Testing many treatments within a single protocol over 10 years at MRC CTU at UCL

Multi-arm, multi stage platform, umbrella and basket protocols

Mahesh KB Parmar, Matthew R Sydes, Fay H Cafferty Babak Choodari-Oskooei, Ruth E Langley, Louise Brown, Patrick PJ Phillips, Melissa R Spears,, Sam Rowley, Richard Kaplan, Nicholas D James, Timothy Maughan, Nicholas Paton, Patrick J Royston

Abstract

There is real need to change how we do some of our clinical trials, as currently the testing and development process is too slow, too costly and too failure-prone - often we find that a new treatment is no better than the current standard. Much of the focus on the development and testing pathway has been in improving the design of phase I and II trials. In this paper we present examples of new methods for improving the design of phase III trials (and the necessary lead up to them) as they are the most time consuming and expensive part of the pathway. Key to all these methods is the aim to test many treatments and/or pose many therapeutic questions within one protocol.

Introduction

There are a number of challenges in the development and testing of new therapies. These include: (i) the development and testing process is too slow, takes too long and too often shows that new is not better than our current standard [1]; (ii) in some diseases the number of new therapies and combinations of therapies demanding evaluation is large, much larger than it has been for many years, due, in large part, to a revolution in biology resulting in a rational selection and synthesis of agents directed at specific targets; (iii) some diseases are being classified into smaller subsets, often using molecular characterisation; and (iv) the process of developing and starting a new trial is costly and time consuming, and as a consequence there is often too long a gap between phase II and phase III trials, and also between separate phase III trials, particularly in the academic world.

Much has been written about some of these problems and a number of solutions have been proposed. However almost all of these solutions have been proposed for the phase I and particularly phase II part of the testing process [2] [3] [4]. In this paper we present some solutions in phase III testing, with the aim of speeding this part of the process, and also increasing the chance, within an individual protocol, of reliably identifying at least one new treatment that is superior to the control treatment [5]. The rationale for this focus is that the phase III part of the process is usually the longest and the most expensive part of the testing process. The details of some of these solutions have been described elsewhere [6] [7] [8]. Here we give a brief summary and examples. The four principles underlying these solutions are:

1) evaluate many primary hypotheses/treatments within the same protocol; this maximises the chance of identifying a new treatment which is better than the current standard [5]; aim to make the research arms as different as possible from each other;
2) if there is a pilot/feasibility/phase II part of the process, aim to have a seamless run through to the phase III part of the process and, if at all possible, include in the phase III evaluation the information from patients in the early evaluation;

3) whenever possible and appropriate, conduct an adaptive trial, with the adaptations being major ones, such as ceasing randomisation to a research arm or introducing new research arms;
4) in situations where we are considering subgroups of a specific disease, often biomarker-defined, aim to include questions testing new treatments in all (or most) subgroups, using an adaptive approach to allow: (i) refinement of the subgroups; (ii) introduction of new subgroups; (iii) the ability to stop testing specific treatments and introduce new treatments; and (iv) evaluation of the link between the biomarker and that treatment.

Below we present some examples of the application of these principles to particular areas, and show how they have led us to develop and launch multi-arm, multi-stage platform, umbrella and basket protocols. We also address the major statistical issue when there are many primary hypotheses, the probability of making at least one false positive discovery when multiple hypothesis tests are performed and the null hypothesis is true. This is called the family wise error rate (FWER). We also consider the implication of adding arms.

Multi-arm, multi-stage platform protocols

STAMPEDE (launched in 2005)

Prostate cancer is one of the most common cancers globally, accounting for 1.1 million cases and 0.3 million deaths each year [9]. A substantial proportion of these men are diagnosed with, or progress to, high-risk prostate cancer. This includes men whose cancer has already spread to other parts of the body, usually the bones, or which is restricted to the pelvis but has other adverse prognostic features. These patients are often referred to as having 'hormone sensitive disease'. In the early 2000s, when STAMPEDE was being designed, the standard-of-care for these men was a treatment which had been introduced more than 40 years before - long-term hormone therapy; aiming to control the disease by reducing the level of the male hormone, testosterone. Despite this treatment, outcomes for these men remained poor, with a five year survival of 40% which had not changed over these 40 years.

At the time of designing STAMPEDE (www.stampedetrial.org) there were a number of treatments showing promise in later stages of prostate cancer and, working with clinical experts in the UK (the National Cancer Research Institute Prostate Cancer Clinical Studies Group and its predecessor committees), these were systematically considered. Rather than, somewhat arbitrarily, selecting a single treatment in which to invest all efforts and resources, it was agreed that it would be better to develop a multi-arm, multi-stage platform clinical protocol so that a small number of the most promising of these treatments could be tested simultaneously. A key criterion in selecting agents to test was their mode of action, with the aim of choosing treatments which had differing modes of action. There was a greater chance in improving outcomes if, in the different research arms, quite different (rather than similar) treatments were tested. Furthermore, treatment arms that were successful could then potentially be combined to achieve even greater improvements in survival. In this sense, the new treatments would not be considered to be competing with each other.

Ultimately, three new treatments were selected: a third-generation bisphosphonate (zoledronic acid, made by Novartis); a taxane-based chemotherapy (docetaxel, Sanofi-Aventis) and an oral Cox-2 inhibitor (celecoxib, Pfizer). Pre-clinical data were also available to support inclusion of two combinations of pairs of these agents. Therefore, a protocol with one control arm (standard of care) and 5 research arms was launched, each of these 5 arms supplementing the standard-of-care with one or two of these agents. Figure 1a shows the design of the protocol at initiation in 2005. One way of viewing this protocol is that it is just a series of 2-arm trials, with each research arm being compared against a (common) control arm for overall survival, all of which happen to be in the same protocol.

Calculations showed that to answer these 5 research questions it would take us approximately 7 years to randomise up to 3,500 patients (with a further minimum follow-up of 3 years) to provide a fully-reliable answer on the primary outcome measure, overall survival. This was clearly ambitious and involved a larger number of patients than any previous trial in prostate cancer had targeted. To help mitigate this challenge, and recognising the fact that it was very unlikely that all the new treatment regimens would be effective, a lack of benefit analysis at a number of stages was planned, comparing each research arm against the control arm in terms of the intermediate outcome measure of failure-free survival. For a multi-arm trial with k research treatments the optimal randomisation ratio between the control and each research arm is $k^{½}$ [10]. In STAMPEDE we started with 5 research arms, therefore as an approximation we randomised patients in the ratio of 2:1:1:1:1:1 control to each research arm.

The biology of the disease and the drugs being tested meant that it was anticipated that any advantage in survival would be preceded by a reduction in markers of prostate cancer activity. Therefore, an intermediate outcome measure of failure-free survival was defined, primarily driven by rises in prostate-specific antigen (a blood marker for prostate cancer), assuming that any advantage in overall survival would be preceded by a benefit in failure-free survival; if there was little or no benefit in failure-free survival there was very unlikely to be a benefit in overall survival. Interim analyses were therefore set up to consider failure-free survival with the aim that recruitment could stop early to any research arm that did not show sufficient activity (compared to the control arm) on this interim outcome measure, rather than waiting for overall survival data. Note that it was not assumed that a benefit in failure-free survival would necessarily translate to an advantage in overall survival. Full details of the design of the trial are given in [11].

Given the ambitious nature of the design and the fact that some of these regimens, particularly the combinations, were being used for the first time in this setting, the trial was initially started in just a few sites in a pilot stage. The Independent Data Monitoring Committee monitored the first 210 patients to confirm the safety of all the research arms before the trial was opened to all sites. Ultimately, more than 100 sites in the UK and 7 sites in Switzerland have randomised at least one patient to the trial.

In practice, 3,585 patients were recruited to these 6 arms. Randomisation to two of the research arms, both of those containing celecoxib, was stopped on the advice of the Independent Data Monitoring Committee at the second interim activity analysis in April 2011 and data supporting this decision were published [12]. The other three research arms successfully recruited past all three interim activity stages and recruitment was completed in March 2013.

All randomised patients were followed until the pre-planned number of deaths (approximately 400 in the control arm) were observed in 2016 [13]. These data, together with other relevant randomised data [14], showed clear evidence of a survival advantage of the chemotherapy drug docetaxel when added to standard of care. Practice in many countries, including the UK, changed rapidly afterwards to allow chemotherapy in this population of patients. Adding zoledronic acid to standard of care showed no evidence of a survival advantage, and whilst the combination of docetaxel and zoledronic acid did show evidence of a survival advantage, there was no evidence that this was better than adding just docetaxel alone.

As a consequence of these results (and the contemporaneous meta-analysis including these results [15]) the design of STAMPEDE was changed immediately in January 2016 to allow patients to receive the drug docetaxel as part of their standard of care. In recent months, four fifths of metastatic patients have planned docetaxel as part of their standard care. STAMPEDE may be a unique trial in which the standard of care for the trial has been changed as a consequence of results from the trial itself.

Adding Arms to STAMPEDE

In 2010, five years into STAMPEDE recruitment, the drug abiraterone was showing very promising results in patients with prostate cancer at more advanced stages than those entering STAMPEDE – i.e. in patients with hormone refractory disease [13]. At this time there were three realistic options: (i) choose to not test abiraterone in this setting and leave its assessment to others; (ii) open a competing trial, which could hamper recruitment to both STAMPEDE and the abiraterone trial and mean that more men were allocated to a control arm across the two trials; or (iii) incorporate abiraterone as a new research arm within STAMPEDE. The last of these options is clearly the most efficient in terms of the total numbers of patients required, recruitment of patients, time taken to initiate and approve a protocol, the workload for initiation of sites, and overall costs.

To incorporate abiraterone involved a new industry partner, Janssen. This new arm, and the 'continued' control arm, was planned to recruit to a specific target number of patients during and after the recruitment period for the original research arms. Patients in this new research arm are only compared with contemporaneously randomised controls. The abiraterone research arm opened to recruitment in November 2011 (four years before the docetaxel results emerged) with an initial target of 1500 patients. It was added through a protocol amendment (and not the development and approval of a whole new trial), providing huge efficiency and speed. It closed to recruitment 9 months earlier than planned, despite an expanded target of 1800 patients, with 1917 patients recruited by January 2014 (again, before the primary results of the original comparisons were mature). This showed clearly that adding an arm into an ongoing assessment could speed the initiation of, and recruitment to, a new comparison. Results from this comparison have recently been presented and have shown that adding abiraterone improves overall survival; the second positive result to emerge from STAMPEDE [16].

The clear success of adding the abiraterone arm subsequently led to the incorporation of three additional research arms into STAMPEDE (figure 1b). Each of these arms has been developed with the traditional approach in terms of scientific rigour. In particular, a persuasive scientific case has to be made, and successful independent, international peer-review, as well as funding, has to be obtained. One helpful way to think about these new research arms is that a new comparison is being

launched, which has been added to the STAMPEDE protocol. The three arms which have been added are: (i) the addition of radiotherapy to the prostate in the subgroup of men whose disease had already spread; (ii) the combination of abiraterone and enzalutamide (bringing a further drug company into STAMPEDE- Astellas); and (iii) the addition of metformin, seeking to repurpose a drug widely used to manage diabetes. A further two research arms will be initiated in 2017. Figure 1b shows the history of initiation, recruitment and analysis activities to the protocol to 2024 [17][18][19].

Family Wise Error (FWER) and Adding Arms

The FWER for the three original research arms that reached their planned target was 6.75% [13] [20]. For the added research arms, ongoing work has shown that the implications for the FWER are dependent on the proportion of overlap in control arm patients between the research arms. If there was overlap of only one patient in the control groups for two different research arms, this would be (almost) like doing two independent trials with one common patient. In this instance, there would be no practical change to the type 1 error for these two comparisons..

In Figure 1b we can observe that the only significant overlap of control arm patients is between research arms H and J, where some allowance for multiple testing may need to be considered – for the remaining new research arms there is relatively little overlap and thus they can be regarded, in practical terms, as independent trials. This is confirmed, for example, by calculating the correlation between the test statistics comparing arm G with contemporaneous controls and the original 6 arms with their contemporaneous controls. The estimated correlation between these test statistics is 0.12 – emphasising the lack of overlap and their relative independence.

By 2020, STAMPEDE will have answered 8 major questions in the 15 years since the first patient was entered; a series of sequential 2-arm trials would have taken many tens of years longer to achieve this. Results from arms B, C, D, E, F and G have already been reported and standard of care has changed as a consequence (figure 1b) [12][13][16][21]. In the years 2015 to 2020, STAMPEDE will have produced a major new result on average every 18 months. With this approach there is a real opportunity to improve outcomes for patients with hormone sensitive prostate cancer in a major way over this period. Future plans for STAMPEDE include testing further treatments, particularly those which are targeted at specific groups of patients in whom they are likely to be most effective.

TRUNCATE-TB (launched in 2017)

Tuberculosis (TB) is one of the oldest known diseases and was considered to have been brought under control with the introduction of very effective treatment more than 30 years ago [22]. Today TB kills more people than any other single infectious disease and, in 2015 alone, TB caused 1.8 million deaths. TB disproportionately affects poorer communities; low income countries that represent only 42% of the world's population account for 65% of TB cases and 71% of deaths. The standard length of treatment for the vast majority of patients (with drug sensitive disease) is 6 months with a cocktail of 4 drugs. In clinical trials, 95% of patients are cured. However, in these trials all patients are treated and followed to a strict protocol and therapy is given in the clinic. In routine use, such close supervision is not possible and adherence to treatment is consequently poorer –

some patients stop taking the medication when they feel better, often after a few months. As a result, the observed cure rate is usually 85% or less.

A shorter treatment regimen that has better adherence in routine use may provide better overall outcomes, even if the efficacy might be slightly lower when tested in a clinical trial. Reporting in 2014, three major phase III randomised trials demonstrated unacceptable relapse rates in excess of 10% in regimens of 4 months in length (compared to the standard 6 months of treatment) and pre-specified non-inferiority criteria were not met. However, if a new treatment regimen is very short, say of 2 months duration, then a relapse rate of 10%, or even higher, might be acceptable, since patients who experience a relapse after such treatment can, in general, be treated successfully with 6 months of therapy. Such a strategy may indeed be more successful than a first-line treatment of 6 months of therapy because trial results would more readily translate into routine practice due to improved adherence. There may be other benefits in that a 2-month regimen would likely be more cost-effective, more attractive to patients, reduce the risk of toxicity and fit well with the initiation of treatment for HIV (as many patients with TB will be diagnosed with HIV at the same time).

There are a number of new and repurposed drugs in clinical development leading to a large number of potential 3-4 drug combinations that could be combined to form a 2-3 month first-line treatment strategy. Similar to STAMPEDE, rather than choose one regimen somewhat arbitrarily, the TRUNCATE-TB trial is evaluating, in the first instance, four novel 2-3 month regimens using a MAMS platform design (www.sprinttb.org/theme-3-clinical-trials). Investigators have used the best available evidence and expert opinion to decide on the combination regimens that have the highest chance of success in a clinical trial. The process resulted in four high priority regimens that are included in the TRUNCATE-TB trial, with further regimens identified should the opportunity arise to initiate new arms (again a process which has been successful in STAMPEDE). TRUNCATE-TB is being conducted by hospitals in East Asia and is a collaboration between the MRC Clinical Trials Unit at UCL and the National University of Singapore. Accrual will start in mid-2017.

The design of TRUNCATE-TB is shown in figure 2 and details are given in [23]. In short, each research arm will be compared against the control arm to assess whether it is non-inferior (by a pre-specified amount) in terms of the proportion of patients with an unsatisfactory clinical outcome at 96 weeks (approximately 2 years) from randomisation. Two interim analyses are planned and the final analysis will be conducted with a one-sided 1.25% significance level (using a Bonferroni correction and assuming that two arms will reach the final stage) to maintain the overall one-sided FWER at 2.5% .

Umbrella protocol: FOCUS 4 (launched in 2014)

Oncology is, in the vanguard of personalised or stratified medicine, an area which requires new and more efficient trial designs [24]. One such design is FOCUS4 [7] (www.focus4trial.org) (figure 3). FOCUS4 is an umbrella design aimed at to identifying effective treatments in patients with metastatic colorectal cancer (CRC). An umbrella protocol is one which aims to include patients with a single disease divided into subsets, often with different therapies being explored in each of these subsets.

FOCUS4 stratifies all patients with metastatic CRC into a hierarchical structure of predominantly genetic mutations that are known to play an important role in the tumour biology of advanced CRC. Within each of these strata, randomised comparisons are opened to test therapies that have been designed to target these specific molecular alterations. Each trial is placebo-controlled whenever possible, and each uses multi-stage methodology to test for drug activity on progression-free-survival at pre-defined time points during each sub-trial. Each sub-trial has its own control arm and has been designed, and will be analysed separately, meaning that there is no issue regarding multiple testing.

The design of FOCUS4 follows seven specific principles [7]:

1) Enables the testing of multiple treatment/biomarker combinations at the same time

2) Uses an enrichment strategy that, at first, only tests the targeted treatment in the biomarker positive group

3) Uses randomised comparisons within each therapy/biomarker evaluation to allow for any prognostic effects of the biomarker (the magnitude of which may often be uncertain)

4) Uses a multi-stage statistical design with pre-defined interim analyses for assessment of drug activity such that drugs failing to provide adequate levels of activity are dropped, whilst more promising drugs can move seamlessly from a phase II to a phase III setting without having to start a new trial

5) For treatment/biomarker combinations that are demonstrating promising drug activity in early interim analyses, the design allows testing of the specificity of the biomarker by opening up trial recruitment to patients who are negative for that biomarker.

6) Allows new treatment/biomarker combinations to be dropped, adapted or added to the umbrella protocol platform in response to an ever moving scientific landscape. As these are essentially separate sub-trials, this is not complex (in terms of the statistical design and operating characteristics).

7) Provides a clinical trial opportunity for all patients, regardless of their biomarker status. Rare biomarkers can be evaluated in the platform alongside more common ones, and patients whose tests have shown no clear result for any biomarker can enter a trial testing a non-stratified research question.

FOCUS4 opened to recruitment in January 2014 and is currently evaluating three treatments in four biomarker-defined groups. One further treatment has already been dropped due to lack of sufficient activity. Over the period of the trial it is hoped that at least 6 new drugs will be evaluated with a number of them advancing to both phase III testing and assessment of biomarker specificity.

Basket protocol: Add-Aspirin (launched in 2015)

The re-evaluation of an already established medicine for a new clinical indication, a process known as re-purposing, is a potentially attractive way to improve outcomes in many diseases. With re-

purposed drugs many traditional aspects of the drug development pathway do not have to be undertaken, speeding the pathway to ultimate phase III testing. For example, some data on efficacy and safety is often already available, typically through epidemiological and/or post-marketing data. Similarly, their use and co-administration with existing therapies is sometimes known, and toxicity and interaction profiles may already be established. Finally, and importantly, generic medicines are often more readily available in resource-limited settings, where much of the burden of some diseases is found – for example 70% of the new cancer patients are found in the resource poor world. Effective repurposing of drugs, therefore, offers the potential for a major impact on a number of diseases worldwide.

In cancer aspirin may be the exemplar of a repurposing agent – preclinical data over 40 years has shown its anti-cancer properties and, in particular, the potential to reduce the formation of metastases [25]. Initial clinical investigation focussed on primary prevention of cancer, particularly colorectal, and it is only with the more recent evaluation of long-term cancer outcomes from randomised trials in the prevention of vascular disease that the full potential of aspirin as an anticancer therapy has been recognised. In these trials, allocation to aspirin resulted in individuals developing fewer cancers across a range of tumour sites. Individuals taking aspirin who did develop cancer were less likely to have metastases at diagnosis and less likely to develop them subsequently [26].

This body of evidence has led to the development of the Add-Aspirin trial – a basket protocol encompassing 4 individually powered, randomised, double-blind, phase III adjuvant studies in breast, colorectal, gastro-oesophageal and prostate cancer (www.addaspirintrial.org). One definition of a basket protocol is one which includes patients with many different diseases, but the patients are linked because there is a treatment which targets a pathway which is common to all these patients.

In the Add-Aspirin trial all potential participants have undergone primary potentially curative therapy (surgery +/- adjuvant/neoadjuvant chemotherapy/radiotherapy or primary chemoradiation). Following a run-in phase where adherence and toxicity on aspirin are assessed, participants are randomised to 100mg aspirin, 300mg aspirin or matched placebo to be taken daily for at least 5 years (figure 4). Full details of the protocol can be found in [8]. In brief, for each cancer type the principal comparison is (for both doses combined) of aspirin vs placebo. In all those cancers showing a positive result, the groups receiving 300mg against 100mg of aspirin will also be compared. Since the latter comparison is a closed test, conditional on seeing a positive result in the primary comparison, there is no impact on the FWER. The trial also has a very long-term co-primary outcome measure of overall survival (analysed at approximately 15 years) across the four tumour cohorts encompassing effects on disease recurrence, potential vascular (and other) benefits and second malignancies. This is important as positive results across all four malignancies might suggest the use of aspirin as an adjuvant treatment for many other cancers not studied within the Add Aspirin protocol – it would be unrealistic to conduct an adjuvant trial in all such malignancies. This analysis will be performed with a more stringent significance level to control the FWER across analyses to be 5% (two-sided). Finally, in the colorectal cancer cohort there will be pre-planned and appropriately powered analysis in the patients whose tumours show PiK3CA mutation, whether or not the overall results in these groups of patients are positive or not. There is some evidence to suggest that individuals with PiK3CA mutation in their tumour might benefit when others do not (or, even when

others do benefit, that these individuals might benefit the most). The trial therefore potentially answers many therapeutic questions beyond those specific to the four tumour types included.

Discussion

There is real need to change how we do some of our clinical trials, as currently the testing and development process is too slow, too costly and too failure-prone - often we find that a new treatment is no better than the current standard. Much of the focus on the development and testing pathway, to date, has been in improving the design of early phase trials. In this paper, we have concentrated on new methods for improving the design of late phase trials (and the necessary lead up to them), as they are the most time consuming and expensive part of the pathway. Key to all these methods is the aim of testing many treatments and/or posing many therapeutic questions within one protocol together with an adaptive approach allowing introduction of new research questions as they arise.

The most mature example is STAMPEDE which was launched in 2005. By 2020, the trial will have produced 8 major results (arms B to J in figure 1b). Standard of care for these patients has already changed as a result of findings from the trial, and the trial has adapted to this changed standard of care. With more traditional designs, a number of separate trials evaluating each of the new therapies might have been conducted sequentially or, if running in parallel, would have used separate control groups. Thus, it would have taken many more decades and required many more patients to achieve such outputs. STAMPEDE continues to recruit well, as do the two trials launched more recently, FOCUS4 and Add-Aspirin. TRUNCATE-TB will be launched in the summer of 2017.

There are nonetheless clear challenges when embarking on such protocols. They include the need: to garner large scale collaboration bringing large parts of the research community together; to obtain significant and long-term funding; to obtain long term commitment from the key research leaders; to ensure that responsibilities (and also acclaim) are shared as widely as possible, and to have operational structures and systems which allow the implementation of such long-term adaptive protocols. These challenges need to be addressed when the protocol is at the design stage, as they will need to be resolved before any funding is likely to be approved and released.

A detailed discussion of the timing and frequency of interim analyses is beyond the scope of this paper. However, it should be noted that the criteria for, and implications of, stopping for lack of benefit are quite different to stopping for overwhelming benefit. In the former the emphasis is usually on the current estimate of the 'treatment effect' on an intermediate outcome measure, and if it is small (or null/negative) then we may conclude the likelihood of a worthwhile treatment effect on the primary outcome measure is also likely to be small. Usually, stopping further randomisations to a research arm for lack of benefit has no implications for either the control arm or other research arms in the trial. In contrast stopping an arm for overwhelming benefit usually focuses on the need for a small p-value for the 'treatment effect' on the primary outcome measure. Further, stopping for an overwhelming benefit has direct implications for the control arm in particular, and potentially all of the other research arms.The designs above are themselves evolving and will continue to do so, each using the platform that has been developed to, for example, test further therapies in subsets or the whole population , or by evaluating the link between a biomarker and a treatment. For example, STAMPEDE is developing randomised questions of targeted therapies for patients with specific genetic mutations, taking aspects of the FOCUS4 design into its design. The Add-Aspirin

protocol is already doing the same and is also considering adding further randomisations testing other re-purposed therapies, either by adding arms or by adding a factorial randomisation. Software to aid the design of these more complex studies is also being developed alongside the trials themselves – for example, the Stata program nstage for MAMS designs[27] and nbinstage for MAMS trials with a binary outcome (freely available from the Statistical Software Components archive at www.repec.org). These will support the development of such trials by other research groups.

A question often posed is: are there an optimal number of arms for a multi-arm trial? The short answer to this question has to be no, as perhaps the biggest drivers of the number of arms are the number of patients available, the number of research treatments that are ready and available for testing and the cost of undertaking a protocol.

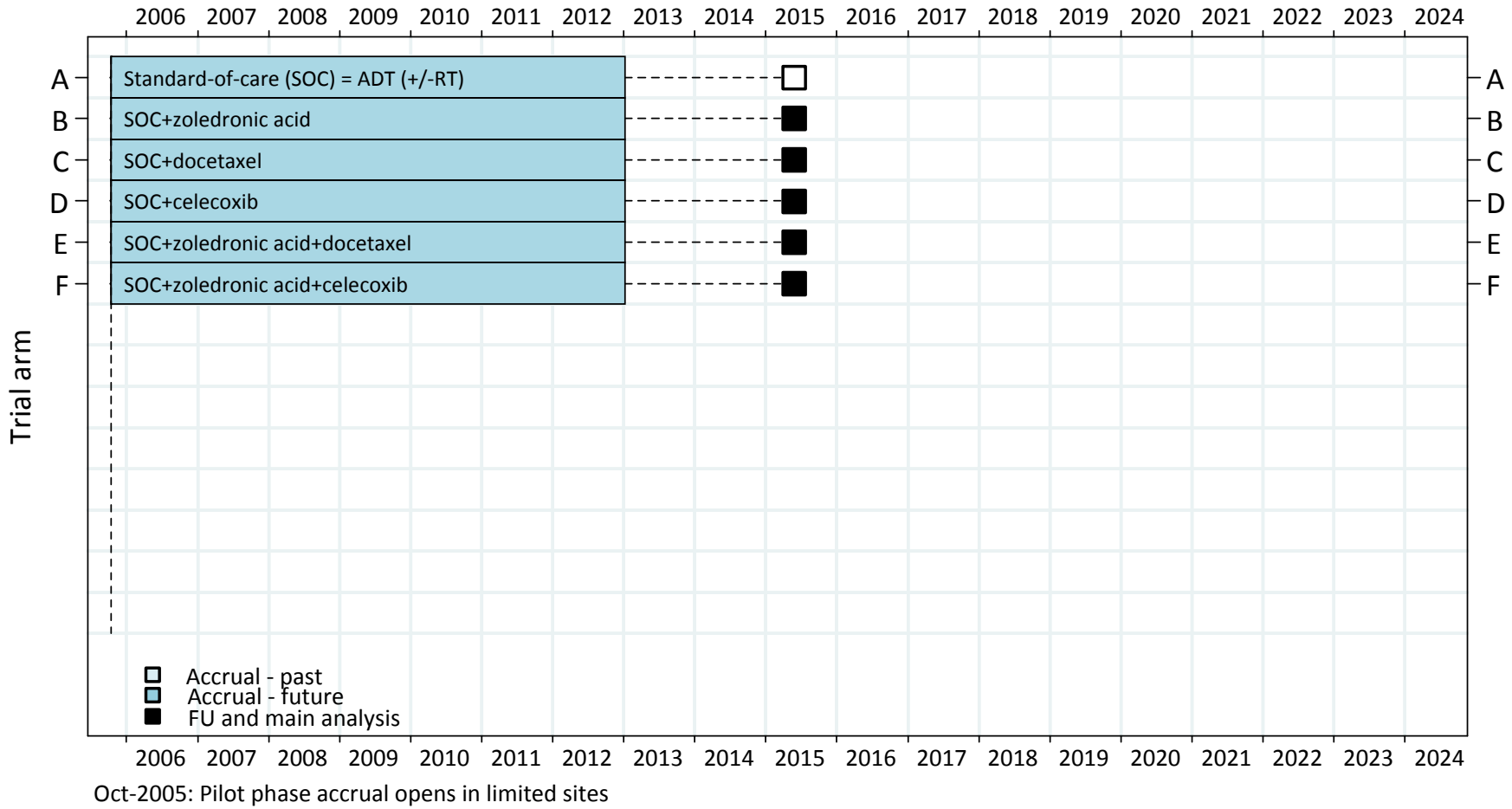
The designs presented in this paper do not use adaptive randomisation methods (other than ceasing accrual to some research arms and initiating new research arms) as deployed in other multi-arm early phase trials [2,3]. The reasons for this are as follows. First, and perhaps most importantly, there are conflicting views on whether adapting the randomisation ratio actually achieves the gains some have claimed [28], particularly in situations where a lack of benefit analysis is already planned. Second, in protocols where the intermediate and final outcome measure are different, the intermediate outcome measure is used only a screen to reject research arms which are unlikely to be effective on the final outcome measure, not as means of assessing whether a research treatment is clinically more effective . In this situation it would be inappropriate to weight the randomisation based on an interim analysis on the intermediate outcome measure. Finally, it is complex and challenging to implement and explain adaptive randomisation methods into large-scale multi-centre randomised late phase trials. Given the lack of clear benefits and these challenges, we have not used them in any our late phase randomised trials.

Our aim in this research programme is to change the fundamental goal of a single clinical protocol from the typical 2-arm randomised trial of 'testing a new treatment' to 'testing many treatments and hence improving outcomes as rapidly as possible'. Clearly not all trials can run in this way. Nevertheless, in situations where: (i) there are a number of new therapies that might be tested; (ii) where the same therapies may be used in multiple related diseases, and/or (iii) where there is sub-categorisation of a single disease into subgroups who might benefit from different new therapies, there is real need to change how we think about a single protocol.

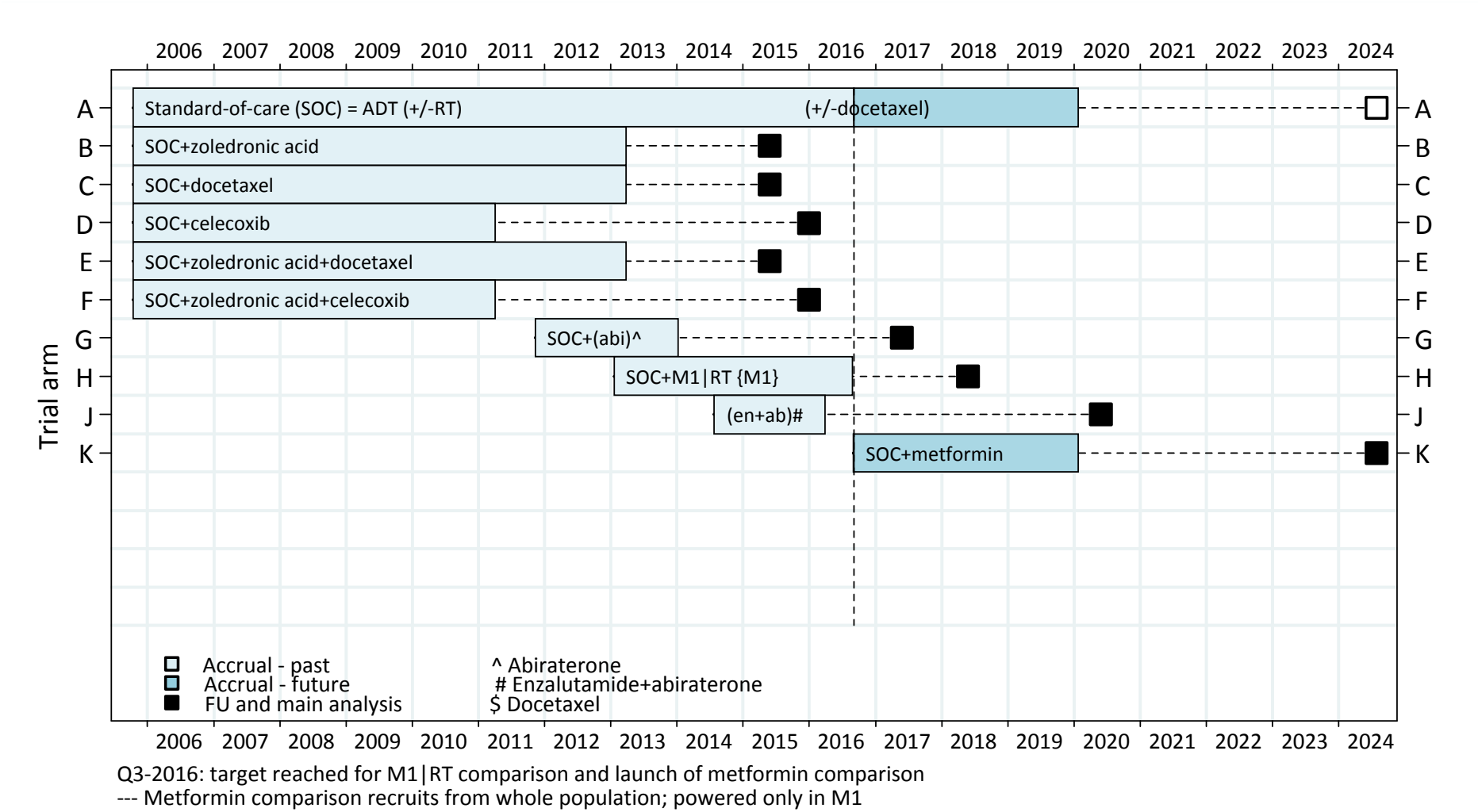
We are exploring opportunities to develop such protocols in areas such as Alzheimer's disease, multiple sclerosis, HIV prevention, hepatitis C and wound healing, to speed the process of testing new therapies, and we urge others involved in the design of new trials to follow this way of thinking.

Figure 1: STAMPEDE Protocol (a) at initiation (figure produced in Oct-2005) and (b) adapted protocol from 2005 to 2024 (figure produced in Sep-2016)

(a)

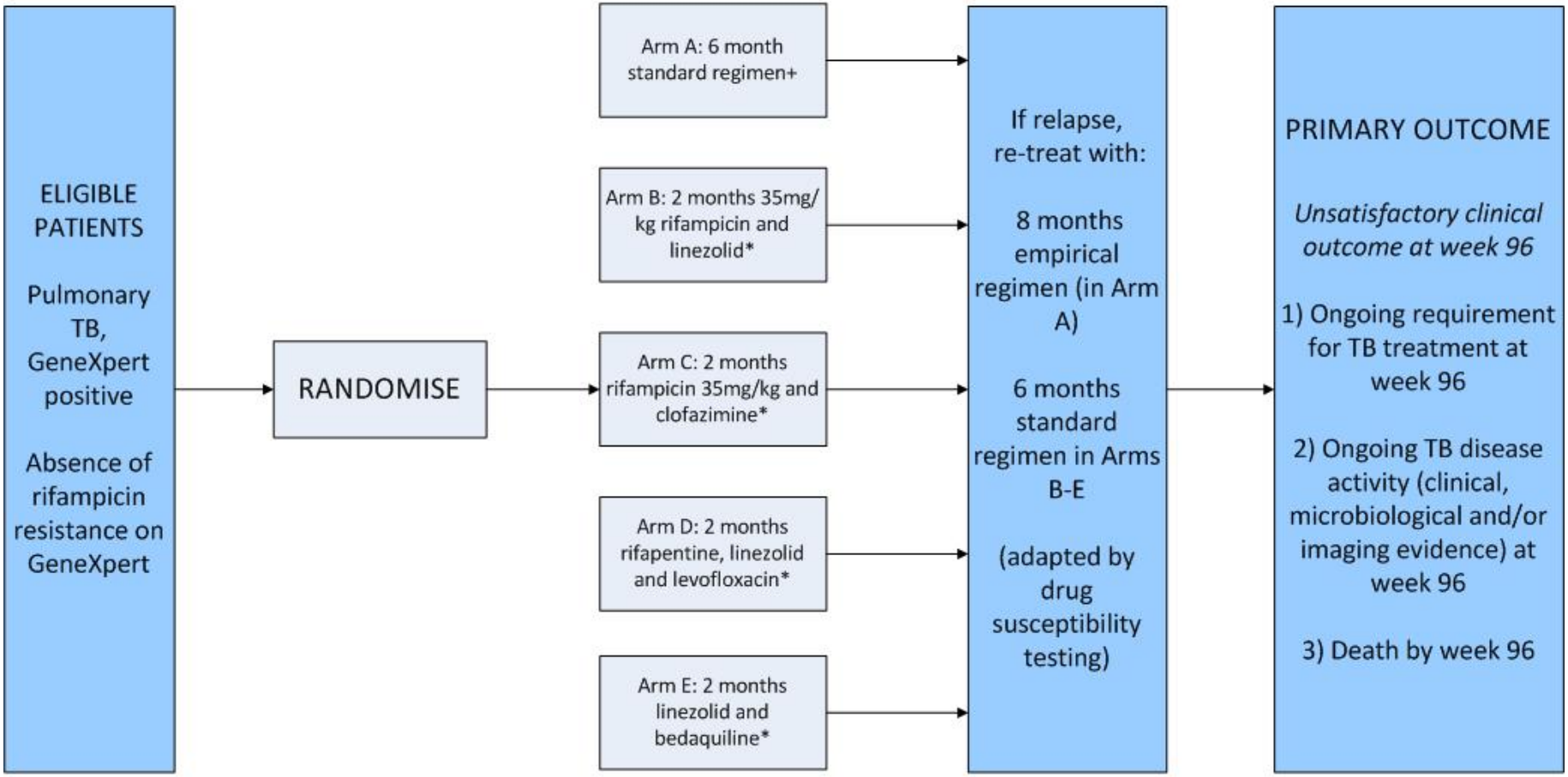


(b)



Q3-2016: target reached for M1|RT comparison and launch of metformin comparison
--- Metformin comparison recruits from whole population; powered only in M1

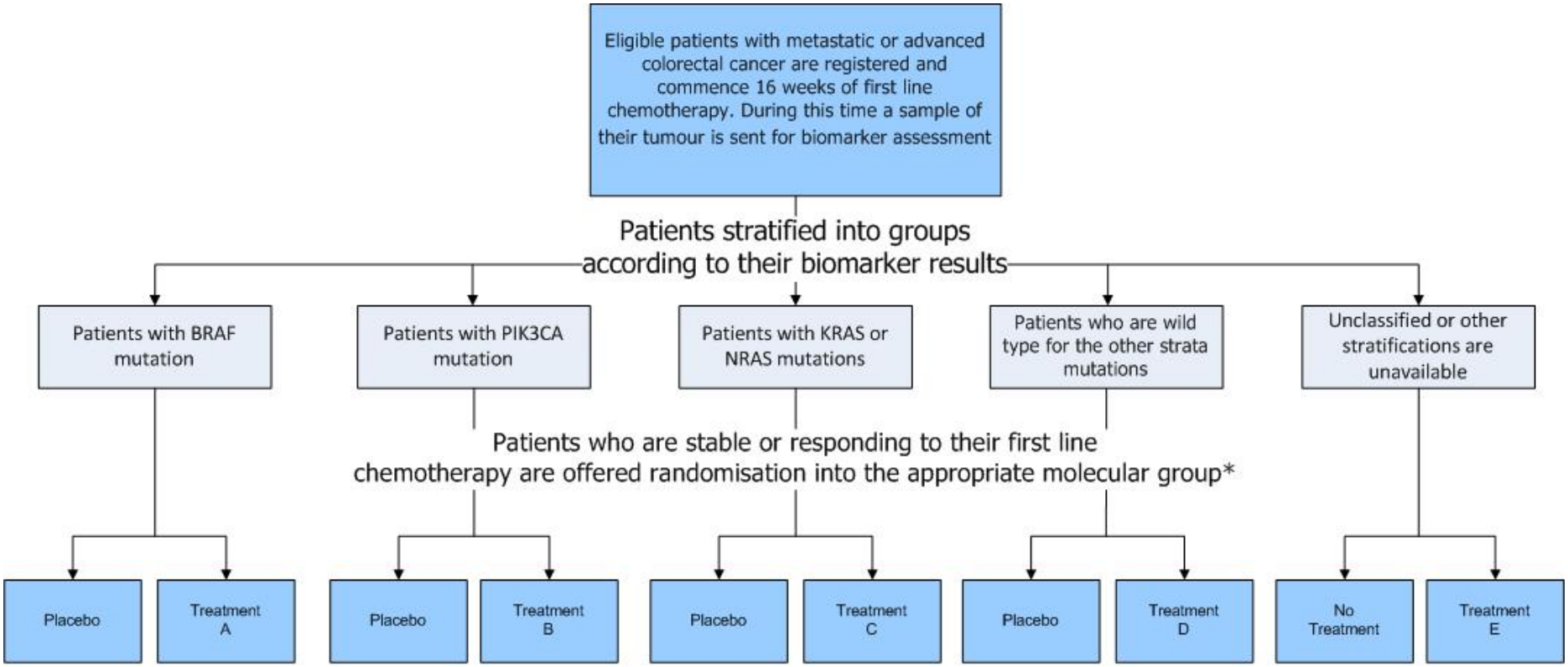
Figure 2: Protocol Schema for TRUNCATE TB



+Arm A is 6 months of 10mg/kg rifampicin and isoniazid, with pyrazinamide and ethambutol added for the first two months

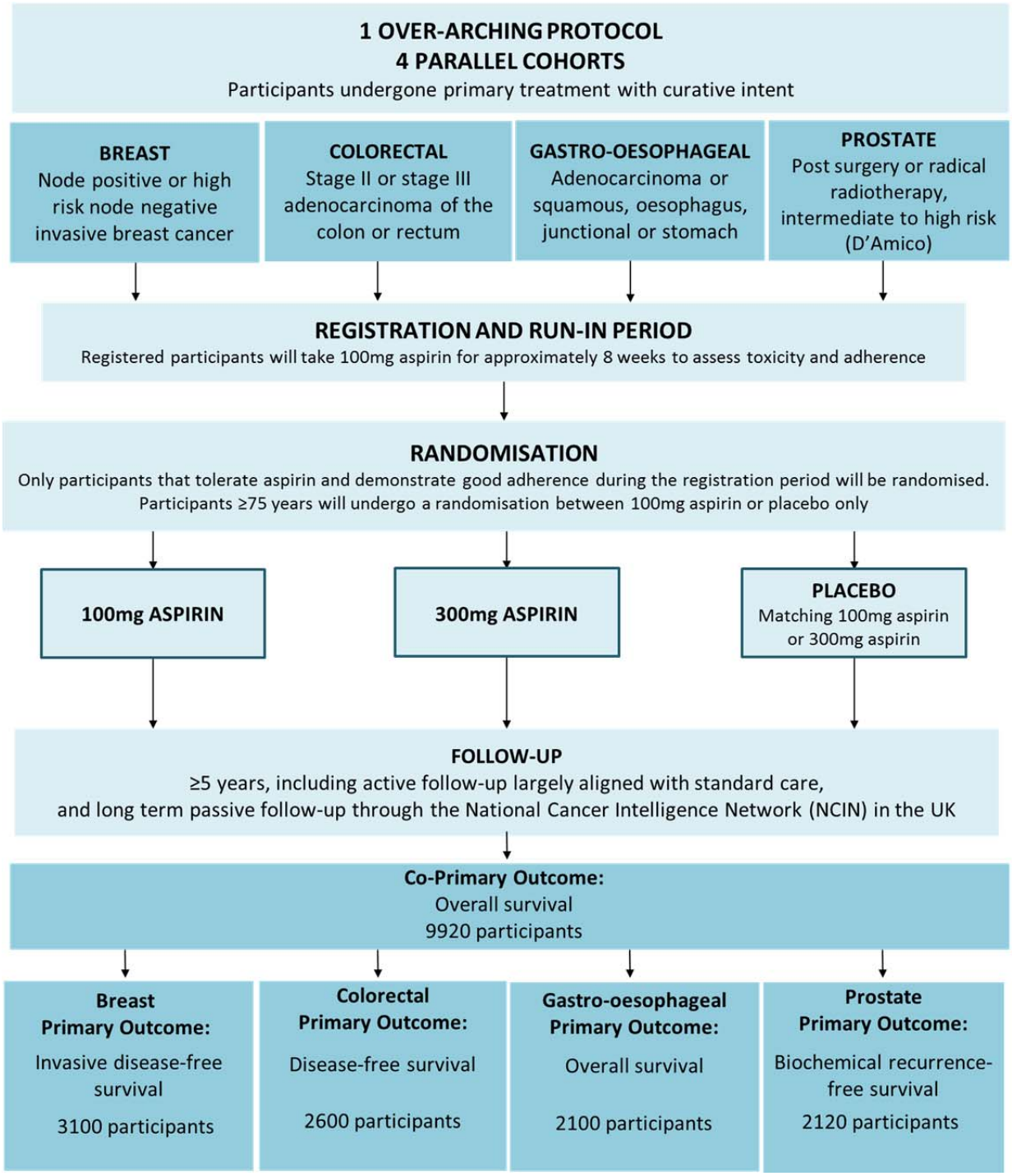
*Extended up to 3 months for missed doses or if symptomatic and positive smear at 2 months. All arms also include isoniazid and pyrazinamide, with ethambutol added to all arms apart from Arm D.

Figure 3: Protocol Schema for FOCUS4



* Patients are stratified according to a hierarchy from left to right. For example, patients with both PIK3CA and KRAS mutation would be stratified into the PIK3CA cohort.

Figure 4: Protocol Schema for Add-Aspirin



References

1. Djulbegovic B, Kumar A, Glasziou P, Miladinovic B, Chalmers I. Trial unpredictability yields predictable therapy gains. Nature. 2013; 500(7463):395-6. doi: 10.1038/500395a.
2. Berry DA. The Brave New World of clinical cancer research: Adaptive biomarker-driven trials integrating clinical practice with clinical research. Mol Oncol. 2015;9(5):951-9.
3. Liu S, Lee JJ. An overview of the design and conduct of the BATTLE trials. Chin Clin Oncol. 2015;4(3):33.
4. O'Quigley J, Conaway M. Continual Reassessment and Related Dose-Finding Designs. Statistical Science. 2010; 25(2):202-216.
5. Parmar MKB, Carpenter J, Sydes MR. More multiarm randomised trials of superiority are needed. 2014; 384(9940): 283-284.
6. Parmar MKB, Barthel FMS, Sydes MR, Langley R, Kaplan R, Eisenhauer E, Brady M, James N, Bookman MA, Swart AM, Qian W, Royston P. Speeding up the Evaluation of New Agents in Cancer. J Natl Cancer Inst. 2008; 100(17): 1204 – 1214.
7. Kaplan R, Maughan T, Crook A, Fisher D, Wilson R, Brown L, Parmar MKB. Evaluating Many Treatments and Biomarkers in Oncology: A New Design. Journal of Clinical Oncology. 2013; 31(36):4562-8. doi: 10.1200/JCO.2013.50.7905.
8. Coyle C, Cafferty FH, Rowley S, MacKenzie M, Berkman L, Gupta S, Pramesh CS, Gilbert D, Kynaston H, Cameron D; Wilson RH, Ring A, Langley RE. ADD-ASPIRIN: A phase III, double-blind, placebo controlled, randomised trial assessing the effects of aspirin on disease recurrence and survival after primary therapy in common non-metastatic solid tumours. Contemporary Clinical Trials. 2016; 51:56-64. DOI: 10.1016/j.cct.2016.10.004
9. Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J. and Jemal, A. (2015), Global cancer statistics, 2012. CA: A Cancer Journal for Clinicians, 65: 87–108. doi:10.3322/caac.21262 http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx?cancer=prostate
10. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association. 1955; 50: 1096-1121
11. Sydes MR, Parmar MK, James ND, Clarke NW, Dearnaley DP, Mason MD, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. Trials. 2009;10(1):39.
12. James, ND, Sydes, MR, Mason, MD et al. Celecoxib plus hormone therapy versus hormone therapy alone for hormone-sensitive prostate cancer: first results from the STAMPEDE multiarm, multistage, randomised controlled trial. Lancet Oncol. 2012;
13. James ND, Sydes MR, Clarke NW, Mason MD, Dearnaley DP, Spears MR, 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(10024):1163-77.

14. Vale CL, Burdett S, Rydzewska LHM, Albiges L, Clarke NW, Fisher D, Fizazi K, Gravis G, James ND, Mason MD, et al.. Addition of docetaxel or bisphosphonates to standard of care in men with localised or metastatic, hormone-sensitive prostate cancer: a systematic review and meta-analyses of aggregate data. Lancet Oncology 2016; 17(2):243–256.

15. de Bono JS, Logothetis CJ, Molina A, Fizazi K, North S, Chu L, Chi KN, Jones RJ, Goodman OB Jr, Saad F, Staffurth JN, Mainwaring P, Harland S, Flaig TW, Hutson TE, Cheng T, Patterson H, Hainsworth JD, Ryan CJ, Sternberg CN, Ellard SL, Flechon A, Saleh M, Scholz M, Efstathiou E, Zivi A, Bianchini D, Loriot Y, Chieffo N, Kheoh T, Hagg CM, Scher HI, COU-AA-301 Investigators: Abiraterone and increased survival in metastatic prostate cancer. N Engl J Med 2011, 364:1995–2005.

16. James, N D . J, de Bono S. J, Spears R. M, Clarke W. N, Mason D. Malcolm, Dearnaley P. David, et al for the STAMPEDE Investigators. Abiraterone for prostate cancer not previously treated with hormone herapy. N Engl J Med. 2017 doi: 10.1056/NEJMoa1702900 (Epub ahead of print)

17. Parker CC, Sydes MR, Mason MD, Clarke NW, Aebersold D, de Bono JS, et al. Prostate radiotherapy for men with metastatic disease: a new comparison in the Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy (STAMPEDE) trial. BJU Int. 2013;111(5):697-9.

18. Sydes MR, Parmar MK, James ND, Clarke NW, Dearnaley DP, Mason MD, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. Trials. 2009;10(1):39.

19. Sydes MR, Parmar MK, Mason MD, Clarke NW, Amos C, Anderson J, et al. Flexible trial design in practice - stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: a multi-arm multi-stage randomized controlled trial. Trials. 2012;13(1):168.

20. Bratton, D.J., Parmar, M.K.B., Phillips, P.P.J., Choodari-Oskooei, B. Type I error rates of multi-arm multi-stage clinical trials: Strong control and impact of intermediate outcomes (2016) Trials, 17 (1), art. no. 309

21. Mason MD, Clarke NW, James ND, Dearnaley DP, Spears MR, Ritchie AW, Attard G, Cross W, Jones RJ, Parker CC, Russell JM, Thalmann GN, Schiavone F, Cassoly E, Matheson D, Millman R, Rentsch CA, Barber J, Gilson C, Ibrahim A, Logue J, Lydon A, Nikapota AD, O'Sullivan JM, Porfiri E, Protheroe A, Srihari NN, Tsang D, Wagstaff J, Wallace J, Walmsley C, Parmar MK, Sydes MR; STAMPEDE Investigators. 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 (Epub ahead of print)

22. World Health Organisation. Global tuberculosis report. Geneva, 2016.

23. Papineni P, Phillips P, Lu Q, Cheung YB, Nunn A, Paton N. TRUNCATE-TB: an innovative trial design for drug-sensitive tuberculosis. International Journal of Infectious Diseases. 2016; 45(S1): 404.

24. Renfro LA, Mallick H, An MW, Sargent DJ, Mandrekar SJ. Clinical trial designs incorporating predictive biomarkers. Cancer treatment reviews. 43:74-82

25. Langley RE, Burdett S, Tierney JF, Cafferty F, Parmar MK, Venning G. Aspirin and cancer: has aspirin been overlooked as an adjuvant therapy? Br J Cancer. 2011;105(8):1107-13.

26. Rothwell PM, Wilson M, Price JF, Belch JFF, Meade TW, Mehta Z. Effect of daily aspirin on risk of cancer metastasis: a study of incident cancers during randomised controlled trials. Lancet. 2012; 379:1591–1601

27. Bratton DJ, Choodari-Oskooei B, Royston P. A menu-driven facility for sample size calculation in multi-arm multi-stage randomised controlled trials with time-to-event outcomes: update. Stata J. 2015;15(2):350–68.

28. Korn EL, Freidlin B. Outcome adaptive randomisation. Is it useful? Journal of Clinical Oncology. 2011; 29: 771-776.