Comparison of measurement methods with a mixed effects procedure accounting for replicated evaluations (COM₃PARE): method comparison algorithm implementation for head and neck IGRT positional verification

Anuradha Roy¹†, Clifton D. Fuller²⁺†, David I. Rosenthal² and Charles R. Thomas Jr.³

Abstract

Purpose: Comparison of imaging measurement devices in the absence of a gold-standard comparator remains a vexing problem; especially in scenarios where multiple, non-paired, replicated measurements occur, as in image-guided radiotherapy (IGRT). As the number of commercially available IGRT presents a challenge to determine whether different IGRT methods may be used interchangeably, an unmet need conceptually parsimonious and statistically robust method to evaluate the agreement between two methods with replicated observations. Consequently, we sought to determine, using an previously reported head and neck positional verification dataset, the feasibility and utility of a Comparison of Measurement Methods with the Mixed Effects Procedure Accounting for Replicated Evaluations (COM₃PARE), a unified conceptual schema and analytic algorithm based upon Roy’s linear mixed effects (LME) model with Kronecker product covariance structure in a doubly multivariate set-up, for IGRT method comparison.

Methods: An anonymized dataset consisting of 100 paired coordinate (X/) measurements from a sequential series of head and neck cancer patients imaged near-simultaneously with cone beam CT (CBCT) and kilovoltage X-ray (KVX) imaging was used for model implementation. Software-suggested CBCT and KVX shifts for the lateral (X), vertical (Y) and longitudinal (Z) dimensions were evaluated for bias, inter-method (between-subject variation), intra-method (within-subject variation), and overall agreement using with a script implementing COM₃PARE with the MIXED procedure of the statistical software package SAS (SAS Institute, Cary, NC, USA).

Results: COM₃PARE showed statistically significant bias agreement and difference in inter-method between CBCT and KVX was observed in the Z-axis (both \( p \)-value < 0.01). Intra-method and overall agreement differences were noted as statistically significant for both the X- and Z-axes (all \( p \)-value < 0.01). Using pre-specified criteria, based on intra-method agreement, CBCT was deemed preferable for X-axis positional verification, with KVX preferred for superoinferior alignment.

Conclusions: The COM₃PARE methodology was validated as feasible and useful in this pilot head and neck cancer positional verification dataset. COM₃PARE represents a flexible and robust standardized analytic methodology for IGRT comparison. The implemented SAS script is included to encourage other groups to implement COM₃PARE in other anatomic sites or IGRT platforms.

*Correspondence: cdfuller@mdanderson.org
†Equal contributors
²Department of Radiation Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA
Full list of author information is available at the end of the article

© 2015 Roy et al. Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Background

Method comparison is a frequent problem encountered whenever different measurement devices/techniques are implemented in the absence of a gold standard [1–7]. Method comparison in radiological science is often a vexing issue [8–15], and is especially notable when competing imaging methodologies are used without establishment of the technical superiority in terms of accuracy of one platform. In a specific example, the explosion in applications of image-guided radiation therapy (IGRT), which necessitates repeated and exceedingly accurate spatial localization in order to carefully deliver conformal radiation dose, places a premium on both reproducibility and accuracy [14, 16–25]. Furthermore, the large number of divergent manufacturer-supported mechanisms for achieving image-guided target localization/positional verification (e.g., 2D- and 3D-ultrasound [26–34], 2-D radiography [35–39], megavoltage and kilovoltage 3-D tomography [40–44]) have arisen in the absence of a gold-standard, and thus have been the impetus for a large number of inter-modality comparative studies, which themselves often utilize a wide array of statistical methods to report between method measurement differences [22, 24, 25, 27, 33, 34, 44–47]. In an effort to more formally assess both inter- and intra-method bias, as well as to streamline comparatively time- and effort-intensive graphical and statistical analysis inherent in many method comparison statistical techniques, we sought to devise an algorithm to explore agreement between two methods of image-guided radiotherapy, using a novel linear mixed effects (LME) model with Kronecker product covariance structure in a doubly multivariate approach [48]. This integrated approach has great potential utility, formally evaluating inter-method bias, inter-subject variability and the intra-subject variability (i.e., agreement between the repeatability coefficients) of the two imaging methods/devices. Testing of all three aspects is crucial, as inter-subject variability is of import when estimating the difference between the two methods giving different measurements on the same subject, while intra-subject variability affords calculation of the random error among the replications taken by the same method on the same subject [49]. We use a doubly multivariate set-up (i.e., measurement data for each subject is considered at two levels, incorporating both the number of measurement methods and number of replicated measurements). This specific LME-based technique, which we shall refer to as COM$_3$PARE (Comparison of Measurement Methods with Mixed Effects Procedure Accounting for Replicated Evaluations) is robust with regard to number of replicates, and is easily performed using SAS software (vide infra). LME models have improved fidelity in scenarios wherein observations are not fully independent and can more correctly models correlated errors, compared to general linear models (GLM), which includes typical statistical analyses (e.g., t-test, ANOVA, linear regression). LME includes multiple random effect components, compared to single element evaluation in most GLM models, affording improved analysis of continuous variables where random effects, multiple hierarchical data levels, and replicated measurements are concerned. The number of replicated measurements on each patient or subject may not be equal, and also the number of replications of the two methods on the same subject may not be equal. The specific aims for this study included:

- First known application of LME-based COM$_3$PARE hypothesis testing protocol for method comparison using imaging data.
- Demonstration of feasibility and utility of COM$_3$PARE using an established head and neck positional verification dataset, previously presented with standard method comparison approaches.

Methods

Datasets

A previously presented dataset consisting of a series of 100 paired measures using two distinct positional verification techniques in a series of 28 sequential head and neck squamous cell carcinoma patients was utilized. As this manuscript is designed to specify a novel statistical methodology, interested readers are referred to the previous manuscript [50], wherein imaging parameters have been previously detailed. Briefly, CBCT and stereoscopic kV X-ray were acquired near-simultaneously at approximately biweekly intervals throughout a patient’s course of treatment (dependent upon the scheduling exigencies in the department) for a series of patients with head and neck cancers. CBCT/kV X-ray analyses were performed using the attached on-board imager (Varian Medical Systems, Palo Alto, CA). Positional verification was performed with manufacturer-supplied software (Varian OBI 1.3/Varian Vision, Varian Medical Systems, Palo Alto, CA) for 3D-3D (CBCT-simulation CT) and 2D-2D (kV X-ray-DRR) automated matching using the aforementioned software. Recorded shifts represent the coregistration/allineation software derived values taken without physician/observer modification. For each paired-method positional acquisition, the origin was defined as the point in space identified by the initial isocenter position using immobilization-mask-based markers. Utilizing a three-dimensional Cartesian coordinate system, this spatial location was designated as a ‘zero point’ with X/Y/Z coordinates of (0, 0, 0). Software-derived shifts for each system were recorded in centimeters, specified as X-, Y- or Z-axis. Software-derived shifts were characterized as X- (lateral or left/right), Y- (vertical or anteroposterior) and Z- (longitudinal or superoinferior) axes, respectively, for both kV X-ray and CBCT IGRT techniques. For the purposes of clarity, we proposed the
following three conditions be met to verify whether two methods for measuring a variable (in this specific case, IGRT-suggested spatial shifts) can be considered interchangeable:

1. No significant bias (i.e., no difference between the means of the two methods under a pre-specified threshold nor a statistically significant difference between said means).
2. No statistically significant difference in the inter-subject (between-subject) variability of the two methods.
3. No statistically significant difference in the intra-subject (within-subject) variability (i.e., repeatability) of the two methods.

For this study, we pre-specified a bias threshold of an absolute value of < 0.1 cm, with a statistically significant difference designated by \( a < 0.05 \). To assess the aforementioned criteria, we implemented the LME methodology proposed by Roy\(^4\), referred to as COM\(_3\)PARE (see Appendix A).

**Statistical analysis with COM\(_3\)PARE**

As mentioned in the introduction the number of replicated measurements on each patient or subject may not be equal, and also the number of replications of the two methods on the same subject may not be equal. Let \( p^\text{KVK}_i \) and \( p^\text{CBCT}_i \) be the number of replications on subject \( i \) by the established method (KVX), and a new method (CBCT) respectively. Let \( p_i = \max(p^\text{KVK}_i, p^\text{CBCT}_i) \), and \( n_i = 2p_i \). Therefore, the number of observations on the \( i \)th subject is \( n_i \) under the assumption that the \( i \)th subject has \( |p^\text{KVK}_i - p^\text{CBCT}_i| \) missing values.

Let \( y_{it}^\text{KVK} \) and \( y_{it}^\text{CBCT} \) be the responses by the established method and a new method of the \( i \)th subject at the \( t \)th replicate, \( i = 1,2,\ldots,N, t = 1,2,\ldots,p_i \). Let \( y_{it} = (y_{it}^\text{KVK}, y_{it}^\text{CBCT})' \) be the \( 2 \times 1 \) vector of measurements corresponding to the \( i \)th subject at the \( t \)th replicate.

Let \( y_i = (y_{i1},y_{i2},\ldots,y_{ip})' \) be the \( (n_i \times 1) \)-dimensional random vector corresponding to the \( i \)th subject. That is, the vector \( y_i \) is obtained by stacking the responses of the KVX method, and the CBCT method at the first replication, then stacking the responses of the KVX method and the CBCT method at the second replication and so on. We write all responses \((y_i)\) of the \( i \)th subject in a matrix equation as

\[
y_i = X_i\beta + Z_i\epsilon_i + \epsilon_i,
\]

with \( b_i \sim N_{n_i}(0,D) \), and \( \epsilon_i \sim N_{n_i}(0,R_i) \),

where \( b_1, b_2, \ldots, b_N, \epsilon_1, \epsilon_2, \ldots, \epsilon_N \) are independent, and \( y_1, y_2, \ldots, y_N \) are also all independent. LME model allows for the explicit analysis of between-subject \((D)\) and within-subject \((R_i)\) sources of variation of the two methods. We define the two methods by a vector variable \( M_{\text{var}} \); \( M_{\text{var}}=1 \) for the KVX method and \( M_{\text{var}}=2 \) for the CBCT method. We choose the intercept and the vector variable \( M_{\text{var}} \) as fixed effects, thus the design matrix \( X_i \) has three columns, and consequently \( \beta = (\beta_0, \beta_1, \beta_2)' \) is a 3-dimensional vector containing the fixed effects. We also choose the vector variable \( M_{\text{var}} \) as random effects, i.e., \( M_{\text{var}} \) is random across individual subjects; thus the design matrix \( Z_i \) has two columns. Therefore, \( b_i = (b_{i1}, b_{i2})' \) is a 2-dimensional vector containing the random effects.

The solution for \( \beta \) gives the means of the two methods \( \mu_{\text{KVX}} \) and \( \mu_{\text{CBCT}} \). The between-subject variance-covariance matrix \( D \) of the KVX method and the CBCT method is a general \((2 \times 2)\)-dimensional matrix, and \( R_i \) is a \((n_i \times n_i)\)-dimensional covariance matrix which depends on \( i \) only through its dimension \( n_i \). The marginal density function of \( y_i \sim N_{n_i}(X_i\beta, Z_iDZ_i'+R_i) \). Suppose the matrix \( \Sigma \) represents the within-subject variance-covariance matrix of the KVX method and the CBCT at any replicate; also, suppose \( V \) represents the \( p \times p \)-dimensional correlation matrix of the replicated measurements on a given method, where \( p = \max(p_i) \). It is assumed that the \( 2 \times 2 \) within-subject variance-covariance matrix \( \Sigma \) is same for all replications, and the correlation matrix \( V \) is assumed to be the same for both the methods. We assume \( R_i = \text{dim}(V \otimes \Sigma), \) where \( V \) and \( \Sigma \) respectively are positive definite matrices as described above, and \( \otimes \) represents the Kronecker product structure. The notation \( \text{dim}(V \otimes \Sigma) \), represents a \((n_i \times n_i)\)-dimensional submatrix obtained from the \((2p \times 2p)\)-dimensional matrix \((V \otimes \Sigma)\), by appropriately keeping the columns and rows corresponding to the \( n_i \)-dimensional response vector \( y_i \). Since the equicorrelated or compound symmetry (CS) structure assumes equal correlation among all replicated measurements, we assume that the correlation matrix \( V \) of the replicated measurements has equicorrelated correlation structure. For the above design matrix \( Z_i \) and between-subject \( D \) and within-subject \( R_i \) sources of variation, the observed \((n_i \times n_i)\)-dimensional overall variance-covariance matrix \( \Omega_i \) for the \( i \)th individual is given by

\[
\text{Cov}(y_i) = \Omega_i = Z_iDZ_i'+R_i = Z_iDZ_i'+\text{dim}(V \otimes \Sigma).
\]

Thus, the covariance matrix has the same structure for each subject, except that of the dimension. The \( 2 \times 2 \) block diagonals Block \( \Omega_i \) in the overall variance-covariance matrix \( \Omega_i \) represent the overall variance-covariance matrix between the two methods. Similarly, the \( 2 \times 2 \) block
diagonals in the overall correlation matrix $\Omega_i$. Correlation represent the overall correlation matrix between the two methods. Thus, the off-diagonal element in this $2 \times 2$ overall correlation matrix gives the overall correlation between the two methods. It can be easily seen that the overall variability is the sum of between-subject variability and within-subject variability (see Roy\textsuperscript{48} for detail). Therefore, we see that if there is a disagreement in overall variabilities, then it may be due to the disagreement in either between-subject variabilities or within-subject variabilities or both.

**MIXED procedure of SAS**

We use MIXED procedure (PROC MIXED) of SAS to get the maximum likelihood estimates (MLEs) of $\beta$, $D$, $R_i$, and $\Omega_i$. METHOD=ML specifies MIXED procedure to calculate the maximum likelihood estimates of the parameters. The COVTEST option requests hypothesis tests for the random effects. CLASS statement specifies the categorical variables. DDFM=KR specifies the Kenward-Roger\textsuperscript{21} correction for computing the denominator degrees of freedom for the fixed effects. The Kenward-Roger correction is suggested whenever one has replicated or repeated measures data; also for missing data. The SOLUITION (S) option in the MODEL statement provides the estimate of the difference between the two mean readings (bias) of the two methods. RANDOM and REPEATED statements specify the structure of the covariance matrices $D$ and $R_i$. See the sample program in Appendix A that demonstrates the use of RANDOM and REPEATED statements. PROC MIXED calculates the $(n_i \times n_i)$-dimensional submatrix $R_i$ of the $i$th subject from the $(2p \times 2p)$-dimensional matrix $(V \otimes \Sigma)$, and eventually calculates $(n_i \times n_i)$-dimensional submatrix $\Omega_i$. When the number of replications on each subject by respective methods is unequal, PROC MIXED considers the case as missing value situation. Options V=3 and VCORR=3 in the RANDOM statement give the estimate of the overall variance-covariance matrix $\Omega_3$ and the corresponding $\Omega_3$-Correlation matrix, i.e., for the third subject. The option G in the RANDOM statement gives the estimate of the between-subject variance-covariance matrix $D$. Option R in the REPEATED statement gives the estimate of the variance-covariance matrix $R_1$ for the first subject. One can get the $\Omega_i$ variance-covariance matrix and the corresponding $\Omega_i$-Correlation matrix for all subjects by specifying $V=1$ to $N$, and VCORR=1 to $N$ in the RANDOM statement. When the correlation matrix $V$ on the replicated measurements assumes equicorrelated structure and $\Sigma$ as unstructured, we use the option TYPE=UN along with SUBJECT=REPLICATE(PATIENT) in the REPEATED statement. This gives the $2 \times 2$ within-subject variance-covariance matrix $\Sigma$. See Appendix A.

**Related hypotheses testings to test the disagreement between KVX and CBCT**

If there is a disagreement between the two methods, it is important to know whether it is due to the bias, due to the difference in between-subject variabilities or due to the difference in within-subject variabilities of the two methods. If it is due to the bias between the two methods, it is easy to correct. The output of PROC MIXED always gives the bias, its $t$-value and its $p$-value. Nonetheless, it is not straightforward to check the agreement or disagreement in between-subject variabilities and in within-subject variabilities of the two methods. We will accomplish these by the indirect use of PROC MIXED in two steps (described below) by using likelihood ratio tests.

**Testing of hypothesis of difference between the means of KVX and CBCT**

We are interested in testing the following hypothesis for bias:

$$H_d : \text{the two methods do not have the same mean,}$$

vs. $$K_d : \text{the two methods have the same mean.}$$

Output of PROC MIXED (Solution for Fixed Effects) gives the bias and the corresponding $t$-value and $p$-value.

**Testing of hypothesis of difference in between-subject variabilities of KVX and CBCT**

Here we are interested in testing the following hypothesis:

$$H_d : \text{the two methods do not have the same between-subject variabilities,}$$

vs. $$K_d : \text{the two methods have the same between-subject variabilities.}$$

We apply the likelihood ratio test for this hypothesis testing. To compute the test statistic $-2 \ln \Lambda_d$, where

$$-2 \ln \Lambda_d = \left[ -2 \ln \max_{H_d} L \right] - \left[ -2 \ln \max_{K_d} L \right].$$

The log likelihood function under both null hypothesis and alternating hypothesis must be maximized separately. We do this by setting the option METHOD=ML in PROC MIXED statement. The option TYPE=UN in the RANDOM statement, along with the option TYPE=UN in the REPEATED statement, is used to calculate the “-2 Log Likelihood” for the covariance structure under $H_d$. Similarly, the option TYPE=CS in the RANDOM statement, along with the option TYPE=UN in the REPEATED statement, is used to calculate the “-2 Log Likelihood” for the covariance structure under $K_d$.

PROC MIXED calculates “-2 Log Likelihood” under the heading of “Fit Statistics”, see Appendix B. The above
testing statistic $-2 \ln \Lambda_d$ under $K_d$ follows a chi-square distribution with degrees of freedom (d.f.) $v_d$, where $v_d$ is computed as

$$v_d = \text{LRT df (under}$H_d$) - \text{LRT df (under}$K_d$).$$

PROC MIXED calculates "LRT df" under the heading of "Null Model Likelihood Ratio Test", see Appendix B.

**Testing of hypothesis of difference in within-subject variabilities of KVX and CBCT**

We test the difference between the repeatability coefficients of the two methods by testing the following hypothesis:

$$H_o : \text{ the two methods do not have the same within-subject variabilities},$$

$$\text{vs. } K_o : \text{ the two methods have the same within-subject variabilities},$$

As before here also we apply the likelihood ratio test for this hypothesis testing, and maximize the log likelihood function under both null hypothesis and alternating hypothesis separately to compute the test statistic $-2 \ln \Lambda_o$, where

$$-2 \ln \Lambda_o = \left[ -2 \ln \max L \right]_{K_o} - \left[ -2 \ln \max L \right]_{H_o}.$$

The option $\text{TYPE=UN}$ in the RANDOM statement, along with $\text{TYPE=UN}$ in the REPEATED statement, is used to calculate the "$2 \text{ Log Likelihood}$" for the covariance structure under $H_o$. $\text{TYPE=UN}$ in the RANDOM statement, along with $\text{TYPE=CS}$ in the REPEATED statement, is used to calculate the "$2 \text{ Log Likelihood}$" for the covariance structure under $K_o$. The test statistic $-2 \ln \Lambda_o$ under $K_o$ follows a chi-square distribution with d.f. $v_o = \text{LRT df (under}$H_o$) - \text{LRT df (under}$K_o$).$

**Results**

Selected parts of the SAS output to test the within-subject variabilities are given in Appendix B. We present the sample SAS code (see Appendix A) to test within-subject variabilities by fitting the linear mixed effects model to our KVX and CBCT shifts for the lateral (X). We see that

$$-2 \ln \max L = 273.7 \quad \text{and} \quad -2 \ln \max L = 239.7,$$

with

$$\text{LRT df (under}$H_o$) = 5 \quad \text{and} \quad \text{LRT df (under}$K_o$) = 4.$$

Therefore,

$$-2 \ln \Lambda_o = \left[ -2 \ln \max L \right]_{K_o} - \left[ -2 \ln \max L \right]_{H_o} = 273.7 - 239.7 = 34.0,$$

with

$$v_o = \text{LRT df (under}$H_o$) - \text{LRT df (under}$K_o$) = 5 - 4 = 1.$$

The $p$-value for testing the within-subject variabilities of the two methods by using IML procedure of SAS is calculated at the third stage (see Appendix A). The $p$-value = 5.5112E–9 (see Appendix B).

Inter-method bias, inter-method agreement, intra-method agreement, overall agreement and correlation results from COV3PARE are presented in Tables 1, 2, 3, 4 and 5. Using COV3PARE, in this specific head and neck positional verification demonstration dataset, while inter-method bias was < 1 mm for all axes, a statistically significant between method bias was noted in the $Z$-axis (superoinferior axis). Also, it was evidenced that there was a statistically significant difference between CBCT and KVX inter-subject variation in the $Z$-axis (Table 1). Intra-subject variability was noted to be statistically significant for $X$- and $Z$-axes, as was overall variation. Correlation coefficient calculation estimation was performed using a mixed effects model (as per Roy [52]).

Using the aforementioned criteria, automated shifts from CBCT and $kV$ X-ray, acquired and processed in the manner detailed are interchangeable only for measurements of the $Y$-axis (anteroposterior), and for example,

| Table 1 | Between-method bias |
|---------|---------------------|
| X       | 0.0335              | 0.6077 |
| Y       | -0.0428             | 0.2836 |
| Z       | -0.0942             | 0.0253 |
should not be used on alternating days in facilities with both systems in either X- or Z-axis. Additionally, our method suggests that, with lower intra-method variability in the X-axis (lateral), CBCT is the preferred measurement method, while in the Z-axis (superoinferior) kV X-ray measurement is preferable.

Discussion

The necessity for quantitative evaluation of competing measurement devices, in cases where on device has not been found to be superior, is a significant need in science generally [1, 2, 5–7], and particularly within the radiological sciences community. Specifically, this issue is encountered when comparing distinct positional verification methods for image-guided radiotherapy [34, 53–56]. The difficulty of assessing competing platforms is particularly vexing, as it impedes efforts at cross platform comparison. Our group [24, 50] and others have implemented several distinct methods for presenting such analysis [25, 27, 33, 34, 45, 57]. Our previous efforts have utilized several extant method comparison statistical presentations (including Bland-Altman, Lin’s concordance [58], Deming orthogonal regression [59, 60]); however, what was gained in completeness was lacking in parsimony. To this end, we sought to define an improved algorithm for practical comparison of distinct imaging methodologies, with a non-fixed number of repeated measurements per patient, in the absence of a “gold standard”. Often, inappropriate statistical analyses are implemented in lieu of formal method comparison statistics. The analysis of different measurement devices is not as straightforward as the initial observer may suppose. Bland and Altman demonstrated that mean comparison and linear regression are insufficient for comparison of differing measurement techniques [1]. The Bland-Altman method is succinct and easily interpretable, making it a classic of medical literature. In a series of seminal papers [1–7], Bland and Altman defined the standard methodology for comparing differing measurements, as well as establishing effective techniques accounting for inter- and intra-method variability/repeatability. However, while the Bland-Altman methodology remains the current benchmark, it fails (by design, one should note) to include generation of a formalized $p$-value, instead recommending that a clinically meaningful difference between measures be utilized. Additionally, though repeatability estimation is a recommended component of accurate method comparison, the calculation for greater than two replicates is somewhat unwieldy using the methodology proposed by Bland and Altman. Since many IGRT datasets span > 30 repeated daily measures, the utility of a statistical methodology which can readily integrate large replicate numbers is desirable. The COM$_3$PARE methodology presented herein represents an attempt to integrate several desirable methodological attributes into a unified, readily performed statistical process. COM$_3$PARE has several advantages over existing method comparison statistical analyses. Specifically, compared to general linear model [61, 62] (GLM)-based approaches (such as the $t$-test, linear regression, and ANOVA [63]), which fail to account for multiple sources of random variance, the linear mixed effects (LME)-based COM$_3$PARE platform integrates variation estimation at multiple hierarchical levels (i.e., between- and within- measurement methods/subjects) [48]. From a practical point of view, this allows factor-wise assessment of procedural or technical variability of each of the two methods rather than a combined assessment, so that there is the capacity to determine the exact source of disagreement. COM$_3$PARE is also resilient with regard to uneven numbers of replicates per device, a feature of great practical utility in a clinical setting, such as daily IGRT recording, where the number of IGRT fractions received for each patient may differ based on fractionation regimen of clinical exigency. Additionally, since COM$_3$PARE has the capacity to fit differences in said variability to a hypothesis testing-friendly Bonferroni-corrected $p$-value output, while still implementing clinician-determined thresholds

| Table 2 Inter-method agreement |
|--------------------------------|
|                   | KVX (cm) | CBCT (cm) | $p$-value |
|---------------------|---------|-----------|-----------|
| X                   | 0.0413  | 0.0670    | 0.4795    |
| Y                   | 0.0511  | 0.0464    | 0.7518    |
| Z                   | 0.0273  | 0.0848    | 0.0010    |

| Table 3 Intra-method agreement |
|--------------------------------|
|                   | KVX (cm) | CBCT (cm) | $p$-value |
|---------------------|---------|-----------|-----------|
| X                   | 0.3396  | 0.1047    | 5.5 x 10$^{-3}$ |
| Y                   | 0.1687  | 0.1693    | 1.0       |
| Z                   | 0.0485  | 0.0825    | 0.0034    |

| Table 4 Overall agreement |
|----------------------------|
|                   | KVX (cm) | CBCT (cm) | $p$-value |
|---------------------|---------|-----------|-----------|
| X                   | 0.3809  | 0.1717    | 3.7 x 10$^{-8}$ |
| Y                   | 0.2198  | 0.2157    | 0.9512    |
| Z                   | 0.0758  | 0.1673    | 1.3 x 10$^{-5}$ |

| Table 5 Mixed effects estimated correlation coefficient Correlation coefficient |
|---------------------------------|------------------|
| X                               | 0.5329           |
| Y                               | 0.8038           |
| Z                               | 0.7336           |
for agreement there is greater interpretability of statistical output, with no loss of clinical relevance. For instance, one could specify a priori that measurement differentials > 1 mm would represent a lack of interchangeability globally. Data presentation was performed in this study in an effort to illustrate potential applications of COM3PARE for replicated image-based measurements of the kind frequently encountered in radiation oncology. The specific dataset included have been previously presented using standard method approaches. By revisiting these data using compare we hope to illustrate implementation of what we perceive to be a more usable and parsimonious approach to conceptualizing method comparison for IGRT applications, expanding upon, rather than obviating the previous work. With regard to the specific dataset presented herein, our analysis points to the difficulties possible when comparing IGRT platforms. For instance, having set our criteria pre-analysis, we were surprised to note that differing measurement methods proved preferable in distinct axes (e.g., CBCT in X-axis, kV X-ray for the Z-axis), while appearing by said criteria interchangeable in the Y-axis. A possible explanation of this phenomenon may appear as a feature of the imaging methodologies themselves. For CBCT, before three-dimensional reconstruction, data is acquired as axial slices (X-axis), while, previous to DRR referencing, the kV X-ray system uses orthogonal projections at oblique angles, parallel to the superoinferior plane (Z-axis). Consequently, method intra-subject repeatability may be tied to the reference plane of image acquisition, though this remains conjecture based on a single dataset. To our knowledge this technique represents the first formal hypothesis testing approach to integrate inter-method bias, inter-subject variability, and intra-subject variability of two methods with any number of replicated measurements for image-guided radiotherapy. As modeled on the aforementioned conceptual schema presented in the “Methods” section, we postulate that the following criteria be formally evaluated as feature of future image-guided radiotherapy measurement comparison studies comparing two imaging platforms, where multiple repeated observations

![Flowchart of COM3PARE schema for IGRT device comparison](image-url)
on the same subject is possible. To meet our criteria for interchangeability [48]:

1. The bias and overall agreement must fall within a pre-specified range (e.g., bias/agreement of < 0.1 cm between IGRT devices).
2. There should be no statistically significant, using a pre-specified threshold (e.g., < 0.05) difference in the inter-subject variability of the two methods.
3. There should be no statistically significant difference in the intra-subject variability (i.e., repeatability) of the two methods.
4. In cases where criteria 2 and 3 are NOT met, the preferred IGRT technique is the one exhibiting the lower intra-subject variability (i.e., greater repeatability).

These criteria are presented as a graphical schema (Fig. 1); notably analysis of criteria 1–3 is easily incorporated in a single step using the COM3PARE SAS Code (Appendix A). The a priori criteria set we specified for interchangeability represented what we considered a reasonable metrics for the given application (i.e. fractionated radiotherapy of 30+ fractions for head and neck cancer) with a standardized PTV margin. The COM3PARE methodology, however, allows specification of any specified difference/p-value combination. Consequently, if a scenario arose whereby either tighter tolerances are desirable (e.g. 3-fraction SBRT), such parameters can be easily defined as an acceptability criteria.

Conclusion
COM3PARE represents an attempt at a unified conceptual schema and analytic algorithm for method comparison of IGRT platforms. Initial application in a head and neck positional verification dataset shows feasibility and utility.

Appendix A
SAS code
Below we provide the sample SAS code to test within-subject variabilities by fitting the linear mixed effects model to our KVX and CBCT shifts data under the null hypothesis that the two methods do not have the same within-subject variabilities.

```sas
/* Here we give the SAS code for fitting the linear mixed effects model to the our KVX and CBCT shifts data under the null hypothesis that the two methods do not have the same within-subject variabilities. */
options ls=64 ps=68 nocenter;
proc import datafile="c:\CBCTKVX.xls" DBMS=EXCELSAS out=sasuser.max2 replace;
/* proc print data=sasuser.max2; run; */
data KVXCBCTdataX; set sasuser.max2; keep pat replicate xm1kvx xm2cbct;
data b; SET KVXCBCTdataX; y=xm1kvx; mvar='1'; output; y=xm2cbct; mvar='2'; output; /* proc print data=b; run; */
proc mixed data=b method=ml covtest; classes pat mvar replicate; model y=mvar / s ddfm=kr; random mvar / type=un subject=pat v vcorr g; repeated mvar/type= cs subject=replicate (pat) r rcorr; run;
/* Here we give the SAS code for fitting the linear mixed effects model to the same data under the alternative hypothesis that the two methods have the same within-subject variabilities. */
proc mixed data=b method=ml covtest; classes pat mvar replicate; model y=mvar / s ddfm=kr; random mvar / type=un subject=pat v vcorr g; repeated mvar/type= cs subject=replicate (pat) r; run;
/* Calculating the p-value to test the within-subject variabilities of the two methods */
proc iml;
pval=1- probchi(34.0, 1);
print pval; run;
```
Appendix B
SAS output for covariance structure under the null and the alternating hypotheses
Below we provide the selected portions of the output of the above program.
/* The following output is for the null hypothesis */

Estimated V Matrix for Pat 3
Row Col1 Col2 Col3 Col4 Col5 Col6
1 0.3809 0.1363 0.04126 0.03663 0.04126 0.03663
2 0.1363 0.1717 0.03663 0.06695 0.03663 0.06695
3 0.04126 0.03663 0.3809 0.1363 0.04126 0.03663
4 0.03663 0.06695 0.1363 0.1717 0.03663 0.06695
5 0.04126 0.03663 0.04126 0.03663 0.3809 0.1363
6 0.03663 0.06695 0.03663 0.06695 0.1363 0.1717

Covariance Parameter Estimates
Cov Parm Subject  Estimate  Standard Error  Z Value  Pr Z
UN(1,1) Pat  0.04126  0.03079  1.34  0.0902
UN(2,1) Pat  0.03663  0.02384  1.54  0.1244
UN(2,2) Pat  0.06695  0.02669  2.51  0.0061
UN(1,1) Replicate(Pat)  0.3396  0.05238  6.48 <.0001
UN(2,1) Replicate(Pat)  0.09963  0.02359  4.22 <.0001
UN(2,2) Replicate(Pat)  0.1047  0.01688  6.20 <.0001

Fit Statistics
-2 Log Likelihood  239.7
AIC (smaller is better)  255.7
AICC (smaller is better)  256.5
BIC (smaller is better)  266.4
Null Model Likelihood Ratio Test
DF Chi-Square Pr > ChiSq
5  74.98 <.0001

Solution for Fixed Effects
Effect  mvar  Estimate  Standard Error  DF  t Value  Pr > |t|
Intercept 0.03022  0.06212  28.7  0.49  0.6304
mvar 1  0.03349  0.06447  26.5  0.52  0.6077
mvar 2  0

/* The following output is for the alternating hypothesis */

Fit Statistics
-2 Log Likelihood  273.7
AIC (smaller is better)  287.7
AICC (smaller is better)  288.2
BIC (smaller is better)  297.0
Null Model Likelihood Ratio Test
DF Chi-Square Pr > ChiSq
4  41.03 <.0001

/* The following output is for the p-value to test the within-subject variabilities of the two methods */
pval
5.5112E-9
References
1. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307–10.
2. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet. 1995;346:1085–7.
3. Bland JM, Altman DG. Comparing two methods of clinical measurement: a personal history. Int J Epidemiol. 1995;24(Suppl 1):S7–14.
4. Bland JM, Altman DG. Measurement error. BMJ (Clinical research ed.) 1996;312:1654.
5. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8:135–60.
6. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol. 2003;22:85–93.
7. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat. 2007;17:571–82.
8. Biswas S, Konstantinou G, Surlan-Popovic K, Khoshnevazadeh A, Baghi M, Vogl TJ, et al. Dynamic contrast-enhanced CT of head and neck tumors: comparison of first-pass and permeability perfusion measurements using two different commercially available tracer kinetics models. Acad Radiol. 2008;15:1580–9.
9. Cronin P, Saab A, Kelly AM, Gross BH, Patel S, Kazerooni EA, Carlos RC. Measurements of pulmonary vein ostial diameter and distance to first bifurcation: A comparison of different measurement methods. Eur J Radiol. 2009;71(1):61–8.
10. Kanze RE, Higashino K, Kido T, Kurata A, Saito M, Sugawara Y, Mochizuki T. Quantitative assessment of regional left ventricular wall thickness and thickening using 16 multidetector-row computed tomography: comparison with cine magnetic resonance imaging. Radiat Med. 2007;25:119–26.
11. Kwee TC, Takahara T, Koh DM, Nievieinstein RA, Luitjen PR. Comparison and reproducibility of ADC measurements in breathhold, respiratory triggered, and free-breathing diffusion-weighted MRI imaging of the liver. Magn Reson Imaging. 2008;26:1141–8.
12. Mähnchen AH, Spuentrup E, Niethammer M, Buecker A, Boese J, Wildberger JE Flohr T, et al. Quantitative and qualitative assessment of left ventricular volume with ECG-gated multislice CT: value of different image reconstruction algorithms in comparison to MRI. Acta Radiol. 2003;44:604–11.
13. Martin KE, Helvie MA, Zhou C, Roubidoux MA, Bailey IE, Paramagul C, et al. Mammographic density measured with quantitative computer-aided method: comparison with radiologists’ estimates and BI-RADS categories. Radiology. 2006;240:656–65.
14. Murakami R, Uozumi H, Hirai T, Nishimura R, Katsuragawa S, Shitashii S, Toya R, Tashiro K, Kawanaka K, Oya N, Tomiguchi S, Yamashita Y. Impact of FDG-PET/CT fused imaging on tumor volume assessment of head-and-neck squamous cell carcinoma: intermethod and interobserver variations. Acta Radiol. 2008;49:693–9.
15. Nieman K, Shapiro MD, Ferec Mir, Nomura CH, Abbara S, Hoffmann U, et al. Reperfused myocardial infarction: contrast-enhanced 64-Slice CT in comparison to MRI imaging. Radiology. 2008;247:49–56.
16. Agarzayan N, Tenn SE, Desettes AA, Selch MT. Image-guided radiosurgery for spinal tumors: methods, accuracy and patient inraction motion. Phys Med Biol. 2008;53:1715–27.
17. Dawson LA, Jaffray DA. Advances in image-guided radiation therapy. J Clin Oncol. 2007;25:938–946.
18. Grills IS, Hugo G, Kerstn L, Galaneri AP, Chao KK, Wloch J, et al. Image-guided radiotherapy via daily online cone-beam CT substantially reduces margin requirements for stereotactic lung radiotherapy. Int J Radiat Oncol Biol Phys. 2008;70:1045–56.
19. Ippolito E, Mertens I, Haustermans K, Gambacorta MA, Pasini D, Valentinii V. IGR in rectal cancer. Acta oncologica (Stockholm, Sweden). 2008;47:1317–24.
20. Kupelian PA, Langen KM, Willoughby TR, Zeidan OA, Meeks SL. Image-guided radiotherapy for localized prostate cancer: treating a moving target. Semin Radiat Oncol. 2008;18:58–66.
21. Wertz H, Lohr F, Dobler B, Mai S, Wenz F. Dosimetric impact of image-guided translational isocenter correction for 3D conformal radiotherapy of the prostate. Strahlenther Onkol. 2007;183:203–10.
22. Wiersma RD, Mao W, Xing L. Combined Kv and MV imaging for real-time tracking of implanted fiducial markers. Med Phys. 2008;35:1191–8.
23. Yorke ED, Keall P, Verhaegen F. Anniversary paper: Role of medical physicists and the AAPM in improving geometric aspects of treatment accuracy and precision. Med Phys. 2008;35:828–39.
24. Fuller CD, Thomas CR, Schwartz S, Golden N, Ting J, Wong A, et al. Method comparison of ultrasound and kilovoltage x-ray fiducial marker imaging for prostate radiotherapy targeting. Phys Med Biol. 2006;51:4981–93.
25. Moseley DJ, White EA, Wiltshire KL, Rosewall T, Sharpe MB, Siewersen JH, et al. Comparison of localization performance with implanted fiducial markers and cone-beam computed tomography for on-line image-guided radiotherapy of the prostate. Int J Radiat Oncol Biol Phys. 2007;67:942–53.
26. Antignan K, Smitsmans MH, Lebesque JV, Jaffray DA, van Her M, Bartelink H. Online ultrasound image guidance for radiotherapy of prostate cancer: impact of image acquisition on prostate displacement. Int J Radiat Oncol Biol Phys. 2004;59:595–601.
27. Cury FL, Shonouda G, Souhami L, Duclos M, Faria SL, David JH, et al. Comparison of localization performance with implanted fiducial markers and cone-beam computed tomography for on-line image-guided radiotherapy of the prostate. Int J Radiat Oncol Biol Phys. 2007;67:942–53.
28. Fung AY, Ayyangar KM, Dajaputra D, Nehru RM, Enke CA. Ultrasound-based guidance of intensity-modulated radiation therapy. Int J Radiat Oncol Biol Phys. 2006;66:1562–7.
29. Fuss M, Salter BJ, Cavanaugh SX, Fuss C, Sadeghi A, Fuller CD, et al. Daily ultrasound-based image-guided targeting for radiotherapy of upper abdominal malignancies. Int J Radiat Oncol Biol Phys. 2004;59:1245–56.
30. Langen KM, Pouliot J, AnzORIZEDnoc C, Aubin M, Gottschalk AR, Hsiu IC, et al. Evaluation of ultrasound-based prostate localization for image-guided radiotherapy. Int J Radiat Oncol Biol Phys. 2003;57:635–44.

31. Patel RR, Orton N, Tome WA, Chappell R, Ritter MA. Rectal dose sparing with a balloon catheter and ultrasound localization in conformal radiation therapy for prostate cancer. Radiother Oncol. 2003;67:285–94.

32. Choi M, Fuller CD, Wang SJ, Siddiqi A, Wong A, Thomas Jr CR, Fuss M. Effect of body mass index on shifts in ultrasound-based image-guided intensity-modulated radiation therapy for ablative malignancies. Radiolther Oncol. 2009;91(1):114–9.

33. Gayou O, Milten F. Comparison of mega-voltage cone-beam computed tomography prostate localization with online ultrasound and fiducial markers methods. Med Phys. 2008;35:531–8.

34. Johnston H, Hilsit M, Bedkhalm W, Berthelet E. 3D ultrasound for prostate localization in radiation therapy: a comparison with implanted fiducial markers. Med Phys. 2008;35:2403–13.

35. Aubry JF, Beaulieu L, Girouard LM, Aubin S, Tremblay D, Lapierre J, et al. Measurements of intrafraction motion and interfraction and intrafraction rotation of prostate by three-dimensional analysis of daily portal imaging with radiopaque markers. Int J Radiat Oncol Biol Phys. 2004;60:30–9.

36. Bailer JM, Sandler HM, Lam K, Bree RL, Lichter AS, Haken TENRK. Measurement of prostate movement over the course of routine radiotherapy using implanted markers. Int J Radiat Oncol Biol Phys. 1995;31:113–8.

37. Evans PM. Anatomical imaging for radiotherapy. Phys Med Biol. 2008;53:R151–191.

38. Lometti MW, Thurston D, Aubin M, Bock A, Verhey L, Lockhart JM, Bland R, Pouliot J, Roach 3rd M. Are lateral electronic portal images adequate for accurate on-line daily targeting of the prostate? Results of a prospective study. Med Dosim. 2008;33:22–9.

39. Somsen de Kostevan JR, Cuijpers JP, Geest de Fijn, Lagenward FJ, Slotman BJ, Senan S. Verifying 4D gated radiotherapy using time-integrated electronic portal imaging: a phantom and clinical study. adiat Oncol (London, England). 2007;2:32.

40. Morin O, Gils A, Chen J, Aubin M, Bucc MK, Roach 3rd M, et al. Megavoltage cone-beam CT: system description and clinical applications. Med Dosim. 2006;31:51–61.

41. Pang G, Bani-Hashemi A, Au P, O'Brien PF, Rowlands JA, Morton G, et al. Megavoltage cone-beam digital tomosynthesis (MV-CBDT) for image-guided radiotherapy: a clinical investigational system. Phys Med Biol. 2008;53:999–1013.

42. Pisani L, Lockman D, Jaffrey D, Yan D, Martinez A, Wong J. Setup error in radiotherapy: on-line correction using electronic kilovoltage and megavoltage radiographs. Int J Radiat Oncol Biol Phys. 2000;47:825–39.

43. Pouliot J, Bani-Hashemi A, Chen J, Svatos M, Ghelmansasai F, Mitsche M, et al. Low-dose megavoltage cone-beam CT for radiation therapy. Int J Radiat Oncol Biol Phys. 2005;61:552–60.

44. Serago CT, Buskirk SJ, Igel TC, Gale AA, Serago NE, Earle JD. Comparison of daily megavoltage electronic portal imaging or kilovoltage imaging with marker seeds to ultrasound imaging or skin marks for prostate localization and treatment positioning in patients with prostate cancer. Int J Radiat Oncol Biol Phys. 2006;65:1585–92.

45. Wu QJ, Godfrey DJ, Wang Z, Zhang J, Zhou S, Yoo S, Brizel DM, Yin FF. On-board patient positioning for head-and-neck IMRT: comparing digital tomosynthesis to kilovoltage radiography and cone-beam computed tomography. Int J Radiat Oncol Biol Phys. 2007;69:598–606.

46. Scarbrough TJ, Golden NM, Ting JY, Fuller CD, Wang A, Kupelian PA, Thomas Jr CR. Comparison of ultrasound and implanted seed marker prostate localization methods: implications for image-guided radiotherapy. Int J Radiat Oncol Biol Phys. 2006;65:378–87.

47. Stutzel J, Oelke U, Nill S. A quantitative image quality comparison of four different image guided radiotherapy devices. Radiat Oncol. 2008;6:20–9.

48. Roy A. An application of linear mixed effects model to assess the agreement between two methods with repeated observations using mixed effects model. Biometrical J. 2006;48:286–301.

49. Cui Y, Galvin JM, Straube WL, Bosch WR, Purdy JA, Li XA, Xiao Y. Multi-System Verification of Registrations for Image-Guided Radiotherapy in Clinical Trials. Int J Radiat Oncol Biol Phys. 2011;81(1):305–312.

50. Mohammed N, Kestin L, Grills I, Shah C, Glode-Hurst C, Yan D, et al. Comparison of IGRT Registration Strategies for Optimal Coverage of Primary Lung Tumors and Involved Nodes Based on Multiple Four-Dimensional CT Scans Obtained Throughout the Radiotherapy Course. Int J Radiat Oncol Biol Phys. 2012;82(4):1541–8.

51. Shi W, Li JG, Zlatecki RA, Yeung A, Newlin H, Palta J, et al. Evaluation of kV cone-beam CT performance for prostate IGRT: a comparison of automatic grey-value alignment to implanted fiducial-marker alignment. Am J Clin Oncol. 2011;34(1):116–21.

52. Thongpseud W, Wu QJ, Lee WR, Chankong V, Yoo S, Mahonora R, Yin FF. Comparison of online IGRT techniques for prostate IMRT treatment: adaptive vs repositioning correction. Med Phys. 2009;36:6515–62.

53. Borst GR, Sonke JJ, Bergten A, Remeijer P, Herk van M, Lebesque JV. Kilo-voltage cone-beam computed tomography setup: measurements of tolerance and comparison with electronic portal-imaging device. Int J Radiat Oncol Biol Phys. 2007;68:555–61.

54. Lin Li A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45:255–68.

55. Konings H. Use of Deming regression in method-comparison studies. Surv Immunol Res. 1982;1:771–74.

56. Martin RF. General deming regression for estimating systematic bias and its confidence interval in method-comparison studies. Clinical Chem. 2000;46:100–4.

57. Milliken G, Graybill F. Extensions of the General Linear Hypothesis Model. J Am Stat Assoc. 1970;65:797–807.

58. Lahlott L. A Canonical Form for the General Linear Model. Ann Stat. 1975;7:587–9.

59. Fennessey J. The General Linear Model: A New Perspective on Some Familiar Topics. Am J Sociol. 1968;74:1–27.

60. Martin RF. General deming regression for estimating systematic bias and its confidence interval in method-comparison studies. Clinical Chem. 2000;46:100–4.

61. Milliken G, Graybill F. Extensions of the General Linear Hypothesis Model. J Am Stat Assoc. 1970;65:797–807.

62. Lahlott L. A Canonical Form for the General Linear Model. Ann Stat. 1975;7:587–9.

63. Fennessey J. The General Linear Model: A New Perspective on Some Familiar Topics. Am J Sociol. 1968;74:1–27.