Disease Modification and Other Trials in Systemic Sclerosis Have Come a Long Way, But Have to Go Further

JANET E. POPE,[1] DINESH KHANNA,[2] SINDHU R. JOHNSON,[3] AND PHIL CLEMENTS[4]

In this issue of *Arthritis Care & Research*, the View article by Mendoza et al about clinical trial design in systemic sclerosis (SSc; scleroderma) suggests that there cannot be a universal design for clinical trials of disease modification in SSc because the disease has so many different presentations (1). However, it is still important in any randomized controlled trial (RCT) to use proper statistics, including sample size calculations, to be able to recruit subjects (feasible), and ideally to have subjects enrolled that are similar to SSc patients with comparable organ-based disease activity (generalizable). There will not be a single design for every trial, but basic principles must be adhered to. We have written this article as a counterview to the views by Mendoza et al (1), but most of the opinions are in agreement. The headings are ordered in a way that would mimic the process of clinical trial design and conduct (general principles, end points, and discussion of organ-specific trials in SSc skin and lung). Examples from trials of various rheumatic diseases have been provided to illustrate various aspects of trial design.

[1]Janet E. Pope, MD, MPH, FRCPC: St. Joseph's Health Care London, University of Western Ontario, London, Ontario, Canada; [2]Dinesh Khanna, MD, MS: University of Michigan Scleroderma Program, Ann Arbor; [3]Sindhu R. Johnson, MD: Toronto Western Hospital, Mount Sinai Hospital, and University of Toronto, Toronto, Ontario, Canada; [4]Phil Clements, MD, MPH: University of California, Los Angeles.

Address correspondence to Janet E. Pope, MD, MPH, FRCPC, St. Joseph's Health Care London, 268 Grosvenor Street, London, Ontario, Canada N6A 4V2. E-mail: janet.pope@sjhc.london.on.ca.

General principles of trial design

For a trial to be successful, a design with an adequate sample size that is not overly optimistic and an analysis plan developed before the trial starts will increase the likelihood of success as measured by a statistically significant primary end point. Having an experienced statistician is recommended, as the analysis plan is critical for trial development and the reporting of results. All randomized trials should be registered, and one web site for this is http://register.clinicaltrials.gov.

Ethics of clinical trials. It is ethical to perform trials when there is equipoise about a treatment (not being certain if it is effective or not, or if the risk/benefit ratio is favorable). Best available treatment (standard of care [SOC]) can be offered in some studies. For instance, a new drug can be compared to SOC in a blinded head-to-head study or as an "add on" to SOC. These designs should be double blind where possible. Some trials are performed where SOC is denied and informed consent provides the participants with the information that during the trial, a specific SOC is not provided. This is the case in some rheumatoid arthritis (RA) trials where, after failure of methotrexate, a biologic agent is compared to a placebo even though many approved treatments are available. Withholding SOC may make recruitment very difficult but it is not unethical, provided patients are informed of treatment options, understand some care is denied, and offer consent. Participation in a trial is optional, not mandatory.

RCTs versus single-arm studies. Many diseases may have heterogeneity, including systemic lupus erythematosus, where there are various antibodies and differences in organ activity and severity. For instance, lupus nephritis has drugs demonstrated to be useful in randomized trials. In SSc, disease modification (or softening of the skin), interstitial lung disease, and pulmonary arterial hypertension (PAH) are complications that have been studied in well-conducted RCTs. SSc treatment of one organ such as the skin should not be extrapolated to the treatment of other organs.

The role of single-arm studies is to determine if there

could be a reasonable expectation of benefit with acceptable risks in a future randomized trial. Most treatment in SSc looks positive in open-label studies, but not in RCTs. The main problems are the biases of open-label data, i.e., no comparison group, no blinding, and publication bias.

Reason for randomization. Fundamentally, randomization distributes disease characteristics and potential confounders equally among the treatment groups. This is done to decrease bias. Not all confounders are known, but randomization should make the groups equal for both known and unknown characteristics that could be associated with the outcome of the trial. Stratification at the time of randomization can be used for some important characteristics to fully ensure that the treatment groups are balanced with respect to those variables. For instance, in a digital ulcer study, there could be stratification for the number of baseline ulcers, such as one versus multiple ulcers (2 strata). There will be equal allocation of subjects receiving study treatments within each stratum, irrespective of the number of patients who have multiple ulcers compared to those with a single ulcer.

Parallel versus crossover design. Most RCTs use a parallel design (comparing one treatment to another or a placebo). Crossover trials are less common (where treatment allocation is changed at least once at a time point, usually in random treatment order). The advantage of a crossover study is that fewer subjects are needed (because the patient is compared to him/herself, which improves trial efficiency). However, the disadvantages include a carryover effect (where the first treatment has not worn off when the second treatment begins and/or the patient is not back at a baseline state) and an order effect (where the results are affected by the order in which various treatments are received). A true crossover design may be used in Raynaud's phenomenon (where the treatment effect starts and stops quickly). This design is unlikely to occur in an SSc skin trial, since it has been demonstrated that it takes many months to see a clinical effect with most of the agents studied to date.

Innovative trial designs. Perhaps combination treatment in some SSc trials could be considered, such as in PAH treatment. Combination therapy has been shown to be successful in RA. Side effects of combining 2 treatments may be higher, but potentially manageable when safety rules such as dose reduction are written into the protocol. Trial designs could include: A versus B (2-arm study), A versus A + B (2-arm study), or A versus B versus A + B (3-arm study). Strategies may compare active versus placebo or active versus active in the setting of adding to SOC or comparing to SOC. In trials of psoriatic arthritis, an add on of a tumor necrosis factor inhibitor to stable methotrexate gives the same magnitude of benefit as an add on to no background methotrexate. A less common design is randomized withdrawal, which has been used for biologic agent studies in juvenile inflammatory arthritis. The problem is (and it is true in many instances in SSc) that if the drug takes a long time to wash out or cause a rebound (and

the effect takes many months to be evident), then there may be a Type II error.

Sample size calculation. It is imperative for the investigator to understand and account for the natural history of the organ system he/she is interested in studying and to identify and account for predictors of worsening during the design phase of the study. This is true for any chronic disease. In SSc trials, some patients get better, some get worse, and some stay the same. Such trajectories have been explored and demonstrated in a meta-analysis of RCTs in diffuse cutaneous SSc (dcSSc) (2). Therefore, when performing a sample size calculation, the variability of the baseline characteristic of interest and the expected change compared to the natural history of the disease within the timeframe of the trial must be estimated. An assumption needs to be made about how much the outcome will change with treatment. Being overly optimistic can lead to a negative trial (such as expecting 50% improvement in the active treatment group compared to no response in the placebo group).

Trial duration: risk versus benefit tradeoff. The study needs to be long enough to determine a between-group difference given the expectations of both natural history and how treatment may alter this. If a drug is potentially harmful, the trial should be as short as required to detect a true difference so that patients are not exposed to unproven/harmful treatments for longer than necessary. When studying stem cell transplantation in SSc, the procedure may rapidly improve skin (early benefit) but increase mortality after the transplant. Later, there may be sustained improvement in major organs compared to patients not receiving this intervention. The duration of observation posttransplant must be sufficiently long to determine if the benefit is sustained and if there is an eventual survival advantage over the comparator. Therefore, the Scleroderma: Cyclophosphamide or Transplantation (SCOT) trial was designed to be long enough to determine a future survival benefit and also to identify any recurrence of disease.

Analysis of rare disease trial data. A disease is often considered rare when the prevalence is less than 1 in 1,500 or less than 1 in 2,500 (as definitions vary between countries). Recruitment for trials may be less feasible in a rare disease, particularly if there are strict inclusion and exclusion criteria. The latter can also affect study generalizability (3). There are analyses that may be helpful in rare diseases with relatively small sample sizes. Johnson et al reanalyzed the methotrexate trial by Pope et al using Bayesian analysis and included imputation for missing values. They found that methotrexate has a 94% probability (16:1 odds) of a beneficial effect on the modified Rodnan skin score (MRSS). There was an 88% probability (7:1 odds) of a beneficial effect of methotrexate on physician global assessment compared to placebo (4). These analyses suggest that there are other methods of analysis that can evaluate a difference in effects of the drug of interest versus the comparison or placebo group that use the avail-

able data, and are less susceptible to study power. However, Bayesian analyses can never make a negative trial positive. It is a different way to analyze data: allowing inferences to be made using the available data. Bayesian methods are increasingly being used in observational studies (5,6) and clinical trials (7,8). Indeed, the US Food and Drug Administration has guidelines for the use of Bayesian methods in medical device clinical trials (9). There are also more sophisticated analytical techniques for handling missing or incomplete data other than the last observation carried forward or intent-to-treat, which tends to harshly penalize the test drug. Data that are randomly missing can be analyzed using multiple imputation methods. Every effort should be made to keep all patients returning for their study visits, even if they have discontinued the study medication.

Breaking the blind and adverse events (AEs). It may or may not be possible to determine worsening SSc from AEs related to the study medication. However, that is why there is a comparison arm (to compare efficacy but also AEs). If a serious unexpected AE occurs, there may be consideration of "breaking the code" (unblinding the subject's treatment allocation). This may not be mandatory if patients will be managed the same irrespective of what study treatment they received.

Escape from treatment allocation (rescue), mandatory withdrawal, and open-label extension. Mendoza et al state that early escape or use of salvage arms in the study design for patients with worsening lung and skin involvement in SSc trials is strongly recommended (1). This design in RA trials has led to problems with the interpretation of results, where in some studies large numbers of patients escape to an active treatment and randomization after that time point is no longer in effect (10). There are other ways to maintain treatment allocation. An escape can be an "add-on therapy" to the study drug, which maintains blinding and study treatment. Strict "treatment failure" end points should be written into the protocol where escape or adding other treatment is allowed (such as ≥25% worsening of skin score or ≥15% worsening of forced vital capacity [FVC] % predicted). Such patients could be eligible to break the code, but only if that would help in deciding future treatment. In the Scleroderma Lung Study (SLS), for example, only 6 of the 158 entrants were declared "treatment failures," defined as a decrease in FVC % predicted by more than 15% of the predicted value on 2 consecutive tests 30 days apart (11). Treatment failures were taken off the study medication and were eligible to have the drug code broken. Only 6 of the 158 patients requested a breaking of the drug code for medical decision making. Once the code was broken, the patient and the treating physician could decide on future therapy, but the future therapy was not dictated by the study protocol. Treatment decisions were implemented outside the treatment protocol. In the SLS, only 2 patients elected to go on to cyclophosphamide (open therapy) after stopping study medication. All of the patients were encouraged to return for the 12-, 18-, and 24-month evaluations, even though they might have discontinued the study medication earlier or started another therapy for lung disease. Escape arms are a problem both scientifically and statistically, especially when there is no SOC for some aspects of SSc. Therefore, salvage treatment does not salvage the trial.

Some trials will have a mandatory dropout, but do not offer the study drug. The subject returns to usual care. Trials enrolling patients where SOC was not effective or contraindicated (e.g., where a new agent is compared to placebo) should not have a rescue arm, as there is no SOC being denied. The ethics of this are clear. Perhaps a better way of determining long-term safety and efficacy is offering an open-label extension (providing the active study drug once the trial is completed) or re-randomizing the patients who are still blinded to receive treatment or placebo in a later time phase of a trial. Open-label extensions may encourage recruitment but have potential risks, especially if there is no proven benefit for the condition under study (12).

Outcome measures/end points

Outcome Measures in Rheumatology has recommendations for validating outcome measurements, and this is an ongoing resource for those designing trials in SSc (13). Ideally, outcome measures should be valid, reliable, and responsive to change. A number of outcome measures have been formally evaluated in SSc (13,14).

Mortality is unlikely to be an end point in SSc trials unless the patients included have a high anticipated mortality or the trial is very long or extremely large. Mortality data from the SCOT trial will soon be available comparing stem cell transplant and cyclophosphamide in SSc patients who at entry had a risk of high mortality.

Objective measures. In open nonrandomized trials, measures based on rater interpretation, such as the MRSS or visual analog scales, are vulnerable to observation bias. That is why a control group and double blinding are needed. A single investigator must do all of the MRSS evaluations on a given patient throughout a trial in order to decrease error (eliminate interrater differences). A trained experienced investigator is even better for reducing intra-rater error.

Sample size calculations do not necessarily consider the minimum important difference (MID) of the end point, but secondary or exploratory outcomes may be based on the MID (such as the proportion in each treatment arm that obtains at least the MID for an outcome of interest). An example of this is the change in a pulmonary function parameter such as diffusing capacity or lung volume, where there could be a statistically significant between-group difference, but the MID is not achieved as a mean difference. However, there could be a higher proportion that achieves the MID in the active treatment group.

A mean difference in an outcome measure may be below the measurement error or precision of the test. For instance, osteoporosis treatment may increase the bone mineral density above placebo by 3%, but the precision of the dual x-ray absorptiometry machine is 3%. Measurement

error should be equal between treatment groups, so the clinical relevance of the results is not related to precision of the measurement. The same would be true for FVC % predicted in a scleroderma lung study.

Organ-specific trials: skin and lung as examples

Treatment to improve SSc skin. Two randomized placebo-controlled methotrexate trials yielded a positive and a negative study (*P* value close to 0.05) (15,16). An overly optimistic sample size calculation can lead to an underpowered trial. The SLS, which was powered appropriately, showed a significant change in the MRSS at 12 months in the cyclophosphamide arm (11). Therefore, the claim that "it can be accurately concluded that SSc skin disease has no proven effective therapy" is not totally accurate.

Other drugs have been studied in SSc to improve skin. Mycophenolate mofetil has been tried in several case series, and in 1 study that provided matched controls acquired by a chart audit, the changes in skin score may not be different from historical controls (17). The use of historical controls will not prove that treatment is effective, but may inform the probability of a positive randomized trial (e.g., yield preliminary data, particularly about safety). As such, it is unlikely that imatinib will enter a phase III trial in SSc because of its poor tolerability and lack of superiority to historical controls (ref. 18 and Distler O, et al: unpublished observations). Even so, well-designed trials can help to determine if a treatment has an advantage for AEs (19).

Use of a small dose of the study drug as a "placebo." Placebo-controlled trials should be performed wherever possible, even in phase II trials. However, sometimes a placebo cannot be made that looks, smells, and tastes identical. The mini-dose of D-penicillamine used in low-versus high-dose D-penicillamine in an early dcSSc trial may have preserved blinding because the metallic taste that can be seen with D-penicillamine was distributed equally in both arms (20). The conclusion of the study could be that a subclinical dose of the drug was equal to usual dosing, or that the drug does not work. The latter conclusion is widely accepted.

Consideration for lung trials: induction and maintenance. Trials of interstitial lung disease in SSc could have 2 parts: induction and maintenance. Inclusion criteria and end points must be defined, such as "alveolitis." In a lung trial, there should be mandatory parameters such as FVC % predicted and high-resolution chest computed tomography evidence of pulmonary fibrosis, ground glass, and honeycombing.

Discussion

There have been many trials in rheumatic diseases with several successful designs. Many end points in SSc trials have been validated. Some successful treatments have been demonstrated in RCTs in SSc. However, there is a large unmet need in SSc because it has the worst prognosis of the rheumatic diseases. Therefore, despite the rarity of SSc, we have come a long way, but need to go further.

REFERENCES

1. Mendoza FA, Keyes-Elstein LL, Jimenez SA. Systemic sclerosis disease modification clinical trials design: quo vadis? Arthritis Care Res (Hoboken) 2012;64:945–54.
2. Merkel PA, Silliman NP, Clements PJ, Denton CP, Furst DE, Mayes MD, et al, for the Scleroderma Clinical Trials Consortium. Patterns and predictors of change in outcome measures in clinical trials in scleroderma: an individual patient meta-analysis of 629 subjects with diffuse scleroderma. Arthritis Rheum 2012. E-pub ahead of print.
3. Villela R, Yuen SY, Pope JE, Baron M, and the Canadian Scleroderma Research Group. Assessment of unmet needs and the lack of generalizability in the design of randomized controlled trials for scleroderma treatment. Arthritis Rheum 2008;59:706–13.
4. Johnson SR, Feldman BM, Pope JE, Tomlinson GA. Shifting our thinking about uncommon disease trials: the case of methotrexate in scleroderma. J Rheumatol 2009;36:323–9.
5. Johnson SR. Bayesian inference: statistical gimmick or added value? J Rheumatol 2011;38:794–6.
6. Johnson SR, Granton JT, Tomlinson GA, Grosbein HA, Hawker GA, Feldman BM. Effect of warfarin on survival in scleroderma-associated pulmonary arterial hypertension (SSc-PAH) and idiopathic PAH: belief elicitation for Bayesian priors. J Rheumatol 2011;38:462–9.
7. Wijeysundera DN, Austin PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. J Clin Epidemiol 2009;62:13–21.e5.
8. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health care evaluation. Chichester (UK): John Wiley & Sons; 2004.
9. Campbell G. Guidance for the use of Bayesian statistics in medical device clinical trials. Rockville (MD): US Food and Drug Administration; 2010.
10. Keystone E, van der Heijde D, Mason D Jr, Landewe R, van Vollenhoven R, Combe B, et al. Certolizumab pegol plus methotrexate is significantly more effective than placebo plus methotrexate in active rheumatoid arthritis: findings of a fifty-two–week, phase III, multicenter, randomized, double-blind, placebo-controlled, parallel-group study. Arthritis Rheum 2008;58:3319–29.
11. Tashkin DP, Elashoff R, Clements PJ, Goldin J, Roth MD, Furst DE, et al. Cyclophosphamide versus placebo in scleroderma lung disease. N Engl J Med 2006;354:2655–66.
12. Khanna D, Clements PJ, Furst DE, Korn JH, Ellman M, Rothfield N, et al, for the Relaxin Investigators and the Scleroderma Clinical Trials Consortium. Recombinant human relaxin in the treatment of systemic sclerosis with diffuse cutaneous involvement: a randomized, double-blind, placebo-controlled trial. Arthritis Rheum 2009;60:1102–11.
13. Furst D, Khanna D, Matucci-Cerinic M, Clements P, Steen V, Pope J, et al. Systemic sclerosis: continuing progress in developing clinical measures of response. J Rheumatol 2007;34:1194–200.
14. Johnson SR, Hawker GA, Davis AM. The Health Assessment Questionnaire disability index and Scleroderma Health Assessment Questionnaire in scleroderma trials: an evaluation of their measurement properties. Arthritis Rheum 2005;53:256–62.
15. Van den Hoogen FH, Boerbooms AM, Swaak AJ, Rasker JJ, van

Lier HJ, van de Putte LB. Comparison of methotrexate with placebo in the treatment of systemic sclerosis: a 24 week randomized double-blind trial, followed by a 24 week observational trial. Br J Rheumatol 1996;35:364–72.

16. Pope JE, Bellamy N, Seibold JR, Baron M, Ellman M, Carette S, et al. A randomized, controlled trial of methotrexate versus placebo in early diffuse scleroderma. Arthritis Rheum 2001; 44:1351–8.

17. Le EN, Wigley FM, Shah AA, Boin F, Hummers LK. Long-term experience of mycophenolate mofetil for treatment of diffuse cutaneous systemic sclerosis. Ann Rheum Dis 2011;70: 1104–7.

18. Pope J, McBain D, Petrlich L, Watson S, Vanderhoek L, de Leon F, et al. Imatinib in active diffuse cutaneous systemic sclerosis: results of a six-month, randomized, double-blind, placebo-controlled, proof-of-concept pilot study at a single center. Arthritis Rheum 2011;63:3547–51.

19. Furst DE, Tseng CH, Clements PJ, Strange C, Tashkin DP, Roth MD, et al, for the Scleroderma Lung Study. Adverse events during the Scleroderma Lung Study. Am J Med 2011;124: 459–67.

20. Clements PJ, Furst DE, Wong WK, Mayes M, White B, Wigley F, et al. High-dose versus low-dose D-penicillamine in early diffuse systemic sclerosis: analysis of a two-year, double-blind, randomized, controlled clinical trial. Arthritis Rheum 1999;42:1194–203.