EBMT

EDITORIAL

Clinical trials: design, endpoints and interpretation of outcomes

Bone Marrow Transplantation (2022) 57:338–342; https://doi.org/10.1038/s41409-021-01542-0

SERIES EDITORS INTRODUCTION

The ability to properly analyze results of clinical trials, especially randomized controlled trials (RCT), is a needed skill for every physician. This is especially so for those involved in haematopoietic cell transplants. Although seemingly straightforward, correct interpretation of clinical trials data is in reality complex and not for the fainthearted. When a RCT reports intervention A is safer and more effective than intervention B do we simply accept the authours' conclusion or is more detective work needed. The answer: call in Inspector Clouseau! In this article Prof. Megan Othus and us discuss complexities in clinical trials interpretation including the challenge of false-positive error control, endpoints, power and sample size estimates (more often guesses), how to analyze competing events such as graft-*versus*-host disease (GvHD) and relapse, what to do when a study has > 1 primary endpoint, analyses of multi-arm trials, how to interpret analyses other than the primary endpoint and what do data from non-inferiority trials tell us. Lastly, we consider, the evil which will not die (the statistical Rasputin): reporting survival outcomes by response. We hope this article will be of practical use to clinicians facing the challenge of correctly interpreting clinical trials data. The good news: only one relatively simple equation. And remember, we can be reached 24/7 on Twitter #BMTStats. Our operators are standing by.

Robert Peter Gale MD, PhD, DSc(hc), FACP, FRCP, FRCPI(hon), LHD, DPS Mei-Jie Zhang PhD

INTRODUCTION

There are those who reason well, but they are greatly outnumbered by those who reason badly
Galileo Galilei

Clinical trials, especially randomized controlled trials, are typically designed to facilitate straightforward interpretation [1]. However, despite randomization, a formal protocol document and clinical trials registries such as clinicaltrials.gov, it remains challenging to appropriately evaluate reports of clinical trials. Herein, we review several issues regarding critical interpretation of clinical trial results.

FALSE-POSITIVE ERRORS

The first topic to discuss is false-positive errors, also called α (alpha) errors. Many potentially convoluted choices are made in the design and presentation of clinical trials data to the end of trying to "control" or hold the false-positive error rate below a specified threshold. Every statistical analysis reporting a p value (or confidence interval, though we will focus on p values for simplicity) and which interprets this value as "significant" (at or below some threshold, often $p < 0.05$) or "without evidence of significance" or "not significant" (above this threshold) is potentially subject to an incorrect conclusion (summarized in Table 1).

When reporting a p value we can only comment on whether it is "statistically significant" or not. We do not know whether this conclusion is correct or not. But by thoughtful construction of the test and calculations used to derive the p value we can quantify the probability of error. Over many years conventions in clinical research (partly driven by regulatory agency standards which themselves might be driven by legislation) have converged on some typical error rates in clinical trials. In most trials the false-negative error rate is typically selected to be 10–20% resulting in a power of 80–90%. False positive error rates in phase-3 trials are typically controlled to be <5% [2–6] or even <2.5% [7]. Randomized phase-2 trials often "relax" the false-positive rate to 10–20% [8–10].

Any one analysis or test has an associated false-positive rate. If more than one test is done, each test has its own rate and we can quantify the overall false-positive rate (the rate of having ≥1 test with a false-positive conclusion). The overall false-positive rate is related to numbers of tests done and false-positive rate of each test. If each test uses the same false positive rate (α), we can write the overall false positive rate as:

$$\text{Probability of making} \geq 1 \text{ false-positive conclusion} = 1 - (1 - \alpha)^{\text{number of tests}}.$$

If we take the common $\alpha = 5\%$ (0.05) then with two tests the overall false positive rate is 9.75%, with 10 tests, 40%, with 20 tests, 64% and with 50 tests, 92%. In an analysis reporting many p values each interpreted individually as significant or not the probability of a false-positive conclusion quickly becomes high. This is why false-positive error control is a major concern in clinical trials design.

If a trial has only one primary endpoint and only one analysis of that endpoint is done the alpha level for that one test will match the overall false positive rate for the trial. However, many clinical trials pre-specify ≥1 endpoints and/or ≥1 analyses. Moreover, often a variety of manipulations are used to control for the false-positive rate across analyses. False-positive errors are further discussed below as different topics intersect with error control in clinical trials.

Received: 8 November 2021 Revised: 12 November 2021 Accepted: 22 November 2021
Published online: 7 January 2022

Table 1. Statistical error summary.

	Reality—drug does not improve outcomes	Reality—drug improves outcomes
Statistical test—significant association between drug and outcome	False-positive error; α (alpha)	Power; equal to 1-false negative error

Table 2. Common types of data used in transplant clinical trial endpoints.

Type of data	Example endpoints	Example statistics	Example statistical tests
Time-to-event	• Survival (time to death) • Event-free survival (time to first of events such as death or relapse or treatment failure) • Time to acute graft-versus-host disease	• Kaplan–Meier estimate of X-months or Y-years survival • Cumulative incidence at X-months or Y-years • Hazard ratio	• Log-rank test • Cox regression model • Gray test of cumulative incidence • Cause-specific regression model • Fine and Gray sub-distribution hazards model
Categorical	Response (typically defined by consensus guidelines)	• Proportions	• Chi-square test • Fisher exact test
Binary (categorical with only two categories)	Complete remission (yes/no)	• Proportions	• Chi-squared test • Fisher's exact test • Logistic regression
Quantitative	Quality-of-life[a], Gene expression	• Mean • Median • Correlation	• Student *t*-test • Wilcoxon test • Linear regression

[a]Some are quantitative, others, categorical.

ENDPOINTS

Endpoints are measures which can be observed or calculated for each subject on a trial. Often these measures are combined together mathematically in various ways to estimate a statistic. All statistics have associated measures of uncertainty. The combination of the statistic and the measure of uncertainty can be used to calculate confidence intervals and *p* values which we typically use to interpret clinical trials results. There are many possible endpoints but those commonly used in clinical trials of haematopoietic cell transplants are summarized in Table 2.

Censoring is what distinguishes time-to-event from quantitative endpoints. Quantitative endpoints should be measurable or observed on every subject in a clinical trial whereas time-to-event endpoints may not. For example, if a clinical trial collects data on subjects for 5 years after study-entry and a subject does not die during that interval the trial will not observe the *time-to-death* for that subject. We know this subject lived ≥5 years and this can be used to evaluate and estimate survival up to 5 years. After 5 years the subject cannot contribute data for estimating or quantifying survival and they are termed censored. Different statistical analyses are needed for time-to-event versus quantitative data to account for censoring.

Regression analyses are an important element of randomized trials analyses even when the primary analysis is not based on regression models. Regression models provide estimates of effect sizes (e.g., odds ratios or hazard ratios), which are important when interpreting the results of trials. In addition, regression analyses allow for adjustments for co-variates not used in randomization stratification. Although randomization is likely to balance most factors across arms it does not guarantee balance without stratification. Regression analyses can allow for more precise estimation of effect sizes when there is an imbalance in a prognostic co-variate across arms.

Power, sample size, and endpoints

In a clinical trial protocol the sample size should have the associated power reported (typically 80–90%). For categorical and quantitative co-variates power is directly related to numbers of subjects enrolled onto the trial. For *time-to-event* endpoints power is driven by the number of "events" (e.g., for survival, the event is death; for *time-to-relaps*e the event is relapse). Numbers of events are driven by the rate at which they occur, the interval subjects were accrued and how long each subject was observed since study-entry. Typically, clinical trials with *time-to-event* endpoints specify analyses will be done after a specified number of events are observed. When developing a protocol best efforts are made at making reasonable assumptions (guesses is often a more accurate descriptor) at how soon the event(s) under consideration will be observed. But if the assumptions are wrong for any reason the timing calculated in the protocol will be incorrect and analyses may be done sooner or later than pre-specified. The issues with post-hoc or retrospective power calculations have been well-described elsewhere, but in short, such calculations are not appropriate and should rarely (potentially never) be performed [11].

Competing events

When a subject can experience >1 event (say relapse and death) and the clinical trial is only interested in the time to one of those events, say *time-to-relaps*e, the other event is called a "competing event." For most *time-to-event* endpoints like relapse, death before relapse is a competing event. For example, in a *time-to-relapse* analysis if a subject dies without relapse we cannot assume they would never have relapsed had they not died. But the subject is also not just censored at time of death as one would do in a survival endpoint analysis because there may be a non-random relationship between death and relapse. For example, there exists a correlation between severity of GvHD and relapse risk (reviewed in Horowitz et al. [12]). To account for this possibility different analyses are needed to analyze such time-to-event endpoints. The Kaplan–Meier method should not be used [13]. Instead cumulative incidence rates should be estimated [14–17]. Log-rank tests should not be used but rather alternative tests which account for competing risks [18–22].

Multiple primary endpoints

It is increasingly common for clinical trials to specify >1 primary endpoint [3, 4]. Why? Clinical trials are expensive and time-consuming and it can be disappointing to complete a trial and

conclude there was no benefit in the investigational cohort because the wrong endpoint was specified. To mitigate this concern multiple primary endpoints can be specified before the study begins. However, as we discuss above, testing >1 endpoint "inflates" or increases the overall false-positive rate above the false positive (α) rate for each test.

There are several strategies to evaluate >1 endpoint. In order to interpret a trial as "positive" if ≥1 endpoint is significant, the α should be "split" (allocated) across endpoints. The split can be done evenly; for example for a trial with overall α of 5% and two primary endpoints, each could be tested with an α of 2.5% [3]. However, the split need not be even. For example, a trial could allocate 4% of the α to the 1st endpoint and use the formula (1-0.04)*$(1 - \alpha_2) = 1 - \alpha = 0.95$ to calculate that $\alpha_2 = 0.0104$ and allocate 1.04% to a 2nd endpoint. Again this must be done before the trial starts. The gain from using this formula versus a simple split of 4 and 1% is small enough such that many trials simply use the simple split [4]. Alpha can also be split between cohorts or sub-cohorts [23]. For example, 4% alpha could be allocated to a survival analysis of amongst all subjects in a trial, with the remaining alpha allocated to a biomarker-positive cohort, say a cohort which has a *FLT3* mutation in a trial of midostaurin. The "remaining alpha" could be set at 1% but because the biomarker-positive cohort is included in the analysis of the full trial population, the results of the analyses are not independent. Because the analyses are not independent we can test the biomarker-positive cohort at an α level >1% and still control the overall α level at 5%. The correlation depends on the proportion of all events observed in the biomarker-positive cohort. Formulae for this calculation can be implemented in statistical programmes [24].

An alternative to this α splitting is a fixed-sequence approach. The sequence of tests is pre-specified and each endpoint is tested at the same α level. Testing continues along the sequence until there is a test with a p value > α, at which point testing stops and no further endpoints in the sequence should be evaluated. Sometimes these tests are described as "carrying forward" the alpha after a significant test. All of the α is "spent" at the first test with p value > α [25].

A combination of α splitting and fixed sequence testing can also be done. As numbers of endpoints increases the numbers of ways to allocate the α across endpoints also increases. The specific α allocation can vary between trials.

It is uncommon in transplant studies to have >1 primary endpoint or to require all primary endpoints to be significant in a trial to declare success [26]. For example, for a design to require a significant association with complete remission and also with survival. The false-positive rate is not inflated in this design because there is only one way to have a positive trial, i.e., in a trial requiring all endpoints to be significantly associated with intervention, each endpoint can be tested at the same α level. Because these designs have increased false-negative error rates compared with designs with a one primary endpoint they have less power and require larger samples.

A single composite endpoint including multiple potential "events" is not uncommon across transplant studies. For example, the endpoint GvHD-relapse-free-survival (GRFS) measures the time until the first event: GvHD, relapse or death. GRFS and similar composite endpoints weight the contributory events equally. If equal weighting of these endpoints is not appropriate, alternative statistics can be used to compare arms in a trial including the win ratio [27] which evaluates composite endpoints in a fixed hierarchy between matched pairs of subjects and tallies in how many pairs the experimental therapy dies first. If neither subject in the pair dies the second event is compared and so forth. Confidence intervals and p values can be calculated for the win ratio like other statistics. Win ratios can be calculated for individual events and composite lists of events and compared to understand the role each event has in the composite win ratio (see Fig. 2 of Pocock et al. [27], for an example). We note that acute GvHD alone or as a component of a composite endpoint is problematic because of the lack of definitive diagnostic criteria with substantial inter-observer discordances. Consequently, a clinical trial with acute GvHD as the primary endpoint (either alone or within a composite endpoint) is only definitive when a masked (blinded) randomized design is used.

Multi-arm clinical trials. Multi-arm clinical trials are an efficient way to conduct >1 investigation/evaluation within a protocol. Multi-arm trials can have increased false-positive error rates like trials with >1 primary endpoint because of multiple comparisons. Strategies like those discussed above can be used to control the false-positive rate (e.g., α splitting; fixed-sequence tests). Some multi-arm trials choose not to control α and use the same α for each comparison. When all comparisons in a multi-arm trial are reported in one report it is straightforward to count numbers of comparisons and calculate the overall false-positive rate [28]. However, it is unfortunately common for multi-arm trials to report each comparison in separate publications [29, 30]. As such, readers need to be aware of the general design when reading and interpreting results of only one comparison within a multi-arm trial.

When comparing two or more interventions added to a backbone, sometimes placebo if there is no *standard-of-care*, factorial designs can be used to evaluate potential synergy or interactions between the interventions. For example, a multi-arm study of two therapies designated X and Y added to a backbone designated B could have four arms: X + B, Y + B, X + Y + B, and B [31]. This design allows quantification of the "interaction" between X and Y, namely, are the therapies better or worse together or do they have individual benefits which are additive? [32, 33]. These designs are uniquely able to evaluate multiple therapies in this way but can quickly become large and expensive. Some factorial designs assume X and Y are "independent" in the sense that any benefit of X can be evaluated ignoring whether a subject received Y or not. Analyses will then pool data across arms to evaluate X and Y separately. For example, to evaluate X, X + B and X + Y + B are combined and compared with B and Y + B. If the assumption of independence is true this design can lead to a substantial decrease in sample size compared with running separate trials of B + X and B + Y. But if there is a positive or negative synergy or interaction between X and Y results of the trial may be uninterpretable. As such, this trial design assumes no synergy and/or interaction, typically an unproved hypothesis. There will also be too little power to separately evaluate the cohorts because the sample size was selected assuming the cohorts could be pooled [34].

All the other analyses reported with a clinical trial. Clinical trials typically report analyses other than the pre-specified primary objective or endpoint. Such analyses are often labeled secondary, exploratory, subgroup, or translational analyses. Because of the increased probability of a false-positive conclusion discussed above all secondary objectives and analyses in a clinical trial should interpreted as non-definitive or hypothesis generating. When many "secondary" analyses are provided after the primary endpoint of a trial is not met, the results of any "significant" findings should be viewed with strong skepticism or outright ignored.

Sub-group analyses are common in clinical trials data reporting. Because the power of a comparison is related to the sample size, sub-group comparisons have less power than comparisons of the entire population. Lack of significance (p value > α) in a sub-group does not mean there is no association in the sub-group. It can be a false-negative result because the sample size is too small or for many other reasons [35]. Interpretation of p values is challenging in general, especially in the context of evaluating multiple

subgroups [35]. In these analyses, reviewing point-estimates and confidence intervals should be the focus. Interpretation of confidence intervals is also challenging; Greenland et al. [35] provide guidance. As noted above, retrospective or post-hoc power analyses are never appropriate for a subgroup or any analysis [11]. Sub-group analyses can only be used to assess if there appears to be significant heterogeneity across sub-groups compared with the entire trial population [36–38]. Forest plots are a way to visualize this. If there appears to be heterogeneity (some subgroups have a benefit and others, not), a definitive evaluation of such a sub-group effect requires validation in a new trial.

Sub-group analyses which are not pre-specified should be viewed skeptically or ignored. If someone evaluates 100 different non-pre-specified subgroups each with an α of 5% we would expect five of these to have a p value < 0.05 even when there is no difference in any of the sub-groups analyzed. This feature of statistical significance testing means that if enough tests are conducted, a significant p value is very likely to be found. Analyses conducted until finding a result with a significant p value are sometimes described as "fishing expeditions." As noted above, subgroup analyses typically lack power which leads many significant subgroup results to be false-positive results. These issues are why so much emphasis is put on pre-specifying subgroup and other secondary analyses in clinical trials.

Non-inferiority. Randomized clinical trials evaluating whether one therapy is better than another nearly always analyze results using the *intent-to-treat* (ITT) principle. Subjects are analyzed in their assigned/randomized cohort regardless of the intervention they received. ITT analyses are considered "conservative" in that subjects receiving the alternative (non-assigned) intervention skew or "bias" towards showing no difference between the cohorts. In this instance an ITT analysis may result in the incorrect conclusion an intervention is ineffective, a false-negative. When the primary objective of a trial is to evaluate non-inferiority, an ITT analysis skews the data towards potentially showing non-inferiority. Because of this it is typical in non-inferiority trials to use an as-treated analysis as the primary analysis [39, 40].

A critical element of a non-inferiority design is the non-inferiority margin the design will exclude. There are no specific rules on what threshold would warrant a conclusion of "non-inferiority," though some regulatory agencies have provided guidance in some situations, and margins vary widely across endpoints, patient populations, and trials [39–44]. Any interpretation of a non-inferiority trial requires the reader to evaluate whether they find the non-inferiority margin selected convincing and of clinical import.

The interpretation of non-inferiority trials (and most clinical trials) is further complicated when endpoints are measured beyond the intervention period, which is (nearly) always the case with a survival endpoint. After the intervention period, there is typically less information on how patients are being treated and followed, and later therapies or interventions are often not randomly or equally allocated across arms. For example, when patients were randomized between lenalidomide and placebo for post-transplant maintenance therapy for multiple myeloma, therapy after failure varies by randomized arm. In many ways, clinical trial analyses with longer-term endpoints should be reviewed as essentially observational database analyses, with the associated caveats in analysis and interpretation [1].

Survival by response. Difficult as it is to believe, analyses comparing survival of responders versus non-responders remains common despite wide-spread knowledge such analyses are subject to diverse biases [45]. A critical bias is that a subject must live long enough to respond. This is referred to as *guarantee-time* or *immortal-time* bias. Statistical remedies of these bias are described [46].

CONCLUSION

This is an education-orientated review of design and correct interpretation of clinical trials data. We discuss issues including multiple endpoints and subgroup analyses. Many of the issues discussed are relevant to the correct interpretation of data from clinical trials of haematopoietic cell transplants.

Megan Othus[1]✉, Mei-Jie Zhang[2] and Robert Peter Gale[3]
[1]*Division of Public Health, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.* [2]*Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI, USA.* [3]*Haematology Research Centre, Department of Immunology and Inflammation, Imperial College London, London, UK.* ✉*email: mothus@fredhutch.org*

REFERENCES

1. Zheng C, Dai R, Gale R, Zhang M. Causal inference in randomized clinical trials. Bone Marrow Transpl. 2019;55:4–8.
2. Jabbour J, Manana B, Zahreddine A, Al-Shaar L, Bazarbachi A, Blaise D, et al. Vitamins and minerals intake adequacy in hematopoietic stem cell transplant: results of a randomized controlled trial. Bone Marrow Transpl. 2021;56:1106–15.
3. Kantarjian HM, DeAngelo DJ, Stelljes M, Martinelli G, Liedtke M, Stock W, et al. Inotuzumab ozogamicin versus standard therapy for acute lymphoblastic leukemia. N Engl J Med. 2016;375:740–53.
4. DiNardo CD, Jonas BA, Pullarkat V, Thirman MJ, Garcia JS, Wei AH, et al. Azacitidine and venetoclax in previously untreated acute myeloid leukemia. N Engl J Med. 2020;383:617–29.
5. Kantarjian H, Stein A, Gökbuget N, Fielding AK, Schuh AC, Ribera J-M, et al. Blinatumomab versus chemotherapy for advanced acute lymphoblastic leukemia. N Engl J Med. 2017;376:836–47.
6. Sanchorawala V, Wright DG, Seldin DC, Falk RH, Finn KT, Dember LM, et al. High-dose intravenous melphalan and autologous stem cell transplantation as initial therapy or following two cycles of oral chemotherapy for the treatment of AL amyloidosis: results of a prospective randomized trial. Bone Marrow Transpl. 2004;33:381–8.
7. Garderet L, Iacobelli S, Moreau P, Dib M, Lafon I, Niederwieser D, et al. Superiority of the triple combination of bortezomib-thalidomide-dexamethasone over the dual combination of thalidomide-dexamethasone in patients with multiple myeloma progressing or relapsing after autologous transplantation: the MMVAR/IFM 2005-04 Randomized Phase III Trial from the Chronic Leukemia Working Party of the European Group for Blood and Marrow Transplantation. J Clin Oncol. 2012;30:2475–82.
8. Deininger MW, Kopecky KJ, Radich JP, Kamel-Reid S, Stock W, Paietta E, et al. Imatinib 800 mg daily induces deeper molecular responses than imatinib 400 mg daily: results of SWOG S0325, an intergroup randomized PHASE II trial in newly diagnosed chronic phase chronic myeloid leukaemia. Br J Haematol. 2014;164:223–32.
9. Rubinstein L, Crowley J, Ivy P, LeBlanc M, Sargent D. Randomized phase II designs. Clin Cancer Res. 2009;15:1883–90.
10. Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP, Smith MA. Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol. 2005;23:7199–206.
11. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat. 2001;55:19–24.
12. Horowitz MM, Gale RP, Sondel PM, Goldman JM, Kersey J, Kolb H-J, et al. Graft-versus-leukemia reactions after Bone Marrow Transpl. 1990;75:555–62.
13. Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Stat Med. 1999;18:695–706.
14. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. Proc Natl Acad Sci. 1975;72:20–2.
15. Mori T, Kikuchi T, Koh M, Koda Y, Yamazaki R, Sakurai M, et al. Cytomegalovirus retinitis after allogeneic hematopoietic stem cell transplantation under cytomegalovirus antigenemia-guided active screening. Bone Marrow Transpl. 2021;56:1266–71.
16. Al-Kadhimi Z, Gul Z, Abidi M, Lum L, Deol A, Chen W, et al. Low incidence of severe cGvHD and late NRM in a phase II trial of thymoglobulin, tacrolimus and sirolimus for GvHD prevention. Bone Marrow Transpl. 2017;52:1304–10.
17. DeFilipp Z, Li S, Avigan D, Armand P, Ho VT, Koreth J, et al. A phase II study of reduced intensity double umbilical cord blood transplantation using fludarabine, melphalan, and low dose total body irradiation. Bone Marrow Transpl. 2020;55:804–10.

18. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat. 1988;16: 1141–54.
19. Inoue Y, Nakano N, Fuji S, Eto T, Kawakita T, Suehiro Y, et al. Impact of conditioning intensity and regimen on transplant outcomes in patients with adult T-cell leukemia-lymphoma. Bone Marrow Transpl. 2021;31:1–11.
20. Shimomura Y, Hara M, Konuma T, Itonaga H, Doki N, Ozawa Y, et al. Allogeneic hematopoietic stem cell transplantation for myelodysplastic syndrome in adolescent and young adult patients. Bone Marrow Transpl. 2021;56:1–8.
21. Inoue Y, Okinaka K, Fuji S, Inamoto Y, Uchida N, Toya T, et al. Severe acute graft-versus-host disease increases the incidence of blood stream infection and mortality after allogeneic hematopoietic cell transplantation: Japanese transplant registry study. Bone Marrow Transpl. 2021;56:1–12.
22. Jepsen C, Turkiewicz D, Ifversen M, Heilmann C, Toporski J, Dykes J, et al. Low incidence of hemorrhagic cystitis following ex vivo T-cell depleted haploidentical hematopoietic cell transplantation in children. Bone Marrow Transpl. 2020;55:207–214.
23. Hoering A, LeBlanc M, Crowley JJ. Randomized phase III clinical trial designs for targeted agents. Clin Cancer Res. 2008;14:4358–4367.
24. Spiessens B, Debois M. Adjusted significance levels for subgroup analyses in clinical trials. Contemp Clin trials. 2010;31:647–656.
25. Center for Drug Evaluation and Research (CDER) CfBEaRC. Multiple Endpoints in Clinical Trials: Guidane for Industry. U.S. Department of Health and Human Services Food and Drug Administration.
26. Malladi R, Ahmed I, McIlroy G, Dignan FL, Protheroe R, Jackson A, et al. Azacitidine for the treatment of steroid-refractory chronic graft-versus-host disease: the results of the phase II AZTEC clinical trial. Bone Marrow Transpl. 2021;56:2948–55.
27. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J. 2012;33:176–82.
28. Sekeres MA, Othus M, List AF, Odenike O, Stone RM, Gore SD, et al. Randomized phase II study of azacitidine alone or in combination with lenalidomide or with vorinostat in higher-risk myelodysplastic syndromes and chronic myelomonocytic leukemia: North American Intergroup Study SWOG S1117. J Clin Oncol. 2017;35:2745–53. https://doi.org/10.1200/JCO.2015.66.2510.
29. Winter SS, Dunsmore KP, Devidas M, Wood BL, Esiashvili N, Chen Z, et al. Improved survival for children and young adults with T-lineage acute lymphoblastic leukemia: results from the Children's Oncology Group AALL0434 methotrexate randomization. J Clin Oncol. 2018;36:2926.
30. Dunsmore KP, Winter S, Devidas M, Wood BL, Esiashvili N, Eisenberg N, et al. COG AALL0434: a randomized trial testing nelarabine in newly diagnosed t-cell malignancy. J Clin Oncol. 2018;36:10500.
31. Pettengell R, Uddin R, Boumendil A, Johnson R, Metzner B, Martín A, et al. Durable benefit of rituximab maintenance post-autograft in patients with relapsed follicular lymphoma: 12-year follow-up of the EBMT lymphoma working party Lym1 trial. Bone Marrow Transpl. 2021;56:1413–21.
32. Milligan DW, Wheatley K, Littlewood T, Craig JI, Burnett AK. Group NHOCS. Fludarabine and cytosine are less effective than standard ADE chemotherapy in high-risk acute myeloid leukemia, and addition of G-CSF and ATRA are not beneficial: results of the MRC AML-HR randomized trial. Blood. 2006;107:4614–22.
33. Morgan GJ, Gregory WM, Davies FE, Bell SE, Szubert AJ, Brown JM, et al. The role of maintenance thalidomide therapy in multiple myeloma: MRC Myeloma IX results and meta-analysis. Blood J Am Soc Hematol. 2012;119:7–15.
34. Green S, Liu P-Y, O'Sullivan J. Factorial design considerations. J Clin Oncol. 2002;20:3424–30.
35. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31:337–50.
36. Lagakos SW. The challenge of subgroup analyses-reporting without distorting. N Engl J Med. 2006;354:1667.
37. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet. 2005;365:176–86.
38. Hernández AV, Boersma E, Murray GD, Habbema JDF, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? Am Heart J. 2006;151:257–64.
39. Jeker B, Farag S, Taleghani BM, Novak U, Mueller BU, Li Q, et al. A randomized evaluation of vinorelbine versus gemcitabine chemotherapy mobilization of stem cells in myeloma patients. Bone Marrow Transpl. 2020;55:2047–51.
40. Schrappe M, Bleckmann K, Zimmermann M, Biondi A, Möricke A, Locatelli F, et al. Reduced-intensity delayed intensification in standard-risk pediatric acute lymphoblastic leukemia defined by undetectable minimal residual disease: results of an International Randomized Trial (AIEOP-BFM ALL 2000). J Clin Oncol. 2017;36:244–53.
41. Johansson J-E, Bratel J, Hardling M, Heikki L, Mellqvist U-H, Hasséus B. Cryotherapy as prophylaxis against oral mucositis after high-dose melphalan and autologous stem cell transplantation for myeloma: a randomised, open-label, phase 3, non-inferiority trial. Bone Marrow Transpl. 2019;54:1482–8.
42. Kanda Y, Kobayashi T, Mori T, Tanaka M, Nakaseko C, Yokota A, et al. A randomized controlled trial of cyclosporine and tacrolimus with strict control of blood concentrations after unrelated bone marrow transplantation. Bone Marrow Transpl. 2016;51:103–9.
43. Center for Drug Evaluation and Research (CDER) CfBEaRC. Non-inferiority clinical trials to establish effectiveness: guidance for industry. U.S. Department of Health and Human Services Food and Drug Administration.
44. Oncology Center of Excellence CfDEaRC, Center for Biologics Evaluation and Research (CBER). Clinical Trial endpoints for the approval of cancer drugs and biologics: guidance for industry. Silver Spring: U.S. Department of Health and Human Services Food and Drug Administration; 2018.
45. Kröger N, Sockel K, Wolschke C, Bethge W, Schlenk RF, Wolf D, et al. Comparison between 5-Azacytidine treatment and allogeneic stem-cell transplantation in elderly patients with advanced MDS according to donor availability (VidazaAllo study). J Clin Oncol. 2021;39:3318–27.
46. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. J Clin Oncol. 1983;1:710–9.

ACKNOWLEDGEMENTS

MO acknowledges support from the National Cancer Institute (NCI) grant U10CA180819. MJZ acknowledges support from the National Institutes of Health (NCI, NHLBI) and Health Resources and Services Administration (HRSA). RPG acknowledges support from the National Institute of Health Research (NIHR) Biomedical Research Centre funding scheme.

AUTHOR CONTRIBUTIONS

MO wrote the initial typescript,. MJZ and RPG reviewed and provided comments. All authours accept responsibly for the content of the final typescript and agree to submit for publication.

CONMPETING INTERESTS

MO is a consultant for Daiichi Sankyo, Biosight, and Merck and is on independent data safety monitoring boards for Celgene and Glycomimetics. RPG is a consultant to BeiGene Ltd., Fusion Pharma LLC, LaJolla NanoMedical Inc., Mingsight Parmaceuticals Inc. CStone Pharmaceuticals, NexImmune Inc. and Prolacta Bioscience; advisor to Antengene Biotech LLC, Medical Director, FFF Enterprises Inc.; partner, AZAC Inc.; Board of Directors, Russian Foundation for Cancer Research Support; and Scientific Advisory Board: StemRad Ltd.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Megan Othus.