Seminars in Arthritis and Rheumatism 51 (2021) 1331–1341

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

journal homepage: www.elsevier.com/locate/semarthrit

Outcome measurement instrument selection for lung physiology in systemic sclerosis associated interstitial lung disease: A systematic review using the OMERACT filter 2.1 process

David Roofeh[a,&,+], Shaney L. Barratt[b,&,+], Athol U Wells[c,+], Leticia Kawano-Dourado[d,+], Donald Tashkin[e,+], Vibeke Strand[f,+], James Seibold[g,+], Susanna Proudman[h,+], Kevin K Brown[i,+], Paul F Dellaripa[j,+], Tracy Doyle[k,+], Thomas Leonard[l,+], Eric L Matteson[m,+], Chester V Oddis[n,+], Joshua J Solomon[i,+], Jeffrey A Sparks[j,+], Robert Vassallo[o,+], Lara Maxwell[p,#], Dorcas Beaton[q,#], Robin Christensen[r,#], Whitney Townsend[s], Dinesh Khanna[a,+,*]

[a] Department of Internal Medicine, Division of Rheumatology, Scleroderma Program, University of Michigan, Ann Arbor, Michigan, USA
[b] Academic Respiratory Unit, School of Clinical Sciences, University of Bristol, UK; Bristol Interstitial Lung Disease Service, North Bristol NHS Trust, Southmead, Bristol, UK
[c] Department of Internal Medicine, Division of Pulmonology, Royal Brompton Hospital and National Heart and Lung Institute; London, UK
[d] HCor Research Institute, Hospital do Coração, São Paulo, Brazil; Pulmonary Division, Heart Institute (InCor), University of Sao Paulo Medical School, São Paulo, Brazil
[e] Department of Medicine, Division of Pulmonary and Critical Care Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California, USA
[f] Division of Immunology/Rheumatology, Stanford University, Palo Alto, California, USA
[g] Scleroderma Research Consultants, Aiken, South Carolina, USA
[h] Rheumatology Unit, Royal Adelaide Hospital and Professor Discipline of Medicine, University of Adelaide, Adelaide, AUS
[i] Department of Medicine, Division of Pulmonary, Critical Care and Sleep Medicine, National Jewish Hospital, Denver, Colorado, USA
[j] Department of Medicine, Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[k] Department of Internal Medicine, Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[l] From Clinical Development and Medical Affairs, Specialty Care Boehringer Ingelheim Pharmaceuticals, Inc. USA
[m] Department of Internal Medicine, Division of Rheumatology, Mayo Clinic, Rochester, Minnesota, USA
[n] Division of Rheumatology and Clinical Immunology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
[o] Department of Internal Medicine, Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, Minnesota, USA
[p] Faculty of Medicine, University of Ottawa, Ottawa, CA
[q] Institute for Work & Health and Institute for Health Policy Management and Evaluation, University of Toronto, Toronto, CA
[r] Musculoskeletal Statistics Unit, the Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, & Research Unit of Rheumatology, Department of Clinical Research, University of Southern Denmark, Odense University Hospital, Odense, Denmark
[s] Taubman Health Sciences Library, University of Michigan, Ann Arbor, MI, USA

ARTICLE INFO

Keywords:
Systemic sclerosis interstitial lung disease
OMERACT
Outcome measures
Core set
Forced vital capacity
Diffusion capacity of carbon monoxide

ABSTRACT

Objective: The Outcome Measures in Rheumatology (OMERACT) is a research organization focused on improving health care outcomes for patients with autoimmune and musculoskeletal diseases. The Connective Tissue Disease-Interstitial Lung Disease (CTD-ILD) Working Group on Lung Physiology is a group within OMERACT charged with identifying outcome measures that should be implemented in studies of patients with CTD-ILD. The OMERACT Filter 2.1 is an evidence-based algorithm used to identify outcome measures that are truthful, feasible, and able to discriminate between groups of interest. Our objective was to summate evidence (published literature, key opinion leader input, patient perspectives) that would influence the CTD-ILD Working Group's vote to accept or reject the use of two measures of lung physiology, the forced vital capacity (FVC) and the diffusion capacity of carbon monoxide (DLco) for use in randomized controlled trials (RTCs) and longitudinal observational studies (LOSs) involving patients with systemic sclerosis associated ILD (SSc-ILD).

Methods: Patient Research Partners (those afflicted with SSc-ILD) and the CTD-ILD Working Group on Lung Physiology were polled to assess their opinion on the FVC and DLco in terms of feasibility; the CTD-ILD

* Corresponding author.
E-mail address: khannad@umich.edu (D. Khanna).
& Both contributed equally
+ Working Group Member
# Technical Advisory Group Member

https://doi.org/10.1016/j.semarthrit.2021.08.001

Working Group was also queried on these instruments' face and content validity. We then conducted a systematic literature review to identify articles in the SSc-ILD population that assessed the following measurement properties of FVC and DLco: (1) construct validity, (2) test-retest reliability, (3) longitudinal construct validity, (4) clinical trial discrimination/sensitivity to detect change in clinical trials, and (5) thresholds of meaning. Results were summarized in a Summary of Measurement Properties (SOMP) table for each instrument. OMERACT CTD-ILD Working Group members discussed and voted on the strength of evidence supporting these two instruments and voted to endorse, provisionally endorse, or not endorse either instrument.
Results: Forty Patient Research Partners reported these two measures are feasible (are not an unnecessary burden or represent an infeasible longitudinal assessment of their disease). A majority of the 18 CTD-ILD Working Group members voted that both the FVC and DLco are feasible and have face and content validity. The systematic literature review returned 1,447 non-duplicated articles, of which 177 met eligibility for full text review. Forty-eight studies (13 RCTs, 35 LOSs) were included in the qualitative analysis. The FVC SOMP table revealed high quality, consistent data with evidence of good performance for all five measurement properties, suggesting requisite published evidence to proceed with endorsement. The DLco SOMP table showed a lack of data to support test-retest reliability and inadequate evidence to support clinical trial discrimination. There was unanimous agreement (15 [100%]) among voting CTD-ILD Working Group members to endorse the FVC as an instrument for lung physiology in RCTs and LOSs in SSc-ILD. Based on currently available evidence, DLco did not meet the OMERACT criteria and is not recommended for use in RCTs to represent lung physiology of SSc-ILD. The OMERACT Technical Advisory Group agreed with these decisions.
Conclusion: The OMERACT Filter 2.1 was successfully applied to the domain of lung physiology in patients with SSc-ILD. The FVC was endorsed for use in RCTs and LOSs based on the Working Group's vote; DLco was not endorsed.

1. Introduction

Several connective tissue diseases may be complicated by interstitial lung disease (CTD-ILD): systemic sclerosis (SSc), idiopathic inflammatory myopathies, rheumatoid arthritis, mixed-CTD, Sjogren's syndrome, systemic lupus erythematosus, and undifferentiated-CTD. Systemic sclerosis associated interstitial lung disease (SSc-ILD) is the inflammatory and fibrotic result of autoimmune-driven thickening of the pulmonary interstitium. As a result of its prevalence and impact on morbidity and mortality, SSc-ILD has garnered the greatest amount of high quality research among the CTDs [1–4]. Although up to 80% of patients have interstitial changes on high resolution computerized tomographic (HRCT) of the chest, not all patients with SSc-ILD require treatment for ILD [5]. Its severity is clinically heterogeneous [6,7] and may be defined by its impact on lung physiology measured by pulmonary function testing. These measures have been used to predict SSc-ILD outcomes [8–11], develop treatment algorithms [12–14], and are used as outcome measures in SSc-ILD randomized clinical trials (RCTs) [15–17].

The OMERACT Filter 2.0 framework standardizes the measurable aspects of a health condition into core domain sets (what should be measured) and core outcome measurement sets (how to measure a domain) [18,19]. Lung physiology is a core domain in the study of CTD-ILD [20]. The CTD-ILD Working Group is now tasked with instrument selection to measure this domain, employing the OMERACT Filter 2.1. In this implementation of the Filter, two measures of pulmonary physiology were assessed as outcome measurement instruments: forced vital capacity (FVC) and the diffusion capacity of carbon monoxide (DLco).

Pulmonary physiology has been routinely used to measure the severity of lung disease in SSc-ILD since the 1950s with the intent to monitor disease and treatment effects. Abnormalities in lung function reflect the effects of interstitial inflammation and/or scarring, resulting in a restrictive ventilatory defect and impaired gas exchange, as well as reduced compliance [21]. The underlying pathophysiology leads to having smaller, stiffer lungs requiring more work to inflate with impaired oxygen diffusion caused by pathologic interstitial and alveolar changes and pulmonary vascular defects [22].

1.1. Forced vital capacity

The FVC is a measurement of expiration and specifically is an indirect measure of the flow-resistive properties of the lung [23]. It is the total amount of air that can be forcefully expelled from the lungs beginning at total lung capacity and is the sum of the tidal volume, inspiratory reserve volume, and expiratory reserve volume. The relationship between expired volume and time during an FVC maneuver informs about airflow during expiration and about the volume of air in the lungs expired within designated time intervals. The reduction in vital capacity in the context of infiltrative diseases, such as SSc-ILD, generally reflects having fewer functional alveolar units [24]. Ensuring accurate results (distinguishing the effects of disease from normal variability) requires consistent, maximal patient effort, technician training, and calibration of equipment [25]. Reference values are determined by reference equations (using factors like height, sex, age, race/ethnicity) and reference subjects (which should include asymptomatic non-smokers with no known exposures or respiratory diseases) [26].

Severe restriction (FVC % predicted $<$50%) at diagnosis and short-term progressive reduction in FVC (over 2 years) have been associated with increased mortality in SSc-ILD [27,28]. FVC has become the primary endpoint in landmark trials in SSc-ILD including the Scleroderma Lung Study I and II (ClinicalTrials.gov Identifier: NCT00004563 and NCT00004563, respectively) and the SENSCIS trial that led to the first FDA-approved medication in SSc-ILD (ClinicalTrials.gov Identifier: NCT03313180) [29–31]. Experts have proposed this outcome measure to have validity, feasibility, reliability, and sensitivity to change with treatment in SSc-ILD [16,32,33]. In a recent systematic review, FVC was the primary outcome measure in approximately 70% of 169 outcome studies on SSc-ILD [29]. Experts have argued that serial change in FVC % predicted (defined as percentage change from baseline) may be the best single outcome measure in lung disease in SSc-ILD [34].

1.2. Diffusion capacity of carbon monoxide (Table 4)

The DLco can be thought of as an assay for pulmonary parenchymal and vascular health [22]. It requires the participant to breathe in a dilute mixture of a known amount of carbon monoxide (CO) and hold his/her breath at total lung capacity for 10 s. The quantity of CO transferred from the alveoli to the pulmonary circulation in a 10 s period is calculated by subtracting the amount of CO exhaled after the 10 s from the initial amount inhaled. A normal diffusion capacity requires a normal pulmonary gas-exchanging surface, normal capillary blood volume and hemoglobin, and homogeneous regional ventilation-perfusion relationships [25]. Assuming proper technique,

defects in any of these or their combination produce deficits in the reported DLco. As this is a calculated measure and dependent upon hemoglobin level, it must be interpreted with caution in the presence of a co-existent anemia [35].

The advantage of measuring DLco may be early detection of ILD. Impaired surface area for the transfer of gases from the alveoli to the pulmonary capillaries can be one of the earliest features seen in ILD [36]. Its decline correlates with the severity of pulmonary fibrosis as measured by HRCT. According to findings from the Scleroderma Lung Studies I and II, the DLCO was the pulmonary function measure that correlated the best with the extent of ILD on HRCT [37]. The disadvantage is that DLco is not specific; it is affected by pulmonary vascular disease (which often co-occurs with ILD in patients with SSc-ILD), airflow limitation, and chronic obstructive disease/emphysema. Further, DLco may be normal in one third of patients despite impaired cardiopulmonary exercise testing and documented fibrosis by lung biopsy or CT scan [38].

Reference values for DLco have not been published by the American Thoracic Society due to the wide divergence of available reference values [26]. Table 6 in MacIntyre et al., 2005 describes an acceptable test and this reference outlines the standards of equipment use, techniques, calculations, and evaluations of the DLco measurements [39]. The European Respiratory Society has assembled the largest collection of normative data (12,660 individuals in 14 countries, white-only) to establish reference equations from age 5 to 85 years [40]. DLco has also been proposed to have validity, feasibility, reliability, but there is a paucity of data to show its ability to discriminate between treatment arms [32,33,41,42].

1.3. OMERACT filter 2.1

To date, there has not been an effort to systematically evaluate the measurement properties of these instruments in SSc-ILD. The OMERACT Filter 2.1 offers a process framework to assess an instrument's feasibility, truth, and discrimination. Feasibility refers to the degree to which an instrument is reasonable to perform as an outcome measure, without posing an undue burden to the patient. Truth refers both to the instrument's face and content validity (its subjective ability to match its target domain and to measure all facets of the target domain) [43]. These two screening elements must be agreed upon as they represent the fundamental basis for using the instrument as an outcome measure.

If agreed upon by Patient Research Partners and Working Group members, the instrument is evaluated in terms of five key measurement properties of the instrument: construct validity, test and retest reliability, longitudinal construct validity, clinical trial discrimination, and threshold of meaning. A systematic literature review (SLR) is employed to evaluate the evidence supporting these five properties. In the final stages of the Filter 2.1, the working group performs a vote on each measurement property of the instrument and ultimately decides to endorse, provisionally endorse with caution, or not endorse the instrument.

2. Materials and methods

2.1. Assessment of feasibility and truth

2.1.1. Patient research partner assessments

Forty Patient Research Partners with ILD were invited to assess these pulmonary measurements in terms of comfort, feasibility (time and effort), and any anxiety related to their test performance. Specifically, patients were asked about procedural comfort, the burden of their time requirement (both travel to the pulmonary testing lab and time to complete the studies), whether they understood their test results, and if the test and results prompted them to have anxiety. These patients were selected from two major research institutions (University of Michigan Health System and the Bristol Interstitial Lung Disease Service at North Bristol NHS Trust), and completed two surveys, one for FVC and one for DLco, with questions regarding their experiences with testing. The results of the patient surveys were not shared with the Working Group until after the Working Group voted on feasibility.

2.1.2. Working group assembly

The Working Group comprised three groups: 1) Co-chairs, 2) Junior Members, 3) Experts in SSc-ILD and Lung Physiology. Group 1 (DK, AUW, SP) fulfilled the OMERACT requirement to have three co-chairs from three continents; Group 2 (DR, SLB) were investigational co-leads and shared work equally in conducting the SLR and voting sessions under the supervision of a Co-chair (DK); Group 3 (LKD, DT, VS, JS, KKB, PFD, TD, TL, ELM, CVO, JJS, JAS, and RV) were invited based on their significant contribution to published work in lung physiology and SSc-ILD. All those invited to participate in the Working Group agreed to be members; there were no invited members that declined. Group 2 organized and prepared material for the Working Group votes; Group 1 offered recommendations and revisions to that material in anticipation of Working Group votes. All Working Group members were allowed to vote.

2.1.3. Working group vote: feasibility and truth

Two screening elements, feasibility and truth (face validity and content validity), were evaluated by the Working Group for both FVC and DLco. A survey was sent via Qualtrics in October 2019 to members of the Working Group and invited them to participate in on-line voting. Working Group members were given definitions of feasibility (if the instrument takes a reasonable amount of time to complete and does not provide an undue burden on the patient), face validity (if the instrument is subjectively viewed as covering the concept it purports to measure), and content validity (if the instrument represents all facets and important elements of a given construct). Respondents were asked to vote 'yes' or 'no' for these three features of each instrument, and also asked to provide comments to qualify their answers. Two separate 1 h teleconferences were then conducted so that the Working Group could discuss the results and review comments made by the experts. During the interactive discussions, all Working Group members had the opportunity to revise their answer after discussing each other's qualifying statements. If the Working Group achieved $\geq$70% agreement, the instrument's candidacy moved forward through the Filter 2.1 process.

2.2. Systematic literature review of measurement properties

2.2.1. Identifying literature

The study protocol was prepared in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. An expert librarian at University of Michigan was consulted to facilitate this process (WT). Table 1 shows the outline of the search strategy to identify relevant articles. Only articles that provided an assessment of primary data using these instrument measures were considered for eligibility. Primary data in the form of longitudinal observational studies (LOSs) with $n < 100$ were excluded, cultivating a uniform caliber of study quality with larger, more robust and better designed published studies. Systematic reviews, literature reviews, and clinical trials with single-arm designs were excluded. These refinements in study inclusion and exclusion were approved by the Technical Advisory Group.

2.2.2. Good methods assessment

Each article that met criteria for full text review was independently assessed by two raters (DR & DK or SLB & DK) using a rubric provided by the OMERACT Technical Advisory Group (see Beaton et al., 2019 for further details of this rubric, as well as the OMERACT

Table 1
Summary of the database search strategy for both selected instruments.

Database Search	Search Property
Population	Systemic Sclerosis associated interstitial lung disease
Selected Instrument(s)	Forced Vital Capacity Diffusion Capacity of Carbone Monoxide
Measurement Properties	Construct Validity Test-Retest Reliability Longitudinal Construct Validity Discrimination in Randomized Controlled Trials Thresholds of Meaning
Publication Language(s)	English
Databases	Pubmed (Medline) Embase.com Scopus Web of Science Core Collection CINAHL CENTRAL Clinicaltrials.gov

Website (https://omeract.org/resources) to ensure that only those articles with low or limited risk of methodological bias were included as potential evidence [43]. Studies on FVC and DLco were considered separately. Articles that were deemed to use "Good Methods" (a Green or Amber rating on 'good methods assessment') were then selected for further review of performance assessment.

2.2.3. Review of the performance of the instruments

Each article with "Good Methods" was assessed for numerical evidence that the instrument demonstrates at least adequate performance for a given measurement property.

2.2.3.1. Construct validity. The FVC and DLco were assessed for performance on construct validity, or the strength of correlation between the instrument and another instrument measuring the same or similar domains reflecting morbidity. Performance of observational studies was classified as "+" (for positive performance), "+/-" (for equivocal performance), and "-" (for negative performance). These classifications were based on a ratio of the observed strength of relationship in a given study compared to the expected strength of relationship. That is, the strength of correlation coefficients ('r') between the instrument and another index of measurement (e.g., symptoms, HRCT imaging, cardiopulmonary functional testing, lung physiology morbidity/mortality) was interpreted against the expected results based on an *a priori* (before consideration of the analysis) relationship between two variables.

The correlation coefficients reported in each study was interpreted as negligible (0.0–0.19), small (0.2–0.39), moderate (0.4–0.59), strong (0.6–0.79), or very strong (0.8–1.0) [44]. Table 2 shows the *a priori* expected relationships of the FVC and DLco to other assessments. The strength of these *a priori* relationships were chosen based on the previously reported associations in the published literature; these were agreed upon by the Working Group members.

2.2.3.2. Test-retest reliability. Test-retest reliability was measured using intra-class correlation coefficient (ICC), with >0.75 considered adequate for performance ("+" was assigned if the ICC was >0.75 [47]). Data were obtained from the screening and baseline FVC% in the Scleroderma Lung Study-I and Scleroderma Lung Study-II, as measurements at these two time points in these two studies were considered a situation of no change over a mean of 34 days between testing [45,46]. A potential systematic error was not taken into account when calculating the ICC.

2.2.3.3. Longitudinal construct validity. Longitudinal construct validity (responsiveness) is the ability to detect change in situations of change; that is, the instrument is able to capture the direction and magnitude of disease progression patients experience when disease progression is present. Standardized estimates of change (the effect size and the standardized response mean) were used to gauge the responsiveness of the instruments to change, and performance was determined by evaluating the *a priori* (expected) vs observed change over time: "+" (for positive performance), "+/-" (for equivocal performance), and "-" (for negative performance).

The *a priori* assessment of disease behavior without treatment is limited by a paucity of data detailing the natural course of disease over time. Factors influencing this interpretation included progression seen in different subsets of SSc-ILD (expected rate of loss of FVC or DLco in patients with limited vs diffuse SSc) and duration of disease from its onset (expected rate of loss early vs late in the disease). Often studies do not report an *a priori* hypothesis of change. For those articles with no *a priori* hypotheses reported, we assessed if the study's authors used an anchor to estimate that change has occurred (e.g., a bigger effect size seen in those with worse patient-reported outcomes like dyspnea, cough, health-related quality of life over time, or high-resolution chest CT imaging demonstrating change over time, cardiopulmonary function testing change over time, or morbidity/mortality measurements).

2.2.3.4. Clinical trial discrimination. Sensitivity to detect change in the context of a randomized controlled trial was evaluated by the magnitude of treatment effect, relative to the anticipated treatment effect. We used the sample size calculations (when reported) to understand the *a priori* estimated mean and SD changes in the FVC. This was not possible for the DLco because none of the clinical trials used DLco as the primary endpoint. An overall performance score was based on observed/expected effect size ratio, however most studies did not explicitly state its expected effect size.

Historically, the expected rate of loss of lung function in FVC or DLco in patients with SSc-ILD were lacking when the first large trials were designed. The estimates of FVC or DLco changes were taken from observational cohorts and the decline seen in placebo groups. The estimates for the active treatment were based on expert input and consensus. Similar to longitudinal construct validity, these estimations are limited by factors related to the SSc disease process (expected rate of loss of FVC or DLco in patients with limited vs diffuse SSc; expected rate of loss early vs late in the disease) and concurrent treatment (the effect of concomitant medications on FVC or DLco loss).

Table 2
Expected *a priori* relationships between instruments and outcome measures.

Instrument	*A Priori* Outcome Measure Correlation*					
	Dyspnea	Cough	Heath Related Quality of Life	HRCT	Hallwalk	Morbidity/Mortality
FVC	Small to Moderate	Small to Moderate	Negligible to Small	Small to Moderate	Small to Moderate	Small to Moderate
DLco	Small to Moderate	Small to Moderate	Negligible to Small	Small to Moderate	Small to Moderate	Small to Moderate

* Negligible (0.0–0.19), small (0.2–0.39), moderate (0.4–0.59), strong (0.6–0.79), or very strong (0.8–1.0)

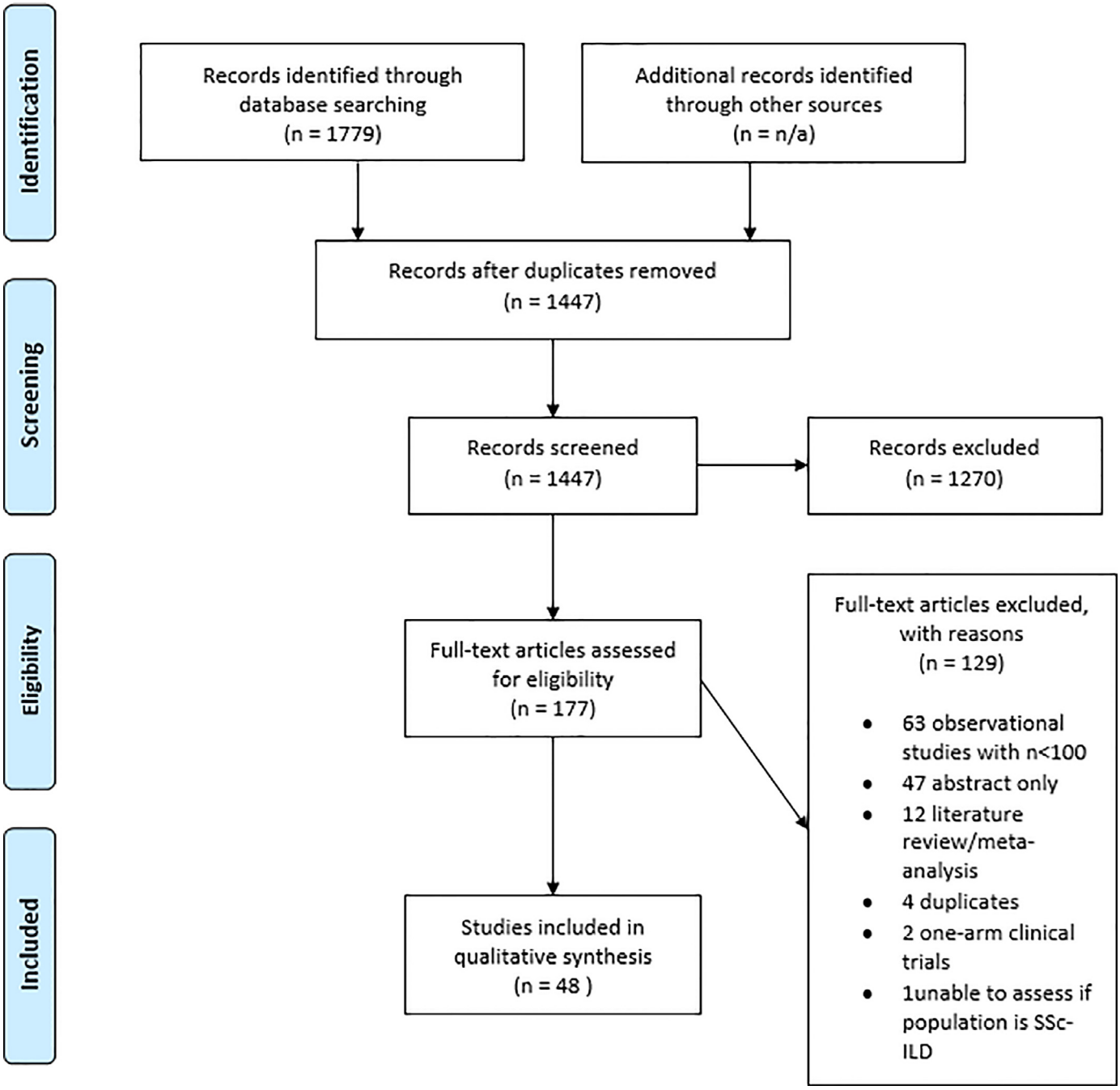

Fig. 1. PRISMA Systematic literature review flow diagram.

For each RCT reporting baseline and follow-up mean and standard deviation values, the relative change from baseline was estimated as the standardized response mean. Values of 0.20, 0.50, and 0.80 indicate small, moderate, and large effect sizes, respectively [48]. For studies not reporting these data (e.g., reporting percentages of patients with FVC improvement of 10% or rate of decline in FVC milliliters over time) but meeting the pre-specified outcomes (that is, author-determined meaningful differences in the proportion of patients in the treatment arm vs the control arm) received a "+". If those conditions were not met, then the study received a "-" (inadequate performance).

2.2.3.5. Thresholds of meaning. Thresholds of meaning are parameters identifying clinically significant changes in health status; this property identifies the amount of measurement change on a continuous instrument scale that provides a meaningful improvement or decline at a group level. The minimal important difference (MID) is a threshold of meaning [49]. We extracted the characteristics from relevant studies, described the results, and identified studies calculating MID based on several anchors and clinically significant benchmark scores. Articles achieving MID based on anchors and clinically significant benchmark scores received a "+". MID based on distribution (i.e., effect size) without anchors received a "-".

2.3. Criteria for final ratings and endorsement

Once each article was filtered through a lens of good methods assessment and performance evaluation, it was included in the summary of measurement properties (SOMP) table. The SOMP table provided a quantifiable estimate of quality studies and an ability to

Table 3
Forced Vital Capacity Summary of Measurement Properties. [56–98]

	Author/year	Ref	Truth	Feasibility	Truth	Discrimination			
			Domain match		Construct validity	Test retest reliability	Longitudinal construct validity	Clinical trial discrimination	Thresholds of meaning
Clinical Trials	Working Group Appraisal		+	+					
	Patient Research Partners		+	+					
	Burt 2011	56						+	
	Daoussis 2017	57						+	
	Distler 2019	58						+	
	Domiciano 2011	59						-	
	Fraticelli 2014	60							
	Hoyles 2006	61						+	
	Khanna 2011	62							
	Khanna 2018	63						+	
	Pakas 2002	64						+	
	Sircar 2018	65						+	
	Sullivan 2018	66						+	
	Tashkin 2006	30						+	
	Tashkin 2016	31						+	
Observational Studies	Assassi 2010	67							
	Baron 2008	68							
	Buch 2007	69			-				
	Goh 2008	70			+				+
	Goh 2017	71					+		+
	Goldin 2008	72			+				
	Goldin 2009	73			+		+/-		
	Goldin 2018	74					+/-		
	Guler 2018	75					+/-		
	Hax 2017	76							+
	Hoffmann-Vold 2015	77			+		+		
	Kafaja 2018*	50				+ +			+
	Khanna 2005	78			+/-				
	Khanna 2009	79					+		
	Khanna 2015	80					+		
	Mango 2018	10							
	Moore 2013	81			-		-		+/-
	Moore 2015	82			+		+		
	Morisset 2017	83							
	Ross 2019	84					+		+
	Roth 2011	85					+/-		
	Ryerson 2015	86			+				
	Salaffi 2016	87			+				
	Showalter 2018	88							
	Steen 1997	89			+		+		+
	Suliman 2015	90							
	Tashkin 2007	91						+	
	Tashkin 2016	92			+/-				
	Tashkin 2017	93			+/-		-		
	Theodore 2012	94			-				
	Volkmann 2015	95							
	Volkmann 2017	96							
	Volkmann 2019	42							
	Volkmann 2019	97			+		+		
	Wallace 2015	98			+		+		
	Total available studies for each property		2	2	30	2	18	15	8
	Total studies available for synthesis		2	2	16	2	15	12	7
	Synthesis Rating		GREEN From Working Group	GREEN From Working Group	GREEN	GREEN	GREEN	GREEN	GREEN
	OMERACT Endorsement		This instrument is endorsed for use in randomized clinical trials and longitudinal observational studies						

* This article references two separate data confirming test-retest reliability from the Scleroderma Lung Study I and the Scleroderma Lung Study II

'+' denotes positive performance, '+/-' denotes equivocal performance, '-' denotes inadequate performance

Low risk of bias

Some cautions, but this will be used as evidence

Don't use as evidence

Table 4
Diffusion Capacity of Carbon Monoxide Summary of Measurement Properties. [56–98]

	Author/year	Ref	Truth	Feasibility	Truth	Discrimination			
			Domain match		Construct validity	Test retest reliability	Longitudinal construct validity	Clinical trial discrimination	Thresholds of meaning
	Working Group Appraisal		+	+					
	Patient Research Partners		+	+					
Clinical Trials	Burt 2011	56						-	
Clinical Trials	Burt 2011	57						-	
Clinical Trials	Distler 2019	58						-	
Clinical Trials	Domiciano 2011	59						-	
Clinical Trials	Fraticelli 2014	60						-	
Clinical Trials	Hoyles 2006	61						-	
Clinical Trials	Khanna 2011	62						-	
Clinical Trials	Khanna 2018	63						-	
Clinical Trials	Pakas 2002	64						-	
Clinical Trials	Sircar 2018	65						-	
Clinical Trials	Sullivan 2018	66						-	
Clinical Trials	Tashkin 2006	30						-	
Clinical Trials	Tashkin 2016	31						-	
Observational Studies	Assassi 2010	67			-				
Observational Studies	Baron 2008	68			-				
Observational Studies	Buch 2007	69			-				
Observational Studies	Goh 2008	70			+				
Observational Studies	Goh 2017	71			-		+		+
Observational Studies	Goldin 2008	72			+				
Observational Studies	Goldin 2009	73			-		-		
Observational Studies	Goldin 2018	74					+/-		
Observational Studies	Guler 2018	75					+/-		
Observational Studies	Hax 2017	76			-				
Observational Studies	Hoffmann-Vold 2015	77			+				
Observational Studies	Kafaja 2018	50			-				
Observational Studies	Khanna 2005	78			+/-				
Observational Studies	Khanna 2009	79			-				
Observational Studies	Khanna 2015	80					-		
Observational Studies	Mango 2018	10			-				
Observational Studies	Moore 2013	81			-		-		+/-
Observational Studies	Moore 2015	82			+		+		+
Observational Studies	Morisset 2017	83			+				
Observational Studies	Ross 2019	84					+		+
Observational Studies	Roth 2011	85					-		
Observational Studies	Ryerson 2015	86			+/-				-
Observational Studies	Salaffi 2016	87			+				
Observational Studies	Showalter 2018	88			+/-				
Observational Studies	Steen 1997	89			+				-
Observational Studies	Suliman 2015	90			-				
Observational Studies	Tashkin 2007	91						-	
Observational Studies	Tashkin 2016	92			+				
Observational Studies	Tashkin 2017	93			+/-				
Observational Studies	Theodore 2012	94			-				
Observational Studies	Volkmann 2015	95			-				
Observational Studies	Volkmann 2017	96			-				
Observational Studies	Volkmann 2019	42					-		
Observational Studies	Volkmann 2019	97			+		+		
Observational Studies	Wallace 2015	98			+		+		
	Total available studies for each property		2	2	28	0	12	14	6
	Total studies available for synthesis		2	2	18	0	10	9	4
	Synthesis Rating		GREEN From Working Group	GREEN From Working Group	GREEN	WHITE	GREEN	RED	AMBER
	OMERACT Endorsement		This instrument is not endorsed for use in randomized clinical trials and longitudinal observational studies						

'+' denotes positive performance, '+/-' denotes equivocal performance, '-' denotes inadequate performance

Low risk of bias

Some cautions, but this will be used as evidence

Don't use as evidence

Table 5
Working Group Vote Showing Unanimous Endorsement of the FVC for Use in Randomized Clinical Trials and Longitudinal Observational Studies (>70%).

Vote Participants		
Working Group Members	18	
Working Group Voters	15/18 (83%)	
FVC Endorsement Vote	*Votes Casted*	*Consensus*
Endorse	15 (100%)	Green (Yes)
Provisionally Endorse	0	-
Do not Endorse	0	-

Table 6
Working Group Vote Showing Failure to Endorse DLco Use in Randomized Clinical Trials and Longitudinal Observational Studies (<70%).

Vote Participants		
Working Group Members	18	
Working Group Voters	15/18 (83%)	
DLco Endorsement Vote	*Votes Casted*	*Consensus*
Endorse	3 (20%)	-
Provisionally Endorse	2 (13%)	-
Do not Endorse	10 (67%)	-

gauge the consistency of findings in those studies to yield a final rating for each measurement property. An instrument received a final rating of "Green" if there were at least two pieces of evidence for each measurement property, with good methods, each with at least adequate performance, consistent across studies. If those criteria were not met, the final rating could be "White" (no evidence to support the measurement property), "Red" (insufficient quantity of good quality studies or inadequate performance), or "Amber" (situations where red, green, or white are not appropriate). Provisional endorsement would be assigned if there was a mix of "Green" and "Amber" for the measurement properties. For measurement properties with an "Amber" rating, a research agenda would be discussed to obtain data needed to achieve full endorsement in the future. No endorsement was to be given if any of the measurement properties were given a final rating of "Red" or "White".

2.3.1. Working group vote: truth, discrimination, and endorsement

Several months after the initial vote on feasibility and truth, the Working Group was asked to vote on endorsement based on the SOMP tables. Members were provided with SOMP tables for the FVC and DLco, summarizing the results of the SLR, and providing evidence-based final ratings for each of the 5 key measurement properties (see results section). The vote was preceded by two 1-hour teleconference sessions where the OMERACT Filter 2.1 method and results were presented. Members were encouraged to critically appraise the methodology with regards to the Good Methods and Performance assessments. The Working Group was then asked to complete an on-line vote on the following questions: (1) if the Working Group agrees with the final ratings for each of the 5 key measurement properties and (2) if the Working Group endorses the instrument for use in clinical trials. After >80% of Working Group members voted, a narrative summary of the methods, results, and discussion were distributed electronically to Working Group members and the Technical Advisory Group for critical feedback.

3. Results

3.1. Patient research partner assessments

The majority of patients found the FVC and DLco to be acceptable in terms of procedural comfort (33/40 and 31/40), to require minimal time to come to the pulmonary testing lab (35/40 and 36/40), and to require 10 min or less to perform the studies (32/40 and 23/40). The majority understood their test results (32/40 and 31/40) and testing resulted in >70% reporting no anxiety associated with their results (30/40 and 29/40).

3.2. Working group vote on feasibility and truth

Fifteen out of 18 members completed the on-line voting. All members agreed that FVC is feasible and has face validity. Thirteen of 15 voting members agreed that FVC has content validity, sufficient to attain agreement (≥70%). Those dissenting cited concern that the measurement does not fully represent lung physiology. Both dissenting opinions regarding the FVC's content validity were based on a concern that it does not represent all facets of the construct in question (disease impact on lung physiology). One vote did not change with subsequent discussion; the second dissenting vote changed to affirmation after reviewing comments with other Working Group members during a subsequent teleconference; the change in vote followed a discussion acknowledging that FVC captures the important elements of the construct, if not all facets, with a dearth of alternative instruments at this time.

Fourteen of 15 voting members agreed that the DLco has face validity and feasibility. Concerns regarding the DLco included inconsistent standardization between institutions and potential burden for patient participation. Twelve of 15 voting members agreed the DLco had content validity; concerns centered around its inability to distinguish parenchymal lung damage due to ILD versus pulmonary vascular disease due to pulmonary hypertension. No votes changed after Working Group discussions.

3.3. Systematic literature review

The systematic literature review returned 1,447 articles (after removal of duplicates), with an *a priori* review protocol (unpublished), explicit, transparent, peer-reviewed search strategy, with the use of a standardized data extraction form. Fig. 1 shows the PRISMA flow diagram. Of the 177 articles that met eligibility for a full text review, 48 studies (13 RCTs, 35 LOSs) were included in the descriptive analysis.

3.4. Good methods check and performance across the measurement properties

Articles included in this phase could assess more than one measurement property per article (e.g., construct validity as well as longitudinal construct validity). For FVC, most articles (30 of the 48) assessed construct validity; 1 assessed test-retest reliability, 18 assessed longitudinal construct validity, 15 assessed discrimination in clinical trials, and 8 assessed thresholds of meaning. After applying Good Methods Assessment (articles determined to have a Green or Amber rating), there remained 16 for construct validity, 1 for test-retest reliability, 15 for longitudinal construct validity, 12 for discrimination in clinical trials, and 7 for thresholds of meaning. Finally, Performance (judgement of the adequacy of the results of these articles) revealed a majority achieved a '+' or '+/-' for construct validity (13 of 16 articles), 2 for test-retest reliability (this one publication focused on two separate clinical trials (SLS-I, SLS-II)), 13 of 15 for longitudinal construct validity, 11 of 12 for clinical trial discrimination, and 7 of 7 for thresholds of meaning (with improvement ranging from 3.0% to 5.3% and worsening from -3.0% to -3.3%) [50]. For DLco, 28 of the 48 articles assessed construct validity; no articles assessed test-retest reliability, 12 assessed longitudinal construct validity, 14 assessed clinical trial discrimination, and 6 assessed thresholds of meaning. After applying Good Methods Assessment, there remained 18 for

construct validity, 10 for longitudinal construct validity, 9 for discrimination in clinical trials, and 4 for thresholds of meaning. A '+' or '+/-' was achieved in 14 of the 18 articles for construct validity, 7 of 10 for longitudinal construct validity, 0 of 9 for discrimination in clinical trials, and 4 of 4 for thresholds of meaning.

3.5. Summary of measurement properties tables and final ratings

Tables 3 (FVC) and 4 (DLco) represent the sum of the literature review articles achieving a "Green" (low risk of bias) or "Amber" (some cautions, but will be used as evidence) rating on Good Methods Assessment and their associated performance. These SOMP tables show that FVC achieved a Final Rating of "Green" for each of the five measurement properties. This indicates that after review of the risk of bias and performance of each study, the five key measurement properties demonstrated at least two pieces of good quality evidence, with consistent findings of at least adequate performance to support a Final Rating of "Green". In contrast to the FVC, there was an absence of data assessing the test-retest reliability of DLco (final rating is "White") and inadequate performance of DLco in clinical trial discrimination ("Red").

3.6. Final ratings vote

Final Ratings were submitted to the Working Group as part of an online questionnaire. Participants were asked to vote if they agreed with the Final Ratings. Of the 18 members, 15 voted (83%); the Working Group vote showed agreement (>70%) with the final rating regarding each measurement property, for both instruments.

3.7. Endorsement

Working group members voted to endorse, provisionally endorse, or not to endorse each instrument separately. The FVC achieved unanimous endorsement (Table 5); the DLco did not achieve endorsement (Table 6).

4. Discussion

Decades of research support the face, content, and construct validity of FVC and DLco as instruments measuring lung function in SSc-ILD. Key properties for these instruments, like truth, feasibility, and discriminatory capacity have been recognized as far back as 2003 [51]. This study formally evaluated these instruments using the OMERACT Filter 2.1 to determine if they demonstrate the requisite evidence to support their use in SSc-ILD clinical trials and longitudinal observational studies.

Input from Patient Research Partners, alongside an international panel of experts in lung physiology and SSc-ILD, and backed by peer-reviewed evidence, this process identifies the FVC as the first instrument in the domain of lung physiology to be endorsed by OMERACT for use in RCTs and LOSs. At this time, published data on test-retest reliability of DLco are lacking and there are no randomized clinical trials that demonstrate DLco's discriminatory capacity.

The American Thoracic Society and the European Respiratory Society (ATS/ERS) have established technical standards for the acceptability, repeatability, and quality control of both measurements [52,53]. These standards work to maximize accuracy by ensuring equipment integrity, competency in operator training, criteria for the quality of the test results, and a grading system for measurement repeatability to reduce the within and between maneuver variability. The methods section of each article included in the Filter 2.1 process (both clinical trials and longitudinal observational studies) describes spirometry and gas exchange, but only a minority explicitly stated adherence to ATS/ERS standards.

Extrapulmonary SSc-specific considerations contextualize these results. Patients with SSc may have significant skin involvement of the face causing lip recession, a small oral aperture, and advanced sicca symptoms; these factors can cause the instrument's mouthpiece not to fit appropriately and impair the tight seal required to ensure an accurate measurement. The FVC maneuver is comprised of four distinct phases (maximal inspiration, a "blast" of expiration, continued complete expiration, and inspiration back to maximum lung volume); any one of those phases may be compromised by scleroderma-related chest wall tightness in those with severe cutaneous disease. In addition, muscle weakness may affect expiration despite normal absolute lung volumes and capacities. Finally, pain and other SSc disease features contributing to the burden of extrapulmonary disease may affect body position, predisposing the lungs to be hypocompliant. Compared to the FVC, the DLco's variability is larger and reproducibility is less reliable, despite laboratory quality control procedures [39]. The DLco is a calculated measure: it is the product of measured Kco (the amount of carbon monoxide taken up in a known volume) and measured total alveolar volume; as a result, measurement variation of either or both of these two measurements may impact its accuracy and reproducibility. Importantly, DLco is subject to confounding due to pulmonary vascular disease in SSc and interpretation is challenging in patients with several pathologic factors that impact the DLco (e.g., SSc-ILD, co-occurring pulmonary hypertension, emphysematous changes from tobacco abuse).

These SSc-specific, extrapulmonary concerns have previously been reported [54]. These are difficult to control for in both the research and clinical use of these two measurements; to optimize measurement integrity, care should be taken to adhere to ATS/ERS recommendations to minimize all other sources of variability [55].

In sum, the OMERACT Filter 2.1 provides an evidence-based instrument selection method that avoids selective outcome reporting bias and standardizes measurement in rheumatologic disease, allowing for accurate comparisons of interventions[43]. This study has identified the FVC, in the domain of lung physiology, to be endorsed for use consistently in the setting of SSc-ILD clinical trials and longitudinal observational studies.

Author credit statement

David Roofeh: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing- Original Draft, Writing- Review & Editing, Visualization, Project Administration, Funding Acquisition;

Shaney L. Barratt: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing- Original Draft, Writing- Review & Editing, Visualization, Supervision, Project Administration;

Athol U Wells: Conceptualization, Methodology, Data Curation, Writing- Review & Editing, Supervision;

Leticia Kawano-Dourado: Conceptualization, Methodology, Data Curation, Writing- Review & Editing;

Donald Tashkin: Conceptualization, Methodology, Data Curation, Writing- Review & Editing, Supervision;

Vibeke Strand: Conceptualization, Methodology, Data Curation, Writing- Review & Editing, Supervision;

James Seibold: Conceptualization, Methodology, Data Curation, Writing- Review & Editing, Supervision;

Susanna Proudman: Conceptualization, Methodology, Data Curation, Writing- Review & Editing, Supervision;

Kevin K Brown: Conceptualization, Methodology, Data Curation, Writing- Review & Editing, Supervision;

Paul Dellaripa: Conceptualization, Methodology, Data Curation, Writing- Review & Editing, Supervision;

Tracy Doyle: Conceptualization, Methodology, Data Curation, Writing- Review & Editing;

Thomas Leonard: Conceptualization, Methodology, Data Curation, Writing- Review & Editing;

Eric Matteson: Conceptualization, Methodology, Data Curation, Writing- Review & Editing, Supervision;

Chet Oddis: Conceptualization, Methodology, Data Curation, Writing- Review & Editing, Supervision;

Josh Solomon: Conceptualization, Methodology, Data Curation, Writing- Review & Editing;

Jeff Sparks: Conceptualization, Methodology, Data Curation, Writing- Review & Editing;

Robert Vassallo: Conceptualization, Methodology, Data Curation, Writing- Review & Editing;

Lara Maxwell: Conceptualization, Methodology, Writing- Review & Editing;

Dorcas Beaton: Conceptualization, Methodology, Writing- Review & Editing;

Robin Christensen: Conceptualization, Methodology, Writing- Review & Editing;

Whitney Townsend: Methodology, Software, Validation, Formal Analysis, Resources, Data Curation;

Dinesh Khanna: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing- Original Draft, Writing- Review & Editing, Visualization, Supervision, Project Administration, Funding Acquisition.

Funding support

Dr. Khanna was supported by the NIH/NIAMS K24AR063120
Dr. Roofeh was funded by the NIH/NIAMS T32 grant (AR007080).
The funding sources had no involvement in the project.

Declaration of Competing Interest

Please see each author's ICJME form.

References

[1] Denton CP, Khanna D. Systemic sclerosis. Lancet 2017;390(10103):1685–99.
[2] Wells AU. Interstitial lung disease in systemic sclerosis. Press Medicale 2014;43(10):e329–43.
[3] Steen VD, Medsger TA. Changes in causes of death in systemic sclerosis, 1972-2002. Ann Rheum Dis 2007;66(7):940–4.
[4] Tyndall AJ, Bannert B, Vonk M, et al. Causes and risk factors for death in systemic sclerosis: a study from the EULAR scleroderma trials and research (EUSTAR) database. Ann Rheum Dis 2010;69(10):1809–15.
[5] Wallace B, Vummidi D, Khanna D. Management of connective tissue diseases associated interstitial lung disease: a review of the published literature. Curr Opin Rheumatol 2016.
[6] Wells AU, Denton CP. Interstitial lung disease in connective tissue disease—mechanisms and management. Nat Rev Rheumatol 2014.
[7] Cottin V, Brown KK. Interstitial lung disease associated with systemic sclerosis (SSc-ILD). Respir Res 2019;20(1):1–10.
[8] Wu W, Jordan S, Becker MO, et al. Prediction of progression of interstitial lung disease in patients with systemic sclerosis: the SPAR model. Ann Rheum Dis 2018;77(9):1326–32.
[9] Ryerson CJ, Vittinghoff E, Ley B, et al. Predicting survival across chronic interstitial lung disease: the ILD-GAP model. Chest 2014;145(4):723–8.
[10] Mango RL, Matteson EL, Crowson CS, Ryu JH, Makol A. Assessing mortality models in systemic sclerosis-related interstitial lung disease. Lung 2018;196(4):409–16.
[11] Nikpour M, Baron M. Mortality in systemic sclerosis: lessons learned from population-based and observational cohort studies. Curr Opin Rheumatol 2014;26(2):131–7.
[12] Mackintosh JA, Stainer A, Barnett JL, Renzoni EA. Systemic sclerosis associated interstitial lung disease: a comprehensive overview. Semin Respir Crit Care Med 2019;40(2):208–26.
[13] Roofeh D, Jaafar S, Vummidi D, Khanna D. Management of systemic sclerosis-associated interstitial lung disease. Curr Opin Rheumatol 2019;31(3):241–9.
[14] Amjadi SS, Roofeh D, Namas R, Khanna D. Management of systemic sclerosis-associated interstitial lung disease in the current era. Int J Rheum Dis 2020;23:137–9.
[15] Volkmann ER, Tashkin DP, Li N, Furst DE, Clements PJ, Elashoff RM. Development of a composite outcome measure for systemic sclerosis related interstitial lung disease. Rheumatol 2015;5(2 PG-).
[16] Khanna D, Seibold J, Goldin J, Tashkin DP, Furst DE, Wells A. Interstitial lung disease points to consider for clinical trials in systemic sclerosis. Rheumatology (Oxford) 2017;56(5):v27–32.
[17] Roofeh D, Distler O, Allanore Y, Denton CP, Khanna D. Treatment of systemic sclerosis–associated interstitial lung disease: lessons from clinical trials. J Scleroderma Relat Disord 2020;5(2S):61–71.
[18] Maxwell LJ, Beaton DE, Shea BJ, et al. Core domain set selection according to OMERACT filter 2.1: the OMERACT methodology. J Rheumatol 2019;46(8):1014–20.
[19] Boers M, Beaton DE, Shea BJ, et al. OMERACT filter 2.1: elaboration of the conceptual framework for outcome measurement in health intervention studies. J Rheumatol 2019;46(8):1021–7.
[20] Khanna D, Mittoo S, Aggarwal R, et al. Connective tissue disease-associated interstitial lung diseases (CTD-ILD) - report from OMERACT CTD-ILD working group. J Rheumatol 2015;42(11):2168–71.
[21] Behr J. Approach to the diagnosis of interstitial lung disease. In: Collard HR, Richeldi L, editors. Interstitial lung disease. Elsevier; 2018. p. 87–96.
[22] Landsberg JW. Pulmonary function testing. Clinical practice manual for pulmonary and critical care medicine. Elsevier; 2018. p. 22–35.
[23] Grippi MA, Tino G. Pulmonary Function Testing. In: Grippi MA, Elias JA, Fishman JA, Kotloff RM, Pack AI, Senior RM SM, ed.*Fishman's Pulmonary Diseases and Disorders*. 5th ed. New York; 2015.
[24] O'Donnell D. Physiology of interstitial lung disease. In: Schwartz M, King T, eds. *Interstitial Lung Disease*.; 1998:51-70.
[25] Scanlon PD. Respiratory function: mechanisms and testing. In: *Goldman-Cecil Medicine*. 25th ed.; 2016:539-545.
[26] Culver BH, Graham BL, Coates AL, et al. Recommendations for a standardized pulmonary function report. an official american thoracic society technical statement. Am J Respir Crit Care Med 2017;196(11):1463–72.
[27] Steen VD, Conte C, Owens GR, Medsger TA. Severe restrictive lung disease in systemic sclerosis. Arthritis Rheum 1994;37(9):1283–9.
[28] Volkmann ER, Tashkin DP, Sim M, et al. Short-term progression of interstitial lung disease in systemic sclerosis predicts long-term survival in two independent clinical trial cohorts. Ann Rheum Dis 2019;78(1):122–30.
[29] Caron M, Hoa S, Hudson M, Schwartzman K, Steele R. Pulmonary function tests as outcomes for systemic sclerosis interstitial lung disease. Eur Respir Rev 2018;27(148):170102.
[30] Tashkin DP, Elashoff R, Clements PJ, et al. Cyclophosphamide versus placebo in scleroderma lung disease. N Engl J Med 2006;354(25):2655–66.
[31] Tashkin DP, Roth MD, Clements PJ, et al. Mycophenolate mofetil versus oral cyclophosphamide in scleroderma-related interstitial lung disease (SLS II): a randomised controlled, double-blind, parallel group trial. Lancet Respir Med 2016;4(9):708–19.
[32] Khanna D, Seibold JR, Wells A, et al. Systemic sclerosis-associated interstitial lung disease: lessons from clinical trials, outcome measures, and future study design. Curr Rheumatol Rev 2010;6(2):138–44.
[33] Khanna D, Brown KK, Clements PJ, et al. Systemic sclerosis-associated interstitial lung disease - Proposed recommendations for future randomised clinical trials. Clin Exp Rheumatol 2010;28:55–62.
[34] Wells AU, Behr J, Silver R. Outcome measures in the lung. Rheumatology 2009;47(SUPPL. 5):48–50.
[35] Pellegrino R, Viegi G, Brusasco V, et al. Interpretative strategies for lung function tests. Eur Respir J 2005;26(5):948–68.
[36] Gold WM, Koth LL. Pulmonary function testing. Murray and Nadel's textbook of respiratory medicine. Elsevier; 2016. p. 407–35.
[37] Tashkin DP, Volkmann ER, Tseng CH, et al. Relationship between quantitative radiographic assessments of interstitial lung disease and physiological and clinical features of systemic sclerosis. Ann Rheum Dis 2016;75(2 PG-374-81):374–81.
[38] Rienmüller RK, Behr J, Kalender WA, et al. Standardized quantitative high resolution ct in lung diseases. J Comput Assist Tomogr 1991;15(5):742–9.
[39] 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(4):720–35.
[40] Stanojevic S, Graham BL, Cooper BG, et al. Official ERS technical standards: global lung Function Initiative reference values for the carbon monoxide transfer factor for Caucasians. Eur Respir J 2017;50(3).
[41] Khanna D, Mittoo S, Aggarwal R, et al. Connective tissue disease-associated interstitial lung diseases (CTD-ILD) - report from OMERACT CTD-ILD working group. J Rheumatol 2015;42(11):2168–71.
[42] Volkmann ER, Tashkin DP, Sim M, et al. Cyclophosphamide for systemic sclerosis-related interstitial lung disease: a comparison of scleroderma lung study I and II. J Rheumatol 2019 jrheum.180441.
[43] Beaton DE, Maxwell LJ, Shea BJ, et al. Instrument selection using the OMERACT filter 2.1: the OMERACT methodology. J Rheumatol 2019;46(8):1028–35.
[44] Swinscow T. Ch 11: Correlation and regression. In: Swinscow T, editor. Statistics at square one. 9th ed. BMJ Publishing Group Ltd and European League Against Rheumatism; 1997 https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression.
[45] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155–63.
[46] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–8.
[47] Liljequist D, Elfving B, Roaldsen KS.*Intraclass Correlation – A Discussion and Demonstration of Basic Features*. Vol 14.; 2019.
[48] Luis Ronir Raggio, Almeida RMVR. On the measurement of change in medical research. Int J Stat Med Res 2012;1:144–7.

[49] Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61(2):102–9.
[50] Kafaja S, Clements PJ, Wilhalme H, et al. Reliability and minimal clinically important differences of FVC results from the scleroderma lung studies (SLS-I and SLS-II). Am J Respir Crit Care Med 2018;197(5):644–52.
[51] Merkel PA, Clements PJ, Reveille JD, Suarez-Almazor ME, Valentini G, Furst DE. Current status of outcome measure development for clinical trials in systemic sclerosis Report from OMERACT 6 J Rheumatol 2003;30(7):1630–47.
[52] Graham BL, Steenbruggen I, Barjaktarevic IZ, et al. Standardization of spirometry 2019 update an official American thoracic society and European respiratory society technical statement. Am J Respir Crit Care Med 2019;200(8):E70–88.
[53] Graham BL, Brusasco V, Burgos F, et al. Executive summary: 2017 ERS/ATS standards for single-breath carbon monoxide uptake in the lung. Eur Respir J 2017;49 (1).
[54] Sumphao-Ngern P, Foocharoen C, Boonsawat W, et al. Causes and prevalence of inadequate pulmonary function testing among patients with systemic sclerosis. Arch Med Sci 2015;11(6):1255–60.
[55] Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. Eur Respir J 2005;26(2):319–38.
[56] Burt RK, Shah SJ, Dill K, et al. Autologous non-myeloablative haemopoietic stem-cell transplantation compared with pulse cyclophosphamide once per month for systemic sclerosis (ASSIST): an open-label, randomised phase 2 trial. Lancet 2011;378(9790):498–506.
[57] Daoussis D, Melissaropoulos K, Sakellaropoulos G, et al. A multicenter, open-label, comparative study of B-cell depletion therapy with rituximab for systemic sclerosis-associated interstitial lung disease. Semin Arthritis Rheum 2017.
[58] Distler O, Highland KB, Gahlemann M, et al. Nintedanib for systemic sclerosis-associated interstitial lung disease. N Engl J Med 2019;380(26):2518–28.
[59] Domiciano DS, Bonfá E, Borges CTL, et al. A long-term prospective randomized controlled study of non-specific interstitial pneumonia (NSIP) treatment in scleroderma. Clin Rheumatol 2011;30(2):223–9.
[60] Fraticelli P, Gabrielli B, Pomponio G, et al. Low-dose oral imatinib in the treatment of systemic sclerosis interstitial lung disease unresponsive to cyclophosphamide: a phase II pilot study. Arthritis Res Ther 2014;16(4):1–10.
[61] 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(12):3962–70.
[62] Khanna D, Saggar R, Mayes MD, et al. A one-year, phase I/IIa, open-label pilot trial of imatinib mesylate in the treatment of systemic sclerosis-associated active interstitial lung disease. Arthritis Rheum 2011;63(11):3540–6.
[63] Khanna D, Denton CP, Lin CJF, et al. Safety and efficacy of subcutaneous tocilizumab in systemic sclerosis: results from the open-label period of a phase II randomised controlled trial (faSScinate). Ann Rheum Dis 2018;77(2 PG-212-220):212–20.
[64] Pakas I, Ioannidis JPA, Malagari K, Skopouli FN, Moutsopoulos HM, Vlachoyiannopoulos PG. Cyclophosphamide with low or high dose prednisolone for systemic sclerosis lung disease. J Rheumatol 2002;29(2):298–304.
[65] Sircar G, Goswami RP, Sircar D, Ghosh A, Ghosh P. Intravenous cyclophosphamide vs rituximab for the treatment of early diffuse scleroderma lung disease: Open label, randomized, controlled trial. Rheumatol (United Kingdom) 2018;57 (12):2106–13.
[66] Sullivan KM, Goldmuntz EA, Keyes-Elstein L, et al. Myeloablative autologous stem-cell transplantation for severe scleroderma. N Engl J Med 2018;378(1):35–47.
[67] Assassi S, Sharif R, Lasky RE, et al. Predictors of interstitial lung disease in early systemic sclerosis: A prospective longitudinal study of the GENISOS cohort. Arthritis Res Ther 2010.
[68] Baron M, Sutton E, Hudson M, et al. The relationship of dyspnoea to function and quality of life in systemic sclerosis. Ann Rheum Dis 2008;67(5):644–50.
[69] Buch MH, Denton CP, Furst DE, et al. Submaximal exercise testing in the assessment of interstitial lung disease secondary to systemic sclerosis: reproducibility and correlations of the 6-min walk test. Ann Rheum Dis 2007;66(2 PG-169-73):169–73.
[70] 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(11 PG-1248-54):1248–54.
[71] Goh NS, Hoyles RK, Denton CP, et al. Short-term pulmonary function trends are predictive of mortality in interstitial lung disease associated with systemic sclerosis. Arthritis Rheumatol 2017;69(8 PG-1670-1678):1670–8.
[72] Goldin JG, Lynch DA, Strollo DC, et al. High-resolution CT scan findings in patients with symptomatic scleroderma-related interstitial lung disease. Chest 2008;134 (2):358–67.
[73] 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(5):1333–40.
[74] Goldin J, Keyes-Elstein L, Crofford L, et al. Changes in quantitative scleroderma lung CT measures in patients treated with cyclophosphamide or transplantation. Arthritis Rheumatol 2018;70(Suppl 10) https://acrabstracts.org/abstract/changes-in-quantitative-scleroderma-lung- ct-measures-in-patients-treated-with-cyclophosphamide-or-transplantation/.
[75] Guler SA, Winstone TA, Murphy D, et al. Does systemic sclerosis-associated interstitial lung disease burn out? :specific phenotypes of disease progression. Ann Am Thorac Soc 2018;15(12):1427–33.
[76] Hax V, Bredemeier M, Didonet Moro AL, et al. Clinical algorithms for the diagnosis and prognosis of interstitial lung disease in systemic sclerosis. Semin Arthritis Rheum 2017;47(2):228–34.
[77] Hoffmann-Vold AM, Aaløkken TM, Lund MB, et al. Predictive value of serial high-resolution computed tomography analyses and concurrent lung function tests in systemic sclerosis. Arthritis Rheumatol 2015;67(8):2205–12.
[78] Khanna D, Furst DE, Clements PJ, et al. Responsiveness of the SF-36 and the health assessment questionnaire disability index in a systemic sclerosis clinical trial. J Rheumatol 2005;32(5):832–40.
[79] 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 2009;48(12):1537–40.
[80] Khanna D, Nagaraja V, hong Tseng C, et al. Predictors of lung function decline in scleroderma-related interstitial lung disease based on high-resolution computed tomography: Implications for cohort enrichment in systemic sclerosis-associated interstitial lung disease trials. Arthritis Res Ther 2015;17(1):1–10.
[81] Moore OA, Goh N, Corte T, et al. Extent of disease on high-resolution computed tomography lung is a predictor of decline and mortality in systemic sclerosis-related interstitial lung disease. Rheumatol 2013;52(1 PG-155-60):155–60.
[82] Moore O, Proudman S, Goh N, et al. Quantifying change in pulmonary function as a prognostic marker in systemic sclerosis-related interstitial lung disease. Clin Exp Rheumatol 2015;33:111–6.
[83] Morisset J, Vittinghoff E, Elicker BM, et al. Mortality risk prediction in scleroderma-related interstitial lung disease: the SADL model. Chest 2017;152(5):999–1007.
[84] Ross L, Stevens W, Wilson M, et al. Can patient-reported symptoms be used to measure disease activity in systemic sclerosis? Arthritis Care Res (Hoboken) 2019:3.
[85] Roth MD, Tseng C-H, Clements PJ, et al. Predicting treatment outcomes and responder subsets in scleroderma-related interstitial lung disease. Arthritis Rheumatol 2011;23(1):2797–808.
[86] Ryerson CJ, O'Connor D, Dunne JV, et al. Predicting mortality in systemic sclerosis-associated interstitial lung disease using risk prediction models derived from idiopathic pulmonary fibrosis. Chest 2015;148(5):1268–75.
[87] Salaffi F, Carotti M, Di Donato E, Di Carlo M, Ceccarelli L, Giuseppetti G. Computer-aided tomographic analysis of interstitial lung disease (ILD) in patients with systemic sclerosis (SSc). Correlation with pulmonary physiologic tests and patient-centred measures of perceived dyspnea and functional disability. PLoS One 2016;11(3):1–13.
[88] Showalter K, Hoffmann A, Rouleau G, et al. Performance of forced vital capacity and lung diffusion cutpoints for associated radiographic interstitial lung disease in systemic sclerosis. J Rheumatol 2018;45(11):1572–6.
[89] Steen VD, Medsger Jr TA. The value of the health assessment questionnaire and special patient-generated scales to demonstrate change in systemic sclerosis patients over time. Arthritis Rheum 1997;40(11 PG-1984-91):1984–91.
[90] Suliman YA, Dobrota R, Huscher D, et al. Pulmonary function tests: high rate of false-negative results in the early detection and screening of scleroderma-related interstitial lung disease. Arthritis Rheumatol 2015;67(12):3256–61.
[91] 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(10):1026–34.
[92] Tashkin DP, Volkmann ER, Tseng CH, et al. Relationship between quantitative radiographic assessments of interstitial lung disease and physiological and clinical features of systemic sclerosis. Ann Rheum Dis 2016;75(2):374–81.
[93] Tashkin DP, Volkmann ER, Tseng CH, et al. Improved cough and cough-specific quality of life in patients treated for scleroderma-related interstitial lung disease: results of scleroderma lung study II. Chest 2017;151(4 PG-813-820):813–20.
[94] Theodore AC, Tseng CH, Li N, Elashoff RM, Tashkin DP. Correlation of cough with disease activity and treatment with cyclophosphamide in scleroderma interstitial lung disease: findings from the scleroderma lung study. Chest 2012;142(3 PG-614-621):614–21.
[95] Volkmann ER, Tashkin DP, Li N, Furst DE, Clements PJ, Elashoff RM. Development of a composite outcome measure for systemic sclerosis related interstitial lung disease. Rheumatology 2015;5(2):6072–8.
[96] Volkmann ER, Tashkin DP, Li N, et al. Mycophenolate mofetil versus placebo for systemic sclerosis-related interstitial lung disease: an analysis of scleroderma lung studies I and II. Arthritis Rheumatol 2017;69(7 PG-1451-1460):1451–60.
[97] Volkmann ER, Tashkin DP, Sim M, et al. Short-term progression of interstitial lung disease in systemic sclerosis predicts long-term survival in two independent clinical trial cohorts. Ann Rheum Dis 2019;78(1 PG-122-130):122–30.
[98] Wallace B, Kafaja S, Furst DE, et al. Reliability, validity and responsiveness to change of the Saint George'S Respiratory Questionnaire in early diffuse cutaneous systemic sclerosis. Rheumatol (United Kingdom) 2015;54(8):1369–79.