TITLE:

What design and analysis features are used in small population and rare disease trials: A targeted review

Giles Partington[a], Suzie Cro[a], Alexina Mason[b], Rachel Phillips[a], Victoria Cornelius[a]

[a] Imperial Clinical Trials Unit, Imperial College London, 1st Floor Stadium House, 68 Wood Lane, London, United Kingdom W12 7RH

[b] Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, Keppel Street, London, United Kingdom WC1E 7HT

ABSTRACT

Objective

Frequentist trials in Rare disease/small population trials often require unfeasibly large sample size to detect minimum clinically important differences. A targeted review was performed investigating what design and analysis methods these trials use when facing restricted recruitment.

Study Design and Setting

Targeted Review searching EMBASE and MEDLINE for Phase II-IV RCTs reporting ‘rare’ disease or ‘small population’ within title or abstract, since 2009.

Results

A total of 6,128 articles were screened with 64 trials eligible (4 Bayesian, 60 frequentist trials). Frequentists trials had planned power ranging 72-90% (median: 80%) but reported recruiting a mean of 6.6% below the planned sample size (n=38) [median 0%, IQR (-5%, 5%)], most used standard type 1 error (52 used 5% and 1 used 1%), and the average standardised effect was high (0.7) with 50% missing their assumed level. Of the 4 Bayesian trials, 3 used informed priors, 2 and 1 trials performed sensitivity analysis for the impact of priors on design and analysis respectively. Historical data, expert consensus, or both were used to construct informative priors. Bayesian trials required 30%-2400% less participants than using frequentist frameworks.

Conclusion

Bayesian trials required lower sample size through use of informative priors. Most frequentists didn’t achieve their target sample size. Bayesian methods offer promising solutions for such trials but are underutilised.

Word count: 197

Key words: Rare Diseases; Small Populations; Bayesian Methods; Randomised trials; Trial Design

What is new?

- Few small population/rare disease trials are using Bayesian methods and those that don't are ending up under sized and underpowered.
- Researchers in rare disease and small population trials should be striving to use Bayesian methods where possible or follow recommendations suggested by Parmar et al. (2016)

1. Background

Rare disease trials face design challenges due to highly restricted recruitment. Participants to trials may also be limited if the focus is for a highly specialised sub-group; for example, a subset of asthmatic patients that do not respond to established therapeutic strategies **[REF1]** or trials where intervention was prohibitive, such as with expensive treatments or surgeries. With traditional parallel arm randomised controlled trials conducted in a frequentist framework, the required sample size is calculated through: desired power, significance level (type I error), and the minimal clinically important difference (MCID). When the target population size is restricted the resulting sample size will often be larger than achievable. As type I error is typically fixed at an accepted threshold of 0.05 and the MCID should be treated as a fixed quantity, the consequence for rare disease and small population trials with small sample sizes is that they will likely have to lower their power to be under the recommended 80%. Sample size also depends on between participant variation in parallel arm trials.[REF2] Different trial designs can be used to overcome between participant variation such cross-over or n-of-1 trials, which enable participants to act as their own controls. [REF3]

The International Rare Diseases Research Consortium (IRDRC)**[REF4]**, European Medicines Association (EMA)**[REF5]** and Federal Drugs Agency (FDA)**[REF6]** have all produced guidelines for how small population and rare disease trials should be conducted. They suggest less simple trial designs such as crossover and adaptive designs as well as using historical data through Bayesian methods.

Parmar et al (2016)**[REF7]** introduced a framework for the design of smaller population trials covering different areas including: recruiting as many participants as possible (e.g. broaden eligibility criteria, multicentre international trials, lengthening timelines), increasing data from fewer participants (e.g. rerandomization, crossover/sequential or n-of-1 trials), ensuring information rich outcomes (e.g. continuous outcomes rather than binary, sensible differences between trial arms, stratification), or including available information from other sources through Bayesian methods.

Bayesian methods can combine existing information, incorporated through informed prior distributions on model parameters, with trial data to strengthen the evidence on the treatment effect of interest. Bayesian trials can provide a solution for frequentist sample size calculation conundrums when there is restricted recruitment. Instead of being bound to a hypothesis testing framework which requires a set high power and low significance level, Bayesian trials result in posterior distributions which can be used to calculate posterior probabilities for specific outcomes providing clinicians with useful information to guide treatment decisions.[REF8] Despite these benefits, Bayesian methods can be overlooked due to the acceptability of including subjective elements and the complexities of designing and conducting Bayesian trials. The use of a prior cannot be avoided and can lead to concerns as this is based on judgement and hence introduces a degree of subjectivity into results, beyond the observed trial data.[REF9] For some disease areas there may be a lack of information which will inhibit development of an informative prior, negating the advantages of the Bayesian approach. Bayesian methods are more computationally intensive, often requiring simulation through Markov Chain Monte Carlo methods to obtain estimates of the posterior distribution. However, in recent years the accessibility of Bayesian methods in software programs and computational power has vastly improved.

This review aims to investigate what trial design and analysis methods small population and rare disease trials use when faced with restricted recruitment. We were specifically interested to examine how often Bayesian methods were used, and when used, how well Bayesian approaches and priors were reported.

2. Methods

2.1. Literature Search

Articles were searched for in EMBASE and MEDLINE from 2009 to 2020, as the IRDRC started in 2009;[REF10] representing the start of better sharing of methods for rare disease trials across the research community. We searched for articles which self-referred as rare disease or small population trials using the terms [randomized controlled trial] AND ([rare/uncommon

disease/condition] OR [low incidence] OR [small population]). The search term [randomized controlled trial] included “random*”, “control*”, “assign*” and “allocat*” terms. A full search strategy can be found in Appendix 1.

2.2. Inclusion criteria

Any completed or in progress (published protocols) phase II-IV Randomised Controlled Trial (RCT) with clinical endpoints published since 2009 that self-referred as small population or rare disease trials within their title or abstract was considered eligible. Published conference abstracts were included if an associated published trial could be identified. There were no restrictions on disease or intervention.

All published materials associated with the primary article were sought, including protocols, Bayesian elicitation protocols, and statistical analysis plans.

Exclusion criteria were research letters, pilot studies, feasibility studies, review articles, exploratory analyses, phase I studies and trials with non-clinical primary outcomes. Studies where no English translation was available were also excluded. In instances where no protocol was available online, a single attempt was made to contact the first author.

2.3. Eligibility Screening

Search results were exported into Covidence[REF11], where duplicates were removed. Titles and abstracts were screened by one reviewer (G.P.) to determine eligibility with a random 10% sample check by a second reviewer (S.C. & V.C.). Full texts of the shortlisted trials were reviewed to confirm their eligibility by the primary reviewer, with a random 10% checked by a second reviewer.

2.4. Data extraction

A standardised data extraction form was developed (Appendix 2), collecting information on first author, publication year and journal, and characteristics of each trial, such as population, clinical area, blinding and study design. This extraction form was piloted with all Bayesian papers and a random sample of 10% of frequentist papers to ensure it collected all necessary

information. Other extracted data included, where relevant: type I and II error, assumed proportional and absolute difference (for binary outcomes), and assumed effect sizes (for continuous outcomes) used to calculate sample size (i.e. MCIDs); information on priors was collected for Bayesian trials. All data was extracted by GP and double data extraction was performed for all papers by a second reviewer (V.C., S.C. or R.P.). Any differences between extracts were aligned through discussion between reviewers.

2.5. Outcomes

The primary outcome was the proportion of trials using Bayesian approaches in design or analysis. Secondary outcomes were separated for the different types of trials. For Bayesian trials: the proportion using informed priors and methods used to elicit them; the proportion of trials conducting sensitivity analysis to assess the impact of the prior distribution in the design; and the proportion of trials proposing sensitivity analysis to assess the impact of the prior distribution in the analysis. For frequentist trials: the average power and type I error ; and the average assumed and actual standardised effect for the trials primary outcome (collected from the sample size calculation and results).

2.6. Statistical Analysis

Outcomes were summarised descriptively using frequencies and percentages for categorical data, with means and standard deviations or median and interquartile range for continuous data. Standardised effect sizes were calculated by dividing means by standard deviations if not expressly given. Where possible, both assumed and actual standardised effect sizes were calculated based on sample size and reported results respectively. The percentage of randomised participants who dropped out of the study before analysis was also calculated. Analysis was done through Stata16.[REF12]

3. Results

3.1. Search Results

As shown in Fig1, the search identified 7141 records, 1279 from Medline and 5862 from Embase. Of these 1013 were identified as duplicates. The remaining 6128 were assessed on their abstracts and titles for eligibility, 5605 were ineligible. Of the resultant 523 papers, 459 were excluded after full text reviews, the main reason being the trials were not rare disease or small population (n=378).

3.2. Trial Characteristics

Of the 64 eligible trial articles, most evaluated drug interventions (57, 89%). Four (6%) articles reported use of Bayesian designs or analysis, all of which were drug intervention trials.

Table1 shows that most trials used parallel arms, with 80% of frequentist (48) and 75% of Bayesian (3) using this design. Two (3%) frequentist and one (25%) Bayesian trials used n-of-1 designs. N-of-1 trials have been proposed to find optimal intervention for individuals, and involve randomising the sequence of interventions a participant is given.[REF3] This design was recommended by both the IRDRC and the EMA for rare disease trials. Seven (12%) frequentist trials used crossover designs, whilst one frequentist trial used a 2-stage Simon's within their parallel arm trial design, an adaptive method allowing for early stopping criteria should a strong enough result be reached for either success or futility, potentially reducing the number of recruited participants. Most primary outcomes were continuous (59%), with only 20% binary and 19% with time-to-event outcomes.

3.3. Frequentist trials

Table2 showed that most trials were multicentre (77%) often international. Type 1 errors infrequently went above the 5% level (8% of trials at 10% type 1) and most confidence intervals were two sided (82%). Many trials could have considered whether their entry criteria were more restrictive than necessary, with 39% of trials restricting potential participants via eligibility criteria restricting more than age and disease of interest. Few trials accounted for covariate information in their sample size calculations which can reduce the required sample size, the 5 that did (8%) accounted for the correlation of baseline and outcome measure. No trials were

designed to rerandomize participants. Rerandomization can only occur for disease trials where enrolment in one arm did not prevent enrolment into another; or where participants did not reach trial endpoints in non-chronic diseases.[REF13]

Only 77% of trials reported power calculations; most used 80% power and above, 5% of trials used power below 80% (between 77% and 72%). The average assumed standardised effect for frequentist trials was 0.7 (SD 0.9), denoting a moderate to large effect size.**[REF14]** Fig2 compares the assumed and actual standardised effect for each trial with a line of equality. This shows actual observed standardised effects often did not reach assumed levels, especially with larger assumed standardised effects. Trials often struggle to achieve planned recruitment targets.**[REF15]** Of the 38 (66%) trials reporting both planned and actual sample sizes (38, 66%), mean actual was 6.6% below planned, [median 0%, IQR (-5%, 5%)].

3.4. Bayesian trials

Of the four Bayesian trials, Wheatley (2011)**[REF16]** used information from a historic trial to help frame simulations to investigate the trial's operating characteristics, but no expert opinion, Stunnenberg (2018)**[REF17]** and Park (2019)**[REF18]** used experts to elicit MCIDs and Hampson (2015)**[REF19]** used experts to elicit priors for their model parameters. Of the four trials, three used informed priors, two used sensitivity analysis to assess prior impact on the design and one for the impact on the analysis. These papers demonstrate how bespoke trials that answer specific clinical questions can be designed using all available information. Wheatley planned to investigate chemotherapy options for Ewing sarcoma with a two stage RCT where Bayesian approaches were used to analyse the first randomisation between two strategies of induction chemotherapy. Conventional sample size calculations of 5% two-sided alpha and 80% power required 2500 recruited participants to detect 5% absolute difference in the primary outcome measure, 3-year event-free survival (EFS). A more plausible recruitment of 600 participants over 5 years was agreed, and the investigators specified a Bayesian model allowing them to make probability statements about treatment effects. The trial operating characteristics were investigated through simulation, fixing EFS for one arm based on a previous trial whilst varying

EFS for the other, and using priors that ensured no arm was favoured. They reported posterior probabilities that the Hazard Ratio occurred in different ranges. Stunnenberg performed a trial investigating the effectiveness of mexiletine for non-dystrophic myotonia, using a Bayesian hierarchical model to aggregate multiple N-of-1 trials. The MCID was set by a consensus meeting of 3 experts. The results of each new trial were combined with the results of the previous trials in a hierarchical model to build an updated Bayesian analysis. Once the posterior probability of having a meaningful clinical treatment effect had exceeded 80%, making it highly unlikely continuing the trial would change the outcome, or clear treatment failure was shown, the N-of-1 trials were stopped. Therefore, fewer participants were exposed to potentially inefficient treatment and moved to effective treatment faster. The treatment plan assumed all 30 participants completing the 4 treatment sets, equating to a standard RCT with 120 participants performing a crossover trial, giving a reduction of 90 participants. Only 4 participants entered their second N-of-1 trial whilst the other 23 participants entered only one (2 lost to follow up and 1 discontinued treatment). This trial demonstrated the possible savings from combining Bayesian methodology with N-of-1 trials to set early stopping rules based on posterior probabilities and elicited MCIDs. Park investigated whether specific clinical questions were answerable using group-sequential Bayesian adaptive design. Twenty-one paediatric stroke experts were surveyed about their preconceptions of using corticosteroids for focal cerebral arteriopathy (FCA), asking what probability of efficacy they would need to treat FCA with corticosteroids, and what level of efficacy/futility would stop them from randomising a subsequent participant. The results were used to specify an MCID and stopping criteria for the study design; data from a previous trial were used to model efficacy outcomes for the control arm. Simulations for 42 participants on a range of intervention effects produced safety and efficacy analysis and found that a trial would on average stop after 20-36 participants, dependent on intervention effect. This trial exemplifies how Bayesian methods can be used to set sensible stopping criteria and elicit parameters for the model. Hampson aimed to maximise the information available from a rare disease trial with 20 participants per arm through Bayesian methods. Fifteen clinicians attended an elicitation meeting performed over 2 days.

They were asked for their opinions about the probability of success, according to the primary endpoint, for participants treated with a drug for polyarteritis nodosa, and about the relative merits of a second drug. Opinions about the probability of success for the second drug were then derived. On day 1, experts reviewed current evidence for treatment options, then completed individual questionnaires of their beliefs for the most likely value for parameters and their uncertainty around them. Clinicians discussed individual priors to reach weighted consensus. On day 2, without seeing final consensus, clinicians were given information on another similar study with 70 participants per arm and asked to weight this trial against their consensus, this weighted information was used to form the final priors. A frequentist trial addressing the same clinical question would need 513 participants per arm to reach 90% power with a 2.5% one-sided alpha and taken over 30 years to recruit. This study demonstrates how expert priors can be developed and updated with information from new trials. Each paper reported priors differently, Wheatley only stated the prior value used for control EFS and that simulation used vague priors. Hampson fully detailed the elicitation process and which experts attended the elicitation meeting. Stunnenberg gave all simulation details, prior functions and distributions in supplementary material along with data from the N-of-1 trials. Park described how they would implement their prior in the paper, full details of the prior and associated sensitivity analysis were in supplementary material.

All papers stated how many participants were required for a frequentist trial, with 2400% (Hampson), 287% (Stunnenberg), 217% (Wheatley), and 30-55% (Park) more participants needed to run each. Hampson stated the effective sample size gain (the equivalent knowledge gained from information within the priors) was an extra 17 and 48 participants to each arm, 85% and 240% increases respectively.

4. Discussion

Few rare disease or small population trials used recommended methods[REF5, REF6, REF7] to overcome reduced sample size problems; including limited use of Bayesian trials requiring lower sample sizes than under frequentist frameworks. Use of Bayesian methods has been

limited due to trialists being more confident with frequentist methods and lacking expertise in Bayesian approaches, with few published Bayesian trials to refer to. There are also barriers to creating well-founded, transparent and informative priors, as this is time consuming and requires elicitation tools and understanding of their correct usage. There can also be concerns that priors and posteriors represent subjective rather than objective knowledge based exclusively on observed data.

Most trials used parallel arm design and frequentist methods. Some frequentist trials used design elements recommended in guidelines from Parmar, the EMA and the IRDRC to avoid reducing assumed power, yet other guideline suggestions including rerandomization where appropriate and adding covariate information to the design went unused or were used only infrequently. However, half did not reach their assumed standardised effect, potentially due to large average standardised effects used in these trials sample size calculation (0.7) or possibly because these treatments were ineffective. Whilst most frequentist trials used reasonable power levels, a substantial proportion did not use or report power in their sample size calculations. Lack of clear reporting occurred frequently within these trials.

Trials using Bayesian methods sidestep these power issues. They start from what is already known about a clinical question, and update this with new evidence about treatment effects. The four reviewed Bayesian papers reported full information about the priors used and their creation, allowing reproducibility. The posteriors from these trials can become priors for future Bayesian trials, as part of an on-going cycle of learning about treatments. To avoid incorrect conclusions, it is crucial that there is transparent reporting of all priors, and prior sensitivity is carried out and reported, so the reader can fully understand the effects of including different information through the prior on the posterior results.

Overall, most trials used continuous outcomes, whilst 20% used binary outcomes. Continuous outcomes were an improvement over binary outcomes due to power loss and residual confounding caused in using binomial, especially for dichotomised continuous outcomes.[REF20] Trials using n-of-1 designs benefitted from being able to test treatments repeatedly on a single

participant which makes them highly statistically efficient as it reduces the between arm variability present in parallel arm trials and so reduces the amount of participants needed overall.

Limitations of this research included the inability to check trials against some guideline recommendations (e.g. increasing timeframes, furthering differentiation between arms). Similarly, this set of trials is not comprehensive, as it only included trials self-reporting as rare; there were also only 4 Bayesian papers in our search limiting our scope. This report could be enriched by looking at specific rare diseases for a representative view from that area. However, this report provides a generalised overview across all medical conditions.

5. Conclusions

Rare disease and small population trials were found to lack recommended design and analysis approaches to increase statistical power. We agree with recommendations to implement Bayesian methods with informed priors to avoid underpowered trials and recommend following the framework for designing trials in smaller populations by Parmar et al. By attaining expert consensus on the required efficacy levels to change clinical practice, rare disease trials can become more feasible to run and more useful interventions can be found.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Sources

Giles Partington, NIHR pre-doctoral research methods fellow is funded by the National Institute for Health Research (NIHR) for this research project. SC is supported by an NIHR advanced fellowship (NIHR300593). The views expressed in this publication are those of the authors and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care.

References

1. Saglani S, Bush A, Carroll W, Cunningham S, Fleming L, Gaillard E, et al. *Biologics for Paediatric Severe Asthma: Trick or TREAT?* Lancet Resp Med. 2019; 7(4): 294-6

2. Evans SR, *Clinical trial structures.* J Exp Stroke Transl Med. 2010 Feb; 3(1): 8-18

3. Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ. *The n-of-1 clinical trial: the ultimate strategy for individualizing medicine?* Per Med. 2011 Mar; 8(2): 161-73

4. Day S, Jonker AH, Lau LPL, Hilgers RD, Irony I, Stallard N, et al. *Recommendations for the design of small population clinical trials.* Orphanet J Rare Dis. 2018; 13: 195

5. European Medical Agency. Small population trials guidelines, https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-trials-small-populations_en.pdf; 2006 <European medical agency small population trials guidelines> [accessed 14 April 2021]

6. Institute of Medicine. *Regulatory Framework for Drugs for Rare Diseases.* In: *Rare Diseases and Orphan Products: Accelerating Research and Development.* Washington, DC: National Academies Press; 2010. p. 73-110.

7. Parmar M, Sydes M, Morris T. *How do you design randomised trials for smaller populations? A framework.* BMC Med. 2016; 14: 183

8. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis. 3rd Ed.* New York: Chapman and Hall; 2013.

9. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. *Bayesian methods in health technology assessment: a review.* Health Technol Assess; 2000. 4(38): 1-130

10. International Rare Diseases Research Consortium, about us, https://irdirc.org/about-us/history/; 2021 [accessed 14 April 2021]

11. Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia. Available at www.covidence.org

12. StataCorp. 2019. *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC.

13. Dunning AJ, Reeves J. *Control of type 1 error in a hybrid complete two-period vaccine efficacy trial.* Pharm Stat. 2014; 13(6): 397-402

14. Rothwell JC, Julious SA, Cooper CL. *A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal.* Trials. 2018; 19(544)

15. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Snowdon C, et al. *What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies.* Trials. 2006; 7(9)

16. Anderton J, Moroz V, Brennan B. *International randomised controlled trial for the treatment of newly diagnosed EWING sarcoma family of tumours – EURO EWING 2012 Protocol*. Trials. 2020; 21*(*96)

17. Hampson LV, Whitehead J, Eleftheriou D, Tudur-Smith C, Jones R, Brogan PA, et al. *Elicitation of Expert Prior Opinion: Application to the MYPAN Trial in Childhood Polyarteritis Nodosa*. PLOS ONE 2015; 10(3)

18. Stunnenberg B, Raaphorst J, Groenewoud HM, Statland JM, Griggs RC, van der Wilt GJ, et al. *Effect of Mexiletine on Muscle Stiffness in Patients With Nondystrophic Myotonia Evaluated Using Aggregated N-of-1 Trials*. J Amer Med Assoc 2018; 320(22)

19. Park Y, Fullerton HJ, Elm JJ. *A pragmatic, adaptive clinical trial design for a rare disease: The Focal Cerebral Arteriopathy Steroid (FOCAS) trial*. Contemp Clin Trials 2019 Nov; 86: 105852

20. Royston P, Altman DG, Sauerbrei W. *Dichotomizing continuous predictors in multiple Regression: a bad idea.* Stat Med 2006 Jan; 25(1): 127-41

KEY WORDS:

EFS	Event Free Survival
EMA	European Medicines Agency
EMBASE	Excerpta Medica database
FCA	Focal Cerebral Arteriopathy
FDA	Federal Drugs Agency
IRDRC	International Rare Disease Research Consortium
IQR	InterQuartile Range
MCID	Minimal Clinically Important Difference
MEDLINE	Medical Literature Analysis and Retrieval System Online
RCT	Randomised Controlled Trial
SE	Standard Error
TREAT	TREATing severe paediatric asthma; a randomised controlled trial of mepolizumab and omalizumab