Inter- and intra-software reproducibility of computed tomography lung density measurements

Miranda Kirby[a)]
Department of Physics, Ryerson University, Toronto, ON, Canada

Charles Hatt
IMBIO, Minneapolis, MN, USA
Deparment of Radiology, University of Michigan, Ann Arbor, MI, USA

Nancy Obuchowski
Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA

Stephen M. Humphries
Department of Radiology, National Jewish Health, Denver, CO, USA

Jered Sieren
VIDA Diagnostics Inc., Coralville, IA, USA

David A. Lynch
Department of Radiology, National Jewish Health, Denver, CO, USA

Sean B. Fain and on behalf of the QIBA Lung Density Committee*
Deparment of Medical Physics, University of Wisconsin, Madison, WI, USA

(Received 5 November 2019; revised 2 March 2020; accepted for publication 2 March 2020;
published 31 March 2020)

Purpose: Multiple commercial, open-source, and academic software tools exist for objective quantification of lung density in computed tomography (CT) images. The purpose of this study was to evaluate the intersoftware reproducibility of CT lung density measurements.
Methods: Computed tomography images from 50 participants from the COPDGene^{TM} cohort study were randomly selected for analysis; n = 10 participants across each global initiative for chronic obstructive lung disease (GOLD) grade (GOLD 0–IV). Academic-based groups (n = 4) and commercial vendors (n = 4) participated anonymously to generate CT lung density measurements using their software tools. Computed tomography total lung volume (TLV), percentage of the low attenuation areas in the lung with Hounsfield unit (HU) values below −950HU (LAA_{950}), and the HU value corresponding to the 15th percentile on the parenchymal density histogram (Perc15) were included in the analysis. The intersoftware bias and reproducibility coefficient (RDC) was generated with and without quality assurance (QA) for manual correction of the lung segmentation; intrasoftware bias and RDC was also generated by repeated measurements on the same images.
Results: Intersoftware mean bias was within ±0.22 mL, ±0.46%, and ±0.97 HU for TLV, LAA_{950} and Perc15, respectively. The RDC was 0.35 L, 1.2% and 1.8 HU for TLV, LAA_{950} and Perc15, respectively. Intersoftware RDC remained unchanged following QA: 0.35 L, 1.2% and 1.8 HU for TLV, LAA_{950} and Perc15, respectively. All software investigated had an intrasoftware RDC of 0. The RDC was comparable for TLV, LAA_{950} and Perc15 measurements, respectively, for academic-based groups/commercial vendor-based software tools: 0.39 L/0.32 L, 1.2%/1.2%, and 1.7 HU/1.6 HU. Multivariable regression analysis showed that academic-based software tools had greater within-subject standard deviation of TLV than commercial vendors, but no significant differences between academic and commercial groups were found for LAA_{950} or Perc15 measurements.
Conclusions: Computed tomography total lung volume and lung density measurement bias and reproducibility was reported across eight different software tools. Bias was negligible across vendors, reproducibility was comparable for software tools generated by academic-based groups and commercial vendors, and segmentation QA had negligible impact on measurement variability between software tools. In summary, results from this study report the amount of additional measurement variability that should be accounted for when using different software tools to measure lung density longitudinally with well-standardized image acquisition protocols. However, intrasoftware reproducibility was deterministic for all cases so use of the same software tool to reduce variability for serial studies is highly recommended. © *2020 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.14130]

Key words: computed tomography, COPD, emphysema, imaging biomarker, lung density

1. INTRODUCTION

Computed tomography (CT) lung density is an imaging biomarker used to objectively and noninvasively quantify the extent of emphysema in the lung. Over the last three decades, numerous studies in patients with chronic obstructive pulmonary disease (COPD) have demonstrated that CT lung density measurements are correlated with emphysema measured in excised lungs by histology,[2,3] are associated with mortality[5] and exacerbations,[6] and can identify subgroups of patients with better responses following lung-volume-reduction surgery[7] and endobronchial valve implantation.[8] Furthermore, in patients with alpha 1-antitrysin deficiency, a significant response to augmentation therapy was shown using CT lung density as a surrogate of emphysema, but not with conventional spirometry measurements.[9] These findings all highlight the potential role of quantitative CT for COPD patient management, such as longitudinal monitoring of disease progression and assessing treatment response.

Maintaining standardized image acquisition parameters, however, is critically important for serial assessments that aim to quantify CT lung density. It is well-established that there are technical challenges for generating reproducible CT measurements. Submaximal inspiration breath-hold volume,[10] dose[11,12] as well as image reconstruction parameters, including slice thickness[13,14] and reconstruction kernel,[15–17] have all been shown to impact CT measurements. However, several large, multicenter, longitudinal cohort studies, such as COPDGene^{TM},[18] have utilized breath-hold coaching and dedicated lung phantoms to standardize image acquisition and reconstruction parameters across all sites to minimize variability introduced by image acquisition related parameters.

Another factor that has the potential to impact the reproducibility of CT measurements is the specific software used to generate the measurements. Lung density measurements are derived from the parenchymal density histogram of CT Hounsfield unit (HU) values and thus are deterministic computations and are directly computed given an accurate lung segmentation mask.[1–4] However, measurement variability may be introduced by differences in the thoracic cavity segmentation, as well as segmentation of the large airways and pulmonary vessels, even when consistent image acquisition and reconstruction settings are utilized. Previous studies investigating the influence of different software tools have shown conflicting results, and in some studies high intersoftware variation for CT lung density measurements have been reported.[19–21]

In an effort to standardize methodology, the Lung Density Committee of the quantitative imaging biomarker alliance (QIBA) has released for public comment a profile regarding the CT lung density measurement.[22] Given the multitude of software tools used by different commercial, open-source, and academic research laboratories, an evaluation of the intersoftware variability of CT lung density measurements is warranted to support this profile, particularly in the context of serial investigations. Furthermore, quantifying intersoftware CT measurement reproducibility requires a cohort with minimal variability introduced by image acquisition parameters. Therefore, here our objective was to investigate and report CT lung volume and lung density measurement intersoftware bias and reproducibility using CT images from the COPDGene^{TM} cohort study, with various academic groups and commercial vendors participating in the reproducibility study.

2. MATERIALS AND METHODS

2.A. Details of the software comparison

Computed tomography images from 50 participants from the COPDGene^{TM} cohort study[18] were selected for analysis; n = 10 participants across each COPD GOLD grade (GOLD 0–IV) were randomly selected. Participation was solicited from academic groups and commercial vendors, and the solicitation letter indicated that the results would be anonymized (ie the software packages were provided on the condition they would not be individually identified). The anonymization was performed by The Radiological Society of North America (RSNA) that acted as a neutral broker between all participating groups and the QIBA Lung Density committee, to ensure that the committee was blinded to the participants' identity. The CT datasets used in this study are accessible in the quantitative imaging data warehouse (QIDW): https://qidw.rsna.org/.

All vendors indicated if their software tool was for academic use only or commercial. Vendors were instructed to generate measurements: (a) without segmentation quality assurance (QA) or manual correction to evaluate intersoftware reproducibility; (b) a repeated set of measurements on the same images, to evaluate intrasoftware reproducibility; and, (c) a third set of measurements repeated on the same images following segmentation QA and manual correction.

2.B. CT image acquisition

Computed tomography images were acquired using CT systems of various makes and models, including GE, Siemens and Philips models, with the participant supine at suspended full-inspiration from apex to base of the lung as previously described.[18] In general, CT images were reconstructed with smooth convolution kernels (Siemens B31f, GE STANDARD, or Philips B) and slice thicknesses and intervals between 0.625 and 0.75 mm. The full-dose protocol used an effective dose of 200 mAs without dose modulation. A more detailed description of the CT image acquisition protocol is described elsewhere.[18]

2.C. CT image analysis

Computed tomography images were processed using academic and commercial CT lung density software. All groups were instructed to generate CT measurements for each image dataset using none or a minimal amount of manual software interaction. We also requested no image auto-calibration or preprocessing (eg noise reduction filtering). All vendors were asked to perform the following steps for lung segmentation:

1. Segmentation of the lung parenchyma from the rest of the thoracic cavity;
2. Removal of airways from the segmentation (no strict definition of which airways were required to be removed was provided, but the software was required to at least remove the trachea and major bronchi from the air-space prior to computing the CT lung density metrics);
3. Blood vessel removal (no instruction was provided on the amount of acceptable blood vessel exclusion from the lung volume).

Next, groups were instructed to repeat each of these steps on the same image dataset in order to assess the intrasoftware repeatability. Finally, the vendors were asked to perform QA by reviewing and manually correcting any lung segmentation errors to generate a third set of CT measurements using the corrected segmented lung volume.

The measurements generated include: the total lung volume (TLV), percentage of the low attenuation areas in the lung with HU values below −950 (LAA_{950}),[1-3] and the HU unit value corresponding to the 15th percentile on the parenchymal density histogram (Perc15).[4]

2.D. Statistical analysis

All statistical analysis was performed using SAS 9.4 software (Cary, NC, USA) and MATLAB R2018a (Natick, MA, USA). A one-way analysis of variance (ANOVA) with a Tukey test for multiple comparison correction was performed for statistical comparison between GOLD groups for age; for sex and race, a Fisher's Exact test was used. MATLAB was used for Bland-Altman analysis to compare measurements generated by each possible pair of software tools; measurements include TLV, LAA_{950} and Perc15 without QA. The reproducibility coefficient (RDC)[23] was calculated for each software tool, as described below, to compare between the different software tools for each lung measurement with and without QA, and by group type (academic-based, commercial). The RDC is the value under which the difference between repeated measurements on the same participant acquired under different conditions (ie. different software tools) should fall within 95% probability. To estimate the RDC for any given software tool, we must estimate the variance relative to the other K − 1 software tools in the comparison (K = 8 in our study). Therefore, for a specific software tool, l, we calculated the mean variance, σ_l^2, for the measurements, subscript i, across the 50 image sets, where $M_{i,l}$ represents measurement i of software l and $\sigma_{i,k,l}^2$ represents the variance between software l and software k for measurement i:

$$\sigma_{i,k,l}^2 = \frac{1}{2}\left(M_{i,k} - M_{i,l}\right)^2$$

Next, $\sigma_{k,l}^2$ represents the variance between software k and software l averaged over all measurements N:

$$\sigma_{k,l}^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_{i,k,l}^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(M_{i,k} - M_{i,l}\right)^2 = \frac{1}{2N}\sum_{i=1}^{N}\left(M_{i,k} - M_{i,l}\right)^2$$

Then, the average variance over the other $K - 1$ software tools is calculated to generate the average variance for software l:

$$\sigma_l^2 = \frac{1}{K-1}\sum_{k=1}^{K-1}\sigma_{k,l}^2$$

The average RDC for software l is then given by:

$$RDC_l = 1.96 * \sqrt{2\sigma_l^2}$$

Low RDC values indicate high reproducibility between software tools. The 95% confidence intervals for the RDC were constructed using bootstrapping with 5000 resamples.

Multivariable linear regression models were built to assess whether group type (academic-based, commercial) was a predictor of the within-subject standard deviation of TLV, LAA_{950} and Perc15 measurements. If group type was found to be a significant predictor, it would indicate that the standard deviation between software tool measurements is different for commercial vendors and academic groups; in other words, it would indicate that CT measurements are more similar between commercial vendors or academic groups. Generalized estimating equations (GEEs) were used to account for the clustered nature of the data.

3. RESULTS

A total of 50 participants were investigated: n = 10 in each GOLD grade. As shown in Table I, there were no differences between the groups for age, sex or race. A total of nine software tools participated in the study; software tools 1–4 were from academic-based groups and software tools 5–9 were from commercial vendors. A single commercial vendor withdrew from the study and therefore a total of eight software tools, n = 4 research-based and n = 4 commercial, were included in the analysis. All eight software tools were able to generate measurements for all images provided. A total of three of eight software tools reported some manual editing of the segmentation masks for some of the CT images as part of the QA step.

Figure 1 shows an example of the CT lung volume (in blue) and LAA_{950} segmentation masks (in red) for two different software tools. The differences observed for exclusion of airways and vessels from the lung volume segmentation mask between the two software tools are subtle and representative of the type of differences that would be expected given

TABLE I. Subject demographics.

Parameter[a]	GOLD 0 (n = 10)	GOLD I (n = 10)	GOLD II (n = 10)	GOLD III (n = 10)	GOLD IV (n = 10)
Age, yr	68 (8)	69 (9)	63 (10)	68 (9)	62 (6)
Female sex, n (%)	4 (40)	4 (40)	4 (40)	3 (30)	5 (50)
Race, n (%)					
Non-hispanic white	10 (100)	9 (90)	8 (80)	8 (80)	6 (60)
African American	0 (0)	1 (10)	2 (20)	2 (20)	4 (40)

[a] All parameter values are mean (±SD) unless otherwise noted.

acceptable segmentation quality for both images (i.e. no major segmentation errors).

3.A. Bland-Altman analysis

Bland-Altman analysis was performed for TLV, LAA_{950}, and Perc15 measurements for each software tool compared with all other software tools. Table II provides the summary of the Bland-Altman analysis for measurements generated by each software tool with the average of all the other software tools for TLV, LAA_{950}, and Perc15 measurements. There was negligible bias for all software tools to within ±0.22 L, ±0.46%, and ±0.97 HU, for TLV, LAA_{950}, and Perc15 respectively.

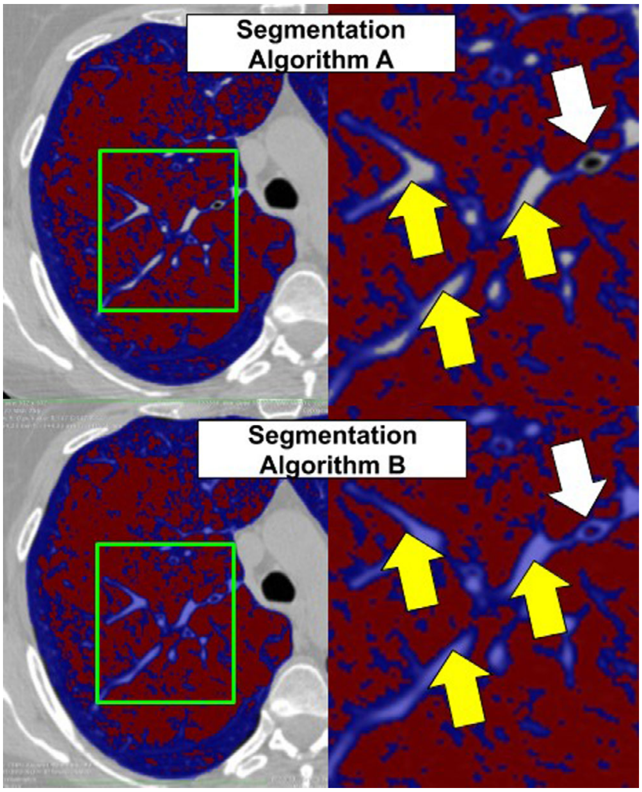

FIG. 1. Computed tomography (CT) lung and emphysema segmentation generated by two different software tools. Shown above are two examples of CT lung segmentation images from two different software tools. Areas of the lung greater than or equal to − 950 HU are colored in blue, areas <−950 HU are colored in red. Differences in the inclusion of blood vessels (yellow arrows) and airways (white arrows) can impact lung volume and low-attenuation area calculations. Note that the CT slice in this figure was the slice with the largest disagreement in segmentation volume over the entire image series. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE II. Bland-Altman analysis for each software compared to the average of all other software tools.

	Mean bias	Median bias	SD of bias	Upper 95% CI	Lower 95% CI
TLV (L)					
Software 1	−0.15	−0.14	0.05	−0.06	−0.25
Software 2	0.04	0.02	0.04	0.12	−0.04
Software 3	0.22	0.23	0.05	0.32	0.13
Software 4	0.05	0.04	0.02	0.09	0.01
Software 5	−0.04	−0.04	0.03	0.01	−0.09
Software 6	−0.01	−0.01	0.02	0.03	−0.06
Software 7	−0.21	−0.19	0.05	−0.10	−0.31
Software 8	0.10	0.09	0.04	0.18	0.02
LAA_{950} (%)					
Software 1	0.33	0.22	0.37	1.05	−0.40
Software 2	−0.24	−0.18	0.28	0.31	−0.80
Software 3	−0.29	−0.14	0.34	0.37	−0.95
Software 4	−0.42	−0.39	0.29	0.15	−0.98
Software 5	−0.34	−0.34	0.19	0.03	−0.71
Software 6	0.42	0.39	0.20	0.82	0.02
Software 7	0.46	0.26	0.49	1.42	−0.50
Software 8	0.09	0.10	0.12	0.32	−0.15
Perc15 (HU)					
Software 1	−0.33	−0.39	0.47	0.58	−1.24
Software 2	0.20	0.14	0.36	0.90	−0.51
Software 3	0.97	0.95	0.46	1.87	0.07
Software 4	0.49	0.54	0.39	1.25	−0.27
Software 5	0.24	0.17	0.36	0.95	−0.47
Software 6	−0.88	−0.80	0.27	−0.35	−1.40
Software 7	−0.58	−0.57	0.57	0.54	−1.70
Software 8	−0.11	−0.18	0.32	0.52	−0.74

TLV, total lung volume; Perc15, 15th percentile on the parenchymal density histogram.

3.B. Reproducibility coefficients

Table III shows the RDC for TLV, LAA_{950}, and Perc15 measurements for eight different software tools with and without QA using manual correction of the lung volume segmentation. Overall, intersoftware RDC was 0.35 L, 1.2% and 1.8 HU for TLV, LAA_{950} and Perc15, respectively. Intersoftware RDC remained unchanged following QA: 0.35 L, 1.2% and 1.8 HU for TLV, LAA_{950} and Perc15 respectively. Intrasoftware RDC was generated by performing repeated measurements using the same software tool without QA; all software had an intrasoftware RDC of 0, indicating that image processing workflows were deterministic for all software tools.

Table III. The reproducibility coefficient (RDC) for total lung volume (TLV), LAA_{950} and 15th percentile on the parenchymal density histogram (Perc15) for all software tools with and without quality assurance (QA).

	Intersoftware RDC without QA		Intersoftware RDC with QA	
Parameter	RDC	95% CI	RDC	95% CI
TLV (L)				
Total	0.35	0.32–0.37	0.35	0.32–0.37
Software 1	0.38	0.35–0.41	0.38	0.35–0.42
Software 2	0.26	0.24–0.27	0.26	0.24–0.28
Software 3	0.26	0.24–0.29	0.26	0.24–0.29
Software 4	0.48	0.46–0.51	0.48	0.45–0.51
Software 5	0.25	0.23–0.27	0.25	0.23–0.27
Software 6	0.46	0.43–0.49	0.46	0.43–0.49
Software 7	0.31	0.28–0.34	0.31	0.28–0.34
Software 8	–	–	–	–
Software 9	0.26	0.24–0.28	0.26	0.24–0.28
LAA_{950} (%)				
Total	1.2	1.0–1.4	1.2	1.0–1.4
Software 1	1.2	1.0–1.5	1.2	1.0–1.5
Software 2	1.1	0.9–1.2	1.1	0.9–1.2
Software 3	1.1	0.9–1.2	1.1	0.9–1.2
Software 4	1.2	0.9–1.4	1.2	0.9–1.4
Software 5	1.2	1.0–1.3	1.2	1.0–1.3
Software 6	1.5	1.2–1.8	1.5	1.2–1.8
Software 7	0.9	0.7–1.0	0.9	0.7–1.0
Software 8	–	–	–	–
Software 9	1.2	1.0–1.4	1.2	1.0–1.4
Perc15 (HU)				
Total	1.8	1.6–2.0	1.8	1.6–2.1
Software 1	1.6	1.4–1.9	1.7	1.4–1.9
Software 2	1.5	1.3–1.7	1.6	1.3–1.8
Software 3	1.5	1.3–1.6	1.5	1.3–1.6
Software 4	2.3	2.1–2.6	2.3	2.1–2.6
Software 5	2.1	1.9–2.3	2.1	1.9–2.3
Software 6	2.0	1.6–2.3	2.0	1.6–2.4
Software 7	1.4	1.2–1.7	1.4	1.2–1.6
Software 8	–	–	–	–
Software 9	1.7	1.5–1.9	1.7	1.5–1.9

Table IV shows the RDC for TLV, LAA_{950}, and Perc15 measurements for software tools by group type (academic or commercial) with and without QA. Academic groups and commercial vendor's software tools generated comparable RDC measurements for TLV, LAA_{950}, and Perc15: 0.39 L/ 0.32 L, 1.2%/1.2%, and 1.7 HU/1.6 HU respectively. As shown in Table IV, QA had negligible impact on measurement reproducibility between software.

3.C. Multivariable linear regression models

Table V shows multivariable linear regression models for within-subject standard deviation of TLV, LAA_{950}, and Perc15 measurements generated by the different software tools with group type (academic, commercial) as a predictor. In the multivariable linear regression model for within-subject standard deviation of TLV, group type (academic = 1, commercial = 2) was a significant predictor ($P < 0.0001$); this indicates that academic vendors had greater within-subject standard deviation of TLV measurements than commercial vendors. However, group type was not a significant predictor for within-subject standard deviation in the multivariable linear regression model for LAA_{950} ($P = 0.46$) or Perc15 measurements ($P = 0.24$).

Table IV. The reproducibility coefficient (RDC) for total lung volume (TLV), LAA_{950} and 15th percentile on the parenchymal density histogram (Perc15) for academic-based and commercial software tools.

Parameter	Intersoftware RDC without QA	95% CI	Intersoftware RDC with QA	95% CI
TLV (L)				
Academic	0.39	0.36–0.41	0.39	0.36–0.41
Commercial	0.32	0.29–0.34	0.32	0.29–0.35
LAA_{950} (%)				
Academic	1.2	0.9–1.4	1.2	0.9–1.4
Commercial	1.2	1.0–1.3	1.1	1.0–1.3
Perc15 (HU)				
Academic	1.7	1.5–1.9	1.7	1.5–1.9
Commercial	1.6	1.3–1.9	1.6	1.3–2.0

Table V. Multivariable linear regression analysis for software tool type with standard deviation of total lung volume (TLV), LAA_{950} and 15th percentile on the parenchymal density histogram (Perc15).

	Estimate	Standard error	Significance of difference (P)
TLV [SD]	−0.03	0.004	<0.0001
LAA_{950} [SD]	−0.009	0.01	0.46
Perc15 [SD]	−0.04	0.03	0.24

Software type (academic = 1, commercial = 2).

4. DISCUSSION

There have been numerous clinical and research studies demonstrating that quantitative CT lung density measurements are related to important outcomes in COPD patients[5-8] and in patients with alpha 1-antitrysin deficiency.[9] Potential clinical applications include patient selection for treatment (eg by lung volume reduction surgery or endobronchial valves), or for evaluating treatment response over time. However, in order for CT lung density measurements to be used as a surrogate of emphysema in clinical applications, the variability of the CT measurements must be carefully controlled. Several large, multicenter, longitudinal cohort studies, including COPDGene,[18] SPIROMICS,[24] ECLIPSE,[25] MESA[26], and CanCOLD,[27] have implemented standardized image acquisition protocols to carefully control for known factors that impact CT measurements. However, the number of software tools developed by academic groups and commercial vendors to generate CT lung density measurements is increasing, with several well-established commercial and prototype software packages now available, and each has their own proprietary segmentation algorithms. For serial assessments or longitudinal evaluations where there is potential to change software tools at different time-points the reproducibility of CT measurements generated for various software tools must be evaluated.

In this study, we evaluated reproducibility for eight different software tools, including well-established software from both academic groups and commercial vendors. We evaluated participants without COPD and participants with a range of COPD severities. Our results indicate relatively high reproducibility across the different software tools for TLV, LAA_{950}, and Perc15 measurements. Although the Bland-Altman analysis and Fig. 1 indicate that there are clear differences for total lung volume segmentation between some of the vendors, which may result in the slight deviations observed in the Bland-Altman analysis for LAA_{950}, the bias overall was quite low and for LAA_{950} the bias was less than 1% between all vendors. This bias is much less than reported previously by Wielputz et al.[20] who investigated five software tools (two academic and three commercial) for lung density measurements in COPD. The more reproducible findings reported here may be related to several factors: the wider range of severity of the patients investigated (the patients evaluated by Wielputz et al.[20] were mainly end-stage COPD); the fact that a more standardized image acquisition protocol was used for COPDGene; or potentially improvements in image processing techniques over the last several years leading to more reproducible measurements between software tools.

In addition to assessing intersoftware agreement for CT measurements, we also generated RDC to determine how much variability may be introduced by using different software tools when repeated measurements are made on the same patient. Again, although the measurements generated by some software tools agreed slightly better than others, the RDC values were low, and overall the RDC between all software tools was only 1.2% for LAA_{950}. For example, this indicates that if the software tool was changed during a longitudinal study, whereby there were repeated measurements on the same patient but measurements were made using different software, the variability attributed to the software would be 1.2% for LAA_{950}. In other words, to detect real emphysema progression, the variability due to intersoftware reproducibility measured in this study is 1.2% for LAA_{950}. However, to determine the true overall RDC, the intersoftware reproducibility would need to be combined with expected test/retest measurement repeatability arising from differences in patient positioning, scanner model, scanner calibration, breath hold volumes, etc., and a detection of progression would need to be greater than the combined variability to be considered significant. Obuchowski et al.[28] has described the RDC calculations required to compute measurement reproducibility and repeatability. In general, however, we recommend that the same software be used for sequential measures during a longitudinal study, especially given that all methods showed deterministic intrasoftware reproducibility.

Intrasoftware reproducibility was evaluated by having all groups run their software tool on the same CT images a second time. The RDC for the intrasoftware comparison was zero. We also requested that each vendor run their software a third time and perform more rigorous QA. Although three of eight vendors reported that manual edits were required in some of the participants evaluated (eg lung volume edits or airway and vessel removal), the RDC did not change between the first run when there was no QA and the third run when QA was performed. This finding suggests that the results generated between the software tools were similar regardless of whether QA was performed. This may indicate that lung segmentation and airway and vessel removal algorithms generate similar results between vendors, before manual editing.

Finally, we investigated the RDC for CT measurements stratified by whether the software was developed by academic-based groups or commercial vendors. Although based on the RDC we found that the lung volume segmentation results tended to agree slightly better within commercial vendors than academic groups, the difference was very small and the RDC for LAA_{950} was 1.2% for both commercial and research vendors. This observation was consistent with the results of the multivariable linear regression analysis in which we investigated group type as a predictor of the standard deviation between the CT measurements generated by the different software tools. We found that commercial vendors had lower within-subject standard deviation of TLV than academic groups, but no difference was found for LAA_{950} or Perc15 measurements. These findings indicate that for CT lung density measurements, the reproducibility within academic-based and commercial vendors is similar.

Although efforts must be made to standardize CT measurements, including image acquisition protocols and image analysis software, there are other sources of variability that may impact CT measurements that were not considered in our study that must be acknowledged. For studies that acquire multiple CT image series over a short period of time, there is the potential for variability to be introduced due to physiological or patient-related factors, but not disease related factors, such as the patient orientation in the bore, slightly different lung inflation volumes at breath-hold, etc. Previous studies have investigated the short-term repeatability of CT lung density measurements within the same-day,[29] over 2-weeks,[30] and over a 1-yr period[31] in healthy volunteers and COPD patients. Although all studies report high short-term repeatability for CT measurements, these patient related factors may also impact how the software performs, and may add additional variability between groups. Therefore, an important limitation in our study is that we did not investigate both the reproducibility and short-term repeatability of the CT measurements between software tools. Our study is also limited by the fact that assessment of CT lung segmentation accuracy is ultimately subjective, and therefore we were only able to compare measurement reproducibility between the various software tools rather than accuracy, as ground truth segmentation is not available. Another factor that should be considered is the potential for individual commercial or academic groups to upgrade their software over time. For serial and longitudinal studies, even when the same software tool is used for CT analysis, CT measurement reproducibility may need to be reassessed. Furthermore, we note that we did not acquire CT measurements by lung lobe from software tools and therefore we did not investigate CT measurement reproducibility at the lobar level. Lobar segmentation algorithms between software tools may be more variable than whole lung segmentation. Reporting CT lung volume and density measurements by lobe is relevant for lung volume reduction applications, and therefore should be investigated in future studies. We also acknowledge that instruction was provided to the academic-based groups and commercial vendors using their software tools for performing the analysis, including how much manual intervention was permitted and that there should be no preprocessing of the images. This may or may not mimic how these vendors generate CT measurements routinely. However, the goal of our study was to assess the reproducibility of their software for generating CT lung density measurements under standardized conditions. Finally, as a result of the well-standardized CT image acquisition parameters used in this study, these findings may only be applicable to other well-standardized studies, or to clinical trials. Further investigation is required to determine CT measurement reproducibility between software tools for studies involving a wider range of CT acquisition parameters, such as those used in clinical practice.

In conclusion, we evaluated CT lung volume and lung density measurement reproducibility between eight different software tools using CT images acquired with standardized image acquisition protocols. The bias was negligible and measurement reproducibility was high between software tools, and was comparable for software developed by academic-based groups and commercial vendors. While using the same software tool for serial studies is highly recommended, these findings report how much added measurement variability will be introduced should it be necessary to include different software tools in serial studies with standardized image acquisition parameters, and provides guidance on how to incorporate such information into longitudinal studies.

ACKNOWLEDGMENTS

MK acknowledges support from the Parker B. Francis Fellowship Program and the Canada Research Chair Program (Tier 2). SF acknowledges research funding from GE Healthcare to develop MRI techniques. We thank Julie Lisiecki and QIBA for their organizational support, and all the academic groups and commercial vendors that participated in this study.

FUNDING

The COPDGene project described is supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene® project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, and Sunovion.

CONFILICTS OF INTEREST

JS was an employee and shareholder of VIDA Diagnostics Inc.; MK is a consultant at VIDA Diagnostics Inc.; CH is an employee of Imbio.

*QIBA Lung Density Committee Co-chairs: Sean B. Fain PhD, David A. Lynch MB, Charles Hatt PhD

[a] Author to whom correspondence should be addressed. Electronic mail: miranda.kirby@ryerson.ca; Telephone: 416-979-5000 (ext. 544418).

REFERENCES

1. Muller NL, Staples CA, Miller RR, Abboud RT. “Density mask”. An objective method to quantitate emphysema using computed tomography. *Chest*. 1988;94:782–787.
2. Gevenois PA, De Vuyst P, de Maertelaer V, et al. Comparison of computed density and microscopic morphometry in pulmonary emphysema. *Am J Respir Crit Care Med*. 1996;154:187–192.
3. Gevenois PA, Zanen J, de Maertelaer V, De Vuyst P, Dumortier P, Yernault JC. Macroscopic assessment of pulmonary emphysema by image analysis. *J Clin Pathol*. 1995;48:318–322.
4. Dirksen A, Dijkman JH, Madsen F, et al. A randomized clinical trial of alpha(1)-antitrypsin augmentation therapy. *Am J Respir Crit Care Med*. 1999;160:1468–1472.
5. Johannessen A, Skorge TD, Bottai M, et al. Mortality by level of emphysema and airway wall thickness. *Am J Respir Crit Care Med*. 2013;187:602–608.
6. Han MK, Kazerooni EA, Lynch DA, et al. Chronic obstructive pulmonary disease exacerbations in the COPDGene study: associated radiologic phenotypes. *Radiology*. 2011;261:274–282.
7. Fishman A, Martinez F, Naunheim K, et al. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med*. 2003;348:2059–2073.
8. Sciurba FC, Ernst A, Herth FJ, et al. A randomized study of endobronchial valves for advanced emphysema. *N Engl J Med*. 2010;363:1233–1244.
9. Chapman KR, Burdon JGW, Piitulainen E, et al. Intravenous augmentation treatment and lung density in severe 1 antitrypsin deficiency (RAPID): a randomised, double-blind, placebo-controlled trial. *Lancet*. 2015;386:360–368.
10. Madani A, Van Muylem A, Gevenois PA. Pulmonary emphysema: effect of lung volume on objective quantification at thin-section CT. *Radiology*. 2010;257:260–268.
11. Yuan R, Mayo JR, Hogg JC, et al. The effects of radiation dose and CT manufacturer on measurements of lung densitometry. *Chest*. 2007;132:617–623.
12. Zaporozhan J, Ley S, Weinheimer O, et al. Multi-detector CT of the chest: influence of dose onto quantitative evaluation of severe emphysema: a simulation study. *J Comput Assist Tomogr*. 2006;30:460–468.
13. Madani A, De Maertelaer V, Zanen J, Gevenois PA. Pulmonary emphysema: radiation dose and section thickness at multidetector CT

quantification–comparison with macroscopic and microscopic morphometry. *Radiology*. 2007;243:250–257.

14. Gierada DS, Bierhals AJ, Choong CK, et al. Effects of CT section thickness and reconstruction kernel on emphysema quantification relationship to the magnitude of the CT emphysema index. *Acad Radiol*. 2010;17:146–156.
15. Boedeker KL, McNitt-Gray MF, Rogers SR, et al. Emphysema: effect of reconstruction algorithm on CT imaging measures. *Radiology*. 2004;232:295–301.
16. Kim V, Davey A, Comellas AP, et al. Clinical and computed tomographic predictors of chronic bronchitis in COPD: a cross sectional analysis of the COPDGene study. *Respir Res*. 2014;15:52.
17. Ley-Zaporozhan J, Ley S, Weinheimer O, et al. Quantitative analysis of emphysema in 3D using MDCT: influence of different reconstruction algorithms. *Eur J Radiol*. 2008;65:228–234.
18. Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD*. 2010;7:32–43.
19. Lim H, Weinheimer O, Wielpütz MO, et al. Fully automated pulmonary lobar segmentation: influence of different prototype software programs onto quantitative evaluation of chronic obstructive lung disease. *PLoS ONE*. 2016;11:e0151498.
20. Wielpütz MO, Bardarova D, Weinheimer O, Kauczor H-U, Eichinger M. Variation of densitometry on computed tomography in COPD-influence of different software tools. *PLoS ONE*. 2014;9:112898.
21. Shen M, Tenda ED, McNulty W, et al. Quantitative evaluation of lobar pulmonary function of emphysema patients with endobronchial coils. *Respiration*. 2019;98:70–81.
22. Lung Density Committee. QIBA Profile: Computed Tomography: Lung Densitometry [Internet]. Available from qibawiki.rsna.org/images/c/c9/QIBA_CT_Lung_Density_Profile_062619-appendix-resolved.pdf
23. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat Methods Med Res*. 2015;24:27–67.
24. Sieren JP, Newell JD Jr., Barr RG, et al. SPIROMICS protocol for multicenter quantitative computed tomography to phenotype the lungs. *Am J Respir Crit Care Med*. 2016;194:794–806.
25. Vestbo J, Anderson W, Coxson HO, et al. Evaluation of COPD longitudinally to identify predictive surrogate end-points (ECLIPSE). *Eur Respir J*. 2008;31:869–873.
26. Multi-Ethnic Study of Atherosclerosis (MESA) Lung Study [Internet]. Available from http://www.cumc.columbia.edu/dept/medicine/generalmed/epi_copd.htm
27. Bourbeau J, Tan WC, Benedetti A, et al. Cohort obstructive lung disease (CanCOLD): fulfilling the need for longitudinal observational studies in COPD. *COPD*. 2014;11:125–132.
28. Obuchowski NA, Reeves AP, Huang EP, et al. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons. *Stat Methods Med Res*. 2015;24:68–106.
29. Iyer KS, Grout RW, Zamba GK, Hoffman EA. Repeatability and sample size assessment associated with computed tomography-based lung density metrics introduction. *Chronic Obstr Pulm Dis*. 2014;1:97–104.
30. Shaker SB, Dirksen A, Laursen LC, et al. Short-term reproducibility of computed tomography-based lung density measurements in alpha-1 antitrypsin deficiency and smokers with emphysema. *Acta Radiol*. 2004;45:424–430.
31. Shin JM, Kim TH, Haam S, et al. The repeatability of computed tomography lung volume measurements: Comparisons in healthy subjects, patients with obstructive lung disease, and patients with restrictive lung disease. *PLoS ONE*. 2017;12:e0182849.