HHS Public Access

Author manuscript

J Cyst Fibros. Author manuscript; available in PMC 2017 March 07.

Published in final edited form as:

J Cyst Fibros. 2017 March ; 16(2): 175–185. doi:10.1016/j.jcf.2016.12.008.

Chest Imaging in CF Studies: What Counts, and Can be Counted?

Rhonda Szczesniak, Ph.D.,
Division of Biostatistics & Epidemiology and Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center; Cincinnati, OH United States

Lidija Turkovic, Ph.D.,
Telethon Kids Institute, West Perth, Australia

Eleni-Rosalina Andrinopoulou, Ph.D., and
Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands

Harm A. W. M. Tiddens, M.D., Ph.D.
Department of Pediatric Pulmonology and Allergology, Department of Radiology Erasmus MC-Sophia Children's Hospital, Rotterdam, The Netherlands

Abstract

Background—The dawn of precision medicine and CFTR modulators require more detailed assessment of lung structure in cystic fibrosis (CF) clinical studies. High-resolution chest computed tomography (CT) scoring has yielded sensitive markers for the study of CF disease progression and therapeutic effectiveness. Similarly, Magnetic Resonance Imaging (MRI) is in development to generate structural as well as functional markers.

Results—The aim of this review is to characterize the role of CT and MRI markers in clinical studies, and to discuss study design, data processing and statistical challenges unique to these endpoints in CF studies. Suggestions to overcome these challenges in CF studies are included.

Conclusions—To maximize the potential of CT and MRI markers in clinical studies and advance treatment of CF disease progression, efforts should be made to develop data repositories, promote standardization and conduct reproducible research.

Keywords

cystic fibrosis; endpoints; imaging analysis; outcome measures; reliability; surrogate endpoints

1. Introduction

High-resolution chest computed tomography (CT) has produced promising outcomes for the clinical study of cystic fibrosis (CF) lung disease progression(1, 2). Markers from CT imaging can convey severity related to mild and regional CF lung disease by quantifying degrees of bronchiectasis, air trapping, and other attributes related to structural lung damage.

Correspondence should be addressed to: Harm A.W.M. Tiddens, M.D., Ph.D., Erasmus MC-Sophia Children's Hospital, Department of Pediatric Pulmonology and Allergology, P.O. Box 2060, 3000CB Rotterdam, The Netherlands. h.tiddens@erasmusmc.nl.

It has been shown that CT markers have higher sensitivity to detect pulmonary disease progression than FEV_1%, an established outcome in CF(3–7). More recently, magnetic resonance imaging (MRI) techniques have emerged to provide radiation-free markers that quantify structural and dynamic aspects of the CF lung(8). Given the challenges related to pulmonary function testing in younger individuals with CF and the need for assessments of early-stage CF lung disease, clinical studies have incorporated CT markers as outcomes for structural lung disease(9). However, there remain obstacles related to study design, data acquisition/processing, and statistical analysis that have not been comprehensively addressed, hindering more complete adaptation of established CT markers as endpoints(10). The aim of this review is to provide information on the utility of imaging markers as endpoints for CF studies and to describe the accompanying statistical considerations. CT serves as the exemplar for the review, given its status as the gold standard for lung structure assessment; however, considerations shared by MRI are also described and accompanied by recommendations.

2. CT Imaging Analysis

2.1 Evolution of scoring systems

The clinical study of imaging markers has evolved from qualitative evaluation to quantitative lung image analysis. Although seminal work involved the study of markings from chest radiographs, this modality has largely been excluded from clinical intervention studies due its poor sensitivity for monitoring CF disease progression (11) the advent of CT scoring systems. Over the last few decades it has been shown that bronchiectasis, airway wall thickening, mucous impaction, and trapped air are the most important markers to quantify on chest CTs. It becomes clearer that the term “trapped air” is probably a misnomer, as hypodense regions on expiratory CT can result both from hypoperfusion and trapped air. For the purpose of this review, we will continue to use the term trapped air, acknowledging it represents a mix of hypoperfusion and trapped air(12).

Semi-quantitative image analysis of CT scans to assess these attributes has yielded a variety of scoring systems that have been used for clinical studies (Table 1). The Brody I system was developed in 1999 using CTs of 8 patients aged 5–16 years. At that time, CF lung disease progressed more rapidly, compared to the present day(13). The Brody II scoring system (14, 15) followed and was frequently used until a decade ago. De Jong and colleagues compared various scoring systems and found that they were reproducible and correlated with pulmonary function data(16); however, these scoring systems were not well-standardized.

To improve standardization and training, the CF-CT scoring system, based on the Brody II system, was developed in 2011. The CF-CT scoring system consists of a large training module and 7 training sets that were scored by Brody and de Jong (the most experienced observers at that time) to define the ‘gold standard’ ratings. To date, over 20 observers have been trained in the Erasmus MC LungAnalysis Core Laboratory using the CF-CT method. It has been used in multiple studies to validate chest CT as an outcome (6, 17–20). An advantage of scoring systems like CF-CT is that the lung volume level during CT acquisition is not very critical for the magnitude of the scores (21) (22).

The CF-CT scoring system still has a number of disadvantages. Firstly, it is insensitive for quantifying early changes (23), only detecting relatively large structural changes over time. Secondly, the clinical value of the numbers generated is difficult to understand. Thirdly, the method is time consuming and observer dependent. Hence, further development of more sensitive methods was required.

For the development of a more sensitive and quantitative method, a morphometric approach was created using a grid projected over the CT image. This approach was used first in a group of 411 patients with end stage lung disease (24, 25) and later to compute volume fractions of trapped air on expiratory scans(26). Next, this method was further developed into the Perth-Rotterdam Annotated Grid Morphometric Analysis for CF (PRAGMA-CF) scoring system to quantify early structural changes(18). This system allows expression of key structural changes, i.e. airway abnormalities and regions of low density, as a fraction of total lung volume. This system can also be applied in more advanced disease(24). A disadvantage of PRAGMA-CF is that it requires two weeks of training and takes around 30 minutes per CT to execute for an experienced observer. Fortunately, it is likely that the system can be automated using a machine learning approach. More recently, the Airway-Artery (AA) method was developed for the sensitive and automated analysis of all visible airway artery pairs. It is likely that this system eventually will take over the scoring of airway abnormalities (Kuo and colleagues, 2016, work in press and other under review).

Image analysis systems are at an early stage of development for MRI-based quantification (27, 28). Failo and colleagues, among others, have reported that CT and MRI modalities produce similar Brody scores(29, 30). A small study of MRI perfusion markers obtained on non-CF adults suggested that scores might be highly dependent upon observer(31). The sensitivity, extent of reproducibility, and repeatability of MRI-based scoring systems needs further study.

2.2 Imaging markers as surrogate outcomes

Quantitative image analysis paved the way for reproducible, reliable CT scoring systems that can be used to produce outcomes for clinical studies. Currently, pulmonary exacerbation, health-related quality of life, pulmonary function and survival are the only recognized clinical endpoints for CF studies(10, 32). As surrogate endpoints, imaging markers do not directly measure how an individual with CF “functions, feels or survives”(33). CT markers, meant to assess structure, are often considered as intermediate endpoints in CF studies due to their ability to predict established clinical endpoints(2, 19, 24). MRI is a more promising modality than CT to assess functional aspects of the lung, such as lung perfusion, pulmonary hemodynamics, central airway dynamics and ventilation of the lung(31, 34).

Given the pathophysiology of CF, it is likely that the imaging marker is measured (perhaps repeatedly) with a particular therapy being applied at some point in the disease process (Figure 1). Other surrogates, such as FEV_1%, are repeatedly collected throughout this process and may impact clinical endpoints independently of the imaging marker. A CT marker, for example, is a useful surrogate, provided it is consistently i) predictive of future events; ii) reflective of a therapeutic response(35). Loeve and colleagues summarized over 20 studies validating CT markers as surrogate endpoints for presence and severity of CF

lung disease, therapeutic responsiveness, reproducibility, and associations with respiratory exacerbations, quality of life, survival and other outcomes(26). As indicated, future validation studies should focus on criterion (ii), in order to further demonstrate surrogacy.

CT scans yield a variety of structural markers, and the choice of which is used for a particular CF study will depend upon several factors, such as the age and severity of the population being studied, therapy being evaluated, and duration of the study. It is important to determine the extent to which a prospective CT marker is predictive of the clinical endpoint and whether the therapeutic response of the CT marker (as a surrogate) is predictive of the therapeutic response detected by the clinical endpoint. CT markers that have scoring systems independently validated prior to the study, have been utilized in previous CF clinical studies and are most closely aligned with the therapeutic aim and causal pathway should be considered as surrogates. Similar factors should be considered when selecting MRI markers as monitoring tools for clinical studies. Timing of study visits could also be considered, as MRI may be utilized to assess short-term changes in therapeutic studies(31, 36).

3. Addressing Sources of Bias Specific to CT Markers

3.1 Protocol, technology and data processing standards

An issue that has received little attention until recently is the impact of standardizing image acquisition techniques. Controlling lung volume is an important issue in the acquisition phase. In children of 6 years of age and above, lung volume can best be controlled for using a spirometer(37). Participants below the age of 6 years may have difficulty following spirometer-based protocols; in these instances, general anesthesia and a pressure-controlled protocol can be used for lung volume control(21). Children aged 3–6 years can be trained to execute a breath hold after taking a deep breath or at the FRC level. Alternatively, CT can be acquired for very young or non-cooperative children while the child is free breathing. For these children, scans acquired at a volume level near functional residual volume are less sensitive, compared to inspiratory scans for the detection of airway disease. Equally important factors to consider include radiation dose, pitch, reconstruction kernels, and slice thickness; all of which are known to influence image quality(17) and can be accounted for in prospective studies.

The selection of the CT protocol is closely linked to the image analysis methods used. When a (semi) automated analysis method is selected, tighter control of the CT protocol is required, but scoring methods are known to be less sensitive to choice of CT protocol and the aforementioned issues with lung volume control(22). To track disease over time, ideally the same volume and CT protocol should be used when follow up CTs are compared to baseline CTs. Recent work highlights the importance of standardization in multicenter studies using CT(17). In addition to previously mentioned longitudinal validation studies, assessment of standardization techniques are also needed for MRI. However, this issue is considered an important technical challenge for MRI.

Age can be a substantial confounder as the chest CT resolution is an important determinant for the smallest structures that can still be observed. This is especially important in the first

two years of life. Minimizing confounding through study design is ideal. It may be possible to use methods based on restriction or matching. Examples include subgroup or covariate adjustment to determine the extent of technology, processing or other effects on analysis results. These stratification and multivariable modeling approaches rely on measured confounders (i.e. the variables are recorded in the database).

3.2 Assessing observer reliability

Despite advancements toward automation, CT scoring systems still require observer evaluation to quantify degree of structural lung disease, and there is no gold standard metric to which observers can be compared. As such, observer agreement and reliability are indistinguishable and often used synonymously in the CF imaging literature. Observer scoring introduces variation known as measurement error (38). It is worth noting that, in contrast, automated scoring might introduce systematic error (e.g. over- or underestimation of airway wall thickness). The following recommendations on observer reliability apply generally to CT and MRI markers.

At minimum, it is recommended that any study with scored CT markers include reliability statistics within observer. Intra-rater reliability refers to the extent to which a single observer can replicate his previous scores on a series of scans. In terms of experimental design, the scans to be repeatedly scored should be selected using random sampling stratified by age and disease severity. This approach should be employed in CF studies, because of the heterogeneous nature of disease progression. For example a study may yield high reliability in the subcohort with severe disease and low reliability in the subcohort with mild disease.

The proportion of between-subject variation relative to the total variation is commonly used to estimate reliability. Shoukri and colleagues provide sample size calculations based on precision with which reliability can be estimated for test-retest data from one or more observers(39). This approach uses the confidence interval (CI) width as a measure of precision, level of confidence (i.e. alpha value), and specification of a “planning value” for the reliability estimate. For example, if high intra-rater reliability is anticipated (proportion: 0.9) with sufficient precision (95% CI width: 0.2), the minimum number of scans to be rescored is 15. Sample size requirements will increase for the following inputs: lower values of planned reliability, higher precision and a higher confidence level. It is recommended that test-retest sample size, like other sample sizes described later, be calculated in the study design phase, as it will heavily depend upon the formula inputs, funding and other resource constraints.

Agreement between the primary observer (an experienced or certified observer whose scores are considered the benchmark) and other observer(s) is needed to check consistency. CT studies typically involve a primary observer and at least one other independent observer who is trained in the scoring system as part of the study. Depending on the extent to which a scoring system has been validated for the target population, it is important to have well-calibrated scores. This process is often qualitative and consists of a subset of scans, chosen by the primary observer, being iteratively scored by the novice observer who discusses results with the primary observer. It is recommended that the calibration subset be chosen via stratified randomization. The scans should be selected through random sampling

stratified according to age and lung disease severity, to ensure calibration is broadly achieved. Sample size requirements for estimating inter-rater reliability can be formed as previously described. For example, if we consider two independent observers to have excellent agreement on their scores (proportion: 0.8) with sufficient precision (95% CI width: 0.2), the minimum number of scans to be scored by each observer is 52. If acquiring high numbers of subjects for imaging is more costly than utilizing multiple observers, then the sample size can be recalculated to include a higher number of observers. Assuming the same agreement and precision as before but utilizing three observers, for example, requires a minimum of 36 scans to be scored.

There are numerous inter-rater reliability indices available for CT studies (Table 2). In the majority of CF studies, the index of choice is the intra-class correlation coefficient (ICC) (40). There are many versions of ICC, but the most commonly utilized version in the CF literature is the original based on the one-way random effects ANOVA model, which assumes that raters are interchangeable. In studies in which there is an experienced rater, the one-way random effects ANOVA can be modified to assess consistency across the different raters. Drawbacks to use of ICC and its categorical analogue, the kappa statistic (41), have been well described in the statistics literature, with a recent illustration involving cardiovascular imaging markers(42). A high estimate of ICC for an overall study sample does not always indicate strong agreement. For example, a CF study with participants who vary in age will likely yield CT markers with broad ranges (i.e., high between subject variability). Regardless of observer agreement, the ICC ratio will unfairly leverage the high between-subject variability relative to within-subject variability to produce a value close to 1. Analogously, a study with a narrow subject age range may yield small between subject variability for each CT marker with ICC estimates that are incorrectly low. For these reasons, it is important to stratify ICC estimates by age and disease severity, which are known sources of heterogeneity in CF clinical studies. Poor ICC estimates within strata imply that additional calibration is necessary.

An alternative reliability statistic that relaxes the ANOVA assumption in the ICC is the Concordance Correlation Coefficient (CCC) (43); however, this statistic does not account for chance agreement and has not been widely used to assess reliability of imaging markers. Threshold-based approaches, such as Bland-Altman analysis(44) and coverage probability (CP), allow more targeted identification of discrepancies. The Bland-Altman approach consists of plotting the mean of differences between observer measurements of a given CT marker against the average of measurement pairs from the observers; limits of agreement on the plot can be used to identify systematic differences in observers across the range of measurement values, or instances of increased variation (i.e. lower precision) between observers. These limits are determined by mean difference, standard deviation of the differences and sample size. CP has the most advantages of the methods described (Table 2) for all stages of reliability assessment, but requires a priori identification of an acceptable threshold for the paired difference in observer scores.

Given the breadth of available reliability indices for CT markers, it is recommended that researchers report estimates for their selected reliability index and descriptive statistics for each type of CT marker, both overall and stratified by age and lung disease severity. The

additional estimates can be used to calculate other reliability indices, such as CP, enabling ‘apples-to-apples’ comparisons of reliability estimates across CF studies and populations.

3.3 Acquiring baseline CT measurement

Baseline CT collection in clinical trials is necessary for assessing randomization, safety and quality of procedure/protocol adherence. Even in modified protocols described previously, young children (aged below 5 years) will not have the same participant performance in CT studies at baseline as they do in follow-up. It is recommended that the mean and variability of baseline CT data be examined prior to inclusion in efficacy analyses. The coefficient of variation, which is the SD divided by the mean in the sample, can be used to examine the quality of baseline data. A large value for this ratio would imply that the baseline data may not be reliable to assess efficacy; however, these data could still be used for phenotyping the patient and checking randomization. These issues do not exist for child participants who are 5 years of age or older; change scores or other longitudinal variables have been validated in previous CF studies(3, 5, 19) and can be used in efficacy analyses.

As an individual grows, the sensitivity of a given CT marker will increase, due to airway development, improved ability to perform protocol steps (e.g. breath holds) and other factors. Growth introduces subtle changes related to the CT scanner performance that have not been thoroughly explored in CF studies. Discarding any baseline data is undesirable from a statistical standpoint. Depending on the coefficient of variation, it may be appropriate to include data from very young cohorts by adjusting for growth using established metrics. This could possibly be achieved using metrics as covariates or by performing a calibration analysis. The effectiveness of these approaches to salvage baseline CT data is unknown but could be investigated with the advent of young CF cohorts with imaging data.

4 Modeling CT Markers as Clinical Endpoints

4.1 Model Assumptions

The collection of CT marker data on an individual subject will consist of overall and region-specific scores. We will need to assume that the outcome variable, given a treatment and possibly other exposure variables, forms a model that follows some type of statistical distribution. Most often, we assume a (multivariate) normal distribution. Distributions of CT markers and modeling assumptions should be assessed and reported in clinical studies. A normal distribution may be reasonable for CT markers observed at later stages of disease severity, but it is plausible in early-stage CF to encounter subjects whose CT markers are long-tailed (e.g. log-normal) or have zero values (e.g., bronchiectasis). Zeroes may be indicative of minimal or absent structural lung disease, an outcome that may be of interest itself. The data may appear “clumped” at zero, as there could be a proportion of subjects who at a young age have not experienced an insult in addition to their CFTR dysfunction resulting in bronchiectasis or other structural lung changes at a young age. Examples of such additional insults can be viral infection or acquisition of *Pseudomonas aeruginosa*. The remaining subjects who have developed structural lung disease will have continuous values for CT markers. The resulting data forms a skewed distribution that may not be well

approximated with a symmetric, bell-shaped curve like the normal distribution. Data of this nature are sometimes referred to as zero-inflated or semicontinuous data(45).

Examples of non-normally distributed data have been recently encountered in the Australian Respiratory Early Surveillance Team for CF (AREST CF) cohort (Figure 2). PRAGMA-CF markers for disease and bronchiectasis, obtained with permission from the previously published cohort data(9), have substantial lack of fit under the normal assumption. The mean and SD are both overestimated for % disease and % bronchiectasis. The fit to the % disease data is improved by accounting for the skewness using a lognormal distribution. Special care is required to fit the % bronchiectasis data, as roughly 42% of the data are zeroes. Min and Agresti reviewed and discussed practical strategies to combat zero-inflated and semi-continuous data(45), which can be extended for longitudinal studies through use of random effects(46). Assuming a normal distribution in either scenario could yield biased estimates and misleading results about the extent of structural lung disease in the cohort being studied, and about the effects of treatment or associations with other exposure variables. Approaches to model CT markers and their relationships with covariates using alternative distributions should be considered for CT as well as for MRI marker data analysis.

4.2 Differential recruitment and missing data in CT studies

In prospective CT studies, participation will depend upon the eagerness of the individual subject. If the study has an intervention arm, participation will generally be high for CF subjects, regardless of whether CT scans are part of the protocol. Subject retention may be complicated in longitudinal settings in which it is expected that a subset of participants could drop out, miss scans, or receive scans at times that are not commensurate with the protocol (creating mistimed measurements). The nature of missing data in CT studies can also include technical failures, protocol violations (e.g. inappropriate data storage or slice thickness), and subject noncompliance. At the very least, missing data can reduce efficiency, thereby limiting statistical power; at worst, missing data can seriously bias study findings. In CT studies, this could imply incorrect conclusions about structural lung disease progression over time, or limited ability to detect efficacy in CF therapeutic studies. Practical methods to limit missing data and address the potential bias via statistical analyses, accounting for the missing data mechanism, are available(47).

In retrospective CT analyses, caution should be applied when selecting individual scans for inclusion. It has been shown in US CF registry analyses that sicker patients tend to have more clinical encounters. Because these patients are sicker, they tend to have worse outcomes(48). Such issues are avoided when only routine clinical scans are included in the analysis. It is recommended to randomly select scans for inclusion, if the study is retrospective. In prospective studies, such as clinical trials, the impact of including clinically-indicated scans could be assessed through sensitivity analyses. Ignoring this source of sampling bias could produce misleading results about associations between treatment and progression of structural lung disease as measured by CT.

4.3 CT marker selection and sample size for models

Several CT markers have been developed over the years and have varying levels of sensitivity (Table 1). CT marker variability depends upon the type of cohort being studied, the scan protocol and scoring algorithm. Minimizing these sources of variability, in turn, maximizes the precision with which treatment effects or associations can be examined using a particular CT marker. Bronchiectasis and trapped air can be established with great precision and are well validated as clinically relevant endpoints. Scoring of airway thickness is more system dependent. The recently objective AA method allows assessment of both bronchiectasis and airways wall thickness with great precision; however, this method is time-consuming and not yet automated. The less intensive PRAGMA-CF scoring system has been shown to correlate well with the AA method and thus is currently the best available method as shown in published work (18) and studies now under review

CT marker effect sizes have not been thoroughly described in the literature, although estimates may be gleaned from completed studies (Table 1). None of the current CT markers have a designated minimally important clinical difference, a threshold used to indicate the smallest change in outcome that a patient would still identify as clinically important. Percent reduction for a given CT marker has been proposed as a biologically plausible outcome for CF clinical studies(18). For example, a 30% reduction in trapped air would be considered clinically relevant. It is worth noting that whatever particular feature the CT marker is intended to measure also determines clinical meaning of the % reduction. For instance, extent of bronchiectasis is a monotonically increasing attribute of structural lung disease, whereas trapped air may be reversed, to some extent, over time.

Both the choice of CT marker and threshold indicating clinically meaningful % reduction will impact sample size required for an interventional study (Figures 2c–2d). Effect sizes were formulated based on mean PRAGMA % disease and % bronchiectasis in children aged 0–5 in AREST CF cohort (9) and were calculated as a % difference between means of two hypothetical treatment arms, assuming a traditional two group clinical trial design (i.e. mean % disease in children aged 0–5 is 1.90, 30% reduction would imply that in the interventional trial arm mean % disease would be reduced to 1.33, 50% would reduce it to 0.95, etc.). Higher magnitudes of the relative difference in % disease (Figure 2c) and % bronchiectasis (Figure 2d) allow for lower sample size requirements per group. Given the relatively low numbers of CF participants, it is likely that only large magnitudes of relative difference will be detectable in interventional studies using scoring systems such as PRAGMA-CF. More sensitive outcome measures, such as the AA-ratio, have the potential to improve precision, thereby enabling detection of relative difference with lower sample sizes.

5. Recommendations

With the number of imaging studies in CF and the advent of (semi) quantitative methods, there are several ways to improve the utility of CT and MRI markers. CT markers have been well validated in the literature, but their role in the randomized controlled trial setting still needs to be proven. MRI markers for studying CF lung disease progression are at an early stage; research to date indicates that this modality tends to overestimate the extent of early disease and underestimate advanced disease. Given the current findings on standardization

for each modality, CT and MRI markers may be feasible for Phase III and single center Phase II studies, respectively. There is now a CTN-TDN task force, composing guidelines that will further harmonize study protocol recommendations, complimenting existing efforts in the European Union that have been published (17) and are under review for CT. Similar efforts have recently been initiated for MRI. Furthermore, minimal barriers exist to implementing these standards for routine clinical scans. Clinical standardization at CTN-TDN sites is expected to decrease bias inherent in the historical (less standardized) CT scanning and data collection methods, thereby enabling data pooling. Another strategy to minimize bias in MRI and CT studies is training and certification of observers. Additional standards for scoring systems should include randomized ordering, de-identification of scans, uniform lighting conditions, and well-defined analysis time. The advent of automated image analysis systems is also expected to decrease inherent bias in CT and MRI data processing.

Reproducibility is essential to elevate pre-processing and statistical analysis standards for imaging studies. Researchers should use online supplements to provide detailed data summaries, including formulas for pre-processing, aggregate calculations and statistical considerations, as in the recent study by Ramsey and colleagues (9). Accessing completed studies via data repositories could also streamline development of novel and robust imaging analysis methods. Such repositories could be leveraged to gain historical control data for examining novel therapies, similar to the use of CF patient registries to gain historical control data as comparisons of clinical trial findings. As reproducibility improves, annual reporting of summary statistics or trends in clinical data, such as lung function, BMI, and other markers reported by the US CF Foundation(49), could be expanded to include CT markers. With multiple modalities being used in parallel, efforts to use CT in conjunction with MRI and other markers can be considered to effectively monitor CF lung disease.

Although challenges with reliability, model assumptions and missing data are still at the forefront of imaging marker analysis, new challenges will emerge with more automated scoring systems, requiring development of spatial data models to understand detailed lung structure and consideration of multiple comparison adjustments as debated in the neuroimaging literature(50). Additional prospective longitudinal studies that include both imaging markers and functional outcomes will be helpful to examine associations between the two evolve over time, improving our understanding of both imaging markers and of the more traditionally used functional markers.

Acknowledgments

The authors are grateful to Tim Rosenow, BSc, Grad Cert PaedRespSci, PhD Candidate, and Mekibib Altaye, PhD, for their critical reading of the manuscript and valuable comments. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH/NHLBI.

Conflict of Interest

Author RS received support for this work from the National Heart, Lung and Blood Institute (NHLBI) of the National Institutes of Health (NIH) under award number K25 HL125954. Author HT acted in the last 4 years as consultant for the Sophia BV of the Erasmus MC Sophia Children's Hospital on advisory boards for Gilead Sciences, Novartis Pharmaceuticals, Insmed, Vertex, and PTC. In addition the Sophia BV received speaker fees for presentations by HT from Gilead, Vertex, and Roche. He does not own any stock. The Sophia BV received unconditional research grants from Roche, Chiesi, Novartis and Gilead for research supervised by HT. He is the

founder and director of Erasmus MC LungAnalysis core laboratory for image analysis under the co-supervision of the research bureau of the department of Radiology.

References

1. Brody AS, Molina PL, Klein JS, Rothman BS, Ramagopal M, Swartz DR. High-resolution computed tomography of the chest in children with cystic fibrosis: support for use as an outcome surrogate. Pediatr Radiol. 1999; 29(10):731–735. [PubMed: 10525778]
2. Loeve M, Krestin GP, Rosenfeld M, de Bruijne M, Stick SM, Tiddens HA. Chest computed tomography: a validated surrogate endpoint of cystic fibrosis lung disease? Eur Respir J. 2013; 42(3):844–857. [PubMed: 23258780]
3. De Jong P, Nakano Y, Lequin M, Mayo J, Woods R, Pare P, et al. Progressive damage on high resolution computed tomography despite stable lung function in cystic fibrosis. European Respiratory Journal. 2004; 23(1):93–97. [PubMed: 14738238]
4. Brody AS, Tiddens HA, Castile RG, Coxson HO, de Jong PA, Goldin J, et al. Computed tomography in the evaluation of cystic fibrosis lung disease. Am J Respir Crit Care Med. 2005; 172(10):1246–1252. [PubMed: 16100011]
5. Owens CM, Aurora P, Stanojevic S, Bush A, Wade A, Oliver C, et al. Lung Clearance Index and HRCT are complementary markers of lung abnormalities in young children with CF. Thorax. 2011; 66(6):481–488. [PubMed: 21422040]
6. Tepper LA, Caudri D, Utens EM, van der Wiel EC, Quittner AL, Tiddens HA. Tracking CF disease progression with CT and respiratory symptoms in a cohort of children aged 6–19 years. Pediatr Pulmonol. 2014; 49(12):1182–1189. [PubMed: 24574038]
7. Tiddens HA. Chest computed tomography scans should be considered as a routine investigation in cystic fibrosis. Paediatr Respir Rev. 2006; 7(3):202–208. [PubMed: 16938643]
8. Puderbach M, Eichinger M. The role of advanced imaging techniques in cystic fibrosis follow-up: is there a place for MRI? Pediatr Radiol. 2010; 40(6):844–849. [PubMed: 20432002]
9. Ramsey KA, Rosenow T, Turkovic L, Skoric B, Banton G, Adams AM, et al. Lung Clearance Index and Structural Lung Disease on Computed Tomography in Early Cystic Fibrosis. Am J Respir Crit Care Med. 2016; 193(1):60–67. [PubMed: 26359952]
10. Rosenfeld M. An overview of endpoints for cystic fibrosis clinical trials: one size does not fit all. Proc Am Thorac Soc. 2007; 4(4):299–301. [PubMed: 17652489]
11. Bennett TI. Discussion on the stethoscope versus X-rays. Proceedings of the Royal Society of Medicine. 1945; 355:7–9.
12. Ciet P, Wielopolski P, Li S, Andrinopolou EA, van Der Wiel E, Morana G, TIddens HAWM. Mosaic pattern in Cystic Fibrosis lung disease: trapped air or hypoperfusion? (submitted).
13. Stephenson AL, Tom M, Berthiaume Y, Singer LG, Aaron SD, Whitmore GA, et al. A contemporary survival analysis of individuals with cystic fibrosis: a cohort study. Eur Respir J. 2015; 45(3):670–679. [PubMed: 25395034]
14. Brody AS, Klein JS, Molina PL, Quan J, Bean JA, Wilmott RW. High-resolution computed tomography in young patients with cystic fibrosis: distribution of abnormalities and correlation with pulmonary function tests. J Pediatr. 2004; 145(1):32–38. [PubMed: 15238903]
15. Brody AS, Kosorok MR, Li Z, Broderick LS, Foster JL, Laxova A, et al. Reproducibility of a scoring system for computed tomography scanning in cystic fibrosis. J Thorac Imaging. 2006; 21(1):14–21. [PubMed: 16538150]
16. de Jong PA, Ottink MD, Robben SG, Lequin MH, Hop WC, Hendriks JJ, et al. Pulmonary disease assessment in cystic fibrosis: comparison of CT scoring systems and value of bronchial and arterial dimension measurements. Radiology. 2004; 231(2):434–439. [PubMed: 15064392]
17. Kuo W, Kemner-van de Corput MP, Perez-Rovira A, de Bruijne M, Fajac I, Tiddens HA, et al. Multicentre chest computed tomography standardisation in children and adolescents with cystic fibrosis: the way forward. Eur Respir J. 2016; 47(6):1706–1717. [PubMed: 27076593]
18. Rosenow T, Oudraad MC, Murray CP, Turkovic L, Kuo W, de Bruijne M, et al. PRAGMA-CF. A Quantitative Structural Lung Disease Computed Tomography Outcome in Young Children with Cystic Fibrosis. Am J Respir Crit Care Med. 2015; 191(10):1158–1165. [PubMed: 25756857]

19. Tepper LA, Utens EM, Caudri D, Bos AC, Gonzalez-Graniel K, Duivenvoorden HJ, et al. Impact of bronchiectasis and trapped air on quality of life and exacerbations in cystic fibrosis. European Respiratory Journal. 2013; 42(2):371–379. [PubMed: 23314900]
20. Wainwright CE, Vidmar S, Armstrong DS, Byrnes CA, Carlin JB, Cheney J, et al. Effect of bronchoalveolar lavage-directed therapy on Pseudomonas aeruginosa infection and structural lung injury in children with cystic fibrosis: a randomized trial. JAMA. 2011; 306(2):163–171. [PubMed: 21750293]
21. Mott LS, Graniel KG, Park J, de Klerk NH, Sly PD, Murray CP, et al. Assessment of early bronchiectasis in young children with cystic fibrosis is dependent on lung volume. Chest. 2013; 144(4):1193–1198. [PubMed: 23681147]
22. Loeve M, Lequin MH, de Bruijne M, Hartmann IJ, Gerbrands K, van Straten M, et al. Cystic fibrosis: are volumetric ultra-low-dose expiratory CT scans sufficient for monitoring related lung disease? Radiology. 2009; 253(1):223–229. [PubMed: 19710003]
23. Stick SM, Brennan S, Murray C, Douglas T, von Ungern-Sternberg BS, Garratt LW, et al. Bronchiectasis in infants and preschool children diagnosed with cystic fibrosis after newborn screening. J Pediatr. 2009; 155(5):623–628. e1. [PubMed: 19616787]
24. Loeve M, Hop WC, de Bruijne M, van Hal PT, Robinson P, Aitken ML, et al. Chest computed tomography scores are predictive of survival in patients with cystic fibrosis awaiting lung transplantation. Am J Respir Crit Care Med. 2012; 185(10):1096–1103. [PubMed: 22403801]
25. Loeve M, van Hal PT, Robinson P, de Jong PA, Lequin MH, Hop WC, et al. The spectrum of structural abnormalities on CT scans from patients with CF with severe advanced lung disease. Thorax. 2009; 64(10):876–782. [PubMed: 19541686]
26. Loeve M, de Bruijne M, Hartmann IC, van Straten M, Hop WC, Tiddens HA. Three-section expiratory CT: insufficient for trapped air assessment in patients with cystic fibrosis? Radiology. 2012; 262(3):969–976. [PubMed: 22357896]
27. van Beek EJ, Hill C, Woodhouse N, Fichele S, Fleming S, Howe B, et al. Assessment of lung disease in children with cystic fibrosis using hyperpolarized 3-Helium MRI: comparison with Shwachman score, Chrispin-Norman score and spirometry. Eur Radiol. 2007; 17(4):1018–1024. [PubMed: 16941089]
28. Wielputz MO, Puderbach M, Kopp-Schneider A, Stahl M, Fritzsching E, Sommerburg O, et al. Magnetic resonance imaging detects changes in structure and perfusion, and response to therapy in early cystic fibrosis lung disease. Am J Respir Crit Care Med. 2014; 189(8):956–965. [PubMed: 24564281]
29. Failo R, Wielopolski PA, Tiddens HA, Hop WC, Mucelli RP, Lequin MH. Lung morphology assessment using MRI: a robust ultra-short TR/TE 2D steady state free precession sequence used in cystic fibrosis patients. Magn Reson Med. 2009; 61(2):299–306. [PubMed: 19165879]
30. Ciet P, Serra G, Bertolo S, Ros M, Assael BM, Morana G, et al. Comparison of chest-MRI to chest-CT to monitor cystic fibrosis lung disease. Pediatr Pulmonol. 2010
31. Ley-Zaporozhan J, Molinari F, Risse F, Puderbach M, Schenk JP, Kopp-Schneider A, et al. Repeatability and reproducibility of quantitative whole-lung perfusion magnetic resonance imaging. J Thorac Imaging. 2011; 26(3):230–239. [PubMed: 20818278]
32. VanDevanter DR, Konstan MW. Outcome measures for clinical trials assessing treatment of cystic fibrosis lung disease. Clin Investig (Lond). 2012; 2(2):163–175.
33. Group NDW. Biomarkers and surrogate endpoints in clinical research: definitions and conceptual model. Amsterdam. 2000
34. Ciet P, Wielopolski P, Manniesing R, Lever S, de Bruijne M, Morana G, et al. Spirometer-controlled cine magnetic resonance imaging used to diagnose tracheobronchomalacia in paediatric patients. Eur Respir J. 2014; 43(1):115–124. [PubMed: 23598953]
35. Weintraub WS, Lüscher TF, Pocock S. The perils of surrogate endpoints. European heart journal. 2015; 36(33):2212–2218. [PubMed: 25975658]
36. Woodhouse N, Wild JM, van Beek EJ, Hoggard N, Barker N, Taylor CJ. Assessment of hyperpolarized 3He lung MRI for regional evaluation of interventional therapy: a pilot study in pediatric cystic fibrosis. J Magn Reson Imaging. 2009; 30(5):981–988. [PubMed: 19856418]

37. Salamon E, Lever S, Kuo W, Ciet P, Tiddens HA. Spirometer guided chest imaging in children: It is worth the effort! Pediatr Pulmonol. 2016
38. Carroll, RJ. Measurement error in epidemiologic studies. In: Armitage, PTC., editor. Encyclopedia of Biostatistics. 3. New York: John Wiley & Sons; 1998. p. 2491-2519.
39. Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. Statistical Methods in Medical Research. 2004; 13(4):251–271.
40. de Jong PA, Tiddens HA. Cystic Fibrosis-Specific Computed Tomography Scoring. Proceedings of the American Thoracic Society. 2007; 4(4):338–342. [PubMed: 17652497]
41. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–174. [PubMed: 843571]
42. Barnhart HX, Yow E, Crowley AL, Daubert MA, Rabineau D, Bigelow R, et al. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. Stat Methods Med Res. 2014
43. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989; 45(1): 255–268. [PubMed: 2720055]
44. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476):307–310. [PubMed: 2868172]
45. Min YAA. Modeling Nonnegative Data with Clumping at Zero: A Survey. JIRSS. 2002; 1(1–2):7–33.
46. Gupta R, Szczesniak RD, Macaluso M. Modeling repeated count measures with excess zeros in an epidemiological study. Ann Epidemiol. 2015; 25(8):583–589. [PubMed: 25887702]
47. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012; 367(14):1355–1360. [PubMed: 23034025]
48. VanDyke RD, McPhail GL, Huang B, Fenchel MC, Amin RS, Carle AC, et al. Inhaled tobramycin effectively reduces FEV1 decline in cystic fibrosis. An instrumental variables analysis. Ann Am Thorac Soc. 2013; 10(3):205–212. [PubMed: 23802816]
49. Registry CFFP. 2014 Annual Data Report. Bethesda, MD: 2015.
50. Bennett CM, Wolford GL, Miller MB. The principled control of false positives in neuroimaging. Soc Cogn Affect Neurosci. 2009; 4(4):417–422. [PubMed: 20042432]
51. de Jong PA, Lindblad A, Rubin L, Hop WC, de Jongste JC, Brink M, et al. Progression of lung disease on computed tomography and pulmonary function tests in children and adults with cystic fibrosis. Thorax. 2006; 61(1):80–85. [PubMed: 16244089]
52. Gustafsson PM, De Jong PA, Tiddens HA, Lindblad A. Multiple-breath inert gas washout and spirometry versus structural lung disease in cystic fibrosis. Thorax. 2008; 63(2):129–134. [PubMed: 17675316]
53. Loeve M, Gerbrands K, Hop WC, Rosenfeld M, Hartmann IC, Tiddens HA. Bronchiectasis and pulmonary exacerbations in children and young adults with cystic fibrosis. Chest. 2011; 140(1): 178–185. [PubMed: 21148242]
54. Sanders DB, Li Z, Brody AS, Farrell PM. Chest computed tomography scores of severity are associated with future lung disease progression in children with cystic fibrosis. Am J Respir Crit Care Med. 2011; 184(7):816–821. [PubMed: 21737586]
55. Bortoluzzi CF, Volpi S, D'Orazio C, Tiddens HA, Loeve M, Tridello G, et al. Bronchiectases at early chest computed tomography in children with cystic fibrosis are associated with increased risk of subsequent pulmonary exacerbations and chronic pseudomonas infection. J Cyst Fibros. 2014; 13(5):564–571. [PubMed: 24726420]

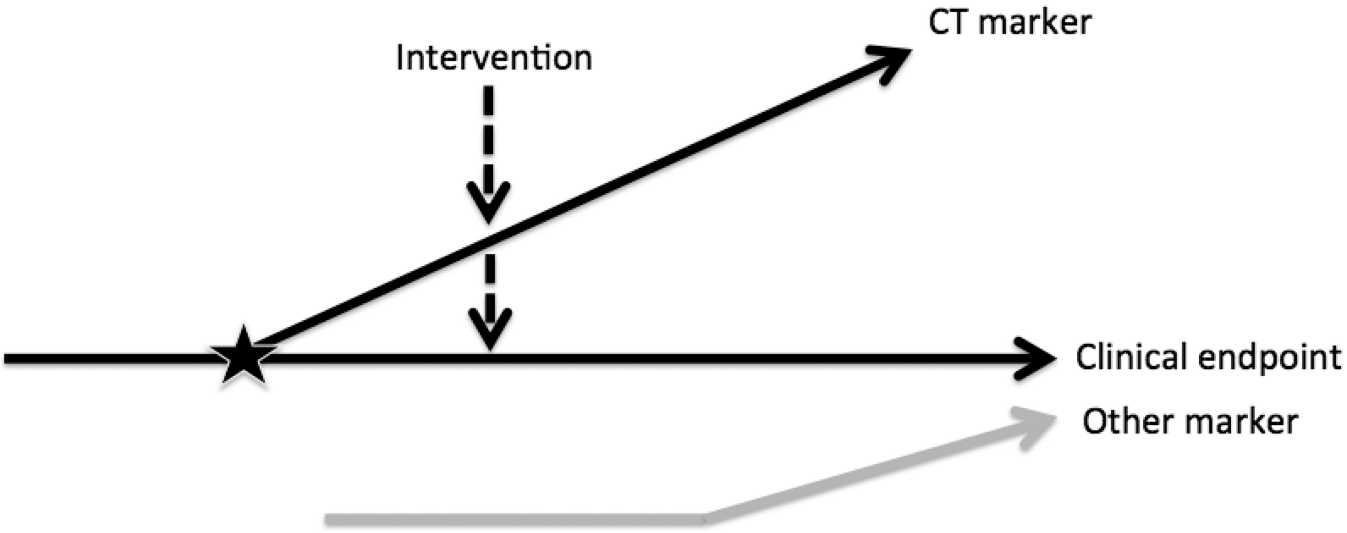

Figure 1. Conceptual Model of Relationship between CT and Clinical Endpoints

Conceptual model adapted from Weintraub et al (35) reflects the uncertainty regarding the nature of the relationship between a given CT marker (surrogate), the clinical endpoint and other markers. The horizontal solid, black line represents the causal pathway of the CF disease process on which the clinical endpoint is situated (e.g., survival). The sloped, solid black line represents the CT marker in relation to the causal pathway. If these two lines intersect as indicated by the star, then the CT marker is a true surrogate. An intervention is likely applied downstream in the CF disease process, affecting the CT marker and/or the clinical endpoint, as noted by the dashed line with a downward arrow on each of the sloped and horizontal lines. It is also possible that another marker (e.g. FEV_1%) affects the clinical endpoint independently of the CT marker as indicated by the gray line.

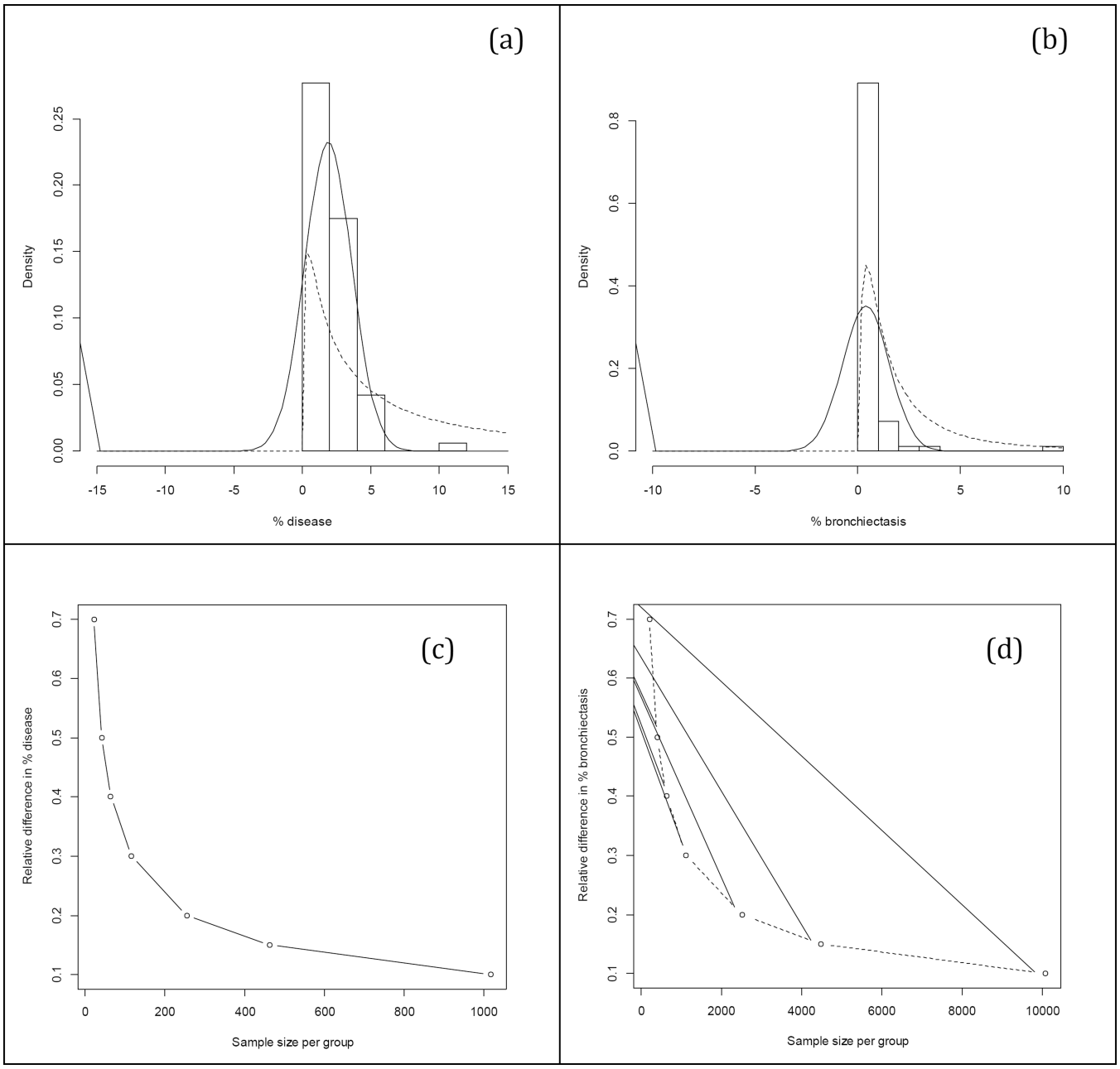

Figure 2. CT Marker Distributions and Sample Size

PRAGMA data from AREST CF cohort (N=83 scans) in children less than 5 years of age. Histograms of % disease (a) and % bronchiectasis (b) with overlaying normal and lognormal distributions (solid and dashed curves, respectively); the “clump” of data in the leftmost vertical bar represents instances in which zero was the observed marker value; the sample mean (SD) for (a) and (b) were 1.90 (1.72) and 0.40 (1.14), respectively. Relative differences in % disease versus sample size (c) and in % bronchiectasis versus sample size (d); differences were based on mean and % reduction for each CT marker for two hypothetical treatment arms, assuming 80% power and type I error rate of 5%.

Table 1

Summary of chest-computed tomography endpoints and association studies

First Author year	Study Design and Data	Number of patients* (N)	Cohort age range at enrollment*	Follow-up duration	Scoring method(s)	Primary findings
Brody (1999) (1)	Retrospective cohort with hospital admissions	8	5–16 years	42 days (maximum)	Brody	CT scores (total airway disease, peribronchial thickening, mucous plugging, and overall appearance) improved from hospital admission to discharge
Brody (2004) (14)	Baseline data from Pulmozyme Early Intervention Trial	60	6–10 years	Cross-sectional	Brody	Introduced weighted CT scoring, demonstrated fair to moderate correlations between subscores and pulmonary function outcomes (FEV_1, FVC and $FEF_{25–75}$)
De Jong (2004) (3)	Retrospective clinical data from annual PFT and biennial CT scans	48	11.1 (mean) years	2 years	Brody II, Castile, Helbich, Santamaria, Bhalla	Each CT scoring method provided more sensitive markers of CF disease progression than pulmonary function outcomes
Kuo (2016) (JCF)	Randomly-selected CT scans of CF patients and pediatric pulmonology referrals	11; 12 Controls	11 (median), 7–16 years; 13.9, 6–16 years for Controls	Cross-sectional	CF-CT, PRAGMA-CF, AA-method	Presented AA ratio method; CF-CT and PRAGMA-CF were sensitive methods to score bronchiectasis and airway wall thickness; PRAGMA-CF more accurately detected bronchiectasis, compared to CF-CT; CF-CT was a more accurate method than PRAGMA-CF to detect airway wall thickness
de Jong (2006) (51)	Retrospective clinical cohort	119	5–52 years	6 years (maximum)	Brody II	Bronchiectasis score declined more quickly over time, compared to pulmonary function outcomes
Gustafsson (2007) (52)	Retrospective study using annual clinical scans	44	5–19 years	Cross-sectional	Brody II	Lung clearance index was associated with CT scores; some patients with abnormal lung clearance index measurements had normal CT
Stick (2009) (23)	Baseline data from AREST CF study	96	0–6 years	Cross-sectional	Adapted CF-CT	Probability of bronchiectasis increased with age and was associated with *Pa* infection, neutrophil count and elastase concentration
Loeve (2011) (53)	Retrospective clinical cohort	115	5–20 years	0–2 years	CF-CT	CT scores were predictive of subsequent pulmonary exacerbation frequency
Sanders (2011) (54)	Longitudinal data from Wisconsin CF Neonatal	81	11.5 (mean) years	7.5 years (mean)	Brody	CT scores are more strongly associated with chest radiograph and subsequent lung disease

J Cyst Fibros. Author manuscript; available in PMC 2017 March 07.

First Author year	Study Design and Data	Number of patients* (N)	Cohort age range at enrollment*	Follow-up duration	Scoring method(s)	Primary findings
	Screening Project					severity measures, compared to spirometry measures
Bortoluzzi (2014) (55)	Retrospective center cohort	83	5–7 years	6 years	CF-CT	Higher bronchiectasis score was associated with increased number of respiratory exacerbations and Pa infection at follow up; bronchiectasis score had higher sensitivity than FEV_1 for respiratory exacerbation prediction
Rosenow (2015) (18)	CT scans from AREST CF study	30	0–6 years	2 years	PRAGMA-CF/ CF-CT	New scoring method (PRAGMA CF) improved repeatability and sensitivity to early-stage disease progression, compared to CF-CT
Tepper (2013) (19)	Clinical cohort with CT and CFQ-R performed on same day	72	6–20 years	1 year	CF-CT	Bronchiectasis, trapped air and CFQ-R RSS were associated with pulmonary exacerbations
Wainwright (2011) (20)	ACFBAL Randomized controlled trial	170	3.6 (mean); 1.6 (SD) months	5 years (maximum)	CF-CT	BAL-directed therapy did not result in lower total CF-CT score, compared to standard therapy
Ramsey et al (2016) (9)	Annual surveillance and clinically-indicated chest CT at center	119	0–16 years	Cross-sectional	PRAGMA-CF	Total disease extent was associated with lung clearance index; bronchiectasis and air trapping were associated with lung clearance index in children>3 years of age

Abbreviations: Airway Artery (AA); Australasian Cystic Fibrosis Bronchoalveolar Lavage (ACFBAL); Australian Respiratory Early Surveillance Team for CF (AREST CF); Cystic fibrosis (CF); Cystic Fibrosis Questionnaire Revised (CFQ-R); High-resolution chest computed tomography (CT); Perth-Rotterdam Annotated Grid Morphometric Analysis for Cystic Fibrosis (PRAGMA-CF)

Table 2

Overview of select reliability indices for chest-computed tomography markers

Index	Overview	Range	Interpretation	Advantage(s) for CT	Disadvantage(s) for CT
Pearson's *r*	Measures strength of linear association between scores from two independent observers	−1 to 1	Values closer to −1 or 1 indicate strong negative or positive relationship, respectively	Easy to compute with standard software; method is well known and coefficient is easy to interpret	Cannot be used to assess intra-rater reliability because it requires exactly 2 independent observers; inappropriate, generally, for reliability because it measures linearity: a perfect, positive correlation ($r = 1$) could result if two observers have systematically different scores; cannot be used for categorical CT scores in CF (e.g. presence/absence of bronchiectasis)
Bland-Altman analysis	CI (typically 95%) for mean difference Δ between two sets of observer scores	Depends on Δ estimate	When used with Bland-Altman plot, LOA can show systematic differences and variability	Easy to interpret plot; statistical approach is straightforward to acquire mean difference and CI	Influenced by normality assumption; only valid for continuous scores; cannot be used if >2 observers; assumes one method of measurement is the 'gold standard' and is therefore known; requires understanding of what is acceptable/unacceptable difference in observer scores
ICC/weighted Kappa statistic	Ratio of between-subject variability to total variability, where total variability is the sum of between- and within-subject variability	−1 to 1	Scaled coefficient ranges adopted in the literature (not recognized widely in statistics) indicate fair (0.4 to 0.6), moderate (0.6 to 0.8) and excellent agreement (0.8 to 1).	Unit-less value; has continuous (ICC) and categorical (Kappa) versions to accommodate different CT subscores and interpretation is consistent for both data types; can be used to assess intra-rater reliability; can accommodate >2 observers	Interpretation for the same CT marker across different populations and/or studies is not consistent; high estimate does not always reflect excellent agreement; requires that ANOVA assumptions are met
CCC	1 minus the ratio of within-subject squared deviation to total deviation	−1 to 1	Scaled coefficient may be thought of as a standardized estimate of the mean squared difference between observers	Same advantages as ICC; relaxes the ANOVA assumption required for ICC; currently the only reliability statistic endorsed by the Metrics Champion Consortium for Imaging	Estimates often correspond to ICC and therefore have same issues with interpretation across multiple studies/populations
CP	Proportion of scans with differences in scores that fall within an acceptable threshold	0 to 1	Unscaled index expressed as a probability estimate, where values close to 1 indicate higher reliability	Can be computed using nonparametric approach for any data type; can be used for > 2 observers; has consistent interpretation across multiple studies/ populations; can be used to pinpoint specific instances of poor reliability	A threshold must be set a priori defining a meaningful difference between observer scores, which can be difficult when studying novel CT markers or systems; this threshold will impact the probability estimate

Abbreviations: Analysis of variance (ANOVA); chest-computed tomography (CT); concordance correlation coefficient (CCC); coverage probability (CP); cystic fibrosis (CF); intra-class correlation coefficient (ICC); limits of agreement (LOA)