Systemic sclerosis-associated interstitial lung disease – proposed recommendations for future randomised clinical trials

D. Khanna[1], K.K. Brown[5], P.J. Clements[1], R. Elashoff[2], D.E. Furst[1], J. Goldin[4], J.R. Seibold[6], R.M. Silver[7], D.P. Tashkin[3], A.U. Wells[8]

[1]Division of Rheumatology, [2]Biomathematics, [3]Pulmonary and Critical Care, Department of Medicine, and [4]Department of Radiology, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA; [5]Department of Medicine, National Jewish Health, Denver, CO, USA; [6]Division of Rheumatology, University of Connecticut Health Center, Farmington, CT, USA; [7]Division of Rheumatology, Medical University of South Carolina, Charleston, SC, USA; [8]Royal Brompton Hospital and National Heart and Lung Institute, London, U.K.

Authorship arranged alphabetically (except for first author).

Dinesh Khanna, MD, MS
Kevin K. Brown, MD
Philip J, Clements, MD, MPH
Robert Elashoff, PhD
Daniel E. Furst, MD
Jonathan Goldin, MD, PhD
James R. Seibold, MD
Richard M. Silver, MD
Donald P. Tashkin, MD
Athol U. Wells, MD

Please address correspondence to:
Dinesh Khanna, MD, MS,
David Geffen School of Medicine at UCLA,
90095 Los Angeles, CA, USA.
E-mail: dkhanna@mednet.ucla.edu

Received on May 10, 2010; accepted on May 11, 2010.

Clin Exp Rheumatol 2010: 28 (Suppl. 58): S55-S62.

Key words: Systemic sclerosis, interstitial lung disease.

Conflict of interest: Dr Clements receives less than $10,000 per year as member of the PAH advisory board for Gilead; Dr Furst has received consultancy fees and grant support from Actelion and Gilead, and honoraria from Actelion; Dr Goldin is the director of MedQIA Imaging Core; Dr Seibold has funded research and/or consultancy relationships with FibroGen, Actelion, Pfizer, Gilead and United Therapeutics; the other co-authors have declared no competing interests.

ABSTRACT
Pulmonary disease is the leading cause of morbidity and mortality in patients with systemic sclerosis (SSc). Recent well-designed trials in SSc-associated interstitial lung disease (SSc-ILD) have provided important insights regarding outcome measures and trial design. Recent investigations into the pathogenesis of SSc-ILD have led to a renewed interest in assessing targeted therapies in SSc-ILD. With this in mind, we propose recommendations for the design of future SSc-ILD studies in this review.

Introduction

Pulmonary disease is the leading cause of morbidity and mortality in patients with systemic sclerosis (SSc) (1). Lung involvement in SSc can be separated into two distinct entities: pulmonary arterial hypertension (PAH) and interstitial lung disease (ILD) although many patient have elements of both. Recent well-designed trials in SSc-ILD have provided useful information regarding outcome measures and trial design. This, coupled with active investigation into the pathogenesis of SSc-ILD, has led to an interest in assessing targeted therapies in SSc-ILD. Reliable trial design is a requisite to assessment of these potential therapies. With that in mind, we propose recommendations for the design of future SSc-ILD studies (Table I).

1. Influence of disease classification

A number of aspects of the natural course of SSc-ILD have been elucidated. Steen *et al*. have demonstrated that the major loss of forced vital capacity (FVC) occurred within the first 4-6 years of SSc: patients who developed severe restrictive disease (%FVC ≤50% of predicted) had lost 32% of their remaining FVC each year for the first 2 years, 12% of remaining FVC for each of the next 2 years, and 3% of remaining FVC for each of the following 2 years. Severe ILD (found in 13% of their large cohort of SSc patients) was seen only slightly more commonly in diffuse SSc than in limited SSc. This observation was replicated by White *et al*. suggesting that the frequency of clinically significant ILD in limited SSc is only slightly less than the frequency in diffuse SSc (2). The Scleroderma Lung Study (SLS-I), a multi-centre, double-blind, randomised controlled study, evaluated the effectiveness and safety of oral CYC (up to 2.0 mg/kg/day) administered for one year in 158 patients with symptomatic SSc-ILD (3). SLS-I provided further evidence that the change in FVC (in either the cyclophosphamide or placebo groups) was virtually identical in limited and diffuse SSc (4).

2. Trial duration

SLS-I (3) showed that a 1-year parallel trial design for the primary outcome (FVC% predicted) was sufficient to demonstrate statistically significant differences between cyclophosphamide *vs*. placebo. Sub-analysis failed to show a significant difference at 6 months between cyclophosphamide *vs*. placebo. On the other hand, an additional 1-year of observation off of medication following the 1-year double-blind treatment period demonstrated a further improvement in outcomes with cyclophosphamide at 18 months. This benefit was subsequently lost, with no differences noted at 24 months. In a supportive study, cyclophosphamide was administered by intravenous infusion monthly for 6 months followed by oral azathioprine (n=23), compared with placebo infusions and oral placebo (n=22), in patients with active SSc-ILD (5). A benefit in the primary endpoint of decline in FVC% predicted trended toward a favourable outcome in the actively treated group (*p*=0.08). These data suggest that treatment trials should be designed for a minimum of one year. Longer studies may be necessary to ascertain the durability of the treatment effect beyond 1 year (especially when medication is stopped at 1 year) or if outcomes such as progression-free survival are primary outcome measures (discussed below).

Table I. Proposed recommendations for future SSc-ILD studies.

1) Limited or diffuse cutaneous SSc.

2) 1-year RCT.

3) Placebo (or active) controlled RCT.

4) The presence of ILD should be based upon the detection of appropriate abnormalities on HRCT, with this process subject to quality control.

5) Rigorous quality control in the accuracy of PFT estimation

6) Presence of dyspnea should not be required as an inclusion criterion.

7) The primary goal of RCT should be to prevent disease progression as defined by stabilisation of FVC% predicted (rather than to achieve regression of fibrotic abnormalities)

8) Progression-free survival is an important secondary end-point.

9) A systematic attempt should be made to enhance the sensitivity of this end-point by means of:
a) Cohort enrichment: the selection of patients at greater risk of progression, based upon disease severity, observed progression or a short duration of systemic disease. Although cohort enrichment might be based on any single one of the criteria, it is recommended that patients attain severity threshold based on HRCT and/or FVC%. In addition, they should meet one of the two additional criteria: a) short duration of systemic disease (defined as first 6 years after onset of signs or symptoms attributable to SSc from first non- Raynaud's signs or symptoms) or b) Observed progression (> 10% of FVC over the past 3-12 months)

10) Composite indices: Exploratory analyses in which marginal reductions in FVC (5-10%) are evaluated alone and in combination(as a composite measure) along with a) changes in dyspnea; and b) changes in HRCT extent.

11) Evaluation of biomarker signal, to inform subsequent study design to establish
a) whether a biomarker signal at baseline predicts a treatment effect;
b) whether short term improvement in a biomarker signal predicts a longer term treatment effect.

3. Placebo or active controls

At this stage, cyclophosphamide has been shown to be superior to placebo in SSc-ILD in a single double-blind placebo controlled trial. A follow up trial (SLS-II) is ongoing, comparing mycophenolate mofetil (test drug) for two years to cyclophosphamide (established drug) for one year followed by placebo for one year. (http://sls.med.ucla.edu/). Unless future investigators are sure that their active arm is likely to be superior to an active control, comparing their active drug to cyclophosphamide or to placebo is recommended.

4. Chest imaging

Progress toward specific treatments for the lung disease in SSc might be accelerated by moving beyond measurements of lung function alone to the precise assessment of the specific parenchymal abnormalities. There are two key roles of thoracic computer tomography (CT) imaging in clinical trials of SSc: (1) detection and staging baseline severity that can be effectively used for i) cohort enrichment and ii) adjusting for baseline severity in key treatment effect analyses (as it is likely that a treatment effect may differ in cases with mild rather than extensive lung disease and discussed in section 8a); and (2) as a surrogate end point or more accurate measure of serial change. In addition, baseline CT is important in the determination of patient eligibility for excluding other significant disease.

Detection and staging baseline severity

CT imaging provides a means of accurately characterising the nature and extent of lung parenchymal changes. SSc-ILD typically manifests on CT as predominantly ground glass appearance (GGO) with an admixture of pulmonary fibrosis, consistent with the non-specific interstitial pneumonia (NSIP) pattern. Patients with SSc-ILD develop honeycomb cystic change (reported in 11-37% of cases), unlike other patients with NSIP who have little or no cystic change (6).

Several visual scales of varying complexity have been developed for scoring the severity and extent of SSc-ILD disease. All of these scales include parenchymal ground glass opacities, reticulation, and honey comb cysts. Some scales include emphysema, architectural distortion, and ground glass with or without architectural distortion. They differ with respect to how the severity of the features is assessed and how features are anatomically localised.

There are two more common visual systems in practice. The first was originally shown to correlate well with pathological specimens (7). This system divides each lung into three zones (lung apex to aortic arch, aortic arch to inferior pulmonary veins, inferior pulmonary veins to lung bases). The reader estimates and grades the amount and the extent of lung abnormality (for ground glass, reticulation and honeycomb cysts) in each zone, scoring involvement using a scale of 0 to 4 (0= absent; 1=1 to 25%; 2=26 to 50%; 3=51 to 75%; 4=76 to 100%) (7). The use of this scoring system was recently found to be both predictive of treatment outcome and a key variable in the model to standardise patient severity in the analysis of treatment efficacy in the SLS I study.

More recently, Goh and colleagues have proposed a simple algorithm that incorporates extent of disease (as defined by presence of ground glass opacity or fibrosis on visual read) and FVC% predicted (8). HRCT involvement >20% and a baseline FVC <70% were associated with increased mortality risk (hazard ratio [HR] 2.48 and HR 2.11).

Computer-based approaches have also been investigated and are based upon measurement of the density or texture features of each pixel and assignment of a measure of the amount of abnormal lung tissue present (9, 10). Computer based models correlate well with visual scoring techniques for the detection of fibrosis (area under the curve [AUC]=0.86) and with the assessment of extent of disease (AUC=0.96 for estimating lung involvement >25%) without the intra-reader variation encountered with visual scoring (9).

Measuring treatment effect

The role of CT has not been extensively studied in longitudinal SSc cohorts. The biggest challenges are: a) uncertainty of when to repeat CT scan and b) how to compare two different CT scans in which, by virtue of the non contiguous image sampling, comparing the same anatomic level is usually not possible. In addition, when anatomic registration is achieved, it is sometimes difficult to ascertain the clinical significance of minor CT changes due to reader variation inherent in visual scoring systems (for SSc-ILD, the inter and intra reader agreement [Kappa statistic] ranges between 0.6 and 0.8, respectively). However, In SLS-I, CT were scored at follow-up as "worse" or "not worse" with respect to reticulation (fibrosis). Cyclophosphamide was associated with a treatment effect (71% of cyclophosphamide group remained "not worse" *vs*. 47% in the placebo group; *p*=0.01) and was associated with an improvement in dyspnea and FVC% improvement (10).

Another approach to score CT scans is a quantitative computer-assisted diagnosis (CAD) scoring system that has the inherent advantage of no intra- or inter-reader variation. Therefore, if quantitative scoring system is applied to the same anatomic data that is acquired in a standardised fashion, any change measured may be a real change. In SLS-I, a quantitative approach by CAD was applied prospectively to serial CT data to assess for interval change (11). The quantitative estimation by CAD demonstrated the same treatment effect (as measured by absolute reduction in percentage of reticulation [fibrosis]) as the semi quantitative visual assessment of "worse *vs*. not worse"). In addition to being able to measure the absolute change, CAD can also assess the rate of change in fibrosis score since CAD can be scored as a continuous measure that is not possible with visual approaches. This may result in better power to detect differences between the 2 groups and a smaller sample size. In SLS-I, the number of cases required to show a treatment effect by CAD scoring was 2/3 the number of cases needed to show treatment efficacy when FVC was the endpoint, and was less than half the number of cases needed to show treatment efficacy when the visual assessment of change on HRCT was the endpoint.

Thus it would seem that visual and computer-based quantitative scoring systems are synergistic rather than competitive. One potential strategy is to use a visual approach to assess the presence and extent of disease on the baseline CT (to enrich study cohorts at entry) and to use the computer-based techniques to assess treatment efficacy. Both of these approaches would offer the added potential advantage of decreasing the sample size.

Quality control

To achieve the roles defined for HRCT, quality control of the chest imaging techniques is critical. At this time, the methodology has not been standardised but guidelines for HRCT evaluation in ILD have been generated (12). In the ongoing scleroderma stem cell transplant study (SCOT) and SLS II study, as well as in two industry trials, a rigorous standardised imaging protocol has been implemented across multiple sites and over several time points. In these studies rigorous quality control is applied to the imaging protocol with ongoing quality control performed after each study by a central imaging core. The use of a phantom at the time of credentialing the scanners has shown that there is significant variation across scanners which can be at least partially ameliorated by modifying protocol sequences to have a similar noise and image quality across the sites. The use of the phantom also allows for handling changes of technology platforms that occur during longitudinal studies. Another important aspect is the training of technologists with respect to breathing instructions given at the time of scanning to ensure the same level of inspiration at each scan examination.

A number of issues remain unresolved. First, there needs to be control of or correction for variability in inspiratory level. Second, there are issues with the optimum image acquisition protocol. The often used 1mm axial HRCT acquired image every 10 to 40 mm makes it difficult to achieve exact anatomical comparability between initial and follow-up CT evaluation, due to the interspaced nature of high-resolution CT evaluation. Volume acquisition techniques overcome this problem but at the cost of increased numbers of images to review and a major increase in the radiation dose considerations. CT attenuation is dependent on scanner type, model, object positioning within the scanner gantry and various physical factors (*e.g.* kilovoltage, current-time product, slice thickness and reconstruction algorithm) (13, 14). Furthermore, it is recognised that spatial uniformity of CT numbers over the entire subject area may only be achievable for certain combinations of parameters. Nevertheless with centralised overview of standardised image acquisition and interpretation the image quality can be optimised, akin to PFT tests. It is commonly recommended that interpretation be performed by expert thoracic radiologists rather than by less experienced observers (15, 16).

With regard to patient safety, CT examination does entail exposure to ionising radiation. Historically considered to be a high-dose examination, modern scanners and protocols have allowed a great reduction in effective dose delivery without loss of fidelity (17, 18). Conventional HRCT (1.5-mm images at 10-mm intervals with 140 kVp and 175 mAs) delivers an effective dose of 0.98 mSv, which is 12 times that of a posteroanterior and lateral chest radiograph. Volumetric protocols entail several times this dose but the use of low-dose multidetector imaging, employing tube currents as low as 100 mAs, reduces the dose to below that recommended in studies of mild-to-moderate risk (3–10 mSv).

5. Pulmonary physiology

Pulmonary function tests are always included as key inclusion/exclusion criteria and endpoints in SSc-ILD clinical trials (3, 5, 19). Of these, vital capacity (either FVC or slow vital capacity [SVC]) is usually selected as the primary endpoint, and total lung capacity (TLC) and the DLCO are usually included as important secondary endpoints. All of these tests are subject to within-subject variability both within each testing session and between sessions (20-22). This variability is related to varying performance characteristics of the equipment used for the test, varying proficiency of the technician in conducting the tests in accordance with standard recommendations (*e.g.* the American Thoracic Society/European Respiratory Society recommendations) and varying ability or motivation of the patient to follow the technician's instructions. Clearly, in order to maximise the ability to detect the impact of an intervention on a significant change in lung function in a multi-centre study population, it is important to minimise the variability of test results within each subject (who serves as his/her own control), as well as between subjects both within and between centres. The optimal method of accomplishing this objective involves the use of a centralised quality control program, ideally consisting of standardised equipment across the different centres, centralised training and certification of the study technicians on study-specific equipment and central review of the pulmonary function data that are electronically transmitted to the reading centre with timely feedback to the technicians concerning the quality of the data (20). Because of the expense of centrally standardised equipment, particularly for DLCO and plethysmographically determined lung volumes, alternative quality control programs are more feasible for studies in SSc-ILD. Such programs rely on: 1) the use of equipment at each participating site that meets recommended criteria for performance characteristics (23-25) as determined and affirmed by the equipment manufacturer, 2) standardised methods of calibrating the equipment with documentation that such calibrations have been carried out, 3) certification of technicians based on local or central review of the results of tests performed on an initial sample of subjects and 4) ongoing review by a core reading centre of the graphic and numeric results of tests performed on all subjects at each site with periodic feedback to the technicians at each site concerning the quality of the tests performed and a plan for remedial action should test qualify fall below acceptable standards. Since measurements of diffusing capacity and plethysmographic lung volumes are subject to instrument "drift", their accuracy and reliability should be verified through the use of biologic standards. Such standards involve repeated measurements at quarterly intervals on two standard normal subjects; measurements of lung volumes and diffusing capacity in the same healthy individual should agree within 10% of the mean of measurements performed 3 months earlier (21).

Quality control programs were implemented in SLS-I with satisfactory results with respect to achievement of acceptable test quality, which is likely to have contributed to the finding of a statistically significant effect of the active study drug (oral cyclophosphamide) on both FVC and TLC as a percent of predicted (3).

6. Respiratory symptoms

Although dyspnea is usually the presenting symptom of SSc-ILD, some patients with moderate-severe loss of lung function/advanced fibrosis on HRCT do not report dyspnea, possibly due to an adjustment of their lifestyle and level of physical activity to compensate for their functional impairment. Moreover, the presence of dyspnea and/or diminished effort tolerance in SSc may be related to physical deconditioning, myocardial involvement and/or pulmonary vascular disease (26). Dyspnea should be included as a patient reported outcome but only as a secondary outcome measure in SSc-ILD trials. At present, only the Mahler dyspnea index meets the requirements for validation of the OMERACT filter (27). Baseline scores of the Mahler dyspnea index depend on ratings for 3 different categories: functional impairment, magnitude of task, and magnitude of effort. In the SLS, baseline scores were able to discriminate between moderate and severe physiological parameters of breathing (FVC and DLCO) and correlated well with the baseline breathing visual analogue scale (r=0.61) (28, 29). Mahler's transition dyspnea score (TDI), which represents a change score from baseline, was able to differentiate between patients on cyclophosphamide and placebo in SLS-I at 1-year, demonstrating its sensitivity to change (30); significant improvements in TDI correlated with both parallel improvements in FVC% predicted and stability or improvement in visual assessments of pulmonary fibrosis on HRCT (10, 30).

7. Primary endpoints

The pathophysiology of SSc postulates a vasculopathy and immune activation resulting in activation of myofibroblasts to produce fibrosis (31). Most patients are not seen until the fibrotic phase is well established. The underlying process of fibrosis is not known to be rapidly reversible. As an example, fibrosis on skin biopsies reversed approximately 3 years after stem cell transplantation (32). Data from SLS-I showed that cyclophosphamide was associated with stabilisation/ improvement of HRCT fibrosis score compared to placebo group (*p*=0.014) 33 and provides preliminary data that CT measures of fibrosis may be at least partially reversible. However, neither the minimally detectable differences (*i.e.* percentage of measurement error on visual reads) nor the minimally important differences for HRCT fibrosis scores have been determined. In practice, RCTs cannot usually be continued for three or four years, thus, with currently available therapies, reversal of fibrosis does not appear to be a practical endpoint.

On the other hand, stabilisation of fibrosis appears a more viable short-term goal. In the SLS-I, patients on cyclophosphamide had stabilisation of their FVC% while the HRCT showed less progression of fibrosis in the cyclophosphamide than in the placebo group (10). The second RCT assessing pulse cyclophosphamide also supported stabilisation of the FVC% (5). These support the concept that prevention of progression is a viable strategy. Large retrospective cohort studies have suggested that mean behavior of FVC is stability over periods of two years, even in patients with otherwise severe and early disease (34). Other fibrosing lung diseases such as IPF tend to be more predictably progressive and spontaneous or treatment reversal of fibrosis is not known to occur. Given the available clinical data and our current relative lack of knowledge about pathophysiology, reversal is currently not a plausible goal. Trials in SSc should strive to slow or prevent disease progression.

8. Progression – free survival as a candidate endpoint

Candidate measures of efficacy have been proposed for IPF that attempt to address the inconsistent relationship of surrogate markers such as FVC with important outcomes such as survival. A "time to worsening" definition has been proposed in IPF that measures time to clinically meaningful events including time to acute IPF exacerbation, IPF-related death, lung transplantation, or hospitalization for respiratory decompensation. The excellent short-term survival in SSc-ILD (see below) and the rarity of performance of lung transplantation reduce the utility of this definition of outcome and response.

An intermediate measure of poor clinical course is termed "progression free survival", specifically the time to first occurrence of either 10% absolute decline in FVC% or 15% absolute decline in DLCO% or death. These measures of outcome are reasonable in the setting of relatively homogeneous progressivity of IPF and the predictably high event rates in a syndrome with median survival 3-5 years (35).

The basic issue becomes one of parametric versus nonparametric analysis and the evolving understanding of the clinical, research and statistical utility of "responder frequency analysis". To the clinician making bedside decisions, it is probably more "in-life" to consider what percentage of patients beginning a therapy might enjoy some level of an a priori definition of response rather than projecting mean responses in large patient groups into interpretation of outcome in the individual. More simply, it permits judging whether or not the individual patient is responding or not. The course of SSc-ILD is far less predictable than that of IPF. SSc-ILD includes many patients with extraordinarily stable disease as well as those with rapid deterioration. In the first year of the SLS-I, there were only two deaths and three patients who met the definition of "treatment failure" (an absolute 15% decline in FVC% predicted repeated once at least 30 days later) out of 79 randomised to drug and three deaths and five "treatment failures" in 79 subjects assigned to placebo (3). In the second year of observation, there were two deaths and four deaths in subjects who previously had received cyclophosphamide versus placebo while worsening was seen in only four and one patient respectively (19). In the larger trial of bosentan, there were no deaths over one year in 163 patients. FVC% and DLCO% were strikingly stable over one year (36). However, 22.5% and 25.6% of patients in the bosentan and placebo groups respectively experienced "progression" thresholds for FVC and DLCO.

Responder frequency analysis thus appears practicable in future studies of ILD in SSc but will require robust standards for cohort enrichment, *i.e.* strategies to identify subjects at high risk of significant deterioration as presented below.

9. Cohort enrichment

Cohort enrichment describes strategies for selection of patients at greater risk of progression, based upon disease severity, observed progression or a short duration of systemic disease. Although cohort enrichment might be based on any single one of the criteria, it is recommended that patient meet the attainment of a primary criterion of severity threshold based on HRCT and FVC% and include one or both of two minor criteria: a) short duration of systemic disease (defined as first 6 years after onset of signs or symptoms attributable to SSc from first non-Raynaud's signs or symptoms) or b) observed progression (>10% of FVC over the past 3-12 months).

In the SLS-I, patients, placebo patients with baseline FVC% of ≤70% had a greater decline in the FVC% predicted at 18 months. Moreover, extent of fibrosis on the baseline HRCT scan was a significant predictor of worsening FVC% in the placebo group and of response to cyclophosphamide in the active treatment group, as indicated by a significant interaction of fibrosis with treatment (p=0.009) (3). In contrast, BAL cellularity at baseline was not a predictor of response (37). In another recent analysis of 215 SSc patients followed for 10 years (8), baseline PFTs and HRCTs were predictive of mortality risk. As described earlier, when HRCT involvement were combined in a simple dichotomous staging system, patients with extensive disease (HRCT extent >20% on rapid evaluation, or FVC <70% when HRCT extent was intermediate) had a much higher mortality risk (HR 3.40–3.80, p< 0.0005) than patients with mild disease (HRCT extent obviously <20%, or FVC >70% when HRCT extent was intermediate).

The enrolment of patients in therapeutic trials is subject to major selection bias, especially when a studied treatment can be routinely prescribed without trial participation. It can then be expected that referring physicians and patients will prefer open therapy when disease is more extensive or progressive, rather than accepting the possibility of inactive treatment in a placebo-controlled study (38). Thus, trial populations may consist of patients with milder disease and a lower likelihood of disease progression during the trial period: interventions directed towards the prevention of disease progression will be applicable only to a minority of patients with progressive disease. Tellingly, the treatment effect of SLS-I was confined to a minority with more extensive disease on HRCT. Moreover, when the management of patients in SLS cohort was passed to local physicians, at the end of the study period, active therapy was thought necessary in less than 15% of cases (19). Essentially, it appears likely that patients most likely to benefit may have been under-represented in the SLS I study (38). The similarity in patient baseline characteristics between the SLS I and the BUILD2 cohorts suggests that selection biases (both investigator and subject) present in the SLS study are likely to be common to treatment studies in SSc-ILD.

The inevitable consequence of underrepresentation of progressive disease is that large numbers of patients must be enrolled in order to power the detection of a treatment effect. Equally important, the average treatment effect across a cohort of patients with non-progressive disease will be small, leading to uncertainty as to the "clinical significance" of the intervention. The process of "cohort enrichment" consists of the selective enrolment of patients at a higher risk of disease progression in treatment studies, reducing the patient numbers required to demonstrate a treatment effect and increasing the average amplitude of such a benefit.

Three factors are considered to predict a higher likelihood of pulmonary disease progression in SSc. Of these, we believe more extensive pulmonary involvement to be the most important. The link between baseline pulmonary disease severity in SSc, as judged by the severity of pulmonary function impairment, and mortality has been quantified in many studies. However, linkage between disease severity at baseline and subsequent worsening of pulmonary function was not explored. However, in the SLS-I, disease progression in the placebo arm was seen in more extensive disease. Recently, a simple staging system, in which rapid HRCT quantification and FVC levels are integrated, has identified a subset of approximately 40% of patients with SSc-ILD, in which more extensive disease is linked to a much higher likelihood of disease progression within the next year (8). Other important enrichment variables include:

1) A shorter duration of systemic disease. The risk of progression is highest in the first four years of systemic disease, based on the observations of Steen in a large patient cohort (39).
2) Observed disease progression. It is generally accepted that, by analogy to reported findings in IPF (40-42), observed progression of lung disease, as judged by significant decline in FVC or DLCO, predicts a higher risk of future progression. It should be stressed that the link between observed pulmonary function trends and future disease progression has not been definitively studied in SSc-ILD.

In recommending that these criteria are met, as entry criteria, we acknowledge that in some patients, there will be insufficient longitudinal data to make an immediate assessment of disease progression. Moreover, it is not yet clear whether disease severity, observed disease progression and a short duration of systemic disease are all independent predictors of more progressive disease. However, a requirement that all three criteria be met, prior to enrolment in a treatment trial, would seem draconian.

One difficulty lies in a priori definitions of response. For example, consider a level of 5% decline in FVC and the influence of Bayes theorem. In a progressive disease like IPF, with a much higher likelihood of real decline, a 5% decline is more likely to represent real decline and relatively less likely to represent measurement variation.

Based on reproducibility (reliability) data, one would expect that in a group of 40 normal individuals, 6 of 40 (15%) would have a reduction in FVC by 5% due to measurement variation. In IPF, one would expect that 20 of 40 patients would have a decline to a threshold of 5%. Of the remaining 20, three would have spurious decline to 5% due to measurement variation (15%). Therefore, the likelihood that 5% represents true decline is 20:3 or roughly 85%. In SSc, without cohort enrichment, we might expect that perhaps five of 40 patients would reach the 5% threshold due to disease progression, with five of the remaining 35 patients (15%) having spurious decline. The likelihood that a 5% decline is real is just 5:5 or 50%. However, with cohort enrichment, again applying Bayes theorem to what is now a more progressive SSc cohort, a 5% decline becomes proportionally more likely to be real (perhaps 70% likely) – and this will be sufficient to provide robust whole cohort statements, even if a 5% change remains uncertain in an individual patient. In a nutshell, cohort enrichment allows us to use the 5% threshold.

10. Composite indices

SLS-I was associated with a variety of positive outcomes that spanned physiologic (FVC and TLC), anatomic (HRCT imaging), and patient-centred (HRQoL, transition dyspnea index [TDI]) parameters. While FVC was a priori defined as the primary outcome, one could question whether the most meaningful response to cyclophosphamide was its modest impact on pulmonary function or its more clinically-meaningful impact on patient-centred outcomes. Furthermore, it is possible that information obtained from one outcome might yield informative data related to other outcomes, resulting in a more robust measure of the overall treatment response. The identification of fibrosis score on HRCT as an essential covariate in the SLS confirmed this hypothesis. To move beyond the covariate analysis, a latent variable model has been considered (43). This approach provides an easy and intuitive mechanism for modelling the correlation among multiple responses and is especially useful for characterising the "overall treatment effect".

The model has three parts. First, a measurement model relating each outcome to important prognostic factors and to the latent variable (or generalisation to a very small number of latent variables). In addition, a random intercept is assumed that models the within subject correlation over time; a random term (usually called random error term) is also added to the measurement model. The second part of the model is the latent variable model, which relates the latent variable to the treatment and covariates, such as time and treatment by time interaction. In addition, a random intercept term is included to describe the heterogeneity of underlying disease status at 3 months. Also included in this linear mixed model for the latent variable is a random slope for time that models the individual trajectory of the disease status. The third part of the model is a missing data model that models non-ignorable missing observations (such as drop-outs due to disease progression, adverse events or death) on the outcomes separately.

For the SLS1 study, the measurement model was developed for FVC, TLC, TDI and the prognostic factors in the measurement model included baseline terms separately for each outcome. The measurement model is being expanded to include quantitative fibrosis as an additional outcome measure. All the factors in the measurement model are positively correlated. A composite index was analytically derived from this three-part model and shown to be normally distributed. In the SLS1 study the composite index combined information as a weighted average of the treatment effects among the three outcomes and led to highly significant treatment effects. The novelty of this three-part approach is the addition of the missing data model and the development of an analytical expression for the composite index.

11. Biomarkers

There is a critical need for non-invasive, clinically applicable biomarkers to improve the evaluation of patients with SSc-ILD, particularly in prospective, randomised controlled trials. Surfactant Protein D (SP-D) and Krebs von den Lungen-6 (KL-6) are glycoproteins secreted by type II pneumocytes that have emerged as possible surrogate markers for ILD, including SSc-ILD. Serum levels of SP-D and KL-6 rise when there is damage to the alveolar epithelium, including alveolar injury as seen in SSc- ILD (44-46). In relatively small series of patients, elevated levels of serum SP-D and/or KL-6 have been associated with decreases in FVC and DLCO and correlated with fibrosis on HRCT (45).

SP-D and KL-6 were measured in sera samples from patients screened for the SLS-I, including patients with and without alveolitis, and normal controls (44). Overall, SSc patients had significantly higher SP-D and KL-6 levels than did normal controls (p<0.0001 for each). SSc patients with alveolitis had significantly higher levels of both SP-D and KL-6 compared with SSc patients without alveolitis (p<0.001 for each). Receiver operating curve analysis demonstrated fairly good sensitivity and specificity for each glycoprotein in the assessment of alveolitis as defined in SLS-I (increase in BAL fluid polymorphonuclear cells and/or ground glass opacities on HRCT).

Other biomarkers are being studied. A recent study in 53 patients with SSc showed that serum levels of polymorphonuclear neutrophilic leukocyte (PMN) elastase were increased in patients with SSc, especially those with SSc-ILD, and that PMN elastase was positively correlated with SP-D and KL-6, making it a novel serologic marker for SSc-ILD (47).

Future studies, including the ongoing SLS II, will gather baseline and serial serum samples to determine whether biomarker signal at baseline predicts any treatment effect and whether short-term improvement in biomarker signal predicts a longer-term treatment effect. One important future role of biomarkers might lie in the selection of patients for trial enrolment, with biomarker estimation supplementing the staging of baseline severity, in identifying patients likely to deteriorate and/or to respond.

Other lessons learnt from previous studies

1. Change in FVC is a superior end-point compared to change in DLCO or the six-minute walk distance and can be regarded as a preferred primary end-point for future studies. Change in DLCO did not mirror treatment effects on FVC and other variables in 2 large RCTs (3, 5) and is an unsatisfactory primary end-point, influenced equally by changes in interstitial and vascular disease In addition, DLCO has marked measurement variability. We recommend not using DLCO as a primary end point.
2. Although the 6-minute walk test showed reliability during screening and baseline visits in the BUILD study, change in the six-minute walk distance exhibited striking long-term variability during the trial and is an unsatisfactory primary end-point. In addition, six-minute walk distance is confounded by co-existing pulmonary hypertension, cardiopulmonary deconditioning, and musculoskeletal limitations (48).
3. Significant pulmonary hypertension should be excluded by right heart catheterisation (defined as mean pulmonary arterial pressure of greater than 25 mmHg) or echocardiogram with Doppler (estimated right ventricular pressure of >=45 mmHg or other signs suggestive of right atrial/ventricular strain) before enrolling in SSc-ILD studies as it may interfere with assessment of dyspnea and DLCO.

Conclusion

We present preliminary recommendations on design of future SSc-ILD RCTs. We see this review as the starting point and welcome comments and criticisms. We anticipate revision as more trial data become available.

Acknowledgements

Dr. Khanna was supported by a National Institutes of Health Award (NIAMS K23 AR053858-04) and the Scleroderma Foundation (New Investigator Award). Dr. Goldin was supported by the National Institutes of Health Award (RO1) and is a Director of MedQIA Imaging Core.

References

1. STEEN VD, MEDSGER TA JR: Changes in causes of death in systemic sclerosis. *Ann Rheum Dis* 2007; 66: 940-4.
2. WHITE B, MOORE WC, WIGLEY FM *et al.*: Cyclophosphamide is associated with pulmonary function and survival benefit in patients with scleroderma and alveolitis. *Ann Intern Med* 2000; 132: 947-54.
3. TASHKIN DP, ELASHOFF R, CLEMENTS PJ *et al.*: Cyclophosphamide versus placebo in scleroderma lung disease. *N Engl J Med* 2006; 354: 2655-66.
4. CLEMENTS PJ, ROTH MD, ELASHOFF R *et al.*: Scleroderma lung study (SLS): differences in the presentation and course of patients with limited versus diffuse systemic sclerosis. *Ann Rheum Dis* 2007; 66: 1641-7.
5. HOYLES RK, ELLIS RW, WELLSBURY J *et al.*: A multicenter, prospective, randomized, double-blind, placebo-controlled trial of corticosteroids and intravenous cyclophosphamide followed by oral azathioprine for the treatment of pulmonary fibrosis in scleroderma. *Arthritis Rheum* 2006; 54: 3962-70.
6. BOUROS D, WELLS AU, NICHOLSON AG *et al.*: Histopathologic subsets of fibrosing alveolitis in patients with systemic sclerosis and their relationship to outcome. *Am J Respir Crit Care Med* 2002; 165: 1581-6.
7. KAZEROONI EA, MARTINEZ FJ, FLINT A *et al.*: Thin-section CT obtained at 10-mm increments versus limited three-level thin-section CT for idiopathic pulmonary fibrosis: correlation with pathologic scoring. *AJR Am J Roentgenol* 1997; 169: 977-83.
8. GOH NS, DESAI SR, VEERARAGHAVAN S *et*

al.: Interstitial lung disease in systemic sclerosis: a simple staging system. *Am J Respir Crit Care Med* 2008; 177: 1248-54.

9. KIM HJ, LI G, GJERTSON D *et al*.: Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study. *Acad Radiol* 2008; 15: 1004-16.
10. GOLDIN J, ELASHOFF R, KIM HJ *et al*.: Treatment of scleroderma-interstitial lung disease with cyclophosphamide is associated with less progressive fibrosis on serial thoracic high-resolution CT scan than placebo: findings from the scleroderma lung study. *Chest* 2009; 136: 1333-40.
11. KIM H, GOLDIN J, TASHKIN D *et al*.: Cyclophosphamide Versus Placebo in Scleroderma Lung Study Using Quantitative Lung Fibrosis Score. 2009 International Conference of the Am. 2009: A3943.
12. American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. This joint statement of the American Thoracic Society (ATS), and the European Respiratory Society (ERS) was adopted by the ATS board of directors, June 2001 and by the ERS Executive Committee, June 2001. *Am J Respir Crit Care Med* 2002; 165: 277-304.
13. KEMERINK GJ, LAMERS RJ, THELISSEN GR *et al*.: CT densitometry of the lungs: scanner performance. *j* 1996; 20: 24-33.
14. KEMERINK GJ, KRUIZE HH, LAMERS RJ *et al*.: CT lung densitometry: dependence of CT number histograms on sample volume and consequences for scan protocol comparability. *J Comput Assist Tomogr* 1997; 21: 948-54.
15. AZIZ ZA, WELLS AU, HANSELL DM *et al*.: HRCT diagnosis of diffuse parenchymal lung disease: inter-observer variation. *Thorax* 2004; 59: 506-11.
16. HUNNINGHAKE GW, ZIMMERMAN MB, SCHWARTZ DA *et al*.: Utility of a lung biopsy for the diagnosis of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2001; 164: 193-6.
17. MAYO JR, ALDRICH J, MULLER NL: Radiation exposure at chest CT: a statement of the Fleischner Society. *Radiology* 2003; 228: 15-21.
18. VAN DER BRUGGEN-BOGAARTS BA, BROERSE JJ, LAMMERS JW *et al*.: Radiation exposure in standard and high-resolution chest CT scans. *Chest* 1995; 107: 113-5.
19. TASHKIN DP, ELASHOFF R, CLEMENTS PJ *et al*.: Effects of 1-year treatment with cyclophosphamide on outcomes at 2 years in scleroderma lung disease. *Am J Respir Crit Care Med* 2007; 176: 1026-34.
20. PUNJABI NM, SHADE D, PATEL AM *et al*.: Measurement variability in single-breath diffusing capacity of the lung. *Chest* 2003; 123: 1082-9.
21. HATHAWAY EH, TASHKIN DP, SIMMONS MS: Intraindividual variability in serial measurements of DLCO and alveolar volume over one year in eight healthy subjects using three independent measuring systems. *Am Rev Respir Dis* 1989; 140: 1818-22.
22. PENNOCK BE, ROGERS RM, MCCAFFREE DR: Changes in measured spirometric indices. What is significant? *Chest* 1981; 80: 97-9.
23. WANGER J, CLAUSEN JL, COATES A *et al*.: Standardisation of the measurement of lung volumes. *Eur Respir J* 2005; 26: 511-22.
24. MACINTYRE N, CRAPO RO, VIEGI G *et al*.: Standardisation of the single-breath determination of carbon monoxide uptake in the lung. *Eur Respir J* 2005; 26: 720-35.
25. MILLER MR, HANKINSON J, BRUSASCO V *et al*.: Standardisation of spirometry. Eur Respir J 2005; 26: 319-38.
26. AU K, KHANNA D, CLEMENTS PJ *et al*.: Current concepts in disease-modifying therapy for systemic sclerosis-associated interstitial lung disease: lessons from clinical trials. *Curr Rheumatol Rep* 2009; 11: 111-9.
27. BOERS M, BROOKS P, STRAND CV *et al*.: The OMERACT filter for Outcome Measures in Rheumatology. J *Rheumatol* 1998; 25: 198-9.
28. KHANNA D, CLEMENTS PJ, FURST DE *et al*.: Correlation of the degree of dyspnea with health-related quality of life, functional abilities, and diffusing capacity for carbon monoxide in patients with systemic sclerosis and active alveolitis: results from the Scleroderma Lung Study. *Arthritis Rheum* 2005; 52: 592-600.
29. KHANNA D, YAN X, TASHKIN DP *et al*.: Impact of oral cyclophosphamide on health-related quality of life in patients with active scleroderma lung disease: results from the scleroderma lung study. *Arthritis Rheum* 2007; 56: 1676-84.
30. KHANNA D, TSENG CH, FURST DE *et al*.: Minimally important differences in the Mahler's Transition Dyspnoea Index in a large randomized controlled trial--results from the Scleroderma Lung Study. *Rheumatology* (Oxford) 2009; 48: 1547-40.
31. CHARLES C, CLEMENTS P, FURST DE: Systemic sclerosis: hypothesis-driven treatment strategies. *Lancet* 2006; 367: 1683-91.
32. FLEMING JN, NASH RA, MCLEOD DO *et al*.: Capillary regeneration in scleroderma: stem cell therapy reverses phenotype? *PLoS One* 2008; 3: e1452.
33. YAMANE K, IHN H, KUBO M *et al*.: Serum levels of KL-6 as a useful marker for evaluating pulmonary fibrosis in patients with systemic sclerosis. *J Rheumatol* 2000; 27: 930-4.
34. MERKEL PA, SILLIMAN N, CLEMENTS P *et al*.: Patterns and Predictors of Change in Outcome Measures in Clinical Trials in Scleroderma. 2005.
35. AMERICAN THORACIC SOCIETY: Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. American Thoracic Society (ATS), and the European Respiratory Society (ERS). *Am J Respir Crit Care Med* 2000; 161 (Pt. 1): 646-64.
36. SEIBOLD J, DENTON CP, GUILLEVIN L *et al*.: Randomized, prospective, placebo-controlled trial of bosentan in interstitial lung disease secondary to systemic sclerosis. 2010 (in press).
37. STRANGE C, BOLSTER MB, ROTH MD *et al*.: Bronchoalveolar lavage and response to cyclophosphamide in scleroderma interstitial lung disease. *Am J Respir Crit Care Med* 2008; 177: 91-8.
38. WELLS AU, LATSI P, MCCUNE WJ: Daily cyclophosphamide for scleroderma: are patients with the most to gain underrepresented in this trial? *Am J Respir Crit Care Med* 2007; 176: 952-3.
39. STEEN VD, CONTE C, OWENS GR *et al*.: Severe restrictive lung disease in systemic sclerosis. *Arthritis Rheum* 1994; 37: 1283-9.
40. FLAHERTY KR, MUMFORD JA, MURRAY S *et al*.: Prognostic implications of physiologic and radiographic changes in idiopathic interstitial pneumonia. *Am J Respir Crit Care Med* 2003; 168: 543-8.
41. COLLARD HR, KING TE JR, BARTELSON BB *et al*.: Changes in clinical and physiologic variables predict survival in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2003; 168: 538-42.
42. LATSI PI, DU BOIS RM, NICHOLSON AG *et al*.: Fibrotic idiopathic interstitial pneumonia: the prognostic value of longitudinal functional trends. *Am J Respir Crit Care Med* 2003; 168: 531-7.
43. YAN X, ELASHOFF R: Multivariate latent variable analysis with non-ignorable missing data (In Press). *Computational statistics and data analysis* 2010.
44. HANT FN, LUDWICKA-BRADLEY A, WANG HJ *et al*.: Surfactant protein D and KL-6 as serum biomarkers of interstitial lung disease in patients with scleroderma. *J Rheumatol* 2009; 36: 773-80.
45. ASANO Y, IHN H, YAMANE K *et al*.: Clinical significance of surfactant protein D as a serum marker for evaluating pulmonary fibrosis in patients with systemic sclerosis. *Arthritis Rheum* 2001; 44: 1363-9.
46. SATO S, NAGAOKA T, HASEGAWA M *et al*.: Elevated serum KL-6 levels in patients with systemic sclerosis: association with the severity of pulmonary fibrosis. *Dermatology* 2000; 200: 196-201.