Influence of Calibration on Densitometric Studies of Emphysema Progression Using Computed Tomography

David G. Parr, Berend C. Stoel, Jan Stolk, Peter G. Nightingale, and Robert A. Stockley

Lung Investigation Unit and Wellcome Trust Clinical Research Facility, Queen Elizabeth Hospital, Birmingham, United Kingdom; and Division of Image Processing, Department of Radiology, and Department of Pulmonology, Leiden University Medical Center, Leiden, The Netherlands

The fundamental importance of calibration for any measuring device is indisputable, but computed tomography (CT) calibration in longitudinal lung densitometry studies is largely unexplored. Although the validity of CT as a measure of emphysema has been confirmed in cross-sectional studies, there are limited data on long-term reproducibility, and this is critically important for validating its use as an outcome measure in therapeutic trials. A general understanding of the strengths and pitfalls of CT densitometry is critical for physicians reviewing the published literature using this methodology. In our study of 57 patients with alpha-1 antitrypsin deficiency (phenotype PiZ), progression of voxel index determined from three successive annual scans acquired with a fully calibrated scanner was intimately associated with changes in CT air densitometry, sampled from patient images. Images were therefore reanalyzed, using a correction technique validated in phantom studies that adjusted for changes in measured air density, and the reliability of the voxel index as a measure of emphysema progression was improved. Comparison of adjusted voxel index thresholds indicated the optimum threshold was –950 Hounsfield units. Internal air calibration is therefore critical in longitudinal and multicenter lung densitometry studies of emphysema and incorporation of a correction factor is essential for quantitative image analysis.

Keywords: alpha-1 antitrypsin deficiency; emphysema; lung densitometry

Projected estimates of the global burden of chronic obstructive pulmonary disease have led to recognition of the need for the development of novel therapies to modify disease progression. In routine clinical practice, physiologic measures such as diffusion capacity and airflow obstruction are traditionally employed as surrogate measures of emphysema severity. However, practical difficulties arising from the limitations of these conventional methods necessitate the development of alternative outcome measures for use in clinical trials.

In contrast to chronic obstructive pulmonary disease, which is defined in physiologic terms, emphysema is defined pathologically. Nevertheless, it may now be diagnosed accurately and quantified noninvasively by computed tomography (CT). The loss of lung tissue associated with emphysema can be measured as a reduction in lung density by CT, using various parameters derived from the frequency distribution histogram of lung voxels. The “voxel index” (VI) method calculates the proportion of low-attenuation lung voxels within a desired range, and although various voxel index thresholds have been employed and validated against pathology, it is likely that no single threshold is appropriate universally.

The fundamental importance of calibration in any measuring device is indisputable, yet few studies have addressed the issue of CT calibration in lung densitometry. Short-term reproducibility has been demonstrated (8–10), but there are limited published data on long-term reproducibility (11). It is recognized that X-ray tube ageing and replacement may introduce data acquisition errors (11), but although accurate scanner calibration is mandatory, procedures of standardization have yet to be defined. Previous studies have used water phantoms alone for quality assurance (4, 8), but in a cross-sectional study to evaluate conformity between scanners, Kemerk and coworkers demonstrated that although water densitometry was consistent, the Hounsfield number for air varied between scanners. A method was proposed for correcting the CT numbers for a scanner with poor air calibration and its potential use as a means of standardizing densitometry was described (12). However, although standardization has been recognized to be crucially important (13), the influence of air calibration errors on lung densitometry studies in a clinical setting has yet to be addressed systematically. Such information is crucial to determine the validity and hence interpretation of observational and interventional studies.

The provisional assessment of CT progression in a sample of patients attending our program had unexpectedly suggested an increase in lung density (consistent with resolution of emphysema) in a limited number of serial annual scans acquired in the early months of 2000 (L. Dowson, personal communication). The current study was therefore undertaken to review the consistency of CT densitometry for the measurement of emphysema progression in a cohort of patients with alpha-1 antitrypsin deficiency (AATD) and to standardize densitometry measurements by adjusting for any identifiable scanner errors. Validation of the methodology was performed in studies with an anthropomorphic phantom. In addition, the opportunity was taken to review data from the Dutch–Danish controlled trial relating to scanner assurance (4, 8), but in a cross-sectional study to evaluate conformity between scanners, Kemerk and coworkers (12) demonstrated that although water densitometry was consistent, the Hounsfield number for air varied between scanners. A method was proposed for correcting the CT numbers for a scanner with poor air calibration and its potential use as a means of standardizing densitometry was described (12). However, although standardization has been recognized to be crucially important (13), the influence of air calibration errors on lung densitometry studies in a clinical setting has yet to be addressed systematically. Such information is crucial to determine the validity and hence interpretation of observational and interventional studies.

The provisional assessment of CT progression in a sample of patients attending our program had unexpectedly suggested an increase in lung density (consistent with resolution of emphysema) in a limited number of serial annual scans acquired in the early months of 2000 (L. Dowson, personal communication). The current study was therefore undertaken to review the consistency of CT densitometry for the measurement of emphysema progression in a cohort of patients with alpha-1 antitrypsin deficiency (AATD) and to standardize densitometry measurements by adjusting for any identifiable scanner errors. Validation of the methodology was performed in studies with an anthropomorphic phantom. In addition, the opportunity was taken to review data from the Dutch–Danish controlled trial (14) relating to scanner assurance (4, 8), but in a cross-sectional study to evaluate conformity between scanners, Kemerk and coworkers (12) demonstrated that although water densitometry was consistent, the Hounsfield number for air varied between scanners. A method was proposed for correcting the CT numbers for a scanner with poor air calibration and its potential use as a means of standardizing densitometry was described (12). However, although standardization has been recognized to be crucially important (13), the influence of air calibration errors on lung densitometry studies in a clinical setting has yet to be addressed systematically. Such information is crucial to determine the validity and hence interpretation of observational and interventional studies.

METHODS

Data records of 57 patients with AATD (PiZ) were selected from among those attending our program (see the online supplement) to include all subjects who had completed three consecutive annual assessments over 2 years spanning the months of concern in early 2000 (see INTRODUCTION). Alpha-1 antitrypsin level and phenotype were verified by immunoassay and isoelectric focusing, respectively, using a dried finger-prick blood spot (Heredilab, Salt Lake City, UT). The program was approved by the local research ethics committee, and all subjects gave written informed consent.
Annual CT images were acquired on a General Electric Prospeed scanner (General Electric Medical Systems, Milwaukee, WI), using a high-resolution protocol (120 kVp [kilovolt peak], 200 mA·s [dose rate], 1-mm collimation, bone reconstruction algorithm) on inspiration and expiration in the supine position as described previously (16). The voxel indices (VI) (−910 Hounsfield units [HU]) were calculated for upper (level of the aortic arch) and lower (level of the inferior pulmonary veins) zone images, using the manufacturer’s Density Mask program.

The methodology for the current study comprised two steps before the analysis of patient data with an additional phantom study to validate the technique employed for the patient data.

1. Error identification. Consistency of lung densitometry was explored by plotting annual VI change against scan date. Air measurements acquired retrospectively from patient images, using region of interest (ROI) densitometry plus weekly quality assurance records of water and bone phantom measurements, were related to date. Both data sets were assessed for discontinuity to identify any coincident events.

2. Development of a correction method. Air, water, and bone density values were compared to characterize the behavior of the scanner over time across a wide range of density values. A correction equation was developed to adjust lung densitometry for air calibration errors by predicting the potential influence of these three measures on the VI threshold.

Validation Studies Using a Lung Phantom

Correction technique for longitudinal data. An anthropomorphic lung phantom was scanned, using the same protocol as in the clinical study, using a Marconi MxView scanner (Philips Medical Systems, Andover, MA) with the facility to adjust the CT number for air. Lung densitometry was calculated for a single slice, using Pulmo-CMS software (MEDIS [Medical Imaging Systems], Leiden, The Netherlands) at different calibration settings without the threshold adjustment and applying the correction equation to the data to determine their efficacy.

Correction technique for cross-sectional multicenter data. Calibration and conformity of six third-generation multislice scanners (two General Electric Lightspeed scanners, one Philips MxView, one Philips MX 8000, one Toshiba Asteon, and one Siemens Sensation 16) were explored, using the dog phantom. The scanning protocol employed was that proposed for future clinical densitometric studies (140 kVp; 40 mA; 5-mm collimation; pitch, 1.5; 2.5-mm increment; field of view, 500 mm; and a smooth reconstruction filter) (17). Air calibration measurements were acquired as described above and surrogate measures of blood density were derived from ROI measurements of the phantom heart.

Whole lung VI of −910 HU was calculated for each series with and without the use of an internal calibration procedure and the effect on standardization was assessed by the coefficient of variation.

Analysis of Clinical Images: Employing the Validated Correction Technique

Upper-zone inspiratory images obtained previously were reanalyzed, using the “adjusted” threshold of −910 HU. Comparison was made of the annual VI progression rate, using uncorrected and adjusted thresholds for three groups of consecutive annual scans: (1) scans predating the discontinuity, (2) scans spanning the discontinuity, and (3) scans following the discontinuity.

In addition, the VI was calculated for corrected images, using 20-HU threshold increments from −870 to −950 HU and comparison of progression over 2 years was performed.

Statistical Analysis

Data were analyzed with a statistical software package (Statistical Package for the Social Sciences [SPSS] version 10.0.5; SPSS, Chicago, IL). Air calibration was assessed with discontinuity analysis and regression slopes were compared with the F test. The rate of CT progression in the three subgroups of scan pairs was compared by Kruskal–Wallis test and in the two subgroups of scan pairs, before and after the discontinuity, by Mann–Whitney test. Comparison of the baseline and Year 2 scans at the different thresholds was done with a Wilcoxon signed rank test for paired data. Differences were considered significant at p < 0.05.

RESULTS

Baseline Characteristics

The baseline lung function and voxel indices (uncorrected threshold, −910 HU; expressed as percentage) at baseline are shown in Table 1.

Error Identification–Discontinuity Analysis

A plot of annual VI change against scan date showed a trend occurring in March 2000 of reversed VI progression, implying disease improvement (Figure 1A). The best fitting model for voxel index progression placed this discontinuity (see the online supplement) just before March 6, 2000 and produced an estimated shift of 4.5% (p < 0.001). Quality assurance records of scanner calibration showed that measurements using the manufacturer’s water phantom were well within the manufacturer’s tolerance of 0 ± 10 HU, but a concomitant shift in air density measured from patient images was observed (Figure 2A). Regression analysis of the air calibration data indicated discontinuity with a positive shift in air number that occurred between the observation on March 6 and the following observation on March 11, 2000 (p < 0.001) (Figure 2A). In addition, an opposing drift of decreasing air number also existed throughout the study period, described by the gradient b in the equation y = a + bx + cx (see the online supplement). There was no difference between the gradient of drift before and after the discontinuity (p = 0.77) and, after elimination of the discontinuity, the overall gradient drift was significantly different from zero (p < 0.001) (Figure 2B).

Measurement of blood density in the descending thoracic aorta did not identify any time-related trend (data not shown).

Dutch–Danish Study

A similar drift but no obvious discontinuity was seen by analyzing ROI air measurements in the Dutch patients included in the Dutch–Danish study of augmentation therapy (14). The individual patient data are shown in Figure 3.

### Table 1. Baseline Characteristics

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Absolute Value*</th>
<th>% Predicted*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, yr</td>
<td>57</td>
<td>50 (46–61)</td>
<td>NA</td>
</tr>
<tr>
<td>FEV₁, L</td>
<td>57</td>
<td>1.34 (0.97–2.08)</td>
<td>42 (30–67)</td>
</tr>
<tr>
<td>VC/ L</td>
<td>57</td>
<td>3.82 (3.12–4.49)</td>
<td>98 (90–127)</td>
</tr>
<tr>
<td>RV/L</td>
<td>57</td>
<td>2.66 (2.06–3.44)</td>
<td>135 (100–170)</td>
</tr>
<tr>
<td>Dlco, mmol/min/kPa</td>
<td>57</td>
<td>5.90 (4.62–7.64)</td>
<td>63 (47–84)</td>
</tr>
<tr>
<td>Dlco/VA, mmol/min/kPa/L</td>
<td>57</td>
<td>1.03 (0.73–1.31)</td>
<td>46 (67–83)</td>
</tr>
<tr>
<td>UZI voxel index, −910 HU</td>
<td>57</td>
<td>34.2 (21.6–44.6)</td>
<td>NA</td>
</tr>
<tr>
<td>UZE voxel index, −910 HU</td>
<td>57</td>
<td>19.0 (7.7–34.2)</td>
<td>NA</td>
</tr>
<tr>
<td>LZE voxel index, −910 HU</td>
<td>57</td>
<td>55.3 (33.2–64.5)</td>
<td>NA</td>
</tr>
<tr>
<td>LZE voxel index, −910 HU</td>
<td>57</td>
<td>47.0 (26.4–57)</td>
<td>NA</td>
</tr>
</tbody>
</table>

Definition of abbreviations: Dlco = single-breath diffusing capacity of the lung for carbon monoxide (gas transfer); Dlco/VA = Dlco normalized per liter alveolar volume (transfer coefficient); HU = Hounsfield unit; LZE = lower-zone expiratory high-resolution computed tomography (HRCT) image; UZI = lower-zone inspiratory HRCT image; VA = not applicable; RV = residual volume; UZE = upper-zone expiratory HRCT image; UZI = upper-zone inspiratory HRCT image.

* Median (interquartile range).

† Postdual bronchodilation with nebulized salbutamol (5 mg) and ipratropium bromide (500 μg).
Development of a Correction Method

Changes in air, water, and bone density values indicated that the effect of an increase in air number would be to increase proportionally the Hounsfield number of those voxels with a density intermediate between water and air (see Figure 4A). Therefore, correction for air calibration errors on VI threshold could be achieved with the following formula: corrected threshold = (air calibration value / water calibration value / 1,000) × index threshold (see Figure 4B).

Validation Studies with Lung Phantom

Correction technique for longitudinal data. The distribution of voxels with the dog lung phantom was similar to that seen in patients with AATD (data not shown), suggesting that it was a good model for emphysematous lung. Adjusting the scanner air calibration (see Methods) resulted in a proportional shift in the lung voxel distribution histogram and marked changes to the voxel index (−910 HU) as predicted from the studies described above. Application of the correction technique reproduced the same results at each calibration setting (see Table 2).

Correction technique for cross-sectional multicenter data. Interscanner variability for air densitometry was greater than for blood densitometry and there was wide discrepancy in the measurement of dog lung voxel indices (VI, −910 HU). There was a significant reduction in this discrepancy by standardizing for air calibration, using the method described here and used in the main study, but the use of blood calibration did not reduce variability (see Table 3).

CT Progression: Application of the Correction Method

The observed reversal in annual voxel index progression that occurred in the patient cohort was eliminated on use of the correction technique (see Figure 1B). For the uncorrected data, the progression rate for scan pairs spanning the discontinuity was significantly different (p < 0.001) from that of the other two groups, in which the scan pair either predated or postdated the discontinuity (see Figure 5A) and suggests an overall improvement over this time period. This difference was abolished when the corrected data for the three groups were compared (p = 0.13) (see Figure 5B). However, there was no significant difference in VI progression rate for scan pairs obtained either before or after the discontinuity in both the uncorrected (p = 0.92) and corrected data (p = 0.72) (see Figures 5A and 5B).
Figure 3. Air calibration data for the Philips Tomoscan SR7000 scanner used in the Dutch–Danish trial of alpha-1 antitrypsin augmentation therapy (14). ROI measurements were acquired from patient images as described for the current study (see Methods and online supplement). Air density values (in Hounsfield units) are expressed as the difference from baseline values and results from individual patients are shown by the joined lines.

Adjusting the analysis threshold for air calibration also improved the detection of progression over 2 years as shown by a comparison of Z statistic (see Figure 6). Progression of the voxel index at the uncorrected threshold of −910 HU was statistically significant only for the upper-zone inspiratory scans (see Figure 6A) as shown previously (8). However, after adjustment with correction for air calibration, significant progression at the threshold of −910 HU was confirmed in both upper- and lower-zone scans (see Figure 6B). The annual VI progression rates at a corrected threshold of −910 HU are shown in Table 4, expressed as median and interquartile range, with data from the previous study (8) for comparison.

The upper-zone voxel index showed significant progression at all thresholds used, with marginal statistical superiority of the inspiratory images over the expiratory images (Figure 6A). Furthermore, in the lower-zone scans the corrected expiratory images now also showed significant progression whatever the threshold and the corrected inspiratory images demonstrated significant progression at thresholds of −950, −930, and −910 HU (Figure 6B). Although all thresholds showed highly significant progression of VI, the Z statistic was best for −950 HU,

![Figure 4. Plot of actual CT numbers for air, water phantom, and bone phantom on a single occasion, demonstrating correct water densitometry but incorrect values for air and bone. Perfect calibration values are shown for comparison. (A) All data points; (B) the influence on low-density values, assuming linearity between air and water. The effect of a shift in air calibration of +30 to −970 HU on an intended VI threshold of −910 HU is shown by the arrow, demonstrating how this adjusted threshold can be calculated, assuming a water value of zero, using the following formula: corrected threshold = (air calibration value − water calibration value/−1,000) × index threshold.
suggesting it is the most appropriate for assessing progression among the patients studied here.

**DISCUSSION**

The current study demonstrates the importance of precise scanner and image calibration in longitudinal lung densitometry studies for the measurement of emphysema progression. We found

Table 3. Mean, Standard Deviation, and Coefficient of Variation for Blood and Air Density Values Measured on Six Different Scanners

<table>
<thead>
<tr>
<th></th>
<th>Blood Density (HU)</th>
<th>Air Density (HU)</th>
<th>No Calibration</th>
<th>Blood Calibration</th>
<th>Air Calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>41.09</td>
<td>-1,005.1</td>
<td>36.23</td>
<td>35.56</td>
<td>34.02</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>5.36</td>
<td>13.73</td>
<td>8.95</td>
<td>9.27</td>
<td>3.33</td>
</tr>
<tr>
<td>Coefficient of variation</td>
<td>13.04</td>
<td>1.37</td>
<td>24.70</td>
<td>26.06</td>
<td>9.79</td>
</tr>
</tbody>
</table>

Voxel index values (VI, −910 HU) for whole dog lung are shown, demonstrating the influence of standardization on scanner variability by calibrating on the basis of either blood or air densitometry. The latter resulted in an almost threefold reduction in variability.

Figure 5. Annual progression in upper-zone inspiratory voxel index (threshold, −910 HU) (UZI VI) before (A) and after (B) adjustment for air calibration. Data are shown as the median and interquartile range. P values relate to the Kruskal–Wallis test comparing VI progression rate in the scan pairs spanning the discontinuity with the other two groups and to the Mann–Whitney U test comparing scan pairs before and after the discontinuity. For the unadjusted data (A) there is a significant difference in the rate of VI progression between the scan pairs spanning the discontinuity and the two other groups and to the Mann–Whitney U test comparing scan pairs before and after the discontinuity (p < 0.001). No significant difference between scan pairs before and after the discontinuity (p = 0.92). (B) After adjustment for air calibration there is no significant difference between the scan pairs spanning the discontinuity and the two other groups (p = 0.13).
emphysema progression. Our results confirm that our previous finding of highest VI progression rate in the upper-zone inspiratory images (8) remains valid after adjustment of scanner calibration errors. Furthermore, adjustment for air calibration resulted in improved variability and the detection of statistically significant progression in all images, even at the threshold of −910 HU used in our original study (8). Consequently, this correction method has now been incorporated in the analysis software being used for our clinical trials and includes calibration of the entire image data before lung segmentation and histogram analysis (18).

Adjustment of lung densitometry to correct for errors in air calibration provides a mechanism to improve scanner consistency in longitudinal studies and scanner comparability in multicenter studies as validated in the phantom studies. The voxel distribution histograms for the dog lung phantom compared well with those from patients with emphysema, confirming that the phantom was an appropriate model to validate the correction method. Although the scanning protocols differed between clinical and phantom studies, it was necessary in the multicenter phantom study to use a whole lung volume scanning protocol and whole lung densitometry to reduce sampling errors.

The multiscanner phantom study confirms the findings of Kemerink and coworkers (12) that consistency of water calibration between scanners coexists with great variability in air calibration. It is not surprising, as the density of blood is close to that of water, that interscanner variability in lung densitometry is not improved by an internal calibration method that incorporates blood density values alone whereas variability is improved when adjusting for differences in air calibration. However, the application of blood calibration may be of greater importance for lung densitometry studies of conditions such as pulmonary fibrosis, which produce changes in tissues of greater density than those affected by the emphysematous process (19).

Good correlation has been reported between quantitative CT and other measures of emphysema both for the voxel index method (6, 7, 20–24) and the percentile method (22). Gevenois and coworkers have shown that the optimal VI threshold (using a high-resolution protocol) is −950 HU in cross-sectional studies comparing CT with both macroscopic (20) and microscopic morphometry (6). In the current study, several VI threshold values were used because it was considered that sensitivity to the changes of emphysema progression would be threshold dependent and, although the correction method was expected to improve variability, it was recognized that the effect could also be to change the average VI threshold to one less sensitive than −910 HU. There are limited data on the measurement of em-
expiratory; UZI

Comparisons are made between densitometry performed on different centers, further studies are needed. Details of calibration methods are essential to the interpretation of lung densitometry, and quality assurance procedures for densitometry studies should include the technique of air calibration described here. These data must be reported in longitudinal studies of disease progression and when direct comparison is made between densitometry performed on different scanners to ensure validity.

TABLE 4. ANNUAL PROGRESSION RATES OF SINGLE IMAGE VOXEL INDEX AT A THRESHOLD OF \(-910\) HU, ADJUSTED FOR CHANGES IN AIR CALIBRATION WITH DATA PREVIOUSLY REPORTED AND INCLUDED FOR COMPARISON

<table>
<thead>
<tr>
<th>Image*</th>
<th>Median Adjusted Rate (IQR)*</th>
<th>Mean Annual Rate Reported Previously (SE)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>UZI</td>
<td>1.67 (-.0.23 to 3.28)</td>
<td>2.82 (0.55)</td>
</tr>
<tr>
<td>LZI</td>
<td>0.61 (-.1.63 to 2.72)</td>
<td>1.03 (0.56)</td>
</tr>
<tr>
<td>UZE</td>
<td>0.58 (-.50 to 2.98)</td>
<td>1.67 (0.38)</td>
</tr>
<tr>
<td>LZE</td>
<td>1.17 (-.29 to 3.85)</td>
<td>1.66 (0.77)</td>
</tr>
</tbody>
</table>

Definition of abbreviations: IQR = interquartile range; LZI = lower-zone inspiratory; LZI = lower-zone expiratory; UZI = upper-zone inspiratory.

* See Reference 8.

† n = 57.

‡ n = 43.

Conflict of Interest Statement: D.G.P. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript; B.C.S. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript; J.S. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript; P.G.N. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript; R.A.S. has lectured widely for non-promotional purposes for GlaxoSmithKline, Bayer, and Eli Lilly and acts on Advisory Boards for AstraZeneca (AZ) and Baxter and has two non-commercial research grants—one from AZ (£125,000 since 2002) and one from Bayer (£500,000 per annum since 1996).

Acknowledgment: The authors thank KCARE (Kings College Hospital, London) for the loan of the lung phantom and Derek Tarrant for technical support with the use of the Marconi MxView scanner. R. A. Stockley and J. Stolk are members of AIR (the Alpha-1 International Registry, www.aatregistry.org) and thank the other council members of AIR for valuable discussions on the results of this study.

References

4. Dirksen A, Friis M, Olesen KP, Skovgaard LT, Sorensen K. Progression of emphysema in severe \(\alpha1\)-antitrypsin deficiency as assessed by annual CT. Acta Radiol 1997;38:826–832.


