Respiratory trace feature analysis for the prediction of respiratory-gated PET quantification

Shouyi Wang[1,2], Stephen R Bowen[3,4], W Art Chaovalitwongse[1,2], George A Sandison[4], Thomas J Grabowski[2,3] and Paul E Kinahan[3]

[1] Department of Industrial and Systems Engineering, 3900 Stevens Way, Seattle, WA 98195, USA
[2] Integrated Brain Imaging Center, 1959 NE Pacific St, Seattle, WA 98195, USA
[3] Department of Radiology, 1959 NE Pacific St, Seattle, WA 98195, USA
[4] Department of Radiation Oncology, 1959 NE Pacific St, Seattle, WA 98195, USA

E-mail: srbowen@uw.edu

Abstract

The benefits of respiratory gating in quantitative PET/CT vary tremendously between individual patients. Respiratory pattern is among many patient-specific characteristics that are thought to play an important role in gating-induced imaging improvements. However, the quantitative relationship between patient-specific characteristics of respiratory pattern and improvements in quantitative accuracy from respiratory-gated PET/CT has not been well established. If such a relationship could be estimated, then patient-specific respiratory patterns could be used to prospectively select appropriate motion compensation during image acquisition on a per-patient basis. This study was undertaken to develop a novel statistical model that predicts quantitative changes in PET/CT imaging due to respiratory gating. Free-breathing static FDG-PET images without gating and respiratory-gated FDG-PET images were collected from 22 lung and liver cancer patients on a PET/CT scanner. PET imaging quality was quantified with peak standardized uptake value (SUV_{peak}) over lesions of interest. Relative differences in SUV_{peak} between static and gated PET images were calculated to indicate quantitative imaging changes due to gating. A comprehensive multidimensional extraction of the morphological and statistical characteristics of respiratory patterns was conducted, resulting in 16 features that characterize representative patterns of a single respiratory trace. The six most informative features were subsequently extracted using a stepwise feature selection approach. The multiple-regression model was trained and tested

based on a leave-one-subject-out cross-validation. The predicted quantitative improvements in PET imaging achieved an accuracy higher than 90% using a criterion with a dynamic error-tolerance range for SUV_{peak} values. The results of this study suggest that our prediction framework could be applied to determine which patients would likely benefit from respiratory motion compensation when clinicians quantitatively assess PET/CT for therapy target definition and response assessment.

Keywords: patient classification, respiratory motion, PET/CT, respiratory gating

1. Introduction

Combined positron emission tomography (PET) and computed tomography (CT) imaging has gained prominence in detection and staging of abdominothoracic cancer patients due to its strong association to clinical outcome (Imamura *et al* 2011, Vansteenkiste *et al* 1998). However, the quantitative accuracy of PET/CT for defining target volumes (Bettinardi *et al* 2010b, Caldwell *et al* 2001, Senan and Ruysscher 2005) and assessing response to therapy (Avril and Weber 2005) has been limited in part by respiratory-induced tumor motion (Nehmeh *et al* 2002). Time-averaged PET/CT images acquired under free-breathing conditions, known as static PET/CT, can artificially reduce apparent lesion uptake and increase apparent tracer-avid lesion volumes from motion blurring (Liu *et al* 2009). Methods to remove this blurring include respiratory-gated PET/CT by compensating for tumor motion (Aristophanous *et al* 2011, Nehmeh *et al* 2003).

Despite the potential for significant improvements in quantitative accuracy from respiratory-gated PET/CT (Bettinardi *et al* 2010a, 2012, Guerra *et al* 2012), particularly for the definition of biological targets (Aristophanous *et al* 2011), the gains vary tremendously between individual patients due to numerous patient-specific factors. While respiratory-gated PET/CT has the potential to increase contrast between tracer-avid lesions and background in some patients, it can also lead to increased image noise levels with no contrast improvement in other patients (Liu *et al* 2009). Therefore, a key question is 'can information gained from the respiratory traces help predict the quantitative gains from respiratory-gated PET/CT?' These predictions could aid in deciding between various motion compensation and motion suppression strategies prior to PET/CT acquisition (Bowen *et al* 2012). For example, in patients where gating would yield only small changes in quantitative accuracy, static scans with or without active breathing control devices may be used instead (Keall *et al* 2006, Wong *et al* 1999). As the clinical use of medical devices for motion suppression or control is invasive, a prediction should ideally be made prior to PET/CT acquisition on whether to gate under free-breathing conditions.

Many types of respiratory motion prediction models exist. The majority have focused on predicting the respiratory pattern at predefined time intervals in advance, most commonly with auto-regressive moving average (McCall and Jeraj 2007, Ren *et al* 2007) or support vector regressive techniques (Riaz *et al* 2009). These models have many applications in respiratory-gated radiation therapy, which allow for the prediction of future tumor positions in order to turn the treatment beam on/off during the appropriate gating window, but thus far have not been applied to respiratory-gated imaging. In particular, the ability to predict changes in PET/CT

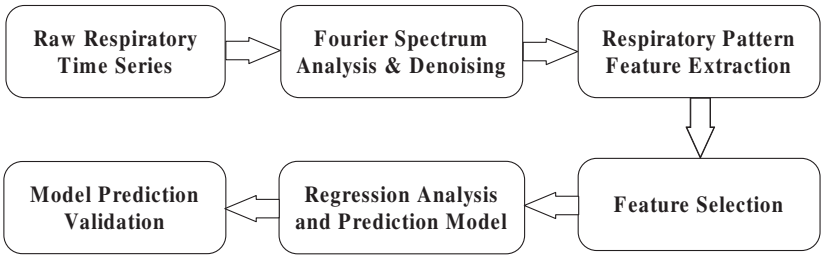


Figure 1. The computational framework for building a prediction model for PET quantification of imaging quality from respiratory trace features.

quantification based on respiratory motion features has not to our knowledge been reported elsewhere.

The purpose of this study was to develop a predictive model of quantitative benefit from respiratory-gated PET/CT that can eventually be used as a clinical decision support tool. Specifically, a new technique was developed to extract features of respiratory patterns, and in turn to construct a model associating these features to changes in PET imaging metrics. Such a model requires quantitative estimates of improvements in respiratory-gated imaging relative to free-breathing static imaging for training, but following validation it should be flexible enough to predict imaging changes based only on respiratory pattern features acquired prior to PET/CT acquisition.

Several quantitative PET imaging metrics have been used clinically, ranging from maximum standardized uptake value (SUV_{max}) to mean SUV (SUV_{mean}) in a region-of-interest. In this study, PET lesion tracer avidity was quantified with peak standardized uptake value (SUV_{peak}) (Wahl *et al* 2009), which was found to be more sensitive to quantitative changes due to respiratory motion than SUV_{mean} but less influenced by increased noise in gated images than SUV_{max}. Relative differences in SUV_{peak} between free-breathing static and respiratory-gated PET images, $\%\Delta SUV_{peak}$, were calculated across patient groups and constructed as a function of several independent respiratory trace features. The final model was then validated for predictive power and robustness to sample size.

2. Materials and methods

In this study, respiratory motion features were characterized by a regression model to predict changes of a PET imaging metric between free-breathing static and respiratory-gated images, i.e., $\%\Delta SUV_{peak}$. The predictive model was built iteratively over several steps, spanning respiratory pattern post-processing, feature extraction and selection, model generation, and final model validation against $\%\Delta SUV_{peak}$. Figure 1 outlines the flowchart for building the prediction model from respiratory trace features to predict the measured differences in SUV_{peak} PET metric.

2.1. Experimental design and data acquisition

Twenty-two lung and liver cancer patients underwent PET/CT examinations, during which a time series of abdominal displacement as a respiratory motion surrogate were collected with the Real-time Position Management™ (RPM) (Varian Inc., Santa Clara, CA) infrared camera and reflective marker block system. Retrospective patient data review and analysis was

conducted under approval from the appropriate Institutional Review Board and in compliance with the Health Information Privacy and Portability Act.

Patients were injected with 10.0 ± 0.8 mCi (MEAN±SD)of 2-deoxy-2-[^{18}F] fluoro-D-glucose (FDG), a surrogate of glucose metabolism, and scanned on a Discovery STE^{TM} PET/CT scanner (GE Healthcare, Waukesha, WI) 60 min post-injection. Two-dimensional whole-body PET list-mode data were acquired over 7 min per bed position. Static sinograms were generated via conventional rebinning of time-averaged list-mode data. Gated sinograms were generated retrospectively by sorting the same list-mode data into adaptive 20% amplitude gates determined between each set of consecutive RPM triggers, known as cycle-based quiescent period gating (QPG) (Liu *et al* 2010).

No trigger rejection was enforced on the respiratory-gated sinograms. The QPG bin utilized roughly half of all detected coincidence events and was designed to compensate for cycle-to-cycle variation in respiratory amplitude. All static and gated images were attenuation-corrected with helical or phase-averaged cine CT, reconstructed with ordered subset expectation-maximization over 2 iterations and 28 subsets, filtered with a 6 mm wide Gaussian post-filter, and sampled onto a grid of 3.65 mm $\times$ 3.65 mm $\times$ 3.27 mm voxels.

Post-acquisition and reconstruction, static and gated PET images were quantified within a region-of-interest completely containing a single lesion to minimize the impact of individual patients biasing the population. In the minority of cases which presented with multiple lesions, only the lesion with highest FDG uptake was selected. The PET measure SUV_{peak} was defined by the average pixel value within a 1 cm^3 sphere. The SUV_{peak} of a given patient image was found by moving the sphere within a user-selected bounding box encompassing entire FDG avid lung lesion until the average voxel uptake across the sphere was maximized (Wahl *et al* 2009). The percentage change in SUV_{peak} between gated and static PET images for each patient was recorded as the dependent variable to which the prediction model was tuned to and compared against.

The raw respiratory traces were sampled at 30 Hz and acquired over times ranging from 15–45 min. The 22 patients displayed a tremendous amount of inter-patient and intra-patient variability, including differing periodic and aperiodic characteristics, which is illustrated in figure 2. In order to explain the variation seen in each patient, predictive associations were sought out between respiratory pattern features and image metric improvement in respiratory-gated PET/CT relative to static PET/CT. The respiratory features were extracted and tested as independent variables for the prediction model.

2.2. Fourier spectrum analysis and signal denoising

Periodicity was a significant distinctive characteristic of respiratory trace patterns. However, raw motion signal of respiratory traces were heavily contaminated by signal drifting and noises, as shown in figure 3(left). In the left-upper plot of the figure, there are prominent signal drift and other sources of noise that distorts the original signal. To remove these artifacts from the respiratory trace signal, harmonic analysis has been widely employed to cleanse and characterize respiratory trace patterns (Hamalainen and Kettunen 2000, Riviere *et al* 2001). Specifically, peaks of the Fourier spectrum were used to determine the dominating periodic behavior of temporal trajectory of time series signal. The Fourier spectrum of a respiratory trace from a patient in our dataset is also shown in figure 3(middle). The dominant breathing frequency was between 0.3–0.6 Hz. Although another dominant frequency below 0.1 Hz was observed, the associated breathing frequency was out of a physiologically reasonable range. Since a normal breathing frequency was generally unlikely to exceed 0.1–1 Hz, this range was used to enforce a minimum and maximum threshold for the respiration frequency. A

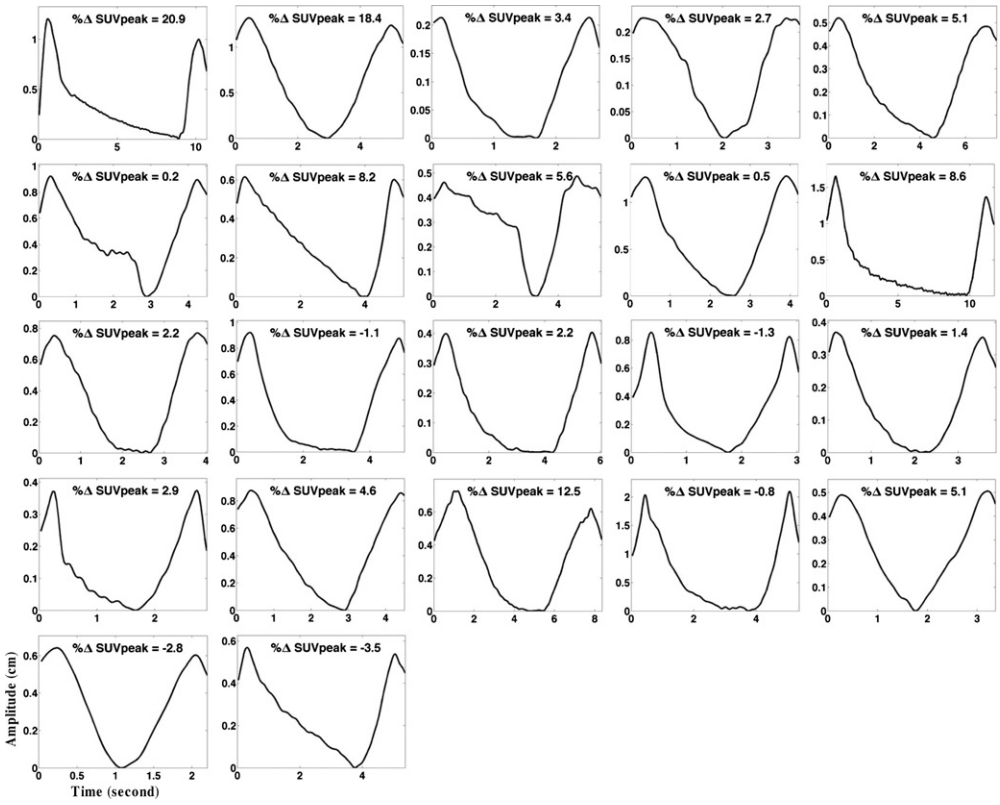


Figure 2. Example of respiratory motion patterns and the corresponding SUV_{peak} percentage changes of the 22 patients. The respiratory pattern shapes vary significantly across the patients.

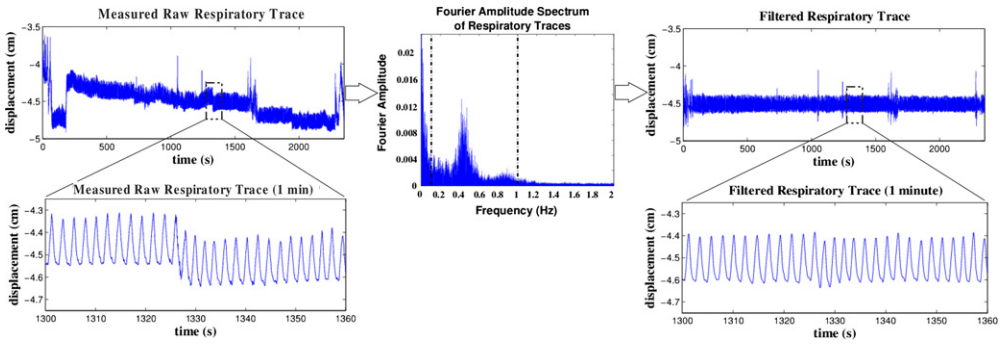


Figure 3. A respiratory trace before and after Fourier transform-based signal cleansing which only retained the signal within the frequency band of [0.1, 1]Hz, which covers the most physiologically reasonable frequencies for respiratory motions. The plot in the middle shows the Fourier spectrum of the respiratory traces. The 'cleansed' respiratory motion traces were reconstructed by an inverse Fourier transform of the frequency components within 0.1–1 Hz.

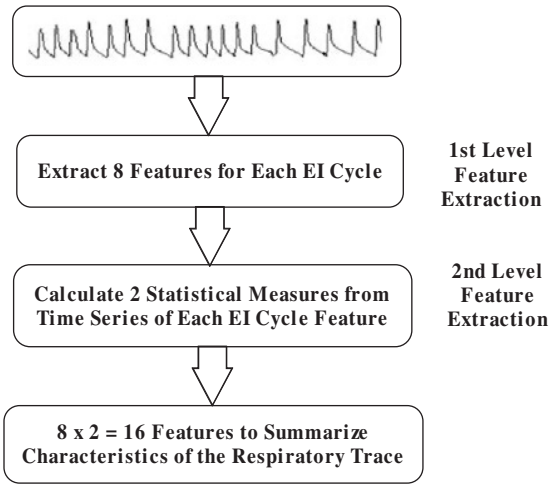


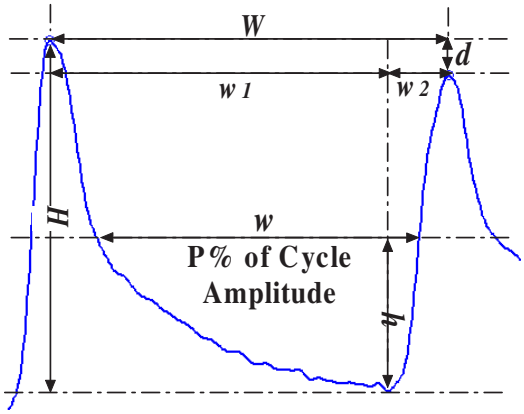
Figure 4. Flowchart of the two-level feature extraction procedure. A total number of 16 features are extracted for each respiratory trace of a patient.

Fourier filter was employed to eliminate physiologically unreasonable frequency components and reconstruct respiratory motion traces. Fourier Transform was first applied to obtain the frequency spectrum of a respiratory motion time series. Subsequently the 'cleansed' respiratory motion trace was reconstructed by an inverse Fourier transform of the frequency components within 0.1–1 Hz, as shown in figure 3(right). This approach removes the low frequency components (e.g., signal drifts and body movements) and high frequency components (e.g., electronics or sensory noises) from the raw respiratory trace, while retaining the useful range of respiratory signal. All subsequent computational experiments of the feature extraction technique and regression analysis in this study were based on the 'cleansed' respiratory traces.

2.3. Feature extraction

Characterization of respiratory trace patterns was a vital step of our approach to build a prediction model for respiratory-gated PET imaging quantification. A two-level feature extraction strategy was used to characterize temporal patterns of respiratory traces. Given a respiratory trace, a set of morphological features of each signal breathing cycle were extracted. Subsequently, the statistical and time-variation properties of the extracted morphological features were quantified. The flowchart of the two-level feature extraction procedure is illustrated in figure 4.

2.3.1. Feature extraction of respiratory morphology. Morphological characteristics of breathing cycles were first analyzed. A respiratory cycle can be defined as a single cycle of inhalation and exhalation. Since respiratory motion trajectory of inhalation and exhalation were not necessarily symmetric, two types of periodic breathing cycles were identified: exhale–inhale (EI) cycle and inhale–exhale (IE) cycle. The EI cycle was defined by a peak-to-peak period, starting from one maximum inspiration to the next maximum inspiration, to complete one respiratory cycle. The IE cycle was defined by a valley-to-valley period, starting from one maximum expiration (exhalation) to the next maximum expiration, to complete one respiratory cycle. Since our QPG approach only performs gating for EI cycles during a PET/CT scan



F1 = *h* (P% cutoff amplitude)
F2 = *w* (P% cutoff period width)
F3 = *h* / *w*
F4 = *w* / *W*
F5 = *w 1* / *w 2* (ratio of exspiration and inpiration time periods)
F6 = *H* (cycle amplitude)
F7 = *w 1* + *w 2* (cycle time period)
F8 = *d* / *H* (ratio of end-inpiration drift and the cycle amplitude)

Figure 5. The eight morphological features of an EI breathing cycle given an amplitude gating threshold P%.

(Liu *et al* 2010), only EI cycle features were considered to be most influential on the efficacy of the QPG approach. Thus in this study, only respiratory pattern features based on EI cycle were extracted and used in our prediction model.

The respiratory gating method extracted PET image data below a certain amplitude threshold in each breathing cycle. The 'gated' respiratory curve patterns of breathing cycles in each respiratory trace were deemed to provide meaningful characteristics that may be linked with improvements of imaging quality. Given an amplitude gating threshold P% of the maximum amplitude during a breathing cycle, four metrics to characterize the gated respiratory motion curve patterns were proposed. In this study, four values of the amplitude gating threshold P% (20%, 30%, 40%, and 50%) were tested, as these values have been used as gating thresholds for the QPG method (Liu *et al* 2010). As illustrated in figure 5, given a gating threshold P%, the four pattern features were defined as follows.

- $F_1(P\%) = h$, amplitude of the cutoff curve at the amplitude threshold of $P\%$.
- $F_2(P\%) = w$, period of the cutoff curve at the amplitude threshold of $P\%$.
- $F_3(P\%) = h/w$, ratio between cutoff amplitude h and cutoff period w.
- $F_4(P\%) = w/W$, ratio between cutoff period and cycle duration, where W was the time period of the breathing cycle.

In addition to the gated-curve pattern features, four additional features quantified the morphology of entire EI cycle patterns. As shown in figure 5, the four morphological features of an entire breathing cycle were defined as follows.

- $F_5 = w_1/w_2$, ratio of expiration and inspiration time.
- $F_6 = H$, cycle amplitude of the breathing cycle.
- $F_7 = w_1 + w_2$, cycle duration of the breathing cycle.
- $F_8 = d/H$, ratio of end-inspiration drift and cycle amplitude.

The relative timing of expiration and inspiration (F_5), cycle amplitude (F_6), and cycle time period (F_7) were commonly used in prior respiratory pattern analysis (Strauss-Blasche *et al* 2000, Tobin *et al* 1983, Tobin 1992). In addition, prior investigations on signal processing of respiratory traces have revealed that end-expiration or end-inspiration displacement can also be an important metric with which to group patients. Based on end-expiration displacement, patients were be grouped into three broad categories: periodic breathers with reproducible

end-expiration displacement, periodic breathers with normal distributions of end-expiration displacement, and chaotic breathers (Liu *et al* 2009). However, it was still unknown whether these patient classification schemes carried associations to quantitative differences in PET image parameters between static free-breathing and respiratory-gated acquisitions. Thus, we introduced the feature F_8, the relative relation of end-expiration/inspiration displacement drift and cycle amplitude, to represent characteristics of end-expiration/inspiration displacement in EI breathing cycles.

2.3.2. Statistical feature extraction. Each respiratory trace contained hundreds of breathing cycles as each PET scan lasted 45–60 min. Eight morphological features were extracted from each EI cycle. Thus for a respiratory trace, each morphological feature had hundreds of observation values. Additionally, statistical analysis on these feature distributions was performed and summarized the characteristics of the entire respiratory time series trace using a much smaller feature dimension. In particular, numerous statistical measures such as mean, standard deviation, skewness, kurtosis, entropy, range, and maximum/minimum value were considered. However, features were not selected directly from all the possible statistical measures together, since that would incur serious problems in feature selection when the number of variables is much larger than the sample size (Fan and Lv 2010). Instead a group selection structure methods was employed to pick up only two statistical measures at a time and then a stepwise feature selection approach was performed, which is discussed in the next section. Two statistical measures, standard deviation and entropy, were selected to report in this paper since they provided clear interpretations of physiological characteristics of respiratory patterns, namely lower order variability due to periodic breathing and higher order variability due to random breathing. Given a time series of a morphological feature $F = [f_1, f_2, \ldots, f_p]$, where the subscript indicates the index of a breathing cycle, the seven statistical measures were calculated as follows.

- Standard deviation: $SF_1 = \sqrt{\frac{\sum_1^p (f_i-\mu)^2}{p-1}}$, where μ was the feature mean with $\mu = \sum_1^p f_i/p$. Standard deviation represented how much variation existed from the average level. Here for each extracted EI feature from a patient, the standard deviation indicates the feature variations over all the breathing cycles in the respiratory trace of the patient. A low standard deviation indicates that the feature values tend to be very close to the average; and a high standard deviation indicates that the EI feature values were spread out over a large range of values.
- Entropy: $SF_2 = -\sum_{i=1}^p P(f_i)\log_b P(f_i)$, where $P(f_i)$, the probability of the value of f_i from the estimated probability mass function $P(F)$ from samples; b was the base of the natural logarithm (i.e. $b = e \approx 2.718\,28$). Entropy was defined as the uncertainty in a random variable in information theory. For an extracted EI feature of a patient, if the feature values were more deterministic (or stable) over breathing cycles in the respiratory trace of the patient, the entropy was low (close to 0). Otherwise, the entropy of this feature was high (close to 1), which indicated the EI feature was more irregular and likely to change significantly from one breathing cycle to another.

2.4. Regression analysis and prediction model

The regression analysis explored the relationships between the imaging variable to be predicted and a set of potential predictor variables from the respiratory pattern. In particular, multiple linear regression with linear functions of a set of predictor variables was used for prediction since it is mathematically well established and usually easy to interpret. In this study, multiple

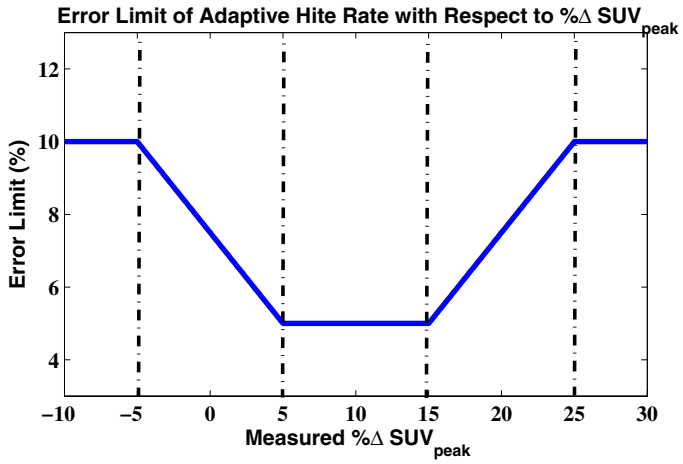


Figure 6. Error limit of the adaptive hit rate with respect to measured $\%\Delta\text{SUV}_{\text{peak}}$.

linear regression was applied to construct a prediction model between quantitative imaging improvement ($\%\Delta\text{SUV}_{\text{peak}}$) and the feature variables extracted from the respiratory motion traces of the 22 patients. To construct a valid regression model for prediction without over-fitting and over-complexity, a feature selection step was utilized to prune unnecessary predictor variables and only keep the most important predictor variables in the model based on certain prediction performance evaluation. The prediction performance measure, the feature selection and the model validation will be discussed in the following subsections.

2.5. Prediction performance measure

A performance measure was introduced, called *hit rate*, to evaluate prediction performance of our regression model. The hit rate was defined as follows:

Definition 1. *Hit rate: given a tolerable error limit, say p%, for each subject, if the predicted value was within actual* $\%\Delta\text{SUV}_{\text{peak}} \pm p\%$*, it was counted as a correct prediction. The hit rate was the percentage of correct predictions over all predictions.*

The error limit was either fixed or adaptive with respect to $\%\Delta\text{SUV}_{\text{peak}} \pm p\%$. Thus, fixed hit rate and adaptive hit rate were defined respectively. A fixed hit rate was defined by a fixed value of error limit. A prediction was considered correct if the predicted value was within the error limit of the measured $\%\Delta\text{SUV}_{\text{peak}} \pm p\%$. In this study, the fixed error limit values were set to 3%, 5%, and 7%, corresponding to hit(3%), hit(5%), and hit(7%), respectively.

In addition, the increases of SUV_{peak} around 10% was used to determine the effectiveness of the gating approach, as this magnitude was considered to reside above the noise threshold (Efron 1983). The gated method was considered to be effective if the increase of SUV_{peak} was higher than 10%. Thus, it was practically important to make accurate predictions around 10% to provide useful decision-making information for physicians to decide whether to use gated approach. To meet this practical consideration, an adaptive hit rate was designed to evaluate prediction performance. Lower error tolerance was assigned between 5% and 15% of $\%\Delta\text{SUV}_{\text{peak}} \pm p\%$ and allowed larger prediction errors in other areas. In particular, the adaptive hit rate was defined as follows (see figure 6),

- for $5\% \leqslant \%\Delta\mathrm{SUV}_{\mathrm{peak}} < 15\%$, the error limit was 5%. A prediction within 5% was considered as a correct prediction;
- for $\%\Delta\mathrm{SUV}_{\mathrm{peak}} \geqslant 25\%$ and $\%\Delta\mathrm{SUV}_{\mathrm{peak}} < -5\%$, the error limit was 10%. A prediction within 10% was considered as a correct prediction;
- for $-5\% \leqslant \%\Delta\mathrm{SUV}_{\mathrm{peak}} < 5\%$ and $15\% \leqslant \%\Delta\mathrm{SUV}_{\mathrm{peak}} < 25\%$, the error limit increased linearly from 5% to 10% in the $\%\Delta\mathrm{SUV}_{\mathrm{peak}}$ range of $[-5\%, 5\%]$ and $[15\%, 25\%]$, respectively.

2.6. Stepwise feature selection

Eight morphological features were extracted for each EI cycle, and each morphological feature had two statistical measures. Thus, each respiratory trace contained $(8) \times 2 = 16$ features as demonstrated in figure 4. Although these 16 features described a respiratory trace in great detail, it was undesirable to use them all in a prediction model in order to avoid over-fitting, as the number of features was larger than the number of samples. In addition, a regression model with too many variables would be hard to interpret and inconvenient to use in clinical practice. Not all of the extract features were necessarily informative to respiratory-gated PET imaging quality quantification. A prediction model only with a few informative features would result in more interpretability, shorter training times, and enhanced generalization with reduced over-fitting.

A stepwise feature selection approach was implemented, which is a popular and widely used approach in statistics to select variables for regression models (Draper and Smith 1998, Miller 2002, Hocking 1976). Starting with no variables (features) in the model, it selects and removes predictive variables in a regression model automatically by a sequence of statistical significant test, such as F-tests, t-test, Akaike information criterion (AIC), etc. The stepwise feature selection procedure was based on the p-value of F-test, which is described as follows.

- Starting with no feature, a variable was added or removed according to the p-value of the F-statistic in F-test of regression coefficients. The maximum p-value for a variable to be added was 0.05; The minimum p-value for a term to be removed was 0.10.
- The procedure continued iteratively until the regression model could not be improved based on the F-test.
- The selected features were those in the final regression model after stepwise selection.

2.7. Predictability evaluation framework by cross-validation

In order to reduce the bias of training and testing data, cross-validation techniques have been extensively used as a method to estimate the generalization error based on 'resampling' by assessing how well the regression model obtained in the training phase perform on future unseen data in the testing phase. The n-fold cross-validation usually divided a dataset into n subsets of (approximately) equal size. Each time a subset was left-out, a regression model was trained on $(n-1)$ subsets and tested on the left-out subset. The procedure repeats until all the subsets have been left-out and tested once (Efron 1983). In this study, due to the limited respiratory traces available (only 22), n-fold cross-validation was not appropriate because the various folds of the training and the fold of the testing would include many respiratory traces that were drawn from the same patients. This situation would increase the likelihood of bias and, thereby, would artificially increase the prediction accuracy. To avoid such a situation, a leave-one-patient-out cross-validation methodology was applied, which trained a regression prediction model using respiratory traces of 21 patients, and tested the obtained

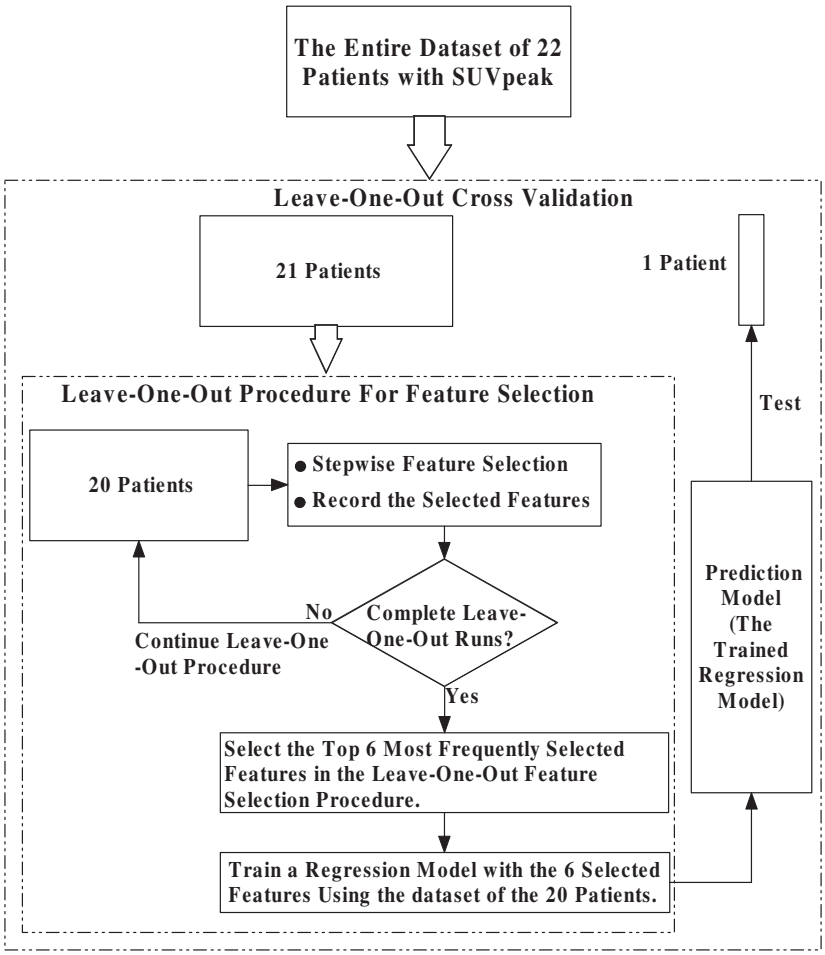


Figure 7. Flowchart of the leave-one-out cross-validation procedure.

prediction model to predict the value of $\%\Delta SUV_{peak}$ of the left-out patient. This procedure was repeated 22 times for the 22 patients. The leave-one-patient-out cross-validation methodology removed the potential for assessment bias. The flowchart of the implemented leave-one-out cross-validation procedure is shown in figure 7.

In addition to training our regression model to optimize regression parameters, respiratory features must also be selected optimally for inclusion in the regression models. Since the patient sample size (22 patients) was not high relative to the number of candidate variables (16 extracted features), the traditional information criteria such as Bayesian information criterion and AIC were not appropriate as the stopping rules in the stepwise searches of small sample sizes would not be strict enough (Kadanea and Lazara 2004). Thus, as illustrated in figure 7, a nested leave-one-out cross-validation was employed, in which a leave-one-out feature selection procedure was performed to validate the goodness of selected features under each cross-validation fold. Particularly, in each fold of leave-one-out cross-validation of the regression model, our training dataset contained 21 subjects, whereby one (subject) cross-validation was utilized to train and test our prediction model 21 times (folds). In each fold, data from 20 subjects were used to select features and train our final prediction model, which

Table 1. The leave-one-patient-out cross-validation results with respect to four cycle-gating threshold levels (20%, 30%, 40%, and 50%).

Cycle gating Threshold (P%)	Leave-one-out cross-validation				
	Correlation	Hit(3%)	Hit(5%)	Hit(7%)	Hit(adapt)
20%	0.76	0.50	0.73	0.95	0.77
30%	**0.88**	**0.59**	**0.95**	**0.95**	**0.95**
40%	0.80	0.41	0.77	0.95	0.95
50%	0.67	0.36	0.73	0.91	0.86

was tested on the one left-out subject. To select features and train our model, a stepwise feature selection approach was performed for each subset of 20 subjects 20 times using a leave-one-out cross-validation. The features selected in different subsets were different, and the numbers of selected features were also different. These sets of selected features were investigated for different subject subsets, and observed that, in most cases, there were six features that were most frequently selected in different sets of selected features. Thus, six features remained as a final robust feature set used to construct regression prediction models in this study.

3. Experimental results

3.1. Evaluation of the predictability of $\%\Delta SUV_{peak}$

The leave-one-out cross-validation results with respect to four cycle-gating threshold levels are summarized in table 1. Four values of amplitude gating threshold (20%, 30%, 40%, and 50%) were employed in feature extraction. The best cross-validation performances were achieved by using the cycle-gating threshold 30%. The hit rates were hit(3%) = 0.59, hit(5%) = 0.95, hit(7%) = 0.95, and hit(adapt) = 0.95. Figure 8 shows the measured and predicted values of $\%\Delta SUV_{peak}$ for the 22 patients. The correlation coefficient between the measured $\%\Delta SUV_{peak}$ and the predicted values was 0.88. The predicted $\%\Delta SUV_{peak}$ values were highly correlated with the actual measured values. Since the trained prediction models in cross-validation did not use any information from the left-out testing patients, the high testing prediction accuracies and high correlation indicated the feasibility of predicting value changes in the imaging quality metric SUV_{peak} using only motion pattern features from respiratory traces.

3.2. Training and robustness analysis of prediction models

The cross-validation results confirmed the possibility to predict SUV_{peak} changes using few variables. However, the trained regression models in different cross-validation runs could produce highly variable results as different subsets of features could be selected. In clinical practice, it would be desirable for physicians to have one prediction model with the most important variables to make predictions and assist their decision-making process. Only using the EI feature set, the selected features in all the 22 cross-validation experiments were recorded to pick the top six most frequently selected features as a robust set of important features for prediction of SUV_{peak} changes. The six selected features are listed in table 2 and include entropy of cutoff cycle amplitude at gating threshold of 30%, entropy of cutoff cycle duration at gating threshold of 30%, standard deviation of the ratio between cutoff amplitude and cutoff period at gating threshold 30%, standard deviation of the ratio between cutoff period at gating threshold 30% and cycle duration, standard deviation of

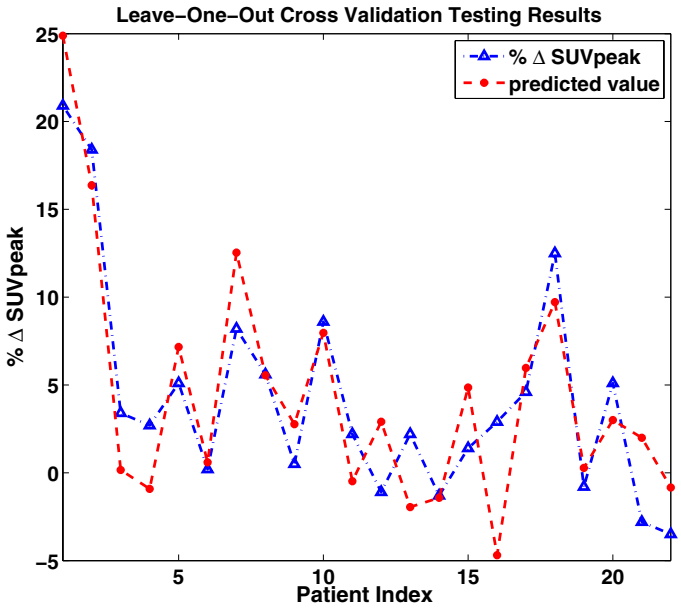


Figure 8. The measured and predicted $\%\Delta SUV_{peak}$ for the 22 patients in the leave-one-patient-out cross-validation. In each cross-validation run, a regression model was obtained from 21 patients and tested on the left-out patient. The testing regression model did not use any information from the left-out patient. The testing performances are hit(3%) = 0.59, hit(5%) = 0.95, hit(7%) = 0.95, hit(adapt) = 0.95.

Table 2. The selected prediction variables by stepwise feature selection from EI cycle feature set.

Selected respiratory motion features from EI cycles	Selected statistical features	Variable denotation
F1: cutoff cycle amplitude at gating threshold of 30%	SF2: entropy	X_1
F2: cutoff cycle duration at gating threshold of 30%	SF2: entropy	X_2
F3: ratio of cutoff amplitude and cutoff period at gating threshold 30%	SF1: standard deviation	X_3
F4: ratio of cutoff period at gating threshold 30% and cycle duration	SF1: standard deviation	X_4
F7: cycle duration	SF1: standard deviation	X_5
F8: ratio of end-inspiration drift and cycle amplitude	SF1: standard deviation	X_6

cycle duration, and standard deviation of the ratio between end-inspiration drift and cycle amplitude. With the 6 selected variables, the regression model was trained from the 22 patients to predict SUV_{peak} value changes. To avoid potential bias in prediction validation, a leave-N-out strategy determined the regression coefficient values in the final prediction model. Specifically, a set of regression coefficients was obtained in each leave-N-out run, where $N = 1$, 2, and 3 in the experiments. The coefficients of the final prediction model were the averaged coefficients over all leave-N-out runs. Table 3 shows three prediction models whose regression coefficients were averaged from leave-one-out, leave-two-out, and leave-three-out experiments, respectively. The testing hit rates of the 3 models on the 22 patients are also reported in the table 3. Though slight variations in regression coefficients were observed, the

Table 3. Three prediction models with the six selected prediction variables obtained by averaging over leave-one-out, leave-two-out, and leave-three-out procedure, respectively.

Model 1: averaged over leave-one-out	$\%\Delta SUV_{amp} = -42.73 + 18.57X_1 + 0.18X_2 - 116.40\,X_3 - 0.11\,X_4 - 17.18\,X_5 + 9.80\,X_6$
Model 2: averaged over leave-two-out	$\%\Delta SUV_{amp} = -42.89 + 18.72X_1 + 0.18X_2 - 116.95\,X_3 - 0.11\,X_4 - 17.19\,X_5 + 9.83\,X_6$
Model 3: averaged over leave-three-out	$\%\Delta SUV_{amp} = -43.00 + 18.86X_1 + 0.18X_2 - 117.54\,X_3 - 0.11\,X_4 - 17.19\,X_5 + 9.86\,X_6$
Testing prediction accuracy	Hit(3%) = 0.95, Hit(5%) = 1, Hit(7%) = 1, Hit(adapt) = 1 (The testing prediction accuracies were same for model 1, 2, and 3.)

Table 4. Testing prediction performance for the 22 patients using the prediction model which was averaged over the leave-one-out procedure: $\%\Delta SUV_{peak} = -42.73 + 18.57X_1 + 0.18X_2 - 116.40X_3 - 0.11X_4 - 17.18X_5 + 9.80X_6$. The hit rates were hit(3%) = 0.95, hit(5%) = 1.00, hit(7%) = 1.00, and hit(adapt) = 1.00.

Patient index	Measured $\%\Delta SUV_{peak}$	Predicted $\%\Delta SUV_{peak}$	Prediction error	Hit (3%)	Hit (5%)	Hit (7%)	Hit (adapt)
1	20.9	20.87	0.03	1	1	1	1
2	18.4	17.25	1.15	1	1	1	1
3	3.4	1.49	1.91	1	1	1	1
4	2.7	−0.27	2.97	1	1	1	1
5	5.1	6.60	−1.50	1	1	1	1
6	0.2	0.47	−0.27	1	1	1	1
7	8.2	10.69	−2.49	1	1	1	1
8	5.6	5.60	0.00	1	1	1	1
9	0.5	2.56	−2.06	1	1	1	1
10	8.6	7.69	0.91	1	1	1	1
11	2.2	0.16	2.04	1	1	1	1
12	−1.1	2.58	−3.68	0	1	1	1
13	2.2	−0.39	2.59	1	1	1	1
14	−1.3	−1.41	0.11	1	1	1	1
15	1.4	2.50	−1.10	1	1	1	1
16	2.9	1.55	1.35	1	1	1	1
17	4.6	5.76	−1.16	1	1	1	1
18	12.5	10.46	2.04	1	1	1	1
19	−0.8	0.01	−0.81	1	1	1	1
20	5.1	3.74	1.36	1	1	1	1
21	−2.8	−0.77	−2.03	1	1	1	1
22	−3.5	−2.37	−1.13	1	1	1	1
Prediction accuracy				0.95	1.00	1.00	1.00

3 models achieved the same prediction accuracy when they were tested on the 22 patients. The hit rates were hit(3%) = 0.95, hit(5%) = 1.00, hit(7%) = 1.00, and hit(adapt) = 1.00. Table 4 shows the measured and predicted values of $\%\Delta SUV_{peak}$ for the 22 patients using the prediction model averaged over the leave-one-out procedure (model 1 in table 3). The table also summarizes each patient's testing prediction results with respect to the four prediction error limits (3%, 5%, 7% and adaptive). A patient was marked as 0 if the prediction error was larger than the error limit; otherwise, the patient was marked as 1. One can see that only patient 12 was mispredicted using the error limit of 3% and all others were correctly predicted with errors within the corresponding error limits. The measured and predicted values of $\%\Delta SUV_{peak}$ are also plotted in figure 9. As shown in the figure, the predicted $\%\Delta SUV_{peak}$ values are

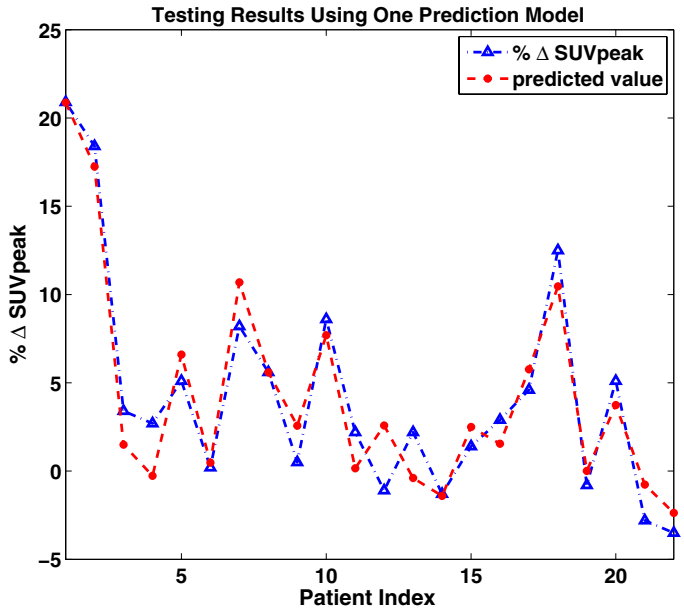


Figure 9. The measured and predicted $\%\Delta\mathrm{SUV}_{\mathrm{peak}}$ for the 22 patients using the prediction model which was averaged over the leave-one-out procedure: $\%\Delta\mathrm{SUV}_{\mathrm{peak}} = -42.73 + 18.57X_1 + 0.18X_2 - 116.40X_3 - 0.11X_4 - 17.18X_5 + 9.80X_6$. The testing prediction performances are hit(3%) = 0.95, hit(5%) = 1, hit(7%) = 1, hit(adapt) = 1.

highly correlated with the measured values. These testing prediction results confirmed that the prediction model with six variables was sufficient to prospectively predict $\mathrm{SUV}_{\mathrm{peak}}$ changes due to use of gating.

For a set of ideal prediction variable, the measured regression coefficients are expected to be robust in different cross-validation runs. Thus, the robustness of the regression coefficients across leave-N-out cross-validation runs was evaluated. The regression coefficients and their standard deviations cross-validation runs are shown in table 5. The variation ratio between standard deviation and the absolute value of the regression coefficients were employed to assess the model stability. The coefficient variation ratios were increasing slightly from leave-one-out to leave-three-out experiments. In the leave-three-out experiment, the variation ratios of all regression coefficients were less or equal than 0.20 except X_4 which reached as high as a variation ratio of 0.28. Overall, the regression coefficient variations were small compared to their amplitudes. This outcome indicated that the selected six prediction variables were a robust subset of features to predict PET imaging quality improvements in terms of $\%\Delta\mathrm{SUV}_{\mathrm{peak}}$.

4. Discussion

This work investigated a high dimensional space of respiratory motion pattern features with the aim of predicting improvements in respiratory-gated PET/CT imaging metrics relative to free-breathing static PET/CT. By using only respiratory pattern characteristics, a model of six variables (features) was developed to predict PET image quality improvements due to gating in terms of relative change in $\mathrm{SUV}_{\mathrm{peak}}$, $\%\Delta\mathrm{SUV}_{\mathrm{peak}}$, which would allow for individual patient assessment prior to PET/CT acquisition.

Table 5. Robustness of the regression coefficients with respect to leave-one, -two, and -three-out model averaging procedures.

	Model 1: averaged over leave-one-out			Model 2: averaged over leave-two-out			Model 3: averaged over leave-three-out		
Prediction variable	Coefficient	Std.	Std./abs(coefficient)	Coefficient	Std.	Std./abs(coefficient)	Coefficient	Std.	Std./abs(coefficient)
Intercept	−42.73	2.02	0.05	−42.89	2.94	0.07	−43.00	3.77	0.09
X_1	18.57	2.04	0.11	18.72	3.00	0.16	18.86	3.84	0.20
X_2	0.18	0.01	0.07	0.18	0.02	0.10	0.18	0.02	0.14
X_3	−116.40	6.57	0.06	−116.95	9.71	0.08	−117.54	12.46	0.11
X_4	−0.11	0.01	0.12	−0.11	0.02	0.20	−0.11	0.03	0.28
X_5	−17.18	0.66	0.04	−17.19	1.00	0.06	−17.19	1.32	0.08
X_6	9.80	0.40	0.04	9.83	0.57	0.06	9.86	0.73	0.07

The averaged prediction accuracy was higher than 90% over the population of 22 patients when using the leave-one-patient-out cross-validation. While other investigators have built predictive models of respiratory motion patterns (McCall and Jeraj 2007, Ren *et al* 2007, Riaz *et al* 2009), particularly for application to respiratory-gated radiation therapy, none have explicitly linked them to changes in quantitative molecular imaging with PET/CT. Our prediction model would enable clinicians to efficiently and accurately evaluate uncertainties due to respiratory motion prior to PET/CT acquisition, which may have an impact on patient care strategies. For example, the following patient-specific workflow could be enacted.

- With patient lying on PET/CT scanner bed, acquire representative sample respiratory trace (e.g., < 5 min).
- Calculate respiratory features and predict changes in SUV_{peak} metric due to motion.
- Determine whether quantitative changes exceed predefined threshold (e.g., 10% change in $SUV_{peak} \pm 95\%$ confidence interval of model prediction).
- If prediction exceeds threshold, patient would likely to benefit from respiratory-gated PET/CT and one could proceed with the image acquisition under free-breathing conditions.
- If predicted change in image metric is not significant, patient is not likely to benefit from respiratory-gated PET/CT due to the respiratory pattern. Instead, one could implement more invasive motion management strategies such as abdominal compression or active breathing control prior to image acquisition.

Interestingly, this workflow relied only on respiratory pattern parameters to make a prediction on changes in PET quantification. When prior CT images were available for review, the model could in principle incorporate additional factors such as lesion location. However, the gains in predictive power when including prior CT factors were not observed to be significant and were therefore omitted from the final model. This may be due in part to sufficient correlation between the external abdominal displacement and the internal tumor motion, which does vary between patients. Nevertheless, the preliminary model validation was robust for this cohort of patients with thoracic and abdominal lesions of differing location, meaning that respiratory patterns alone appeared to account for a large percentage of the variance in observed PET image parameters.

A crucial component of the prediction model was its construction from generalized statistical regression methods. Since it was independent of both the respiratory pattern measuring device and PET/CT scanner, the analysis could be replicated at other clinical and academic centers. Specifically, a different set of prediction model coefficients could be derived when using a different combination of PET/CT system and respiratory measurement device. The extracted respiratory pattern features would potentially vary for predicting imaging changes generated by different respiratory gating algorithms, as quantitative improvements in phase-gated PET images may depend on measures of phase shifts due to variable breathing cycle durations while improvements in fixed amplitude-gated PET images may depend on measures of baseline displacement drifts. However, the sensitivity of such models to accurately predict changes in individual patients would likely remain highest during the end-exhale portion of the breathing cycle at times of lower residual motion and lowest during fast phase transitions to peak inhale at times of high residual motion. Different SUV metrics, such as SUV_{max} or SUV_{mean}, could also be chosen to fit an alternate statistical model. For example, an SUV_{max} model would tend to have high inter-patient variability in the prediction accuracy due to the influence of noise on image parameter changes. On the other hand, an SUV_{mean} model would yield lower correlation between respiratory pattern parameters and image parameters as the influence of motion on SUV_{mean} is dependent upon the region-of-interest that defines the lesion. The SUV_{peak} prediction model balances the impact of noise with sufficient sensitivity

to respiratory motion. Ultimately, the flexibility afforded with this approach enabled the individualization of patient management for respiratory-gated PET/CT that over time could be cross-calibrated between scanners, respiratory motion surrogates, and gating methods in multi-institution trials.

The study was limited primarily by a small patient sample size of only 22 patients. Future work includes a completely independent validation of the prediction model on a second test cohort of patients. This would enable the application of the model as a decision tool, whereby sample respiratory traces for a given a patient could inform the manner by which respiratory motion should be managed during PET/CT acquisition. Alternatively, the tool may be used after a PET/CT acquisition to assess whether a free-breathing static or respiratory-gated image should be reconstructed, as some patient images may not benefit from any form of motion management. Furthermore, the predicted changes in SUV metrics between respiratory-gated acquisition and free-breathing static acquisition may relate an estimate of the expected quantitative uncertainty when assessing routine clinical PET images. For instance, the reported SUV in a region-of-interest could carry a confidence interval over which clinicians could reliably interpret the findings.

5. Conclusion

The results of the prediction model supported the need for improved patient-specific management of respiratory motion during PET/CT acquisition. This work addressed such a need by proposing a comprehensive tool to support clinical decision-making. Increasingly accurate and robust prediction models may pave the way for efficient patient classification and motion uncertainty mitigation, which would assist clinicians when utilizing quantitative PET/CT for therapy target definition and response assessment.

Acknowledgments

This work was supported by NIH grant 5P30 CA015704 and a research contract from GE Healthcare (Waukesha, WI).

References

Aristophanous M, Berbeco R, Killoran J, Yap J, Sher D, Allen A, Larson E and Chen A 2011 Clinical utility of 4D FDG-PET/CT scans in radiation treatment planning *Int. J. Radiat. Oncol. Biol. Phys.* **82** e99–105

Aristophanous M, Yap J, Killoran J, Chen A and Berbeco R 2011 Four-dimensional positron emission tomography: implications for dose painting of high-uptake regions *Int. J. Radiat. Oncol. Biol. Phys.* **80** 900–8

Avril N and Weber W 2005 Monitoring response to treatment in patients utilizing PET *Radiol. Clin. North Am.* **43** 189–204

Bettinardi V, Picchio M, Muzio N D, Gianolli L, Gilardi M C and Messa C 2010a Detection and compensation of organ/lesion motion using 4d-PET/CT respiratory gated acquisition techniques *Radiother. Oncol.* **96** 311–6

Bettinardi V, Picchio M, Muzio N D, Gianolli L, Messa C and Gilardi M C 2010b PET/CT for radiotherapy: image acquisition and data processing *Q. J. Nucl. Med. Mol. Imaging* **54** 455–75

Bettinardi V, Picchio M, Muzio N D and Gilardi M 2012 Motion management in positron emission tomography/computed tomography for radiation treatment planning *Semin. Nucl. Med.* **42** 289–307

Bowen S, Nyflot M, Gensheimer M, Hendrickson K, Kinahan P, Sandison G and Patel S 2012 Challenges and opportunities in patient-specific, motion-managed, and PET/CT-guided radiation therapy of lung cancer: review and perspective *Clin. Transl. Med.* **1** 18

Caldwell C, Mah K, Ung Y, Danjoux C, Balogh J, Ganguli S and Ehrlich L 2001 Observer variation in contouring gross tumor volume in patients with poorly defined non-small-cell lung tumors on CT: the impact of 18FDG-hybrid PET fusion *Int. J. Radiat. Oncol. Biol. Phys.* **51** 923–31

Draper N and Smith H 1998 *Applied Regression Analysis* 3rd edn (Hoboken, NJ: Wiley-Interscience) pp 307–12 (www.wiley.com/WileyCDA/WileyTitle/productCd-0471170828.html)

Efron B 1983 Estimating the error rate of a prediction rule: improvement on cross-validation *J. Am. Stat. Assoc.* **78** 316–31

Fan J and Lv J 2010 A selective overview of variable selection in high dimensional feature space *Stat. Sin.* **20** 101–48 (PMC: 3092303)

Guerra L *et al* 2012 Respiratory gated PET/CT in a European multicentre retrospective study: added diagnostic value in detection and characterization of lung lesions *Eur. J. Nucl. Med. Mol. Imaging* **39** 1381–90

Hamalainen R and Kettunen A 2000 Stability of Fourier coefficients in relation to changes in respiratory air flow patterns *Med. Eng. Phys.* **22** 733–9

Hocking R 1976 The analysis and selection of variables in linear regression *Biometrics* **32** 1–49

Imamura Y *et al* 2011 Prognostic value of SUVmax measurements obtained by FDG-PET in patients with non-small cell lung cancer receiving chemotherapy *Lung Cancer* **71** 49–54

Kadanea J and Lazara N 2004 Methods and criteria for model selection *J. Am. Stat. Assoc.* **99** 279–90

Keall P *et al* 2006 The management of respiratory motion in radiation oncology report of AAPM task group 76 *Med. Phys.* **33** 3874–900

Liu C, Alessio A, Pierce L, Thielemans K, Wollenweber S, Ganin A and Kinahan P 2010 Quiescent period respiratory gating for PET/CT *Med. Phys.* **37** 5037–43

Liu C, Pierce L II, Alessio A and Kinahan P E 2009 The impact of respiratory motion on tumor quantification and delineation in static PET/CT imaging *Phys. Med. Biol.* **54** 7345–62

McCall K and Jeraj R 2007 Dual-component model of respiratory motion based on the periodic autoregressive moving average (periodic ARMA) method *Phys. Med. Biol.* **52** 3455–66

Miller A 2002 *Subset Selection in Regression* 2nd edn (London: Chapman and Hall) (www.crcpress.com/product/isbn/9781584881711)

Nehmeh S A, Erdi Y E, Ling C C, Rosenzweig K E, Schoder H, Larson S M, Macapinlac H A, Squire O D and Humm J L 2002 Effect of respiratory gating on quantifying PET images of lung cancer *J. Nucl. Med.* **43** 876–81 (PMID: 12097456)

Nehmeh S A, Erdi Y E, Rosenzweig K E, Schoder H, Larson S M, Squire O D and Humm J L 2003 Reduction of respiratory motion artifacts in PET imaging of lung cancer by respiratory correlated dynamic PET: methodology and comparison with respiratory gated PET *J. Nucl. Med.* **44** 1644–8 (PMID: 14530480)

Ren Q, Nishioka S, Shirato H and Berbeco R 2007 Adaptive prediction of respiratory motion for motion compensation radiotherapy *Phys. Med. Biol.* **52** 6651–61

Riaz N, Shanker P, Wiersma R, Gudmundsson O, Mao W, Widrow B and Xing L 2009 Predicting respiratory tumor motion with multi-dimensional adaptive filters and support vector regression *Phys. Med. Biol.* **54** 5735–48

Riviere C N, Thakral A, Iordachita I I, Mitroi G and Stoianovici D 2001 Predicting respiratory motion for active canceling during percutaneous needle insertion *EMBS'01: Proc. 23rd Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society* **4** 3477–80

Senan S and Ruysscher D D 2005 Critical review of PET-CT for radiotherapy planning in lung cancer *Crit. Rev. Oncol. Hematol.* **56** 345–51

Strauss-Blasche G, Moser M, Voica M, McLeod D, Klammer N and Marktl W 2000 Relative timing of inspiration and expiration affects respiratory sinus arrhythmia *Clin. Exp. Pharmacol. Physiol.* **27** 601–6

Tobin M 1992 Breathing pattern analysis *Intensive Care Med.* **18** 193–201

Tobin M, Chadha T, Jenouri G, Birch S, Gazeroglu H and Sackner M 1983 Breathing patterns: I. Normal subjects *Chest* **84** 202–5

Vansteenkiste J F *et al* 1998 Lymph node staging in non-small-cell lung cancer with FDG-PET scan: a prospective study on 690 lymph node stations from 68 patients *J. Clin. Oncol.* **16** 2142–9 (PMID: 9626214)

Wahl R, Jacene H, Kasamon Y and Lodge M 2009 From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors *J. Nucl. Med.* **50** (Suppl. 5) 122S–50S

Wong J, Sharpe M, Jaffray D, Kini V, Robertson J, Stromberg J and Martinez A 1999 The use of active breathing control (ABC) to reduce margin for breathing motion *Int. J. Radiat. Oncol. Biol. Phys.* **44** 911–9