Harmonization of chest CT scans for different doses and reconstruction methods

Gonzalo Vegas-Sánchez-Ferrero[a)]
Applied Chest Imaging Laboratory (ACIL), Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Maria Jesus Ledesma-Carbayo
Biomedical Image Technologies Laboratory (BIT) ETSI Telecomunicacion, UPM, and CIBER-BBN, Universidad Politécnica de Madrid, Madrid, Spain

George R. Washko
Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Raúl San José Estépar[a)]
Applied Chest Imaging Laboratory (ACIL), Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

(Received 2 January 2019; revised 25 March 2019; accepted for publication 22 April 2019; published 7 June 2019)

Purpose: To develop and validate a computed tomography (CT) harmonization technique by combining noise-stabilization and autocalibration methodologies to provide reliable densitometry measurements in heterogeneous acquisition protocols.
Methods: We propose to reduce the effects of spatially variant noise such as nonuniform patterns of noise and biases. The method combines the statistical characterization of the signal-to-noise relationship in the CT image intensities, which allows us to estimate both the signal and spatially variant variance of noise, with an autocalibration technique that reduces the nonuniform biases caused by noise and reconstruction techniques. The method is firstly validated with anthropomorphic synthetic images that simulate CT acquisitions with variable scanning parameters: different dosage, nonhomogeneous variance of noise, and various reconstruction methods. We finally evaluate these effects and the ability of our method to provide consistent densitometric measurements in a cohort of clinical chest CT scans from two vendors (Siemens, n = 54 subjects; and GE, n = 50 subjects) acquired with several reconstruction algorithms (filtered back-projection and iterative reconstructions) with high-dose and low-dose protocols.
Results: The harmonization reduces the effect of nonhomogeneous noise without compromising the resolution of the images (25% RMSE reduction in both clinical datasets). An analysis through hierarchical linear models showed that the average biases induced by differences in dosage and reconstruction methods are also reduced up to 74.20%, enabling comparable results between high-dose and low-dose reconstructions. We also assessed the statistical similarity between acquisitions obtaining increases of up to 30% points and showing that the low-dose vs high-dose comparisons of harmonized data obtain similar and even higher similarity than the observed for high-dose vs high-dose comparisons of nonharmonized data.
Conclusion: The proposed harmonization technique allows to compare measures of low-dose with high-dose acquisitions without using a specific reconstruction as a reference. Since the harmonization does not require a precalibration with a phantom, it can be applied to retrospective studies. This approach might be suitable for multicenter trials for which a reference reconstruction is not feasible or hard to define due to differences in vendors, models, and reconstruction techniques. [https://doi.org/10.1002/mp.13578]

Key words: CT scanner, calibration, Hounsfield Unit correction, lung density, quantitative imaging, COPD

1. INTRODUCTION

The characterization of computed tomography (CT) density measures is crucial for the development of image-based biomarkers for disease diagnosis, prognostication, and monitoring.[1–3] The main purpose of quantitative imaging (QI) techniques is to reduce functional, biological, and morphological processes to a measurable quantity employing medical imaging. The necessity of QI is especially important in light of a new healthcare delivery system that requires more personalized treatments and tries to tailor therapies to the underlying pathophysiology. The advancement in techniques to automatically interpret and quantify medical images has been recognized by regulatory agencies that have now proposed guidelines for the qualification of image-based biomarkers to be used as valid endpoints in clinical trials.[4] Nevertheless, the utility of QI is undoubtedly hampered by the discrepancies in the acquisition and reconstruction parameters, and dose

differences in clinical studies. In the specific case of CT images, the quantitative analysis involves dealing with intrinsic disparities in the density measures since image characteristics vary as a function of the chosen scanning parameters.[5,6]

One way to alleviate these discrepancies is to establish an acquisition protocol for multicenter studies to share similar scan parameters, although differences between manufacturers are still present.[7] Additionally, the high utilization of CT scans has raised some concerns about the implications of radiation exposure in clinical populations, suggesting the use of low-dose (LD) and ultra-low-dose techniques in clinical practice.[8,9] Therefore, although acquisition protocols and devices might be the same, the doses may change from subject to subject in both longitudinal and cross-sectional studies. The scenario becomes even more intricate with the advent of iterative reconstruction methods to deal with LD acquisitions. The iterative reconstruction methods affect the CT numbers differently depending on their assumptions and may result in a deviation of the desired calibration as commonly seen in PET attenuation correction techniques using LD CT protocols.[10]

Some efforts have been made to minimize the previously mentioned issues in clinical studies. Spatial discrepancies in the attenuation levels have been largely observed in clinical studies.[11,12] Some approaches using anatomical references like trachea and aorta densities have shown promising results.[13,14] The interscanner deviations are also an important factor that has been studied.[7]

In this work, we present a methodology that effectively combines the stabilization of spatially variant noise, that characterizes and minimizes the effect of noise, with an autocalibration scheme to remove the nonstationary biases. For this purpose, we apply the noise-stabilization method proposed by Vegas-Sánchez-Ferrero et al.[15] to characterize the spatially variant effects of noise in the CT numbers due to the different factors in the acquisition and reconstruction of CT scans. This methodology characterizes the signal/noise relationship through a statistical model that describes the spatially variant nature of noise and retrieves both components of signal and noise. Then, we combine the information obtained from the signal and noise estimates to remove the spatially variant and systematic biases across the scan following the method we proposed.[16] Since this methodology does not consider any combination of parameters for the reconstruction as a preferential standard, we refer to this method as a *harmonization* methodology. The methodological contribution is the nontrivial way of combining the noise-stabilization and the autocalibration techniques to take advantage of the features of both methodologies at the same time.

The assessment of our harmonization methodology is performed in two scenarios: first, with an *in silico* and anatomically realistic simulations with known CT numbers and edge-spread function; and second, with a clinical dataset of subjects from the COPDGene cohort that were scanned in either GE or Siemens scanners with different doses and reconstructed with two filtered backprojection kernels and iterative methods.

2. MATERIALS AND METHODS

2.A. Harmonization

The harmonization of CT scans is performed in four steps: (a) characterization of noise and tissues, (b) estimation of signal and local variance, (c) correction of spatially variant and systematic biases, and (d) preservation of stabilized details. In Fig. 1, we provide a schematic description of the method to illustrate every step.

(a) Characterization of noise and tissues. The estimation of both the signal and noise components of the CT image is performed by adopting a noncentral gamma distribution (nc-Γ) as the probabilistic model that characterizes the intensity values of the reconstructed CT images. It follows a three-parameter distribution:

$$f_\Gamma(x|\alpha,\beta,\delta) = \frac{(x-\delta)^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{\frac{x-\delta}{\beta}}, \quad x>\delta, \alpha>0, \beta>0 \tag{1}$$

where $\Gamma(\cdot)$ is Euler's gamma function, α and β are the shape and scale parameters, respectively, and δ is the location parameter set to the minimum value observed in the image, usually set slightly below -1000 HU. In our case, we set it to -1024 HU since that is usually the lower bound in clinical scans.

This probabilistic distribution has been recently proposed and evaluated as a versatile statistical model of CT numbers for different doses and reconstruction methods.[15] The presence of different tissues can be effectively modeled by a mixture model of nc-Γ distributions to the CT:

$$p(x(\mathbf{r})) = \sum_{j=1}^{J} \pi_j(\mathbf{r}) f_\Gamma(x(\mathbf{r})|\alpha_j(\mathbf{r}),\beta_j(\mathbf{r}),\delta) \tag{2}$$

for J components, where π_j are the weights of the mixture and α_j, β_j the parameters of the j-th component. This model allows us to describe the heterogeneous nature of tissues, the spatially variant response of noise, and its statistical properties described in the literature.[12,15–17]

To ensure that the heterogeneous composition of tissues is properly described, we set J = 9 components with mean values, $\{\mu_j\}_{j=1}^{J}$, ranging from -1000 to 400 HU. This is a reasonable range of attenuations to model tissues including air (-1000 HU), lung parenchyma (-700 HU), fat (-90 HU), vasculature and muscle (50 HU), and bone (>200 HU). The estimation of the parameters for each component is achieved through the expectation–maximization method for known mean values for each component, μ_j, which reduces the problem to solve a nonlinear equation in each iteration at each location.[15,18] The estimation of the shape parameters, $\alpha_j(\mathbf{r})$, are obtained by solving the following nonlinear equation derived from the maximum likelihood estimation in the

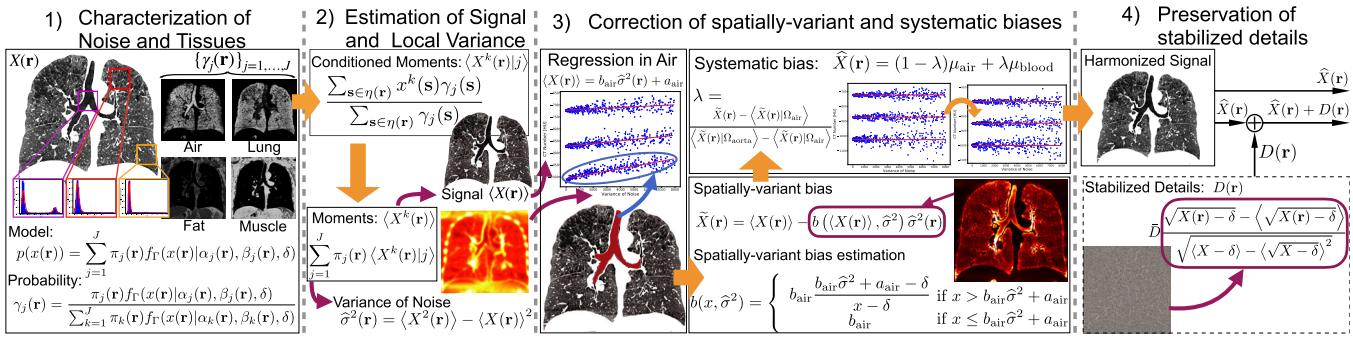

FIG. 1. Scheme of the proposed harmonization method. (a) The image is statistically characterized by a mixture model that provides the probability of belonging to each tissue class (posterior probabilities). (b) The conditioned local statistical moments are calculated through the local characterization per tissue by using the posterior probabilities. Then, they can be aggregated to estimate the moments. The signal and the spatial variance are the first and second order moments, respectively. (c) The functional relationship between signal and variance, $b(x,\hat{\sigma}^2)$, is estimated and the spatially variant bias removed. Then, the systematic bias is corrected considering two anatomical structures (trachea: -1000 HU, and descending aorta: 50 HU), resulting in the harmonized signal estimate: $\hat{X}(\mathbf{r})$. (d) The harmonized signal can be combined with the stabilized residual in order to preserve any details in the structure that might be codified within the noise. The detail, $D(\mathbf{r})$, can be added with an average standard deviation, $\bar{D}$, set as a parameter. [Color figure can be viewed at wileyonlinelibrary.com]

local neighborhood $\eta(\mathbf{r})$ (for clarity, we omit the reference to location):

$$\log(\alpha_j)-\psi(\alpha_j)=\frac{\sum\limits_{\mathbf{s}\in\eta}\gamma_j\frac{x(\mathbf{s})-\delta}{\mu_j-\delta}}{\sum\limits_{\mathbf{s}\in\eta}\gamma_j}-\frac{\sum\limits_{\mathbf{s}\in\eta}^{N}\gamma_j\log\left(\frac{x(\mathbf{s})-\delta}{\mu_j-\delta}\right)}{\sum\limits_{\mathbf{s}\in\eta}\gamma_j}-1 \tag{3}$$

where $\psi(\cdot)$ is the digamma function and $\gamma_j(\mathbf{r})$ is the probability for the j-th tissue class at location $\mathbf{r}$:

$$\gamma_j(\mathbf{r})=\frac{\pi_j(\mathbf{r})f_\Gamma(x(\mathbf{r})|\alpha_j(\mathbf{r}),\beta_j(\mathbf{r}),\delta)}{\sum_{k=1}^{J}\pi_k(\mathbf{r})f_\Gamma(x(\mathbf{r})|\alpha_k(\mathbf{r}),\beta_k(\mathbf{r}),\delta)}. \tag{4}$$

Then, the scale factor is calculated as $\beta_j=(\mu_j-\delta)/\alpha_j$ and the priors π_j are updated as $\pi_j=\frac{1}{|\eta(\mathbf{r})|}\sum_{\mathbf{s}\in\eta(\mathbf{r})}\gamma_j(\mathbf{s})$.

Equations (3) and (4) are iteratively applied until convergence is reached, which is usually achieved in very few iterations due to the constraint imposed by the mean $\{\mu_j\}_{j=1}^{J}$ for each tissue. A suitable initialization of parameters for the iterative optimization is $\pi_j=1/J$, $\alpha_j=2$ and $\beta_j=\mu_j/\alpha_j$ for each component.

(b) Estimation of signal and local variance. The estimation of the signal and the local variance of noise is finally achieved by the calculation of the sample conditioned moments to each tissue class as follows:

$$E\{X^k(\mathbf{r})|j\}\approx\langle X^k(\mathbf{r})|j\rangle=\frac{\sum\limits_{\mathbf{s}\in\eta(\mathbf{r})}x^k(\mathbf{s})\gamma_j(\mathbf{s})}{\sum\limits_{\mathbf{s}\in\eta(\mathbf{r})}\gamma_j(\mathbf{s})} \tag{5}$$

where the conditioned expectation operator, $E\{\cdot|j\}$, is approximated by the conditioned sample mean operator $\langle\cdot|j\rangle$. This formulation provides a more robust estimate of conditioned local moments since it just considers samples belonging to the j-th tissue class.

Finally, the moments for each location can be estimated as the weighted average of the conditioned moments as:

$$E\{X^k(\mathbf{r})\}\approx\langle X^k(\mathbf{r})\rangle=\sum_{j=1}^{J}\pi_j(\mathbf{r})\langle X^k(\mathbf{r})|j\rangle \tag{6}$$

So, the signal and the variance of noise can be directly estimated as $\langle X(\mathbf{r})\rangle$ and $\hat{\sigma}^2(\mathbf{r})=\langle X^2(\mathbf{r})\rangle-\langle X(\mathbf{r})\rangle^2$.

(c) Correction of spatially variant bias. The location-dependent variance of noise induces a bias in the CT numbers. As an example of this effect, in Fig. 2, we represent the bias induced by the spatially variant noise observed in a conventional clinical CT scan acquired with a high-dose (HD) and a soft reconstruction kernel (Siemens Definition, dose 400 mA, kernel B31f). Note that in Fig. 2(b) the samples of the trachea exhibit both a systematic bias (the mean value in the trachea is -938.36 HU) and a location-dependent bias that depends linearly on the variance of noise. We also represent the regression line $\langle X(\mathbf{r})\rangle=b_{\text{air}}\hat{\sigma}^2(\mathbf{r})+a_{\text{air}}$. The regression coefficient depends on tissue density.[16]

One of the advantages of adopting the noise model of Eq. (2) is that it provides a functional relationship between the regression coefficient and the CT number as the density increases: $b(x)=C/(x-\delta)$, where x is the CT number and C is a constant to be determined.[15] Note that $b(x)$ decreases as the tissue becomes denser due to the more symmetric distribution of tissues with higher attenuation value.

For the calculation of C, one can take advantage of the linear regression in the trachea and/or air external to the body as follows:

$$b(x,\hat{\sigma}^2)=\begin{cases}b_{\text{air}}\frac{b_{\text{air}}\hat{\sigma}^2+a_{\text{air}}-\delta}{x-\delta} & \text{if } x>b_{\text{air}}\hat{\sigma}^2+a_{\text{air}}\\ b_{\text{air}} & \text{if } x\le b_{\text{air}}\hat{\sigma}^2+a_{\text{air}}\end{cases} \tag{7}$$

This function sets the highest value of the regression coefficient to the one estimated for air. Then, the coefficient continuously decreases following the functional relationship $1/(x-\delta)$. Therefore, the linear relationship between density and noise variance fits both the regression in air and the decrease law observed.

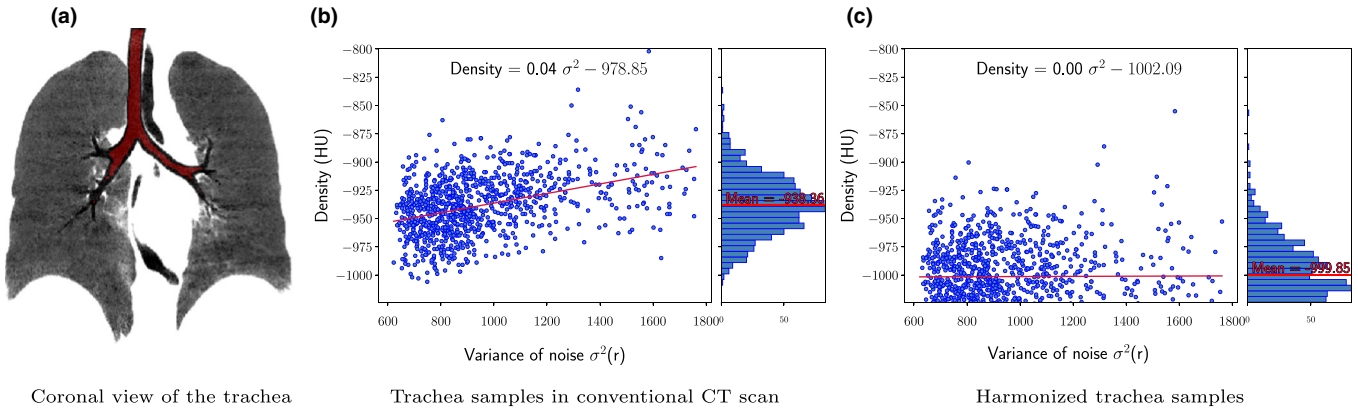

FIG. 2. Example of correction of functional dependence between density and noise variance in the trachea air for a high-dose and soft kernel reconstruction. (a) Coronal view of the minimum intensity projection to perceive the trachea (in red). (b) Regression of the computed tomography numbers and the spatial variance of noise in the trachea. (c) The harmonization corrects the functional dependence and reduces the bias in air. [Color figure can be viewed at wileyonlinelibrary.com]

Now, the spatially variant bias can be removed as follows:

$$\tilde{X}(\mathbf{r}) = \langle X(\mathbf{r})\rangle - b\big(\langle X(\mathbf{r})\rangle, \hat{\sigma}^2\big)\hat{\sigma}^2(\mathbf{r}) \tag{8}$$

It is important to note that a systematic bias still can be present in the image since Eq. (8) removes the linear relationship with the local variance, but not the intercept. We remove the intercept by adjusting the average attenuation levels of well defined anatomical references to their nominal values. The most evident structures are the descending aorta, Ω_{aorta}, where the blood attenuation level $\mu_{\text{blood}} = 50$ HU is usually adopted;[14] and the trachea and/or external air, Ω_{air}, where the air is set to $\mu_{\text{air}} = -1000$ HU by definition. Then, the harmonized image is obtained by linear interpolation for those attenuation levels:

$$\hat{X}(\mathbf{r}) = (1-\lambda)\mu_{\text{air}} + \lambda\mu_{\text{blood}},$$
$$\text{with} \quad \lambda = \frac{\tilde{X}(\mathbf{r}) - \langle\tilde{X}(\mathbf{r})|\Omega_{\text{air}}\rangle}{\langle\tilde{X}(\mathbf{r})|\Omega_{\text{aorta}}\rangle - \langle\tilde{X}(\mathbf{r})|\Omega_{\text{air}}\rangle} \tag{9}$$

In Fig. 2(c), we illustrate the effect of harmonization in the trachea. Note that the linear relationship between the variance of noise and the attenuation level is effectively removed. Furthermore, the intercept is corrected to a more reasonable CT value for air (~ -1000 HU) and the dispersion within the trachea is noticeably reduced providing a more accurate measure of density.

(d) Preservation of stabilized details. At this point, the signal and variance are estimated, and the spatially variant and systematic biases due to the heterogeneous noise corrected. So, if we are interested in a densitometric study, the harmonization methodology provides a suitable calibrated signal $\hat{X}(\mathbf{r})$ [cf. Eq. (1)]. However, if we are interested in the local changes of the signal to analyze specific anatomical structures, we would like to preserve any detail that might remain within the noise as much as possible. Our harmonization methodology is able to estimate those details, $D(\mathbf{r})$, and even incorporate them to the estimated signal as an additive component with a predetermined and stabilized dispersion across the image, $\bar{D}$ (i.e., a spatially homogeneous dispersion of average $\bar{D}$). To do so, we make use of the stabilization transformation derived by Vegas-Sánchez-Ferrero et al.[15] for a Gamma distribution, $\sqrt{X(\mathbf{r}) - \delta}$, that follows a Gaussian-like distribution. Then, the detail component following a zero-mean with homogeneous standard deviation, $\bar{D}$, to be added to the signal estimation $\hat{X}(\mathbf{r})$ becomes:

$$D(\mathbf{r}) = \bar{D} \cdot \frac{\sqrt{X(\mathbf{r}) - \delta} - \langle\sqrt{X(\mathbf{r}) - \delta}\rangle}{\sqrt{\langle X - \delta\rangle - \langle\sqrt{X - \delta}\rangle^2}} \tag{10}$$

2.B. Synthetic CT scans

The evaluation of the proposed methodology is performed in a set of anatomically realistic images with known reference values, spatially variant variance of noise, and edge spread function. The methodology applied to generate these images is the following:

Definition of the reference standard. The definition of an anatomical model with relevant structures for clinical purposes is always a challenging task. One natural option would be to make use of a physical phantom mimicking human structures of interest. Physical phantoms, however, usually present too homogeneous patterns and very localized synthetic lesions (when they are available) that make the analysis of confounding factors, such as spatially variant noise, more intricate and nonrealistic. In our case, with the purpose of having a gold standard to measure the discrepancies due to acquisition parameters, we created an anthropomorphic in-silico *phantom* from a HD and high-resolution real CT scan. We removed the noise and bias components; and assumed the resulting densities as the gold standard over which we will generate the simulated acquisitions.

These synthetic images have some limitations: (a) Reconstruction artifacts are assumed to be signal. However, those artifacts will be equally present across realizations, so they

will not affect the comparative analysis of harmonization. (b) The resolution is limited to that one obtained from the original image. To alleviate this fact, we considered a reference image acquired with a sharper kernel to obtain a representative resolution range as kernels become softer.

The anatomic reference model was obtained from a scan acquired with a Siemens Sensation 64, reconstruction kernel B46f, slice thickness 0.75 mm, and pixel dimensions 0.64 × 0.64 mm, (120 kVp, 400 mA). The noise was removed by means of an anisotropic filter proposed by Vegas-Sánchez-Ferrero et al.[15] to preserve the edges of the anatomical structures. In Fig. 3(a), we show the anatomic synthetic image with two different visualization windows (pulmonary and soft tissues). We selected a subject with evident areas of emphysema to model the bias effect of reconstruction and doses in areas of clear presence of disease.

Noise model. The noise generation was performed following the characteristics of noise described in the literature: (a) right skewness,[15,16,19] (b) linear relationship between mean and variance,[15,16] and (c) spatially variant noise.[12, 15–17] Following these features, the signal is set as the most likely value of a random realization of noise: the mode.

According to the linear relationship between attenuation level and noise, the highly dense structures should show a higher noise response, for example, bones and their neighborhoods, whereas soft tissues and air show a moderate noise response. The simulated nonstationary standard deviation of noise is shown in Fig. 3(b). It was defined as a smooth region-dependent map derived from the image with a low-pass filter. This map of the standard deviation of noise is normalized to an average level of 1 Hounsfield Unit (HU). The simulation of different doses will modify this average level, hereafter called $\bar{s}$, to mimic the increase of noise due to dose reduction.

To provide a realistic model of noise, we adopt the noncentral Gamma distribution since it fulfills the features of noise observed in studies with different devices, doses, and reconstruction algorithms.[15]

Due to the spatially variant nature of noise, the random variable parameters will depend on location $\mathbf{r}$. More specifically, let $m(\mathbf{r})$ be the signal defined at location $\mathbf{r} \in \Omega \subset \mathbb{R}^3$, and $s(\mathbf{r})$ the normalized standard deviation as the one shown in Fig. 3(b). According to the nc-Γ parametrization in Eq. (1), we can write the parameters in terms of the mode, $m(\mathbf{r})$, and the variance, $s^2(\mathbf{r})$, as follows:

$$\alpha(\mathbf{r}) = 1 + \frac{(m(\mathbf{r})-\delta)^2 + \sqrt{(m(\mathbf{r})-\delta)^4 + 4(m(\mathbf{r})-\delta)^2 s^2(\mathbf{r})}}{2s^2(\mathbf{r})}, \tag{11}$$

$$\beta(\mathbf{r}) = \frac{m(\mathbf{r})-\delta}{\alpha(\mathbf{r})-1} \tag{12}$$

We can now produce a realization of an anatomically realistic CT scan with the desired parameters (signal: $m(\mathbf{r})$, spatially variant noise: $s^2(\mathbf{r})$) by generating a nc-Γ($\alpha(\mathbf{r})$,$\beta(\mathbf{r})$) noise realization distributed at location $\mathbf{r}$.

In Fig. 4(a) and 4(b), we show the probability density functions (PDF) of a signal set to −1000 HU (corresponding to the air density) and −700 HU (lung parenchyma) for an increasing standard deviation, σ.

Resolution effects. The effects on the spatial resolution due to the different reconstruction kernels were simulated through isotropic Gaussian filtering for a varying range of scales defined by its standard deviation (κ) in the pixel dimensions.

It is important to note that the mean value of the noise distributions (dots for each distribution in Fig. 4) increases with σ due to the positive skewness of the distribution. This implies that any reconstruction kernel that performs low-pass filtering will induce an average bias in the attenuation level. The bias is more apparent for lower CT numbers such as air emphysema due to the stronger skewness Fig. 4(a). For higher attenuations, the bias is less pronounced Fig. 4(b) as the distribution becomes more symmetric. This effect has

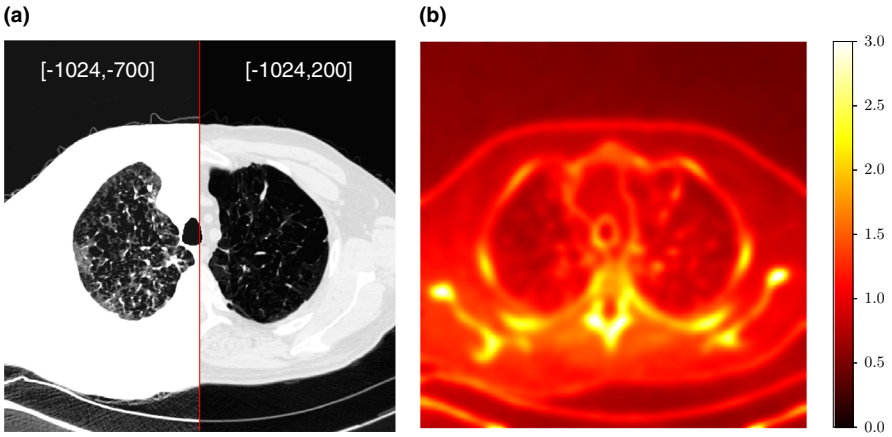

FIG. 3. (a) Synthetic image for two different visualization windows (left: lung parenchyma [−1024,−700] HU; right: soft tissue [−1024,200] HU). (b) Normalized nonstationary standard deviation of noise. [Color figure can be viewed at wileyonlinelibrary.com]

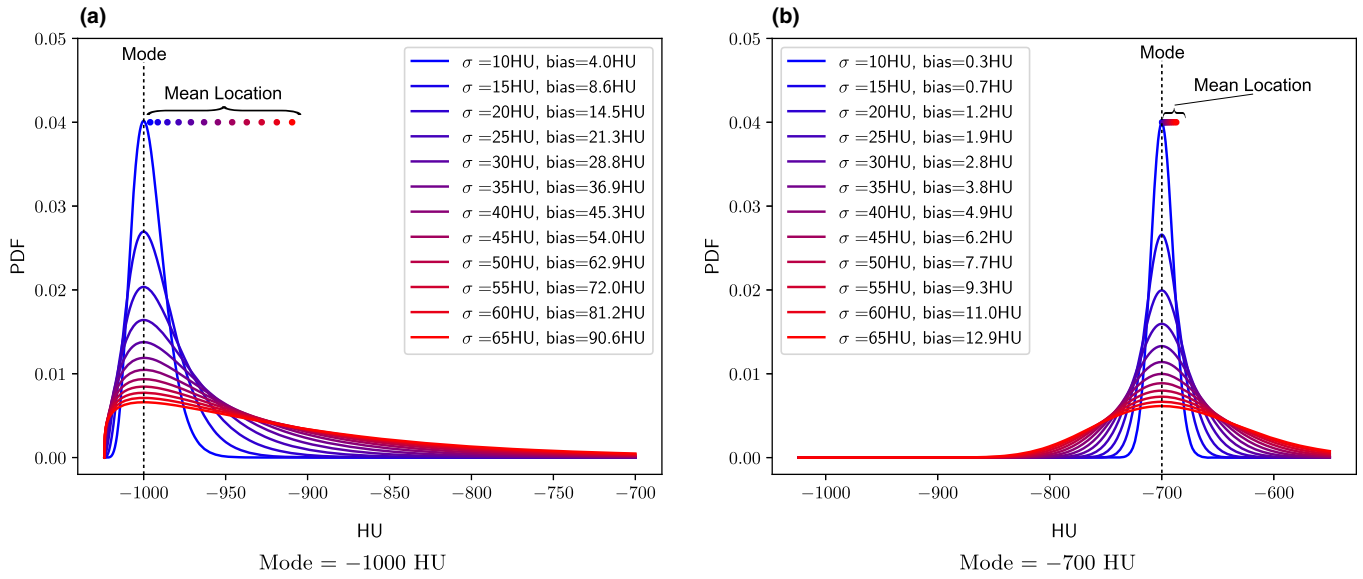

FIG. 4. Probability density function of a nc-Γ distribution known signal value set in the mode and increasing noise. Note that the most likely tissue density relies in the mode of the distribution, whereas the mean (represented as dots) depends on the variance of noise. This interrelationship will generate a bias (difference between mean and mode) that depends on the reconstruction kernel and the dose. [Color figure can be viewed at wileyonlinelibrary.com]

been extensively observed in the literature for clinical CT scans — especially in the trachea — due to the increase of noise variance caused by the surrounding densities of tissues.[13,14,16]

The selection of the scale range was performed considering the resolution provided by the reference image, which was assessed as the difference between 10% and 90% of the amplitude of the edge-spread function (ESF) in a region of the image with stable and known nominal attenuation levels (external air and subcutaneous fat). Since the synthetic image comes from an anatomical CT scan, we selected the edge response by following an approach inspired by the one proposed by Sanders et al.[20], who employed the contour of the subject to get the ESF between external air and the body surface. We set those nominal values to −1000 HU for air and −90 HU for fat. The samples to obtain the ESF were acquired in the perpendicular direction of the body contour. To correct the impairments in the ESF profiles, we approximated a sigmoid function as suggested by Li et al.[21] and centered the samples to a common reference. Finally, the centered ESF samples are used to fit a sigmoid function to calculate the difference between 10% to 90% range analytically. The result obtained for the reference image was $\Delta_{10\%}^{90\%} = 1.93$.

2.C. Clinical CT scans

A set of 104 subjects from the Phase 2 COPDGene study with multiple acquisitions were randomly selected. COPDGene is an observational longitudinal study funded by the NHLBI of 10 300 smokers whose goal is to define the epidemiologic associations and genetic risk factors for the development of COPD. The CT data available from those subjects come from two manufacturers: Siemens (54 subjects) and GE Medical Systems (50 subjects). Each subject was acquired at full inspiration, 120 kVp, with HD and LD protocols during the same session (400 mA and 100 mA, respectively). The CT images were reconstructed with both filtered backprojection (FBP) and iterative methods with a resolution of $\sim 0.65 \times 0.65 \times 0.75$ mm. The FBP reconstructions were performed with a soft and a sharp kernel in HD. The Siemens dataset provides seven different configurations of the same scanned subject that we will describe according to the dose (HD or LD) and the reconstruction method (FBP: soft, sharp; Iterative: Iter. 1, Iter. 2, and Iter. 3). Similarly, five different configurations are available for the GE dataset. A detailed summary of datasets is provided in Table I. Note that the Siemens dataset has iterative reconstruction for 33 of the 54

TABLE I. Summary of vendor parameters for each dataset.

Models	Avg. CTDIvol	Method	Description	#
Siemens (n = 54)				
Definition, Definition AS+, Sensation 64	13.47 mGy	B31f	HD soft	54
		B45f	HD sharp	54
	2.87 mGy	B31f	LD soft	54
		B45f	LD sharp	21
		I31f2	LD Iter. 1	33
		I31f5	LD Iter. 2	33
		I44f2	LD Iter. 3	33
GE (n = 50)				
Discovery CT750 HD, LightSpeed VCT	15.63 mGy	Standard	HD soft	50
		Bone	HD sharp	50
	2.96 mGy	Standard	LD soft	50
		ASIR 40%	LD Iter. 1	50
		ASIR 100%	LD Iter. 2	50

subjects, the other 21 subjects were reconstructed with the sharp kernel.

2.D. Evaluation metrics

We performed two different quantitative assessments of the harmonization technique. First, we focus on the effects over the image resolution, the noise reduction, and bias correction. For this purpose, we use the synthetic images described in Section 2 as the reference standard. We simulated a set of $n = 50$ realizations of independent noisy acquisitions generated from the anatomical reference with a wide range of average noise standard deviations, $\bar{s} \in \{1, 10, 20, 30, 40, 50, 60, 70\}$, and reconstruction kernels, $\kappa \in \{0.0, 0.3, 0.4, 0.5\}$. This is a reasonable range of parameters since HD scans with sharp kernels are equivalent to $\bar{s} \sim 40$ HU and $\kappa \sim 0.0$ by construction, and soft kernels have been reported to be about $\kappa \sim 0.4$.[15]

Then, we assessed the harmonization in the comparison of CT scans from the same subjects but different dose and reconstruction method. The evaluation was performed both in the synthetic images and in the clinical datasets. The assessment involved an analysis of consistency of densitometric measures (consistency of local mean across acquisitions) and an analysis of concordance (consistency of mean and variance across acquisitions). Finally, the assessment of similarity between local structures was studied by analyzing the differences between local probability distributions throughout a paired Kolmogorov–Smirnov test in the clinical dataset.

The specific metrics employed during the evaluation are defined as follows:

Spatial resolution. Distance for which the 10% and 90%, $\Delta_{10\%}^{90\%}$, of the edge response is reached. These margins serve as a robust metric to compare the differences between modulation transfer functions quantitatively.[22]

Noise reduction. The root mean square error (RMSE) with respect to the reference reconstruction.

Bias correction. Difference between the average CT number within the trachea and the nominal value of air −1000 HU. For the higher densities, we measure the bias of the fat surrounding the contour of the chest with respect to the nominal value in the synthetic images (−90 HU).

Consistency. The consistency of densitometric measures across doses and reconstruction methods is studied by modeling the local average of samples (neighborhoods of 7 × 7 voxels in the axial planes) with a two-level hierarchical linear model. The samples were taken in different locations of the volume covering densities from −1000 to 550 HU. The adopted model is the following:

$$y_{i,j,k,r} = A_j + M_k + e_{i,j,k} \tag{13}$$

where A_j is a random effect explaining the contribution of the different density regions and it is assumed to follow a normal independent distribution $\mathcal{N}(\mu, \sigma_A^2)$. M_k models the fixed effects of the reconstruction (dose, kernel, device) indexed as $k = 1, , n_M$ (Siemens: $n_M = 7$; GE: $n_M = 5$). In this model, the level-1 units are the samples acquired $i = 1,\ldots,n$ (Siemens: n = 54, GE: n = 50) and the level-2 units are the $j = 1, , n_A$ density levels (we set $n_A = 213$ levels for both datasets). The residual $e_{i,j,k}$ is assumed to follow a normal independent distribution $\mathcal{N}(0, \sigma_e^2)$.

For the synthetic images, the effects of the reconstruction method are the result of the composition of a reconstruction kernel term, R_r, a dose term, D_l, and their interaction:

$$y_{i,j,r,l} = A_j + \underbrace{R_r + D_l + (R \times D)_{r,l}}_{M_k} + e_{i,j,r,l} \tag{14}$$

So, the indices describe the different doses $l = 1,\ldots,8$ and reconstruction kernels $r = 1,\ldots,4$ (i.e., $k = 1,\ldots,32$).

The goodness of fit of the models is evaluated for both analyses through the root mean square error (RMSE). The effects of the reconstruction methods are described by using two dispersion metrics. First, we use the *range*, defined as $\Delta_{\text{range}} = \max(M_k) - \min(M_k)$ to study the range of effects due to the reconstruction methods. Second, we use the *mean absolute difference*, defined as $\Delta_{\text{MD}} = \sum_i \sum_j |M_i - M_j|/n^2$ to analyze the average deviations between methods. Low values of Δ_{MD} and high values of Δ_{range} would indicate that methods can be clustered. The software applied for this analysis was JMP® (Version 14.0.0, SAS Institute Inc., Cary, NC).

Concordance. We analyze the reliability of measures across acquisitions by means of the *concordance correlation coefficient*, ρ_{ccc}, proposed by Lin — a widely accepted index of agreement in settings with different raters.[23] This measure penalizes biases (differences in the first-order statistics) and deviations in their variances (second-order statistics), assumes a positive correlation between raters, and is defined in [0,1]. Values under 0.9 are commonly considered as a poor concordance.[24] A good harmonization methodology would provide high concordance across doses and kernels.

Statistical similarity. We assess the similarity between scans by comparing the probabilistic distribution of nonoverlapping local neighborhoods of 7 × 7 × 7 voxels through the paired Kolmogorov–Smirnov test. We consider the ratio of patches that are statistically indistinguishable as the statistical similarity metric. To ensure the preservation of local structure within the signal, a detail factor of $\bar{D} = 40$ HU is used, which is approximately the standard deviation of HD and soft kernel images.[15]

3. RESULTS AND DISCUSSION

3.A. Synthetic study

Noise reduction. The visual aspect of harmonized images is represented in Fig. 5 for a prototypical HD acquisition with both sharp and soft kernels. To improve the visualization of the results, we divided the image into two regions (left and right) to show the result of harmonization in different visualization windows, that is, lung window: [−1024,−700] HU, and soft tissue window: [−1024,200] HU. Qualitatively, harmonized images achieve a remarkable increase of contrast in

both lung parenchyma and soft tissues. This effect is due to the combination of noise reduction and bias removal, which we further analyze in the next sections.

Quantitatively, the RMSE results depicted in Fig. 6 increase with $\bar{s}$ as expected. There is also an increasing tendency as the kernel standard deviation, κ, becomes higher (softer). This trend is due to the blurring effect caused by softer reconstruction kernels. The slope of the curves, however, reduces as the kernel becomes softer as a result of the noise reduction by the reconstruction kernel.

The results obtained for the harmonized data show a less pronounced increase with noise, indicating a more robust behavior. Interestingly, both curves intersect at certain points for each reconstruction kernel. This intersection describes the situation in which noise is so low that the harmonization does not contribute positively. It is, nevertheless, important to state that those scenarios are far below the current acquisition conditions for clinical studies, even for HD protocols (clinical studies of HD protocols and soft reconstruction kernels have shown to be equivalent to $\kappa = 0.4$ and $\bar{s} = 40$ HU, for 400 mA, 120 kVp).[15]

Resolution. The results obtained for $\Delta_{10\%}^{90\%}$ are illustrated in Fig. 7. The proximity of results for the noisy and harmonized images indicates no loss of edge response after harmonization. Actually, there is a small improvement in the edge response for softer reconstruction kernels ($\kappa \geq 0.4$). This improvement is due to an increase in contrast achieved by the bias removal effect performed by the harmonization.

Bias correction. Quantitative results of bias obtained in the trachea are shown in Table II. Note that the bias is dramatically reduced by the harmonization for all the kernel and noise configurations. The bias is always below 2 HU after harmonization, which means a reduction of more than 95% for realistic scenarios such as $\kappa = 0.0$ and $\bar{s} = 40$ HU, for which the bias observed in the trachea is about 45 HU. In the case of softer reconstructions, the improvement is even better.

For the subcutaneous fat (−90 HU), the bias is less pronounced, reaching values around 5 HU in the noisiest scenarios. Similarly, the harmonization successfully removes the bias, achieving an average of $-3.04 \cdot 10^{-5} \pm 1.77 \cdot 10^{-5}$ HU for all configurations. For brevity, we omit the specific results.

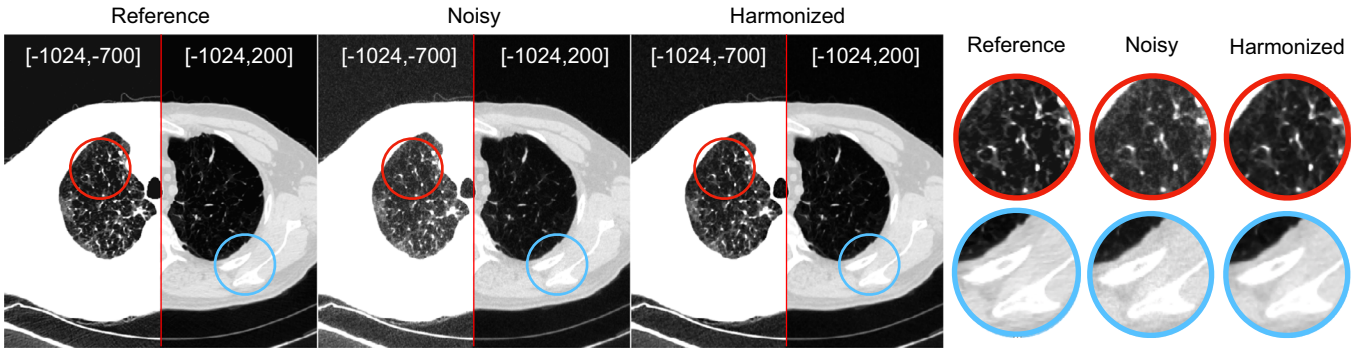

FIG. 5. Visual representation of harmonization of synthetic images reconstructed for a prototypical configuration for high-dose acquisitions with soft kernels ($\kappa = 0.4$ and $\bar{s} = 40$ HU). The image is divided in two regions where different visualization windows are applied to improve the visualization of the effect of noise and harmonization. [Color figure can be viewed at wileyonlinelibrary.com]

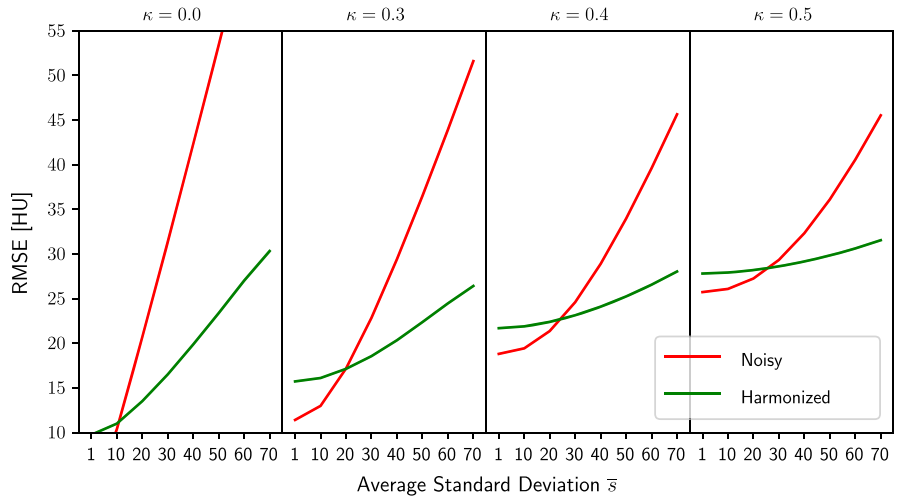

FIG. 6. Root mean square error for different reconstruction kernels and average standard deviation of noise. [Color figure can be viewed at wileyonlinelibrary.com]

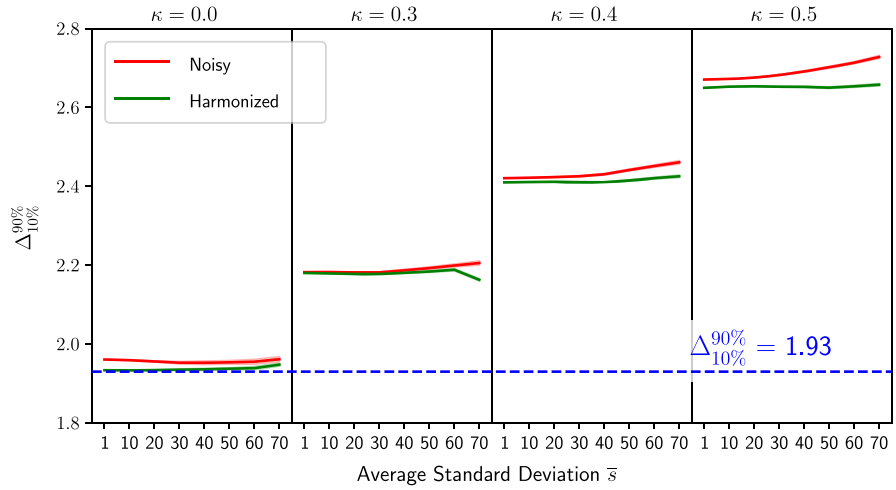

FIG. 7. Analysis of resolution for the width required to raise the edge response from 10% to 90% of reference levels for air (−1000 HU) and surrounding fat (−90 HU). [Color figure can be viewed at wileyonlinelibrary.com]

TABLE II. Bias observed in the noisy and harmonized data in the trachea region.

		Average standard deviation of noise $\bar{s}$															
Kernel	Description	1		10		20		30		40		50		60		70	
0.0	Noisy	1.56	(5.40)	7.25	(53.81)	18.19	(108.92)	30.74	(142.48)	42.76	(248.07)	55.25	(220.54)	68.77	(254.80)	81.56	(288.12)
	Harmonized	**0.90**	(2.90)	**0.70**	(17.66)	**0.06**	(31.74)	**-0.52**	(43.80)	**-0.71**	(81.59)	**-0.30**	(89.47)	**0.10**	(92.34)	**0.98**	(104.90)
0.3	Noisy	1.79	(4.63)	7.34	(41.56)	18.41	(70.02)	30.56	(106.49)	42.85	(149.22)	55.64	(191.28)	68.50	(241.51)	82.37	(281.89)
	Harmonized	**-0.04**	(3.16)	**0.22**	(17.55)	**-0.05**	(29.58)	**-0.73**	(53.34)	**-1.31**	(55.87)	**-1.50**	(70.43)	**-1.51**	(59.02)	**-0.77**	(107.44)
0.4	Noisy	3.54	(3.89)	9.17	(38.34)	20.15	(75.37)	32.37	(98.18)	44.97	(134.67)	57.37	(153.50)	70.34	(172.40)	83.25	(201.92)
	Harmonized	**0.32**	(4.92)	**0.71**	(16.17)	**0.68**	(29.10)	**-0.12**	(39.36)	**-0.88**	(50.51)	**-1.76**	(52.01)	**-2.20**	(72.85)	**-2.59**	(84.75)
0.5	Noisy	5.73	(3.59)	11.27	(24.43)	22.06	(52.03)	34.30	(83.47)	46.94	(112.42)	59.82	(123.47)	72.49	(140.61)	85.42	(155.18)
	Harmonized	**1.68**	(22.40)	**1.91**	(19.60)	**1.81**	(38.79)	**1.33**	(43.50)	**0.38**	(55.50)	**-0.63**	(55.39)	**-1.24**	(71.12)	**-1.64**	(71.69)

Bold indicates the lowest bias. Values are represented as: *mean* (*std* × 10^{-2}) HU.

TABLE III. Analysis of effects of reconstruction parameters for the different datasets (synthetic images, Siemens, GE) for the root mean square error (RMSE), dynamic range of fixed effects estimates, Δ_{range}, and mean absolute difference of the fixed effects, Δ_{MD}.

	Noisy/Reference			Harmonized		
Dataset	RMSE	Δ_{range}	Δ_{MD}	RMSE	Δ_{range}	Δ_{MD}
Synthetic images	9.58	40.44	14.19	6.41 (33.10%)	10.44 (74.18%)	2.65 (81.32%)
Siemens	83.48	40.65	13.56	61.19 (26.70%)	12.83 (68.44%)	3.50 (74.20%)
Siemens* (w/o LD Iter. 3)	70.47	15.86	7.16	48.51 (31.17%)	2.18 (86.27%)	0.90 (87.49%)
GE	78.03	5.56	2.45	59.16 (24.19%)	4.85 (12.72%)	1.94 (20.84%)

Values are represented in HU (% reduction with respect to the noisy/reference image).

Consistency. In Table III, we provide the results for the metrics describing the effects due to kernel reconstruction and dosage within the hierarchical model. As expected, the harmonization reduces the RMSE remarkably (33.10%), going from 9.58 to 6.41 HU. This implies an appreciable reduction of the residual noise due to the estimation of the signal. Besides, Δ_{range} is reduced in a 74.18%, from 40.44 to 10.44 HU. This reduction exhibits the excellent suppression of both dose and kernel effects by the harmonization methodology. That result is strongly supported by the 81.32% of reduction for the average absolute differences (from 14.19 to 2.65 HU).

Concordance. The concordance correlation coefficient, ρ_{ccc}, is shown in Fig. 8 for all possible combinations of kernel and doses. Note that the harmonized data obtains a much more homogeneous concordance map than the observed for the original data. The average concordance value obtained for the noise data is $\rho_{ccc} = 0.85$. This value implies that the noisy images show a poor concordance according to the strength-of-agreement proposed by McBride for the Lin's concordance correlation coefficient.[24] Conversely, the harmonized data achieves a $\rho_{ccc} = 0.95$ which is qualified as a substantial concordance and shows that densitometric measures are comparable in studies with

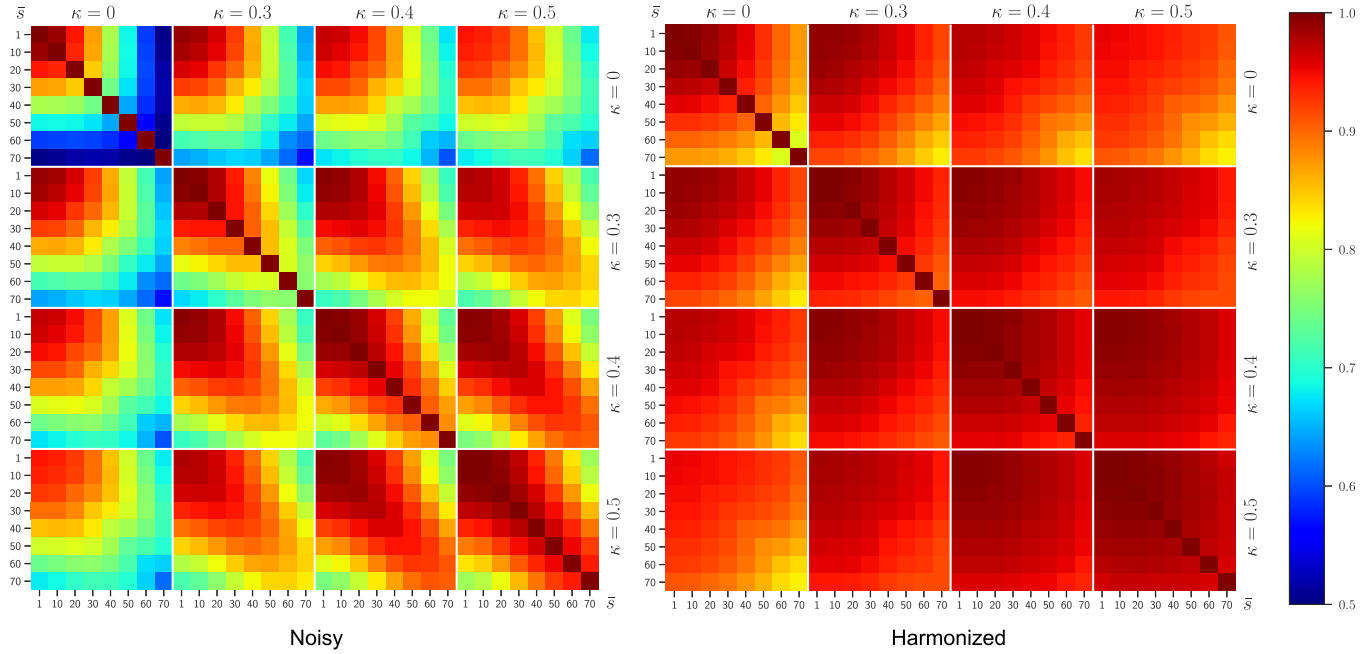

FIG. 8. Concordance correlation coefficient (ρ_{ccc}) for a pair of images with different configurations of dose ($\bar{s}$) and kernels (k). Note that ρ_{ccc} for the noisy data shows important discrepancies as the noise differences of images increase. The harmonized data shows a more homogeneous concordance across doses and kernels. [Color figure can be viewed at wileyonlinelibrary.com]

different doses and reconstruction kernels under the harmonization paradigm.

3.B. Clinical data

In Fig. 9, we show an example of the effect of harmonization for each dataset. We select two subjects with clear emphysema regions to appreciate the reduction of contrast as a consequence of the increase of noise and the reconstruction method for LD acquisitions. Note that the harmonization reduces dramatically the noise in the sharp HD reconstruction without compromising the anatomical structures. The harmonization also increases the contrast of LD acquisitions comparable to HD reconstructions.

Consistency. The statistical significance of all random and fixed terms is detailed in Appendix A, where we show the suitability of the proposed hierarchical linear model and the non-negligible effects of reconstruction methods. In Table III, we summarize those results through the RMSE and the two dispersion metrics employed for the consistency analysis. Interestingly, the estimate of the fixed effect of Iter. 3 for the Siemens dataset described in Appendix A (29.03 HU) shows that the bias dramatically differs from the other iterative methods (Iter 1: 4.09 HU and Iter. 2: 4.24 HU). This lack of consistency between Iter. 3 and the other iterative methods used for LD reconstruction, in combination with the remarkable high bias, indicates that the Siemens Iter. 3 method might not be an ideal reconstruction method for some quantitative analyses in studies with heterogeneous acquisition protocols. For this reason, we also included an analysis of the Siemens dataset without the LD Iter. 3 reconstruction. Nevertheless, that reconstruction technique will show the performance obtained by the harmonization in extreme cases.

The reference data has an RMSE around 80 HU for GE and Siemens datasets. The higher RMSE obtained in clinical data is obtained as a result of the variability introduced by the subject anatomies (the phantom just considered one anatomic reference).

Note that the harmonization reduces the RMSE by 25%. This reduction implies that the consistency among acquisitions that is not explained through the fixed effects M_k or random effects A_j is also considerably reduced. These reductions are a consequence of the noise reduction effect of harmonization and are consistent with the results obtained for the synthetic image dataset.

The Siemens dataset shows a $\Delta_{range} = 40.65$ HU, while $\Delta_{MD} = 13.65$ HU. This indicates that the effects due to reconstruction methods are clustered. To analyze this clustered behavior, we provide a detailed analysis of the pairwise differences for both datasets. Table IV shows the absolute pairwise differences for both the Siemens dataset (upper diagonal: reference; lower diagonal: harmonized). We also represent the connections between reconstruction methods by grouping them with letters (those configurations sharing letters are statistically similar to each other according to the Tukey's range test with significance 0.05). Note that the differences between LD Iter. 3 and the rest of methods are abnormally high (always beyond 24 HU and reaches more than 40 HU) indicating that Iter. 3 introduces large biases. However, even in this extreme case, the harmonization reduces the deviations in more than a 70% in some cases (from 40.65 HU to 10.65 HU when compared to HD soft),

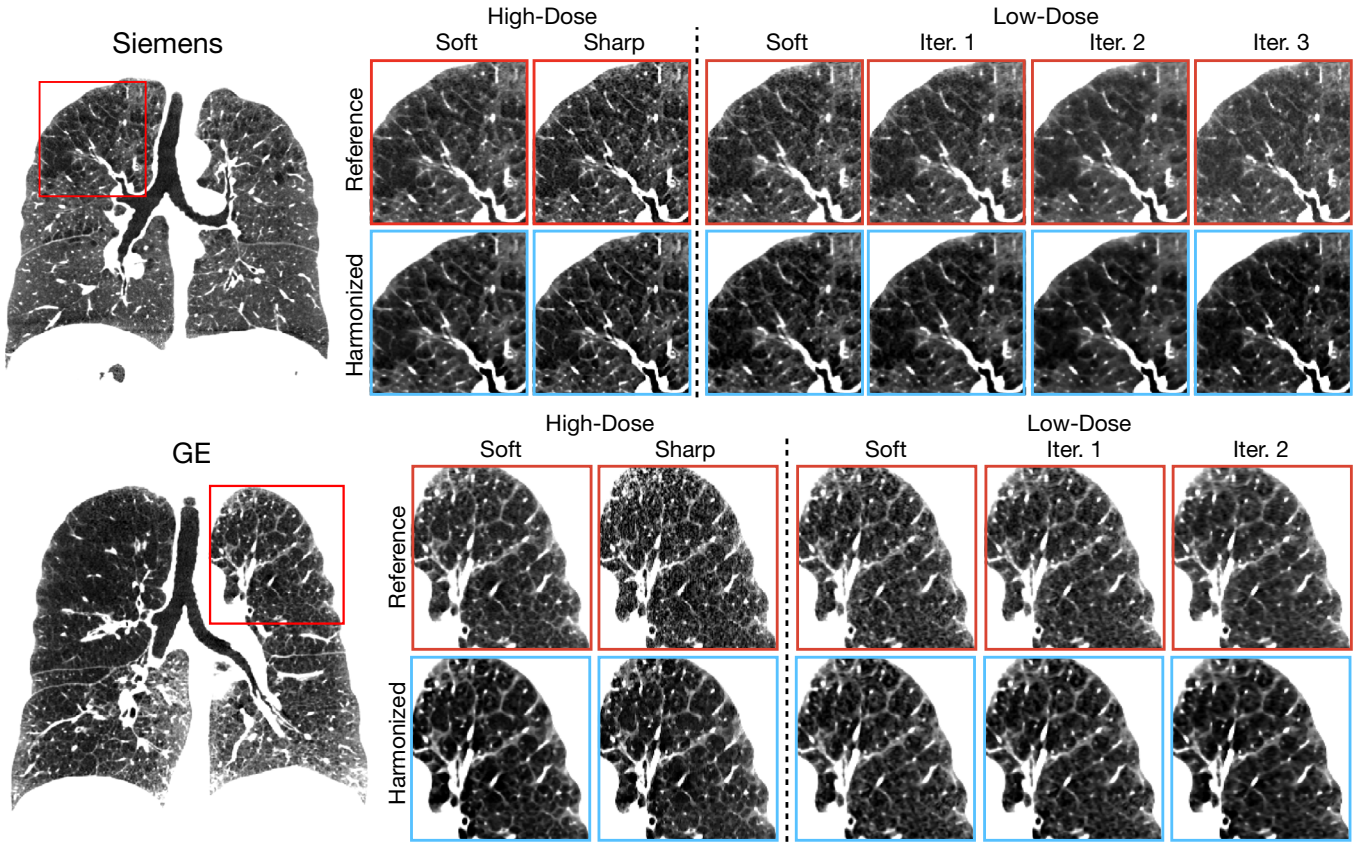

FIG. 9. Examples of harmonization for each dataset and detailed region with emphysema in the lung window [−1024,−700] HU. Note that the effect of noise reduces the visibility of anatomical structures in the sharp HD reconstruction for the reference images, and the reduction of the contrast in the low-dose reconstructions as a result of the noise and reconstruction method. The harmonization reduces the noise, improves the visibility of anatomical structures and increases the contrast in the low-dose reconstructions to the same levels observed in the high-dose reconstructions. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE IV. Absolute differences in the reconstruction effect (average std $\bar{s}$) for the Siemens dataset.

		HD		LD				
		Soft	Sharp	Soft	Sharp	Iter.1	Iter.2	Iter.3
HD	Soft	–	**0.26**	**2.75**	6.07	15.71	15.86	40.65
	Sharp	**0.20****	–	**2.49**	5.81	15.45	15.60	40.38
LD	Soft	**0.42*****	**0.23*****	–	**3.33**	12.96	13.11	37.90
	Sharp	**2.18*****	**1.98*****	**1.75*****	–	9.63	9.78	34.57
	Iter.1	**1.70*****	**1.50*****	**1.27*****	**0.48*****	–	**0.15**	24.94
	Iter.2	**1.17*****	**0.97*****	**0.75*****	**1.01*****	**0.52**	–	24.79
	Iter.3	10.65***	10.85***	11.07***	12.83***	12.34***	11.82***	–
Reference		A	A	A,B	B	C	C	D
Harmonized		A	A	A	A	A	A	B

Upper diagonal: reference data; lower diagonal: harmonized data; bold letters: statistically equivalent; * Difference decrease >10%, ** >20%, *** >30%. Configurations not connected with the same letters indicate are statistically different.

leading to an overall reduction of 68.44% for Δ_{range} (from 40.65 HU to 12.83 HU) and a 74.2.0% for Δ_{MD}.

The Tukey's range test shows three clusters in the reference dataset: (a) HD filtered backprojection methods; (b) LD filtered backprojection methods; (c) LD Iter. 1 and Iter. 2; and (d) Iter. 3. Interestingly, the LD soft method shows a low difference with the HD methods and the Tukey's range test accept their similarity. However, a higher tendency of more than 2.4 HU is observed. After harmonization, the Siemens dataset exhibits a homogeneous behavior for all the reconstruction method (except for Iter. 3). If we exclude Iter. 3, we get $\Delta_{\text{range}} = 2.18$ and $\Delta_{\text{MD}} = 0.90$, showing a reduction of more than 85% for both metrics. This reduction makes all the reconstruction methods statistically comparable and demonstrates the suitability of the harmonization method for the comparison of densitometric measures in chest CT scans.

TABLE V. Absolute differences in the reconstruction effect (average std $\bar{s}$) for the GE dataset.

		HD		LD		
		Soft	Sharp	Soft	Iter.1	Iter.2
HD	Soft	–	**1.08**	4.48	4.21	3.99
	Sharp	**1.65**	–	5.56	5.29	5.07
LD	Soft	4.85	3.20***	–	**0.27**	**0.49**
	Iter.1	4.08	2.43***	**0.78**	–	**0.22**
	Iter.2	**2.09*****	**0.44*****	2.76	**1.99**	–
Reference		A	A	B	B	B
Harmonized		A	A	B	B,C	A,C

Upper diagonal: reference data; lower diagonal: harmonized data; bold letters: statistically equivalent; * Difference decrease >10%, ** >20%, *** >30%. Configurations not connected with the same letters indicate are statistically different.

TABLE VI. Average of concordance correlation coefficient in % for pair comparisons of images acquired with different configurations in the Siemens dataset.

		HD		LD				
		Soft	Sharp	Soft	Sharp	Iter.1	Iter.2	Iter.3
HD	Soft	–	95.64	93.34	88.53	91.90	92.37	88.21
	Sharp	96.11	–	87.76	83.80	87.61	88.15	84.53
LD	Soft	94.56	**90.34**	–	98.12	96.99	96.11	92.09
	Sharp	**91.51**	**88.69**	97.77	–	–	–	–
	Iter.1	**95.19**	**92.10**	**99.77**	–	–	99.61	95.40
	Iter.2	**95.33**	**92.57**	**98.76**	–	99.48	–	95.60
	Iter.3	**90.97**	**89.60**	**94.30**	–	95.00	95.84	–

Upper diagonal: original images, lower diagonal: harmonized images. Bold letters: Improvement >1%.

TABLE VII. Average of concordance correlation coefficient in % for pair comparisons of images acquired with different configurations in the GE dataset.

		HD		LD		
		Soft	Sharp	Soft	Iter.1	Iter.2
HD	Soft	–	95.60	90.39	91.03	91.32
	Sharp	**97.24**	–	84.22	84.59	84.56
LD	Soft	91.17	**88.29**	–	99.80	98.69
	Iter.1	91.77	**88.31**	99.71	–	99.52
	Iter.2	91.91	**87.78**	98.44	99.43	–

Upper diagonal: original images, lower diagonal: harmonized images. Bold letters: Improvement >1%.

The GE dataset shows a more homogeneous behavior (Δ_{range} = 5.56 HU, Δ_{MD} = 2.45 HU), although the lower value of Δ_{MD} also suggests that the effects can be clustered. The reduction of the effects due to reconstruction is more subtle, decreasing Δ_{range} a 12.72% and Δ_{MD} a 20.84%. We also provide a detailed analysis of the pairwise differences for both datasets in Table V. The range test also identified two clusters: (a) HD methods and (b) LD methods. After harmonization, the LD cluster splits into the (b) LD soft method and (c) Iter. 1 and Iter. 2. However, note that now Iter. 2 does not show statistically significant differences in the HD methods.

Concordance. The average concordance correlation coefficient for the Siemens and the GE datasets are represented in Tables VI and VII, respectively. In the case of the Siemens, the concordance levels experience a clear increase after harmonization. Especially, the HD soft reconstruction compared to all the LD reconstruction, which reach substantial concordance levels (>95%).

We observe a similar behavior for the GE dataset, where the most noticeable increase is performed for the sharp HD method compared to the LD reconstructions. This increase is in agreement with the results observed in the consistency analysis. It is worth noting that the concordance increase observed for GE is more modest than for the Siemens dataset, although the densitometric results of both datasets are comparable. This is probably due to a more pronounced spatial variance within the GE scans since the ρ_{ccc} penalizes not only the biases but also the variance differences in the comparisons. This fact will also be perceived in the following section where we compare the similarity for local statistical distributions (and therefore, the biases, and variances).

Statistical Similarity. The ratio of neighborhoods with indistinguishable distributions considering a statistical significance of 0.001 are shown in Tables VIII and IX for each dataset. The increase for all pairs compared is evident (bold letters) except for those pairs that already showed very high values. The conclusions here are similar to the previous analyses: the harmonization remarkably increases the similarity of images across reconstruction scenarios (doses and kernels, and iterative methods). It is worth noting that the HD ratio of statistically equivalent samples was 52.88% and 37.87% for Siemens and GE, respectively. This low ratio is not due to differences in the mean value, as confirmed the consistency analysis for HD in both datasets, but in the variance of the local distribution, as was pointed out in the concordance analysis. The harmonization reduces the impairment of local variance and improves the ratio in more than 30 percentage points (Siemens: 89.24%, GE: 77.25%). Perhaps the most remarkable improvement is that the comparison between harmonized LD-to-HD reconstructions achieves higher ratios than HD-to-HD for the nonharmonized comparisons. This result evidences the suitability of the harmonization technique to alleviate statistical discrepancies between images acquired with different doses.

4. DISCUSSION AND CONCLUSIONS

We proposed a harmonization methodology to reduce the effects of spatially variant noise and biases derived from acquisitions. Our technique is intended to reduce the spatially variant nature of noise. It retrieves both the signal and noise components separately and removes the location-dependent biases induced by noise.

TABLE VIII. Average ratio of samples with statistically equivalent distributions with the Kolmogorov–Smirnov test (significance 0.001) in % acquired with different configurations in the Siemens dataset.

		HD		LD				
		Soft	Sharp	Soft	Sharp	Iter.1	Iter.2	Iter.3
HD	Soft	–	52.88	44.66	27.01	47.73	41.12	34.23
	Sharp	**89.24*****	–	56.51	62.49	43.96	16.24	33.65
LD	Soft	**68.13****	**61.20**	–	71.64	97.35	32.96	65.59
	Sharp	**51.40****	**66.91**	**74.09**	–	–	–	–
	Iter.1	**65.79***	**53.85**	97.22	–	–	64.78	69.59
	Iter.2	**51.79***	**32.52***	**75.84*****	–	**93.01****	–	40.61
	Iter.3	**50.49***	**48.30***	**71.35**	–	**75.81**	**71.12*****	–

Upper diagonal: original images, lower diagonal: harmonized images. Bold letters: Improvement >1%. * >10%, ** >20%, *** >30%.

TABLE IX. Average ratio of samples with statistically equivalent distributions with the Kolmogorov–Smirnov test (significance 0.001) in % acquired with different configurations in the GE dataset.

		HD		LD		
		Soft	Sharp	Soft	Iter.1	Iter.2
HD	Soft	–	37.87	42.95	51.74	46.66
	Sharp	**77.25*****	–	53.44	34.73	15.13
LD	Soft	**59.99***	**61.04**	–	99.98	57.40
	Iter.1	**63.68***	**52.41***	97.20	–	92.04
	Iter.2	**60.66***	**37.21****	**77.79****	**94.64**	–

Upper diagonal: original images, lower diagonal: harmonized images. Bold letters: Improvement >1%. * >10%, ** >20%, *** >30%.

We provided a thorough assessment of the harmonization in two scenarios. First, we evaluated its performance in anatomical *in-silico* simulations for different doses, kernels, and spatially variant noise. Then, we repeated the same analysis of consistency in a clinical dataset acquired in nine different hospitals from two manufacturers: Siemens (50 subjects) and GE (54 subjects). The subjects were scanned at different dose protocols during the same session, and the CT scans were reconstructed with different reconstruction methods including filtered back projection and iterative methods.

Our evaluation was performed in anthropomorphic simulations and real acquisitions to consider better the spatially variant noise that is the result of the interaction between x-ray energy with an inhomogeneous medium like the human body. Although studies based on phantom acquisitions are very relevant when assessing systematic effects of reconstruction techniques and potential signal restoration approaches, our experience is that the spatially variant effects due to photon starvation and beam hardening are hard to mimic in phantom experiments and real acquisition provides a more suitable scenario for the method presented in this work.

The evaluation in the *in-silico* simulations showed that our methodology is robust to noise without compromising the resolution of the original image. The biases induced by noise were successfully removed in low densities (air in the trachea) and high densities (subcutaneous fat). We also tested the consistency of local measurements with a hierarchical linear model showing a reduction of the effects caused by dose and kernel (average absolute reduction of 81.32%, from 14.19 to 2.65 HU). The statistical concordance between acquisitions was also measured, showing a remarkable increase in the strength-of-agreement (from 0.85 to 0.95).

The assessment methodology performed in the clinical datasets was able to detect strong biases introduced by an iterative reconstruction method (I44f2 Siemens iterative method). Even in that case, the harmonization was able to reduce the range of differences between acquisitions from 40.65 to 12.83 HU (a 68.44%). The performance of GE and Siemens scans is similar when the outlier is removed from the analysis (average differences between acquisitions of 7.15 HU for Siemens and 2.45 HU for GE). For both vendors, the harmonization is able to reduce the average differences between acquisitions to <2 HU). More importantly, the Tukey's range test shows that the differences between LD and HD CT scans observed in the original acquisitions are no longer appreciable after harmonization. The analysis of statistical concordance shows improvements in agreement with the consistency analysis (comparisons between LD and HD methods reached >95% of concordance after harmonization). A more detailed comparison of the local statistical distribution of data through the paired Kolmogorov–Smirnov test showed that the harmonization is able to increase the ratios of comparable regions between LD and HD scans to higher levels than high-dose to high-dose comparisons of nonharmonized scans.

In light of these results, we can conclude that the harmonization reduces the bias induced by noise without compromising resolution of the original image, as suggested by the evaluation of the edge response Fig. 7 in our simulations and the visual results of the harmonized image Fig. 9. The unique interaction between the noise and the location-dependent bias is an interesting effect that we adequately address with our methodology. This effect is especially relevant in low-density regions where the deviation between different acquisitions with different noise spectra is most acute. Prior studies using the COPDGene phantom has shown biases in the air compartment consistent with our results for both Siemens and GE as reported in Table III.[19,25] The hierarchical mixed-effect model shows that the effects due to dose and reconstruction

kernel are remarkably reduced after harmonization, making LD and HD reconstructions statistically equivalent for densitometric purposes. The harmonization not only reduces the deviations between reconstructions, concerning local average intensity, but also reduces discrepancies in the second-order statistics (variance) as was confirmed by a concordance correlation analysis between acquisitions.

Despite the positive effect of iterative reconstruction techniques in reducing image noise in LD acquisitions, our results indicate that the harmonization still is necessary and brings a positive effect for LD iterative reconstructions. We also noted that some iterative reconstructions might introduce unexpected deviations that should be considered when designing imaging protocols for specific studies. Nevertheless, the harmonization technique was able to reduce those outlier behaviors to bring them within the same class of performance in terms of absolute differences as shown in Table IV.

A relevant contribution of our work is the evaluation framework presented in Section 2.D. This framework assesses the difference between acquisition according to a panel of metrics for resolution, noise reduction, bias correction, consistency, concordance, and statistical similarity. We believe that this framework could be used for the systematic assessment of different protocols and the ability to produce reliable results for clinical studies that will rely on quantitative imaging biomarkers.

This work has some limitations that are worth noting. First, the presented harmonization technique only models the effect of spatially variant noise and induced biases, and it does not attempt to address other effects as beam hardening

Random Effects Covariance Parameter Estimates

Variance Component	Estimate	Std Error	95% Lower	95% Upper	Wald p-Value
Region	194822.55	18925.239	157729.77	231915.34	<.0001*
Residual	6968.7195	40.285317	6890.4309	7048.3562	
Total	201791.27	18925.282	169287.67	244699.48	

Fixed Effects Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|	95% Lower	95% Upper
Intercept	-578.2527	30.245487	212.0	-19.12	<.0001*	-637.8731	-518.6323
Method[Iter. 1 LD]	4.0942448	0.915286	59847	4.47	<.0001*	2.3002809	5.8882087
Method[Iter. 2 LD]	4.245363	0.915286	59847	4.64	<.0001*	2.4513991	6.0393269
Method[Iter. 3 LD]	29.031599	0.915286	59847	31.72	<.0001*	27.237635	30.825563
Method[Sharp HD]	-11.35232	0.7499019	59847	-15.14	<.0001*	-12.82213	-9.882513
Method[Sharp LD]	-5.539368	1.1146351	59847	-4.97	<.0001*	-7.724057	-3.354679
Method[Soft HD]	-11.6135	0.7499019	59847	-15.49	<.0001*	-13.08331	-10.14369

Reference

Random Effects Covariance Parameter Estimates

Variance Component	Estimate	Std Error	95% Lower	95% Upper	Wald p-Value
Region	220345.38	21403.125	178396.03	262294.74	<.0001*
Residual	3744.1685	21.644582	3702.1055	3786.9559	
Total	224089.55	21403.136	187415.32	272752.71	

Fixed Effects Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|	95% Lower	95% Upper
Intercept	-607.1729	32.164499	212.0	-18.88	<.0001*	-670.5761	-543.7697
Method[Iter. 1 LD]	-2.407203	0.6709004	59847	-3.59	0.0003*	-3.72217	-1.092236
Method[Iter. 2 LD]	-1.882972	0.6709004	59847	-2.81	0.0050*	-3.197939	-0.568004
Method[Iter. 3 LD]	9.9368806	0.6709004	59847	14.81	<.0001*	8.6219134	11.251848
Method[Sharp HD]	-0.910324	0.5496746	59847	-1.66	0.0977	-1.987688	0.1670404
Method[Sharp LD]	-2.889539	0.8170223	59847	-3.54	0.0004*	-4.490905	-1.288172
Method[Soft HD]	-0.711402	0.5496746	59847	-1.29	0.1956	-1.788766	0.3659624

Harmonized

FIG. 10. Details and significance of the hierarchical linear model for the Siemens dataset. [Color figure can be viewed at wileyonlinelibrary.com]

Random Effects Covariance Parameter Estimates

Variance Component	Estimate	Std Error	95% Lower	95% Upper	Wald p-Value
Region	191939.87	18645.214	155395.92	228483.82	<.0001*
Residual	6089.0844	37.393309	6016.4544	6163.0438	
Total	198028.96	18645.251	166022.11	240327.67	

Fixed Effects Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|	95% Lower	95% Upper
Intercept	-591.7664	30.020668	212.0	-19.71	<.0001*	-650.9436	-532.5891
Method[Iter. 1 LD]	1.8884065	0.6763106	53033	2.79	0.0052*	0.5628319	3.2139812
Method[Iter. 2 LD]	1.6726595	0.6763106	53033	2.47	0.0134*	0.3470848	2.9982341
Method[Sharp HD]	-3.400755	0.6763106	53033	-5.03	<.0001*	-4.72633	-2.07518
Method[Soft HD]	-2.320755	0.6763106	53033	-3.43	0.0006*	-3.64633	-0.99518

Reference

Random Effects Covariance Parameter Estimates

Variance Component	Estimate	Std Error	95% Lower	95% Upper	Wald p-Value
Region	215428.69	20925.644	174415.18	256442.2	<.0001*
Residual	3499.7901	21.49235	3458.045	3542.2993	
Total	218928.48	20925.655	183075.81	266511.51	

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|	95% Lower	95% Upper
Intercept	-604.9164	31.803585	212.0	-19.02	<.0001*	-667.6082	-542.2247
Method[Iter. 1 LD]	1.5438394	0.5127329	53033	3.01	0.0026*	0.5388784	2.5488004
Method[Iter. 2 LD]	-0.442395	0.5127329	53033	-0.86	0.3882	-1.447356	0.5625662
Method[Sharp HD]	-0.885737	0.5127329	53033	-1.73	0.0841	-1.890698	0.1192243
Method[Soft HD]	-2.534657	0.5127329	53033	-4.94	<.0001*	-3.539618	-1.529696

Harmonized

FIG. 11. Details and significance of the hierarchical linear model for the GE dataset. [Color figure can be viewed at wileyonlinelibrary.com]

or other reconstruction artifacts that might have a differential behavior across vendors and reconstruction software versions. Additionally, although the resolution of harmonized images is not compromised, the differences in resolution between soft and sharp reconstructions still persist. This limits the improvement of concordance between reconstructions with too different resolutions. Second, the presented evaluation is confined to commons reconstruction techniques for Siemens and GE. Although the spatially variant noise is implicit to the physics of the CT acquisition, we do not have data from other vendors that could lead to more general conclusions related to the effects of harmonization in multicenter studies using imaging. Finally, the data used in this study were collected as part of phase 2 of the COPDGene study and is limited to the protocol characteristics defined during the study design.

To conclude, we believe that the results obtained with our harmonization method evidence its suitability to alleviate the statistical discrepancies between images acquired under heterogeneous acquisition protocols in a vendor-neutral fashion — a particularly desirable property for multicenter studies. The reduction of the impact of kernels and protocol differences is an important property that might contribute to defining more robust features in emphysema and gas trapping densitometric biomarkers for COPD and radiomics features in lung cancer studies.[26–28] Besides, since our technique is built upon the statistical models that characterize the signal-to-noise relationship, it does not require any special

precalibration phantom and, therefore, can be applied retrospectively to re-analyze already-acquired data.

DISCLOSURES

The authors have no conflicts to disclose.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health NHLBI awards R01HL116931, R01HL116473, R21HL14042; DoD award PR171782; and a sponsored research award by Boehringer-Ingelheim Pharmaceuticals. The COPDGene study (NCT00608764) is supported by NHLBI U01 HL089897 and U01 HL089856, and the COPD Foundation through contributions made to an Industry Advisory Committee comprised of AstraZeneca, Boehringer-Ingelheim, GlaxoSmithKline, Novartis, and Sunovion. This work has been partially funded by the Spanish Ministry of Science,Innovation and Universities through project RTI2018-098682-B-I00 [Correction added on June 20, 2019, after first online publication: Acknowledgment statement updated to include partial funding by the Spanish Ministry of Science, Innovation and Universities.]

APPENDIX A

HIERARCHICAL LINEAR MIXED MODEL FOR CONSISTENCY ANALYSIS OF CLINICAL SCANS

The results of the adjusted hierarchical linear mixed model is shown in Figs. 10 and 11 for the Siemens and GE scans, respectively. Note that the model significantly explains the effects of kernels and acquisitions in the reference datasets of both scanner brands. After harmonization, the Siemens dataset shows a remarkable reduction in the fixed effects for all reconstruction methods. In some cases, the reduction is so effective that the fixed effects become nonstatistically significant (soft HD and sharp HD). For the GE scans, the harmonization achieves a smaller reduction, but also makes some of the fixed effects nonstatistically significant (sharp HD and Iter. 2).

[a)]Authors to whom correspondence should be addressed. Electronic mails: gvegas@bwh.harvard.edu; rsanjose@bwh.harvard.edu.

REFERENCES

1. Buckler AJ, Bresolin L, Dunnick NR, et al.Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. *Radiology.* 2011;259:875–884
2. Abramson R, Burton K, Yu J, et al. Methods and challenges in quantitative imaging biomarker development. *Acad Radiol*. 2015;22:25–32
3. Wu AC, Kiley JP, Noel PJ, et al. Current status and future opportunities in lung precision medicine research with a focus on biomarkers. An American Thoracic Society/National Heart, Lung, and Blood Institute Research Statement. *Am J Respir Crit Care Med*. 2018;198:e116–e136
4. FDA-NIH BiomarkerWorking Group. BEST (Biomarkers, EndpointS, and other Tools) Resource; 2016.
5. Cagnon CH, Cody DD, McNitt-Gray MF, Seibert JA, Judy PF, Aberle DR. Description and implementation of a quality control program in an imaging-based clinical trial. *Acad Radiol*. 2006;13:1431–1441
6. Mulshine JL, Gierada DS, Armato SG III, et al. Role of the quantitative imaging biomarker alliance in optimizing CT for the evaluation of lung cancer screen-detected nodules. *J Am Coll Radiol*. 2015;12:390–395
7. Chen-Mayer HH, Fuld MK, Hoppel B, et al. Standardizing CT lung density measure across scanner manufacturers. *Med Phys*. 2017;44:974–985
8. Brenner DJ, Hall EJ. Computed tomography — an increasing source of radiation exposure. *N Engl J Med*. 2007;357:2277–2284
9. Mayo-Smith WW, Hara AK, Mahesh M, Sahani DV, Pavlicek W. How I do it: managing radiation dose in CT. *Radiology* 2014;273:657–672
10. Xia T, Alessio AM, De Man B, Manjeshwar R, Asma E, Kinahan PE. Ultra-low dose CT attenuation correction for PET/CT. *Phys Med Biol*. 2012;57:309–328
11. Zeng GL. Nonuniform noise propagation by using the ramp filter in fan-beam computed tomography. *IEEE Trans Med Imaging*. 2004;23:690–695
12. Kim JH, Chang Y, Ra, JB. Denoising of polychromatic CT images based on their own noise properties. *Med Phys*. 2016;43:2251–2260
13. Choi S, Hoffman EA, Wenzel SE, Castro M, Lin CL. Improved CT-based estimate of pulmonary gas trapping accounting for scanner and lung-volume variations in a multicenter asthmatic study. *J Appl Physiol*. 2014;117:593–603
14. Kim SS, Seo JB, Kim N, et al. Improved correlation between CT emphysema quantification and pulmonary function test by density correction of volumetric CT data based on air and aortic density. *Eur J Radiol*. 2014;83:57–63
15. Vegas-Sánchez-Ferrero G, Ledesma-Carbayo MJ, Washko GR, San José Estépar R. Statistical characterization of noise for spatial standardization of CT scans: enabling comparison with multiple kernels and doses. *Med Image Anal*. 2017;40:44–59
16. Vegas-Sánchez-Ferrero G, Ledesma-Carbayo MJ, Washko GR, Estépar RSJ. Autocalibration method for non-stationary ct bias correction. *Med Image Anal*. 2018;44:115–125
17. Vaishnav JY, Jung WC, Popescu LM, Zeng R, Myers KJ. Objective assessment of image quality and dose reduction in CT iterative reconstruction. *Med Phys*. 2014;41:071904
18. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol*. 1977;39:1–38
19. Rodriguez A, Ranallo FN, Judy PF, Fain SB. The effects of iterative reconstruction and kernel selection on quantitative computed tomography measures of lung density. *Med Phys*. 2017;44:2267–2280
20. Sanders J, Hurwitz L, Samei E. Patient-specific quantification of image quality: an automated method for measuring spatial resolution in clinical CT images. *Med Phys*. 2016;43:5330–5338
21. Li, T, Feng H, Xu Z, Li, X, Cen Z, Li Q. Comparison of different analytical edge spread function models for MTF calculation using curve-fitting. Number June; 2009: 74981H.
22. Smith SW. Special imaging techniques. In: *Digital Signal Processing*. Amsterdam: Elsevier; 2003:423–450
23. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45:255
24. McBride G. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. *NIWA Client Report*. 2005;45:307–310
25. Sieren JP, Hoffman E, Fuld MK, Chan KS, Guo J, Newell JD. Sinogram affirmed iterative reconstruction (SAFIRE) versus weighted filtered back projection (WFBP) effects on quantitative measure in the COPDGene 2 test object. *Med Phys*. 2014;41:091910
26. Gierada DS, Bierhals AJ, Choong CK, et al. Effects of CT section thickness and reconstruction Kernel on emphysema quantification. *Acad Radiol*. 2010;17:146–156
27. Mackin D, Fave X, Zhang L, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol*. 2015;50:757–765
28. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys*. 2018;102:1143–1158