

Determining sample size for progression criteria for pragmatic pilot RCTs: The Hypothesis test Strikes Back!

Martyn Lewis (✉ a.m.lewis@keele.ac.uk)

Keele University <https://orcid.org/0000-0001-5290-7833>

Kieran Bromley

Keele University

Christopher J Sutton

The University of Manchester

Gareth McCray

Keele University

Helen Lucy Myers

Keele University

Gillian A Lancaster

Keele University

Methodology

Keywords: Outcome and Process Assessment; Pilots; Sample size, Statistics

Posted Date: April 29th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-24939/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published on February 3rd, 2021. See the published version at <https://doi.org/10.1186/s40814-021-00770-x>.

Abstract

Background The current CONSORT guidelines for reporting pilot trials do not recommend hypothesis testing of clinical outcomes on the basis that a pilot trial is under-powered to detect such differences and this is the aim of the main trial. It states that primary evaluation should focus on descriptive analysis of feasibility/process outcomes (e.g. recruitment, adherence, treatment fidelity). Whilst the argument for not testing clinical outcomes is justifiable, the same does not necessarily apply to feasibility/process outcomes, where differences may be large and detectable with small samples. Moreover, there remains much ambiguity around sample size for pilot trials.

Methods Many pilot trials adopt a 'traffic light' system for evaluating progression to the main trial determined by a set of criteria set up a priori. We construct a hypothesis-testing approach focused around this system that tests against being in the RED zone (unacceptable outcome) based on an expectation of being in the GREEN zone (acceptable outcome) and choose the sample size to give high power to reject being in the RED zone if the GREEN zone holds true. Pilot point estimates falling in the RED zone will be statistically non-significant and in the GREEN zone will be significant; the AMBER zone designates potentially acceptable outcome and statistical tests may be significant or non-significant.

Results For example, in relation to treatment fidelity, if we assume the upper boundary of the RED zone is 50% and the lower boundary of the GREEN zone is 75% (designating unacceptable and acceptable treatment fidelity, respectively), the sample size required for analysis given 90% power and one-sided 5% alpha would be around $n=30$ (intervention group alone). Observed treatment fidelity in the range of 0-15 participants (0-50%) will fall into the RED zone and be statistically non-significant; 16-22 (51-74%) fall into AMBER and may or may not be significant; 23-30 (75-100%) fall into GREEN and will be significant indicating acceptable fidelity.

Discussion In general, several key process outcomes are assessed for progression to a main trial; a composite approach would require appraising the rules of progression across all these outcomes. This methodology provides a formal framework for hypothesis-testing and sample size indication around process outcome evaluation for pilot RCTs.

Background

The importance and need for pilot and feasibility studies is clear: "A well-conducted pilot study, giving a clear list of aims and objectives ... will encourage methodological rigour ... and will lead to higher quality RCTs" (1). The CONSORT extension to external pilot and feasibility trials was published in 2016 (2) with the following key methodological recommendations: (i) investigate areas of uncertainty about the future definitive RCT; (ii) ensure primary aims/objectives are about feasibility, which should guide the methodology used; (iii) include assessments to address the feasibility objectives which should be the main focus of data collection and analysis, and (iv) build decision processes into the pilot design whether or how to proceed to the main study. Given that many trials incur process problems during

implementation – particularly with regards to recruitment (3-5) - the need for pilot and feasibility studies is evident.

One aspect of pilot and feasibility studies that remains unclear is the required sample size. There is no consensus but recommendations vary from 10-12 per group through to 60-75 per group (or, at least 20-30 overall) depending on the main objective of the study. Sample size may be based on: precision of a feasibility parameter (6,7); precision of a clinical parameter which may inform main trial sample size – particularly the standard deviation (SD) (8-11) but also event rate (12) and effect size (13,14); or, to a lesser degree, for clinical scale evaluation (9,15). Billingham et al. (16) reported that the median sample size of pilot and feasibility studies is around 30-36 per group but there is wide variation. Herbert et al. (17) reported that targets within internal as opposed to external pilots are often slightly larger and somewhat different, being based on percentages of the total sample size and timeline rather than any fixed sample requirement.

The need for a clear directive on sample size of studies is of utmost relevance. The CONSORT extension (2) reports that: “Pilot size should be based on feasibility objectives and some rationale given”, and states that a “confidence interval approach may be used to calculate and justify the sample size based on key feasibility objective(s)”. Specifically, item 7a (How sample size was determined: Rationale for numbers in the pilot trial) qualifies: “Many pilot trials have key objectives related to estimating rates of acceptance, recruitment, retention, or uptake ... for these sorts of objectives, numbers required in the study should ideally be set to ensure a desired degree of precision around the estimated rate”. Item 7b (When applicable, explanation of any interim analyses and stopping guidelines) is generally an uncommon scenario for pilot and feasibility studies and is not given consideration here.

A key aspect of pilot and feasibility studies is to inform progression to the main trial, which has important implications for all key stakeholders (funders, researchers, clinicians and patients). The CONSORT extension (2) states that: “decision processes about how to proceed needs to be built into the pilot design (which might involve formal progression criteria to decide whether to proceed, proceed with amendments, or not to proceed)” and authors should present “if applicable, the pre-specified criteria used to judge whether or how to proceed with a future definitive RCT; ... implications for progression from pilot to future definitive RCT, including any proposed amendments”. Avery et al. (18) published recommendations for internal pilots emphasising a traffic light (stop-amend-go / red-amber-green) approach to progression with focus on process assessment (recruitment, protocol adherence, follow-up) and transparent reporting around the choice of trial design and the decision-making processes for stopping, amending or proceeding to a main trial. The review of Herbert et al. (17) reported that the use of progression criteria (including recruitment rate) and traffic-light stop-amend-go as opposed to simple stop-go is increasing for internal pilot studies.

A common misuse of pilot and feasibility studies has been the application of hypothesis testing for clinical outcomes in small underpowered studies. Arain et al. (19) claimed that pilot studies were often poorly reported with inappropriate emphasis on hypothesis-testing. They reviewed 54 pilot and feasibility

studies published in 2007-8, of which 81% incorporated hypothesis-testing of clinical outcomes. Similarly, Leon et al. (20) stated that a pilot is not a hypothesis testing study: safety, efficacy and effectiveness should not be evaluated. Despite this, hypothesis testing has been commonly performed for clinical effectiveness/efficacy without reasonable justification. Horne et al., (21) reviewed 31 pilot trials published in physical therapy journals between 2012-5 and found that only 4/31 (13%) carried out a valid sample size calculation on effectiveness/efficacy outcomes but 26/31 (84%) used hypothesis testing. Wilson et al. (22) acknowledged a number of statistical challenges in assessing potential efficacy of complex interventions in pilot and feasibility. The CONSORT extension (2) re-affirmed many researchers' views that formal hypothesis testing for effectiveness/efficacy is not recommended in pilot/feasibility studies since they are underpowered to do so. Sim's commentary (23) further contests such testing of clinical outcomes stating that treatment effects calculated from pilot or feasibility studies should not be the basis of a sample size calculation for a main trial.

However, when the focus of analysis is on confidence interval estimation for process outcomes this does not give a definitive basis for acceptance/rejection of progression criteria linked to formal powering. The issue in this regard is that precision focuses on alpha (α , type I error) without clear consideration of beta (β , type II error), and may therefore not reasonably capture true differences if a study is under-powered. Further, it could be argued that hypothesis testing of feasibility outcomes is justified on the grounds that moderate-to-large differences ('process-effects') may be expected. Moore et al. (24) previously stated that some pilot studies require hypothesis testing to guide decisions about whether larger subsequent studies can be undertaken, giving the following example of how this could be done for feasibility outcomes: asking the question "Is taste of dietary supplement acceptable to at least 95% of the target population?" they showed that sample sizes of 30, 50, and 70 provide 48%, 78%, and 84% power to reject an acceptance rate of 85% or lower if the true acceptance rate is 95% using 1-sided $\alpha=0.05$ binomial test. Schoenfeld (25) advocates that, even for clinical outcomes, there may be a place for testing at the level of clinical 'indication' rather than 'clinical evidence'. He suggested that preliminary hypothesis testing for efficacy could be conducted with high alpha (up to 0.25), not to provide definitive evidence but as an indication as to whether a larger study should be conducted. Lee et al. (14) also reported how type 1 error levels other than the traditional 5% could be considered to provide preliminary evidence for efficacy, although they did stop short of recommending doing this by concluding that a confidence interval approach is preferable.

Current recommendations for analysis of pilot/feasibility studies are non-definitive, have a single rather than a multi-criterion basis and do not necessarily link directly to formal progression criteria. The purpose of this article is to introduce a simple methodology that allows sample size derivation and formal testing of proposed progression cut-offs, whilst offering suggestions for multi-criterion assessment, thereby giving clear guidance and sign-posting for researchers embarking on a pilot / feasibility study to assess uncertainty in feasibility parameters prior to a main trial. The suggestions within the article do not directly apply to internal pilot studies built into the design of a main trial, but given the similarities to external randomised pilot and feasibility studies, many of the principles outlined here for external pilots might also extend to some degree to internal pilots of randomised and non-randomised studies.

Methods And Results

The proposed approach focuses on estimation and hypothesis testing of progression criteria for feasibility outcomes that are potentially modifiable (e.g. recruitment, treatment fidelity/ adherence, level of follow up). Thus, it aligns with the main aims and objectives of pilot and feasibility studies and with the progression stop-amend-go recommendations of Eldridge et al. (2) and Avery et al. (18).

Hypothesis concept

The concept is to set up hypothesis-testing around progression criteria that tests against being in the RED zone (designating lower/unacceptable feasibility outcome – ‘**STOP**’) based on an expectation of being in the GREEN zone (designating higher/acceptable feasibility outcome – ‘**GO**’). Specifically, testing is against the upper “RED” (**stop**) cut-off (denoted, R_{UL}) (on the assumption that the “GREEN” (**go**) threshold – lowest value of the GREEN zone (denoted, G_{LL}) – is true), as:-

- Null hypothesis: True feasibility not greater than the upper “RED” **stop** limit (R_{UL})
- Alternative hypothesis: True feasibility is greater than R_{UL}

The test is a 1-tailed test with suggested customary values for alpha (α) of 0.1, 0.05 or 0.01 and beta (β) of 0.1 or 0.2, dependent on the required strength of evidence of the test.

Progression rules

Let E denote the observed point estimate (ranging from 0 to 1 for proportions, or for percentages 0-100%); R_{UL} , the upper limit for the RED zone; G_{LL} , the lower limit for the GREEN zone. Simple 3-tiered progression criteria would follow as:-

- $E \leq R_{UL}$ [P-value non-significant ($P \geq \alpha$)] -> RED (unacceptable - STOP)
- $R_{UL} < E < G_{LL}$ -> AMBER (potentially acceptable - AMEND)
- $E \geq G_{LL}$ [P-value significant ($P < \alpha$)] -> GREEN (acceptable - GO)

In this case, we express progression criteria as three distinct progression signals, as illustrated in Figure 1(a).

Sample size

Table 1 displays a quick look-up grid for sample size across a range of anticipated proportions for R_{UL} and G_{LL} for one-sample one-sided 1% and 5% alpha with typical 80% and 90% power (see Appendix 1 for corresponding mathematical expression; derived from Fleiss et al. (27)). Clearly, as the difference between proportions R_{UL} and G_{LL} increases the sample size requirement is reduced.

Table 1: Sample size and significance cut-points for (G_{LL} - R_{UL}) differences, power (80%/90%) and 1-tailed significance levels (1%/5%).

R_{UL} %	G_{LL} %	a(0.05) b(0.2)			a(0.05) b(0.1)			a(0.01) b(0.1)		
		n	A_{sig}	$A_R\%$	n	A_{sig}	$A_R\%$	n	A_{sig}	$A_R\%$
10	20	79	16.6	66.1	112	15.6	55.5	157	16.9	68.6
15	25	101	21.5	65.5	141	20.5	55.4	202	21.7	67.3
	30	50	24.7	64.8	69	23.3	55.1	97	25.3	68.4
20	30	119	26.5	65.3	166	25.5	55.3	241	26.6	66.4
	35	57	29.7	64.9	79	28.3	55.1	113	30.1	67.4
	40	34	32.9	64.6	47	31.0	55.0	66	33.7	68.4
25	35	135	31.5	64.9	186	30.5	55.3	272	31.6	66.0
	40	64	34.6	64.2	87	33.3	55.1	126	35.0	66.7
	45	37	37.9	64.5	51	36.0	54.9	73	38.5	67.4
	50	25	40.9	63.7	34	38.7	54.6	48	42.0	68.0
30	40	146	36.5	64.9	201	35.5	55.3	297	36.6	65.6
	45	69	39.6	63.9	94	38.2	54.8	136	39.9	66.2
	50	40	42.7	63.7	54	41.0	54.8	78	43.4	66.8
	55	26	45.9	63.8	35	43.7	55.0	51	46.8	67.2
	60	20*	48.3	61.0	26	46.0	53.5	36	50.5	68.3
35	45	155	41.5	64.7	213	40.5	55.2	316	41.5	65.3
	50	72	44.6	63.9	98	43.2	54.8	144	44.8	65.6
	55	42	47.6	63.1	56	45.9	54.7	82	48.2	66.0
	60	27	50.8	63.2	36	48.7	54.8	53	51.6	66.5
	65	20*	53.4	61.3	26	51.1	53.8	37	55.3	67.6
40	50	161	46.4	64.5	220	45.5	55.2	327	46.5	65.1
	55	74	49.5	63.7	100	48.2	54.8	148	49.8	65.3
	60	43	52.5	62.7	57	50.9	54.5	84	53.1	65.5
	65	28	55.5	62.1	37	53.5	54.0	54	56.5	65.8
	70	20*	58.3	61.0	26	56.0	53.5	38	59.9	66.3
45	55	164	51.4	64.2	222	50.5	55.2	333	51.5	64.8
	60	75	54.5	63.2	100	53.2	54.8	149	54.8	65.1
	65	43	57.5	62.4	57	55.8	54.2	84	58.0	65.2
	70	28	60.4	61.5	36	58.6	54.2	53	61.5	65.8
	75	20#*	63.0	60.1	25	61.1	53.7	37	64.9	66.2
50	60	163	56.4	64.1	221	55.5	55.1	331	56.5	64.7
	65	74	59.5	63.0	99	58.2	54.5	147	59.7	64.9
	70	42	62.4	62.2	55	60.9	54.3	82	63.0	65.0
	75	27	65.3	61.3	35	63.5	53.8	52	66.3	65.1
55	65	159	61.4	63.9	215	60.5	55.0	323	61.5	64.5
	70	72	64.4	62.6	95	63.2	54.5	143	64.7	64.5
	75	40	67.4	62.0	53	65.8	53.9	79	67.9	64.6
60	70	152	66.4	63.6	205	65.5	54.8	309	66.4	64.3
	75	68	69.3	62.3	90	68.1	54.1	135	69.6	64.3
	80	38	72.2	61.1	49	70.8	53.8	74	72.9	64.3
65	75	143	71.3	63.0	190	70.5	54.7	288	71.4	64.0
	80	63	74.3	61.7	82	73.1	54.1	124	74.6	64.1
	85	35	77.0	60.2	44	75.7	53.7	67	77.8	64.1
70	80	129	76.3	62.7	171	75.4	54.5	260	76.4	63.8
	85	57	79.1	60.7	73	78.0	53.6	111	79.5	63.6
75	85	113	81.2	61.9	147	80.4	54.3	225	81.4	63.6
	90	49#	83.9	59.5	61	83.0	53.4	94	84.5	63.3
80	90	93	86.1	60.9	119	85.4	53.8	183	86.3	63.3

R_{UL} =upper limit of RED zone; G_{LL} =lower limit of GREEN zone; A_{sig} =AMBER-statistical significance threshold (within the AMBER zone) where an observed estimate equal or below the cut-point will result in a non-significant result ($p \geq 0.05$ or ≥ 0.01) and figures above the cut-point will be significant ($p < 0.05$ or < 0.01); $A_R\%$ =percent of AMBER zone yielding a non-significant test result (% within $AMBER_R$ sub-zone).

Sample sizes were derived using the normal approximation to the binomial distribution (with continuity correction) formula given in the Appendix, which by convention is stable for $np > 5$ and $n(1-p) > 5$ (this is the case for the scenarios in the above table except where indicated by # where $n(1-p) = 4.9$ and 5.0). * Derived sample size is less than 25 which may give overly wide confidence intervals (if $n \geq 25$ (recommended) then the standard error for purposes of 1-sided interval estimation will be no greater than 0.1).

Multi-criteria assessment

We recommend that progression for all key feasibility criteria should be considered separately, and hence overall progression would be determined by the worst-performing criterion e.g. RED if at least one signal is RED; AMBER if none of the signals fall into RED but at least one falls into AMBER; GREEN if all signals fall into the GREEN zone. Hence, the GREEN signal to 'GO' across the set of individual criteria will give indication that progression to a main trial can take place without any necessary changes. A signal to 'stop' and not proceed to a main trial is recommended if any of the observed estimates are 'unacceptably' low (i.e. fall within the (RED) zone). Otherwise, where neither 'GO' nor 'STOP' are signalled, the design of the trial will need amending by indication of subpar performance on one or more of the criteria.

Sample size requirements across multi-criteria will vary according to the designated parameters linked to the progression criteria, which may be set at different stages of the study on different numbers of patients (e.g. those screened, eligible, recruited and randomised, allocated to the intervention arm, total followed up). This is illustrated in Box 1 for statistical testing requirement at different levels e.g. number needed to be targeted/screened (for potential recruitment); number to be randomised; number to be randomised specifically to the intervention arm. The overall size needed will be dictated by the requirement to power each of the multi-criteria statistical tests. Since these tests will yield separate conclusions in regards to the decision to 'GO' across all individual feasibility criteria there is no need to consider a multiple testing correction with respect to alpha. However, researchers may wish to increase power (and hence, sample size) to ensure adequate power to detect 'GO' signals across the collective set of feasibility criteria. For example, powering at 90% across three (assumed independent) criteria will ensure a collective power of 73% (i.e. 0.9^3), which may be considered reasonable; but, 80% power across five criteria will give an overall probability of only 33% for identifying 'GO' signals across all five criteria.

Box 1: A two-arm parallel design (1:1 allocation to intervention and control arms) with three key feasibility objectives, to assess: (i) recruitment uptake (percent of screened patients recruited); (ii) treatment fidelity; (iii) participant retention (follow up). Hypothesis-testing incorporates α (1-sided)=5% and power=90%.

Assume the progression criteria (and affiliated sample size requirements) for each are as follows:-

(i) Recruitment uptake: $\leq 20\%$ (RED zone), $\geq 35\%$ (GREEN zone) $\{R_{UL} = 20\% / G_{LL} = 35\%\}$

1. Required sample size $n = 79$ [total screened patients]

(ii) Treatment fidelity: $\leq 50\%$ (RED zone), $\geq 75\%$ (GREEN zone) $\{R_{UL} = 50\% / G_{LL} = 75\%\}$

1. Required sample size $n = 35$ [intervention arm only]

(iii) Follow up: $\leq 65\%$ (RED zone), $\geq 85\%$ (GREEN zone) $\{R_{UL} = 65\% / G_{LL} = 85\%\}$

1. Required sample size $n = 44$ (total randomised participants with 22 per arm)

The sample sizes across criteria (i)-(iii) are at different levels - (i) is at the level of screened patients, whereas (ii)-(iii) are at the level of all randomised patients. To meet criteria (i) we need $n_{\text{screened}} \geq 79$ (although we anticipate $n_{\text{screened}} = 200$ (i.e. $(1/0.35) \times n_{\text{randomised}}$ where 0.35 is the expected proportion uptake of the total number screened), and for (ii)-(iii) we need $n_{\text{randomised}} = 70$ (35 per arm, based on (ii)).

Taking each of the objectives in turn (and the updated sample sizes to meet the multi-criteria objectives), we express progression criteria for the three objectives as follows:-

(i) Recruitment uptake [$n_{\text{screened}} \geq 79$; expected $n_{\text{screened}} = 200$; maximum $n_{\text{screened}} = 350$ (i.e. $(1/0.2) \times n_{\text{randomised}}$)]

- $E \leq 0.2$ [$P \geq 0.05$] -> RED (STOP)
- $0.2 < E < 0.35$ -> AMBER (AMEND)
- $E \geq 0.35$ [$P < 0.05$] -> GREEN (GO)

(with the following signals for expected $n_{\text{screened}} = 200$: 0 - 40 (RED); 41 - 69 (AMBER); 70 - 200 (GREEN)) [i.e. $\leq 0.2 \times 200 = 40$ and $\geq 0.35 \times 200 = 70$]

(ii) Treatment fidelity [$n_{\text{intervention-arm}} = 35$ (intervention arm only)]

- $E \leq 0.5$ [$P \geq 0.05$] -> RED (STOP)
- $0.5 < E < 0.75$ -> AMBER (AMEND)
- $E \geq 0.75$ [$P < 0.05$] -> GREEN (GO)

(with the following signals for $n_{\text{intervention-arm}} = 35$: 0 - 17 (RED); 18 - 26 (AMBER); 27 - 35 (GREEN)) [i.e. $\leq 0.5 \times 35 = 17$ and $\geq 0.75 \times 35 = 27$]

(iii) Follow up [$n_{\text{randomised}} = 70$ (intervention and control arms)]

- $E \leq 0.65$ [$P \geq 0.05$] -> RED (STOP)
- $0.65 < E < 0.85$ -> AMBER (AMEND)
- $E \geq 0.85$ [$P < 0.05$] -> GREEN (GO)

(with the following signals for $n_{\text{randomised}} = 70$: 0 - 45 (RED); 46 - 59 (AMBER); 60 - 70 (GREEN)) [i.e. $\leq 0.65 \times 70 = 45$ and $\geq 0.85 \times 70 = 60$]

In accordance with the multi-criteria aim, the decision to proceed would be based on the worst-signal

1. If signal=RED for (i) or (ii) or (iii) -> overall signal is RED

2. Else, if no signal is RED but signal=AMBER for (i) or (ii) or (iii) -> overall signal is AMBER

3. Else, if signals=GREEN for (i) and (ii) and (iii) -> overall signal is GREEN

Further expansion of AMBER zone

Within the same sample size framework the AMBER zone may be further split to indicate whether 'minor' or 'major' amendments are required according to the significance of the p-value. Consider a 2-way split in the AMBER zone (designated AMBER_R (region of Amber zone adjacent to RED zone) and AMBER_G (region of AMBER zone adjacent to GREEN zone). This would draw on two possible levels of amendment (major amend and minor amend) (as shown in Figure 1(b)). Hence, the re-configured approach would follow as:-

- $E \leq R_{UL}$ [P-value non-significant ($P \geq a$)] -> RED (unacceptable - STOP)
- $R_{UL} < E < G_{LL}$ -> AMBER (potentially acceptable - AMEND)
 - $R_{UL} < E < G_{LL}$ and $P \geq a$ -> AMBER_R (major AMEND)
 - $R_{UL} < E < G_{LL}$ and $P < a$ -> AMBER_G (minor AMEND)
- $E \geq G_{LL}$ [P-value significant ($P < a$)] -> GREEN (acceptable - GO)

Figure 2 illustrates the signals according to this expanded approach. The worked example for this extension is presented in Box 2.

Box 2:

Taking each of the objectives in turn, we re-express the progression criteria for the three objectives according to the 4-tiered approach, as follows:-

(i) Recruitment uptake [expected $n_{\text{screened}}=200$]

- $E \leq 0.2$ [$P \geq 0.05$] -> RED (STOP)
- $0.2 < E < 0.35$ -> AMBER (AMEND)
 - $0.2 < E < 0.256$ [$P \geq 0.05$] -> AMBER_R (AMEND-major)
 - $0.256 < E < 0.35$ [$P < 0.05$] -> AMBER_G (AMEND-minor)
- $E \geq 0.35$ [$P < 0.05$] -> GREEN (GO)

(with the following signals for $n_{\text{screened}}=200$: 0-40 (RED); 41-51 (AMBER-major); 52-69 (AMBER-minor); 70- 200 (GREEN))

(ii) Treatment fidelity [$n_{\text{intervention-arm}}=35$ (intervention arm only)]

- $E \leq 0.5$ [$P \geq 0.05$] -> RED (STOP)
- $0.5 < E < 0.75$ -> AMBER (AMEND)
 - $0.5 < E < 0.643$ [$P \geq 0.05$] -> AMBER_R (AMEND-major)
 - $0.643 < E < 0.75$ [$P < 0.05$] -> AMBER_G (AMEND-minor)

- $E \geq 0.75$ [$P < 0.05$] -> GREEN (GO)

(with the following signals for $n_{\text{intervention-arm}}=35$: 0-17 (RED); 18-22 (AMBER-major); 23-26 (AMBER-minor); 27-35 (GREEN))

(iii) Follow up [$n_{\text{randomised}}=70$ (intervention and control arms)]

- $E \leq 0.65$ [$P \geq 0.05$] -> RED (STOP)
- $0.65 < E < 0.85$ -> AMBER (AMEND)
 - $0.65 < E < 0.742$ [$P \geq 0.05$] -> AMBER_R (AMEND-major)
 - $0.742 < E < 0.85$ [$P < 0.05$] -> AMBER_G (AMEND-minor)
- $E \geq 0.85$ [$P < 0.05$] -> GREEN (GO)

(with the following signals for $n_{\text{randomised}}=70$: 0-45 (RED); 46-51 (AMBER-major); 52-59 (AMBER-minor); 60-70 (GREEN))

In accordance with the multi-criteria aim, the decision to proceed would be based on the worst-signal (as in Box 1)

Discussion

The methodology introduced in this article provides an innovative formal framework and approach to sample size derivation, aligning sample size requirement to progression criteria with the intention of providing greater transparency to the progression process and full engagement with the standard aims and objectives of pilot/feasibility studies. Through the use of both alpha and beta parameters (rather than alpha alone), the method ensures rigour and capacity to address the progression criteria by ensuring there is adequate power to detect an acceptable threshold for moving forward to the main trial. As several key process outcomes are assessed in parallel and in combination, the method embraces a composite multi-criterion approach that appraises signals for progression across all the targeted feasibility measures. The methodology extends beyond the requirement for “sample size justification but not necessarily sample size calculation” (28).

The focus of the strategy reported here is on process outcomes, which align with the recommended key objectives of primary feasibility evaluation for pilot and feasibility studies (2,24) and necessary targets to address key issues of uncertainty (29). The concept of justifying progression is key. Charlesworth et al. (30) developed a checklist for intended use in decision-making on whether pilot data could be carried forward to a main trial. Our approach builds on this philosophy by introducing a formalised hypothesis test to address the key objectives. Though the suggested sample size derivation focuses around the key process objectives, it may also be the case that other objectives are also important e.g. assessment of precision of clinical outcome parameters. In this case, researchers may also wish to ensure that the size of the study suitably covers the needs of those evaluations e.g. to estimate the SD of the intended clinical outcome, then the overall sample size may be boosted to cover this additional objective (10). This tallies with the review by Blatch-Jones et al. (31) who reported that testing recruitment, determining the sample

size and numbers available, and the intervention feasibility were the most commonly used targets of pilot evaluations.

Hypothesis-testing in pilot studies, particularly in the context of effectiveness/efficacy of clinical outcomes, has been widely criticized due to the improper purpose and lack of statistical power of such evaluations (2,20,21,23). Hence, pilot evaluations of clinical outcomes are not expected to include hypothesis testing. Since the main focus is on feasibility the scope of the testing reported here is different and importantly relates back to the recommended objectives of the study whilst also aligning with nominated progression criteria (2). Hence, there is clear justification for this approach.

We provide recommended sample sizes within a look-up grid relating to perceived likely progression cut-points to aid quick access and retrievable sample sizes for researchers. For a likely set difference in proportions of 0.15 to 0.2 when $\alpha=0.05$ and $\beta=0.1$ the corresponding required sample sizes take the range of around 50 to 100 (25 to 50 per arm) for a two-arm comparison, and around 35 to 75 in total for studies with $\beta=0.2$. Note, for treatment fidelity particularly, the marginal difference could be higher e.g. $\geq 25\%$, but since this relates to an arm-specific objective (relating to evaluation of the intervention only) then a usual 1:1 pilot will require twice the size; hence, the arm-specific sample size powered for detecting a $\geq 25\%$ difference from the null would be about 30 – as depicted from our illustration (and $n=60$ overall for a 1:1 pilot; intervention and control arms). Hence, we expect that typical pilot sizes of around 30-40 randomised per arm (16) would likely fit with the proposed methodology within this manuscript (the number needed for screening being extrapolated upward of this figure).

Importantly, the methodology outlines the necessary multi-criterion approach to the evaluation of pilot and feasibility studies. If all progression criteria are performing as well as anticipated (highlighting 'go' according to all criteria) then the recommendation of the pilot/feasibility study is that all criteria meet their desired levels with no need for adjustment and the main trial can proceed without amendment. However, if the worst signal (across all measured criteria) is an amber signal; then adjustment will be required against those criteria that fall within that signal. Consequently, there is the possibility that the criteria may need subsequent re-assessment to re-evaluate processes in line with updated performance for the criteria in question. If one or more of the feasibility statistics fall within the RED zone then this signals 'stop' and concludes that a main trial is not feasible based on those criteria. This approach to collectively appraising progression based on the results of all feasibility outcomes assessed against their criteria will be conservative as the power of the collective will be lower than the individual power of the separate tests; hence, it is recommended that the power of the individual tests is set high enough (for example, 90-95%) to ensure the collective power is high enough (e.g. at least 70 or 80%) to detect true 'go' signals across all the feasibility criteria.

In this article we also expand the possibilities for progression criterion and hypothesis testing where the AMBER zone is sub-divided arbitrarily based on the significance of the p-value. This may work well when the AMBER zone has a wide range and is intended to provide a useful and workable indication of the level of amendment ('minor' (non-substantive) or 'major' (substantive)) required to progress to the main trial.

Examples warranting substantial amendment include study re-design including possible re-appraisal and change of statistical parameters, inclusion of several additional sites, adding further data recruitment methods, significant reconfiguration of exclusions, major change to the method of delivery of trial intervention, additional mode of collecting and retrieving data. Minor amendments include small changes to the protocol and methodology e.g. addition of one or two sites for attaining a slightly higher recruitment rate, and adding a further reminder process for boosting follow up. For the most likely parametrisation of $a=0.05/ b=0.1$, the AMBER zone division will be roughly at the midpoint. Other ways of providing an indication of level of amendment could include evaluation and review of the point and interval estimates or by evaluating posterior probabilities via a Bayesian approach (14).

The methodology illustrated here focuses on feasibility outcomes presented as percentages/proportions, which is the most common form for progression criteria under consideration. However, the steps that have been introduced can be readily adapted to any feasibility outcomes taking a numerical format e.g. rate of recruitment per month per centre.

Issues relating to progression criteria for internal pilots may be different to those for external pilots and non-randomised feasibility studies. The consequence of a 'stop' within an internal pilot may be more serious for stakeholders (researchers, funders, patients) as it would bring an end to the planned continuation into the main trial phase, whereas there would be less at stake for a negative external pilot. By contrast, the consequence of a 'go' signal may work the other way with a clear and immediate gain for the internal pilot whereas for an external pilot, the researchers would still need to apply and get the necessary funding and approvals to undertake an intended main trial. The chances of falling into the different traffic-light zones are likely to be quite different between the two designs. Possibly external pilot and feasibility studies are more likely to have estimates falling in and around the RED zone than for internal pilots, reflecting the greater uncertainty in the processes for the former and greater confidence in the mechanisms for trial delivery for the latter. However, to counter this, there are often large challenges with recruitment within internal pilot studies where the target population is usually spread over more diverse sites than may be expected for an external pilot. Despite this possible imbalance the interpretation of zonal indications remains consistent for external and internal pilot studies. As such, our focus with regards to the recommendations in this article are aligned to requirements for external pilots; though, application of this methodology to a degree may similarly hold for internal pilots (and further, to non-randomised studies that can include progression criteria - including longitudinal observational cohorts with the omission of the treatment fidelity criterion).

Conclusions

We propose a novel framework that provides a paradigm shift towards formally testing feasibility progression criteria in pilot and feasibility studies. The outlined approach ensures rigorous and transparent reporting in line with CONSORT recommendations for evaluation of stop-amend-go criteria and presents clear progression signposting which should help decision-making and inform stakeholders. Targeted progression criteria are focused on recommended pilot and feasibility objectives, particularly

recruitment uptake, treatment fidelity and participant retention, and these criteria guide the methodology for sample size derivation and statistical testing. This methodology is intended to provide a more definitive and rounded structure to pilot and feasibility design and evaluation than currently exists. Sample size recommendations will be dependent on the nature and cut-points for multiple key pre-defined progression criteria but the typical overall size of around 30-40 per arm is likely to be sufficient to fulfil the necessary requirements for evaluation against these criteria, whilst also ensuring a sufficient sample size for other feasibility outcomes such as review of the precision of clinical parameters to better inform main trial size.

List Of Abbreviations

Alpha (α) = Significance level (Type I error probability)

AMBER_G = AMBER sub-zone split adjacent to the GREEN zone (within 4-tiered approach)

AMBER_R = AMBER sub-zone split adjacent to the RED zone (within 4-tiered approach)

A_R% = Percent of AMBER zone yielding a non-significant test result (% within AMBER_R sub-zone)

A_{sig} = AMBER-statistical significance threshold (within the AMBER zone) where an observed estimate equal or below the cut-point will result in a non-significant result ($p \geq 0.05$ or ≥ 0.01) and figures above the cut-point will be significant ($p < 0.05$ or < 0.01)

Beta (β) = Power (1 – Type II error probability)

E = Best-estimate (the observed point estimate)

G_{LL} = Lower Limit of GREEN zone

n = Sample size (n_{screened} = number of patients screened; $n_{\text{randomised}}$ = number of patients randomised; $n_{\text{intervention-arm}}$ = number of patients randomised to the intervention arm only)

R_{UL} = Upper Limit of RED zone

Declarations

Ethical approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: Not applicable.

Competing interests: The authors declare that they have no competing interests.

Funding: KB was supported by a UK 2017 NIHR Research Methods Fellowship Award (ref: RM-FI-2017-08-006).

Authors' contributions: ML and CJS conceived the original methodological framework for the paper. ML prepared draft manuscripts. KB and GMcC provided examples and illustrations. All authors contributed to the writing and provided feedback on drafts and steer and suggestions for article updating. All authors read and approved the final manuscript.

Acknowledgements: We thank Professor Julius Sim, Dr Ivonne Solis-Trapala, Dr Elaine Nicholls and Marko Raseta for their feedback on the initial study abstract.

References

1. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract* 2004; 10(2):307-12
2. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *Pilot Feasibility Stud* 2016; 2:64
3. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. *Trials* 2006; 7:9
4. Sully BG, Julious SA, Nicholl J. A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies. *Trials* 2013; 14:166
5. Walters SJ, Bonacho Dos Anjos Henriques-Cadby I, Bortolami O, Flight L, Hind D, Jacques RM, et al. Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom Health Technology Assessment Programme. *BMJ Open* 2017; 7(3):e015276. *BMJ Open* 2017; 7:e015276.
6. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. *Pharm Stat* 2005; 4:287-291
7. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol* 2010; 10:1
8. Browne RH. On the use of a pilot sample for sample size determination. *Stat Med* 1995; 14:1933–1940
9. Hertzog MA. Considerations in determining sample size for pilot studies. *Res Nurs Health* 2008; 31(2):180-91.
10. Sim J and Lewis M. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. *J Clin Epidemiol* 2012; 65(3):301-8.
11. Whitehead AL, Julious SA, Cooper CL, Campbell MJ. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Stat Methods Med Res* 2016; 25(3):1057–73.

12. Teare MD, Dimairo M, Shephard N, Hayman A, Whitehead A, Walters SJ. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*. 2014; 15:264.
13. Cocks K and Torgerson DJ. Sample size calculations for pilot randomized trials: a confidence interval approach. *J Clin Epidemiol* 2013; 66(2):197-201.
14. Lee EC, Whitehead AL, Jacques RM, Julious SA. The statistical interpretation of pilot trials: should significance thresholds be reconsidered? *BMC Med Res Methodol* 2014; 14:41
15. Johanson GA and Brooks GP. Initial Scale Development: Sample Size for Pilot Studies. *Edu Psychol Measurement* 2010; 70(3):394–400
16. Billingham SA, Whitehead AL, Julious SA. An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. *BMC Med Res Methodol* 2013; 13:104.
17. Herbert E, Julious SA, Goodacre S. Progression criteria in trials with an internal pilot: an audit of publicly funded randomised controlled trials. *Trials* 2019; 20(1):493.
18. Avery KN, Williamson PR, Gamble C, O'Connell Francischetto E, Metcalfe C, Davidson P, et al. Informing efficient randomised controlled trials: exploration of challenges in developing progression criteria for internal pilot studies. *BMJ Open* 2017; 7(2):e013537
19. Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Med Res Methodol* 2010; 10:67
20. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. *J Psychiatr Res*. 2011; 45(5):626-9. Moore CG, Carter RE, Nietert PJ, Stewart PW. Recommendations for planning pilot studies in clinical and translational research. *Clin Transl Sci* 2011; 4(5):332-7
21. Horne E, Lancaster GA, Matson R, Cooper A, Ness A, Leary S. Pilot trials in physical activity journals: a review of reporting and editorial policy. *Pilot Feasibility Stud* 2018; 4:125
22. Wilson DT, Walwyn RE, Brown J, Farrin AJ, Brown SR. Statistical challenges in assessing potential efficacy of complex interventions in pilot or feasibility studies. *Stat Methods Med Res* 2016; 25(3):997-1009.
23. Sim J. Should treatment effects be estimated in pilot and feasibility studies? *Pilot Feasibility Stud* 2019; 5:107.
24. Moore CG, Carter RE, Nietert PJ, Stewart PW. Recommendations for planning pilot studies in clinical and translational research. *Clin Transl Sci* 2011; 4(5):332-7.
25. Schoenfeld D. Statistical considerations for pilot studies. *Int J Radiat Oncol Biol Phys* 1980; 6 (3): 371–374
26. File:Saudi Arabia - Road Sign - Traffic signals ahead.svg by ⚔⚔⚔⚔⚔⚔⚔⚔⚔ licensed under CC BY-SA 4.0. Source = Wikimedia Commons. Web-link, <https://search.creativecommons.org/photos/6314f39f-a9a0-4fb7-a9b1-dd86baabb67e>. Accessed 18 Apr 2020.

27. Fleiss JL, Levin B, Paik MC. (2003). Statistical Methods for Rates and Proportions, Third Edition, John Wiley & Sons, New York.
28. Julious SA. Pilot Studies in clinical research. Stat Methods Med Res 2016; 25(3):995-6
29. Lancaster GA. Pilot and feasibility studies come of age! Pilot Feasibility Stud 2015; 1(1):1
30. Charlesworth G, Burnell K, Hoe J, Orrell M, Russell I. Acceptance checklist for clinical effectiveness pilot trials: a systematic approach. BMC Med Res Methodol 2013; 13:78
31. Blatch-Jones AJ, Pek W, Kirkpatrick E, Ashton-Key M. Role of feasibility and pilot studies in randomised controlled trials: a cross-sectional study. BMJ Open 2018; 8(9):e022233.

Figures

(a)



(b)

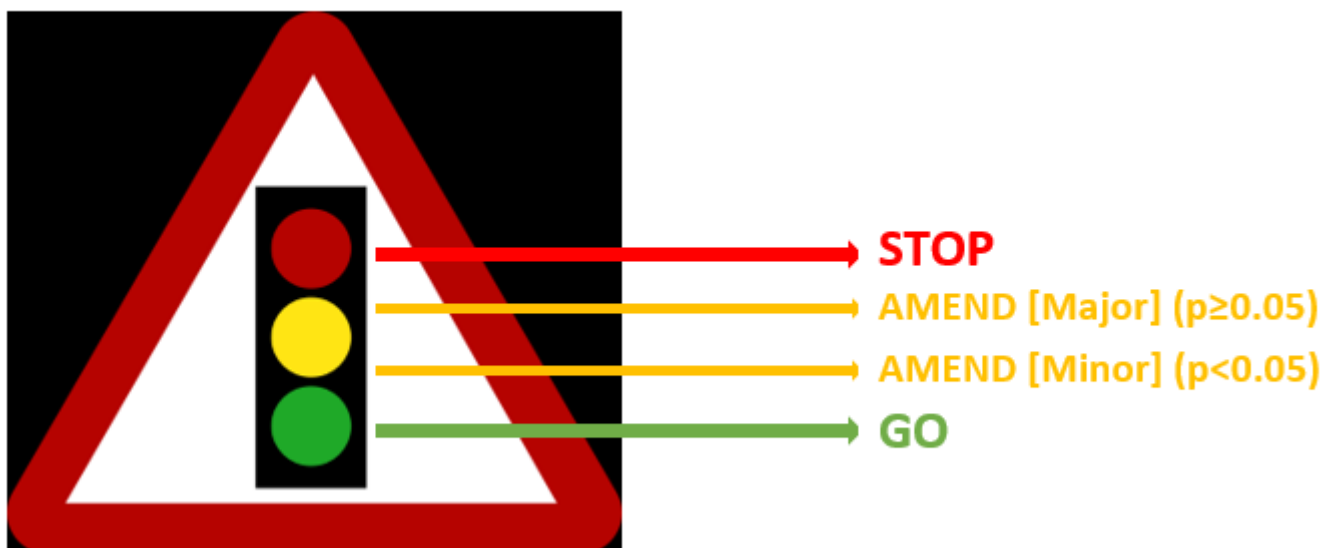


Figure 1

Illustration of the traffic-light system for progression indicators (a) Simple 3-tiered indicators (b) Extended 4-tiered indicators Illustrations here were extracted from the following web-link reference (modified to include STOP-AMEND-GO signposting) (26)

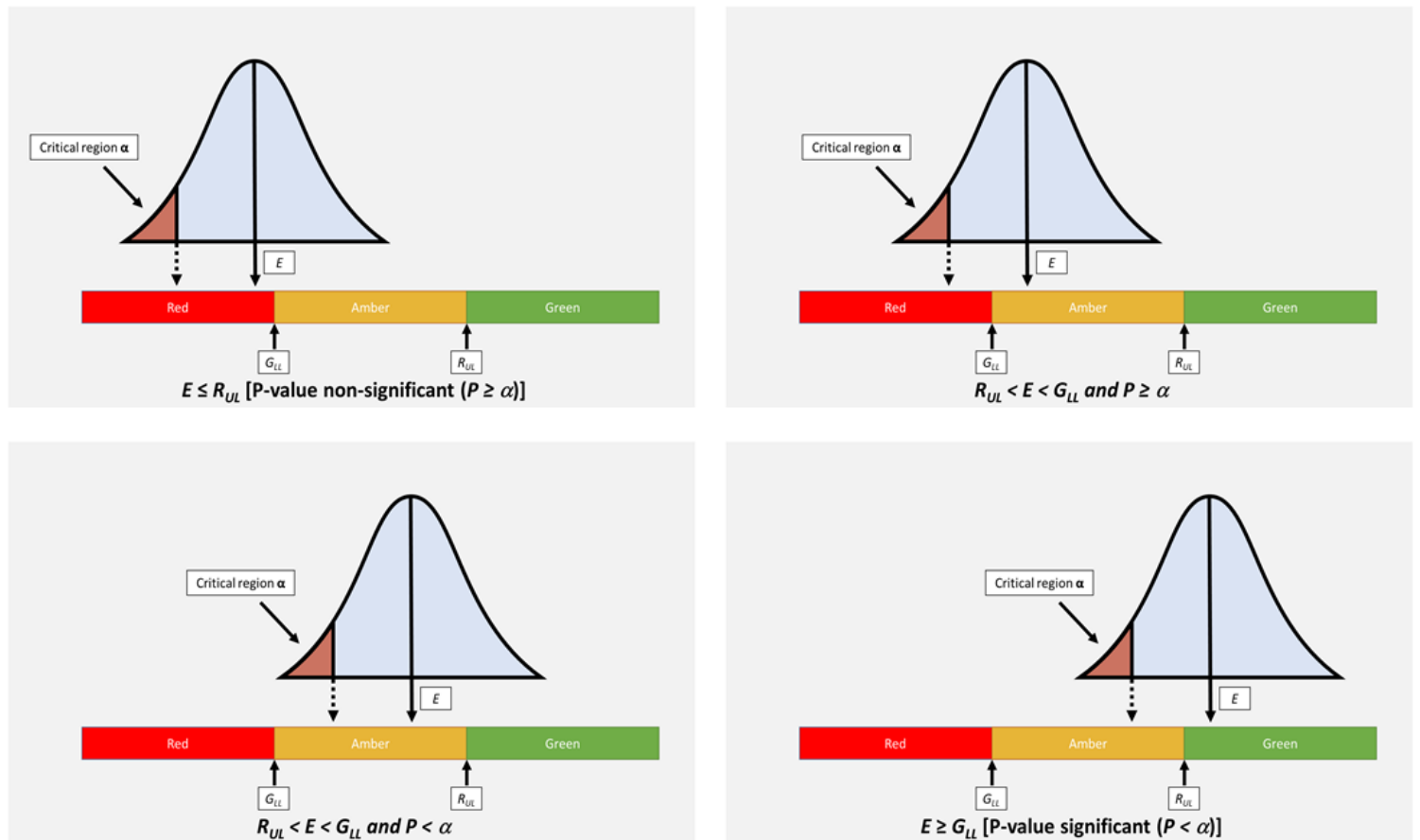


Figure 2

Illustration of signals according to the expanded 4-tiered system E= Observed point estimate; RUL=upper limit of RED zone; GLL=lower limit of GREEN zone.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Appendix.pdf](#)