Published in final edited form as:
J Ultrasound Med. 2010 June ; 29(6): 891–901.

Volume of preclinical xenograft tumors is more accurately assessed by ultrasound imaging than manual caliper measurements

Gregory D. Ayers, MS[*,1], **Eliot T. McKinley, BSE**[*,2,3], **Ping Zhao, BS**[3], **Jordan M. Fritz, BS**[3], **Rebecca E. Metry, BS**[3], **Brenton C. Deal, BS**[3], **Katrina M. Adlerz, BS**[3], **Robert J. Coffey, MD**[5], and **H. Charles Manning, Ph.D**[2,3,4,6,7]

[1]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, 37232.

[2]Department of Biomedical Engineering, Vanderbilt University Medical Center, Nashville, TN, 37232.

[3]The Vanderbilt University Institute of Imaging Science (VUIIS), Vanderbilt University Medical Center, Nashville, TN, 37232.

[4]Department of Radiology and Radiological Sciences, Vanderbilt University Medical Center, Nashville, TN, 37232.

[5]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, 37232.

[6]Program in Chemical and Physical Biology, Vanderbilt University Medical Center, Nashville, TN, 37232

[7]Department of Neurosurgery, Vanderbilt University Medical Center, Nashville, TN, 37232

Abstract

Objective—Volume of subcutaneous xenograft tumors is an important metric of disease progression and response to therapy in preclinical drug development. Non-invasive imaging technologies suitable for measuring xenograft volume are increasingly available, yet manual calipers, which are susceptible to inaccuracy and bias, are routinely employed. The goal of this study was to quantify and compare the accuracy, precision, and inter-rater variability of xenograft tumor volume assessment by caliper measurements and ultrasound imaging.

Methods—Subcutaneous xenograft tumors derived from human colorectal cancer cell lines (DLD1, SW620) were generated in athymic nude mice. Experienced independent reviewers segmented three-dimensional ultrasound data sets and collected manual caliper measurements resulting in tumor volumes. Imaging- and caliper-derived volumes were compared to tumor mass, the gold standard, determined following resection. Bias, precision and inter-rater differences were estimated for each mouse among reviewers. Bootstrapping was used to estimate mean and confidence intervals of variance components, intra-class correlation coefficients (ICC's) and confidence intervals for each source of variation.

Results—Average deviation from true volume and inter-rater differences were significantly lower for ultrasound volumes compared to caliper volumes. Reviewer ICC's for ultrasound and

Corresponding Author: H. Charles Manning, Ph.D., Vanderbilt University Institute of Imaging Sciences (VUIIS), Vanderbilt University Medical School, 1161 21st Ave. S., AA 1105 MCN, Nashville, TN 37232-2310, Tel: (615) 322 -3793, Fax: (615) 322-0734. (henry.c.manning@vanderbilt.edu).
*These authors contributed equally to this work.

caliper measurements were similarly low (1%), yet caliper volume variance was 1.3-fold higher than ultrasound.

Conclusions—Ultrasound imaging more accurately, precisely, and reproducibly reflects xenograft tumor volume than caliper measurements. These data suggest that preclinical studies utilizing xenograft burden as a surrogate endpoint measured by ultrasound imaging require up to 30% fewer animals to reach statistical significance compared to analogous studies utilizing caliper measurements.

Keywords

tumor burden; ultrasound imaging; colorectal cancer; preclinical mouse models; xenograft

Introduction

Longitudinal measurement of subcutaneous xenograft tumor volume is a central component of numerous preclinical studies utilizing mouse models of human cancer. Tumor volume is used as a metric to assess growth and disease progression, as well as to quantify response to therapeutic regimens. Meaningful assays of tumor volume are highly accurate, precise, and possess the requisite sensitivity to detect subtle differences between experimental arms, while utilizing as few experimental animals as possible. Measurements of xenograft tumors collected with manual calipers are rapid, non-invasive and inexpensive (1,2). However, caliper measurements of subcutaneous xenografts are affected by contributions to the measure from epidermis and adipose tissue, as well as fur if present, each of which introduces error and variability into volume determinations. Furthermore, caliper measurements are commonly collected along the longest two dimensions of the tumor x/y plane only, with the z-axis dimension assumed to be the same as the shortest dimension (1).This practice contributes to the expediency of the method but hinders accuracy because volume estimation with this approach assumes ellipsoidal shaped xenografts, which is frequently incorrect. These weaknesses highlight that improved methods to accurately and reproducibly determine xenograft tumor volumes on a routine basis are needed in preclinical cancer research.

A number of alternative methods to measure the volume of xenograft tumors have been reported, each varying widely with respect to the time required to make the measurement, cost, and accuracy. Physical methods, such as cast modeling, have been shown superior to caliper measurements for determining xenograft tumor volume in mouse models (3). Cast modeling, however, is extremely time and labor intensive. Furthermore, cast modeling requires xenografts to be placed in a limited number of anatomical locations, such as the ventral chest wall, which provide suitable resistance to pressure as to enable the cast to properly form (3). Several imaging methods suitable for assessing xenograft tumor volume exist and are quite attractive due to their non-invasive nature and potential for highly-resolved measurement. Computed tomography (4) has been reported more reliable for assessing tumor volume than calipers in rat models of mammary cancer (5). However, the availability and expense of small animal CT scanners, the requirement of ionizing radiation, and the inherently poor soft tissue contrast of CT without the use of exogenous contrast material limit the routine use of CT for xenograft volume measurements in small animals. Magnetic Resonance Imaging (MRI) methods provide reliable tumor volume measurements without exogenous contrast materials (6,7), yet without specialized animal holders accommodating numerous animals simultaneously (8), MRI is generally too expensive and time consuming for routine xenograft tumor measurements when studies include more than a few cohorts. Bioluminescence imaging (BLI) has been used as a measure of relative tumor burden (9) and can be rapid. However, in most cases BLI is acquired in two-dimensional

planar format and is thus unable to provide an absolute tumor volume measurement. BLI also has the added requirement that xenografts must be generated from tumor cells engineered to express luciferase and such models require injection of luciferin substrate, limiting the breadth of preclinical models that can be studied.

Ultrasound imaging in the preclinical setting has emerged as an inexpensive, non-invasive method for measuring xenograft tumor volume. Ultrasound imaging boasts excellent soft tissue contrast without the use of exogenous contrast agents or ionizing radiation and offers considerably higher throughput than CT or MRI (up to 30 animals/hr). Previous studies have reported that ultrasound imaging provides reliable tumor volumes in longitudinal studies (10–12) and in *in vitro* tissue samples(13,14).. However, studies to date have not directly compared the accuracy and reliability of xenograft volumes determined by ultrasound imaging to those obtained using external caliper measurements. In this study, we show that xenograft tumor volume determination is significantly more accurate and reproducible using ultrasound imaging than external caliper measurements. Accordingly, we demonstrate that preclinical studies employing ultrasound imaging for volume determination of xenograft tumors require significantly fewer animals to reach statistical significance than analogous studies relying upon standard caliper measurements.

Materials and Methods

Preclinical mouse models

Studies involving mice were conducted in accordance with federal and institutional guidelines. SW620 and DLD1 human colorectal cells were cultured in DMEM supplemented with 10% fetal bovine serum at 37°C in an atmosphere of 5% CO_2. For *in vivo* studies, xenograft tumors were generated as described (15). Briefly, 4×10^6 cells were injected subcutaneously on the right flank of 5–6 week old female athymic nude mice (Harlan Sprague-Dawley). Using this method, palpable tumors were typically observed within 2 weeks following injection of cells and were allowed to progress until at least 400 mm^3 for these studies.

Caliper measurements of subcutaneous xenografts

The two longest perpendicular axes in the x/y plane of each xenograft tumor were measured to the nearest 0.1 mm by three independent observers (reviewers 5–7) familiar with collecting caliper measurements of xenograft tumors in mice. The depth was assumed to be equivalent to the shortest of the perpendicular axes, defined as y. Measurements were made using a digital vernier caliper while mice were conscious and were calculated according to Equation 1 as is standard practice (1, 2):

$$\text{Xenograft volume} = xy^2/2 \quad (1)$$

Ultrasound imaging and data analysis

Immediately following caliper measurements, three-dimensional ultrasound imaging data sets were collected for each xenograft using a Vevo 770 ultrasound microimaging system (VisualSonics Inc.) designed for small animal imaging. For imaging acquisition, mice were initially anesthetized using 2% isofluorane in oxygen followed by placement on a heated stage during the course of imaging. Anesthesia was maintained during imaging using 2% isoflurane in oxygen. Xenografts were coated in warmed (37°C) Aquasonic 100 ultrasound gel (Parker Laboratories) and centered in the imaging plane. Three-dimensional B-mode data was acquired by automated translation of the 30 MHz ultrasound transducer along the entire length of the xenograft. The resulting data sets had a 17mm × 17mm field of view

with an in-plane pixel resolution of 33.2 × 33.2 μm and an interslice spacing of 101.6 μm, resulting in 33.2 × 33.2 × 101.6 μm voxels.

For analysis of ultrasound data, images were imported into Amira 5.2 (Visage Imaging) for volumetric analysis. Tumor tissue exhibited photopenia compared to non-tumor tissue allowing the tumor tissues to be manually segmented by four trained observers (reviewers 1–4) to obtain a volume for each xenograft. Tumor volume was determined by summation of the in-plane segmented regions and multiplying this quantity by the inter-slice spacing as described(12).

Validation of xenograft tumor volumes

Animals were sacrificed immediately following ultrasound imaging and xenograft tumors excised and stripped of non-tumor tissue if present. Tumor mass has been shown to directly correlate with volume measured by water displacement (r=1.0000)(2). Mass was determined to the nearest 0.1 mg using a calibrated analytical balance. Xenograft tumor volume was calculated from tissue mass assuming a density of 1 mg/mm^3. This value was used as the true tumor volume (TTV) for comparison purposes.

Data and statistical analysis

Volumes derived from the mass of excised xenograft tumors were established as the "gold standard" value for volume. Overall bias was estimated separately for each measurement type using an intercept only mixed models analysis of variance assuming normal errors containing random effect terms for reviewers and mice. Inter-rater variability was assessed using the average of the absolute value of inter-rater differences over mice by measurement type. Among the 4 ultrasound reviewers, these averages were comprised of $\frac{4!}{2!(4-2)!}=6$ observations per mouse compared to $\frac{3!}{2!(3-2)!}=3$ observations per mouse for the 3 caliper reviewers. Per mouse averages were compared using the Wilcoxon signed rank test for paired observations. Five thousand bootstrap replicates of the data were generated under the model described above to estimate nonparametric confidence intervals for sources of variability, inter-class correlation coefficients (ICC), and the total variance. This number of replicates ensured consistent precision to 3 decimal places for the confidence intervals of point estimates. The ICC is defined as the ratio of variance components; specifically as the ratio of variability among mice to the sum of variability due to mice, reviewers, and error. An analogous ICC was also estimated for reviewers placing reviewer variability in the numerator. Reviewer 3 and 6 were a common reviewer, though observations from this individual were considered independent between measurement types in all analyses as the observer was blinded to which animal was being measured.

Results

The mean (standard deviation) TTV measured on 14 mice was 1117 mm^3 (587 mm^3). TTV ranged from 460 mm^3 to 2323 mm^3 with a median of 951 mm^3. Typical reconstructed three-dimensional tumor volumes derived from ultrasound imaging data are shown in Figure 1, where the segmented tumor volume is shown in purple for display purposes. As displayed, tumors are ranked by TTV. Non-spheroid tumors, which comprised approximately half of the tumors studied, are denoted with an asterisk. Table 1 summarizes the cell line and the measurements of tumor volume by ultrasound and calipers as well as the TTV.

Figure 2 depicts the distributions of bias (experimental measure minus TTV) for each independent rater. Overall average bias (± s.e.) among ultrasound measurements was −53

mm^3 (± 43 mm^3) compared to 96 mm^3 (± 88 mm^3) for caliper measurements. Deviations per reviewer compared to the overall bias were relatively small. The coefficient of variation for bias and its standard deviation for ultrasound and caliper measurements were 0.81 and 0.92, respectively. Reviewer variability was relatively small compared to the variability among mice for both ultrasound and caliper measurements as assessed by the reviewer intraclass correlations of 1% for both modalities.

Figure 3 depicts Bland-Altman plots for reviewer 1 (ultrasound reviewer, 3A) and 5 (caliper reviewer, 3B). The Bland-Altman plots from these reviewers were representative of the other reviewer plots, which are all shown in Supplemental Data. Importantly, these plots illustrate more precise inter-rater agreement between ultrasound imaging measurements and TTV than for caliper measurements performed on the same mice. These data also illustrate that accurate ultrasound measurement may be limited to tumors <1500 mm^3, as ultrasound estimates appeared consistently smaller than the TTV in tumors larger than 1500 mm^3. Underestimation of volume in very large tumors was not observed for caliper measurements, but the variability in caliper estimates was found to increase with tumor volume.

Median (range) inter-rater variability, as assessed by the average of the absolute value of inter-rater differences, was significantly lower among ultrasound measurements 73 mm^3 (25 mm^3 to 138 mm^3) compared to 147 mm^3 (66 mm^3 to 408 mm^3) for caliper measurements ($p=0.001$). The reviewer differences between measurement type were not highly correlated (spearman $r = 0.17$), suggesting that tumor characteristics that result in large inter-rater deviations for caliper measurements are not the same as those that result in the larger inter-rater differences for ultrasound (Figure 4A). Shown another way (Figure 4B), inter-rater differences plotted against the rank of TTV for each mouse illustrates a trend toward increasing disagreement as true volume increases for caliper measurements ($p=0.063$) and ultrasound ($p=0.004$). For caliper measurements, the increase was an average 0.09 mm^3 per 1 mm^3 increase in true tumor volume compared to 0.04 mm^3 among ultrasound reviewers. The discrepancy in p-values is likely associated with the higher variability among reviewers making caliper measurements (residual standard error for caliper reviewers = 93 mm^3 versus 23 mm^3 for ultrasound reviewers).

Bootstrapped estimates of the sources of variability confirmed results from previous analyses (Table 2). Interestingly, high mouse ICC and low reviewer ICC for both types of measurements indicate that multiple reviewers are not necessary to establish precise estimates of tumor volume whether measured via ultrasound or calipers. Four percent of the total variance was in the error term for caliper measurements, which in this model may suggest an interaction between reviewers and mice. We interpret the four-fold greater ICC error among caliper measurements compared to ultrasound imaging measurements to stem from the observation that inter-reviewer variability apparently increases with tumor size among caliper measurements but not for ultrasound measurements.

Discussion

Studies to elucidate the biological effects and therapeutic potential of candidate drugs in oncology may begin in an *in vitro* setting, but promising strategies are ultimately advanced to *in vivo* mouse models. Though elegant transgenic mice have been developed which enable study of subtle biological and clinical traits of human cancer, a majority of *in vivo* assays designed to assay the efficacy of novel agents utilize simple measurement of tumor growth and/or regression. A natural progression from *in vitro* screening, these studies routinely employ subcutaneous human cell line xenografts grown in athymic nude mice and rely upon accurate and precise measurement of xenograft volume.

In this study, we compared the accuracy and precision of ultrasound imaging volumes of subcutaneous xenograft tumors and volumes determined using caliper measurements to the true volume of the tumor. Accuracy can be defined as proximity between the measure of an object and its true value. We quantified the accuracy of volumetric measurements as the average difference between the TTV and measured volume using either ultrasound image volumes or caliper measurements of the tumor. Precision is the average squared distance of measurements from their mean (variance or standard deviation). It is important to note that a group of measurements can be accurate, precise, both, or neither. Bias is defined as a systematic difference from the true value. In multi-arm experiments, the impact of bias is negligible because bias, usually assumed to be constant in all treatment groups, is negated by subtraction. An exception occurs when bias is not constant across the range of values measured in the study. For example, if one treatment (e.g., control) elicits no response with concomitant high values of a measurement and the active treatment group yields small response values on average when systematic bias is present, the estimated treatment difference will contain much of the larger bias of the control treatment. In short, valid measurement systems are both accurate and precise and contain minimal bias.

Ultrasound measurements ($-53\ mm^3 \pm 43\ mm^3$) of tumor volume were less biased compared to tumor measurements taken with calipers ($96\ mm^3 \pm 88\ mm^3$). Locally weighted scatterplot smoothing of data points in Bland-Altman plots show ultrasound measurements may underestimate tumor volume for true tumor volumes greater than 1500 mm^3. While no systematic bias was detected among caliper measurements, the variability of caliper measurements increased with true tumor volume. This suggests that data transformations (e.g., natural log) may be necessary to stabilize the variability of caliper measurements before statistical methods assuming Gaussian (normal) errors can be used appropriately.

The bootstrap estimated standard deviation of caliper and ultrasound measurements (i.e., square root of the total variance excluding reviewer sources) were 583 mm^3 and 508 mm^3, respectively. As rule of thumb, sample size increases in direct proportion to the ratio of variances of alternative scenarios. The ratio of caliper versus ultrasound variances was 1.32. In comparable randomized studies, the sample size necessary to have the same power to detect the same difference in tumor volume among treatment groups will be 1.32 times higher using caliper measurements compared to using ultrasound to estimate tumor volume. To illustrate this further, we conducted sample size calculations for a hypothetical two-arm mouse study where the primary endpoint is post-treatment minus pre-treatment change in tumor volume (Table 3). In the hypothetical study, treatment begins when tumors are palpable at 100 mm^3. We assume an average increase in tumor volume to 1100 mm^3 among control mice and the treatment effect results in a decreased average tumor volume to 1000 mm^3, 850 mm^3, or 600 mm^3. Change in tumor volume is compared using a two-sided, two sample t-test that is considered statistically significant when $p < 0.05$. Without loss of generality, we assume pre- and post-treatment tumor measurements are uncorrelated. For the purposes of this exercise, we used the mixed model estimates of standard deviations among mice of 563 mm^3 for caliper-measured mice and 490 mm^3 for ultrasound-measured mice as determined in the present study. For significant results to be achieved, the number of animals required necessarily increases as the true treatment effect decreases, desired power increases, and/or the standard deviation among animals increases. This latter effect is constant, regardless of a treatment effect or desired power, and directly proportional to the ratio of the variances between measurement types. Note that deviations from a ratio of 1.32 are due to rounding of sample size to integer values, the effect of which increases with smaller sample sizes.

In addition to focusing on accuracy and repeatability, both of which affect the conduct of basic and clinical research, the third component of this study evaluated measurement reproducibility. Reproducibility is concerned with the precision of repeated measurements within (intra) and among (inter) reviewers. We did not assess intra-reviewer precision in this study. Inter-reviewer variability was measured as the average absolute difference between reviewer measurements by type. Thus each difference is always positive, as is the average. With 4 ultrasound reviewers and 3 caliper reviewers, there were 6 and 3 differences per mouse, respectively, comprising the average for each mouse. Average absolute differences among caliper measurers had a median (range) of 147 mm^3 (66 mm^3 to 408 mm^3). Ultrasound reviewers had a median value of 73 mm^3 (25 mm^3 to 138 mm^3). Lines that match mice across measurement type in Figure 3 support a low Spearman correlation of 0.17 between measurement types. Importantly, from this result we conclude that tumor characteristics that elicit low reproducibility in caliper measurements do not greatly affect ultrasound measurements. The average absolute differences tended to increase in both groups as tumor size increased with the rate in the caliper group approximately double that of the ultrasound group. Ordered by true tumor size, mouse number 6, 10, 13, and 14 had the highest disagreement values in the caliper group. Although our results suggest that ultrasound may underestimate tumor volume when tumors are above 1500 mm^3, these plots show that the ultrasound reviewers were consistent in their assessments regardless of tumor size. We hypothesize that attenuation of ultrasound signal at depths approaching 15 mm may cause a degree of uncertainty in segmenting the basal surface of very large xenograft tumors. If true, this effect could be minimized at lower frequencies where ultrasound penetration is less affected by these determinants, but would decrease the spatial resolution of the ultrasound images. In either case, variance component analysis revealed very high ICC of 0.95 and 0.98 for caliper and ultrasound measurements suggesting that reviewer contributions to variability are very small relative to the total variance. Consequently, the expense of using multiple reviewers in mouse studies using either measurement type seems unwarranted.

A key determinant in the discrepancy between TTV and caliper-derived volumes, as opposed to ultrasound-derived volumes, is the assumption of spheroid shaped tumors in the case of caliper measurements. Figure 5 shows representative examples for tumors of similar size that were determined to be spheroid (5A, tumor 9) and non-spheroid (5B, tumor 10) as determined by three-dimensional visualization of the ultrasound data set. For the spheroid tumor, both ultrasound and caliper measurements accurately and precisely represent the TTV. For the non-spheroid tumor, the ultrasound measurements are both accurate and precise, while the caliper measurements are neither.

In this study, we have shown that ultrasound measurements of subcutaneous xenograft tumor volume exhibit significantly more accuracy, precision, and reproducibility than measurements made with standard calipers. Though the benefits outweigh the costs in the long-term, ultrasound necessarily requires access to imaging instrumentation, and like caliper measures, removal of fur if present. We conclude that use of ultrasound to measure tumor volume in mouse studies will result in more sensitive and reproducible measures of volume change at endpoints and throughout longitudinal studies. This is expected to translate into more efficient studies requiring fewer animals to obtain statistically meaningful results and requiring less infrastructure to support them.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by funding from the National Cancer Institute (NCI): P50 95103 (Vanderbilt GI SPORE Program), U01 084239 (Mouse Models of Human Cancers Consortium), U24 CA126588 (South-Eastern Center for Small-Animal Imaging), 1RO1 CA46413, 1R01 CA140628, 1RC1 CA145138, K25 CA127349, and 1P50CA128323 (Vanderbilt ICMIC Program). ETM was supported by pre-doctoral training grant in imaging science T32 EB003817.

References

1. Tomayko MM, Reynolds CP. Determination of subcutaneous tumor size in athymic (nude) mice. Cancer Chemother Pharmacol 1989;24:148–154. [PubMed: 2544306]
2. Euhus DM, Hudd C, LaRegina MC, Johnson FE. Tumor measurement in the nude mouse. J Surg Oncol 1986;31:229–234. [PubMed: 3724177]
3. Fiennes AG. Growth rate of human tumour xenografts measured in nude mice by in vivo cast modelling. Br J Surg 1988;75:23–24. [PubMed: 3337943]
4. Hector S, Prehn JH. Apoptosis signaling proteins as prognostic biomarkers in colorectal cancer: a review. Biochim Biophys Acta 2009;1795:117–129. [PubMed: 19167459]
5. Ishimori T, Tatsumi M, Wahl RL. Tumor response assessment is more robust with sequential CT scanning than external caliper measurements. Acad Radiol 2005;12:776–781. [PubMed: 15935976]
6. He Z, Evelhoch JL, Mohammad RM, et al. Magnetic resonance imaging to measure therapeutic response using an orthotopic model of human pancreatic cancer. Pancreas 2000;21:69–76. [PubMed: 10881935]
7. Mazurchuk R, Glaves D, Raghavan D. Magnetic resonance imaging of response to chemotherapy in orthotopic xenografts of human bladder cancer. Clin Cancer Res 1997;3:1635–1641. [PubMed: 9815854]
8. Dazai J, Bock NA, Nieman BJ, Davidson LM, Henkelman RM, Chen XJ. Multiple mouse biological loading and monitoring system for MRI. Magn Reson Med 2004;52:709–715. [PubMed: 15389955]
9. Jenkins DE, Oei Y, Hornig YS, et al. Bioluminescent imaging (BLI) to improve and refine traditional murine models of tumor growth and metastasis. Clin Exp Metastasis 2003;20:733–744. [PubMed: 14713107]
10. Cheung AM, Brown AS, Hastie LA, et al. Three-dimensional ultrasound biomicroscopy for xenograft growth analysis. Ultrasound Med Biol 2005;31:865–870. [PubMed: 15936502]
11. Wirtzfeld LA, Wu G, Bygrave M, et al. A new three-dimensional ultrasound microimaging technology for preclinical studies using a transgenic prostate cancer mouse model. Cancer Res 2005;65:6337–6345. [PubMed: 16024636]
12. Graham KC, Wirtzfeld LA, MacKenzie LT, et al. Three-dimensional high-frequency ultrasound imaging for longitudinal evaluation of liver metastases in preclinical models. Cancer Res 2005;65:5231–5237. [PubMed: 15958568]
13. Hughes SW, D'Arcy TJ, Maxwell DJ, Saunders JE, Chinn S, Sheppard RJ. The accuracy of a new system for estimating organ volume using ultrasound. Physiol Meas 1997;18:73–84. [PubMed: 9046539]
14. De Odorico I, Spaulding KA, Pretorius DH, Lev-Toaff AS, Bailey TB, Nelson TR. Normal splenic volumes estimated using three-dimensional ultrasonography. J Ultrasound Med 1999;18:231–236. [PubMed: 10082358]
15. Manning HC, Merchant NB, Foutch AC, et al. Molecular imaging of therapeutic response to epidermal growth factor receptor blockade in colorectal cancer. Clin Cancer Res 2008;14:7413–7422. [PubMed: 19010858]

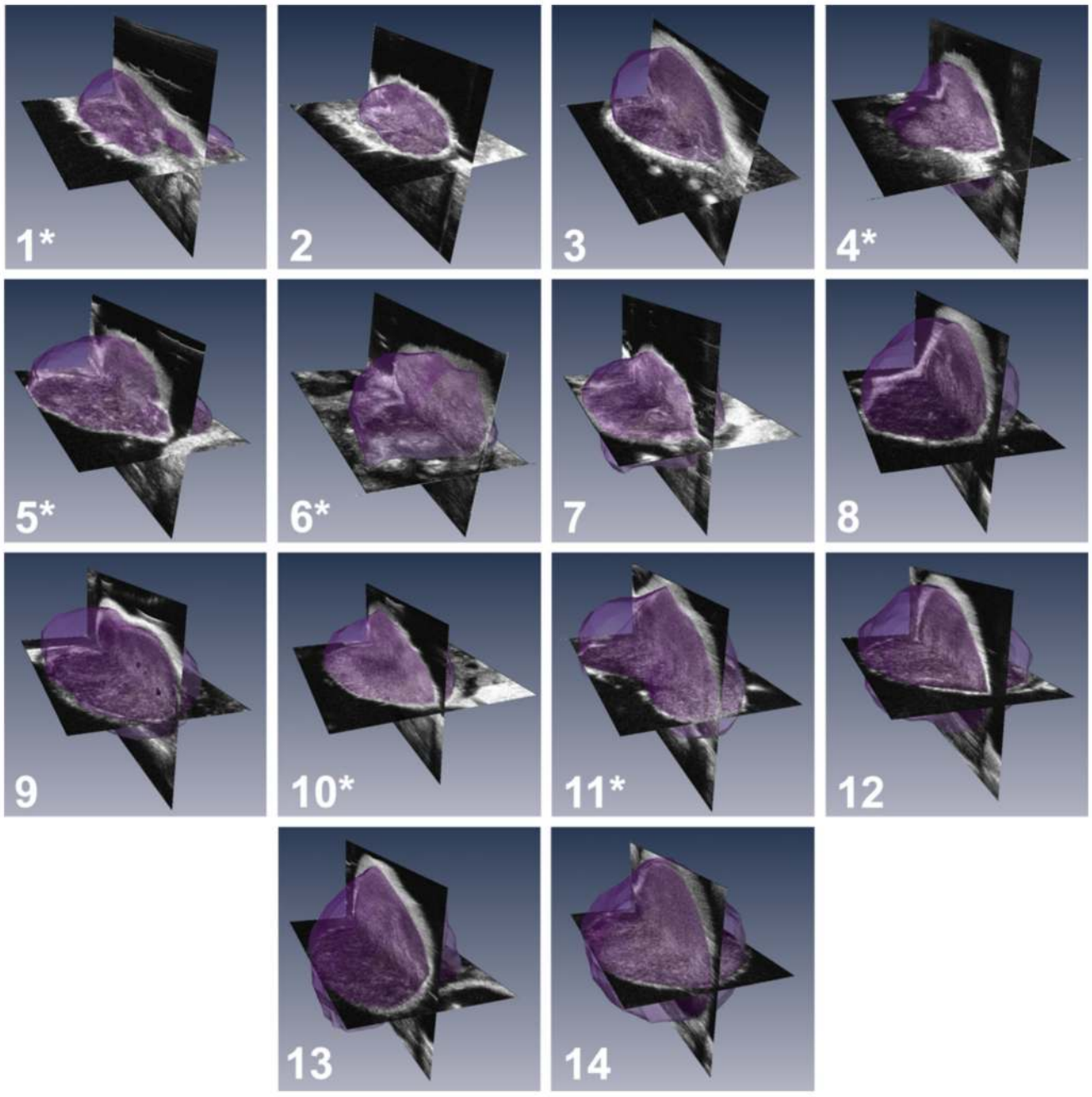

Figure 1.
Three-dimensional reconstructions of ultrasound imaging data for each xenograft tumor evaluated in the study. The segmented xenograft volume is shown in purple for a representative ultrasound reviewer and are ordered by TTV. Non-spheroid (irregularly shaped) xenografts, which are difficult to evaluate with calipers, are denoted by (*). Further details regarding each tumor can be found in Table 1.

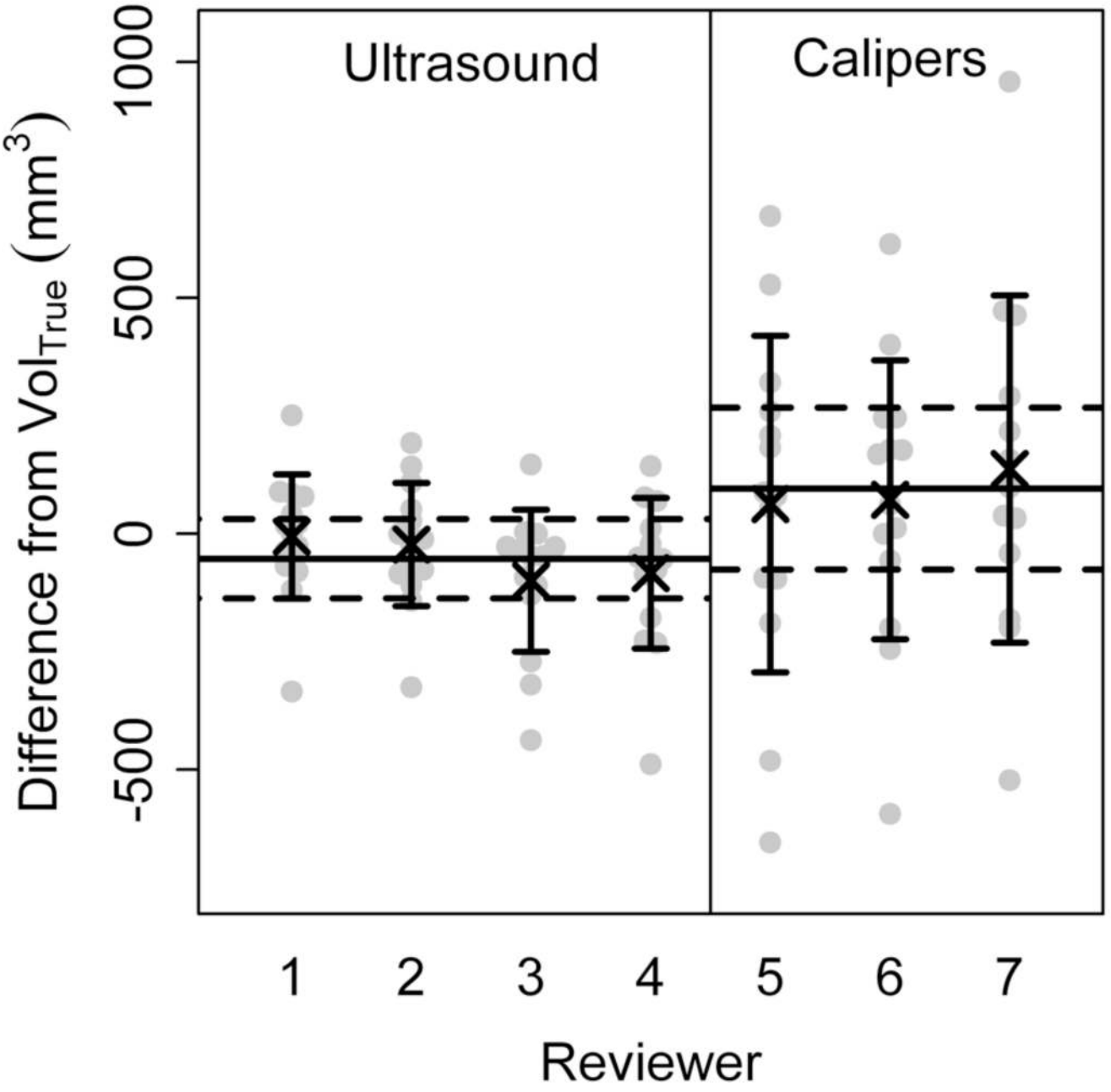

Figure 2.
Overall bias (solid lines) and standard error (dashed lines) are lower for ultrasound measurements (−53 mm^3 ± 43 mm^3) than caliper measurements (96 mm^3 ± 88 mm^3). Average bias for individual reviewers (X) was much lower in ultrasound than caliper, however, in both modalities, deviations per reviewer from the overall bias were small. Differences for each tumor are denoted as (●) for each reviewer.

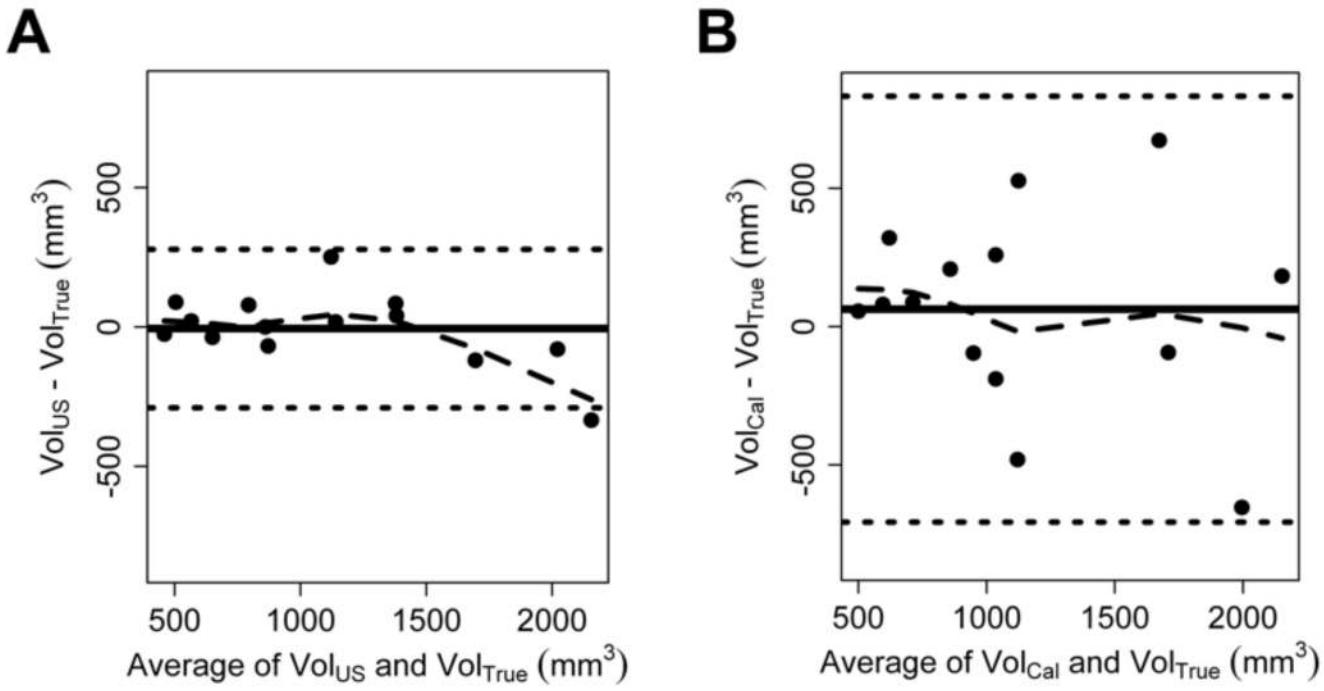

Figure 3.
Bias (solid lines) and 95% confidence intervals (dotted lines) were found to be smaller for ultrasound compared to caliper measurements. Data shown is from ultrasound Reviewer 1 (A) and caliper Reviewer 5 (B), both highly representative of other reviewer data sets. Individual xenograft difference from the true volume (●) was less in ultrasound measurements than caliper measurements. Lines are from nonparametric locally weighted scatterplot smoothing of the data points.

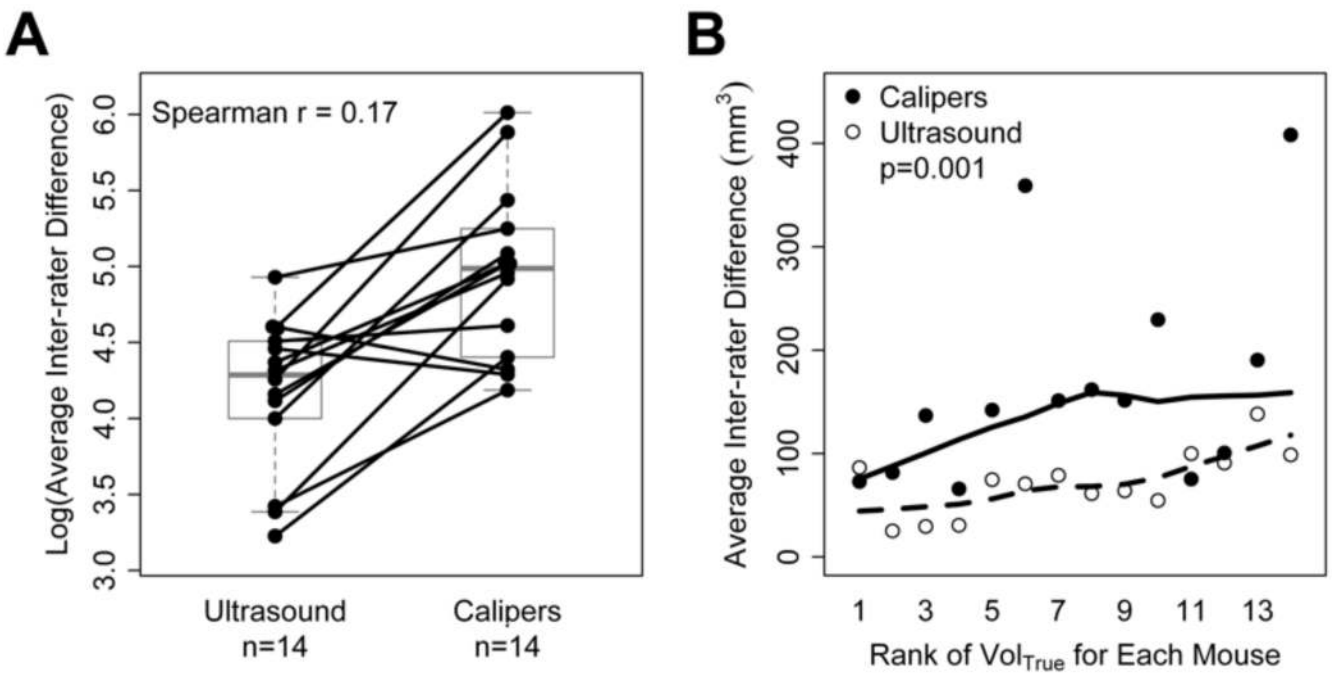

Figure 4.

Inter-rater variability was significantly lower in ultrasound than caliper measurements. Tumors that resulted in large inter-rater differences for ultrasound did not correlate with tumors that resulted in larger inter rater differences for ultrasound (A). Inter-rater differences plotted against rank of true tumor volume for each mouse increased at faster rate for caliper measurements compared to ultrasound measurements as true volume increased (B). Trend lines feature nonparametric locally weighted scatterplot smoothing of the data points.

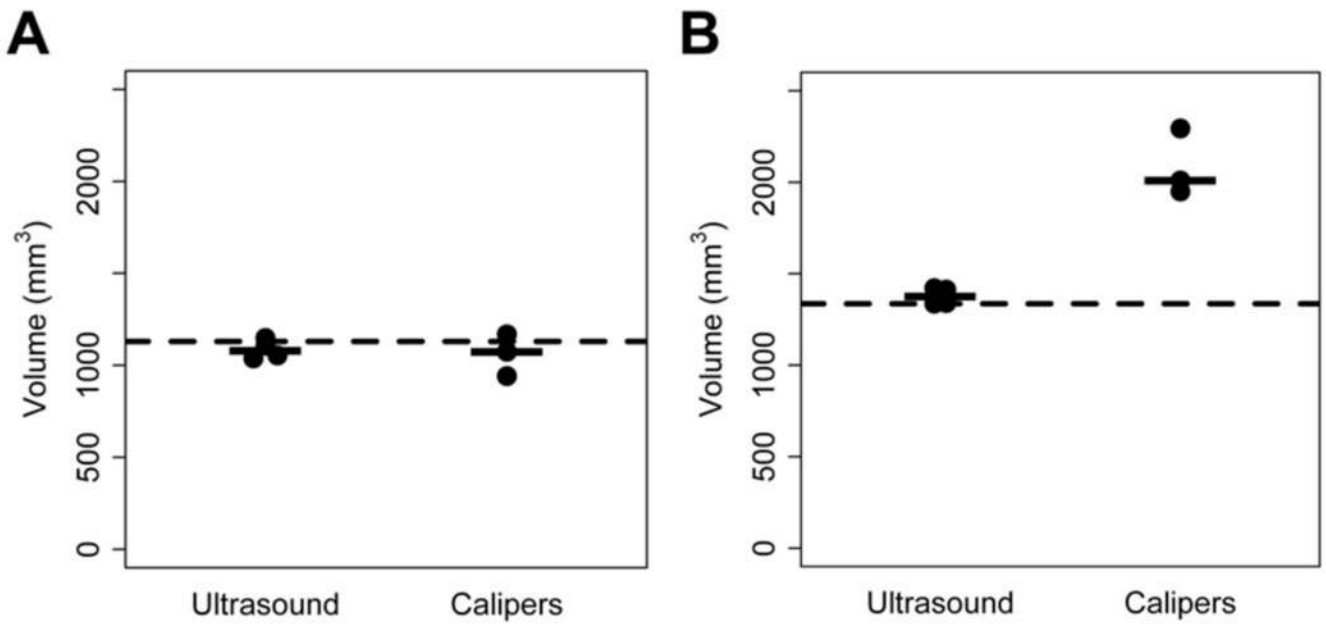

Figure 5.
Median error (solid bars) for an ellipsoidal shaped tumor (xenograft 9) was small for both ultrasound and calipers compared to the true volume (dashed line). Median error for a non-ellipsoidal shaped tumor (xenograft 10) was small for ultrasound but large for calipers. The large error in caliper measurement for non-ellipsoidal tumors derives from the assumption that tumors are ellipsoidal inherent in the equation used to calculate tumor volume with caliper measurements.

Figure 6.

Bland Altman plots for each reviewer illustrate the relationship between measured volume and true volume against the average of measured and true volumes. Consistent trends among reviewers within a measurement type show, 1) larger bias and variability among caliper measurements, 2) a tail trend among ultrasound reviewers for the largest 3 tumors, and 3) increasing variability for caliper measurements with the mean. Trends were consistent among all reviewers within a group.

Table 1

Tumor volume measurements for DLD1 and SW620 xenograft tumors for ultrasound imaging and caliper measurements are ranked by true tumor volume.

Tumor Rank	Tumor Cell Line	Ultrasound Volume (mm^3)	Caliper Volume (mm^3)	True Volume (mm^3)
1*	DLD1	496.78 ± 71.84	797.53 ± 56.61	460.0
2	SW620	436.78 ± 19.80	602.18 ± 65.40	471.9
3	SW620	563.00 ± 24.94	658.00 ± 104.42	555.0
4*	DLD1	612.06 ± 24.71	730.20 ± 54.23	668.3
5*	SW620	773.72 ± 58.53	916.65 ± 113.69	754.5
6*	DLD1	794.26 ± 54.87	1208.33 ± 179.51	860.3
7	SW620	777.79 ± 68.21	1231.60 ± 127.23	906.0
8	SW620	1179.74 ± 50.20	924.77 ± 123.08	996.3
9	DLD1	1086.93 ± 50.47	1061.31 ± 114.18	1130.5
10*	DLD1	1377.89 ± 45.79	2084.49 ± 183.92	1336.2
11*	DLD1	1412.32 ± 78.62	828.76 ± 57.14	1361.1
12	SW620	1564.18 ± 72.09	1576.29 ± 77.40	1755.2
13	SW620	1897.59 ± 114.47	2335.41 ± 163.41	2061.1
14	DLD1	1926.56 ± 79.71	2024.25 ± 317.88	2322.9

Tumor rank corresponds to figure 1, non-ellipsoidal tumors are indicated by (*).

Table 2

Model based estimates and bootstrap 95% confidence intervals (B=5,000) of total variance and source specific intraclass coefficient estimates for calipers (3 reviewers) and ultrasound based tumor volume measurements of the same 14 mice.

Source	Calipers	Ultrasound
Total Variance	340581 (240870 to 383716)	259773 (205978 to 261161)
ICC mouse	0.94 (0.912 to 0.983)	0.98 (0.978 to 0.992)
ICC reviewer	0.004 (0 to 0.040)	0.007 (0.004 to 0.015)
ICC error	0.06 (0.011 to 0.068)	0.006 (0.003 to 0.009)

Table 3

Per group sample size required to have 80% or 90% power to detect a given biologically meaningful difference in a two-arm treatment study assuming a two-side type I error rate of 5% using the two-sample t-test. Standard deviations among mice for caliper and ultrasound measurements were 583 mm^3 and 508 mm^3.

Average Difference in Tumor Volume Between Two Measurements	Power	Per Group Sample Size Using Calipers	Per Group Sample Size Using US	Difference	Ratio
−100 mm^3	80%	534	406	128	1.32
	90%	715	543	172	1.32
−250 mm^3	80%	87	66	21	1.32
	90%	115	88	27	1.31
−500 mm^3	80%	23	18	5	1.28
	90%	30	23	7	1.30