | 0.8                | 0.47           | 28      | 270 |

Abbreviations: CP = conditional power. Calendar years of recruitment were: Study 8 (Dec 1992–Oct 2001), Study 12 (May 2003–Feb 2006), Study 14 (June 2003–Sep 2005), TOPICAL (April 2005–April 2009), and UKHAN (Dec 1992–Oct 2001). Annual costs used here were: full-time co-ordinator (£45,000), full-time data manager (£35,000), half-time administrator (£25,000), regulatory support (£10,000), IT support (£5,000), and running expenses of £7,000. TOPICAL, Study 12 and Study 14 were large, so we allowed for 1.5 data managers. Costs saved, if the trial is stopped early, are rounded to the nearest £1,000.

of insufficient patients or events. We give examples (ACT I and UKHAN_2) where interim HRs are close to or exceed 1.0, with low CP, but the final HR indicated a clinically important effect.

The results and conclusions of three of the trials with no overall effect provided useful information after reaching the target sample size, especially when examining important subgroup analyses. Study 8, whose results were unexpectedly inconsistent with a preceding Canadian trial (despite having the same protocol), led to a systematic review showing that early radiotherapy only improved survival if patients completed chemotherpay (Spiro et al, 2006). A post-hoc subgroup analysis in Study 14 (Lee et al, 2006) indicated that patients with squamous histology who had at least stable disease by chemotherapy cycle 3 had an OS HR of 0.71, and this has led to a randomised phase II trial using another
Stopping clinical trials early for futility

**Box I** Considerations for stopping a clinical trial early for futility

- Futility might not be useful for early-stage cancers that have a good prognosis, where events (e.g., recurrences or deaths) take several years to be seen. By the time lack of benefit is determined to be reliable, recruitment (and probably treatment) is probably close to finishing.
- There should be a low conditional power (e.g., <15%), based on the target effect size. The research team and IDMC should agree what they consider to be low. It is also worth considering estimates of uncertainty (e.g., bootstrapping CIs).
- Effect size should be very close to or above the no effect value (e.g., HR > 1, possibly with lower 95% CI limit ≥0.90 or 0.95).
- The IDMC and trial team should agree that enough patients and, importantly, events have been observed so far to produce a reliable effect (remembering that the trial investigators are likely to want to continue); interim data will be influenced by chance, especially in early analyses.
- There are many more patients left to recruit, or to finish accrual is likely to take many more months (with financial cost considerations).
- Other clinically important end points do not show evidence of a benefit.
- There is no evidence of an effect in important pre-specified subgroups. However, if there is evidence within a subgroup (but not overall), an early effect could be spurious, especially if not based on many events or patients. Continuing to the end should confirm whether there really is an effect in the subgroup, and if there is, obtain a better estimate of it.
- The adverse events profile is acceptable (if there are no safety concerns, one might wish to continue to ensure that a modest effect is not missed).

The problem with this is that the observed HR is likely to be unreliable early on in the trial. However, CP based on the target effect size is relatively insensitive to the early results of a trial. Deciding whether to trigger the interim analysis on proportion of patients recruited or events observed is also important. The observed effect size early on in a trial may fluctuate too much and so be unreliable, especially if there is treatment imbalance (Herson et al., 2011), and regardless of the method or assumptions used. Many researchers use percentage of events to trigger the interim analysis, a reasonable approach given that the statistical analyses are often influenced most by the number of events, and hence might be more reliable than percentage of patients. In the set of trials we examined, futility analyses triggered on events (after 50 or 75%) could stop four out of the five trials with no overall benefit, and only one trial with a moderate effect. Whereas analyses triggered on patients could also stop four out of five studies with no benefit, but 2 trials with a moderate effect. An important consideration is that analyses triggered on events are more likely to be based on longer follow-up, so the potential savings are generally less than analyses triggered on number of patients (Table 5).

Further research using modelling and simulations could examine an appropriate frequency of interim analyses, specifying situations when futility may or may not be appropriate, and which method(s) are appropriate, including whether to trigger the early looks on percentage of events or patients observed. Terminology from medical screening could be useful: detection rate (DR – the proportion of truly negative trials that are stopped early) and false-positive rate (FPR – the proportion of trials with modest treatment effects that are stopped early). A good method will have high DR and low FPR, and these parameters could be examined in relation to trial size, the timing of interim analyses, and different statistical methods. Other authors have discussed futility in relation to falsely stopping studies (Hughes et al., 2009). Methods examining two or more end points could also be developed.

In summary, careful application of futility methods can lead to ineffective treatments not being given to future trial patients, and this could also lead to shorter trial duration and reduced financial costs. However, there are situations when the end of the trial is not far off, so the research team may as well complete it. A major concern is that there are studies with modest treatment effects that could be inappropriately stopped early, and a clinically important effect missed. Therefore, unless there is very clear and sufficient evidence for futility, it is often best to continue to the planned end.

**ACKNOWLEDGEMENTS**

We thank Juan Valle (ABC02), John Northover (ACT I), Michael Baum (Over 50s, ZIPPP), Jeff Tobias (UKHAN) and Stephen Spiro (Study 8) on behalf of their respective trial study collaborators, for use of trial data for which they were principal investigators.
REFERENCES

Barthel FM, Parmar MK, Royston P (2009) How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design—a reanalysis of 4 trials. Trials 10: 1–10
Baum M, Hackshaw A, Houghton J, Rutqvist, Fornander T, Nordenskjold B, Nicolucci A, Sainsbury R, ZIPP International Collaborators Group (2006) Adjuvant goserelin in pre-menopausal patients with early breast cancer: Results from the ZIPP study. Eur J Cancer 42(7): 895–904
DeMets DL (2006) Futility approaches to interim monitoring by data monitoring committees. Clin Trials 3: 522–529
Food and Drug Administration. Guidance for clinical trial sponsors. Establishment and operation of clinical trial data monitoring committees. FDA (2006) http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm127073.pdf
Freidlin B, Korn EL, Gray R (2010) A general inefficacy interim monitoring rule for randomized clinical trials. Clin Trials 7: 197–208
Hackshaw A, Roughton M, Forsyth S, Monson K, Reczko K, Sainsbury R, Baum M (2011) Long-term benefits of 5 years of tamoxifen: 10 year follow up of a large randomised trial in women aged at least 50 years with early breast cancer. J Clin Oncol 29(13): 1657–1663
Halperin M, Lane KKG, Ware JH, Johnson NJ, DeMets DL (1982) An aid to data monitoring in long-term clinical trials. Control Clin Trials 3: 311–323
Herson J, Buyse M, Wittes JT (2011) On Stopping a Randomized Clinical Trial for Futility. In Designs for Clinical Trials: Perspectives on Current Issues, Harrington D (ed) (Applied Bioinformatics and Biostatistics in Cancer Research) Springer: USA
Hughes S, Cuffe RL, Lieftucht A, Garrett Nichols W (2009) Informing the selection of futility stopping thresholds: case study from a late-phase clinical trial. Pharm Stat 8(1): 23–37
Lachin JM (2005) Futility interim monitoring with control of type I and II error probabilities using the interim Z-value or confidence limit. Clin Trials 6: 565–573
Lane KKG, DeMets DL (1983) Discrete sequential boundaries for clinical trials. Biometrika 70: 659–663
Lane KKG, Simon R, Halperin M (1982) Stochastically curtailed tests in long-term clinical trials. Commun Stat C1: 207–219
Lane KKG, Wittes J (1988) The B-value: a tool for monitoring data. Biometrics 44: 579–585
Lee SM, Wall PJ, Rudd R, Ferry D, O’Brien M, Middleton G, Spiro S, James LE, Rudd RM, Khan I, Upadhyay S, Lewanski CR, Falk S, Skailis G, Partridge R, Ngyi Y, Boshoff C (2010) TOPICAL: Randomized phase III trial of erlotinib compared with placebo in chemotherapy-naïve patients with advanced non-small cell lung cancer (NSCLC) and unsuitable for first-line chemotherapy. J Clin Oncol 28(15 suppl): abstr 7504
Northover J, Glynne-Jones R, Sebag-Montefiore D, James R, Meadows H, Wan S, Jitlal M, Ledermann J (2010) Chemoradiation for the treatment of epidermoid anal cancer: 13-year follow-up of the first randomised UKCCCR Anal Cancer Trial (ACT I). Br J Cancer 102(7): 1123–1128
Pocock SJ (2006) Current controversies in data monitoring for clinical trials. Clin Trials 3: 513–521
Proshan MA, Lane KKG, Wittes JT (2006) Statistical Monitoring of Clinical Trials: A Unified Approach. 1st edn. Springer: USA
Royston P, Mahesh KB, Qian W (2003) Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. Stat Med 22: 2239–2256
Snapinn S, Chen MG, Jiang Q, Koutsoukos T (2006) Assessment of futility in clinical trials. Pharm Stat 5(4): 273–281
Spiegelhalter DJ, Freedman LS, Blackburn PR (1986) Monitoring clinical trials: Conditional or predictive power? Control Clin Trials 7(1): 8–17
Sprio SG, James LE, Rudd RM, Trask CW, Tobias JS, Snee M, Gilligan D, Murray PA, Ruiz de Elvira MC, O'Donnell KM, Gower NH, Harper PG, Hackshaw AK, London Lung Cancer Group (2006) Early compared with late radiotherapy in combined modality treatment for limited disease small-cell lung cancer: a London Lung Cancer Group multicenter randomized clinical trial and meta-analysis. J Clin Oncol 24(24): 3823–3830
Tobias JS, Monson K, Gupta N, Macdougall H, Glahom J, Hutchison I, Kadalayil L, Hackshaw A, UK Head and Neck Cancer Trials’ Group (2010) Chemoradiotherapy for locally advanced head and neck cancer: 10-year follow-up of the UK Head and Neck (UKHAN1) trial. Lancet Oncol 11(1): 66–74
UKCCCR Anal Cancer Trial Working Party (1996) Epidermid anal cancer: Results from the UKCCCR randomized trial of radiotherapy alone versus radiotherapy, 5-fluorouracil, and mitomycin. Lancet 348: 1049–1054
Valle J, Wasan H, Palmer DH, Cunningham D, Anthony E, Maraveyas A, Madhusudan S, Iveson T, Hughes S, Pereira SP, Roughton M, Bridgewater J, ABC-02 Trial Investigators (2010) Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. N Engl J Med 362(14): 1273–1281
Whitehead J, Matsushita T (2003) Stopping clinical trials because of treatment ineffectiveness: a comparison of a futility design with a method of stochastic curtailment. Statistics in Medicine 22(5): 677–687

APPENDIX 1

Equation for calculating conditional power (CP) (Proshan et al, 2006) used in Table (2).

The CP at a specific time \( t \) is

\[
CP = 1 - \phi \left[ \frac{Z_{T/2}}{\sqrt{1 - \phi}} \right] / \sqrt{1 - \phi}.
\]

where \( \phi \) is the area under the standard normal distribution associated with what is in the brackets.

- \( Z_{T/2} \) is the Z-value cutoff associated with the target level of statistical significance (we use a \( P \)-value of 0.05, so \( Z_{T/2} \) is 1.96).
- \( B_{0} \) is the transformed \( Z \)-statistic (based on a Brownian motion applied to sequential analyses), i.e., \( B_{0} = Z(t) \times \sqrt{t} \).
- \( E[B_{0}|B_{0}] \) is the expected value of \( B_{0} \) at the end of the trial (when \( t = 1 \)), given the data observed until point \( t \).
- The information fraction, \( t \), is the number of observed events so far, expressed as a proportion of the planned number of events.

\[
B_{0} = Z(t) \times \sqrt{t}
\]

\[
E[B_{0}|B_{0}] = \frac{1}{\sqrt{1 - \phi}} \left[ Z(t) \times \sqrt{t} \right] = \frac{1}{\sqrt{1 - \phi}} \sqrt{\frac{1}{\sqrt{1 - \phi}} \times \left( Z(t) \times \sqrt{t} \right)}
\]

\[
E[B_{0}|B_{0}] = 0.141 + 2.582 = 2.723
\]

\[
CP = 1 - \phi \left[ \frac{Z_{T/2}}{E[B_{0}] \sqrt{1 - \phi}} \right] / \sqrt{1 - \phi}
\]

\[
CP = 1 - \phi \left[ \frac{1.96 + 2.723 \sqrt{0.766}}{\sqrt{0.766}} \right] = 0.81 \text{ (i.e., } 81\% \text{)}
\]
APPENDIX 2

Stata code for calculating conditional power and also for generating 1000 bootstrap samples for a trial

Conditional power

The following variables need to be present in the data set for each interim analysis:

\[ n \]: the number of events observed up until the interim analysis.
\[ N \]: the number of planned events at the end of the trial.
\[ t \]: the information fraction = \[ n/N \].
\[ HR_{O} \]: the observed hazard ratio at the interim analysis.
\[ HR_{E} \]: the planned hazard ratio.

*Conditional Power, based on planned data (the following represents two lines of code):

\[
generate \ con\_power\_plan = \left( 1 - \frac{\Phi\left( \frac{1.96 + \left( \sqrt{\frac{n}{4}} \ln\left( \frac{1}{HR_{O}} \right) \times (1-t) \right)}{\sqrt{t}} \right)}{\frac{\sqrt{t}}{2}} \right) \times 100
\]

label variable con_power_plan Conditional power (%) – planned.

Bootstrap sampling

The bootstrap sampling is based upon data at a particular time point. In our analysis this relates to 25, 50 or 75% events or patients. That is, this restricts the data set to include only those patients who have been entered into the study by the specified time point (see Materials and Methods for further details).

*Bootstrap samples: 1000 replicates, based on generic data (one line of code):

\[
\text{bootstrap } \_b \ N\_\text{fail} = e(N\_\text{fail}), \text{rep(1000) strata(treat)} \text{ saving(trial\_bootstrap, replace): stcox treat}
\]

Table A1 Conditional power based on a fixed percentage of recruited patients or events (assumes future data is consistent with the observed hazard ratio so far

| Trial       | 25%  | 50%  | 75%  | 25%  | 50%  | 75%  |
|-------------|------|------|------|------|------|------|
| No evidence of a benefit |
| Study 8     | 98   | 1    | <0.01| 69   | 9    | <0.01|
| Study 12    | <0.01| 0.1  | <0.01| <0.01| <0.01| <0.01|
| Study 14    | 13   | 3    | 0.02 | 8    | 0.01 | <0.01|
| TOPICAL     | 6    | 26   | 0.8  | 8    | 32   | 1    |
| UKHAN_1     | 30   | 0.3  | 0.3  | 0.1  | 0.01 | 0.01 |
| Moderate treatment effect |
| ACT 1       | 0.07 | 5    | <0.01| 2    | <0.01| 0.09 |
| Over 50s    | 9    | 22   | 0.4  | 6    | 88   | 72   |
| UKHAN_2     | 0.5  | <0.01| 17   | 0.3  | 21   | 38   |
| ZIPPP: EFS  | 99.6 | 90   | 91   | 98   | 64   | 99.8 |
| ZIPPP: OS   | <0.01| <0.01| 0.6  | 3    | 55   | 99.8 |
| Large treatment effect |
| ABC02       | 99.6 | 99.4 | 99.3 | 99.9 | 98   | 100  |

Stopping clinical trials early for futility
M Jitlal et al

Table A1 Conditional power based on a fixed percentage of recruited patients or events (assumes future data is consistent with the observed hazard ratio so far

| Trial       | 25%  | 50%  | 75%  | 25%  | 50%  | 75%  |
|-------------|------|------|------|------|------|------|
| No evidence of a benefit |
| Study 8     | 98   | 1    | <0.01| 69   | 9    | <0.01|
| Study 12    | <0.01| 0.1  | <0.01| <0.01| <0.01| <0.01|
| Study 14    | 13   | 3    | 0.02 | 8    | 0.01 | <0.01|
| TOPICAL     | 6    | 26   | 0.8  | 8    | 32   | 1    |
| UKHAN_1     | 30   | 0.3  | 0.3  | 0.1  | 0.01 | 0.01 |
| Moderate treatment effect |
| ACT 1       | 0.07 | 5    | <0.01| 2    | <0.01| 0.09 |
| Over 50s    | 9    | 22   | 0.4  | 6    | 88   | 72   |
| UKHAN_2     | 0.5  | <0.01| 17   | 0.3  | 21   | 38   |
| ZIPPP: EFS  | 99.6 | 90   | 91   | 98   | 64   | 99.8 |
| ZIPPP: OS   | <0.01| <0.01| 0.6  | 3    | 55   | 99.8 |
| Large treatment effect |
| ABC02       | 99.6 | 99.4 | 99.3 | 99.9 | 98   | 100  |

General form

Example (TOPICAL trial, after 25% of patients have been recruited)

\[ n = \text{Number of events observed so far} \]
\[ N = \text{Target number of events} \]
\[ t = n/N \text{ (information fraction)} \]
\[ HR_{O} = \text{observed hazard ratio} \]
\[ HR_{E} = \text{planned hazard ratio} \]

\[ \xi = \ln(\frac{1}{HR_{O}}) \times \sqrt{t} \]

\[ \phi(\xi) \]

\[ CP = 1 - \phi\left( \frac{1.96 + \left( \sqrt{\frac{n}{4}} \xi \right)}{\sqrt{t}} \right) \times 100 \]