Open Journal of Statistics, 2012, 2, 260-268
http://dx.doi.org/10.4236/ojs.2012.23031 Published Online July 2012 (http://www.SciRP.org/journal/ojs)

An Empirical Bayes Approach to Robust Variance Estimation: A Statistical Proposal for Quantitative Medical Image Testing

Zhan-Qian John Lu[1,2], Charles Fenimore[1], Ronald H. Gottlieb[3], Carl C. Jaffe[4]
[1]National Institute of Standards and Technology, Gaithersburg, USA
[2]Statistical Engineering Division (776), Information Technology Laboratory, NIST, Gaithersburg, USA
[3]University of Arizona, Tucson, USA
[4]Boston University, Boston, USA
Email: john.lu@nist.gov

Received March 30, 2012; revised May 2, 2012; accepted May 12, 2012

ABSTRACT

The current standard for measuring tumor response using X-ray, CT and MRI is based on the response evaluation criterion in solid tumors (RECIST) which, while providing simplifications over previous (WHO) 2-D methods, stipulate four response categories: CR (complete response), PR (partial response), PD (progressive disease), SD (stable disease) based purely on percentage changes without consideration of any measurement uncertainty. In this paper, we propose a statistical procedure for tumor response assessment based on uncertainty measures of radiologist's measurement data. We present several variance estimation methods using time series methods and empirical Bayes methods when a small number of serial observations are available on each member of a group of subjects. We use a publically available database which contains a set of over 100 CT scan images on 23 patients with annotated RECIST measurements by two radiologist readers. We show that despite the bias in each individual reader's measurements, statistical decisions on tumor change can be made on each individual subject. The consistency of the two readers can be established based on the intra-reader change assessments. Our proposal compares favorably with the RECIST standard protocol, raising the hope that, statistically sound decision on change analysis can be made in the future based on careful variability and measurement uncertainty analysis.

Keywords: RECIST; Quantitative Imaging as a Biomarker; Change Analysis; Lung CT Image Measurement; Inter-Reader and Intra-Reader Variability; Time Series Variance Estimation; Estimation of Many Variances; Statistical Decision Rule on Change

1. Introduction

Currently there is much interest in treating quantitative medical imaging as a biomarker, employing medical imaging tools to assess tumor change, especially in the assessment of response to medical therapy. In order to use medical imaging effectively as quantitative measurement tools, a number of questions are raised regarding quantifying cancerous tumor changes over course of time, such as

- What measures should be used in quantifying meaningful change or response of suspicious tumor objects from images, whether it is based on volume (3D) which has attracted a lot of current interest, or the WHO[1] (2D) or RECIST (1D) [1,2]?
- What is the basic variability in these measures, including both intrinsic measurement variability (e.g. repeatability and reproducibility, and effects from different instrument settings), and expert bias in marking up these measures, or biological variability?
- A critical but related question is—given the variability in imaging acquisition analysis, what is the minimum change that can be detected for a given imaging modality and a chosen image processing method and sizing measure. For example, one would like to know with credible statistical accuracy how large a measured size change must be in order to be declared "significant" in a single individual?

In this paper, we present preliminary steps toward a statistical methodology for variance estimation that will help address these questions. Because there are typically few measurements available on each individual subject, even in a longitudinal study, it is crucial that individual variance estimates for many patients be pooled together

[1]Acronym for the World Health Organization tumor measurement technique which assesses size change over time based on two orthogonal dimensions of the object.

to arrive at stable individual estimates. We apply the empirical Bayes method of Herbert Robbins [3] on estimating many variances for this purpose. Our approach is based on the following intuitive rationale: First, we want to have an empirical variance estimation which is not biased (upward) by the presence of signals, while on the other hand we want to avoid underestimation due to failure to account for additional sources of uncertainty. Secondly, because there are only a few observations for each patient, the variance estimation, whatever method being used, is going to be highly variable due to the low degree of freedom, and it is imperative that more stabilized variance estimates be applied in order to achieve higher power. We propose to use time-series-based robust variance estimators and rejection rules based on a series of measurements on each patient for a given reader so that the effect of real trend in the measurements in an individual subject's progress over time may be minimized. The empirical Bayes variance estimation approach [3,4] can then be applied to individual variance estimates by pooling information across subjects (patients) on which a reader (radiologist) has made observations, providing an indirect way of incorporating intra-reader measurement uncertainty. Finally, a statistical decision rule of change analysis can be developed for a single individual, even if an individual reader may commit systematic bias in his or her measurements.

Currently the most common quantitative measure of tumor nodule size is based on the RECIST technique [1], a set of protocols based on the endpoint defined as the sum of largest diameters of all "target" lesions. In addition, RECIST also recommends the following percentage-based decision rules: Partial Response (PR) in which there occurs at least a 30% decrease from initial baseline measurement, Progressive Disease (PD) where there is at least a 20% increase relative to the smallest value of measurement after treatment initiation, and Stable Disease (SD) where there is neither sufficient shrinkage to qualify as PR or sufficient increase to qualify as PD.

There are at least two concerns from the statistical point of view in applying RECIST guidelines to practice: First, the guide fails to address the uncertainty that is associated with the RECIST measurement, such as the effect of various slice thickness or spacing, and effects from experimental factors [6]; Secondly, the guide fails to clarify or ignores the importance of intra-observer and inter-observer variability by radiologists. Several recent studies have indicated the significant variance contribution of the second source and its important effect on the RECIST decisions [7-9]. By focusing on a case study of a small set of CT scan images from the RIDER [10] database on 23 patients on which two expert radiologists have made a series of markings on some single nodule of RECIST measurements, we demonstrate that a variance estimation approach works reasonably well in providing a statistical alternative to the RECIST percentage-threshold decision rules, and in providing an assessment of the reliability of the two observers. For example, we find that even if there is a clear systematic bias between the two observers, statistical decision on the change analysis can be made reliably based on the serial observations from a single observer, by combining information across different subjects, and that the two observers agree with each other more often than expected from a random guess. The results of the statistical decision rules compare favorably with the categorical percentage-based RECIST method. Thus, variance analysis in quantitative imaging measures can provide informed decisions on clinical image change analysis using a statistical approach based on variance estimation and measurement uncertainty analysis.

2. Statistical Methodology for Variance Estimation

Imagine that there are a number of patients under observation at some discrete time points in a given timespan, as in a typical longitudinal therapeutic study. The data can either be some derived measures of nodule volume, area (WHO) or diameter (RECIST), provided by computer-assisted or manual readings by radiologists, and we denote them for a given patient as a time series, $X_1, X_2, \cdots, X_N$, assuming that they are taken at equally spaced time points for each patient, though our methodology does not require equally spaced observational times. Specifically we may write $X_i = y(t_i)$, $i = 1, \cdots, N$. If time is the only covariate of interest-though any other information serves as a covariate—we can assume that the data (by one reader, one computer algorithm) for each patient consist of:

$$y(t) = f(t) + \beta + \nu^{1/2}(t)\varepsilon_t \tag{1}$$

where $f(t)$ models the change (signal) which may reflect growth as well as effects from clinical treatments, β denote the systematic bias, $\nu(t)$ denotes the repeatability variance component, and ε_t is the measurement error with zero mean and unit variance. Both regression model $f(t)$ and variance function $\nu(t)$ in (1) can be extended to include more covariates and even past observations. Such models are widely used in financial volatility modeling; see for example [11].

Our focus is on how estimation of $\nu(t)$ can be made in the presence of $f(t)$, which is usually unknown. The first case, is to assume that $f(t) \equiv c$, some unknown constant. Then if we assume that $\nu(t) \equiv \sigma^2$, constant variance over time, an obvious estimator is

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2 \quad (2)$$

which is exactly the same estimator as $\hat{\sigma}_U^2 = \frac{1}{N(N-1)}\sum_{1\le i<j\le N}\left(X_i - X_j\right)^2$, a U-statistics-based estimator, as suggested in [12]. However, estimators $\hat{\sigma}, \hat{\sigma}_U$ are valid only under the assumption that there is no change, or $f(t)$ is a constant, and will be heavily biased if $f(t)$ changes with time. We present an alternative estimator,

$$\hat{\sigma}_{TS}^2 = \frac{1}{2(N-1)}\sum_{i=2}^{N}\left(X_i - X_{i-1}\right)^2 \quad (3)$$

This difference-based variance estimator can be justified based on the assumption that f is slowly-varying, or locally constant. In addition, some robust statistics measures may be desirable, due to the fact that they are useful for small data set and are resistant to potential outliers in data. For example, we consider the variance estimator,

$$\hat{\sigma}_{Gini} = \frac{\sqrt{\pi}}{N(N-1)}\sum_{1\le i<j\le N}\left|X_i - X_j\right| \quad (4)$$

also called the Gini mean difference. Also related is the Median Absolute Deviation (MAD) measure, defined as

$$\hat{\sigma}_{MAD} = 1.4826\ \text{median}\left\{\left|\mathrm{X_i} - \text{median}\left\{X_j, j = 1,\cdots,N\right\}\right|, \atop i = 1,\cdots,N\right\} \quad (5)$$

References [13,14] gave extensive discussion of the strengths of robust estimation in practice.

Consequently, we propose

$$\hat{\sigma}_{RTS} = \frac{\sqrt{\pi}}{2(N-1)}\sum_{i=2}^{N}\left|X_i - X_{i-1}\right| \quad (6)$$

as a robust version of (3). A few comments on comparing the different variance estimators are in order.

1) The Gini mean difference in (4) $\hat{\sigma}_{Gini}$ is more robust than (2) and has less variance than MAD in (5).

2) In order to reduce potential bias when there *is* change (or when $f(t)$ is not a constant), we recommend a time-differenced based estimators in (3) and (6).

It should be emphasized that variance estimators like (6) are proposed here to address variance estimation when "no change conditions" can be met in incremental time steps. If this condition cannot be met, estimators like (3) or (6) can still contain significant bias due to signals, and this should be adjusted according procedures suggested in Section 4 by pooling information from other patients.

Once reliable variance estimation becomes available, one can use them to make inference on change analysis, we can define t-statistics-like quantity such as

$$\frac{y(t_p) - y(t_1)}{\hat{\sigma}} \quad (7)$$

which gives the standardized overall change for a patient in the study period $\left[t_1, t_p\right]$ and may be compared to standard statistical inference procedure such as significance test or power analysis. Here $\hat{\sigma}$ can be one of the variance estimators proposed here, such as (6). However, we recommend more stabilized variance estimates by pooling information from other patients, as discussed in Section 3.

If there are m patients being monitored over n_i time intervals, $t_{i1},\cdots,t_{in_i}$, i = 1, ···, m, and the number of readers (radiologists) is p, we can generalize model (1) to individual-based model as:

$$\begin{aligned} y_{ij}(t_{ik}) &= f_i(t_{ik}) + \left(\beta_j(t_{ik}) + b_j(t_{ik})\eta_j(t_{ik})\right) \\ &\quad + v_i^{1/2}(t_{ik})\varepsilon_{ij}(t_{ik}) \\ i &= 1,\cdots,m;\ j = 1,\cdots,p;\ k = 1,\cdots,n_i. \end{aligned} \quad (8)$$

The reader difference due to multiple readers or radiologists is modeled by the bias $\beta_j(t)$ and variance $b_j(t)$, and each patient may have his or her own variance function $v_i(t)$ due to measurement uncertainty and his or her own change function $f_i(t)$. In our formulation, to simplify, we have ignored the actual time recordings and treated the time series data as if they are observed on equally spaced time intervals. Because typically there are only a few observations (over a period of 4 to 5 visits by a patient), the variance functions associated with model (8) are assumed to be independent of time t (homoscedastic), and the reader bias is assumed to be constant over time for each reader. In the following section we discuss how variance estimates from different subjects can be combined to provide an improved and more stable statistical test.

3. Pooled Variance Estimates

Recall that we may use a variance estimator like (6) which, however, requires the longitudinal growth to be slowly-varying. We discuss bias reduction by using information from data sets on other patients. If there are many variances to be estimated, the main issue is how information from similar data sets can be combined to obtain improved and more reliable variance estimates.

Robbins [3] discussed a linear empirical Bayes approach to estimation of many variances which share some common mean. Specifically, if we are given a number of data sets to estimate respectively many means and variances, simultaneously. Let x_{ij} be independent and normal for i = 1, ···, m and j = 1, ···, $n_i = 2r_i + 1 \ge 2$, with unknown $\mu_i = Ex_{ij}$ and $\sigma_i^2 = Var\ x_{ij}$. Define

$$\bar{x}_i = \frac{1}{n_i}\sum_j x_{ij},\ s_i^2 = \frac{1}{n_i - 1}\sum_j \left(x_{ij} - \bar{x}_i\right)^2,\ q^2 = \frac{1}{m}\sum_i s_i^2,$$
$$\bar{x} = \frac{1}{m}\sum_i \bar{x}_i,\ d^4 = \left[\frac{1}{m}\sum_i \left(\frac{r_i}{1+r_i}s_i^4 - q^4\right)\right]^+ \quad (9)$$

where $[\ ]^+$ denotes the nonnegative part. Then one of the ways of estimating the σ_i^2 by linear empirical Bayes method (abbreviated as l.e. B) is to use

$$\hat{\sigma}_i^2 = q^2 + \left[1 - \frac{d^4 + q^4}{\left(1+r_i\right)d^4 + q^4}\left(s_i^2 - q^2\right)\right]. \quad (10)$$

Equation (49) of [3], see also [4]. In our applications, for readings of each patient, the robust variance estimate $\hat{\sigma}_{RTS}^2$ will be computed and used in place of s_i^2 in (10).

It is noted that in Robbins's approach, the signal is assumed to be constant over time. As discussed in Section 3, this assumption can be relaxed when variance estimator based on (6) is used, since the latter is still valid when the underlying *f* is locally constant. However, if the latter assumption cannot be met, the variance estimate can be inflated due to the bias from the signals. Bias adjustment procedure can be easily devised by "borrowing" information from variance estimates across many subjects. An implementation is illustrated within a real data example in the next section.

Given the availability of reliable variance estimation, we can define a statistical procedure for change analysis based on the z-type ratio quantity for comparing means of two normal random variables:

$$\frac{y_i\left(t_p\right) - y_i\left(t_1\right)}{\sqrt{2}\hat{\sigma}_i} \quad (11)$$

where $y_i\left(t_p\right) - y_i\left(t_1\right)$ defines the overall change in the study period for patient *i*, for $i = 1, \cdots, m$, and the ratio can be compared to the standard normal distribution for significance test. Here $\hat{\sigma}_i$ is the final variance estimate for a given patient based on annotation data from a given radiologist. A change analysis decision rule can be based on (11), using, say the standard normal distribution as reference for significance test whether there is an increase, or a decrease in the serial measurements of nodule diameters. This statistical proposal for deciding change is in contrast to the recommended RECIST practice [1] which is based on the percentage change

$$\frac{y_i\left(t_p\right) - y_i\left(t_1\right)}{y_i\left(t_1\right)}, \quad (12)$$

if the measurement at the entry time point is taken as the baseline for patient *i*. We approximate the RECIST guideline by classifying progressive disease (PD) or partial response (PR) based on whether the measured tumor percentage change (12) is a greater than or equal to 20% *increase* (*i.e.*, PD), or shows shrinkage by 30% *or more* (*i.e.*, PR).

4. Analysis of the Bias and Corroboration of Expert Annotations in the RIDER Database

The annotation data consisting of single tumor diameter measurements by two expert radiologists on 23 patient cases based on over 100 CT image scans contained from the National Cancer Institute RIDER image archive (NCIA) [10] is the focus of this statistical case study. The RIDER medical image archive [15] is a large collection of CT images of patients undergoing treatment for non-small cell lung cancer. CT scans, de-identified for patient privacy, had their cancer masses measured by RECIST guidelines at approximately 12 week repeated intervals to track tumor response during the course of therapy. The images were acquired by state of the art 16-row multi-detector spiral CT scanners at adjacent 5 mm slice thicknesses and stored in standard DICOM data format[2]. The cases were viewed and the tumor masses measured at each time interval on a standard picture-archiving system (PACS) viewing workstation (Cedara, Merge Healthcare[3]). These time-sequence RECIST readings by multiple radiologists provide a candidate "ground truth" nodule size behavior on each patient. There are 90 observations in total for 23 patients, with longitudinal observations ranging from 2 to 7 visits per patient.

Figure 1 shows the plot of the raw data. **Figure 2** shows the sequential steps of variance estimation process discussed in Section 2 and Section 3. The top figure shows the raw standard deviation based on (6) based on one reader's observations for each patient. One can see that there is a common range for std values among all patients and only for a few patients whose estimates are clearly outlying due to the signal contamination. In the middle figure, a bias adjustment procedure is implemented by replacing the outlying standard deviation (std) by the mean std plus or minus the MAD of stds among all patients. The bottom figure gives the variance estimates based on the Robbins method (*i.e.* Bayes method) applied to the bias-adjusted stds shown in the middle figure. The statistical test statistics are computed for each patient and are shown in **Figure 3**. In the top figure, the test statistics is based on (7) with variance estimate based

[2]Digital Imaging and Communications in Medicine, http://medical.nema.org/

[3]Certain commercial equipment, instrument, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

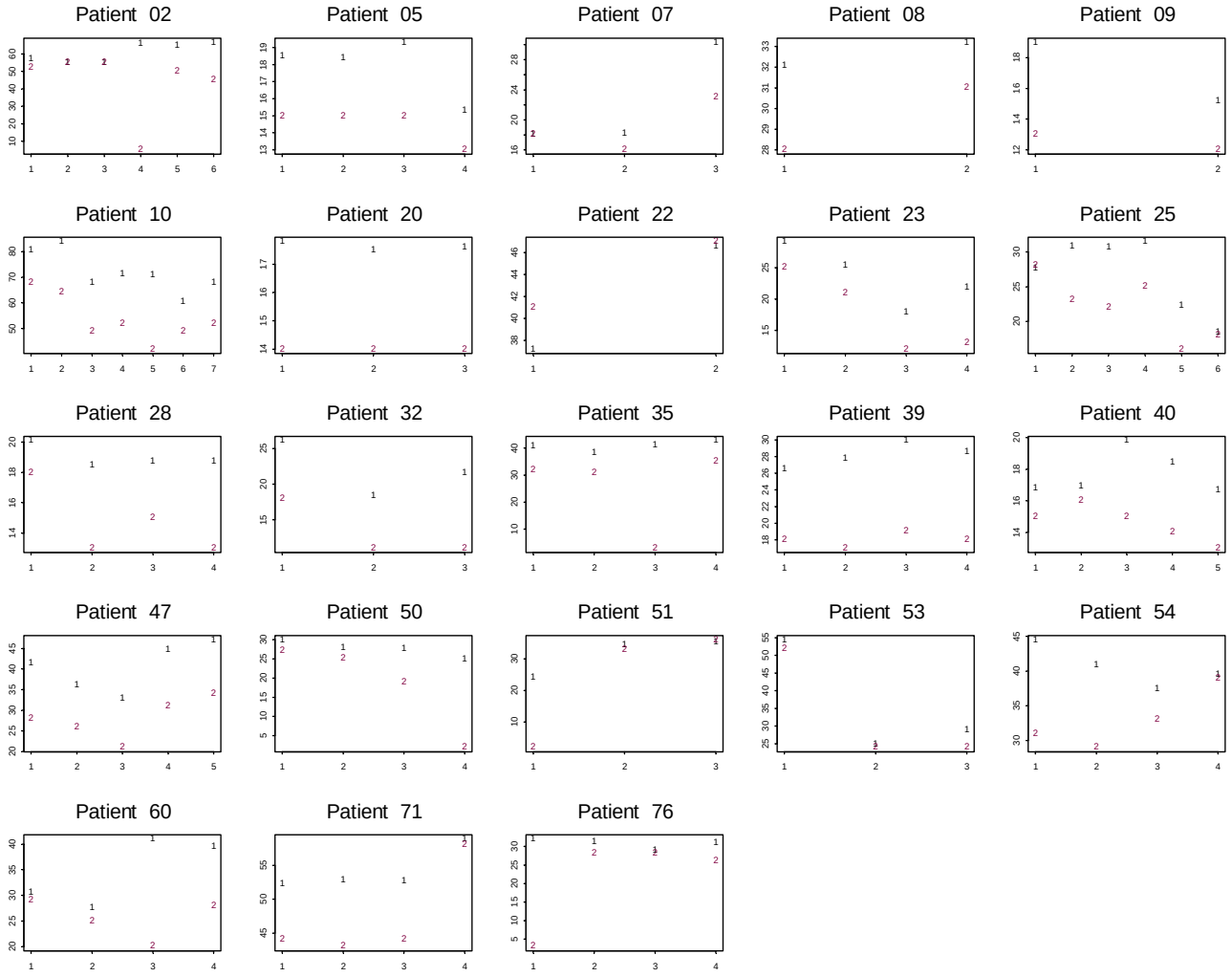

Figure 1. Plots of RECIST readings versus time index for each patient by two radiologists (denoted 1, 2) in a longitudinal study. The RECIST markup data here is the largest diameter of one identified nodule, in millimeters (mm).

on (6) for each patient. The bottom figure, the test statistics is defined similarly as in (11) but with variance estimate as computed given in the bottom figure in **Figure 2**. One can conclude from **Figure 3** that the test results in the bottom figure significantly improve over original results in top figure. For patient number 2, the new test does not find significance, while original raw-variance based test finds strong significance due to a low variance estimate. The new test seems to be more consistent with visual appearance of patient number 2 in **Figure 1**. Similar comments apply to data for patient number 11. The opposite is observed for patient number 18, and patient number 19. Original tests based on raw variance estimate do not find significance due to inflated variance estimates, and this is corrected in the new test. As a result, significant change is observed for both patient numbers 18 and 19 using the improved test. It is found that the two readers agree with each other on their assessment in the direction of change on 19 out of 23 patients. (They disagreed on patient number 1, 20, 21, and 23). A summary of decision results based on the statistical tests is provided in detail in **Table 1**, where we use 10% as the threshold for significant increase and 5% level for significant decrease.

Interestingly, one may compare the statistical test results with the RECIST guidelines (a time-sequence increase of at least 20% defines "progressive disease" (PD), while a decrease of at least 30% defines "partial response" (PR)), which can be inferred from the relative percentage change data plotted in **Figure 4**. Summary of the RECIST analysis is given in **Table 2**. In short, on 4 out of 23 patients, the two experts have given percentage changes of opposite directions (cf. patient numbers 1, 20, 21, and 23). The two experts differ in their computed percentage changes with a mean average difference of 21%. In terms of RECIST decision results based on the radiologists' individual assessments, in addition to the 4 patients on which they totally disagree, they agree on 15

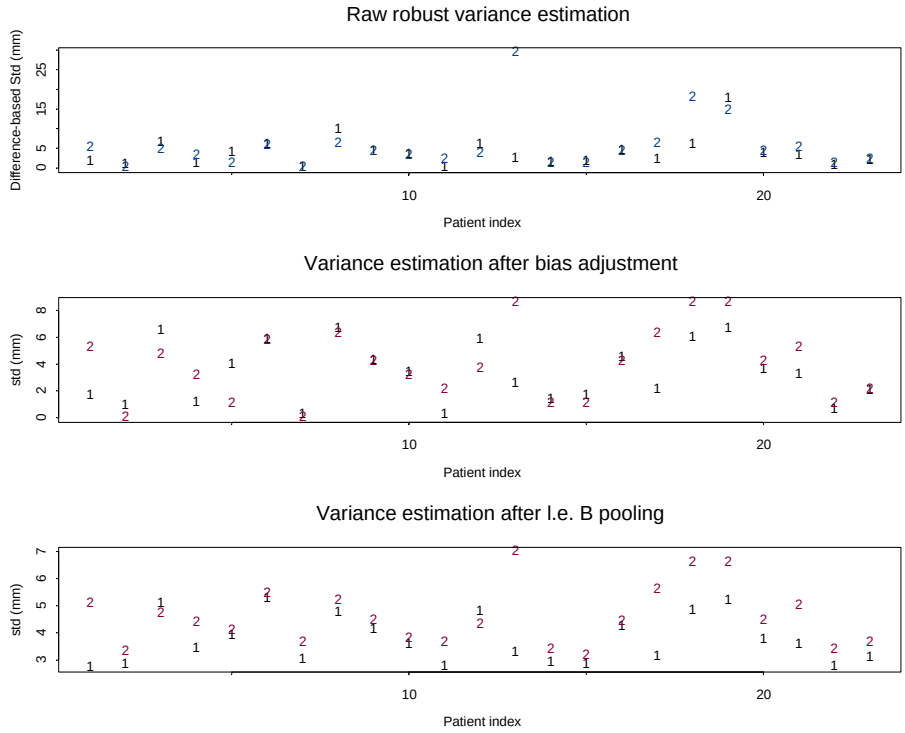

Figure 2. Variance estimation. Top: Raw robust variance estimates for within-patient readings from each of the two experts (denoted by 1 and 2). Middle: after bias adjustment from signal bias (mainly for highest variances). Bottom: after variance pooling to stabilize and to improve underestimated variance estimates (low variances) using Robbins method.

Table 1. *Summary on the statistical results*: Out of 23 patients annotated, two readers totally disagree on 4 patients (Patients 1, 20, 21, 23). They agree on 16 patients, they agree partially on 3 patients (Patients 3, 8, 17).

		Reader 2			
		Significant increase (x)	Increase but not sig (%)	Decrease but not sig (o)	Significant decrease (--)
Reader 1	Significant increase (x)	2	2	0	2
	Increase but not sig (%)	0	4	0	0
	Decrease but not sig (o)	0	0	4	0
	Significant decrease (--)	2	0	1	6

We define the following symbols: x for significant increase at 10% level, -- for significant decrease, o for non-significant decrease, and % for non-significant increase.

Table 2. *Summary on RECIST results*: On 4 out of 23 patients, the two experts have given percentage changes of opposite directions (cf. Patients 1, 20, 21, 23). The two experts differ in their computed percentage changes with a mean average difference of 21%. In terms of RECIST decision results, they agree on 15 patients, and agree partially on 4 patients (patients 8, 9, 17, 22).

		Reader 2			
		Progressive disease (PD)	Increase but below 20% (y)	Decrease but below 30% (y)	Partial recovery (PR)
Reader 1	Progressive disease (PD)	2	1	1	0
	Increase but below 20% (y)	1	4	1	0
	Decrease but below 30% (N)	1	1	7	2
	Partial Recovery (PR)	0	0	0	2

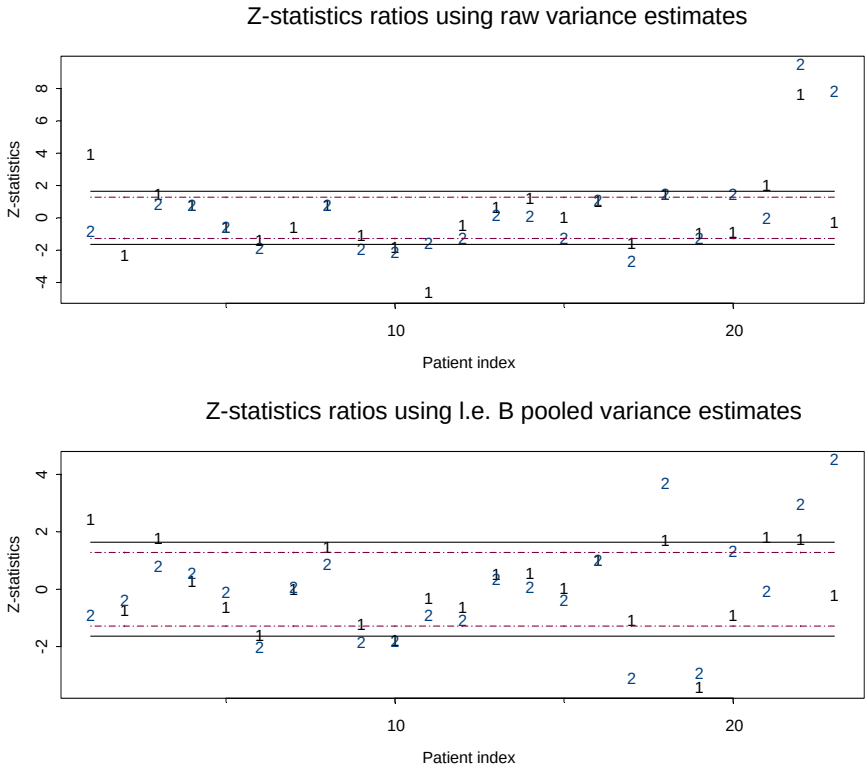

Figure 3. Z-ratio test statistics for change based on readings from two experts (denoted by symbols 1, 2) in a longitudinal study involving 23 Patients. Top: Based on raw variance estimate. Bottom: Based on pooled and stabilized variance estimation as given in Figure 2. The solid lines (black) denote the 0.05 significance test threshold for positive or negative change in the mean, and dashed line (red) denote the 0.10 significance test threshold for change.

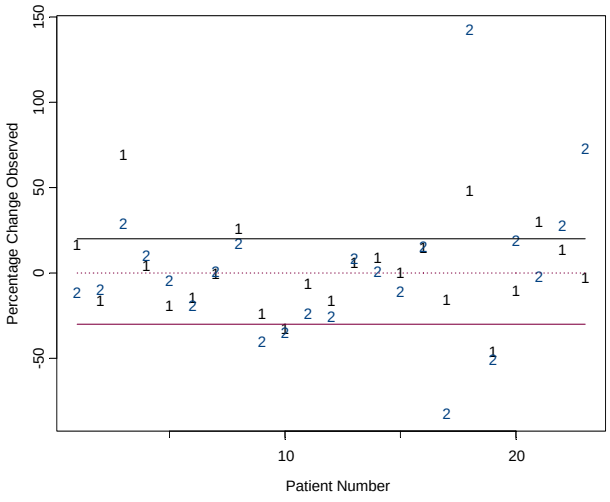

Figure 4. RECIST percentage-based interpretation of annotated RIDER data. X-axis: Patient number from 1 through 23 for 23 patients in the annotated database. Y-axis: Computed percentage changes (measurement at final time minus at entry time, divided by measurement at entry time) for two experts (1s for reader one, 2s for reader two). Data are the RECIST annotations (largest diameter of nodule among all slices at a given time) at different times for all 23 patients by two radiologists. The two solid lines denote the 20% increase and 30% decrease thresholds.

patients, and agree partially on 4 patients (patient numbers 8, 9, 17, and 22) in their categorical classifications (progressive disease (PD), partial response (PR)). In terms of the Kappa measure [16], the statistical tests give

$$\sum_i \theta_{ij} - \sum_i \theta_{i.} \cdot \theta_{.i} \Big/ 1 - \sum_i \theta_{i.} \cdot \theta_{.i} = 0.5861,$$

where θ_{ij} denotes one of the entries in **Table 1**, and $\theta_{.i}$ and $\theta_{i.}$ denote the row and column sums, divided by the total (23). The Kappa number for the statistical test indicates slightly better agreement between the two readers than the Kappa measure (0.5027) based on RECIST-based results in **Table 2** (However, both approaches indicate there is moderate agreement among the two readers [17]).

At the minimum we can ask whether there is any corroboration or dependence between the two readers' assessments. If we only use the signs of the categorical score measurements (so the variance estimation has no impact), there are 7 concordant "increases", 10 concordant "decreases", 4 discordant decisions for "increase" by one reader and "decrease" by another reader, and vice versa. The Chi-squared test for independence by the two readers using the contingency table gives a value of 5.5457 with 1 df, and P-value of 0.0185. Using Fisher's exact test, P-value is 0.0092 for two-sided alternative.

We may also decide on a threshold value, say 1, and assign the decision 1, 0, −1 for "increase", "indecision", "decrease" if the score on a patient is ≥1 between −1 and 1, or ≤−1. The contingency table for the two readers in the column and row order of −1, 0, 1 is: 7, 0, 0; 2, 5, 3; 0, 4, 2. The Pearson's Chi-squared test for independence has value of 16.3215, with df = 4, P-value = 0.0026. Fisher's exact test has a P-value of 0.0012 (for two-sided alternative).

5. Discussion

We believe that there is a strong need to study the reliability and statistical performance of RECIST, or any other time-sequence tumor size measurement regimes such as WHO or 3D volume metrics. Statistical methods suggested in this paper are used to demonstrate the potential of medical decision making by taking into account explicitly the uncertainty in the markings by expert radiologists, and a statistical decision rule for change could potentially be available for the future based on realistic measurement quantification along the lines of [6,18]. In addition, there is a critical need for establishing measurement uncertainty, such as accommodating the effects of protocols and instrument settings [6]. Statistics-based decision rule can easily incorporate the different facets of uncertainty components in therapy response decision making. There are needs to study biological variability and to study the algorithmic factors of computer-assisted measurements in other size measures such as volume metric which is mainly useful for thin slice CT scans (1.0 mm or less) [6].

Partly due to the observation that there is measurement bias in the absolute nodule size measurements, alternative procedures have been investigated for direct change measurements (e.g. [19,20]). However, we caution the readers that the latter approach raises additional issues with the uncertainty in the change measurements themselves and there are still issues on how to assess measurement uncertainty in change-measurement data such as for small nodules. Though there are many developments with RECIST, this important topic has received little attention in the statistical literature (an exception is [21]), we believe there are ample opportunities for statisticians to be engaged in this important medical image decision analysis concerned with assessing therapeutic effectiveness.

6. Acknowledgements

The first two authors would like to thank our colleague Qiming Wang for her work in analyzing and accessing the DICOM images used in this paper, and to our colleague Alden Dima who made the DICOM image database server available to us.

REFERENCES

[1] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe and J. Verweij, "New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (Version 1.1)," *European Journal of Cancer*, Vol. 45, No. 2, 2009, pp. 228-247. doi:10.1016/j.ejca.2008.10.026

[2] C. C. Jaffe, "Measures of Response: RECIST, WHO, and New Alternatives," *Journal of Clinical Oncology*, Vol. 24, No. 20, 2006, pp. 3245-3251. doi:10.1200/JCO.2006.06.5599

[3] H. Robbins, "Estimating Many Variances," In: S. S. Gupta, Ed., *Statistical Decision Theory and Related Topics III*, *Vol*. 2, Academic Press, New York, 1982, pp. 251-261.

[4] H. Robbins, "Some Thoughts on Empirical Bayes Eestimation," *Annals of Statistics*, Vol. 11, No. 3, 1983, pp. 713-723. doi:10.1214/aos/1176346239

[5] L. H. Schwartz, M. Mazumdar, W. Brown, A. Smith and D. M. Panicek, "Variability in Response Assessment in Solid Tumors: Effect of Number of Lesions Chosen for Measurement," *Clinical Cancer Research*, Vol. 9, No. 12, 2003, pp. 4318-4323.

[6] Z. Q. J. Lu, N. Petrick, C. Fenimore, D. Clunie, K. Borradaile, R. Ford, M. F. McNitt-Gray, H. J. G. Kim, R. Zeng, M. A. Gavrielides, B. Zhao and A. J. Buckler, "Statistical Analysis of Reader Measurement Variability in Nodule Sizing with CT Phantom Imaging Data," NIST Interagency Report, 2012.

[7] J. J. Erasmus, G. W. Gladish, L. Broemeling, B. S. Sabloff, M. T. Truong, R. S. Herbst and R. F. Munden, "Interobserver and Intraobserver Variability in Measurement of Non-Small-Cell Carcinoma Lung Lesions: Implications for Assessment of Tumor Response," *Journal of Clinical Oncology*, Vol. 21, No. 13, 2003, pp. 2574-2582. doi:10.1200/JCO.2003.01.144

[8] L. E. Dodd, R. F. Wagner, S. G. Armato III, M. F. McNitt-Gray, S. Beiden, H.-P. Chan, D. Gur, G. McleNnan, C. E. Metz, N. Petrick, B. Sahiner and J. Sayre, "Assessment Methodologies and Statistical Issues for Computer-Aided Diagnosis of Lung Nodules in Computed Tomography: Contemporary Research Topics Relevant to the Lung Image Database Consortium," *Academic Radiology*, Vol. 11, No. 4, 2004, pp. 462-475. doi:10.1016/S1076-6332(03)00814-6

[9] C. R. Meyer, T. D. Johnson, G. McLennan, D. R. Aberle, E. A. Kazerooni, H. MacMahon, B. F. Mullan, D. F. Yankelevitz, E. J. R. van Beek, S. G. Armato III, M. F. McNitt-Gray, A. P. Reeves, D. Gur, C. I. Henschke, E. A. Hoffman, R. H. Bland, G. Laderach, R. Pais, D. Qing, C. Piker, J. Guo, A. Starkey, D. Max, B. Y. Croft and L. P. Clarke, "Evaluation of Lung MDCT Nodule Annotation Across Radiologists and Methods," *Academic Radiology*, Vol. 13, No. 10, 2006, pp. 1254-1265. doi:10.1016/j.acra.2006.07.012

[10] RIDER: Reference Image Database to Evaluate Response,

National Institute of Biomedical Imaging and Bioengineering Institute of NIH. http://www.nibib.nih.gov/Research/Resources/ImageClinData#RIDER

[11] Z. Q. Lu, "Local Polynomial Prediction and Volatility Estimation in Financial Time Series," In: A. S. Soofi and L. Cao, Eds., *Modelling and Forecasting Financial Data*: *Techniques of Nonlinear Dynamics*, Kluwer, Boston, 2002, pp. 115-135.

[12] C. R. Meyer, S. G. Armato III, C. P. Fenimore, G. McLennan, L. M. Bidaut, D. P. Barboriak, M. A. Gavrielides, E. F. Jackson, M. F. McNitt-Gray, P. E. Kinahan, N. Petrick and B. Zhao, "Quantitative Imaging to Assess Tumor Response to Therapy: Common Themes of Measurement, Truth Data, and Error Sources," *Translational Oncology,* Vol. 2, No. 4, 2009, pp.198-210.

[13] P. J. Huber, "Robust Statistics," Wiley, New York, 1981.

[14] D. C. Hoaglin, F. Mosteller and J. W. Tukey, "Understanding Robust and Exploratory Data Analysis," Wiley, New York, 1983.

[15] S. G. Armato III, C. R. Meyer, M. F. McNitt-Gray, G. McLennan, A. P. Reeves, B. Y. Croft and L. P. Clarke, "The Reference Image Database to Evaluate Response to Therapy in Lung Cancer (RIDER) Project: A Resource for the Development of Change-Analysis Software," *Clinical Pharmacology & Therapeutics*, Vol. 84, No. 4, 2008, pp. 448-456. doi:10.1038/clpt.2008.161

[16] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, Vol. 33, No. 1, 1977, pp. 159-174. doi:10.2307/2529310

[17] A. J. Viera and J. M. Garrett, "Understanding the Interobserver Agreement: The Kappa Statist*ics*," *Family Medicine*, Vol. 37, No. 5, 2005, pp. 360-363.

[18] B. Zhao, L. P. James, C. S. Moskowitz, P. Guo, M. S. Ginsberg, R. A. Lefkowitz, Y. Qin, G. J. Riely, M. G. Kris and L. H. Schwartz, "Evaluating Variability in Tumor Measurements from Same-Day Repeat Scans of Patients with Non-Small Cell Lung Cancer," *Radiology*, Vol. 252, No. 1, 2009, pp. 263-272. doi:10.1148/radiol.2522081593

[19] A. P. Reeves, A. B. Chan, D. F. Yankelevitz, C. I. Henschke, B. Kressler, W. J. Kostis, "On Measuring the Change in Size of Pulmonary Nodules," *IEEE Transactions on Medical Imaging*, Vol. 25, No. 4, 2006, pp. 435-450. doi:10.1109/TMI.2006.871548

[20] J. M. Reinhardt, K. Ding, K. Cao, C. E. Christensen, E. A. Hoffman and S. V. Bodas, "Registration-Based Estimates of Local Lung Tissue Expansion Compared to Xenon CT Measures of Specific Ventilation," *Medical Image Analysis*, Vol. 12, No. 6, 2008, pp. 752-763. doi:10.1016/j.media.2008.03.007

[21] L. D. Broemeling, "Bayesian Biostatistics and Diagnostic Medicine," Chapmall & Hll/CRC, Boca Raton, 2007.

