Uncertainties in RECIST as a measure of volume for lung nodules and liver tumors

Zachary H. Levine,[a] Adam L. Pintar, John G. Hagedorn, and Charles P. Fenimore
National Institute of Standards and Technology, Gaithersburg, Maryland 20899

Claus P. Heussel
Chest Clinic, University Hospital Heidelberg, Amalinstraße 5, Heidelberg 69126, Germany

(Received 3 January 2012; revised 13 March 2012; accepted for publication 22 March 2012; published 19 April 2012)

Purpose: The authors wish to determine the extent to which the Response Evaluation Criteria in Solid Tumors (RECIST) and the criteria of the World Health Organization (WHO) can predict tumor volumes and changes in volume using clinical data.
Methods: The data presented are a reanalysis of data acquired in other studies, including the public database from the Lung Image Database Consortium (LIDC) and from a study of liver tumors.
Results: The principal result is that a given RECIST diameter predicts volume to a factor of 16 or 10 for the two data sets, respectively, by examining 95% prediction bounds and that changes in volume are predicted only little better: to within a factor of 7 for the liver data. The WHO criteria reduce the prediction bounds by a factor of 1.3 in all cases. Also, the RECIST threshold of 10 mm to measure a nodule corresponds to a transition zone width of a factor of more than 2 in volume for the nodules in the LIDC database.
Conclusions: While the RECIST diameter is certainly correlated with the volume, and similarly for changes in these quantities, the use of the diameter introduces additional variation assuming volume is the quantity of interest. Exactly how much this reduces the statistical power of clinical drug trials is a key open question for future research. [http://dx.doi.org/10.1118/1.3701791]

Key words: X-ray imaging, RECIST, tumor size, volumetric measurement, LIDC database

I. INTRODUCTION

The principal formal method for determining whether cancerous nodules are growing or shrinking is the Response Evaluation Criteria in Solid Tumors (RECIST).[1,2] In the decade or so since its introduction by a committee of American, Canadian, and European cancer specialists, the capabilities of computed tomography (CT) machines has increased considerably, leading to the possibility of widespread adoption of volumetric methods rather than the 1D RECIST measure. Indeed, even as the initial 1D RECIST were issued, careful volume studies of CT phantoms (i.e., reference objects) and patients were performed indicating the ability to measure volumes to a few percent.[3] In the revision to the RECIST standard, Eisenhauer *et al.*[2] discussed the alternative of using the measurement of tumor volume. They cited several studies in concluding that the RECIST measure was comparable to the volume standard. However, as our group noted earlier,[4] these studies required volume changes to be outside of the range of −66% to +73% to be considered significant. This range was required to match the RECIST criteria of partial response or progressive disease, a −30% or +20% change in RECIST diameter, respectively. A required volume change of about 70% may be too large: for example, Lee and coworkers[5] conclude that a 35.6% decrease in volume in gastric lesions after 8 weeks is sufficient to determine pathogenic responders with 100% sensitivity with a 58.8% specificity.

There is current interest in exploring the question of RECIST vs 3D techniques,[6–9] following a long history of the question of 1D vs 3D techniques.[10] Three-dimensional techniques may be more accurate than RECIST.[11] A consensus statement of the International Cancer Imaging Society[12] noted that tumors do not necessarily grow or shrink in a rounded fashion, so a measurement of the longest diameter may not necessarily represent the true response. Mantatzis and coworkers[7] found that the assumption of uniform growth behind the spherical model was only somewhat applicable to liver tumors. In the case of nasopharyngeal cancers, the irregular tumor shape led to RECIST diameters being poorly correlated with volume, although the two-dimensional analogue[13] due to the World Health Organization (WHO) was possibly sufficient.[14] The nonspherical growth pattern of malignant pleural mesothelioma is also a challenge for RECIST.[15] Additionally, studies of early lung cancer tend to use volumetric methods, if only because the RECIST standard explicitly excludes nodules under 10 mm, whereas 5 mm is a more typical minimum diameter for tumors studied for early lung cancer.[16]

Recently, our group has considered the relation between RECIST and volume measurements principally in the context of physical and mathematical ellipsoidal models.[4,17,18] Here, we characterize the relationship between RECIST and volume for the clinical data for a previously published study on liver tumors[19] as well as from the Lung Image Database

Consortium (LIDC).[20] These two data sets are complimentary in which the liver tumors allow us to study the change in volume and change in RECIST values over time, whereas the LIDC data give us full descriptions of the nodules which allow us to reorient nodules before finding the RECIST values. In this study, we only compare RECIST and WHO values with volume on a nodule-by-nodule basis and we do not sum nodules across a scan.

It must be noted that changes in the character of the tumor, e.g., the density or even necrosis cannot be captured by purely dimensional measurements.[15,21] Another interesting question is the number of tumors which should be measured to get a representative understanding of the total tumor burden.[22] Such issues are beyond the scope of this paper.

II. MATERIALS AND METHODS

II.A. Liver data

As described by Heußel and coworkers,[19] the liver data consist of 82 patients with one to five tumors measured at two times, including 198 tumors at the first time and 180 tumors still present at the second time. The study was conducted retrospectively from CT scans collected in the routine course of treatments from data collected using one CT model, namely, a Philips[23] Somatom Brilliance 40 run at a tube voltage of 120 kV, an exposure of 165 mAs, a slice thickness of 3 mm, and a reconstruction interval of 2 mm. The data were anonymized before analysis. The primary cancer sites and frequencies were hepatocellular ($n=36$), colon ($n=24$), rectum ($n=7$), pancreas ($n=5$), and other metastases ($n=10$).

The measurements were the RECIST diameter d, the product of two diameters w according to the criteria of the WHO, and the volume V as determined by the consensus of two radiologists. The radiologists used ONCOTREAT software version 1.2 (Mevis, Bremen).[23] After a manual identification of a lesion, the segmentation was performed automatically. However, it could be corrected interactively. Lesions at both time points were available to the radiologists simultaneously. After the segmentation was finalized, the three values were determined automatically. The study[19] concluded that the RECIST and WHO criteria are of limited use if volume was considered the primary parameter characterizing tumor development. Our goal in this reanalysis is to construct a prediction interval in which $\log V$ is likely to lie for a given value of $\log d$ or $\log w$.

We model the RECIST data with a mixed linear model as follows. Let d_{ij} denote the RECIST value of nodule j in patient i for either time 1 or time 2. We model d_{ij} as

$$\log d_{ij} = \beta_0 + \beta_1 \log V_{ij} + \epsilon_{ij}, \tag{1}$$

with $i = 1,\ldots,N^{patient}$ and $j = 1,\ldots,N_i^{nodule}$. In Eq. (1), β_0 and β_1 are fixed unknown parameters, V_{ij} is the volume for nodule j in patient i, and ϵ_{ij} is a random error. All logarithms are in base 10. The random error, ϵ_{ij}, can be decomposed further as

$$\epsilon_{ij} = P_i + N_{i(j)}, \tag{2}$$

where P_i is the random effect for patient i and $N_{i(j)}$ is the random effect of nodule j nested within patient i.

The construction of prediction intervals (for predicting the inverse relationship, i.e., predicting $\log V$ given $\log d$) and confidence intervals for β_0 and β_1 based on the model described by Eqs. (1) and (2) is carried out in two ways. First, we assume that $P_i \overset{iid}{\sim} N(0,\sigma_P^2)$ independent of $N_{i(j)} \overset{iid}{\sim} N(0,\sigma_N^2)$, where $N(\mu,\sigma^2)$ refers to a Gaussian distribution with mean μ and variance σ^2. The notation "iid" means independent and identically distributed. In this case, the model parameters can be estimated using residual maximum likelihood (ReML).[24] The prediction intervals for β_0 and β_1 are constructed using propagation of error and the normal (Gaussian) distribution. One might argue that a Student's t distribution should be used in place of the normal distribution; however, the most conservative degrees of freedom to use for the Student's t distribution would be $N^{\text{patient}} - 1 = 81$. Since a Student's t distribution with 81 degrees of freedom is very near the standard normal distribution, the difference is ignored. When assuming Gaussian random effects, there are four parameters to estimate, two regression coefficients, β_0 and β_1, and two variance components, σ_P and σ_N.

The second approach is based on estimating β_0 and β_1 by least squares; then, a bootstrap algorithm is used for the construction of confidence and prediction intervals. The bootstrap algorithm is used to avoid specific distributional assumptions on P_i and $N_{i(j)}$, and the algorithm is discussed for the more complex LIDC data. For the liver data, there is insufficient evidence to refute the assumption of a normal distribution for the random effects. However, for the LIDC data, it is clear that a normal distribution is inappropriate for at least one of the random effects. To be consistent, both approaches are applied to both data sets. All bootstrap intervals in this paper pertaining the liver data are constructed from 5000 bootstrap samples.

The change data are modeled in exactly the same way as the static data. For the change data, the response is $\log d_{ij}^{(21)} = \log d_{ij}^{(2)} - \log d_{ij}^{(1)}$, where $d_{ij}^{(t)}$ is the RECIST value for nodule j in patient i at time $t=1, 2$. For both lung and liver data, the WHO data are modeled similarly, under $d \to w$, with all subscripts and superscripts applied.

The results are given in Sec. III.A.

II.B. Lung data

The LIDC database consists of data from 1018 helical thoracic scans from 1010 different patients. At the time we accessed the database in June 2011, a pilot database with 399 scans was available for download.[25] In the LIDC study, every CT scan was read by 4 radiologists drawn from a pool of 12, first independently ("blinded") and then a second time with knowledge of the other three radiologists' markup ("unblinded"). The unblinded markup from each of the four radiologists was made available. Nodules were marked by a radiologist if the measured diameter was at least 3 mm and if it was a nodule in the judgment of a given radiologist.

We downloaded 203 patient files using the selection criterion that the slice thickness be no more than 3 mm. We

found 28 of the 203 files did not contain marked nodules, and another 4 files may have been internally inconsistent. The 171 remaining patient files were read with custom-written code in IDL.[23] Using the LIDC rules,[20] the markup was turned into a set of bitmaps for 1252 nodule readings. (The principal rule is that a nodule is the set of voxels inside, but not including, the markup; interior regions were excluded in some cases.) We use the term "nodule reading" to distinguish this data from "physical nodules," i.e., the nodules in a patient.

Because the nodule readings from the same physical nodule are highly correlated, the statistical method we used required associating nodule readings with physical nodules. Although this information was known for the LIDC study,[20] it was not present in the pilot database. Accordingly, we reconstructed the information from a knowledge of the position and second moments of each nodule reading. The LIDC database has a single co-ordinate system for all nodules for a given scan, so the centroid of the bitmap is meaningful across those nodule readings. We found the 3×3 covariance matrix C (also known as the second moment tensor about the centroid) of each bitmap for each nodule reading. For a given scan, among those nodule readings which came from different radiologists, we added their covariance matrices pairwise and formed the Mahalanobis distance, a positive definite quantity $Z^{(ab)}$ given by

$$Z^{(ab)} = \left(\mathbf{r}^{(ab)T} \cdot C^{-1} \cdot \mathbf{r}^{(ab)}\right)^{1/2}, \tag{3}$$

where $\mathbf{r}^{(ab)} = \mathbf{r}^{(a)} - \mathbf{r}^{(b)}$ is the difference between the centroids of nodule readings a and b, and T indicates the transpose is to be taken. The quantity Z is analogous to the "Z-score" used in statistics but adapted to the anisotropic, 3D case. We found that there was a bimodal distribution of Z values with a sharp minimum for $Z \approx 3$ as seen in Fig. 1. We interpreted $Z^{(ab)} \leq 3$ as meaning a and b were two readings from the same physical nodule and the others as two readings from different physical nodules. To make the formal identification of physical nodules, we formed a graph as follows: (1) nodule readings are vertexes in the graph; (2) bidirectional edges were added to the graph whenever the nodule readings a and b obeyed $Z^{(ab)} \leq 3$, were from the same scan, and linked readings by two different radiologists. Under these terms, the 1252 nodule readings formed 511 disconnected graphs. Of these 511 disconnected graphs, 509 were complete graphs, indicating that 1, 2, 3, or 4 radiologists all marked the same physical nodule. (Our implementation used the graph algorithms of MATHEMATICA.[23]) Of the two other cases, in one case one radiologist identified a small nodule nearby and in the other one radiologist marked a single nodule where the other three marked two nodules; these involved a total of 12 nodule readings, allowing the other 1240 nodule readings to be associated with 509 physical nodules. Our finding of 2 incomplete identifications of 511 physical nodules is consistent with the LIDC finding of 6 incomplete identifications in 2669 nodules.[20] The results presented in Table I for the proportions of nodules identified by one to four radiologists are in good agreement with those from the full LIDC study.

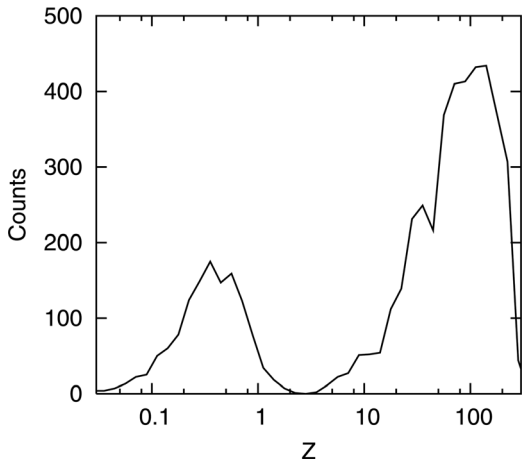

FIG. 1. Histogram of Z, defined in Eq. (3), a normalized distance between two nodule readings. The first peak is interpreted as two nodule readings being associated with the same physical nodule; for the second peak, the two nodule readings are associated with different physical nodules. The peaks are well separated. Counts are for intervals of $\log Z$ of 0.1.

The first task was to derive bitmaps for each nodule from the markup that represents the nodule readings. These bitmaps distinguish voxels that are interior from those that are exterior to a given nodule. The markup for a nodule consists of closed paths in each voxel plane that describe the boundary of the nodule. Within each plane, the bitmap for a nodule is generated using the following steps:

1. Create an initial bitmap that identifies voxels that lie on each markup path. These voxels lie on the boundary of the nodule.
2. Apply the IDL function LABEL_REGION to this bitmap to identify connected regions separated by the boundary voxels.
3. Discard regions that contain known exterior voxels. Any voxel that is beyond the extent of the markup path coordinates is known to be exterior, and because none of the markup paths extend to the edge the voxel plane, we can always find such exterior voxels.
4. Discard all of the boundary voxels identified in the initial bitmap.
5. Label all remaining voxels as interior to the nodule in the final bitmap.

The LIDC markup rules allow the possibility of excluded regions; these regions are identified by the same type of

TABLE I. Proportion of nodules identified by $N = 1,\ldots,4$ radiologists in the LIDC database (Ref. 20) and the present subset of the LIDC database. The two cases in which there was not a 1:1 correspondence between identified nodules are omitted from our data.

N	LIDC	Present
1	29.1%	33.4%
2	18.4%	19.4%
3	17.7%	17.3%
4	34.4%	29.8%
Total	2669	509

closed-path markup. Bitmaps representing these excluded regions are generated exactly as described above. The final nodule bitmaps are modified based on the exclusion bitmaps.

The final result is a set of bitmaps for 1252 nodule readings. The nodule reading was considered to be the union of the included voxels from all of the voxel planes. We made no attempt to consider partial volume effects or to create a smooth surface.

The bitmaps were read by a FORTRAN 95 program written for the present study to determine volume and RECIST values. The volume of each nodule reading was obtained by counting the voxels in the bitmap and multiplying by the volume of each voxel. The RECIST diameter d is defined relative to a scan axis. To calculate this, we find the projection of the lesion's 3D bitmap onto a plane orthogonal to the scan axis, then take the maximum diameter of this projection. A second diameter of the projection, orthogonal to the RECIST diameter, is then derived. The product of the two diameters is the WHO value w.

Different choices of the scan axis result in different RECIST values. Rather than use the single scan axis provided by the measurement, we make the assumption that the orientation of nodules relative to the scan axis is random. Then we found many possible RECIST diameters and WHO values by taking the scan axis to be any of 144, 544, or 2112 directions distributed uniformly throughout a hemisphere. These particular numbers arise because we obtain a uniform sampling of the sphere as follows: each eighth of the sphere is sampled by a triangular array of one quasi-triangular and several quasi-rectangular regions of equal area bounded by constant values of the polar coordinates θ and ϕ. The projections are taken at the midpoints of each region. The array is repeated four times to populate a hemisphere, which is sufficient because each projection represents two points on the sphere. Given a number of rows in the pattern n, the number of points selected is $4T_n = 2n(n+1)$, where T_n is a triangular number. The values 144, 544, and 2112 arise from the choices $n = 8$, 16, and 32, respectively. We present only the results for 2112 directions, although the results with fewer directions are very similar, suggesting that even 144 directions are adequate for a representative sampling of all directions.

We present a few representative distributions of the RECIST diameter for a single tumor in Fig. 2. The distributions have a great variety, an effect which was predicted from a model based on the union of ellipsoids.[18] By rotating each individual nodule, we learn about the possible variation of the RECIST diameter for a given volume much more quickly than if only one measurement per nodule was supplied.

The mixed linear model we use for the lung data is similar to the one for the liver data in Eq. (1) with the exception that we must account for multiple levels of nesting. The random effects are for patient, physical nodule, radiologists' readings within a physical nodule, and the hypothetical orientation of a nodule reading. The parameters of the model, which are "fixed terms" in the language of mixed models, are an intercept in $\log d$ and a slope of $\Delta \log d / \Delta \log V$. Translated to the original variables, the equation $V \sim d^{\alpha}$ includes the point $(d, V) = (0, 0)$ as long as $\alpha > 0$. Let $d_{ijk\ell}$ denote the RECIST value for orientation ℓ of the markup of nodule j by radiologist k in patient i. We model $d_{ijk\ell}$ as

$$\log d_{ijk\ell} = \beta_0 + \beta_1 \log V_{ijk} + \epsilon_{ijk\ell}, \tag{4}$$

with $i = 1, \ldots, N^{\text{patient}}$, $j = 1, \ldots, N_i^{\text{nodule}}$, $k = 1, \ldots, N_{ij}^{\text{reader}}$, and $\ell = 1, \ldots, N_{ijk}^{\text{orientation}}$. In Eq. (4), β_0 and β_1 are fixed unknown parameters, V_{ijk} is the volume for the markup of nodule j by radiologist k in patient i, and $\epsilon_{ijk\ell}$ is a random error. (Although we use some of the same symbols here as for the liver model, the association to a given model will be clear from context.) The random error $\epsilon_{ijk\ell}$ can be decomposed further since patient differences, nodule shape, the radiologists' markup, and the random orientation of the nodule contribute to the error. Specifically,

$$\epsilon_{ijk\ell} = P_i + N_{i(j)} + R_{ij(k)} + O_{ijk(\ell)}, \tag{5}$$

where P_i is the random effect of patient i, $N_{i(j)}$ is the random effect of nodule j nested within patient i, $R_{ij(k)}$ is the random effect of the markup of nodule j by radiologist k in patient i, and $O_{ijk(\ell)}$ is the random effect of orientation ℓ for the markup of nodule j by radiologist k in patient i. While it would be more appropriate for the radiologist effects not to be nested within a patient–nodule combination, radiologists cannot be uniquely identified across patients, so this is the best that can be done. As for the liver data, the model for the WHO two-dimensional product w is the same under the replacement $d \to w$, with all subscripts and superscripts preserved.

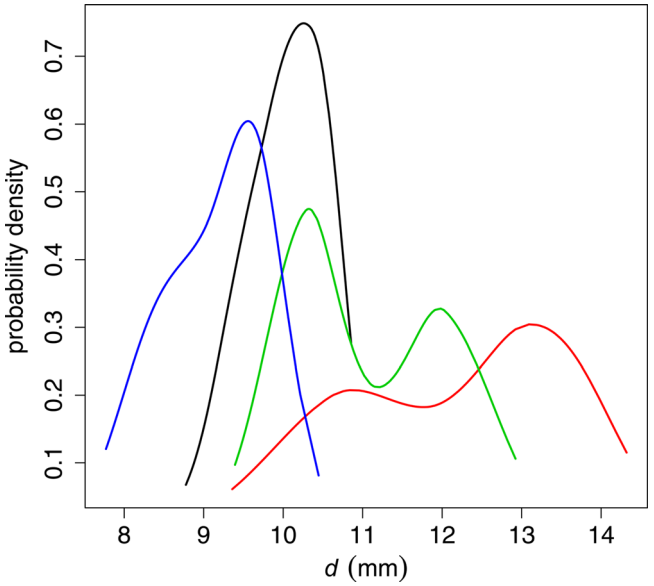

FIG. 2. Probability distributions of RECIST diameter are given for four nodule readings taken from four different patients. The probability distribution assumes a scan axis is picked from a uniform random distribution on the unit sphere. Smoothing is performed (Ref. 29). The selected nodules all have a volume between 750 and 1250 mm^3 with an anisotropy parameter between 0.9 and 1.1. These distributions are representative of the large variety of those linking the RECIST diameter and volume.

As with the liver data, two approaches to inference are considered. The first approach assumes $P_i \overset{iid}{\sim} N(0,\sigma_P^2)$, $N_{i(j)} \overset{iid}{\sim} N(0,\sigma_N^2)$, $R_{k(ij)} \overset{iid}{\sim} N(0,\sigma_R^2)$, and $O_{l(ijk)} \overset{iid}{\sim} N(0,\sigma_O^2)$, with all effects independent of each other. Again, ReML is used for parameter estimation, and the standard normal distribution is used to construct confidence intervals for β_0 and β_1. To construct prediction intervals for $\log V$ at a given $\log d$, both propagation of error and the standard normal distribution are leveraged. When assuming Gaussian random effects, there are six parameters to estimate including two regression coefficients, β_0 and β_1, and four variance components, $\sigma_P, \sigma_N, \sigma_R$, and σ_O.

For the LIDC data, the distribution of $O_{ijk(\ell)}$ can be carefully examined since for each (patient, nodule, reader) triple, there are 2112 observations of the orientation effect. We illustrate a few of those distributions in Fig. 2. Although we have selected nodules with similar volumes which are relatively large and chosen from cases with relatively isotropic voxels, there is still a huge variation in the individual distributions, some of which are far from normally distributed. Such results are consistent with results from mathematical tumor models.[18]

In light of this, and for the purpose of comparison, a second approach to inference is considered. For the second approach, the regression coefficients β_0 and β_1 are estimated via least squares, and confidence and prediction intervals are constructed using a bootstrap algorithm. The bootstrap algorithm is based on resampling residuals and retains the correlation structure of the nested random effects that comprise $\epsilon_{ijk\ell}$, so it retains the correlation structure of the data too. To generate a single bootstrap sample of residuals, the patients are first sampled with replacement. Then, for any sampled patient, say i_0, all nodules in patient i_0 are sampled with replacement. Then, for any sampled nodule, say j_0, in patient i_0, radiologists are sampled with replacement. Last, for any sampled radiologist markup, say k_0, of nodule j_0 in patient i_0, orientations are sampled with replacement. This is repeated until we have $1240 \times 2112 = 2\,618\,880$ resampled residuals. The confidence intervals for β_0 and β_1 are percentile intervals[26] and the prediction intervals for $\log V$ are similar to those in Ref. 27. Any bootstrap intervals in this paper pertaining to the LIDC data are based on 3000 bootstrap samples.

To evaluate our bootstrap method, we used the following cross-validation procedure for the LIDC data. First, the overall data set is split into training and validation data sets. The validation data set is made of all observations from 20 randomly selected patients. The training data set consists of the remaining observations. The bootstrap procedure was then applied to the training data set to form 95% point-wise prediction bounds for $\log V$, and the proportion of the validation data set that falls within the bounds is calculated. If approximately 95% of the validation data set lies within the bounds, the bootstrap procedure can be considered appropriate. The requirement of approximately 95%, instead of exactly 95%, is for two reasons. First, the validation data set is only a sample of patients, so the proportion of the validation data set that falls within the bounds is subject to sampling variability. Second, the prediction bounds are point-wise bounds, not simultaneous bounds. We found 93% in a cross-validation study fell within the 95% prediction intervals for the lung data. Similar cross-validation procedures were performed with the liver data where 98% of the static validation data set and 98% of the change validation data set fell within the bounds. The cross-validation studies show that the bootstrap prediction intervals for $\log V$ at least approximately maintain their stated confidence level, 95%. For the WHO data (either LIDC or liver), and no cross-validation was performed.

Results are given in Sec. III.B.

III. RESULTS

III.A. Liver data

We may assess the effect of change over time using the liver data because each liver tumor (also called malignoma) was read at two times[19] (called here 1 and 2). There are 198 tumors in this data set, of which 18 were not visible on the second reading, and of the remainder, the RECIST diameter decreased by at least 30% for 14 tumors, the RECIST diameter increased by at least 20% for 41 tumors, and the balance of 125 tumors had moderate change. (These values are very similar to the RECIST categories of complete response, partial response, progressive disease, and stable response, although our categories refer to changes in individual tumors, without summation as called for in RECIST.[2]) Our statistical model for this data is described in Sec. II.A.

The results of the Gaussian random effects approach to inference are shown in Table II. The minimum width of the

TABLE II. For the Gaussian random effects model, statistical parameters characterizing differences of the prediction bounds of the $\log V$ and the corresponding ratios of the prediction bounds in the volume shown in Figs. 3, 4, 6, and 8–10. The columns "min" and "max" represent the range of 95% prediction interval for these quantities which vary slowly with $\log d$ or $\log w$ or the 95% confidence interval, as appropriate. The columns labeled "point" are point estimates of β_0 or β_1.

			Log		Ratio		$\hat{\beta}_0$			$\hat{\beta}_1$		
Organ	Method	Type	Min	Max	Min	Max	Min	Point	Max	Min	Point	Max
Lung	RECIST	Static	1.226	1.227	16.81	16.86	0.16	0.17	0.19	0.317	0.321	0.325
Liver	RECIST	Static	1.024	1.057	10.56	11.39	0.01	0.07	0.14	0.332	0.350	0.368
Liver	RECIST	Change	0.881	0.884	7.56	7.65	0.00	0.01	0.02	0.270	0.297	0.324
Lung	WHO	Static	1.136	1.136	13.66	13.69	0.09	0.12	0.15	0.666	0.672	0.679
Liver	WHO	Static	0.930	0.960	8.50	9.12	−0.18	−0.06	0.06	0.684	0.718	0.751
Liver	WHO	Change	0.774	0.791	5.95	6.18	−0.01	0.01	0.04	0.589	0.634	0.691

TABLE III. For the bootstrap algorithm, statistical parameters characterizing differences of the prediction bounds of the $\log V$ and the corresponding ratios of the prediction bounds in the volume for the fit parameters of Figs. 3, 4, 6, and 8–10. The columns min and max represent the range of 95% prediction interval for these quantities which vary slowly with $\log d$ or $\log w$ or the 95% confidence interval, as appropriate. The columns labeled point are point estimates of β_0 or β_1.

			Log		Ratio		$\hat{\beta}_0$			$\hat{\beta}_1$		
Organ	Method	Type	Min	Max	Min	Max	Min	Point	Max	Min	Point	Max
Lung	RECIST	Static	1.2	1.3	16	20	0.19	0.22	0.26	0.29	0.30	0.31
Liver	RECIST	Static	1.0	1.1	10	13	−0.01	0.06	0.13	0.34	0.35	0.36
Liver	RECIST	Change	0.8	1.0	6	10	0.00	0.01	0.02	0.28	0.30	0.35
Lung	WHO	Static	1.1	1.2	13	16	0.17	0.24	0.32	0.59	0.62	0.64
Liver	WHO	Static	0.9	1.1	8	13	−0.22	−0.10	0.03	0.69	0.73	0.76
Liver	WHO	Change	0.7	0.8	5	6	0.00	0.02	0.04	0.60	0.65	0.70

bounds for predicting $\log V$ at time 1 is 1.024, and the minimum width of the bounds for the change data is 0.881. Thus, when the random effects are assumed to be Gaussian, knowledge of d yields knowledge of V to within a factor of 10.56, and knowledge of the ratio d_2/d_1 yields knowledge of V_2/V_1 to within a factor of 7.56.

The results of the bootstrap approach to inference are shown in Table III, when all of the liver data is used to create interval estimates (i.e., no cross-validation). The minimum width of the bounds for $\log V$ vs $\log d$ for time 1 is 1.0, and the minimum width of the bounds for the change data is 0.8. Hence, knowledge of d yields knowledge of V to within a factor of 10, and knowledge of the ratio d_2/d_1 yields knowledge of V_2/V_1 to within a factor of 7. The point-wise prediction intervals for $\log V$ and $\log V_2 - \log V_1$ using both the Gaussian random effects and bootstrap approach are depicted graphically in Figs. 3 and 4, respectively.

Whereas 198 tumors at time 1 were available for Fig. 3, only the 180 tumors visible at both times were plotted in Fig. 4. The fact that the prediction intervals for the change data are only a little smaller than that of the static data suggests the correlation of nodule shape from one time to another is weak. In particular, if nodules changed their volume by preserving some nodule-dependent arbitrary shape and orientation, knowledge of ratio d_2/d_1 would deterministically yield knowledge of V_2/V_1; we do not observe this.

III.B. Lung data

In Fig. 5, we show the distributions of volumes for nodules identified by different numbers of radiologists. It is not surprising that larger nodules are more consistently identified than smaller ones. However, the size ranges have broad overlap, suggesting that size is not the only factor that the radiologists in the LIDC study considered when determining if a nodule-candidate was judged to be a nodule. As emphasized in Ref. 19, during the unblinded reading phase, each radiologist was aware of the marks of the others and nothing was overlooked.

The prediction intervals given for many values of $\log d$ by solid lines in Fig. 6 were calculated using the Gaussian random effects approach. The differences between the upper

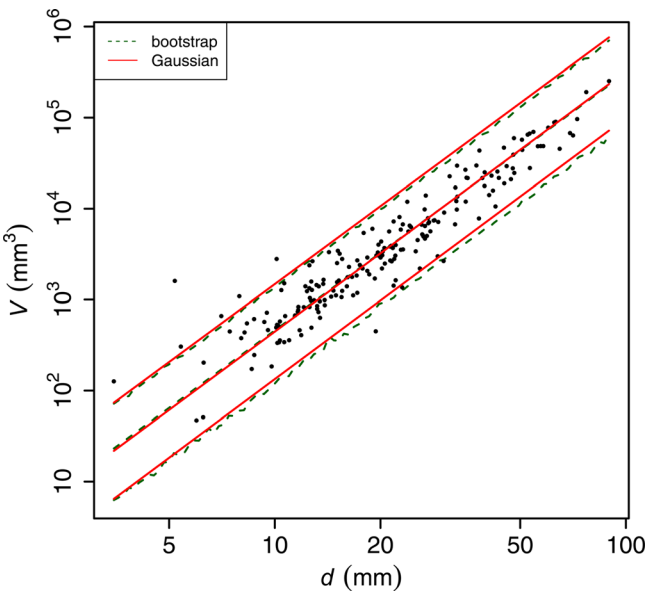

FIG. 3. Fit and 95% prediction intervals for the volume from the RECIST diameter for liver data for two models.

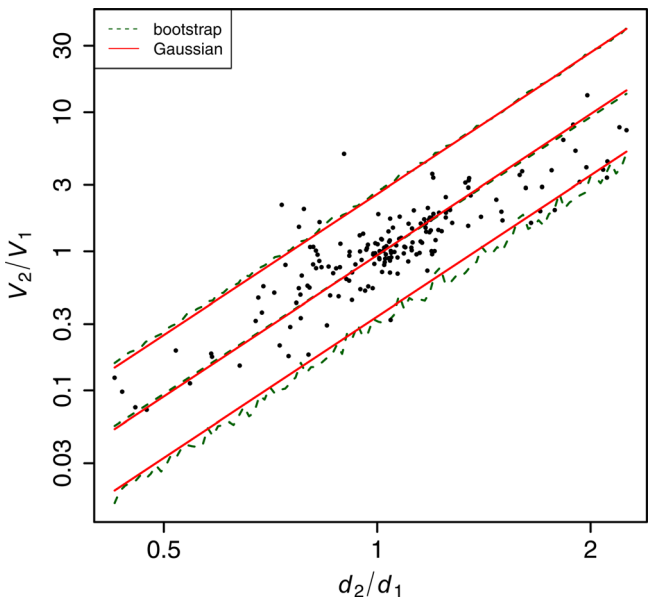

FIG. 4. Fit and 95% prediction intervals for proportional volume from the proportional change in the RECIST diameter for liver data for two models.

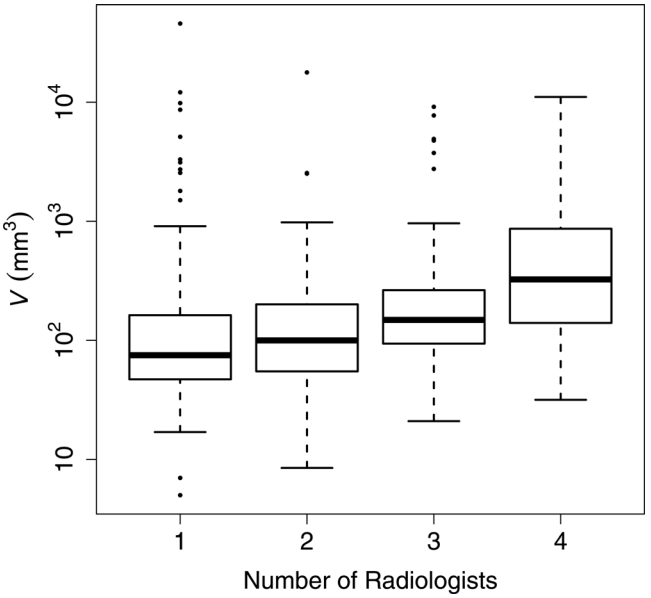

FIG. 5. The distribution of the volume of a physical nodule as a function of the number of radiologists identifying physical nodule as such. (Each volume is found as a harmonic mean.) The middle line in the box represents the median. The top and bottom of the box represent the interquartile range (IQR). The dashed lines ("whiskers") extend to the final data points that lie no more than 1.5 times the IQR beyond the box. All points not encompassed by the whiskers are plotted as points.

and lower bounds are all between 1.226 and 1.227, as seen in Table II. This means that given the RECIST diameter d, the volume V is known within a 95% prediction interval spanning factor of 16.81 or more. The 95% confidence intervals for the fixed effect parameters β_0 and β_1 are [0.16, 0.19] and [0.317, 0.325], respectively.

The prediction intervals, given for many values of $\log d$ by dashed lines in Fig. 6, were calculated using the bootstrap approach, and they vary little in width as $\log V$ changes. The

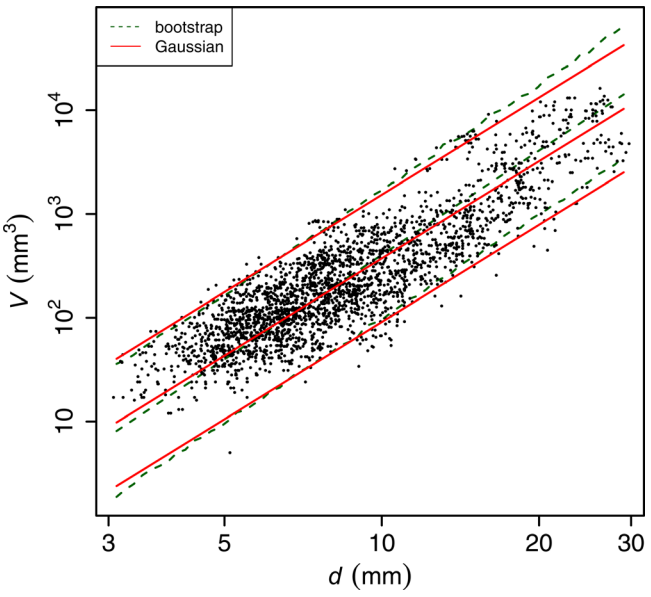

FIG. 6. Fit and 95% prediction intervals for the volume from the RECIST diameter for rotated lung nodule readings for two models, shown with a representative sample of 1–2 orientations per nodule (0.1% of total available).

interval widths, i.e., the difference between the upper and lower bounds, are all between 1.2 and 1.3. This means that given the RECIST diameter d, the volume V is known within a 95% prediction interval spanning factor of 16 or more. This is consistent with the Gaussian random effects approach. The point estimates and 95% confidence intervals for β_0 and β_1 are 0.22 and 0.30 and [0.19, 0.26] and [0.29, 0.31], respectively.

Similar results were presented earlier by Reeves and coworkers.[28] In particular, they give an example that a 10 mm RECIST value corresponds to volume with sphere diameter of 4.16–11.48 mm with 95% confidence. In our terms, this represents a prediction interval which is a factor of $(11.48/4.16)^3 = 21$ wide, which is not far from our result.

One variable which does not appear in our mixed linear model is the voxel anisotropy. All of the LIDC data were isotropic in the scan plane, but the ratio of the voxels along the scan axis to the in-plane length, called here the anisotropy parameter, ranged from 0.68 to 5.30. To understand the importance of anisotropy, we binned the data into two groups, one with a value less than 2 and the balance in the other group. These represent roughly isotropic voxels and prolate voxels. (Highly oblate voxels do not appear in our sample.) When the ratio is less than 2, the 95% confidence interval for β_0 is [0.11, 0.17], and the 95% confidence interval for β_1 is [0.30, 0.32]. When the ratio is greater than or equal to 2, the 95% confidence interval for β_0 is [0.19, 0.26], and the 95% confidence interval for β_1 is [0.31, 0.34]. The confidence intervals for the intercept do not overlap, which implies that there is evidence that ratio has an effect. For anisotropies less than 2, the difference of the upper and lower prediction bounds ranges from 0.94 to 1.03 which represents a prediction interval spanning a factor of 8.7 to 10.7 for volume. The corresponding figures for the group with larger anisotropies are 0.84, 0.94, 6.9, and 8.7. These values are somewhat smaller than the corresponding values in Table III. Hence, the range of anisotropies of the voxels which is present in the lung data, but not the liver data, could account for the larger prediction bounds seen in the lung case, as opposed to some difference in the distribution of nodule shapes. Accounting for voxel anisotropy may narrow the prediction intervals for the volume somewhat, but considering the distributions of Fig. 2, the effect of voxel anisotropy is likely dominated by the effects of nodule shape and orientation.

The RECIST standard indicates a nodule may be measured only if its diameter is 10 mm or more.[2] In Fig. 7, we show the proportion of RECIST diameters of at least 10 mm as a function of the nodule volume under the assumption that the nodule would have been just as likely to have grown with a different orientation. The S-shaped curve illustrates that as the volume of the nodule increases, the chance of a nodule exceeding the threshold increases; moreover, the greatest slope of the S-curve occurs just as a growing sphere would cross the threshold. The feature we wish to emphasize is that the S-curve is quite wide. For example, whereas a nodule with a volume of 270 mm^3 has on average a 25% chance of being measurable under RECIST, a nodule well

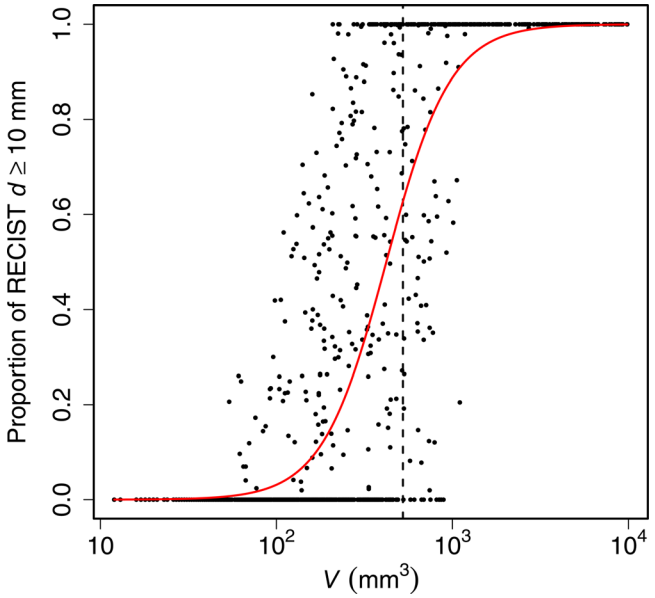

FIG. 7. Proportion of orientations above threshold for lung nodules with a given volume and a fit line. The dashed line at $V = 10^3\pi/6$ mm^3 represents the threshold for the spherical case. The fit is done using the LOGIT functional form $\frac{\exp[-14.422+5.495(\log_{10}(V))]}{1+\exp[-14.422+5.495(\log_{10}(V))]}$.

over twice as large at 670 mm^3 has on average a 75% chance of being classified as measurable, depending on the orientation of the nodule relative to the scan axis.

III.C. Both liver and lung data

Considering the corresponding quantities for $\log V$ vs $\log w$, where w is the WHO bidimensional product, leads to similar results as shown in Figs. 8–10, with the various parameters reported in Tables II and III. We find that $\log w$ leads to prediction bounds which are smaller by 0.1 unit in the logarithm or a factor of 1.3, which is a relatively modest improvement. This is true regardless of whether the Gaussian random effects or the bootstrap approach is considered.

For the LIDC data, the correlation coefficient between RECIST and volume is 0.87. For the liver data at time 1, the correlation coefficient between RECIST and volume is 0.94. For the liver data, the correlation coefficient between the RECIST difference and volume difference is 0.86. Note that all of the correlation coefficients are statistically significantly different from zero no matter which the inference approach is used since none of the 95% confidence intervals for β_1 contain zero.

Under the assumption that nodules grow by increasing their volume without changing the distribution of their shapes, we would expect $d \sim V^{1/3}$. In Figs. 3, 4, and 6, we observe power law dependencies $\hat{\beta}_1$ in Tables II and III. All of the RECIST cases are in reasonable agreement with the scaling prediction of $\frac{1}{3}$. Similarly, we expect $w \sim V^{2/3}$, since w has the dimensions of area. These values, shown graphically in Figs. 8–10, are given numerically in Tables II and III as the WHO cases. Again, reasonable agreement with the scaling prediction of $\frac{2}{3}$ is found in all cases. The lack of a scaling rule for individual nodules discussed earlier does not lead to a lack of scaling for their distributions.

IV. DISCUSSION

We used two statistical methods both based on the same mixed model but making different assumptions about the random effects to obtain our results. In the case of the liver data, the confidence intervals for β_0 and β_1 overlap, and all of the prediction intervals are nearly identical. So our conclusions are the same for both methods. In the case of the lung data, which had a more complicated data structure and likely more highly correlated observations, the results of the two methods are in marginal disagreement for the values of slopes and intercepts for both RECIST and WHO data.

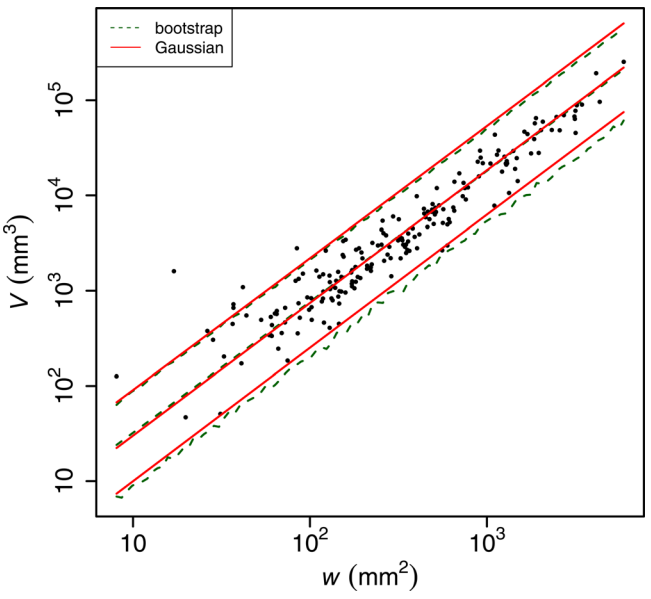

FIG. 8. Fit and 95% prediction intervals for volume from the WHO diameter product for liver data for two models.

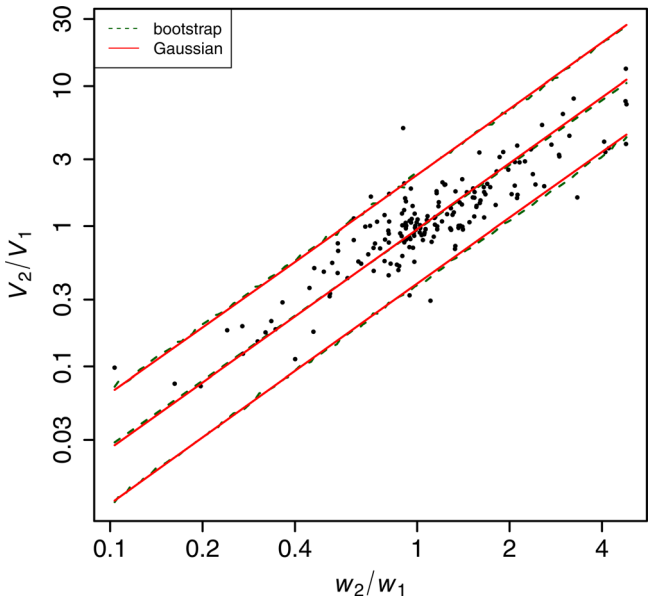

FIG. 9. Fit and 95% confidence limits for the prediction of proportionate change in volume from the change in the WHO diameter product for liver data for two models.

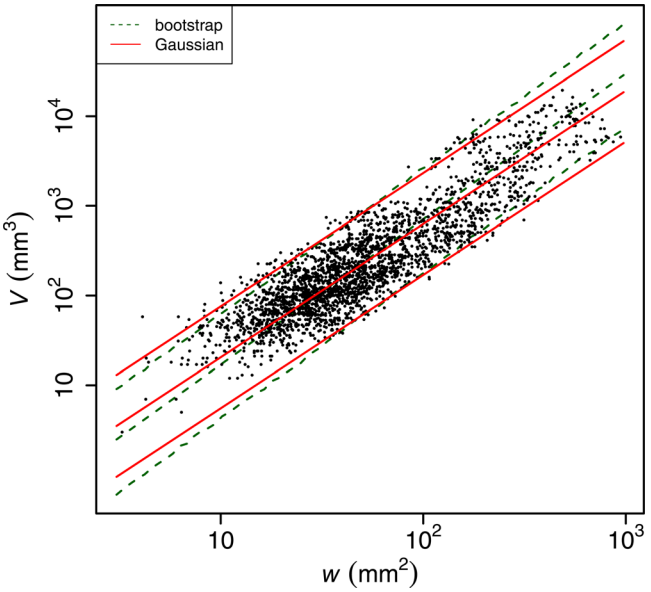

FIG. 10. Fit and 95% point-wise prediction intervals limits for the volume from the WHO diameter product for rotated lung nodule readings for two models.

For the lung data, where the two inference methods slightly disagree about the fixed slope and intercept, one may wish to choose the best of the two methods. However, a best method (between the two considered) does not seem apparent for two reasons. First, the bootstrap procedure uses ordinary least squares to estimate the fixed slope and intercept. It is known that when the observed data contain correlations, the ordinary least squares estimates may be biased, which is undesirable. Second, assuming that $O_{ijk(\ell)}$ follows a Gaussian distribution when it clearly does not (see Fig. 2) is undesirable, and it can also affect the estimates of fixed slope and intercept. To see this more clearly, consider that the likelihood used by ReML is built from the Gaussian assumptions, and that likelihood is maximized to obtain estimates of the variance components. Those estimated variance components are then fed into an expression to get estimates of the fixed slope and intercept. Incorrectly assuming Gaussian random effects can affect the estimate of the fixed slope and intercept through the following chain: the Gaussian assumptions affect the shape of the likelihood which in turn affects the estimates of the variance components which finally affect the estimate of the fixed slope and intercept. Thus, neither approach is ideal, but they both provide a practical solution to a difficult problem.

Interestingly, however, the ratios of the upper and lower bounds of the prediction interval is nearly the same for both the Gaussian and bootstrap approach. This ratio is the key point of interest; although one might like to know exactly how RECIST and volume are related, our result that there is an order of magnitude of variation holds for both approaches.

V. CONCLUSIONS

The RECIST diameter, like the WHO bidimensional product introduced earlier, is used as a practical proxy for volume in studies of the changes of nodule size. While the RECIST diameter and the WHO value are certainly correlated with the volume, and similarly for changes in these quantities, the use of the RECIST diameter introduces additional variation assuming volume is the quantity of interest. In this study, using clinical data, the RECIST diameter determines volume to within a prediction interval which spans an order of magnitude and changes in diameter determine changes in volume only slightly better. The WHO bidimensional product provides only a modest improvement of a factor of 1.3 in the prediction of volumes.

The range of nodule volumes from those that would be almost certainly omitted by the RECIST 10 mm threshold to those that almost certainly would be included spans an order of magnitude. Comparisons of RECIST and volumetrics in the literature do not, to our knowledge, address the question of additional variation due to the random nature of the selection procedure. Although it is implicit that the largest lesions should be followed for growth, RECIST does not necessarily find these lesions, assuming largest means largest in volume.

Exactly how much the use of RECIST rather than volumetrics reduces the statistical power of clinical drug trials[8] is a key open question for future research.

ACKNOWLEDGMENTS

The authors are pleased to acknowledge assistance from Joseph Chen, William Mitchell, and Nicholas Petrick.

[a]Author to whom correspondence should be addressed. Electronic mail: zlevine@nist.gov

[1]P. Therasse *et al.*, "New guidelines to evaluate the response to treatment in solid tumors," J. Natl. Cancer Inst. **92**, 205–216 (2000).
[2]E. A. Eisenhauer *et al.*, "New response evaluation criteria in solid tumors: Revise RECIST guideline (version 1.1)," Eur. J. Cancer **45**, 228–247 (2009).
[3]D. F. Yankelevitz, A. P. Reeves, W. J. Kostis, B. Zhao, and C. I. Henschke, "Small pulmonary nodules: Volumetrically determined growth rates base on CT evaluation," Radiology **217**, 251–256 (2000).
[4]Z. H. Levine, B. R. Galloway, A. P. Peskin, C. P. Heussel, and J. J. Chen, "Tumor volume measurement errors of RECIST studied with ellipsoids," Med. Phys. **38**, 2552–2557 (2011).
[5]S. M. Lee *et al.*, "Usefulness of CT volumetry for primary gastric lesions in predicting pathologic response to neoadjuvant chemotherapy in advanced gastric cancer," Abdom. Imaging **34**, 430–440 (2009).
[6]M. Fabel and H. Bolte, "Automated procedure for volumetric measurement of metastases. Estimation of tumor burden," Der Radiologe **9**, 857–862 (2008).
[7]M. Mantatzis *et al.*, "Treatment response classification of liver metastatic disease evaluated on imaging. Are RECIST unidimensional measurements accurate?," Eur. Radiol. **19**, 1809–1816 (2009).
[8]P. D. Mozley, L. H. Schwartz, C. Bentsen, B. Zhao, N. Petrick, and A. J. Buckler, "Change in lung tumor volume as a biomarker of treatment response: A critical review of the evidence," Ann. Oncol. **21**, 1751–1755 (2010).
[9]S. Steger, F. Franco, N. Sverzellati, G. Chiari, and R. Colomer, "3D Assessment of lymph nodes vs. RECIST 1.1," Acad. Radiol. **18**, 391–394 (2011).
[10]G. Carlsson, B. Gullberg, and L. Hafström, "Estimation of liver tumor volume using different formulas—An experimental study in rats," J. Cancer Res. Clin. Oncol. **105**, 20–23 (1983).
[11]A. K. P. Shanbhogue, A. B. Karnad, and S. R. Prasad, "Tumor response evaluation in oncology: Current update," J. Comput. Assist. Tomogr. **34**, 479–484 (2010).
[12]J. E. Husband *et al.*, "Evaluation of the response to treatment of solid tumors—A consensus statement of the International Cancer Imaging Society," Br. J. Cancer **90**, 2256–2260 (2004).

[13]A. B. Miller, B. Hoogstraten, M. Staquet, and A. Winkler, "Reporting results of cancer treatment," Cancer **47**, 207–214 (1981).
[14]A. D. King *et al.*, "Nasopharyngeal cancers: Which method should be used to measure these irregularly shaped tumors on cross-sectional imaging?," Int. J. Radiat. Oncol., Biol., Phys. **69**, 148–154 (2007).
[15]S. M. Schuetze, L. H. Baker, R. S. Benjamin, and R. Canetta, "Selection of response criteria for clinical trials of sarcoma treatment," Oncologist **13**, 32–40 (2008).
[16]R. J. van Klaveren *et al.*, "Management of lung nodules detected by volume CT scanning," New Engl. J. Med. **361**, 2221–2229 (2009).
[17]Z. H. Levine *et al.*, "RECIST versus volume measurement in medical CT using ellipsoids of known size," Opt. Express **18**, 8151–8159 (2009).
[18]Z. H. Levine, B. R. Galloway, and A. P. Peskin, "Tumor volume measurement errors of RECIST studied with realistic tumor models," J. Res. Natl. Inst. Stand. Technol. **116**, 685–688 (2011).
[19]C. P. Heußel, S. Meler, S. Wittelsberger, H. Götte, P. Mildenberger, and H.-U. Kauczor, "Follow-up CT measurement of liver malignoma according to RECIST and WHO vs. volumetry," Fortschr. Geb. Rontgenstr. Nuklearmed. **179**, 958–964 (2007).
[20]S. G. Armato III *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," Med. Phys. **38**, 915–931 (2011).
[21]A. Forner *et al.*, "Evaluation of tumor response after locoregional therapies in hepatocelluar carcinoma," Cancer **115**, 616–623 (2009).
[22]M. H. S. E. Darkeh, C. Suzuki, and M. R. Torkzad, "The minimum number of target lesions that need to be measured to be representative of the total number of target lesions (according to RECIST)," Br. J. Radiol. **82**, 681–686 (2009).
[23]The mention of commercial products does not imply endorsement by the authors' institutions nor does it imply that they are the best available for the purpose.
[24]C. E. McCulloch and S. R. Searl, *Local Regression and Likelihood* (Wiley, New York, 2001), Chap. 6.9.
[25]"LIDC," National Cancer Institute, https://wiki.nci.nih.gov/display/CIP/LIDC.
[26]B. Efron and R. S. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, Boca Raton, FL, 1993), p. 170.
[27]G. Jones *et al.*, "Application of the bootstrap to calibration experiments," Anal. Chem. **68**, 763–770 (1996).
[28]A. P. Reeves *et al.*, "The Lung Image Database Consortium (LIDC): A comparison of different size metrics for pulmonary nodule measurements," Acad. Radiol. **14**, 1475–1485 (2007).
[29]C. Loader, *Local Regression and Likelihood* (Springer-Verlag, New York, 1999), pp. 79–100.