INSTITUTE OF PHYSICS PUBLISHING PHYSICS IN MEDICINE AND BIOLOGY

Phys. Med. Biol. **51** (2006) 4981–4993 doi:10.1088/0031-9155/51/19/016

Method comparison of ultrasound and kilovoltage x-ray fiducial marker imaging for prostate radiotherapy targeting*

Clifton David Fuller[1,2], Charles R Thomas Jr[3], Scott Schwartz[4], Nanalei Golden[5], Joe Ting[5], Adrian Wong[6], Deniz Erdogmus[3,7] and Todd J Scarbrough[3,5]

[1] Department of Radiation Oncology, University of Texas Health Science Center—San Antonio, San Antonio, TX, USA
[2] Graduate Program in Radiological Sciences, Department of Radiology, University of Texas Health Science Center—San Antonio, San Antonio, TX, USA
[3] Department of Radiation Medicine, Oregon Health & Science University, Portland, OR, USA
[4] Department of Mathematics, Trinity University, San Antonio, TX, USA
[5] MIMA Cancer Center, Melbourne, FL, USA
[6] Department of Diagnostic and Interventional Imaging, University of Texas Health Science at Houston, Houston, TX, USA
[7] Department of Computer Science and Electrical Engineering, and Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA

Received 2 May 2006, in final form 24 July 2006
Published 18 September 2006
Online at stacks.iop.org/PMB/51/4981

Abstract

Several measurement techniques have been developed to address the capability for target volume reduction via target localization in image-guided radiotherapy; among these have been ultrasound (US) and fiducial marker (FM) software-assisted localization. In order to assess interchangeability between methods, US and FM localization were compared using established techniques for determination of agreement between measurement methods when a 'gold-standard' comparator does not exist, after performing both techniques daily on a sequential series of patients. At least 3 days prior to CT simulation, four gold seeds were placed within the prostate. FM software-assisted localization utilized the ExacTrac X-Ray 6D (BrainLab AG, Germany) kVp x-ray image acquisition system to determine prostate position; US prostate targeting was performed on each patient using the SonArray (Varian, Palo Alto, CA). Patients were aligned daily using laser alignment of skin marks. Directional shifts were then calculated by each respective system in the *X*, *Y* and *Z* dimensions before each daily treatment fraction, previous to any treatment or couch adjustment, as well as a composite vector of displacement. Directional shift agreement in each axis was compared using Altman–Bland limits of agreement, Lin's concordance coefficient with Partik's grading schema, and Deming orthogonal

* Portions of this data were presented at the American Society of Clinical Oncology/American Society for Therapeutic Radiology and Oncology/Society of Surgical Oncology 2006 Prostate Cancer Symposium, San Francisco, CA, USA.

bias-weighted correlation methodology. 1019 software-assisted shifts were suggested by US and FM in 39 patients. The 95% limits of agreement in X, Y and Z axes were $\pm$9.4 mm, $\pm$11.3 mm and $\pm$13.4, respectively. Three-dimensionally, measurements agreed within 13.4 mm in 95% of all paired measures. In all axes, concordance was graded as 'poor' or 'unacceptable'. Deming regression detected proportional bias in both directional axes and three-dimensional vectors. Our data suggest substantial differences between US and FM image-guided measures and subsequent suggested directional shifts. Analysis reveals that the vast majority of all individual US and FM directional measures may be expected to agree with each other within a range of 1–1.5 cm. Since neither system represents a gold standard, clinical judgment must dictate whether such a difference is of import. As IMRT protocols seek dose escalation and PTV reduction predicated on US- and FM-guided imaging, future studies are needed to address these potential clinically relevant issues regarding the interchangeability and accuracy of novel positional verification techniques. Comparison series with multiple image-guidance systems are needed to refine comparisons between targeting methods. However, we do not advocate interchangeability of US and FM localization methods.

M This article features online multimedia enhancements

(Some figures in this article are in colour only in the electronic version)

Introduction

Modern radiotherapy allows the delivery of radiation dose with impressive conformality (Eng *et al* 2005). This conformality necessitates steep dose gradients which allow a substantial dose differential between targeted tumours and adjacent tissues (Byrne 2005, Schaly *et al* 2005). Yet steep dose gradients also provide potential for over- or under-dosing structures with comparatively small geometric positional differences (Keall *et al* 1999, van Herk *et al* 2002, Bos *et al* 2005); these structures include gross tumour volumes (GTVs), clinical tumour volumes (CTVs), and proximal organs at risk (OARs). The major limitation of dose escalation is the dose tolerance of OARs. Delivery of a high dose to GTVs and CTVs is further complicated by the fact that many tumour volumes and organs, the prostate in particular, are known to vary in spatial position relative to bony anatomy and skin alignment marks (Crook *et al* 1995, Purdy 2002). This deviation is accounted for by the addition of a planning target volume (PTV) which compensates for targeting and set-up uncertainty (McKenzie *et al* 2002, Stroom and Heijmen 2002).

In an effort to reduce PTV margins and/or escalate tumour dose without crossing the biological dose thresholds of OARs, several groups have explored utilization of ultrasound (US) guidance for prostate radiotherapy (Patel *et al* 2003, Chandra *et al* 2003, Kuban *et al* 2005). These efforts have demonstrated the localization capabilities of the technique using software-computed table shifts to align the patient isocentre before every radiotherapy fraction. Simultaneously, others have explored several methods of target localization using kilovoltage (kV) x-ray imaging of radio-opaque seed implant fiducial markers (FMs) which are placed within the prostate (Poggi *et al* 2003, Herman *et al* 2003, Schallenkamp *et al* 2005, Van den Heuvel *et al* 2003, Welsh *et al* 2004, Zhang *et al* 2006). These seeds may then be visualized by physicians to align patient anatomy.

While series on US and FM localization individually are numerous within the scientific literature, few series have sought to characterize relational agreement between these methods (Scarbrough *et al* 2006, Van den Heuvel *et al* 2003). To a great degree this may be due to the difficulty of method comparison: when comparing the relationship between two distinct mechanisms of measurement (even when they utilize the same proportional scale) the use of standard statistical approaches is sub-optimal (Bland and Altman 1986). While commonly utilized for other diagnostic measures, as in clinical chemistry, method comparison techniques have been utilized sporadically in the imaging literature (Hilson 2004, Bland and Altman 2003, Bielak *et al* 2001, Mahnken *et al* 2005, Schramm *et al* 2004, Bradley *et al* 1999) and rarely in the radiotherapy literature. The specific aims of this study included

(1) evaluation of distributional normality of daily US and FM software-suggested shifts on a series of patients imaged using both techniques;
(2) examination of US and FM paired software-suggested shifts using accepted strategies for method comparison;
(3) illustration of method comparison approaches in target localization for image-guided radiotherapy;
(4) generation of hypotheses for future prospective research.

Methods

Simulation

For a non-random cohort of 40 men undergoing definitive external beam radiotherapy, four 1.2 × 3 mm gold FMs (NMPE, Inc., Orange City, IA) were placed under sagittal ultrasound guidance via preloaded needles. FMs were implanted by a simple transrectal technique at apex, mid-gland, and left and right base positions without any anaesthesia. At least 3 days were then allowed to pass before computed tomography (CT) simulation (GE LightSpeed Plus, GE Healthcare, Fairfield, CT). For all patients, 2.5 mm slice thicknesses were obtained through the pelvis and prostate for non-contrasted CT simulation. A solid foam 12 inch × 12 inch × 4 inch (30 cm × 30 cm × 10 cm, approximately) block braced between the patient's feet and a soft ring hand holder was used for patient immobilization. The isocentre was placed near the centre of the prostate and marked by anterior and lateral skin marks with alignment to in-room lasers. The original simulation skin marks were not altered during therapy. Target volumes and OARs were contoured with Eclipse (Varian Medical Systems, Palo Alto, CA) treatment planning software. Isocentre location and organ contours were then uploaded to both localization systems. Daily positioning was recorded for both localization systems resulting in 1019 paired measures.

Localization systems

Ultrasound prostate localization. The SonArray system allows coregistration of real-time US images with the CT images/dataset. The device itself consists of an optically-guided 3D-ultrasound target localization system using an ultrasound transducer with affixed infrared (IR) reflective registration markers. Simultaneously, a series of rigidly affixed markers on the treatment couch are imaged using IR. The system calculates probe and couch position in three dimensions and correlates this with the registered US/CT offsets in vertical, longitudinal and lateral axes. For the purposes of this study, seed markers were contoured on CT simulation and visually referenced during set-up. This system is operator-driven insofar as the location of the prostate image in real-time US is determined to be aligned with the CT-derived volume

overlay. A lateral (X), vertical (Y) and longitudinal (Z) shift correction is calculated to align the patient geometrically with the pre-determined isocentre. At each initial patient set-up, the radiation oncologist verified the US/CT match which established a visual template for subsequent users to employ during treatment.

Kilovoltage x-ray localization of fiducial markers. The ExacTrac system consists of an automated kV imaging system designed to utilize radio-opaque anatomic or FMs as points of reference to ascertain optimum alignment of the patient. FMs were localized at the ExacTrac computer station by the radiation oncologist by delineating the FMs as seen on the simulation CT in transverse, sagittal and coronal views. At each subsequent treatment, two non-coplanar (stereoscopic) kV x-ray (XFM) images of the seeds were analysed by the system using its automated seed marker matching function. A set of extrapolated X, Y and Z shift corrections were then calculated and recorded.

Both systems were calibrated daily according to manufacturer-specified protocols. All therapists in the centre were rigorously trained by the manufacturers and had several months of experience with the systems prior to the inception of this retrospective analysis. In addition, the radiation oncologists at the centre verified the therapists' skills prior to study commencement. US and FM registration of all patients was physician-verified and visualized over the course of the study.

Measurement technique

Patients were placed on the treatment couch, immobilized and aligned to skin marks. Trans-abdominal ultrasound images were obtained using a simple one-pass transverse sweep over the suprapubic region. US-derived X, Y and Z shifts were calculated and logged, but the patient was not shifted. Next, kV x-ray FM images were obtained, and FM-derived X, Y and Z shifts were calculated and logged. The patient was then shifted (based on FM analysis) and treated.

Statistical analysis

For a daily patient measure, the origin was defined as the point in space marked by the initial position using skin marks. Using a three-dimensional Cartesian coordinate system, measurement systems register this as the 'zero point' having $X/Y/Z$ coordinates of 0, 0, 0. Outputs from both systems were recorded in mm in a given X, Y or Z axis. The US system returns shifts that have a lateral (left/right) coordinate of X_{US}, vertical (AP/PA) coordinate of Y_{US} and longitudinal (superior/inferior) coordinate of Z_{US}. The FM system yields shifts of X_{FM}, Y_{FM} and Z_{FM}. The absolute 3D shift vectors were calculated and designated as $3D_{US}$ and $3D_{FM}$, respectively.

Statistical analysis was performed using JMP v5 (SAS Institute, Cary, NC, USA) and Analyse-IT (Analyse-It Software, Ltd, Leeds, UK) software packages. Analysis consisted of descriptive moment analyses, distributional analysis of normality using the Shapiro–Wilks test (Shapiro *et al* 1968), agreement by Altman Bland limits of agreement and bias estimation, concordance by Lin's concordance correlation, and constant/proportional bias detection by Deming orthogonal regression technique.

Altman–Bland analysis illustrates method agreement in the cases when an established 'gold-standard' has not been determined (Bland and Altman 1999, 1997, 1995a, 1986). It consists of a graphic plot of the difference between paired measures. The technique calculates a limit of agreement (LOA). The 95% LOA (defined as the mean of the differences between paired measures ± 1.96 times the standard deviation of the differences) is the range within

Table 1. Descriptive statistics for all directional axes and combined 3D vectors, by imaging modality; all values in mm.

	X_{US}	Y_{US}	Z_{US}	$3D_{US}$	X_{FM}	Y_{FM}	Z_{FM}	$3D_{FM}$
Median	1.5	−0.7	2.7	8.7	0.7	0.8	−0.2	7.3
Mean	2.0	−0.8	3.5	9.8	0.8	1.1	0.0	8.1
Upper 95% CI	2.3	−0.4	3.9	10.2	1.1	1.4	0.3	8.4
Lower 95% CI	1.6	−1.2	3.1	9.4	0.5	0.7	−0.3	7.8
Variance	29.0	40.5	50.9	40.7	20.7	36.5	27.0	20.8
SD	5.4	6.4	7.1	6.4	4.6	6.0	5.2	4.6
Minimum	−30.0	−40.4	−105.0	0.3	−21.9	−22.0	−36.7	0.5
Maximum	54.8	27.0	30.7	105.0	51.0	22.0	16.7	64.8
Skewness	1.1	−0.2	−3.1	4.3	1.5	−0.2	−0.2	2.8
Kurtosis	13.5	2.1	52.5	51.8	15.3	1.2	2.7	24.6

which 95% of all differences between measures may be predicted to occur. Additionally, bias or measurement differential component estimates may be illustrated using LOA plots. The systematic bias estimation (SBE; denoted as the mean difference of each paired measure) and random bias estimation (RBE; calculated as the standard deviation of the difference between paired measurements) were calculated. The 95% LOA was then calculated for each axis to represent the total bias estimation between devices. The LOA allows clinicians to determine how interchangeable two distinct techniques are, as the 95% LOA indicates a range of agreement between methods.

Lin's concordance coefficient is an evaluative measure of concordance between techniques (Lin 1989, Arbillaga *et al* 2002, King and Chinchilli 2001a, 2001b). Designed to determine how well a given set of observations reproduce an original series, it is a calculated product of two components: the Pearson correlation coefficient (r) and a bias correction factor (C_b). The methods are regarded as equivalent if the coefficient is greater than a specified threshold (agreement increases as ρ approaches 1.0). The calculations utilized for Lin's coefficient of concordance are given in the appendix. Partik's categorical grading criteria were utilized to grade the coefficient of covariance (Partik *et al* 2002); Partik's criteria grade ρ values larger than 0.95 as 'excellent', >0.90 as 'very good', >0.80 as 'fairly good', >0.70 as 'middling/satisfactory', >0.60 as 'mediocre', >0.50 as 'poor', and ⩽0.50 as 'unacceptable'.

Deming orthogonal regression (Payne 1997, Linnet 1998, 1999b, Stockl *et al* 1998) was used to illustrate bias-weighted correlation between measures and to detect constant and proportional bias. Deming orthogonal regression is often encountered in clinical measure series (Konings 1982). Generally, ordinary linear regression is to be avoided for method comparisons (Bland and Altman 1986) as only a single axis is postulated to exhibit variance; the orthogonal regression plots circumvent this shortcoming by incorporating variance for both measures (Martin 2000). The orthogonal fit ratio was calculated using the variance in bias between measures, and a population variance fit ratio (calculated as $[SD_{US}]^2/[SD_{FM}]^2$) was user-specified for each plot; both fits were illustrated against a plotted line of identity (slope = 1). If the 95% confidence interval for the intercept contains the value 0, the methods do not exhibit significant constant bias; if the 95% confidence interval for the intercept does not include 0, then both measurement methods differ and a constant bias component is extant between methods. Proportional bias may be similarly assessed. If the 95% confidence interval for the slope does not contain the value 1 (with slope = 1 representing a line of identity), significant proportional bias between methods exists.

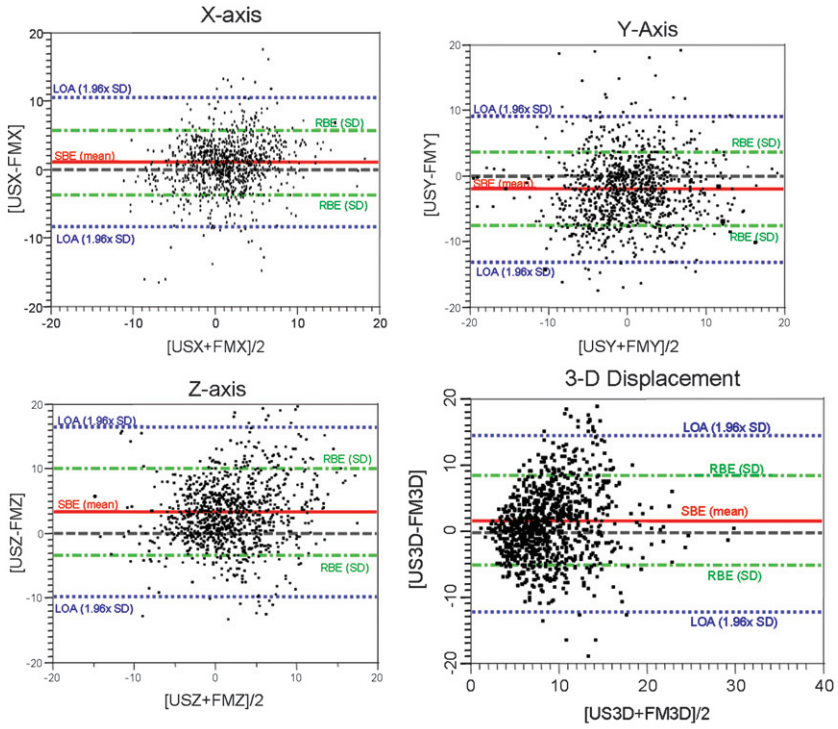

Figure 1. Altman–Bland bias plots of directional axes and vectors showing systematic bias estimate (SBE, solid red line), random bias estimate (RBE, dashed green line), 95% limits of agreement (LOA, dashed blue line) and identity line (dashed grey line).

Table 2. Altman–Bland between method bias and limits of agreement, with 95% CI.

	Systematic bias estimate (mm)	95% CI of systematic bias (mm)	Random bias estimate (mm)	95% CI of Random bias estimate (mm)	95% Limits of agreement (mm)	CI of limits of agreement (mm)
$X_{US}\|X_{FM}$	+1.2	+0.9–1.5	±4.7	±4.5–4.9	±9.4	±9.0–9.8
$Y_{US}\|Y_{FM}$	−1.9	−1.5–2.2	±5.7	±5.4–5.9	±11.3	±10.8–11.8
$Z_{US}\|Z_{FM}$	+3.5	+3.0–3.9	±6.7	±6.0–7.0	±13.4	±12.8–14.0
$3D_{US}\|3D_{FM}$	+1.7	+1.3–2.2	±6.8	±6.5–7.1	±13.4	±12.8–14.0

Results

A consecutive non-randomized cohort of 40 patients across 1019 separate daily radiotherapy fractions was analysed using the statistical methods described. Descriptive distributional data from all FM and US measurements are given in table 1. For $\alpha = 0.05$, if the calculated Wilks value (W) is <0.05, the null hypothesis (H_0: the data are normally distributed) is rejected. If $W \geqslant 0.05$, the null fails to be rejected. The Y_{US} directional axis distribution suggests a normal distribution ($W = 0.985$). The other directional axes were found to exhibit a significantly non-Gaussian distribution pattern ($W \leqslant 0.01$ for all others).

Altman–Bland LOAs revealed SBE for X, Y and Z axes of +1.2 mm, −1.9 mm and +3.5 mm, with positive values indicating US > FM. X_{RBE} was ±4.7 mm, Y_{RBE} was ±5.7 mm

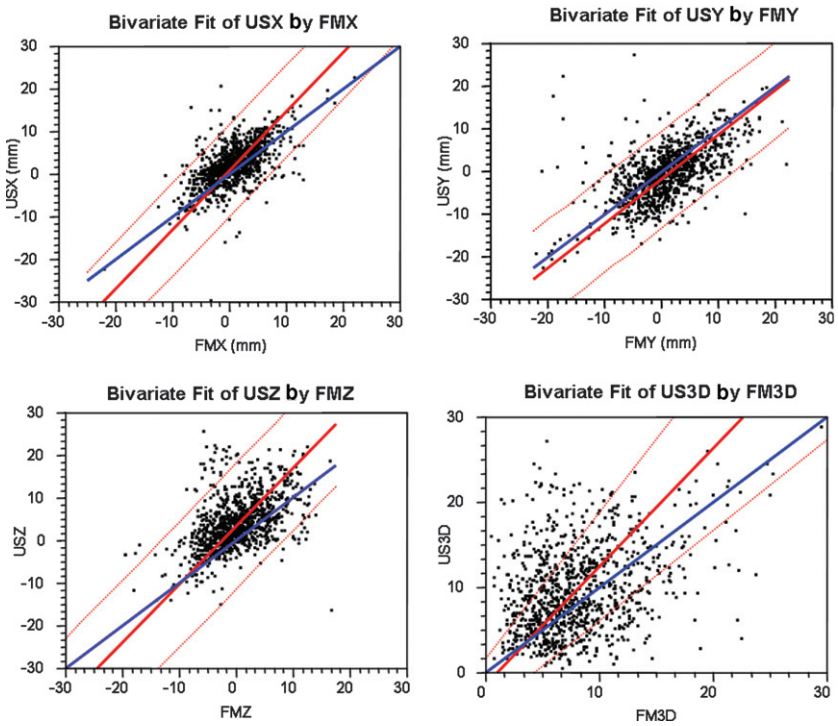

Figure 2. Deming bivariate orthogonal regression for each axis and vector (red line); showing 95% CI (parallel red dashed lines) and line of identity (slope = 1; blue line).

Table 3. Components of Lin's concordance coefficient and evaluation with Partik's criteria; r = correlation coefficient, C_b = bias correction factor, ρ = Lin's concordance coefficient. All values are in millimetres.

	r	C_b	ρ	Concordance
$X_{US}\|X_{FM}$	0.5	0.7	0.4	Poor
$Y_{US}\|Y_{FM}$	0.5	0.9	0.5	Unacceptable
$Z_{US}\|Z_{FM}$	0.4	0.5	0.2	Unacceptable
$3D_{US}\|3D_{FM}$	0.2	0.6	0.1	Unacceptable

and Z_{RBE} ±6.7 mm. The 95% LOAs were ±9.4 mm, ±11.3 mm and ±13.4 mm, in the X, Y and Z axes respectively. Vector differential distribution was noted to exhibit SBE, RBE, and 95% LOAs of +1.7, ±6.8 and 13.4 mm, respectively. Bias estimates and 95% CIs are shown in table 2 and illustrated graphically in figure 1.

Pearson's R (r), the bias correction (C_b) component, and Lin's correlation coefficient (ρ) are listed in table 3, as are the categorical assessment of said level of concordance using Partik's grading schema. Orthogonal regression plots using the specified orthogonal variance fit ratio for each axis and vector with line of identity and 95% CI are illustrated in figure 2 and table 4.

Table 4. Deming orthogonal regression results, all values in mm.

	Intercept	95% CI of intercept	Slope	95% CI of slope
$X_{US}\|X_{FM}$	0.9	0.3–1.4[a]	1.4	1.3–1.5[b]
$Y_{US}\|Y_{FM}$	−1.9	−2.5–−1.5[a]	1.1	1.0–1.1
$Z_{US}\|Z_{FM}$	3.5	2.6–4.3[a]	1.4	1.2–1.6[b]
$3D_{US}\|3D_{FM}$	−1.4	−4.4–1.6	1.4	1.1–1.7[b]

[a] Constant bias component.
[b] Proportional bias component.

Discussion

In radiotherapy and medical physics applications, even minor geometric uncertainties may affect cure rates and normal tissue complication probabilities (Bel *et al* 1996, Bos *et al* 2005, Keall *et al* 1999, Stroom *et al* 1999, Stroom and Heijmen 2002, van Herk *et al* 2000, 2002, 2003). Comparing distinct methods of measurement is a vexing problem. Typical comparative statistical techniques assume operational equivalence with regard to the measurement devices being compared; they are incorrectly applied when measurement devices differ ('apples and oranges', etc). For instance, standard means comparison (*t*-test) assumes that comparison groups are measured using the same scalar system. Thus, for example, if we were to compare the weights of two groups of persons on the same scale, a *t*-test might be an appropriate statistical tool. However, method comparison is more akin to measuring the agreement between two different scales, without knowing which is more accurate.

The necessity for developing accurate measure comparison techniques is neither novel nor obscure (Dunn and Roberts 1999). While tempting, reliance on means comparison (*t*-test), correlation coefficients or least-squares regression are statistically inappropriate for method comparison series, as illustrated by Altman and Bland in a series of some of the most frequently cited papers in the medical literature (Bland and Altman 1986, 1995b, 1995a, 1997, 1999, 2003). Using the standard Student's *t*-test for method comparison is flawed, as mean comparison carries the *a priori* assumption of normal distributions for both groups using a standardized measurement system, and are suspect when proportional error is high (Linnet 1999a). Similarly, standard correlation coefficients account not for agreement, but only for the strength of a bivariate relationship and are insensitive to changes in the scale of measurement (Westgard and Hunt 1973). Least-squares regression is also limited as variance is assumed in only one of the two measured variables and fails to account for outliers within the dataset (Martin 2000). Consequently, the specific statistical techniques that should be utilized for method comparison are not typically encountered in between groups testing. Thus for the purposes of this series we sought to utilize several established techniques in order to illustrate, for future series, accepted techniques for method comparison of image-guidance measures.

In a previous manuscript (Scarbrough *et al* 2006), we compared the independent distributions and inter-method error of US and FM image-acquisition and targeting methods and compared them to extant series (Fung *et al* 2006, Van den Heuvel *et al* 2003, Langen *et al* 2003, Chandra *et al* 2003, Huang *et al* 2002). We also sought to determine the overall systematic and random error of both US and FM imaging independently and compare calculated image-modality-specific margins using recommended 'margin recipes' (Stroom and Heijmen 2002, van Herk 2004)). In this analysis, we sought to explore comparative techniques that afford relational comparison in order to illustrate several established ways to surmount the 'apples and oranges' scenarios which come to the fore when comparing distinct measurement methods. For this reason we have sought to avoid confusion by refraining from utilizing the

term 'error' profligately. Since error typically refers to intra-method measurement uncertainty within the radiotherapy literature (van Herk 2004), its meaning must be carefully noted to avoid obfuscation of results. For such purposes, we have instead utilized the term 'bias'. While in the strict sense bias refers to a systematic deviation of a value from a reference, we utilize it to define 'the amount by which a specified set of values departs from a comparison set', in this case US- versus FM-suggested shifts from an initial position. The results of our series are designed to explore aspects of software-suggested targeting using two distinct systems. Since the systems represent independent measurement devices, we have sought to ascertain not just the magnitude of their difference but also to summarize measures which provide information on their relative relationships. It is important to note that while US and FM use the same scale (mm) and same initial reference point (skin marks) *they are distinct measurement devices with specific physical principals underlying the acquisition of 3D images*. Skilled radiotherapists must acknowledge the impact of our measurement device choice upon image-guided radiotherapy, and subsequently modify prescription parameters with the measurement implements being utilized. US and FM based targeting may in fact necessitate different respective PTV margins, for instance. Several interesting margin calculation formulae have been proposed; almost all are predicated on either standard deviation or variance of corrective shifts. Thus, examination of univariate distributional parameters is of interest.

According to the central limit theorem, measurements are expected to be normally distributed; in any case where measurement bias is of interest, failure to exhibit a normal distribution is a finding which must be addressed. In this dataset, only 1/6 directional components exhibited a normal distribution by Shapiro–Wilks goodness-of-fit. Consequently, we may assume that systematic bias is extant and must be accounted for within our samples. From a practical perspective, there are clinically relevant implications of these distributional findings. While PTV margins are often determined omni-directionally, it is obvious from our data that observed unidirectional variations are rather different. It may be reasonable to include calculations which account for variation in *each axis*, rather than in all directions (Cheung *et al* 2005).

Altman–Bland analysis revealed that a 95% agreement limit of 1–1.5 cm in any given dimension would have to be accepted; that is, it would have to be assumed that 95% of measures would fall within 1.5 cm *or less* for all cardinal axes. Altman–Bland plots do not give a formal threshold for agreement but rather leave the option for determination of whether two devices are in sufficient agreement as a pragmatic clinical judgment. In the context of modern conformal radiotherapy, a method of measure disagreement of the magnitude observed in our series would be clinically significant.

Lin's concordance correlation coefficient is similarly useful insofar as it provides a single number conceptualization of concordance for continuous variables (Barnhart *et al* 2002); this number shows the error weighted strength of relationship. No formally designated thresholds exist for ρ; however, we have chosen to utilize a grading schema proposed by Partik *et al* (2002) for volumetric comparison of ultrasound volumes using Lin's coefficient. It is striking to note that the concordance evaluation using Partik's criteria parameters varies from 'poor' to 'unacceptable' in every axis and composite measure for the calculated ρ values. This provides a considerable rationale against viewing the examined targeting approaches as equivalent measures of inter-fraction motion.

Deming orthogonal regression is superior (for method comparison) to standard linear regression as it affords a correction for bivariate bias, a feature not available with standard regression (Cornbleet and Gochman 1979). Utilization of Deming regression also allows evaluation of constant or proportional bias between measures though the accuracy of input data is of greater importance than the regression technique utilized in method comparison. Our

findings demonstrate that both constant and proportional biases are notable between imaging devices. Additionally, the processes which create this differential are as yet unidentified. However, it should be noted that constant error is much easier to account for, as it represents a linear differential, as opposed to the proportional error component, which may exhibit a different mathematical relationship. Future studies are needed to determine why proportional error was observed in some axes but not others.

The differential between US and FM seen in this series has several possible sources. First, there are inherent physical limitations of both systems. Trans-abdominal US images are known to be impaired by subcutaneous fat, whereas reduction in image quality for kV x-ray FM localization is minimal (Millender *et al* 2004). Similarly, the US procedure requires physically placing the transducer on the patient, which may alter the prostate organ position through pressure on organs of interest (Dobler *et al* 2006). Also, it has been shown that some degree of marker movement and organ deformation may be observed with FM approaches. Additionally, while US target localization is reliant on daily user manual input *during* image acquisition, the FM system is more automated and requires physician input *after* image acquisition and image coregistration. Consequently, the throughput processes are somewhat distinct and may impact on the differential between suggested shifts.

There are several limitations intrinsic in this study. This study is a single-institution experience and has a patient $n = 40$. Additionally, we have not formally accounted for many possible mechanisms which might impact the difference between measures in both imaging systems (such as organ deformation, seed migration and inter-user variability), nor have we attempted to determine more eloquent permutations of the utilized techniques to determine the impact of repeated measures (Barnhart *et al* 2005). Seed migration (Crook *et al* 1995, Poggi *et al* 2003, Welsh *et al* 2004), while impacting the distributional outcomes observed, should nonetheless have little effect on the statistical analysis performed for method comparison, as any change in position should be accounted for proportionally by both imaging devices. Were US visualization to rely only on organ contours, potential seed migration might preferentially affect the centroid calculated by the FM automatic algorithm; however, this effect was not observed in this dataset, as no detectable seed migration was observed by either method. However, the question of repeatability is a prime concern. In a method comparison analysis, if one or both methods has poor repeatability, agreement will be predictably poor (Bland and Altman 1986). We did not record the user for each and every US and FM shift, and thus cannot estimate the impact of intra-user or intra-system repeatability, and/or inter-user variation.

Nonetheless, this series has several notable features of import. To our knowledge, it represents the first exploration of these two image-acquisition and targeting modalities using method comparison analysis. The fact that the same users implemented both devices on all patients is noteworthy. However, the above caveats still apply, and consequently the authors strongly encourage exploration of similar method comparison series in multi-institutional settings. Additionally, this series is not designed to determine which device is 'better' insofar as we may only compare the given procedures to either initial skin-mark alignment or one another. Yet it must be noted that US and FM are far from identical and in our judgment should not be utilized interchangeably. While method comparison techniques present a useful way of conceptualizing differentials between measures, the final clinical decision of which measurement method is preferable remains with the therapy team.

It is imperative however that comparison between image-guidance methods occur; otherwise, we are unable to determine whether a 1 mm US margin is equivalent to a 1 mm FM margin (our data suggest they are *not* interchangeable). Margin recommendations may need adjustment accordingly. Since the analyses included in this paper are straightforward and accessible, we advocate utilization of method comparison analysis rather than standard

correlation or regression. A recent example of such a method comparison study is the article by McNair *et al*, which successfully implements bias plots in analysis of CT and US techniques (McNair *et al* 2006). Such papers should serve as templates for future series. When utilized—ideally in combination (Karlsson *et al* 2002) with systematic and random error calculations—such methods provide more accurate characterization of imaging techniques for target localization.

Image-guided radiotherapy holds great promise compared to conventional radiotherapy in many applications. This promise may only be accomplished by careful comparison of competing systems (Ling *et al* 2006). By delineating and establishing sound practices for method comparison clinicians may then more carefully assess optimum implementation of specific commercial and experimental image-guidance using data-driven approaches.

Conclusion

Method comparison of US and FM data indicates a level of agreement such that 95% measures will agree within approximately 1.5 cm in any room axis, with poor concordance, and detectable proportional and constant bias components. US- and FM-derived data are *not* equivalent and should not be used interchangeably. Direct one-to-one comparison of US and FM shift data must account for observed method differentials. Future studies are needed to ascertain the impact of image-guidance method differences on patient dosimetry and subsequent clinical endpoints.

Appendix

Lin's concordance coefficient:

The coefficient ρ is a comparison of specified measurements, demonstrating deviation from a line through the origin at an angle of 45° (as if the two measurements generated identical results, i.e. a line of identity):

$$\rho = 1 - d_c^2 / d_u^2$$

when d_c^2 is the expected squared perpendicular deviation from the line, and d_u^2 is the expected squared perpendicular deviation from the line when the measurements are uncorrelated.

Alternately, this may be represented as

$$\rho = r \times C_b,$$

where r is the calculated Pearson correlation coefficient, and where C_b is a bias correction factor which is calculated by

$$C_b = 2/(v + 1/v + u^2)$$
$$v = (\mathrm{SD}_1/\mathrm{SD}_2)$$
$$u = (m_1 - m_2)/\sqrt{(\mathrm{SD}_1 \times \mathrm{SD}_2)}$$

where m_i and SD_i $(i = 1, 2)$ are the mean and standard deviation of the ith set of measurements.

References

Arbillaga H O, Montgomery G P, Cabarrus L P, Watson M M, Martin L and Edworthy S M 2002 *BMC Musculoskelet. Disord.* **3** 13

Barnhart H X, Haber M and Song J 2002 *Biometrics* **58** 1020–7

Barnhart H X, Song J and Haber M J 2005 *Stat. Med.* **24** 1371–84

Bel A, van Herk M and Lebesque J V 1996 *Med. Phys.* **23** 1537–45
Bielak L F, Sheedy P F II and Peyser P A 2001 *Radiology* **218** 224–9
Bland J M and Altman D G 1986 *Lancet.* **1** 307–10
Bland J M and Altman D G 1995a *Lancet* **346** 1085–7
Bland J M and Altman D G 1995b *Int. J. Epidemiol.* **24** (Suppl. 1) S7–S14
Bland J M and Altman D G 1997 *Ann. Clin. Biochem.* **34** 570–1
Bland J M and Altman D G 1999 *Stat. Methods Med. Res.* **8** 135–60
Bland J M and Altman D G 2003 *Ultrasound Obstet. Gynecol.* **22** 85–93
Bos L J, van der Geer J, van Herk M, Mijnheer B J, Lebesque J V and Damen E M 2005 *Radiother. Oncol.* **76** 18–26
Bradley A J, Carrington B M, Lawrance J A, Ryder W D and Radford J A 1999 *J. Clin. Oncol.* **17** 2493–8
Byrne T E 2005 *Med. Dosim.* **30** 155–61
Chandra A, Dong L, Huang E, Kuban D A, O'Neill L, Rosen I and Pollack A 2003 *Int. J. Radiat. Oncol. Biol. Phys.* **56** 436–47
Cheung P, Sixel K, Morton G, Loblaw D A, Tirona R, Pang G, Choo R, Szumacher E, Deboer G and Pignol J P 2005 *Int. J. Radiat. Oncol. Biol. Phys.* **62** 418–25
Cornbleet P J and Gochman N 1979 *Clin. Chem.* **25** 432–8
Crook J M, Raymond Y, Salhani D, Yang H and Esche B 1995 *Radiother. Oncol.* **37** 35–42
Dobler B, Mai S, Ross C, Wolff D, Wertz H, Lohr F and Wenz F 2006 *Strahlenther. Onkol* **182** 240–6
Dunn G and Roberts C 1999 *Stat. Methods Med. Res.* **8** 161–79
Eng T Y, Luh J Y and Thomas C R Jr 2005 *Curr. Urol. Rep.* **6** 194–209
Fung A Y, Ayyangar K M, Djajaputra D, Nehru R M and Enke C A 2006 *Med. Dosim.* **31** 20–9
Herman M G, Pisansky T M, Kruse J J, Prisciandaro J I, Davis B J and King B F 2003 *Int. J. Radiat. Oncol. Biol. Phys.* **57** 1131–40
Hilson A 2004 *Radiology* **231** 604–5 (author reply)
Huang E, Dong L, Chandra A, Kuban D A, Rosen I I, Evans A and Pollack A 2002 *Int. J. Radiat. Oncol. Biol. Phys.* **53** 261–8
Karlsson A S, Hedren M, Almqvist C, Larsson K and Renstrom A 2002 *Allergy* **57** 164–8
Keall P J, Beckham W A, Booth J T, Zavgorodni S F and Oppelaar M 1999 *Australas. Phys. Eng. Sci. Med.* **22** 48–52
King T S and Chinchilli V M 2001a *Stat. Med.* **20** 2131–47
King T S and Chinchilli V M 2001b *J. Biopharm. Stat.* **11** 83–105
Konings H 1982 *Surv. Immunol. Res.* **1** 371–4
Kuban D A, Dong L, Cheung R, Strom E and De Crevoisier R 2005 *Semin. Radiat. Oncol.* **15** 180–91
Langen K *et al* 2003 *Int. J. Radiat. Oncol. Biol. Phys.* **57** 635–44
Lin L I 1989 *Biometrics* **45** 255–68
Ling C C, Yorke E and Fuks Z 2006 *Radiother. Oncol.* **78** 119–22
Linnet K 1998 *Clin. Chem.* **44** 1024–31
Linnet K 1999a *Clin. Chem.* **45** 314–5
Linnet K 1999b *Clin. Chem.* **45** 882–94
Mahnken A H, Koos R, Katoh M, Spuentrup E, Busch P, Wildberger J E, Kuhl H P and Gunther R W 2005 *Eur. Radiol.* **15** 714–20
Martin R F 2000 *Clin. Chem.* **46** 100–4
McKenzie A, van Herk M and Mijnheer B 2002 *Radiother. Oncol.* **62** 299–307
McNair H A, Mangar S A, Coffey J, Shoulders B, Hansen V N, Norman A, Staffurth J, Sohaib S A, Warrington A P and Dearnaley D P 2006 *Int. J. Radiat. Oncol. Biol. Phys.* **65** 678–87
Millender L E, Aubin M, Pouliot J, Shinohara K and Roach M III 2004 *Int. J. Radiat. Oncol. Biol. Phys.* **59** 6–10
Partik B L, Stadler A, Schamp S, Koller A, Voracek M, Heinz G and Helbich T H 2002 *Invest. Radiol.* **37** 489–95
Patel R R, Orton N, Tome W A, Chappell R and Ritter M A 2003 *Radiother. Oncol.* **67** 285–94
Payne R B 1997 *Ann. Clin. Biochem.* **34** 319–20
Poggi M M, Gant D A, Sewchand W and Warlick W B 2003 *Int. J. Radiat. Oncol. Biol. Phys.* **56** 1248–51
Purdy J A 2002 *Semin. Radiat. Oncol.* **12** 199–209
Scarbrough T J, Golden N M, Ting J Y, Fuller C D, Wong A, Kupelian P A and Thomas C R Jr 2006 *Int. J. Radiat. Oncol. Biol. Phys.* **65** 378–87
Schallenkamp J M, Herman M G, Kruse J J and Pisansky T M 2005 *Int. J. Radiat. Oncol. Biol. Phys.* **63** 800–11
Schaly B, Bauman G S, Song W, Battista J J and Van Dyk J 2005 *Phys. Med. Biol.* **50** 3083–101
Schramm P, Schellinger P D, Klotz E, Kallenberg K, Fiebach J B, Kulkens S, Heiland S, Knauth M and Sartor K 2004 *Stroke* **35** 1652–8
Shapiro S S, Wilk M B and Chen H J 1968 *J. Am. Stat. Assoc.* **63** 1343–72
Stockl D, Dewitte K and Thienpont L M 1998 *Clin. Chem.* **44** 2340–6

Stroom J C, de Boer H C, Huizenga H and Visser A G 1999 *Int. J. Radiat. Oncol. Biol. Phys.* **43** 905–19
Stroom J C and Heijmen B J 2002 *Radiother. Oncol.* **64** 75–83
Van den Heuvel F, Powell T, Seppi E, Littrupp P, Khan M, Wang Y and Forman J D 2003 *Med. Phys.* **30** 2878–87
van Herk M 2004 *Semin. Radiat. Oncol.* **14** 52–64
van Herk M, Remeijer P and Lebesque J V 2002 *Int. J. Radiat. Oncol. Biol. Phys.* **52** 1407–22
van Herk M, Remeijer P, Rasch C and Lebesque J V 2000 *Int. J. Radiat. Oncol. Biol. Phys.* **47** 1121–35
van Herk M, Witte M, van der Geer J, Schneider C and Lebesque J V 2003 *Int. J. Radiat. Oncol. Biol. Phys.* **57** 1460–71
Welsh J S, Berta C, Borzillary S, Sam C, Shickell D, Nobile L, Greenberg M, Weiss S and Detorie N 2004 *Technol. Cancer Res. Treat.* **3** 359–64
Westgard J O and Hunt M R 1973 *Clin. Chem.* **19** 49–57
Zhang M, Moiseenko V, Liu M and Craig T 2006 *Phys. Med. Biol.* **51** 269–85