“Thursday’s child has far to go” -- Interpreting subgroups and the STAMPEDE trial

Melissa R Spears MRC Clinical Trials Unit at UCL, London

Nicholas D James Institute of Cancer and Genomic Sciences, University of Birmingham

Matthew R Sydes MRC Clinical Trials Unit at UCL

STAMPEDE is a multi-arm multi-stage randomised controlled trial protocol, recruiting men with locally advanced or metastatic prostate cancer who are commencing long-term androgen deprivation therapy. Opening to recruitment with five research questions in 2005 and adding in a further five questions over the past six years, it has reported survival data on six of these ten RCT questions over the past two years. [1-3] Some of these results have been of practice-changing magnitude, [4, 5] but, in conversation, we have noticed some misinterpretation, both over-interpretation and under-interpretation, of subgroup analyses by the wider clinical community which could impact negatively on practice. We suspect, therefore, that such problems in interpretation may be common. Our intention here is to provide comment on interpretation of subgroup analysis in general using examples from STAMPEDE. Specifically we would like to highlight some possible implications from the misinterpretation of subgroups and how these might be avoided, particularly where these contravene the very design of the research question. In this, we would hope to contribute to the conversation on subgroup analyses. [6-11]

For each comparison in STAMPEDE, or indeed virtually any trial, the interest is the effect of the research treatment under investigation on the primary outcome measure across the whole population. Upon reporting the primary outcome measure, the consistency of effect across pre-specified subgroups, including stratification factors at randomisation, is presented; these are planned analyses.

The forest plot is a valuable tool in displaying and assessing treatment effects in subgroups. Forest plots, first used in 1978, were initially used for illustration (in meta-analyses) of the treatment effect *within* studies, offering a sense of consistency of effect *across* studies. [12, 13] The lack of consistency, or heterogeneity, is formally calculated and taken into consideration when interpreting the pooled estimate. Forest plots traditionally offer two distinct vertical lines for reference: no effect (e.g. hazard ratio (HR)=1.00) and the estimated, overall treatment effect. The line of no effect is an important consideration in assessing the overall treatment effect: the confidence interval for any one of the trials may cross this line if it is underpowered or it observed no effect on the outcome of interest, and, for each well-powered study on the forest plot, the confidence intervals should be fairly narrow. This line of no effect also helps in the interpretation of the overall, pooled effect. The second line presents the pooled effect and helps readers to consider how each individual study looks compared to this overall effect.

Illustration of treatment effect across subgroups within one trial is a more recent use of forest plots, and in this scenario the emphasis must undoubtedly be placed on the **consistency** of treatment effect, not the **individual** effect within each subgroup. Unlike the individual trials in a meta-analysis of trials, each subgroup within one trial, usually, **will not** be well-powered at the time of reporting the overall effect and the confidence intervals will, typically, be wide as a result. Therefore, any assessment based solely on whether confidence intervals for treatment effect within a subgroup cross the line of no effect, are unhelpful and potentially harmful. This is particularly the case where an interaction has been tested for and there is no evidence to suggest there is indeed any difference in the overall treatment effect.

From the STAMPEDE perspective, we consider this to be an important distinction, specifically in relation to the translation in to practice of trial results found to show overwhelming benefit in the entire, eligible trial population. Most notably this has been most apparent in relation to the subgroup of metastatic status at randomisation. Patients with metastatic disease (M1) at randomisation have substantially poorer prognosis and therefore, sadly, contribute the higher proportion of events sooner; conversely those patients with non-metastatic disease (M0) at randomisation live longer and as such, subgroup analyses on survival at the time of first reporting can always be expected to appear relatively immature. For reported survival results relating to both the "docetaxel comparison" and the "abiraterone comparison", we observed no evidence of inconsistency in the treatment effect across both subgroups of metastatic status at randomisation, with a compelling overall effect and the hazard ratio (HR) in each group pulling in the direction favouring the research treatment over the standard-of-care; there was no evidence of a lack-of-consistency by metastatic status for either additional treatment in survival nor failure-free survival. [1, 3]

The focus of commenters on both the "docetaxel comparison" and the "abiraterone comparison" from STAMPEDE has most often been on the beneficial effect in the M1 subgroup, despite the underlying design and results being positive for the broader population.

Figure 1 shows three example subgroup analyses from the "abiraterone comparison" in STAMPEDE on overall survival, one already published [3] and two deliberately trawled for. The first section of the forest plot shows the subgroup analysis by metastatic status at randomisation. The interaction p-value of 0.37 shows no good evidence of heterogeneity of treatment effect across these subgroups. There is more evidence at this time in the M1 setting so the confidence intervals are narrower than for M0 but one should take the overall effect from the trial; the point estimate in the M0 setting (HR=0.75) is exactly that targeted by the protocol. However, many in the clinical community, and some commissioners in the case of docetaxel, have only focused on the data from the M1 subgroup.

The second part of **Figure 1** shows the impact on survival of abiraterone based on day of birth. The confidence intervals are broad because the patients are split, fairly evenly, into 7 groups. Like with metastatic status, there is no good evidence of heterogeneity (p=0.33). There is no reasonable clinical hypothesis to underpin a different outcome by day of birth. Therefore, the fact that the point estimates vary by weekday of birth must be by chance. Yet, some of the confidence intervals include the null line and some do not. Under reducto ad absurdum, people uncertain about the impact of abiraterone in M0 patients based on the first part of the graph should also be uncertain about the addition of abiraterone in patients who were born during the latter part of the week (Thursday to Sunday); after all, their confidence intervals also all include the null line e.g. Thursday's HR=0.69 with 95%CI 0.42 to 1.15.

However, we should also be cautious not to over-interpret subgroups. The final part of **Figure 1** is based on weekday of diagnosis, where the point estimate of abiraterone is less favourable for people diagnosed on a Monday than those on other days with striking evidence of heterogeneity (p=0.021). There is no reasonable clinical rationale in a multi-centre multinational clinical trial for this, and even with apparent statistical evidence, this must be a chance finding (note: it *is* a chance finding – we trawled through implausible subgroups before deliberately selecting this one).

Estimation of treatment effect within specific subgroups can of course be desirable, and stratification is the crucial step in moving forward towards "personalised" medicine.

In summary, it is important to consider the reasons behind low event rates and whether these are independent of a treatment effect. For this example, the M0 subgroup is prognostically favourable and does better regardless of adding in the research treatment. However, this does not equate to there being no treatment effect; the evidence here is consistent across both populations. There are clearly circumstances in which over-interpretation of subgroups can be detrimental and access to treatments is arguably one of these. Moving forward we would provide the following recommendations when interpreting trial results:

- Focus should be placed on the design of the trial; what is the primary hypothesis being tested?

- When considering any difference in treatment effect across subgroups the primary assessment should be relative to the direction of overall effect and whether the subgroup effect contests this i.e. is there consistency across subgroups?
- Context can help: consider whether the treatment effect is consistent across other trial endpoints.
- Where there is inconsistency the clinical plausibility of this should be clearly considered.

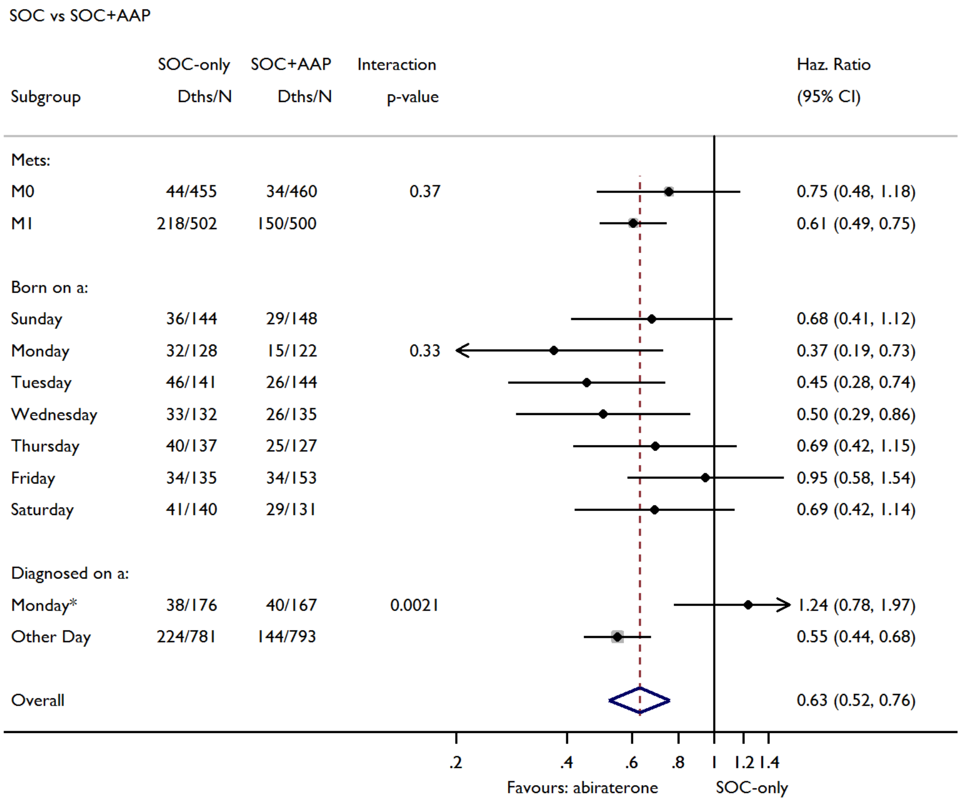
Context is particularly helpful for the “abiraterone comparison” in STAMPEDE. The impact on failure-free survival, the intermediate primary outcome measure for the trial, is very positive for adding abiraterone across the board and positive in each of the subgroups by baseline metastatic status (**Figure 2**). There is some evidence of heterogeneity of the treatment effect in these groups (p=0.085) which is more favourable in the M0 setting than in the M1 setting. Commentators who take that there is an impact on survival in M1 disease but not M0 must also conclude that there is a larger impact on FFS in the M0 setting.

In conclusion, readers must protect themselves equally from over-interpretation and under-interpretation of subgroup effects. The onus on interpretation of results subgroup analyses lies equally with the journal and reviewers who see such details prior to publication and have a role to play in shaping the message of the publication.

References:

1. James ND, Sydes MR, Clarke NW et al. Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. The Lancet 387: 1163-1177.
2. Mason MD, Clarke NW, James ND et al. Adding Celecoxib With or Without Zoledronic Acid for Hormone-Naïve Prostate Cancer: Long-Term Survival Results From an Adaptive, Multiarm, Multistage, Platform, Randomized Controlled Trial. Journal of Clinical Oncology 2017; 35: 1530-1541.
3. James ND, de Bono JS, Spears MR et al. Abiraterone for Prostate Cancer Not Previously Treated with Hormone Therapy. New England Journal of Medicine 2017.
4. England N. Clinical Commissioning Policy Statement: Docetaxel in combination with androgen deprivation therapy for the treatment of hormone naïve metastatic prostate cancer. 2016.
5. NICE. Hormone-sensitive metastatic prostate cancer: docetaxel. 2016.
6. Fletcher J. Subgroup analyses: how to avoid being misled. British Medical Journal 2007; 335: 96
7. Naggara O, Raymond J, Guilbert F, Altman DG. The Problem of Subgroup Analyses: An Example from a Trial on Ruptured Intracranial Aneurysms. American Journal of Neuroradiololgy 2011; 32: 633-36
8. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomised clinical trials. JAMA 1991;266:93-8.
9. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials: a survey of three medical journals. N Engl J Med 1987;317:426–32
10. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992;116:78–84
11. Schulz KF, Grimes DA. Multiplicity in randomised trials. II. Subgroup and interim analyses. Lancet 2005;365:1657–61
12. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial. New England Journal of Medicine 1978; 299: 690-694.
13. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. British Medical Journal 2001; 322: 1479-1480.

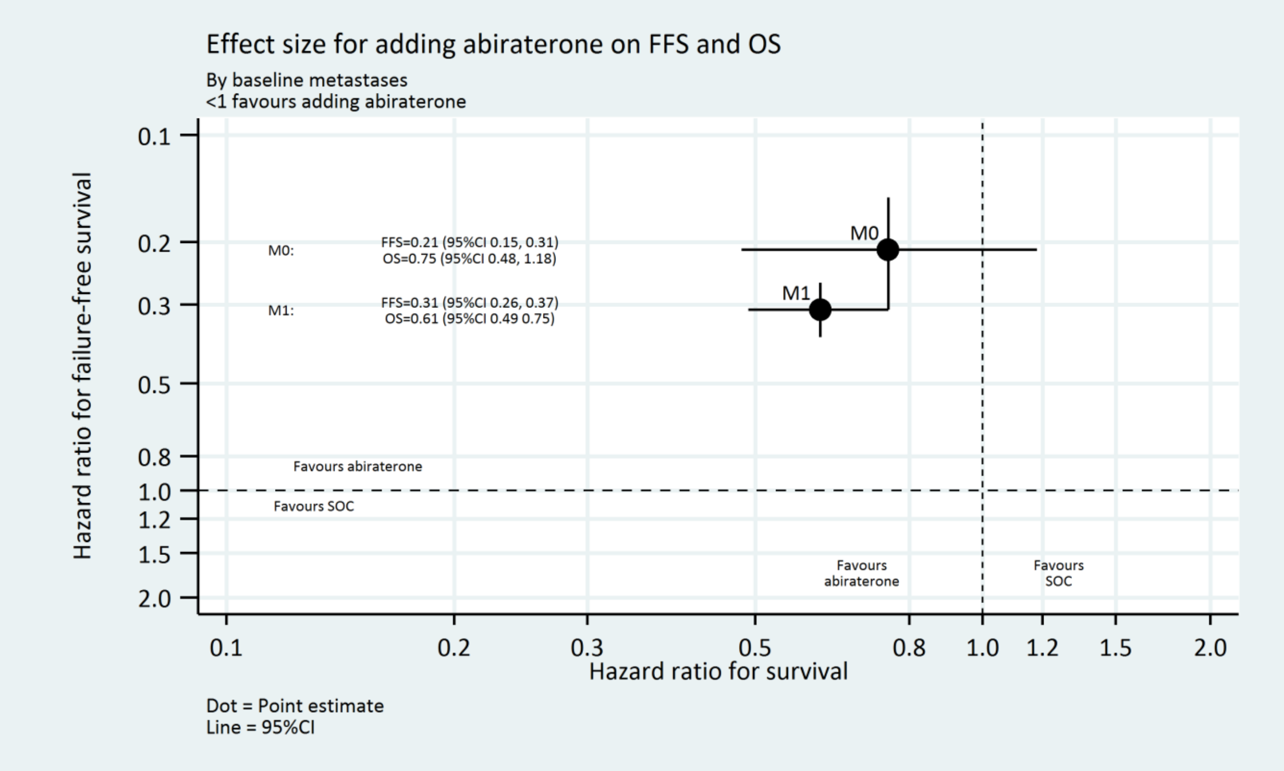
Figure 1: Forest plot from STAMPEDE data showing the effect on survival of adding abiraterone to SOC, within subgroups



SOC vs SOC+AAP

Subgroup	SOC-only Dths/N	SOC+AAP Dths/N	Interaction p-value	Haz. Ratio (95% CI)
Mets:				
M0	44/455	34/460	0.37	0.75 (0.48, 1.18)
M1	218/502	150/500		0.61 (0.49, 0.75)
Born on a:				
Sunday	36/144	29/148		0.68 (0.41, 1.12)
Monday	32/128	15/122	0.33	0.37 (0.19, 0.73)
Tuesday	46/141	26/144		0.45 (0.28, 0.74)
Wednesday	33/132	26/135		0.50 (0.29, 0.86)
Thursday	40/137	25/127		0.69 (0.42, 1.15)
Friday	34/135	34/153		0.95 (0.58, 1.54)
Saturday	41/140	29/131		0.69 (0.42, 1.14)
Diagnosed on a:				
Monday*	38/176	40/167	0.0021	1.24 (0.78, 1.97)
Other Day	224/781	144/793		0.55 (0.44, 0.68)
Overall				0.63 (0.52, 0.76)

Figure 2: Scatter plot of the treatment effect of abiraterone on survival and failure-free survival by baseline metastatic status


Effect size for adding abiraterone on FFS and OS
By baseline metastases
<1 favours adding abiraterone
Hazard ratio for failure-free survival
M0:
FFS=0.21 (95%CI 0.15, 0.31)
OS=0.75 (95%CI 0.48, 1.18)
M1:
FFS=0.31 (95%CI 0.26, 0.37)
OS=0.61 (95%CI 0.49 0.75)
M0
M1
Favours abiraterone
Favours SOC
Favours abiraterone
Favours SOC
Hazard ratio for survival
Dot = Point estimate
Line = 95%CI