'Thursday's child has far to go'—interpreting subgroups and the STAMPEDE trial

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 6 years, it has reported survival data on 6 of these 10 RCT questions over the past 2 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 with 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 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

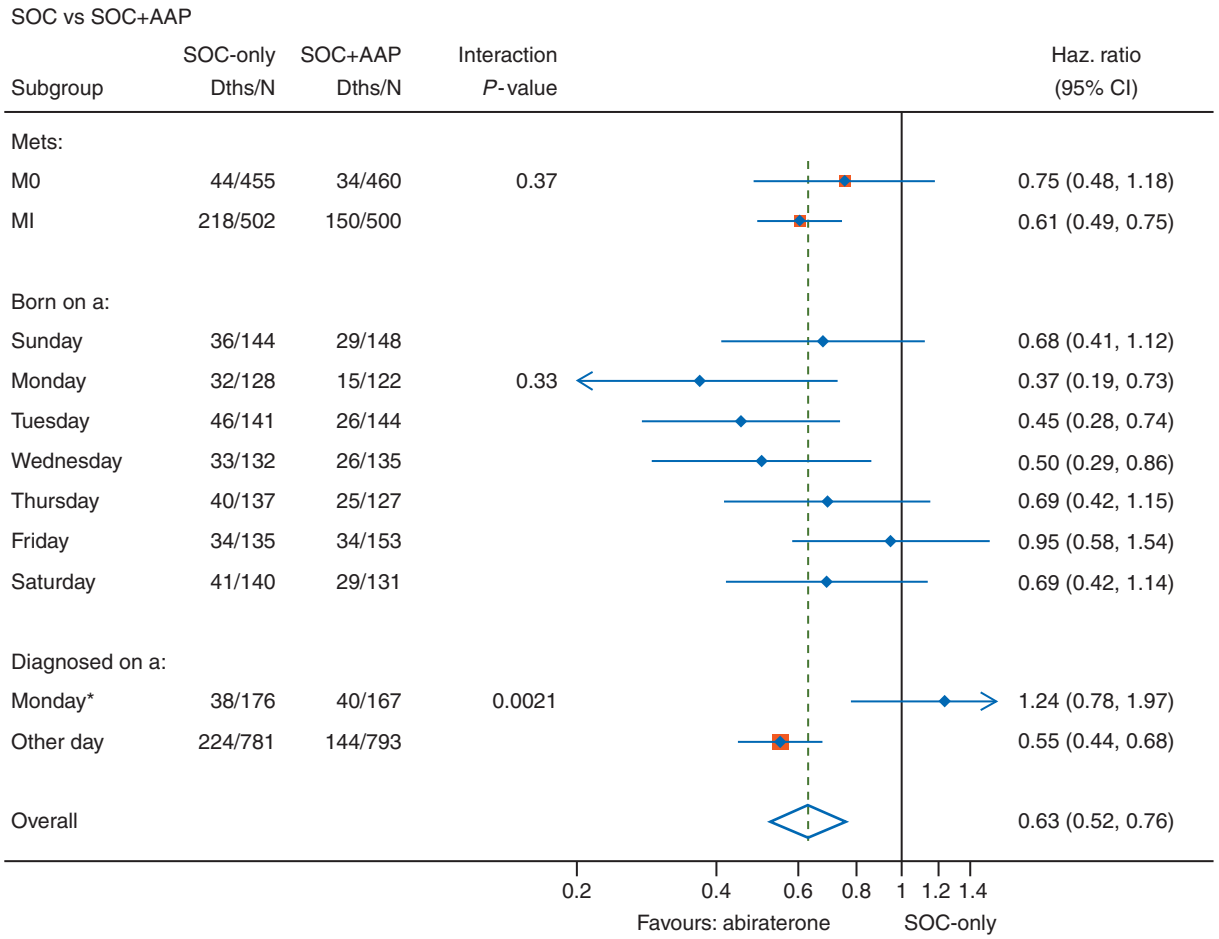

Figure 1. Forest plot from STAMPEDE data showing the effect on survival of adding abiraterone to SOC, within subgroups.

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 0.42–1.15.

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 end points.
- 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.

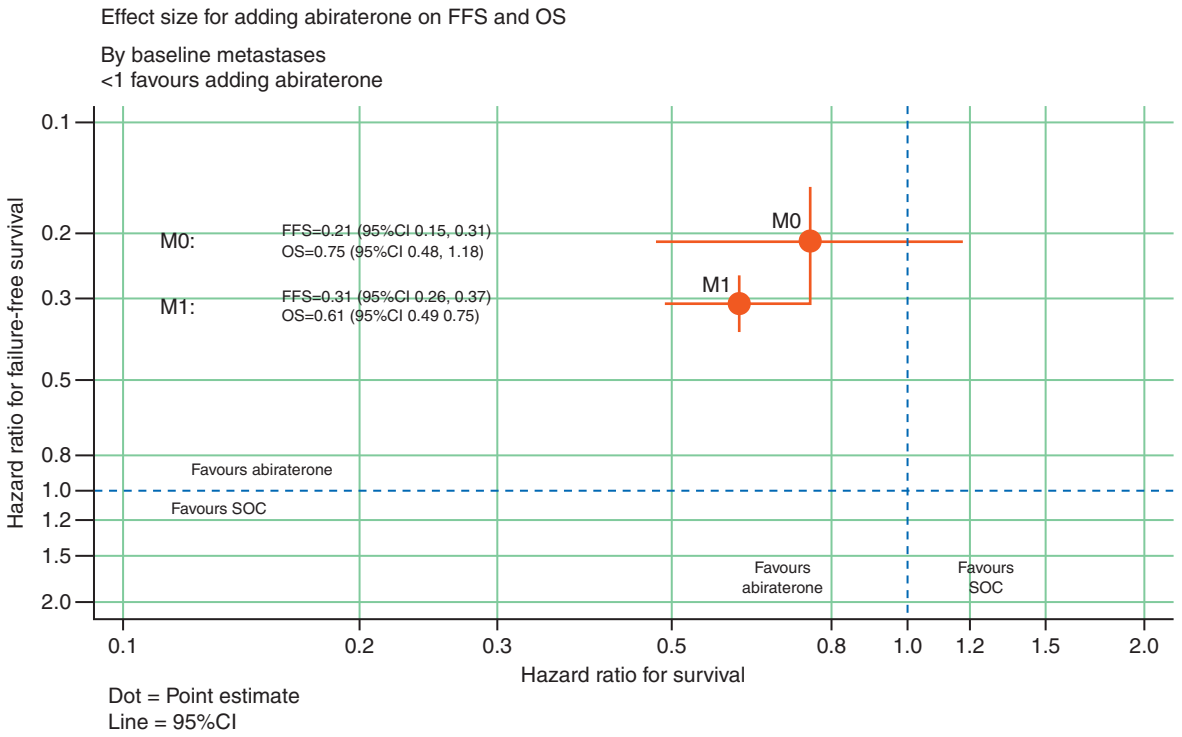

Figure 2. Scatter plot of the treatment effect of abiraterone on survival and failure-free survival by baseline metastatic status.

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 before publication and have a role to play in shaping the message of the publication.

M. R. Spears[1], N. D. James[2] & M. R. Sydes[1*]
[1]MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, University College London, London; [2]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK
(*E-mail: m.sydes@ucl.ac.uk)

Funding

Cancer Research UK (20942); Medical Research Council (MC_UU_12023/25).

Disclosure

MRSp reports grants and non-financial support from Janssen, during the conduct of the study; grants and non-financial support from Astellas, grants and non-financial support from Clovis Oncology, grants and non-financial support from Novartis, grants and non-financial support from Pfizer, grants and non-financial support from Sanofi-Aventis, outside the submitted work.

NDJ reports grants, personal fees and non-financial support from Janssen, during the conduct of the study; grants, personal fees and non-financial support from Astellas, grants and non-financial support from Clovis Oncology, grants, personal fees and non-financial support from Novartis, grants, personal fees and non-financial support from Pfizer, grants, personal fees and non-financial support from Sanofi-Aventis, outside the submitted work.

MRSy reports grants and non-financial support from Janssen, during the conduct of this comparison within STAMPEDE grants and non-financial support from Astellas, grants and non-financial support from Clovis Oncology, grants and non-financial support from Novartis, grants and non-financial support from Pfizer, grants and non-financial support from Sanofi-Aventis, outside this comparison but within STAMPEDE.

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. Lancet 2016; 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. J Clin Oncol 2017; 35: 1530–1541.
3. James ND, de Bono JS, Spears MR et al. Abiraterone for prostate cancer not previously treated with hormone therapy. N Engl J Med 2017; 377: 338–351.
4. NHS England. Clinical Commissioning Policy Statement: Docetaxel in combination with androgen deprivation therapy for the treatment of hormone naÿve metastatic prostate cancer, NHS England, 3019. 2016; (7 July 2017, date last accessed).
5. NICE. Hormone-sensitive metastatic prostate cancer: docetaxel, 2016 Hormone-sensitive metastatic prostate cancer: docetaxel, NICE, 2016 (7 July 2017, date last accessed).
6. Fletcher J. Subgroup analyses: how to avoid being misled. Br Med J 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. Am J Neuroradiol 2011; 32: 633–636.

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: 93–98.
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: 426–432.
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: 1657–1661.
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. N Engl J Med 1978; 299: 690–694.
13. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. Br Med J 2001; 322: 1479–1480.

doi:10.1093/annonc/mdx410
Published online 2 August 2017

Of mice and men: patient-derived xenografts in cancer medicine

A large proportion of oncological care is directed towards patients in whom the fraction of responders is low and the duration of response short with minimal clinical benefit. As acerbically observed by Roy Porter, 'What an ignominious destiny if the future of medicine turns into bestowing meagre increments of unenjoyed life?' [1]. Thus, it is a priority in cancer research to radically shift from therapies which are used indiscriminately across specific tumour types, to personalized therapies which can both improve efficacy and reduce unnecessary toxicities and futile treatments [2].

Much of the work in the area of personalized or precision medicine has focused on identification of genomic alterations in cancers to inform therapeutic decision making. Although there have been major successes using this approach in non-small-cell lung cancer, chronic myeloid leukaemia and other malignancies, the critical problem of intratumoral heterogeneity limits the potential utility of such techniques [3]. Clearly, there are multiple examples of actionable mutations which can be used to make rational decisions for targeted therapies in malignancies, such as non-small-cell lung cancer [4]. However, developing personalized approaches to selection of conventional chemotherapeutic agents has been generally unsuccessful [5]. Even validated biomarkers of responsiveness may not be applicable across tumour types. For example, although methylation status of the MGMT gene is a valuable predictor for determining responses of GBM to temozolomide [6], MGMT methylation may not predict chemosensitivity to temozolomide in non-glioma cancers [7].

To determine tumoral sensitivity to conventional and novel therapeutics, several approaches have been investigated. Organoid technologies have been extremely effective in increasing biological understanding of cancer and have showed great utility in pre-clinical *in vitro* testing of combination therapies, targeted agents and schedules [8]. Tumour-derived organoids can be used for genetic and high throughput drug screening and also to generate xenografts. Although organoids maintain the genomic makeup of the tumours from which they are derived, it is unclear if intratumour heterogeneity is preserved. There is currently insufficient data to validate their routine use as a tool in directly inform treatment decisions. Organoids do not model essential stromal interactions which may profoundly influence the response to therapy. Although co-culture methodologies allow these interactions to be investigated experimentally [9], developing clinically derived organoids with stroma remains a major challenge. Factoring the effects of angiogenesis and immune interactions are also limitations of these models.

An important advance in replication of human tumours has been the use of patient-derived xenografts (PDX) as models for genetic analysis of tumours and expediting therapeutic strategies [10]. In a seminal study by Novartis over 1000 PDX's were used to study potential therapeutic combinations [11]. It was found that PDX's were superior to cell lines in prediction of clinical sensitivity and could provide valuable information on novel targeted combinations for clinical studies. PDX's allowed assessment of inter-patient response heterogeneity, identification of responsive subpopulations and enabled identification of novel predictive biomarkers. They were also valuable in determining novel and previously unappreciated mechanisms of resistance. Two important studies in breast cancer [12, 13] have shown that PDXs demonstrate dynamic clonal evolution, which was measured by deep genome sequencing and single cell analysis, indicating that xenografts, as with the native tumours from which they are derived, are communities of clones. Importantly most of the clonal dynamics occur at implantation and subsequently xenografts remain remarkably stable and polyclonal. A detailed and meticulously curated PDX tumour biobank has allowed detailed characterization of tumour cell genetics, stromal interactions and clonal evolutionary dynamics [13]. This study and others illustrate that the PDX's maintain the essential characteristics of the cancers from which they are derived.

The study by Izumchenko et al. [14] is a comprehensive analysis of the feasibility and utility of deriving large numbers of PDX's from multiple tumour types and included 237 cases. Of note is the high rate of successful engraftment in this series with rates varying from 30% in breast cancers to 85% in colorectal cancers. This is highly encouraging given that samples were obtained from multiple institutions including non-academic facilities and the experience gained from this study will inform further initiatives in this area.

As in other studies [13], propagated tumours preserved histopathological features of the parental tumours and this was maintained through later passages. Whole exome next-generation sequencing was carried out on the 237 early passage PDXs and compared with the TCGA database with high comparability of mutational frequencies. In analysis of the small number of tumours in which genomic analysis was carried out in primary and tumours and matched PDX's, there was generally tight correlation with 88% of mutations identified in parental tumours present in the corresponding PDXs.

Comparison of responses to therapeutic combinations of drugs again showed a good correlation between responses of PDX's and the clinical results. As this is a major rationale for this study the results are of interest. Both positive and negative therapeutic responses in the clinical scenarios could be replicated with PDX's