Eur Respir J 2005; 26: 755–758
DOI: 10.1183/09031936.05.00104105

EDITORIAL

Clinical trials in idiopathic pulmonary fibrosis: a word of caution concerning choice of outcome measures

W.C. Johnson* and G. Raghu#

Idiopathic pulmonary fibrosis (IPF) is a challenging and frustrating disease for clinicians and patients confronted with this clinical entity, since it is associated with a dismal ~3-yr median survival from the time of diagnosis [1] and an effective treatment regimen is yet to be determined. The lack of an effective treatment option has encouraged investigation of new treatment regimens, and promising agents are being investigated in several clinical trials worldwide. Despite aggressive, worldwide collaborations in pursuit of an effective regimen, significant progress is still needed to recommend a specific treatment regimen for IPF [2].

While recent publications have provided insights regarding appropriate outcome measures chosen as end points for IPF clinical trials, results from these studies have either (unfortunately) been negative or merely provided positive signals and generated hypotheses that deserve to be investigated further in well-designed studies [3–6]. Investigators who are involved in planning IPF studies face daunting challenges in selecting outcome measures that will appropriately allow statistical assessment of clinically meaningful effects or outcomes. Clinicians reviewing the IPF literature are confronted with equally daunting challenges in interpreting the clinical and statistical importance of these findings and applying this knowledge in the clinical care of their patients. Both the investigator and the clinician interpreting the results need to understand how IPF disease severity and progression are best measured in order to provide the most convincing and interpretable study evidence.

The 2000 International Consensus statement for IPF diagnosis and treatment suggests measures with promise for measuring IPF disease extent and potential for measuring change, but it is clear that validated outcome measures were not known when the document was drafted [1]. Investigators involved in the largest IPF trial conducted to date recently reported an apparent lack of appropriate IPF outcome measures [6]. They reported that change in forced vital capacity (FVC) could potentially be used as an outcome measure, but suggested that mortality as an end point is more sensitive to a treatment effect in their study than any of the physiological markers of disease progression that they observed. Other investigators have documented that changes in FVC and walk-test measures are associated with survival [7–12]. Thus, there is continued interest and optimism in using these physiological measures and, potentially, other outcome measures to assess disease progression and overall survival.

It should be noted that not all outcome measures ought to be considered with equal importance and that the choice of outcome measures should be appropriate for the goals of each study [13–15]. The appropriateness of an outcome measure should be judged according to its clinical relevance and the likelihood that it can be used to detect statistically significant effects. When considering clinical relevance, it is important to specifically consider what magnitude would be considered clinically meaningful, because this is where statistical aspects of study planning and conduct are directly linked to the clinical aspects. Outcome measures that are able to demonstrate clear benefit to patient well-being should be given primary importance in later definitive trials (*i.e.* phase III trials). Demonstration of a survival benefit would be clinically relevant and indisputably beneficial to patient well-being, but achieving such a result would typically require well-designed, long-term studies with a large number of well-defined IPF patients. The feasibility of patient recruitment, retention and protocol adherence represents a major practical challenge. Therefore, a more feasible approach with corresponding outcome measures is necessary in early studies that are intended to guide decisions concerning design of the later critical definitive trials. Outcome measures that are believed to be in the pathway of disease progression may be referred to as surrogate measures (*i.e.* surrogate end points, surrogate markers or biomarkers). Frequently, the number of participants in nondefinitive studies (*e.g.* phase I, proof of concept or phase II trials) would be fewer than what is needed for a definitive trial due to the need for efficiency and affordability. However, it is imperative that each study is planned to ensure adequate sample size and power to meet the study goals. In a highly fatal disease such as IPF, definitive trials might be expected to make use of a nonsurrogate end point such as survival, whereas an earlier study might make use of surrogate end points such as change at 1 yr in FVC, changes in high-resolution computer tomography (HRCT) findings, or perhaps change at 6 months in measures of the 6-min walk test. It is imperative that investigators take steps to ensure that the outcome measures chosen are rigorously validated and clinically meaningful before embarking on a study. Failure to do so can detract from the study results (even positive ones) and thereby challenge whether limited IPF resources are being

*Collaborative Health Studies Coordinating Center, and #Division of Pulmonary and Critical Care Medicine, University of Washington, Seattle, WA, USA.

CORRESPONDENCE: G. Raghu, Division of Pulmonary and Critical Care Medicine, University of Washington, Campus Box 356522, Seattle, WA, USA 98195. Fax: 1 2065982105. E-mail: graghu@u.washington.edu

efficiently utilised. Investigators considering a surrogate outcome should carefully consider whether the effects of any intervention being studied lie in the causal pathway and whether there might be effects which lie outside the casual pathway [13–15]. Accuracy of interpretation of any observed effect of an intervention will depend on these factors, particularly in IPF research where mechanism of action may not be well understood.

Hierarchy and *a priori* definition of outcome measures should be considered carefully when evaluating reported study results. Failure to do so puts the researcher and the clinician at risk of misinterpreting the importance of spurious results, which occur naturally when many comparisons are being made (*i.e.* multiple comparisons problem). Typically, a type one error rate of $\alpha=0.05$ is used for judging statistical significance of medical study results. This suggests that one might expect, on average, one out of 20 results reported significant at the 0.05 level to be in error. When results of multiple outcome measures and in multiple subgroups are reported, the number of tests being reported can become large and the opportunity for spurious results increases. In studies of rare diseases such as IPF, this point is particularly important because studies are frequently underpowered (*i.e.* a sufficient number of eligible participants or resources for expanding recruitment may not be available). To avoid misinterpretation, it is important that a primary outcome measure is identified prior to study initiation and that the study results are primarily based on this outcome. In addition to the primary outcome measure, other outcome measures may be identified *a priori* and would be considered secondary outcome measures. Results from the secondary measures should be interpreted insofar as they are consistent and supportive of the primary outcome measure, and provide insight into possible mechanisms of action and a more complete picture of the research questions. While researchers should carefully define primary and secondary end points for their studies, exploratory or *post hoc* subgroup analyses are often reported. These exploratory analyses should be interpreted with caution because, scientifically, they can only be expected to be hypothesis generating and would need confirmation in future studies. Exploratory end points are useful in describing an interesting detail or for planning future research, but should not be considered with equal importance as results of the primary and secondary end points.

CATEGORIES OF OUTCOME MEASURES FOR IPF

Mortality

A mortality end point is frequently described as a survival outcome measure and may be more specifically defined as time until death. At this point in IPF clinical trials, one could argue for the choice of a robust end point. Trials which are planned to make a definitive statement concerning true efficacy must involve survival in the primary hypothesis and as the primary outcome measure. Due to the high mortality rate associated with this disease, it is difficult to imagine a situation where a therapy would be considered conclusively efficacious without demonstrating superiority in survival. In addition to survival being directly meaningful to the IPF patient, analysis methods for survival data take into account death and have inherent capabilities for dealing with loss to follow-up missing data issues.

Some investigators may choose an outcome measure that employs survival analysis methodology, but which does not specifically measure mortality. The investigators of one large IPF study chose progression-free survival as their primary outcome measure [4, 6]. Progression-free survival was defined as time until a 5-mmHg increase in the alveolar–arterial pressure difference in oxygen ($P_{A\text{-}a,O_2}$) or a 10% decrease in per cent predicted FVC or until death, in an attempt to maximise the number of observed end points over that of strict mortality. While mortality was taken into account in this end point, this would not be considered a mortality end point and the results would have to be interpreted accordingly.

Direct measures of morbidity

It could easily be argued that a difference in rates of measures of morbidity are relevant to patient well-being and, therefore, may potentially be an appropriate outcome measure even for definitive trials.

Acute exacerbation of IPF

While the possibility of acute exacerbation (AE) of IPF as an objective measurement of IPF exacerbation events has been recently reported by AZUMA *et al.* [3], widely accepted diagnosis criteria for AEs of IPF have not yet been adopted. The importance of this outcome measure would be dependent upon the notion that IPF decline is associated with episodes of acute respiratory illness. In a recent retrospective analysis of a large clinical trial, investigators reported that an apparent rapid respiratory decline preceded death [16]. Thus, demonstration of a reduction in frequency or severity of exacerbations would also be an important finding in its own right. Standardisation and validation of diagnostic criteria for AE of IPF are among the more important outcome measures development needs and a standardised definition could play an important role in tracking IPF patient health in the clinical setting.

Hospitalisation

Number of hospitalisations or number of hospital days during a period of time could prove valuable as an outcome measure, but would also benefit from formal validation for use in IPF [16]. In particular, respiratory hospitalisations may be promising as they could provide similar information to the previously mentioned IPF exacerbation. Further assessment of the relative importance of IPF related hospitalisation is needed.

Need for supplemental oxygen requirement

Need for supplemental oxygen requirement should be based on physiological needs and documentation rather than the patient's perception of using/need for supplemental oxygen. Demonstration of decreased need for supplemental oxygen would be clinically significant and beneficial.

Adverse events

Rates of important adverse events should not be relegated to consideration only as safety issues in this population. Events that are particularly serious in nature could also serve as efficacy outcome measures. Pulmonary embolus, pneumonia and respiratory failure events are observed in IPF patients and could prove to be important efficacy outcome measures as well. One might expect the potential for such outcome

measures to become evident during typical evaluation of safety measures during early clinical studies.

Measurements of dyspnoea

University of California, San Diego (UCSD) Shortness of Breath Questionnaire, transition dyspnoea index, and other dyspnoea scores appear to be of limited use and, at this time, might best be considered as providing supporting information to more promising end points [16].

Physiological measures

Pulmonary function testing methodologies for evaluating lung volumes, capacities and gas transfer are well defined. Strict adherance to standardised proceedures in accordance to current American Thoracic Society and European Respiratory Society task force guidelines are necessary. The importance of these measures is well understood for IPF and other lung diseases. The shortcomings of these measures are that they are typically performed in an ''at rest'' situation and FVC measurements are effort dependent. In addition, the severity based on measures of resting pulmonary function tests (PFTs), such as FVC, forced expiratory volume in one second (FEV_1), total lung capacity (TLC) and diffusing capacity of the lung ($D_{L,CO}$) measurements may not necessarily reflect the true severity and extent of the disease process in IPF. Regardless, these are standardised procedures, validated and have been well utilised in assessing the functional status of a subject's breathing potential based on resting PFTs as follows. 1) FVC: change in FVC has been shown to be a promising outcome measure appropriate for early studies and as a secondary outcome measure in definitive trials [6, 9]. 2) $D_{L,CO}$ (corrected for haemoglobin): change in single-breath $D_{L,CO}$ may be an important outcome measure appropriate for early studies and as a secondary outcome measure in definitive trials in IPF patients, but results with this measure have been somewhat controversial and may be dependent on initial disease severity [6, 17]. 3) Other static measures of lung function such as FEV_1, TLC, and $P_{A\text{-}a,O_2}$ at rest have not proven to be reliable IPF measures of change. 4) Measures of oxygenation such as arterial oxygen saturation measured by pulse oximetry (S_{P,O_2}), arterial blood gases and oxygen use may be relevant clinical measures of disease severity, but have not yet proven useful as an outcome measure, particularly when change over time is of interest. 5) Exercise testing: exercise-induced hypoxia has been shown to be associated with IPF disease severity. However, the exercise test appears not to be the best choice for measuring exercise-induced hypoxia because it was shown to have relatively poor reproducibility [10]. The 6-min walk test may be appropriately sensitive as an outcome measure in some IPF patients, and the walk distance has been shown to be reproducible [10, 11]. An important limitation of the 6-min test, however, is that patients with severe IPF may not be able to complete the requisite 6 min of walking. A modification of the 6-min walk test, known as the timed walk test, attempts to provide methodology for evaluating exercised-induced hypoxia in all IPF patients, regardless of whether they can complete 6 min of walking [12]. Like the 6-min walk test, it is not known how oxygen supplementation and participant effort may affect outcome measures associated with the procedure (distance, velocity, S_{P,O_2}).

Radiographic findings with HRCT

Radiographic findings with HRCT images of the pulmonary parenchyma have been developed for diagnostic purposes. Development and validation of a radiographic measure of IPF disease severity and change could prove important to IPF research [18]. Since HRCT findings are dependent on an expert chest radiologist's interpretation, scoring and scales are somewhat subjective, require replicate scoring and an evaluation of reproducibility.

Quality of life

IPF quality of life (QOL) measures have been developed and adequately measure important attributes of patient life. Given the serious need for a treatment regimen which impacts the excessive mortality rate, it is hard to imagine how QOL will provide more than supporting evidence in studies. None of the studies to date have demonstrated differences in the QOL measures.

Composite end points

Composite end points take two or more of the previously noted outcome measures and combine them into a single outcome measure. Composite end points are typically considered to be a special class of surrogate end points in the literature [15]. In IPF research, investigators have considered a composite end point for evaluating disease progression defined as an FVC decrease of 10% or $D_{L,CO}$ relative decrease of 15%. While the use of composite end points can be tempting, a composite end point may suffer from the same difficulties that a single component measure suffers from and are rarely well validated. A composite end point may be appropriately considered for use as a secondary end point [19].

ANALYTICAL ISSUES IN IPF STUDIES

The selection of analytic techniques needs careful consideration prior to initiation of the study to ensure they are appropriate for the outcome measure and to ensure that missing data are adequately accounted for in the analyses. It is very common to have substantial missing data in IPF research due to participant dropout, lung transplant and death. Dropouts may be related to health deterioration and, therefore, may not be missing at random, which can pose special difficulties for analytical situations when a mortality end point is not used. For valid results in situations where there are missing data, investigators must seek guidance from a statistician with experience in missing data situations because appropriate analyses methods in these situations can be complex and simple methodologies, such as complete case analysis or last observation carried forward analysis, may not be appropriate [20].

In situations where a continuous outcome measure is available (*e.g.* FVC), statistical precision and power are typically best preserved when the analyses make use of the continuous variable. If one recodes the outcome measure into a categorical variable for their analyses, the power to detect a difference or association will almost always be lost. However, the resulting analysis of the categorical variable may provide a clinically meaningful and interpretable presentation of the data and could be considered as an adjunct to the typically more powerful analyses of the continuous variable. AZUMA *et al.* [3] provide an example of this approach by presenting counts of the number of patients who declined, remained stable, or

improved over the course of the study in order to enhance interpretation of their estimates of change in S_{P,O_2} during exercise.

In conclusion, both investigators and clinicians need good insight into the potential problems associated with clinical trials with IPF. It is especially important to avoid over interpretation of results from early studies (*e.g.* retrospective with poorly defined IPF patient populations) or from subgroup and exploratory analyses, as such practice can be misleading. Analyses from *post hoc*/subgroup analyses may generate important hypotheses and, if proven to be biologically plausible, clinically meaningful, testable and feasible, they deserve to be further studied in well-designed studies. Indeed, the ongoing multinational study designed to determine the survival benefit using gamma interferon (INSPIRE trial) stemmed from the hypothesis generated from such a subgroup analyses [4]. In the end, an efficacious IPF treatment regimen will need to be proven through a series of clinical trials that culminate in an adequately powered definitive clinical trial with a statistically significant and clinically meaningful effect being measured by well-defended methodologies.

Thus, much work is still needed in developing new outcome measures and validating existing ones for use in studies. Careful consideration and choice of outcome measures used in idiopathic pulmonary fibrosis studies will help establish effective and achievable drug development programmes and will enable clinicians and investigators to make informed critical decisions in recommending a treatment regimen for the patient suffering from idiopathic pulmonary fibrosis. A critical and careful assessment of analytic methods used in reporting idiopathic pulmonary fibrosis clinical trial results is essential for the investigator and clinician to interpret and appropriately place into context the results of idiopathic pulmonary fibrosis studies.

REFERENCES

1 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: 646–664.

2 Bouros D, Antoniou KM. Current and future therapeutic approaches in idiopathic pulmonary fibrosis. *Eur Respir J* 2005; 25: 1693–1702.

3 Azuma A, Nukiwa T, Tsuboi E, *et al.* Double-blind, placebo-controlled trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2005; 171: 1040–1047.

4 Raghu G, Brown KK, Bradford WZ, *et al.* Idiopathic Pulmonary Fibrosis Study Group. A placebo-controlled trial of interferon gamma-1b in patients with idiopathic pulmonary fibrosis. *N Engl J Med* 2004; 350: 125–133.

5 Du Bois RM. Is idiopathic pulmonary fibrosis now treatable? *Am J Respir Crit Care Med* 2005; 171: 939–940.

6 King TE Jr, Safrin S, Starko KM, *et al.* Analyses of efficacy end points in a controlled trial of interferon-gamma1b for idiopathic pulmonary fibrosis. *Chest* 2005; 127: 171–177.

7 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: 543–548.

8 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: 531–537.

9 Collard HR, King TE Jr, Bartelson BB, Vourlekis JS, Schwarz MI, Brown KK. Changes in clinical and physiologic variables predict survival in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2003; 168: 538–542.

10 Eaton T, Young P, Milne D, Wells AU. Six-minute walk, maximal exercise tests: reproducibility in fibrotic interstitial pneumonia. *Am J Respir Crit Care Med* 2005; 171: 1150–1157.

11 Lama VN, Flaherty KR, Toews GB, *et al.* Prognostic value of desaturation during a 6-minute walk test in idiopathic interstitial pneumonia. *Am J Respir Crit Care Med* 2003; 168: 1084–1090.

12 Hallstrand TS, Boitano LJ, Johnson WC, Spada CA, Hayes JG, Raghu G. The timed walk test as a measure of severity and survival in idiopathic pulmonary fibrosis. *Eur Respir J* 2005; 25: 96–103.

13 Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996; 125: 605–613.

14 Fleming TR. Surrogate endpoints and FDA's accelerated approval process. *Health Aff (Millwood)* 2005; 24: 67–78.

15 Fleming TR. Issues in the design of clinical trials: insights from the trastuzumab (Herceptin) experience. *Semin Oncol* 1999; 26: Suppl. 12, 102–107.

16 Martinez FJ, Safrin S, Weycker D, *et al.* The clinical course of patients with idiopathic pulmonary fibrosis. *Ann Intern Med* 2005; 142: 963–967.

17 Raghu G, Johnson WC, Lockhart D, Mageto Y. Treatment of idiopathic pulmonary fibrosis with a new antifibrotic agent, pirfenidone: results of a prospective, open-label Phase II study. *Am J Respir Crit Care Med* 1999; 159: 1061–1069.

18 Lynch DA, David Godwin J, Safrin S, *et al.* High-resolution computed tomography in idiopathic pulmonary fibrosis: diagnosis and prognosis. *Am J Respir Crit Care Med* 2005; 172: 488–493.

19 Cannon CP. Clinical perspectives on the use of composite endpoints. *Control Clin Trials* 1997; 18: 517–529.

20 Molenberghs G, Thijs H, Jansen I, *et al.* Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 2004; 5: 445–464.