Sample size reporting in protocols: a methodological review of randomised adaptive clinical trials

Zihang Yu

Supervisors: Qiang Zhang and Dr Munya Dimairo

1. Introduction

Design is the most crucial element in any clinical trial, and sample size calculation is a central part of study design (Julious 2009). Inappropriate sample size may cause negative consequences. For example, if the sample size is underestimated, this will lead to an underpowered trial which is unable to address research questions definitively. Huge uncertainties will remain and need to be addressed through more trials. On the other hand, an overestimated sample size may waste many resources, and drug development may be delayed to benefit more patients and the public (Yin 2012). Most importantly, either an excessive or insufficient sample size may expose participants to the potential risk of ethical issues (Altman 1980).

In recent studies, adaptive designs are increasingly considered by investigators when designing clinical trials due to their flexibility (Dimairo, Coates et al. 2018). Compared with a fixed design, trials with an adaptive design are generally more ethical, effective and informative because they usually make better use of resources such as time and money, and may even need fewer participants (Pallmann, Bedding et al. 2018). There are some well-known types of adaptive design, such as group sequential design, adaptive sample size re-estimation design, multi-arm multi-stage design, population enrichment design, adaptive seamless design and so on. Based on the accumulating data, these designs provide controlled chances to modify trial aspects while they are still being conducted while maintaining the validity and integrity of the trial outcomes (Wason, Dimairo et al. 2022). For example, the features of trial adaptation may include early trial stopping, sample size re-estimation, dropping off the futile arm, and selecting the promising arm.

Although adaptive designs are appealing, they bring additional considerations. The sample size calculation for adaptive designs may rely on the trial adaptation so it can be complex. For example, sometimes the actual sample size is unknown at the start of a trial, or it could be modified based on the interim data. Therefore, it brings a challenge to the communication of sample size in adaptive clinical trials. Reporting the sample size improperly may eclipse the benefits of an adaptive design.

Guidelines on essential information to include when reporting sample size calculations in randomised adaptive clinical trials in publications such as journals exist (Dimairo, Pallmann et al. 2020). However, several factors are needed to be considered when determining and

communicating the sample size for randomised adaptive trials to stakeholders at the planning stage in protocols. Therefore, this internship project will review randomly selected protocols for adaptive clinical trials registered on ClinicalTrials.gov to:

1) Characterise the adaptive trials,

2) Understand the regularly occurring patterns for the determination of sample size in protocols of adaptive trials,

3) Understand how sample size is currently communicated in protocols of adaptive trials,

4) Explore and comment on any research gaps in sample size reporting in protocols of adaptive trials.

2. Methods

This work was part of a broad methodological review of randomised trials with adaptive designs in study documents such as protocols and grant applications. This review was undertaken following adapted principles from the PRISMA guidance (Page, Moher et al. 2021) to address the objectives stated in Section 1. This work was restricted to a nested methodological review of randomly selected trial protocols as described in Section 2.1.

2.1 Data sources

This study reviewed adaptive trials registered on the ClinicalTrials.gov trial registry which is one of the most extensive sources of information for clinical trials across the world operated by both industry and the public sector. As shown on the website of ClinicalTrials.gov, the registry contains information on over 425,000 trials across trial phases.

A database that includes clinical trials using an adaptive design was compiled by another researcher (QZ), and he will review all the trials in the database while this study will randomly review 10% of them. Results from two independently reviewing researchers will be compared.

2.2 Criteria for selection

The inclusion and exclusion criteria for this review are listed in Table 1. In summary, randomised trials comparing at least two treatments and using an adaptive design as indicated by the researchers or there is evidence of at least one of the trial adaptations documented in the protocol were included. Included trials were posted between 2010/01/01 and 2022/01/01 to capture more recent research practice.

Table 1. Inclusion and exclusion criteria for study selection

Inclusion criteria
1. Phase II or phase III (include phase II/III seamless) or phase IV (include phase III/IV and phase II/III/IV) clinical trial
2. Randomised trial
3. An adaptive design with at least one or more specific types of adaptations as defined in Table 2
4. Posted between 2010/01/01 and 2022/01/01
5. Designed using either frequentist or Bayesian methods
Exclusion criteria
1. Non-randomised trial (EX1)
2. Study protocol not written in English language (EX2)
3. Not an adaptive design although a terminology on adaptive design was used to describe the trial (EX3)
4. Features for adaptations or sample size estimation part remain unavailable (EX4)

2.3 Data collection

A data extraction form was built based on the study objective using excel, then the key information from the selected trials were extracted in this form. The terms that are closely related to sample size estimation will be analysed and discussed in this paper.

The details of the extraction terms are listed in Table 2.

Table 2. Extraction terms related to sample size

No.	Category	Extraction terms
1	Registered number	The unique registration code of a clinical trial on ClinicalTrials.gov
2	Type of the sponsor	Industry, public sector
3	Geographic location of the study sites.	Asia, Africa, Europe/North America, South America, globe
4	Year	Year of first posed, year of trial started
5	Research phase	Phase II, phase II/III, phase III, phase III/IV, phase IV, phase II/III/IV
6	Number of starting treatment arms	positive integer
7	Type of intervention	Drug, device, surgical treatment, other
8	Type of control	Active, placebo, blank, standard of care

9	Classification of primary endpoint(s)	Binary, continuous, time-to-event, ordinal, etc
10	Type of original primary hypothesis test(s)	Superiority, non-inferiority, equivalence, etc
11	Statistical framework used to design the trial	Bayesian, frequentist, mixed
12	Statistical framework used for interim analysis	Bayesian, frequentist, mixed
13	Nature of trial adaptations	Free text
14	Early stopping category	Futility, efficacy, efficacy and futility, non-inferiority
15	Adaptive types	Group sequential design, adaptive two stage design, adaptive treatment selection, sample size re-estimation design, multi-arm multi-stage design, adaptive platform design, adaptive randomization design, adaptive population enrichment design, adaptive hypothesis design, adaptive basket design, adaptive seamless design, multiple adaptive design
16	Statistical approach to sample size estimation	Analytical, simulation, unclear
17	Sample size calculation	Parameters for sample size calculated well stated (yes, no, partially), Justification for sample size parameters described (yes, no, partially), Sample size details adequately described to allow interpretation and reproducibility (yes, no), Statistical software stated (yes, no)
18	Sample size reporting	Sample size for a fixed design, maximum sample size, expected sample size
19	Operating characteristics	Nominal type I error rate, familywise type I error, nominal power, global power, marginal power, decision making probabilities under certain scenarios
20	Sample size re-estimation blind/unblind	Blinded, unblinded
21	Sample size re-estimated comparative/ non-comparative	Comparative, non-comparative
22	Sample size re-estimation based on	Nuisance parameters, interim treatment effect, both

2.4 Screening of trials for eligibility and quality control

The database of adaptive clinical trials was compiled by QZ using a user-written automated Python toolkit which ensured the comprehensive of the searching since the ClinicalTrial.gov

platform did not provide a whole text searching strategy. The trials reviewed in this study were simply randomly selected from the database compiled by QZ, and then the selected trials were screened for eligibility using the criteria stated in Table 1. Sixty-seven extraction terms are considered in the review process. However, only the terms that are closely related to sample size estimation are analysed and discussed in this paper.

After the review work, the result was compared with that of the other independently working reviewer (QZ). Simple observational errors were fixed. Other disagreements were discussed with Dr Munya Dimairo and reached to an agreement between QZ and the author.

2.5 Statistical analysis methods

After screening of trials for eligibility and quality control, eligible trials were characterised using descriptive statistics. There are no continuous variables analysed in this study. Categorical variables were summarised using numbers and percentages. Four aspects related to the objectives of this study were included in the analysis. The four aspects are basic characteristics of included adaptive trials, characteristics of trials adaptations, characteristics of sample size calculation and the state of reporting sample size calculations. Some of the results presented were stratified by sponsor and type of adaptive design. Data visualisation and data analysis methods such as using pie chart, histogram and tables, and the charts are generated using Excel.

3. Results

This section details results starting with eligibility screening of trials followed by characteristics of eligible trials and their adaptive features. Characteristics of sample size calculations and their reporting are described. Finally, an exemplar of good practice in reporting sample size and adaptive features is highlighted.

3.1 Eligibility screening

Figure 1 is an overview of the screening process including the number of trials that were excluded for reasons as shown in Table 1. Of the 37 randomly selected trials, 10 trials were excluded, and the remaining 27 trials were eligible for inclusion in the analysis.

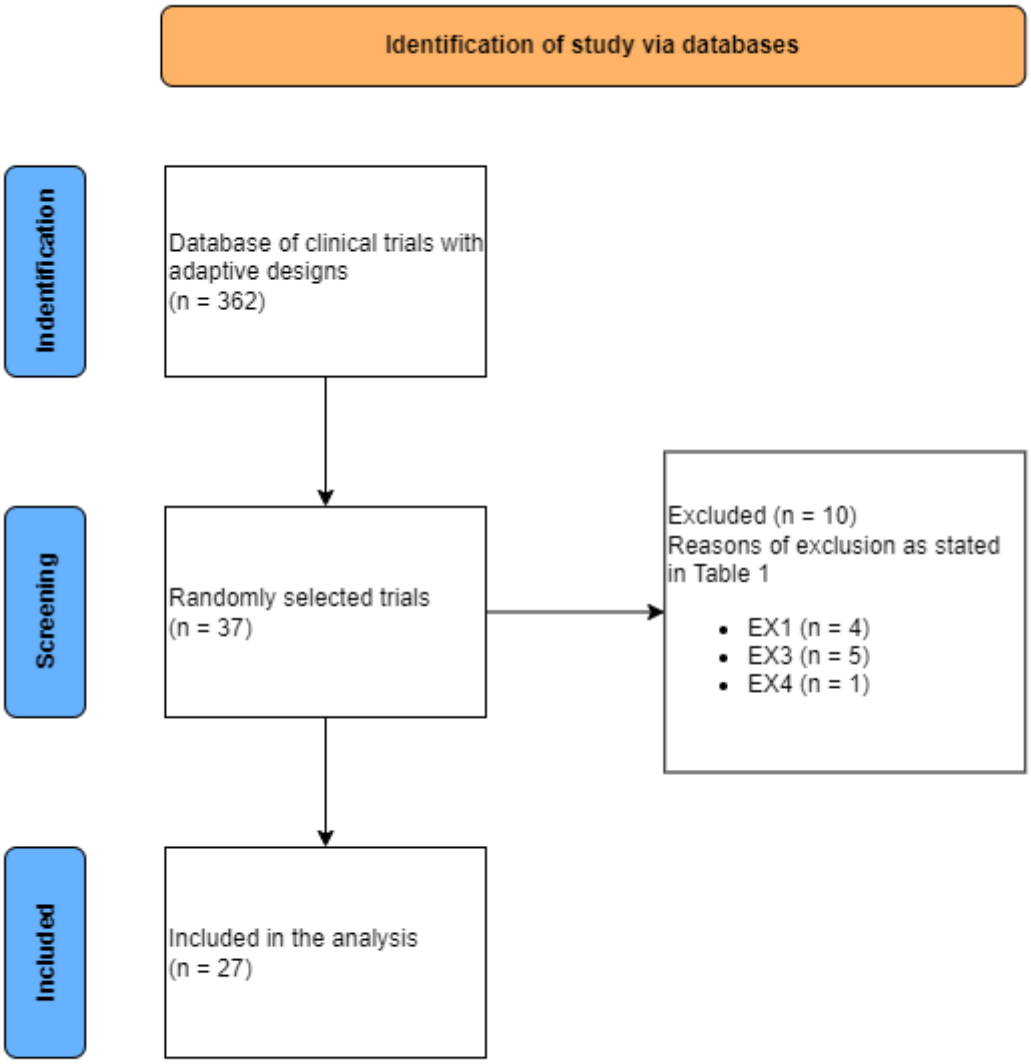

Figure1. Flow chart of the review process

3.2 Basic characteristics of included adaptive trials

Table 3 shows the basic characteristics of the eligible adaptive trials. Of these 27 trials, 21 (78%) were industry-funded trials, and 18 (67%) trials were phase III trials. All the public-funded trials are conducted in Europe and North America. The number of industry-funded trials conducted globally was 13 (48%), slightly more than that in Europe and North America which is 8 (30%). The most common type of intervention and original primary hypothesis test is drug 25 (93%) and superiority 24 (89%) respectively. Time-to-event is the most common type of primary endpoint, which is 15 (56%).

Table 3. Basic characteristics of included adaptive trials

Characteristics	Category	All trials (n=27)	
		Industry funded (n=21)	Public funded (n=6)
Phase	Phase 2	1 (5%)	1 (17%)
	Phase 3	14 (67%)	4 (67%)
	Phase 3/4	1 (5%)	0
	Phase 4	1 (5%)	1 (17%)
Location	Europe/North America	8 (38%)	6 (100%)
	Globe	13 (62%)	0
Start year	2010-2016	11 (52%)	4 (67%)
	2017-2022	10 (48%)	2 (33%)
Type of intervention	Drug	19 (90%)	6 (100%)
	Device	2 (10%)	0
Type of control	Active	13 (62%)	4 (67%)
	Placebo	7 (33%)	2 (33%)
	Both	1 (5%)	0
Type of original primary hypothesis test(s)	Superiority	18 (86%)	6 (100%)
	Non-inferiority	3 (14%)	0
Classification of Primary endpoint(s)	Time-to-event	12 (57%)	3 (50%)
	Binary	6 (29%)	2 (33%)
	Continuous	2 (10%)	1 (17%)
	Continuous and Binary	1 (5%)	0

3.3 Characteristics of trial adaptations

This section details characteristics of adaptive features.

3.3.1 Type of adaptive design

Figure 2 displays the types of adaptive design of eligible trials. Of the 27 trials, the majority 20 (74.1%) used group sequential designs and 17 (85.0%) of them were industry-funded. Moreover, for the 5 multiple adaptive designs which accounted for 18.5% of the trials, all of them used a combination of a group sequential design and sample size re-estimation.

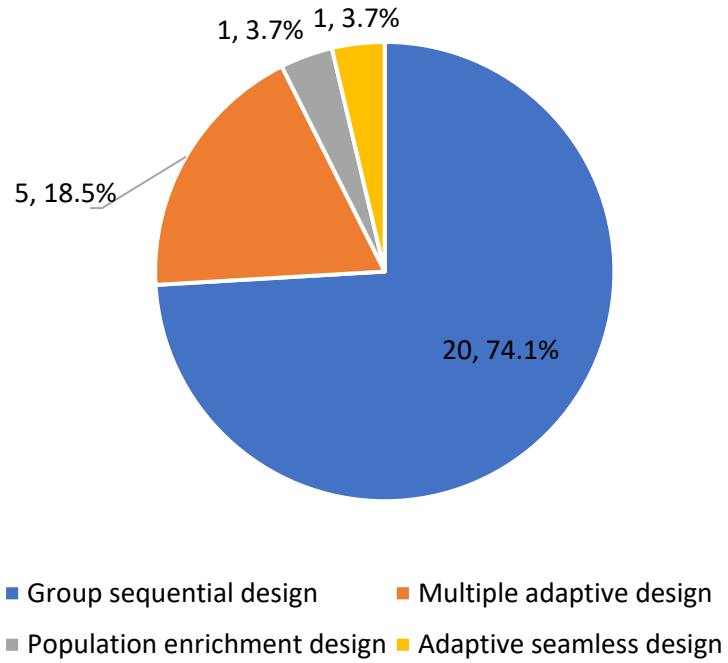

Figure 2. Type of adaptive design used

3.3.2 More characteristics of trial adaptations

Table 4 details more characteristics of trial adaptations in detail. All the 27 trials included at least an early stopping option. Eleven (40.7%) of them are planned to stop early for efficacy, 4 (14.8%) for futility, and 12 (44.4%) for either efficacy or futility. The majority 17 (63.0%) of trials had one planned interim analysis while 9 (33.3%) have two planned interim analyses.

Table 4. More characteristics of trial adaptations

More characteristics	Category	All trials (n=27)
Nature of adaptation rules considered	early trial stopping option only	21 (77.8%)
	early trial stopping option and sample size re-estimation	5 (18.5%)
	early trial stopping option and population enrichment	1 (3.7%)
Type of early stopping category	stop for efficacy	11 (40.7%)
	stop for futility	4 (14.8%)
	stop for efficacy and futility	12 (44.4%)
Was the futility boundary binding or non-binding	Non-binding	9 (33.3%)
	Unclear/not stated	6 (22.2%)
	Not applicable	12 (44.5%)
Number of interim analyses	1	18 (66.7%)
	2	9 (33.3%)

3.4 Characteristics of sample size calculation

Table 5 indicates the number of characteristics of sample size calculation including the parameters, justification, reproducibility and statistical software for sample size calculation. The result is stratified by the type of adaptive design and sponsor.

Table 5. Characteristics of sample size calculation

Sample size calculation	GSD (n = 20)		MAD (n = 5)		APED (n = 1)		ASD (n = 1)		All trials (n= 27)
	Industry (n = 17)	Public (n = 3)	Industry (n = 3)	Public (n = 2)	Industry (n = 1)	Public (n = 0)	Industry (n = 0)	Public (n = 1)	
Parameters stated (yes)	17 (100%)	3 (100%)	3 (100%)	2 (100%)	0	NA	NA	1 (100%)	26 (96.3%)
Parameters justified (yes)	14 (82.4%)	3 (100%)	3 (100%)	1 (50.0%)	0	NA	NA	1 (100%)	22 (81.5%)
Reproducible (yes)	17 (100%)	3 (100%)	3 (100%)	1 (50.0%)	0	NA	NA	1 (100%)	25 (92.6%)
Software stated (yes)	8 (47.1%)	1 (33.3%)	2 (66.7%)	2 (100%)	0	NA	NA	1 (100%)	14 (51.9%)

GSD: Group sequential design, *MAD:* Multiple adaptive design, A*PED:* Adaptive population enrichment design, *ASD:* Adaptive seamless design

Figure 3 presents the number of characteristics of sample size calculation among all types of sectors and adaptive designs. Almost all the trials reported the parameters for sample size calculation except for one which is the population enrichment design funded by the industry sector. The report of justification and reproducibility are generally good, they are 22 (81.5%) and 25 (92.6%) respectively among all 27 trials. Just over half 14 (51.9%) of the trials disclosed statistical software used for sample size calculation and designing the trial.

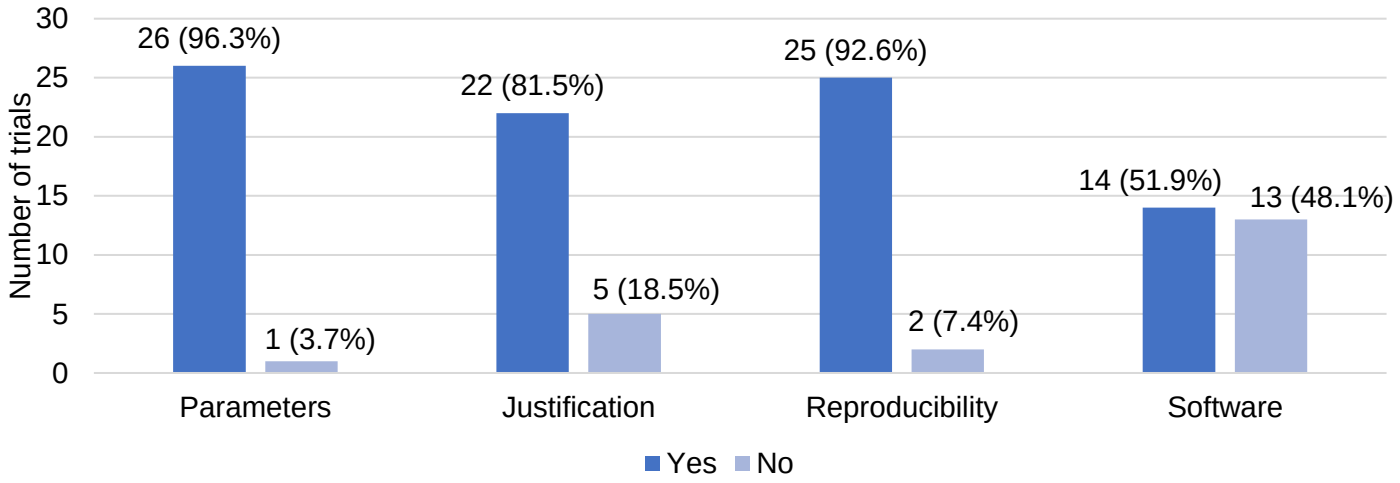

Figure 3. Characteristics of sample size calculation

3.5 The state of reporting sample size calculations

Table 6. indicates the state of reporting sample size calculations. The maximum sample size was reported in most of the trials, 25 (92.6%). There are 5 (83.3%) public-funded trials reported the maximum sample size in their protocols and 16 (76.2%) in industry-funded trials. No trial reported the expected sample size.

Table 6. The state of reporting sample size calculations

Sample size reporting patterns	Public funded (n = 6)	Industry funded (n = 21)	All trials (n=27)
Maximum sample size reported	5 (83.3%)	16 (76.2%)	21 (77.8%)
Sample size for fixed design	0	2 (9.5%)	2 (7.4%)
Sample size for fixed design and maximum reported	1 (16.7%)	3 (14.3%)	4 (14.8%)

3.6 Number of other operating characteristics reported

Table 7 presents the number of other operating characteristics reported for the sample size calculation stratified by the type of adaptive designs. All the included trials clearly stated the overall type I error and the nominal power in the protocol. Familywise type I error and global power is only relevant for adaptive designs with multiple hypothesis testing.

Table 7. Number of other operating characteristics reported

All trials (n = 27)	group sequential design (n=20)	multiple adaptive design (n=5)	population enrichment design (n=1)	adaptive seamless design (n=1)	total
NOTPI	20 (100%)	5 (100%)	1 (100%)	1 (100%)	27
FAMTPI	9 (45.0%)	0	0	0	9
NOMPO	20 (100%)	5 (100%)	1 (100%)	1 (100%)	27
GLOPO	11 (55.0%)	0	0	0	11
MARPO	3 (15.0%)	0	0	0	3
DECPR	2 (10.0%)	0	0	0	2

NOTPI: Nominal type I error rate, *FAMTPI:* familywise type I error, *NOMPO:* Nominal power, *GLOPO:* global power, *MARPO:* marginal power, *DECPR:* Decision making probabilities under certain scenarios

3.7 Exemplars for good reporting of sample size and adaptive features

The trial with registered number NCT03993288 is a good exemplar for the reporting of sample size and adaptive features. In general, the report of sample size calculation in the protocol of this trial is well structured and comprehensive. The source of statistical software and packages

used in the sample size calculation is presented. The parameters of sample size calculation are clearly stated to allow reproducibility. The sample size calculation is properly justified, and the published works on which sample size calculation is based are referenced. The minimum sample size is reported, and the state of reporting is clear. The percentage of total sample size for interim analysis and the alpha-spending function for interim analysis are reported.

4. Discussions

From these preliminary randomly selected adaptive trials with accessible protocols, the group sequential design was the most common type of adaptive design. Although the software for sample size calculation was sometimes not clearly stated, the parameters and justifications of the calculated sample size were generally well presented across sectors and types of adaptive designs. Moreover, the sample size calculation process usually allows reproducibility. For the state of reporting sample size calculation, the most reported value was the maximum sample size and most of these trials were funded by the industry.

4.1 Reflection on the results

The result shows that the group sequential design is the most wildly used type of adaptive design. A possible reason might be it has a strong statistical foundation and is regarded as being well understood by regulators, and trial statisticians might be more likely to use designs they are familiar with (Hatfield, Allison et al. 2016).

The most common pattern to report sample size is the maximum sample size. This might be because of the following considerations. Firstly, take group sequential designs as an example, although the early stopping option may lead to a lower sample size than required for a fixed design, but it could still be inflated to account for repeated hypothesis testing if the trial failed to stop early. To achieve the appropriate power, group sequential methods do not allow the maximum sample size to be changed mid-trial (Yin 2012) but assumes the assumptions made about design parameters are correct. Additionally, the ideal sample size is enough for all important outcomes and making sure the sample size is adequate for more than one intended analysis may be appropriate in some circumstances (Cook, Julious et al. 2018). Above all, reporting the maximum sample size can reduce the uncertainty of sample size in adaptive designs and avoid issues caused by insufficient sample size, therefore it is more likely that the sponsors to fund the trials in which the maximum sample size is reported.

Most of the trials reporting the maximum sample size were funded by the industry. The first reason might be there are more industry-funded trials in this review than public-funded ones. The resulting bias can be reduced by stratified sampling which can pick the same amount of

industry-funded and public-funded trials. Another reason is that public sectors have more consideration for cost and the time spent on the trials. Funding-related issues are more specific to the public sector since personnel are usually hired for a particular study (Pallmann, Bedding et al. 2018). For industry sectors that have sufficient funds, their mean concern might be the time cost of re-recruiting participants and delays in the new drugs coming into the market, therefore an adequate sample size is their preference.

4.2 Suggestions

Based on the result shown in Table 5, the report of other operating characteristics for sample size calculation in protocols should be improved. A sample size calculation given only the power and significance level is meaningless, it is suggested that any underlying assumptions for sample size calculation, such as type I error, power, event rate in the control group, and treatment effect should be presented to readers in the study documents (Schulz and Grimes 2005). The sample size calculation features such as justification and reproducibility are satisfying as shown in the result, but there is space for improvement, especially for the reference of statistical software used in sample size determination. For readers to be able to repeat the sample size calculation, it is important to adequately justify the information that drove the determination of the sample size. It is also precise research practice to cite any software, packages, or code used in the work (Dimairo, Pallmann et al. 2020). Since public and industry sectors may have different concerns on the budget issues and this may be reflected in the state of sample size reporting in protocols. Therefore, future research needs to focus more on the sample size calculation and budget considerations (Bell 2018).

4.3 Limitations and challenges

There are some limitations to this study. First, except for protocols, the review materials of some trials in the database were based on grant applications. However, none of the trials based on grant applications was selected in the simple random selecting process, the information of them was missing in this study. Second, this study only reviewed and analysed 10% of the samples in the database. Since this is just a small amount in the database, there could be bias in the results. However, this bias could be reduced in QZ's work since he will review all 362 trials in the database. It is also notable that the sample size estimation in registered trials in the database tends to be well reported, and there may be other adaptive trials which are poorly reported and not included in the database. Finally, the features for adaptations or sample size estimation part remain unavailable in certain trials so their data are lost in the results. While some of the information in the protocols may not be publicly disclosed for confidentiality reasons, there is no reason why the sample size estimates should be unavailable.

There are three main challenges encountered in this review. The first one is sometimes a sample size is reported without clearly claiming whether it is the minimum, expected or maximum sample size in most of the protocols reviewed in this work. The second one is the terminology of other operating characteristics could be divergent. For example, whether the nominal type I error is the significance level to be achieved across the whole trials or the significance level at each interim analysis. The third one is the features for adaptations or sample size estimation part remain unavailable for certain trials.

5. Conclusions

This paper explores and comments on the issues related to sample size calculation and communication in protocols of adaptive clinical trials. In general, the reporting of sample size needs to be extended when considering trial feasibility and more research is needed. The findings of this study could be used to understand the current patterns of sample size calculation and communication in clinical trial practice. For funders, this study may help them to make better investment decisions. For trial statisticians, this study may help them to form a clear framework of sample size calculating and reporting in protocols. For QZ's work in the future, the challenges and limitations in this study may be used to improve his study. The limitation of this study will be reduced to a large extent, and the actual result will be shown in QZ's work.

6. Acknowledgement

This research internship is funded by the Wellcome Trust Biomedical Vacation Scholarships. I would like to express my gratitude to my supervisor, Qiang Zhang, who guided me throughout this project. I would also like to thank Dr. Munya Dimairo who supported me and offered deep insight into the study. Many thanks to the medical statistics group from the Department of Health and Related Research at the University of Sheffield.

References

Altman, D. G. (1980). "Statistics and ethics in medical research: III How large a sample?" BMJ **281**(6251): 1336-1338.

Bell, M. L. (2018). "New guidance to improve sample size calculations for trials: eliciting the target difference." Trials **19**(1): 605.

Cook, J. A., et al. (2018). "DELTA(2) guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial." BMJ **363**: k3750.

Dimairo, M., et al. (2018). "Development process of a consensus-driven CONSORT extension for randomised trials using an adaptive design." BMC Medicine **16**(1).

Dimairo, M., et al. (2020). "The Adaptive designs CONSORT Extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design." BMJ: m115.

Hatfield, I., et al. (2016). "Adaptive designs undertaken in clinical research: a review of registered clinical trials." Trials **17**(1).

Julious, S. A. (2009). Sample sizes for clinical trials, chapman and hall/CRC.

Page, M. J., et al. (2021). "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews." BMJ: n160.

Pallmann, P., et al. (2018). "Adaptive designs in clinical trials: why use them, and how to run and report them." BMC Medicine **16**(1).

Schulz, K. F. and D. A. Grimes (2005). "Sample size calculations in randomised trials: mandatory and mystical." The Lancet **365**(9467): 1348-1353.

Wason, J. M. S., et al. (2022). "Practical guidance for planning resources required to support publicly-funded adaptive clinical trials." BMC Medicine **20**(1).

Yin, G. (2012). Clinical trial design: Bayesian and frequentist adaptive methods, John Wiley & Sons.