STATISTICAL ANALYSIS

original reports

Two-Stage Single-Arm Trials Are Rarely Analyzed Effectively or Reported Adequately

Michael J. Grayling, PhD[1] and and Adrian P. Mander, PhD[2]

abstract

PURPOSE Two-stage single-arm designs have historically been the most common design used in phase II oncology. They remain a mainstay today, particularly for trials in rare subgroups. Consequently, it is imperative such studies be designed, analyzed, and reported effectively. We comprehensively review such trials to examine whether this is the case.

METHODS Oncology trials that used Simon's two-stage design over a 5-year period were identified and reviewed. They were evaluated for whether they reported sufficient design (eg, required sample size) and analysis (eg, CI) details. Articles that did not adjust their inference for the incorporation of an interim analysis were also reanalyzed.

RESULTS Four-hundred twenty-five articles were included. Of these, just 47.5% provided the five components that ensure design reproducibility. Only 1.2% and 2.1% reported an adjusted point estimate or CI, respectively. Just 55.3% provided the final stage rejection bound, indicating many trials did not test a hypothesis for their primary outcome. Trial reanalyses suggested reported point estimates underestimated treatment effects and reported CIs were too narrow.

CONCLUSION Key design details of two-stage single-arm trials are often unreported. Their inference is rarely performed such as to remove bias introduced by the interim analysis. These findings are particular alarming when considered against the growing trend in which nonrandomized trials make up a large proportion of all evidence on a treatment's effectiveness in a rare biomarker-defined patient subgroup. Future studies must improve the way they are analyzed and reported.

JCO Precis Oncol 5:1813-1820. © 2021 by American Society of Clinical Oncology

ASSOCIATED CONTENT

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on November 9, 2021 and published at ascopubs.org/journal/po on December 16, 2021: DOI https://doi.org/10.1200/PO.21.00276

Preprint version available on arXiv.org.

BACKGROUND

For many types of cancers, randomized trials are becoming more common in phase II.[1] However, recent analyses indicate single-arm designs remain most widely used.[2] Additionally, as more cancer studies investigate treatments targeting particular molecular alterations, it is likely single-arm trials will remain commonly used in oncological drug development, given consensus opinion is that rarer subgroups are one area in which a single-arm trial is a logical design.[1]

In single-arm trials, the primary outcome is often dichotomous,[2] typically chosen as objective response[1] through RECIST.[3] Among the available single-arm designs for a binary outcome, Simon's two-stage design[4] is generally preferred.[5] The habitual use of Simon's design has seen much research be conducted into its effective utilization. Recent work includes methodology to account for deviation from the planned design,[6-8] criteria to simultaneously optimize design and analysis,[9] and evaluations of the value of such trials within wider drug development plans.[10] Indeed, many publications have now addressed how to handle issues that can arise in trials using Simon's design. Nonetheless, it is not known to what extent the advice provided has permeated through to practice.

Several authors have evaluated the reporting of phase II oncology trials without differentiating by design. Grellety et al[11] reviewed 156 phase II oncology trials published in 2011, assessing the quality of reporting using two scores. One of these, the Key Methodological Score (KMS), consisted of three items: provision of a clear (1) definition of a criterion of principal judgment, (2) justification for the number of patients included, and (3) definition of the population on which the principal and/or secondary judgment criteria were evaluated. They found the median KMS was 2/3, whereas only 16.1% of the studies had a KMS of 3/3. Langrand-Escure et al[2] reviewed 557 phase II and phase II/III oncology trials published in 2010-2015 in three high-impact journals, also appraising the quality of reporting using the KMS. They concluded just 26.2% of

CONTEXT

Key objective

Accurate reporting of clinical trial design and analysis is critically important for scientific reproducibility. Simon's two-stage is among the most commonly used designs in cancer research. We use 425 recent reports on the results of phase II oncology trials to determine how the cancer community can improve their communication of such trials.

Knowledge generated

Many important features of the design and analysis of included trials were not adequately described in the reports. Efficient design alternatives to the conventional optimal and minimax designs were rarely used. Numerous papers have now been published that help better analyze Simon's two-stage trials, but we found little evidence of their use in practice.

Relevance

Greater care is needed at the design, analysis, and reporting stages of trials using Simon's two-stage design. This may improve knowledge transfer on estimated patient response rates and is particularly relevant, given the growing trend of non-randomized trials for evaluating treatment effectiveness in rare biomarker-defined patient subgroups.

the articles had a KMS of 3/3. They additionally found a sample size calculation was missing in 66% of the articles.

These findings are concerning, but it is possible they only scratch the surface of the issues in the use of two-stage single-arm designs in practice. No paper has sought to ascertain the degree to which precise components of the design of such trials are included in the published reports. Moreover, no research has evaluated the frequency with which trialists have heeded the recommendations of the many articles that argue for the need for the final analysis to be adjusted to account for the interim analysis. Finally, the extent to which deviation from the planned design occurs in practice, or the impact of this on study error rates, is unknown. Given the extent of the use of two-stage single-arm designs in practice, it is paramount such studies be designed, analyzed, and reported effectively. This is particularly true when a confirmatory randomized trial is unlikely to be possible; the single-arm trial then forms the majority of evidence from which important decisions (eg, around licensing) must be made. With little known about the quality of articles on trials that used a two-stage design, we sought to systematically review a large number of such trials to ascertain issues in design, analysis, and reporting.

METHODS

Simon's Two-Stage Design

We review trials that used Simon's two-stage design. Therefore, we briefly summarize the statistical aspects of such trials.

The design evaluates a binary primary outcome, $x_i \in \{0, 1\}$ from patient i, assumed to be distributed as $X_i \sim Bern(P)$ (ie, $\text{Prob}(X_i = 1) = P$). Thus, P is the probability of success for the primary outcome. The following hypothesis is tested: $H_0 : P \leq P_0$, with a type I error rate of $\alpha \in (0, 1)$ when $P = P_0$. The trial is powered to $1 - \beta \in (0, 1)$ when $P = P_1 > P_0$. Here, P_0 and P_1 are commonly referred to as the maximal success probability that does not warrant further investigation and the minimal success probability that allows further investigation. Often, P_0 is based on the historical success probability for the current standard of care.

The design includes a single interim analysis for futility (a no-go decision) and is indexed by a_1, a, n_1, and n. In stage I, outcomes for n_1 patients are accumulated. Then, a_1 serves as a stopping boundary: if $s_{n_1} = \sum_{i=1}^{n_1} x_i \leq a_1$, the trial terminates for futility, with H_0 not rejected. Otherwise, outcomes for $n_2 = n - n_1$ further patients are gathered. Finally, a is used to determine whether to reject H_0: it is rejected if $s_n = \sum_{i=1}^{n} x_i > a$ and not rejected otherwise. The design parameters, a_1, a, n_1, and n, are chosen to minimize optimality criteria, among the combinations that meet the type I error and power requirements. Simon[4] suggested two optimality criteria: (1) null-optimal, to minimize the expected sample size when $P = P_0$, and (2) minimax, to minimize the maximal sample size n. Other optimality criteria have since been proposed.[12-14]

Post-trial inference could be performed using methods developed for one-sample proportions, eg, a CI could be computed using Clopper-Pearson.[15] Depending on the stage of termination, a point estimate for P could be given as $\hat{P} = s_{n_1}/n_1$ or s_n/n (these are sometimes referred to as naïve estimates within the context of a Simon two-stage trial). However, it is well-known that the inclusion of an interim analysis means that adjusted inference should be performed. This is to ensure computed P values are consistent with the decision on whether to reject H_0, which acquired CIs have the desired coverage, and to reduce point estimate bias.[16] Many adjusted methods have been proposed, including that of Jung et al[17] for P values, Jennison and Turnbull[18] for CIs, and Jung and Kim[19] for point estimates. Several methods for handling deviation from the planned design (ie, scenarios in which the interim or final analysis is conducted with a sample size different from n_1 or n) have also been developed.[6-8,20]

We provide extended details of all methods used later in the Data Supplement (online only). Here, we focus on providing more details of a particular method for computing an adjusted point estimate, which will be used at length later. As noted above, several methods have been proposed for estimating P in a Simon two-stage trial. Each essentially aims to reduce the bias in the estimate. Informally, bias can be thought of as expecting to, on average, incorrectly estimate P. The reason multiple methods have been developed is that no one approach is clearly best.[16] However, some believe the uniform minimum variance unbiased estimator (UMVUE) should be preferred.[16,19] This has the lowest variance among estimators that are always unbiased; low variance is useful as it means that, on average, our estimate $\hat{P}$ should be closer to P. The UMVUE has a more complex form than the naïve estimates given above (Data Supplement) but is still easy to calculate.

Literature Review

See the Data Supplement for further details.

Inclusion criteria. To identify articles, PubMed was searched on February 21, 2018, using the term ("2013/01/01"[Date - Publication]: "2017/12/31"[Date - Publication]) AND Clinical Trial[Publication Type] AND (phase II[Title/Abstract] OR phase 2[Title/Abstract]) AND (cancer[All Fields] OR oncology [All Fields]), returning 5,344 articles for review.

The key inclusion criteria were (1) full-length articles, (2) primary publications on a trial's complete results, and (3) report results for at least one treatment arm that used Simon's two-stage design.

Next, 534 articles (10.0%) were randomly selected for evaluation for inclusion by M.J.G. and A.P.M., with a 10.0% duplicate extraction used to ensure agreement on inclusion could be precisely estimated. The authors agreed on inclusion for 520 articles (97.3%). Given the high-level of agreement, the remaining articles were assessed for inclusion by M.J.G. only, with discussion with A.P.M. where required.

Data extraction. Data on each of the questions listed in Data Supplement were extracted by M.J.G. for each arm, in each article, deemed eligible for inclusion. To establish the reliability of this extraction, data extracted by M.J.G. were compared with those independently extracted on 58 arms by A.P.M. across 14 questions requiring nonbinary value extraction (eg, "Q5. What was the value of P_0?"), the duplicate extractions agreed 96.2% of the time. Across a wider set of 26 questions, including those requiring only binary value extraction, the duplicate extractions agreed 94.3% of the time.

Trial reanalyses. Reanalyses of included articles were conducted to evaluate the possible impact of not using adjusted inferential procedures. The UMVUE (which as discussed may be preferred because of its unbiasedness) was compared with reported naïve point estimates to measure the potential degree of over or underestimation in practice compared with a best practice analysis. We compared the estimated coverage of computed adjusted CIs with those of reported unadjusted CIs to determine whether CIs may be attaining the desired coverage.

Given the absence of evidence is not evidence of absence, the included articles that did not state they reported an adjusted point estimate (Q25) were also reanalyzed (subject to reporting required design components) to evaluate which of seven possible point estimates the reported point estimate (Q26) was consistent with, to the reported number of decimal places. Equivalent computations were conducted for those articles that did not state they reported an adjusted CI (Q30); the reported CI (Q32-34) was compared for consistency with four unadjusted and two adjusted CIs.

Reanalyses were limited to those trials (1) adjudged to have terminated in stage II, as point estimate and CI procedures do not, in general, adjust when a trial terminates in stage I and (2) that reported the number of successes and sample size assumed in the analysis, as these are required to calculate unadjusted point estimates and CIs. To reanalyze using adjusted inferential procedures, a_1 and n_1 must have been reported.

RESULTS

Included Articles

Five hundred articles were deemed eligible for inclusion, with 425 reporting the results of a single eligible treatment arm. The remaining 75 articles reported the results for an additional 204 eligible arms (arms per article: median 2, range [2-15]).

To remove the need to account for skew caused by the quality of articles reporting multiple included treatment arms, we discuss here the findings for only the 425 articles that reported the results of a single eligible treatment arm. Findings for the remaining 75 articles are given in the Data Supplement.

Table 1 provides descriptors on the 425 articles. At least 15.8% of the articles came from each allowed publication year, with included articles being published in 100 journals and considering a wide variety of cancer types.

One hundred ten trials (25.9%) were judged to have terminated in stage I and 298 in stage II (70.1%). Among the 298 judged to have terminated in stage II, only 80 (26.4%) stated the criteria had been met for progression to stage II, indicating this judgment often had to be based on the enrolled sample size. For 17 articles (4.0%), it was not possible to ascertain when the trial terminated; this

TABLE 1. Descriptors on the 425 Included Articles That Reported the Results of a Single Eligible Treatment Arm

Descriptor	Value	No. (%)
Publication year	2013	102 (24.0)
	2014	101 (23.8)
	2015	79 (18.6)
	2016	76 (17.9)
	2017	67 (15.8)
Journal	Cancer Chemother Pharmacol	44 (10.4)
	Ann Oncol	30 (7.1)
	Invest New Drugs	23 (5.4)
	Lung Cancer	17 (4.0)
	Cancer	15 (3.5)
	BMC Cancer	13 (3.1)
	J Clin Oncol	13 (3.1)
	Br J Haematol	11 (2.6)
	Lancet Oncol	11 (2.6)
	Other (91 journals)	248 (58.4)
Cancer	Lung	59 (13.9)
	Lymph	53 (12.5)
	Colon	32 (7.5)
	Breast	28 (6.6)
	Stomach	25 (5.9)
	Head and neck	23 (5.4)
	Blood	22 (5.2)
	Kidney	21 (4.9)
	Other	162 (38.1)
Stage of termination	I	110 (25.9)
	II: Stated the criteria had been met for progression	80 (18.8)
	II: Did not state the criteria had been met for progression	218 (51.3)
	Unclear	17 (4.0)

NOTE. The denominator for computing percentages (given to 1 decimal place) is 425 in all instances.

TABLE 2. Reporting of the Design of the 425 Included Articles That Reported the Results of a Single Eligible Treatment Arm

Criteria	No. (%)
Used the phrase Simon two-stage (or similar) or cited Simon (1989)[4]	357 (84.0)
Clearly stated P_0	380 (89.4)
Gave a justification for P_0	78 (18.4)
Citation given	40 (9.4)
Justification given but no citation	38 (8.9)
Clearly stated P_1	391 (92.0)
Clearly stated α	372 (87.5)
$\alpha = .05$	231 (54.4)
$\alpha = .1$	103 (24.2)
Clearly stated β	382 (89.9)
$\beta = .1$	165 (38.8)
$\beta = .2$	173 (40.7)
Clearly stated the optimality criteria	240 (56.5)
Null-optimal	142 (33.4)
Minimax	93 (21.9)
Admissable	4 (0.9)
Other	1 (0.2)
Clearly stated a_1	349 (82.1)
Clearly stated a	235 (55.3)
Clearly stated n_1	371 (87.3)
Clearly stated n	394 (92.7)
Indicated the recruitment target was greater than n	117 (27.5)
Clearly stated P_0 and P_1	373 (87.8)
Clearly stated P_0, P_1, α, and β	340 (80.0)
Clearly stated a_1, a, n_1, and n	221 (52.0)
Clearly stated P_0, P_1, α, β, and the optimality criteria	202 (47.5)
Clearly stated P_0, P_1, α, β, the optimality criteria, a_1, a, n_1, and n	109 (25.6)

NOTE. The denominator for computing percentages (given to 1 decimal place) is 425 in all instances.

was typically caused by neither of the planned stagewise sample sizes being reported.

Reporting of Design Characteristics

Table 2 summarizes extracted data on reporting of design characteristics. Although 380 articles (89.4%) clearly stated P_0, only 78 (18.4%) provided a justification for its value. The probability P_1 was often reported (391 articles; 92.0%), as were the desired type I (372 articles; 87.5%) and type II error rates (382 articles; 89.9%). The chosen optimality criteria were stated in only 240 articles (56.5%). This drives the fact that only 202 articles (47.5%) reported P_0, P_1, α, β, and the optimality criteria, the five components that ensure easy design reproduction. Although a_1 (349 articles; 82.1%), n_1 (371 articles; 87.3%), and n (394 articles; 92.7%) were all regularly reported, a was given in only 235 articles (55.3%).

Reporting of Inferential Procedures

Table 3 summarizes extracted data on the reporting of inferential procedures. Although point estimates were often reported (372 articles; 87.8%), only five articles (1.2%) stated they had reported an adjusted point estimate. In contrast, P values were rarely reported (four articles; 1.3%). For CIs, just 233 articles (54.8%) reported a CI, with only nine (2.1%) indicating they reported an adjusted CI. All trials that stated they had reported an adjusted point estimate or CI were ones judged to have terminated in stage II; we return to this point in the Discussion.

To evaluate whether articles that reported a point estimate or CI but did not indicate it was adjusted were consistent (to

TABLE 3. Reporting of Inferential Procedures Performed in the 425 Included Articles That Reported the Results of a Single Eligible Treatment Arm, With Additional Stratification by Stage of Termination

Criteria	Stage I, No. (%)	Stage II, No. (%)	All, No. (%)
Reported a point estimate, *P* value, or CI for the primary outcome	72 (65.5)	287 (96.3)	375 (88.2)
Reported a point estimate	70 (63.6)	287 (96.3)	372 (87.8)
Stated the point estimate had been adjusted for the two-stage design	0 (0)	5 (1.7)	5 (1.2)
Reported a *P* value	0 (0)	4 (1.3)	4 (0.9)
Stated the *P* value had been adjusted for the two-stage design	0 (0)	3 (1.0)	3 (0.7)
Reported a CI	40 (36.4)	187 (62.8)	233 (54.8)
Stated the CI had been adjusted for the two-stage design	0 (0)	9 (3.0)	9 (2.1)
Analysis performed assuming a sample equal to that given in the design	27/70 (38.6)	72/278 (25.9)	99/348 (28.4)

NOTE. The denominators for computing percentages (given to 1 decimal place) in the three columns are 110, 298, and 425, respectively, unless stated otherwise.

their reported number of decimal places) with unadjusted or adjusted analyses, the trials were reanalyzed (Table 4). Two hundred seventy (96.1%) reanalyzed articles reported point estimates consistent with an unadjusted estimate. However, 133 of 228 articles (58.3%) for which adjusted point estimates could be calculated were also consistent with at least one adjusted estimate. For the CIs, 116 of 178 reanalyzed articles (65.2%) were consistent with at least one unadjusted interval. Far fewer articles (3/140; 2.1%) for which adjusted CIs could be computed were consistent with an adjusted CI.

To visualize the impact of not using adjusted inferential procedures, Figure 1A displays the unadjusted estimate ($\hat{P}_{\text{naive}}$) against the UMVUE ($\hat{P}_{\text{umvue}}$) for the 233 trials that terminated in stage II where the UMVUE could be computed. The difference between $\hat{P}_{\text{umvue}}$ and $\hat{P}_{\text{naive}}$ is presented as a percentage of $P_1 - P_0$ in Figure 1B. These plots indicate that although the difference between the unadjusted and adjusted estimates may often be small, there are instances in which it is large, in 25 cases more than 25% of the difference $P_1 - P_0$. Furthermore, there were 103 trials for which P_1 was reported, and a UMVUE was computable, in which $\hat{P}_{\text{naive}} < P_1$. Potentially significantly, among these, 4.9% (5/103) trials had $\hat{P}_{\text{umvue}} \geq P_1$ (i.e., the estimated response rate changed from below to above P_1 following adjustment).

Similar visualizations are provided in Figure 2. Figure 2A displays the length of the reported unadjusted CI against the length of the corresponding adjusted CI proposed by Jennison and Turnbull[18] for the 140 trials for which this adjusted CI could be computed. Figure 2B compares the respective coverage of these unadjusted and adjusted CIs when $P = \hat{P}_{\text{umvue}}$ for the 131 trials in which the target coverage was 0.95. In general, the length of the unadjusted CI is shorter than the corresponding adjusted CI, which is reflected in the coverage being below the desired level for the unadjusted procedure in several instances. The adjusted CI procedure guarantees coverage of at least 0.95, but the cost of this is coverage sometimes far above that required.

Note that 348 trials that were judged to have ended in stage I or stage II reported a point estimate, *P* value, or CI, as well as the sample size required by their design. Among these, just 99 (28.4%) performed their analysis using the planned sample size. Differences between planned and analyzed sample sizes are shown in the Data Supplement.

DISCUSSION

A large proportion of all phase II evidence comes from trials using Simon's two-stage design. In addition, for trials in rare molecularly defined patient subgroups, it may often be the case that such two-stage single-arm trials will provide the

TABLE 4. Reanalysis of the Subset of the 425 Articles That Reported a Point Estimate or CI Not Stated to Have Been Adjusted

Criteria	No. (%)
Reported a point estimate not stated as adjusted and clearly reported the number of successes and sample size assumed in the analysis	281/298 (94.2)
Reported point estimate consistent with an unadjusted estimate	270/281 (96.1)
Reported point estimate consistent with at least one adjusted estimate	133/228 (58.3)
Reported a CI not stated as adjusted and clearly reported its level, the number of successes, and sample size assumed in the analysis	178/298 (59.7)
Reported CI consistent with at least one unadjusted interval	116/178 (65.2)
Reported CI consistent with at least one adjusted interval	3/140 (2.1)

NOTE. Consistency is measured in all cases against the reported number of decimal places. The denominators for computing percentages (given to 1 decimal place) are given in each row. Note that the reanalysis is limited to those articles that were judged to have terminated in stage II.

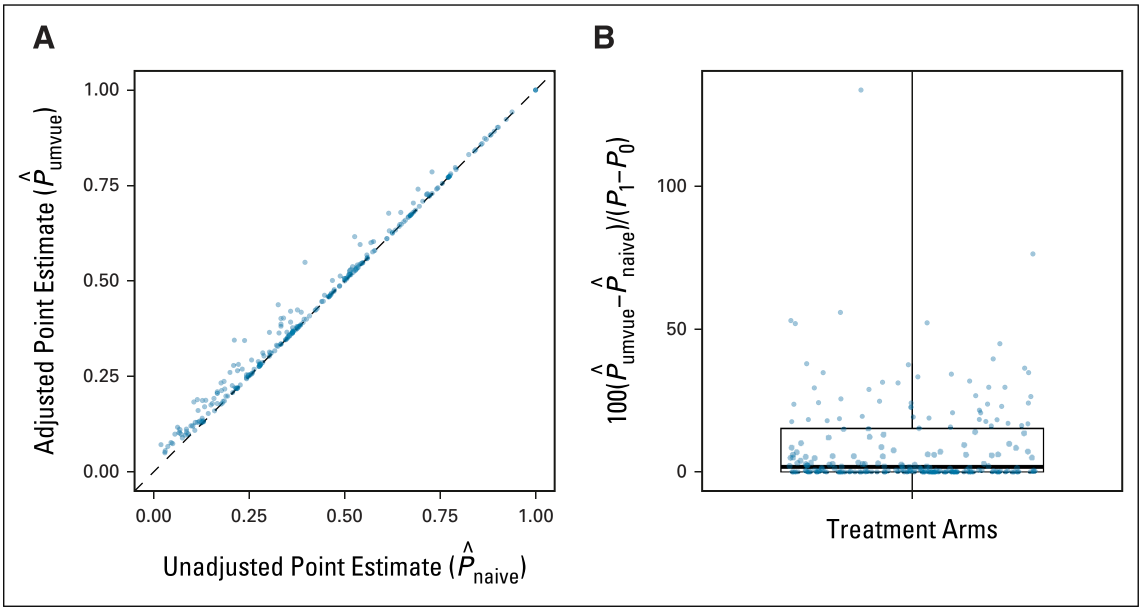

FIG 1. Point estimate comparison. (A) A comparison of the naive unadjusted point estimate ($\hat{P}_{naive}$) and the UMVUE ($\hat{P}_{umvue}$) is given for the 233 trials that terminated in stage II where the UMVUE could be computed. (B) The difference between $\hat{P}_{naive}$ and $\hat{P}_{umvue}$ is presented as a percentage of $P_1 - P_0$ (where P_0 and P_1 are the maximal success probability that does not warrant further investigation and the minimal success probability that allows further investigation, specified at the design stage), along with a boxplot to indicate the distribution of these data. UMVUE, uniform minimum variance unbiased estimator.

majority of evidence ever available on a treatment's efficacy. This necessitates that such studies be designed, analyzed, and reported effectively. We evaluated the degree to which this is true through a comprehensive review.

It is easy to argue reporting of design components was extremely poor. Reproducibility of designs is limited by infrequent reporting of P_0, P_1, α, and β in unison. It is alarming only 18.4% of the trials provided a justification for P_0, considering result interpretation is highly dependent on this value. It may be considered disappointing that most trials chose standard error rates (e.g., $\alpha = .05$), as it has been highlighted small concessions in this regard can lead to notable efficiency gains.[21] Similar statements are true for the optimality criteria.[13,22]

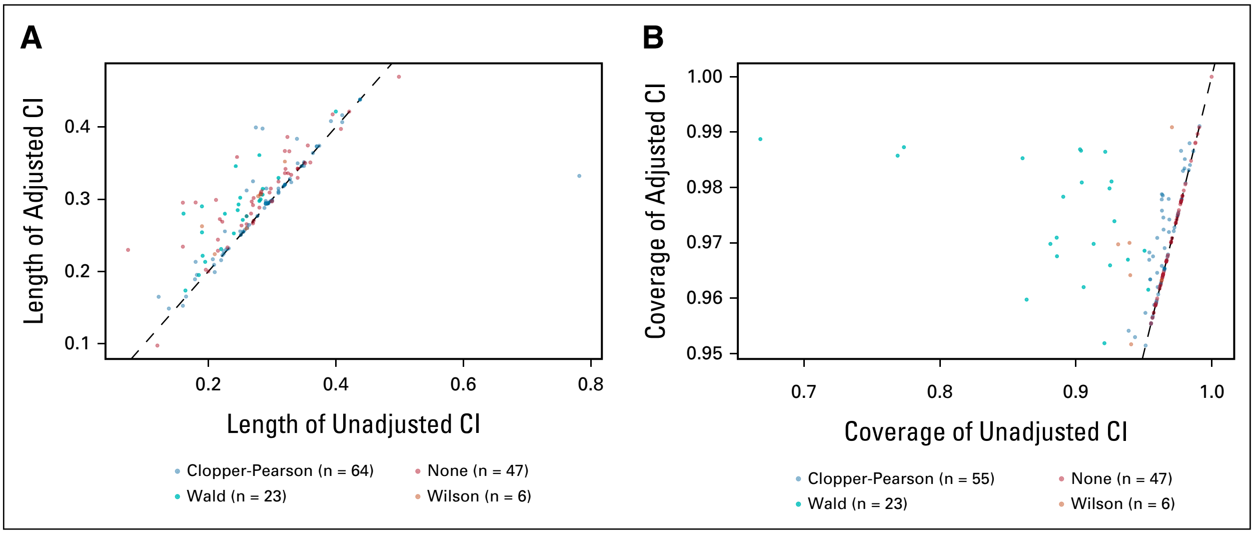

FIG 2. CI comparison. (A) The length of the reported unadjusted CI is compared with the length of the corresponding adjusted CI proposed by Jennison and Turnbull for the 140 articles for which this adjusted CI could be computed. (B) The respective coverage of these unadjusted and adjusted CIs when $P = \hat{P}_{umvue}$ is given for the 131 of these articles in which the target coverage was 0.95. In both cases, points are colored by the unadjusted CI that the reanalysis indicated the reported CI matched with. For those CIs that matched none of the unadjusted CIs, Clopper-Pearson was used to compute the coverage.

Few articles stated they used adjusted inference. Given there is no additional cost to using these methods, this is disappointing. Figures 1 and 2 indicate the result of this may be that trials were conservative in their reported point estimate, but anticonservative in the width of their CI. It is also concerning that only 54.8% of the articles included a CI, given the size of single-arm trials makes uncertainty around a point estimate important to quantify.

Many final analyses were performed with a sample size different from that specified in the design (71.6%). This highlights the need to plan for design deviation and echoes previous findings.[23] We initially hoped to extract data on how trials handled design deviation when interpreting their results. This was unfortunately judged to be too subjective an endeavor, as many studies interpreted findings through informal comparison of their point estimate or CI bounds to P_0 and/or P_1.

Difficulties in attaining the planned sample size may be reflected in only 55.3% of the trials reporting a. Lack of reporting of a also indicates many trials that use Simon's design do not formally test the hypothesis they claim to. It is troubling that so many trials are being published without a formal statistical test being conducted for their primary outcome. We note that methodology to comprehensively handle design deviation is available; its use is depicted in the Data Supplement, which provides the error rates for 45 trials when the methodology of Englert and Kieser[7] is implemented. Using this methodology, trials are assured to conform to their desired type I error rate, and it appears sample sizes that enable power to reach close to the desired level may have been achieved in practice. Without using methodology to account for design deviation, many trials may be interpreting their findings in a manner associated with a high probability of erroneous decision making.

We acknowledge several limitations to our work. Only a 10% duplicate extraction was performed. Given the strength of our findings, though, it is unlikely our conclusions would be altered by additional duplicate extractions. It is also impossible to be certain those trials that did not state they used an adjusted inferential procedure had used an unadjusted method. Our reanalyses (Table 4) provide evidence this may be the case. However, for trials that terminated in stage I, we cannot know whether a plan to use adjusted inference if the trial had continued to stage II went unreported.

Given past work assessing adherence to CONSORT recommendations,[24] our findings should perhaps not be surprising. Nonetheless, it may have been hoped the simplicity of Simon's design would lead to effective reporting. Our results indicate a CONSORT extension for single-arm oncology trials may be warranted.

AFFILIATIONS

[1]Population Health Sciences Institute, Newcastle University, Newcastle, United Kingdom
[2]Centre for Trials Research, Cardiff University, Cardiff, United Kingdom

Preprint version available on arXiv.org.

CORRESPONDING AUTHOR

Michael J. Grayling, PhD, Population Health Sciences Institute, Baddiley-Clark Building, Richardson Rd, Newcastle upon Tyne NE2 4AX, United Kingdom; e-mail: michael.grayling@newcastle.ac.uk.

PRIOR PRESENTATION

Presented at the 5th International Clinical Trials Methodology Conference, Brighton, United Kingdom, October 6-9, 2019 (oral presentation). See Grayling MJ, Mander AP: PS3—O2 Two-stage single-arm oncology trials: More adjusted analyses needed. Trials 20:579, 2019.

AUTHOR CONTRIBUTIONS

Conception and design: All authors
Collection and assembly of data: All authors
Data analysis and interpretation: Michael J. Grayling
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/po/author-center.
Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

No potential conflicts of interest were reported.

REFERENCES

1. Grayling M, Dimairo M, Mander A, et al: A review of perspectives on the use of randomization in phase II oncology trials. J Natl Cancer Inst 111:1255-1262, 2019
2. Langrand-Escure J, Rivoirard R, Oriol M, et al: Quality of reporting in oncology phase II trials: A 5-year assessment through systematic review. PLoS One 12:e0185536, 2017
3. Eisenhauer E, Therasse P, Bogaerts J, et al: New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer 45:228-247, 2009
4. Simon R: Optimal two-stage designs for phase II clinical trials. Control Clin Trials 10:1-10, 1989

5. Ivanova A, Paul B, Marchenko O, et al: Nine-year change in statistical design, profile, and success rates of phase II oncology trials. J Biopharm Stat 26:141-149, 2016
6. Belin L, Broet P, De Rycke Y: A rescue strategy for handling unevaluable patients in Simon's two stage design. PLoS One 10:e0137586, 2015
7. Englert S, Kieser M: Methods for proper handling of overrunning and underrunning in phase II designs for oncology trials. Stat Med 34:2128-2137, 2015
8. Zhao J, Yu M, Feng X: Statistical inference for extended or shortened phase II studies based on Simon's two-stage designs. BMC Med Res Methodol 15:48, 2015
9. Bowden J, Wason J: Identifying combined design and analysis procedures in two-stage trials with a binary end point. Stat Med 31:3874-3884, 2012
10. Grayling M, Mander A: Do single-arm trials have a role in drug development plans incorporating randomised trials? Pharm Stat 15:143-151, 2016
11. Grellety T, Petit-Moneger A, Diallo A, et al: Quality of reporting of phase II trials: A focus on highly ranked oncology journals. Ann Oncol 25:536-541, 2014
12. Hanfelt J, Slack R, Gehan E: A modification of Simon's optimal design for phase II trials when the criterion is median sample size. Control Clin Trials 20:555-566, 1999
13. Jung S, Lee T, Kim K, et al: Admissible two-stage designs for phase II cancer clinical trials. Stat Med 23:561-569, 2004
14. Mander A, Thompson S: Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. Contemp Clin Trials 31:572-578, 2010
15. Clopper C, Pearson E: The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26:404-413, 1934
16. Porcher R, Desseaux K: What inference for two-stage phase II trials? BMC Med Res Methodol 12:117, 2012
17. Jung S, Owzar K, George S, et al: p-value calculation for multistage phase II cancer clinical trials. J Biopharm Stat 16:765-775, 2006
18. Jennison C, Turnbull B: Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. Technometrics 25:49-58, 1983
19. Jung S, Kim K: On the estimation of the binomial probability in multistage clinical trials. Stat Med 23:881-896, 2004
20. Wu Y, Shih W: Approaches to handling data when a phase II trial deviates from the pre-specified Simon's two-stage design. Stat Med 27:6190-6208, 2008
21. Khan I, Sarker SJ, Hackshaw A: Smaller sample sizes for phase II trials based on exact tests with actual error rates by trading-off their nominal levels of significance and power. Br J Cancer 107:1801-1809, 2012
22. Mander A, Wason J, Sweeting M, et al: Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. Pharm Stat 11:91-96, 2012
23. Koyama T, Chen H: Proper inference from Simon's two-stage designs. Stat Med 27:3145-3154, 2008
24. Turner L, Shamseer L, Altman D, et al: Does use of the CONSORT statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. Syst Rev 1:60, 2012