

# Variability in computed tomography diameter measurements of solid renal masses

Lorenzo P. Orton,<sup>1</sup> Richard H. Cohan,<sup>1</sup> Matthew S. Davenport,<sup>1</sup> Robert A. Parker,<sup>2</sup> Aishwarya Parameswaran,<sup>1</sup> Elaine M. Caoili,<sup>1</sup> Ravi K. Kaza,<sup>1</sup> Isaac R. Francis,<sup>1</sup> James H. Ellis,<sup>1</sup> J. Stuart Wolf,<sup>3</sup> Khaled Hafez<sup>3</sup>

<sup>1</sup>Department of Radiology, University of Michigan Hospitals, 1500 East Medical Center Drive, Room BID502G, Ann Arbor, MI 48109-5030, USA

<sup>2</sup>Directory of Biometry, Medical Practice Evaluation Center, Massachusetts General Hospital, Boston, MA 02114, USA

<sup>3</sup>Department of Urology, University of Michigan Hospitals, Ann Arbor, MI 48109-5030, USA

## Abstract

**Purpose:** To determine whether the frequency of intra-observer measurement discrepancies  $\geq 5$  mm for solid renal masses varies by renal mass characteristics and CT contrast phase.

**Materials and methods:** This HIPAA-compliant retrospective study was approved by our IRB. We selected single CT images performed during the nephrographic phase (NP) of renal enhancement in 97 patients, each with a single solid renal mass. Mass location, margin, heterogeneity, and growth pattern were assessed. Six readers measured each mass on two occasions  $\geq 3$  weeks apart. Readers also measured the masses on images in 50 patients who had corticomedullary phase (CMP) images obtained during the same study. Results were assessed using Chi-square/Fisher's exact and Wilcoxon Signed Rank tests, and logistic regression analyses.

**Results:** For NP to NP comparisons, intra-reader measurement differences  $\geq 5$  mm were seen for 3.7% (17/463) of masses  $< 4$  cm, but increased to 16.8% (20/119) for masses  $\geq 4$  cm ( $p < 0.0001$ ). Masses with poorly defined margins (15.9% [22/138] vs. 3.4% [15/444] for well-defined margins,  $p < 0.0001$ ) and heterogeneity (15.3% [22/144], vs. 5.0% [14/282] for minimally heterogeneous, vs. 0.6% [1/156] for homogeneous,  $p < 0.0001$ ), were more frequently associated with measurement differences  $\geq 5$  mm. Differences  $\geq 5$  mm were more frequent when only CMP images were utilized (14% [42/299]), or when CMP images were compared with NP images (26% [77/299]).

**Conclusions:** A  $\geq 5$  mm intra-reader variation in measured size of solid renal masses  $< 4$  cm is uncommon for NP to NP comparisons. Variation increases when masses are  $\geq 4$  cm, poorly defined, or heterogeneous; or when CMP images are utilized.

**Key words:** Renal mass measurements—Computed tomography—Solid renal mass

Despite recent improvements in the detection of small solid renal neoplasms, there has been no corresponding significant decrease in disease-specific mortality from renal cancer [1, 2]. While up to 20% of small solid renal masses are benign [3], it has been speculated that many of the remaining small renal neoplasms, though malignant, are likely indolent and might never grow large enough to cause symptoms, invade local structures, or produce regional or distant metastatic disease. These tumors would therefore probably never affect patient morbidity and mortality, especially in the elderly, many of whom have significant morbidity and mortality from other diseases [4].

As a result of these observations, active surveillance of renal masses has been increasingly used as a management strategy for patients with small incidentally detected renal neoplasms [5–12]. An important aspect of this strategy involves the performance of sequential follow-up imaging with cross-sectional imaging studies (e.g., CT), with more rapidly growing masses shown to be more likely to require eventual treatment [4, 8]. Thus, accurate determination of renal mass size and prompt recognition of rapid growth has become increasingly important in the clinical decision-making process.

There have been only a few studies evaluating repeatability of renal mass measurements on CT and these have included relatively small numbers of patients [13–15]. In one of these studies, inter-observer and intra-observer variability in renal mass measurements was determined to be 3 and 2 mm, respectively [14]. Partly due to this known measurement discrepancy rate, we and others [16] have used a measured change in a solid renal mass of 5 mm or more within a year to indicate rapid growth and an increased need for definitive treatment rather than active surveillance. The purpose of this study was to determine the frequency of intra-observer measurement discrepancies  $\geq 5$  mm for solid renal masses and to assess whether the frequency varies for renal masses having different CT characteristics and when renal masses are imaged during different phases of contrast material enhancement.

## Materials and methods

Prior to the initiation of this investigation, Institutional Review Board approval was obtained. This retrospective study was carried out in compliance with the Health Insurance Portability and Accountability Act (HIPAA, USA). Patient informed consent was not required based on institutional policy and the retrospective nature of this investigation.

### Subjects

We accessed an Interventional Radiology database for a list of all patients who had undergone percutaneous biopsy of a solid renal mass between April 2009 and December 2011. From this list, one radiologist chose a subset of masses imaged with CT that had prospective measurements (by the initial exam interpreters) of 1–5 cm in greatest dimension. The CTs were chosen from this list, even though they were performed using a variety of protocols, because the list included many small biopsy proven solid renal masses of a wide variety of sizes, allowing us to choose a total of 100 scans, which imaged masses equally divided among four size groups of interest. The first twenty-five masses on the list that met each of the following size ranges: 1.0–1.9, 2.0–2.9, 3.0–3.9, and 4.0–4.9 cm were selected. Three of these masses were subsequently excluded (one mass in the 1.0–1.9 cm group, one mass in the 3.0–3.9 cm group, and one mass in the 4.0–4.9 cm group), as they were displayed on images in which additional masses were present, which could have led to confusion (concerning which mass was being measured) when our experimental measurements were made. Our final study population therefore consisted of 97 patients with 97 solid renal masses (58 males [mean age: 64 years, range 59–81 years] and 39 females [mean age: 59 years, range 25–82 years]). Fifty-four masses were in the right kidney and 43 masses were in the

left kidney. Sixty-seven of the masses were subsequently diagnosed as malignant (all of which were renal cancer), 28 as benign, and two were indeterminate (representing oncocytic renal neoplasms in patients who had no further tissue sampling).

### Renal mass protocol for computed tomography

All masses were imaged on 16 or 64 slice helical CT scanners using 0.625–5 mm image thickness and reconstruction intervals (0.625 mm [ $N = 1$ ], 1.25 mm [ $N = 1$ ], 2.5 mm [ $N = 41$ ], 3.0 mm [ $N = 6$ ], 3.75 mm [ $N = 2$ ], 4 mm [ $N = 1$ ], 5 mm [ $N = 44$ ], 5.5 mm [ $N = 1$ ]) with the varying slice thicknesses utilized primarily because many patients had studies performed at outside institutions. Delayed enhanced images were obtained in each patient during the nephrographic phase (NP), with homogeneous nephrograms demonstrated. Our institutional technique consisted of obtaining NP images at least 100 s following the initiation of the contrast material injection of 100 or 125 mL of 300 mg I/mL nonionic iodinated contrast media.

From the included group of 97 masses, 50 patients also had corticomedullary (CMP) images obtained during the same CT when the NP images were acquired. Images for these series obtained beginning at 60–70 s following the initiation of contrast material administration. The CMP masses belonged to the following size groups: 1.0–1.9 cm [ $N = 12$ ], 2.0–2.9 cm [ $N = 17$ ], 3.0–2.9 cm [ $N = 14$ ], 4.0–4.9 cm [ $N = 7$ ]. Image thicknesses for the CMP images ranged between 0.625 and 5.5 mm (0.625 mm [ $N = 1$ ], 1.25 mm [ $N = 0$ ], 2.5 mm [ $N = 5$ ], 3.0 mm [ $N = 3$ ], 3.75 mm [ $N = 2$ ], 4 mm [ $N = 1$ ], 5 mm [ $N = 38$ ], 5.5 mm [ $N = 0$ ]). Three patients were scanned with thinner CMP than NP image collimation (0.625 vs. 5.0 mm [ $N = 1$ ] and 2.5 vs. 5.0 mm [ $N = 2$ ]), and five patients were scanned with thicker CMP than NP image collimation (5.0 vs. 2.5 mm [ $N = 5$ ]). In the remaining 42 patients, CMP and NP image thickness were identical.

### Image measurement protocol

To minimize variation that might be introduced by readers misidentifying a targeted renal mass, or readers making measurements in a non-standard plane of acquisition, the readers were given only a single image for each CT examination contrast phase on which to make their diameter measurements. These representative images were selected by one abdominal radiologist prior to study initiation and chosen to reflect the axial slice demonstrating the largest diameter of the mass in each phase. The location of the mass (with respect to side) was indicated on the reader data sheet. Each mass was then assessed by the same abdominal radiologist (on each selected NP image for the following features: (1) margins

[whether or not the margins of the mass were well-defined or poorly defined], (2) heterogeneity (subjectively graded as homogeneous, mildly heterogeneous [with slight internal variations in attenuation], or very heterogeneous [with pronounced internal variations in attenuation]), (3) location (polar or interpolar based upon whether or not the mass crossed the renal polar lines), and (4) growth pattern (whether the mass was  $>$  or  $\leq 50\%$  exophytic).

Six readers, each of whom is also an experienced abdominal radiologist, measured each mass in two dimensions in the axial plane, recording a maximal diameter measurement and then a second short-axis diameter measurement perpendicular to the first. Measurements were made from outer to outer mass margin, using electronic calipers on a workstation (Horizon Medical Imaging—Version 11.2, McKesson Information Solutions, Richmond, BC, Canada). Windowing and leveling were performed at the discretion of each reader. The readers then repeated measurements of the same masses with a minimum interval time period of three weeks between review sessions. The two sets of CMP measurements were obtained after the NP measurements had been completed.

### *Statistical methods*

For the 97 included images, there were a total of 582 (97 images  $\times$  6 readers) observations of both the maximum size and the orthogonal size measurement for each of two reading sessions. The difference between each of the two maximum diameter measurements was calculated and the difference between each of the two orthogonal measurements was calculated. If either difference was  $\geq 5$  mm, we defined this as a clinically relevant discrepancy. The larger of the two measurement differences (in absolute value) was defined as the magnitude of the discrepancy. The same procedure was followed for comparing the two sets of CMP image measurements in the subset of 50 patients who also had CMP images provided. In order to assess the impact of different types of renal enhancement on the rate of measurement discrepancies, we also compared the first set of NP measurements to the first set of CMP measurements in the 50 patients who had both sets of images obtained.

As the trend of discrepancies did not increase linearly with increases in lesion size across the measured groups, we used an indicator variable for lesions which were 4.0–4.9 cm in average size when comparing them to lesions  $< 4.0$  cm in average size, rather than using the actual size of the lesion in analyses. A cut-off of 4.0 cm was employed, as this is used as a threshold by our urologists to indicate the need for treatment in patients who are undergoing active surveillance. Statistical analysis was performed using SAS V9.3 (SAS Institute, Cary, NC). Chi-square or Fisher's exact test (with extensions) was

used to compare discrepancy rates for groups composed of masses with each individual characteristic (e.g., growth pattern and size category). Logistic regression (both univariate and multivariate) was employed to estimate the odds ratio of a discrepancy for the different mass characteristics. Intrareader discrepancies for the different phase comparisons (NP vs. NP images, CMP vs. CMP images, and NP vs. CMP images) were tested using the Wilcoxon Signed Rank test. Measurement discrepancy rates were then calculated for mm size thresholds other than  $\geq 5$  mm for each of the three sets of comparisons (NP vs. NP, CMP vs. CMP, and NP vs. CMP) to estimate appropriate alternative thresholds.

## Results

### *Renal mass location and morphology on NP images*

Seventy-four of the 97 renal masses were classified as well-marginated and 23 as poorly marginated. Forty-seven renal masses were mildly heterogeneous, 26 renal masses were completely homogeneous, and 24 renal masses were very heterogeneous. Fifty-six renal masses were polar in location, and 41 renal masses were in the mid kidney. Fifty-eight renal masses were endophytic and 39 renal masses were exophytic.

### *Differences in discrepancy rates among the six readers*

Table 1 demonstrates the frequency with which intra-reader measurements differed by  $\geq 5$  mm for each of the three comparisons by each reader (NP vs. NP for 97 masses, CMP vs. CMP for 50 masses, and NP vs. CMP for 50 masses). The six  $\geq 5$  mm intra-reader discrepancy rates were generally low for the NP vs. NP comparisons, ranging from 2% (2/97) for one reader to 10% (10/97) for another. The differences in discrepancy rates were not significantly different among readers ( $p = 0.12$ ). Measurement variations of  $\geq 5$  mm between review sessions were more common when CMP images were evaluated (Fig. 1). When a comparison was made between the two CMP series review sessions, discrepancy rates of  $\geq 5$  mm among the six readers ranged from 8% (4/49) to 22% (11/50); the differences among readers were not significantly different ( $p = 0.36$ ). However, for each of the six readers, there was a significantly higher  $\geq 5$  mm discrepancy rate for the CMP to CMP comparisons than for the NP to NP comparisons ( $p = 0.031$ ). When NP measurements made during one review were compared with CMP measurements made during another review, the frequency of  $\geq 5$  mm discrepancies was even higher, ranging between 20% (10/50) and 32% (16/50) for the different readers. The differences among the readers were again not significant ( $p = 0.69$ ). The discrepancy rates for CMP to NP comparisons for each of the six readers

**Table 1.** Intra-reader measurement discrepancy rate ( $\geq 5$  mm) for solid renal masses in the nephrographic and corticomedullary phases

Reader	NP vs. NP 97 masses	CMP vs. CMP 50 masses	NP vs. CMP 50 masses
1	6% $n = 6$	22% $n = 11$	26% $n = 13$
2	2% $n = 2$	10% $n = 5$	20% $n = 10$
3	9% $n = 9$	14% $n = 7$	32% $n = 16$
4	7% $n = 7$	8% $n = 4^a$	20% $n = 10^a$
5	3% $n = 3$	18% $n = 9$	26% $n = 13$
6	10% $n = 10$	12% $n = 6$	30% $n = 15$
Total comparisons	6% $n = 37/582$	14% = 42/299	26% $n = 77/299$

<sup>a</sup> This reader could not identify one of the masses during one of the review sessions, resulting in only 49 rather than 50 comparisons  
*NP* Nephrographic phase images, *CMP* corticomedullary phase images,  $n$  number of cases in which there was a  $\geq 5$  mm discrepancy

were significantly higher than the discrepancy rates for CMP to CMP comparisons, as well as the discrepancy rates for NP to NP comparisons (both  $p = 0.031$ ).

### *Impact of lesion characteristics on intra-reader NP measurement discrepancy rates*

**Renal mass size.** The intra-reader NP to NP  $\geq 5$  mm discrepancy rate was uncommon for renal masses  $< 4$  cm. For masses  $< 2.0$  cm, the rate was 0.8% (1/132). For masses 2.0–2.9 cm, the rate was 6.0% (11/182), and for masses 3.0–3.9 cm, the rate was 3.4% (5/149) (Fig. 2). These differences were statistically significant ( $p = 0.046$ ). For renal masses  $\geq 4$  cm, the discrepancy rate was much higher (17% [20/119],  $p < 0.0001$ ). Renal masses  $\geq 4$  cm were significantly more likely to be associated with a  $\geq 5$  mm intra-reader discrepancy compared to renal masses  $< 4$  cm ( $p \leq 0.0001$ , OR 5.3 [95% confidence interval (CI) 2.6–10.5]).

**Renal mass margins.** Intra-reader NP to NP measurement discrepancies  $\geq 5$  mm were observed more frequently for poorly defined than for well-defined renal masses (15.9% [22/138] vs. 3.4% [15/444], respectively,  $p < 0.0001$ ; odds ratio [OR] 5.4 [95% CI 2.7–10.8]) (Fig. 3). Masses  $< 4$  cm in diameter were less likely to be poorly defined than those  $\geq 4$  cm (13/73 [18%] vs. 10/24 [42%], respectively) ( $p = 0.017$ ).

**Renal mass heterogeneity.** Heterogeneous renal masses resulted in more NP to NP intra-reader measurement discrepancies of  $\geq 5$  mm than did homogeneous masses. For example, the  $\geq 5$  mm discrepancy rate was 0.6% (1/156) for homogeneous masses, 5.0% (14/282) for mildly heterogeneous masses, and 15.3% (22/144) for very heterogeneous masses (overall  $p < 0.0001$ ). Compared to homogeneous renal masses, mildly heterogeneous and very heterogeneous masses were significantly more likely to result in a  $\geq 5$  mm discrepancy (overall  $p < 0.0001$ ; OR 8.1 [95% CI 1.05–62] and OR 28 [95% CI 3.7–210], respectively) (Fig. 4). There was no significant difference in the percentage of masses  $< 4$  cm in diameter that were either mildly or markedly heterogeneous in comparison to masses  $\geq 4$  cm (51/73 [70%] vs. 20/24 [83%], respec-

tively) ( $p = 0.20$ ). In comparison, renal masses  $< 4$  cm were less likely to be markedly heterogeneous than masses measuring  $\geq 4$  cm (14/73 [19%] vs. 10/24 [42%], respectively) ( $p = 0.026$ ).

**Renal mass location.** Intra-reader NP to NP renal mass measurement discrepancies of  $\geq 5$  mm occurred significantly more frequently for polar masses than for masses located at least partially between the polar lines (8.9% [30/336] vs. 2.8% [7/246], respectively,  $p < 0.005$ , OR 3.3 [95% CI 1.4–7.8]).

**Renal mass growth pattern.** Intra-reader NP to NP renal mass measurement discrepancies of  $\geq 5$  mm were more frequent for exophytic than endophytic masses (8.1% [19/234] vs. 5.2% [18/348]), but this difference was not significant ( $p = 0.15$ ).

**Impact of multiple predictors on rate of measurement discrepancy.** Using multivariate logistic regression, independent predictors of a  $\geq 5$  mm discrepancy included mass margins ( $p = 0.0009$ , OR 3.5 [95% CI 1.6–7.4]), mass heterogeneity ( $p = 0.0012$ , minimal heterogeneity OR 4.3 [95% CI 0.5–35], marked heterogeneity OR 13.7 [95% CI 1.7–107]), and mass size  $\geq 4.0$  cm ( $p = 0.0030$ , OR 3.0 [95% CI 1.4–6.4]). Renal mass location was not a significant independent predictor of discrepancy after adjusting for these other factors ( $p = 0.28$ ).

### *Impact of renal mass characteristics on CMP to CMP measurement discrepancy rates*

Unlike for NP images, specific renal mass characteristics were not independent predictors of  $\geq 5$  mm intra-reader discrepancies for CMP to CMP comparisons. The  $\geq 5$  mm discrepancy rates for the six different readers were similar for all renal mass size groups, ranging from 12.0% (6/50) to 16.2% (16/99) (Fig. 1;  $p = 0.89$ ). Thus, unlike NP to NP comparisons, the likelihood of a  $\geq 5$  mm discrepancy for CMP to CMP comparisons was the same for renal masses  $< 2.0$  cm as for renal masses  $\geq 4.0$  cm. Similarly, there was no significant effect of renal mass margins, renal mass location, renal mass growth pattern,



**Fig. 1.**  $\geq 5$  mm measurement discrepancy on corticomedullary phase to corticomedullary phase comparisons. **A** Nephrographic phase image in a 56-year-old man demonstrates a solid mass measured by the reviewers as being between 2.4 and 3.5 cm in maximal diameter. The greatest measurement discrepancy was 4 mm (by one reviewer). **B** On corticomedullary phase images obtained from the hypoenhancing renal medulla, which likely explains why four reviewers made measurements that disagreed with one another by  $\geq 5$  mm.

or renal mass heterogeneity (all  $p$  values  $> 0.10$ ; Table 2) on CMP to CMP discrepancy rates.

### *Impact of renal mass characteristics on CMP to NP measurement discrepancy rates*

Intra-reader discrepancy rates were highest when comparisons were made between CMP and NP measurements, ranging from 21% (17/81) to 29% (17/59) depending on renal mass size (Fig. 1). The CMP to NP discrepancy rates were not affected by renal mass size ( $p = 0.70$ ). As with the NP to NP comparisons, renal mass margin was an important predictor of intra-reader  $\geq 5$  mm discrepancies (poorly defined: 45% [27/60] vs. well-defined: 21% [50/239],  $p < 0.0001$ ; Table 2). Renal mass heterogeneity was also a predictor of intra-reader  $\geq 5$  mm discrepancies (homogeneous: 18% [18/101] vs. minimally heterogeneous 32% [44/138] vs. very heterogeneous 25% [15/60],  $p < 0.05$ ). Renal mass location was not an independent predictor of measurement discrepancies ( $p = 0.16$ ), but an exophytic growth pattern was (exophytic: 19% [28/150] vs. 33% [49/149] endophytic,  $p < 0.005$ ).

### *Intra-reader discrepancy rates for different discrepancy thresholds*

Intra-reader measurement differences exceeding different thresholds are summarized in Table 3. Repeat size measurements by the same reader were concordant at least 95% of the time when the threshold for discordance was increased to  $\geq 6$  mm for NP vs. NP,  $\geq 9$  mm for CMP vs. CMP, and  $\geq 13$  mm for NP vs. CMP comparisons.

## Discussion

The clinical management of small incidentally detected solid renal masses has trended toward increasing use of active surveillance [5, 7–9, 11, 17]. This is because, many small renal neoplasms, even those that are malignant, do not produce significant morbidity or mortality. Some renal masses that meet histologic criteria for malignancy have an indolent growth pattern, while others arise in the setting of severe comorbidities that may be better predictors than the renal mass of the patient's eventual clinical outcome [4].

In recently published series assessing the role of active surveillance of renal masses, detection of a rapid growth rate has been shown to be a strong indicator that a patient will need subsequent treatment [6, 7, 9, 17]. For example, in one study that included 470 observed renal masses, all seven renal masses that eventually metastasized grew more rapidly prior to the appearance of the metastatic foci in comparison to the renal masses that did

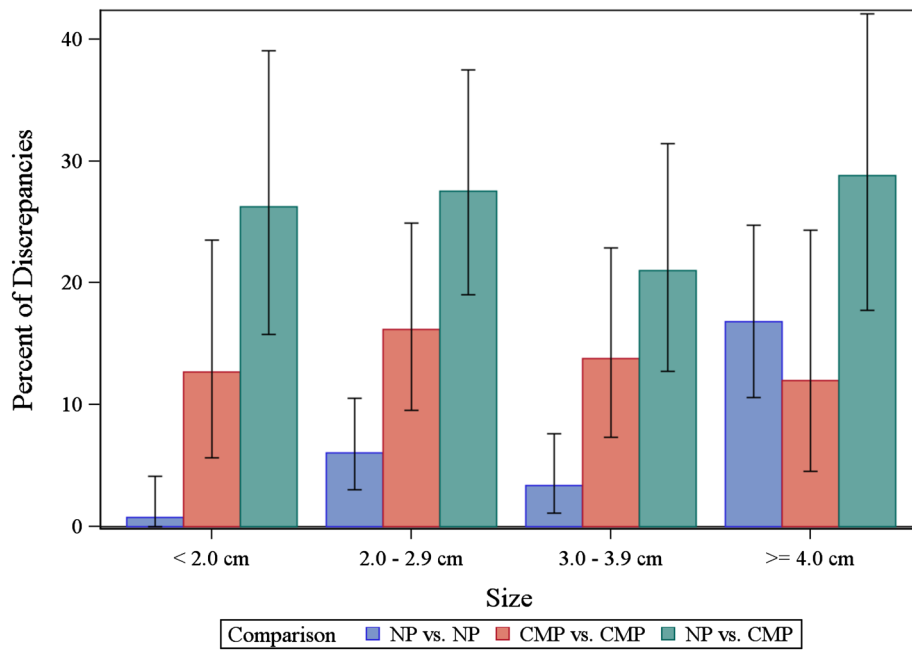


Fig. 2. Percentage of  $\geq 5$  mm intra-reader discrepancies in size measurements in four different size groups; for 97 NP vs. NP comparisons (blue), 50 CMP vs. CMP comparisons (red), and 50 NP vs. CMP comparisons (green) for all six reviewers. For NP vs. NP comparisons, discrepancy rates were highest for masses  $\geq 4$  cm. For CMP vs. CMP and NP vs. CMP comparisons,

there was no relationship between renal mass size and  $\geq 5$  mm discrepancy rate. Comparisons using any CMP images were more likely to vary by  $\geq 5$  mm than comparisons using only NP images ( $p = 0.031$ ). Discrepancy rates were highest for NP vs. CMP comparisons.

not metastasize [18]. Since early detection of rapid growth within a renal mass is important, measurement accuracy on follow-up imaging studies has become increasingly utilized [7]. In order to best determine whether a measured increase in renal mass size indicates true renal mass growth, some quantification of intra-observer and inter-observer variation of renal mass measurement on CT is essential.

Intra-observer and inter-observer variation in CT lesion diameter measurements has been evaluated in the past, in a variety of non-genitourinary tract masses, including the liver [19] and abdominal aorta [20–23]. In these studies, intra-observer and inter-observer variation has been small. In two series, maximal diameter measurements of aortic aneurysms differed by 2 mm or less in 90% or more of patients [20, 21]. In another series, of 25 measured abdominal aortic aneurysms [22], the mean inter-observer measurement difference for a standardized well-defined approach was 2.8 mm; however, when an explicit description on how the aneurysm should be measured was not provided, the mean average measurement difference increased to 4.0 mm. CT measurements have also been observed to be less reproducible in masses that have irregular shapes or that are poorly defined [24].

Only a few studies have assessed consistency of renal mass measurements on CT [13, 14]. In a series of 16 renal tumors, Tann et al. [13] found that there was good intra-

observer and inter-observer agreement for renal tumor volume measurements, although there were considerable differences between the CT and specimen volume measurements. In a study of 29 renal masses in 21 patients assessed by three radiologists, Punnen et al. [14] observed an intra-observer variation in maximal diameter measurement of the same renal mass of 2.3 mm and an inter-observer variation of 3.1 mm. These variations in renal mass measurement are similar to those observed in the aorta [20–22].

Given the known 2 mm intra-observer and 3 mm inter-observer variation that exists for CT diameter measurements, a larger change in renal mass diameter, such as an increase in 4 or 5 mm, is needed to identify a true change in a lesion's size. A 5 mm increase in diameter represents a considerable increase in volume of a small renal mass. For example, a 5 mm increase in size of a 1.0 cm spherical renal mass indicates more than a tripling of the renal mass volume. It has been recommended that a 5 mm change over a year be used as a threshold value to trigger the transfer of patients from active surveillance to treatment with partial nephrectomy or ablation [16]. Our study was designed to evaluate the frequency of intra-observer measurement variation of solid renal masses using this potentially clinically relevant threshold.

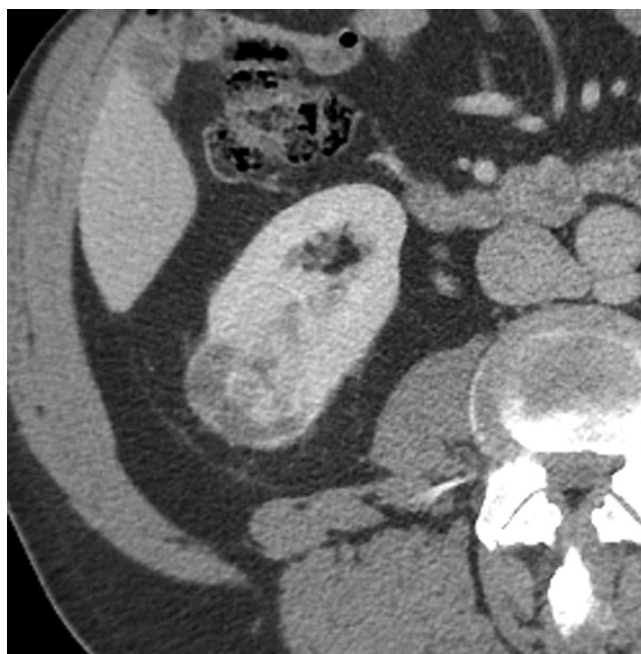
We assessed the frequency with which a change in renal mass dimension of  $\geq 5$  mm might be erroneously



**Fig. 3.** <5 mm measurement discrepancy on nephrographic phase to nephrographic phase comparisons. Nephrographic phase image in a 49-year-old woman demonstrates a well-defined homogeneous solid renal mass that was measured by the reviewers as being between 3.7 and 4.1 cm in maximal diameter. There were no measurement discrepancies of more than 2 mm for any of the six reviewers.

identified under very strict conditions (in which the specific image on which the measurement is to be made is provided). We chose this methodology in order to determine the minimal discrepancy rate that could be obtained under the most ideal circumstances. Using this approach, the principal finding in our study, which included six different experienced readers, is that renal masses <4 cm in maximal diameter could be measured reliably on NP images with <5 mm variance the vast majority of time, although  $\geq 5$  mm measurement discrepancies were more common when certain renal mass characteristics were present (poorly defined margins, heterogeneity, and polar location). Utilization of CMP images was much more problematic and resulted in higher  $\geq 5$  mm discrepancy rates, with the greatest discrepancy rates occurring when NP images were compared to CMP images. Our study indicates that when a radiologist compares a NP CT measurement to a CMP measurement, he or she should expect that a measurement difference of  $\geq 5$  mm can be encountered about 25% of the time, even when a mass has not truly changed in size.

For NP to NP comparisons, we noted that  $\geq 5$  mm measurement discrepancies were more common in larger



**Fig. 4.**  $\geq 5$  mm measurement discrepancy on NP to NP comparisons. NP image in a 55-year-old man demonstrates a very heterogeneous solid renal mass that was measured by the reviewers as being between 4.0 and 6.0 cm in maximal diameter. Three reviewers had  $\geq 5$  mm measurement discrepancies that were likely related to lesion heterogeneity.

masses. The reasons for this finding are uncertain; however, it is possible that there were greater variations in the axes selected for measuring the larger renal masses. Additionally, a significantly higher percentage of masses  $\geq 4$  cm in diameter were markedly heterogeneous and poorly defined, differences that are likely at least partially responsible for the greater inconsistency in the measurements of these lesions.

Our observations that renal mass measurements comparing CMP with either NP or CMP images are prone to greater intrareader variability in contrast to NP to NP comparisons is not surprising. It has been shown that renal masses are best detected and characterized when NP or excretory phase images are used (rather than CMP images) [25, 26]. This is because, on CMP images, it can be difficult to distinguish between hypervascular components of renal tumors and relatively hyperenhancing normal renal cortex, as well as between poorly enhancing cystic or necrotic components of renal tumors and hypoenhancing normal renal medulla. On NP and excretory phase images, normal renal parenchyma is homogeneous. Renal neoplasms are often more easily identified and localized, and probably more easily measured, on these more delayed images.

In a study of 40 solid renal masses, Rosencrantz and colleagues assessed the accuracy of renal mass measurements in detecting growth [15]. The authors of this study

**Table 2.** Effect of solid renal mass characteristics on intra-reader measurement  $\geq 5$  mm discrepancy rate

	NP vs. NP 97 masses (582 comparisons)	<i>p</i> value	CMP vs. CMP 50 masses <sup>a</sup> (299 comparisons) <sup>b</sup>	<i>p</i> value	CMP vs. NP 50 masses <sup>a</sup> (299 comparisons) <sup>b</sup>	<i>p</i> value
Margins						
Well-defined	15/444 (3.4%)	<0.0001	30/239 (12.6%)	0.14	50/239 (20.9%)	<0.0001
Poorly defined	22/138 (15.9%)		12/60 (20.0%)		27/60 (45.0%)	
Heterogeneity						
None	1/156 (0.6%)	<0.0001	15/101 (14.9%)	0.69	18/101 (17.8%)	<0.05
Minimally heterogeneous	14/282 (5.0%)		17/138 (12.3%)		44/138 (31.9%)	
Very heterogeneous	22/144 (15.3%)		10/60 (16.7%)		15/60 (25.0%)	
Location						
Polar	30/336 (8.9%)	0.003	20/150 (13.3%)	0.72	44/150 (29.3%)	0.16
Interpolar	7/246 (2.8%)		22/149 (14.8%)		33/149 (22.1%)	
Growth pattern						
Primarily (up to 50%) endophytic	18/348 (5.2%)	0.15	25/149 (16.8%)	0.18	49/149 (32.9%)	0.0049
Primarily (> 50%) exophytic	19/234 (8.1%)		17/150 (11.3%)		28/150 (18.7%)	

<sup>a</sup>Renal mass characteristics were those assessed on NP images

<sup>b</sup>One of the six readers failed to identify one of the 50 masses during one of the review sessions, resulting in only 299 comparisons

Percentages indicate the intra-reader discrepancy rate between image sets for all renal masses with each of the following characteristics. Renal mass characteristics were defined on nephrographic phase (NP) images

used a summation of areas technique to create a reference standard to determine whether a renal mass had enlarged on serial CT scans. The authors showed that two-dimensional and three-dimensional measurements were more accurate than a subjective impression or a one-dimensional measurement, with good inter-reader measurement agreement [15]. Differences between this study and ours should be emphasized. Our study was designed to assess variability in repetitive renal mass measurements in the absence of growth rather than to determine the ability of such measurements to detect growth when it is present. In contrast, Rosencrantz et al did not attempt to determine error rates in the absence of renal mass growth.

Use of different measurement thresholds will affect the frequency with which size differences will be spuriously detected in stable renal masses. The choice of a 5 mm threshold difference to indicate true growth, although reasonable given the known mean variations of 2 mm for intra-observer and 3 mm for inter-observer measurements, is arbitrary. In fact, our analysis shows that if the goal is to use a threshold for which masses would be correctly identified as being stable more than 95% of the time, measurement thresholds would have to be raised from 5 mm to: 6 mm (NP vs. NP), 9 mm (CMP vs. CMP), and 13 mm (NP vs. CMP). Smaller thresholds could be used if greater discrepancy rates can be tolerated. So, for example, if it is acceptable to exceed a threshold for stable renal masses no more than 10% rather than no more than 5% of the time, NP vs. NP, CMP vs. CMP, and NP vs. CMP thresholds can be lowered to 5, 6, and 9 mm, respectively.

Although many urologists are now making decisions about renal mass management based upon lesion size or growth rates, it is possible that other changes in renal mass morphology (such as increases in heterogeneity or more poorly defined margins) will be identified that can also be used to suggest that a lesion under active surveillance may need to undergo treatment. More research in this area is needed.

Our study has several important limitations. We restricted our reader analysis to only one image per mass, which likely reduced measurement variability. Interpreting radiologists would otherwise have been required to select a representative image from a complete CT and different radiologists might have chosen different images on which to make their comparison measurements. We also chose to focus on intra-observer rather than inter-observer measurement variability. This is because, our methodology was utilized in order to determine the absolute minimum variability that might be expected under the most ideal circumstances. In clinical practice, the encountered variability would likely be larger. Our study is also limited by our reliance upon absolute measurement differences of 5 mm or more between two studies for most of our analyses. A measurement



**Table 3.** Frequency of intra-reader discrepancies by imaging phase comparison using various measurement difference thresholds

mm	NP vs. NP (%) 582 comparisons	CMP vs. CMP (%) 299 comparisons	NP vs. CMP (%) 299 comparisons
≥0	582 (100)	299 (100)	299 (100)
≥1	533 (91.6)	273 (91.3)	290 (97.0)
≥2	275 (47.3)	154 (51.5)	231 (77.3)
≥3	109 (18.7)	77 (25.8)	165 (55.2)
≥4	61 (10.5)	54 (18.1)	109 (36.5)
≥5	37 (6.4)	42 (14.0)	77 (25.8)
≥6	<b>27 (4.6)</b>	28 (9.4)	56 (18.7)
≥7	<b>21 (3.6)</b>	20 (6.7)	43 (14.4)
≥8	<b>18 (3.1)</b>	15 (5.02)	35 (11.7)
≥9	<b>17 (2.9)</b>	<b>12 (4.0)</b>	29 (9.7)
≥10	<b>14 (2.4)</b>	<b>10 (3.3)</b>	24 (8.0)
≥11	<b>12 (2.1)</b>	<b>9 (3.0)</b>	19 (6.4)
≥12	<b>11 (1.9)</b>	<b>7 (2.3)</b>	15 (5.02)
≥13	<b>10 (1.7)</b>	<b>5 (1.7)</b>	<b>12 (4.0)</b>

Size thresholds, which if used, would result in a measurement discrepancy rate of ≤5% are indicated in bold font

reduction of 5 mm would also be considered to be a significant outlier. In clinical practice, such a change would not be interpreted as indicating renal mass growth. We chose this technique merely to determine what the variability in renal mass measurement is between two repeated measurements.

In addition, our study included renal masses imaged axially using a variety of CT techniques with many of the included studies having been performed at outside institutions, with image thicknesses varying as a result from 0.625 to 5.0 mm. This also likely led to some variation in the volume and concentration of contrast material administered, and to some differences in reconstruction technique. Although measurement variations might differ when CT scans of differing techniques are utilized, we believe that this was likely not a substantial factor affecting measurement differences in our study, since images were only chosen on which the renal mass could be clearly depicted and each image was utilized for each of two different measurements with image thickness obviously being the same for both measurements. Still, it is possible that discrepancy rates could vary for images of different thickness. Also, it is possible that discrepancies in the measurements of polar renal masses might have been reduced had coronal or sagittal reconstructed images been available for use. Also, while CMP images were matched to NP images as much as possible, these two different images were obtained at different times. As a result, additional variation may have been introduced for the NP to CMP comparisons, perhaps at least partly explaining why discrepancy rates were highest when NP images were compared to CMP images.

Finally, as previously stated, our study did not assess the ability of readers to detect renal mass growth reliably, which is the most important feature of serial renal mass measurement. Of course, it is difficult, and perhaps even impossible, to determine whether any renal mass has truly grown over a short interval. This is why Tann et al. and Rosenkrantz et al. [13, 15] have relied upon volu-

metric measurements as a gold standard, though these measurements, too, can contain errors.

In summary, solid renal mass size measurement is less prone to error when NP images are used rather than CMP images, but NP measurement variation also increases for larger, more poorly marginated, and heterogeneous masses. In particular, for NP images, an erroneously detected change in size of ≥5 mm occurs uncommonly (3%) in a <4 cm solid renal mass but increases for masses ≥4 cm (17%); or for masses that are poorly defined (16%) or mildly (5%) or very (15%) heterogeneous. Unneeded intervention/treatment might therefore be performed in up to one in six patients with masses possessing these imaging characteristics. For this reason, larger size thresholds could and likely should be employed as an indicator for intervention when these characteristics are present on NP images or when CMP images are utilized. Alternatively, if a 5 mm threshold is still to be utilized even for problematic renal lesions, additional assessment could be considered prior to definitive therapy (such as with mass measurement by another reader, MRI, a third short-interval follow-up CT, or repeat renal mass biopsy). Additional work will be required to assess the validity of any of these other approaches.

## References

1. Chow WH, Devesa SS, Warren JL, Fraumeni JF Jr (1999) Rising incidence of renal cell cancer in the United States. *JAMA* 281:1628–1631
2. Hollingsworth JM, Miller DC, Daignault S, et al. (2006) Rising incidence of small renal masses: a need to reassess treatment effect. *J Natl Cancer Inst* 98:1331–1334
3. Frank I, Blute ML, Cheville JC, et al. (2003) Solid renal tumors: an analysis of pathological features related to tumor size. *J Urol* 170:2217–2220
4. Thomas AA, Campbell SC (2011) Small renal masses: toward more rational treatment. *Clev Clin J Med* 78:539–547
5. Kassouf W, Aprikian AG, LaPlante M, Tanguay S (2004) Natural history of renal masses followed expectantly. *J Urol* 171:111–113

6. Chawla SN, Crispen PL, Hanlon AL, et al. (2006) The natural history of observed enhancing renal masses: meta-analysis and review of the world literature. *J Urol* 175:425–432
7. Kouba E, Smith A, McRackan D, Wallen EM, Pruthi RS (2007) Watchful waiting for solid renal masses: insight into the natural history and results of delayed intervention. *J Urol* 177:466–470
8. Kunkle DA, Kutikov A, Uzzo RG (2009) Management of small renal masses. *Semin Ultrasound CT MR* 30:352–358
9. Rosales JC, Haramis G, Moreno J, et al. (2010) Active surveillance for renal cortical neoplasms. *J Urol* 183:1698–1702
10. Haramis G, Mues AC, Rosales JC, et al. (2011) Neoplasms during active surveillance with follow up longer than 5 years. *Urology* 77:787–791
11. Mason RJ, Abdolell M, Trotter G, et al. (2011) Growth kinetics of renal masses: analysis of a prospective cohort of patients undergoing active surveillance. *Eur Urol* 59:863–867
12. Halvorson SJ, Kunju LP, Bhalla R, et al. (2013) Accuracy of determining small renal mass management with risk stratified biopsies: confirmation by final pathology. *J Urol* 189:441–446
13. Tann M, Sopov V, Croitoru S, et al. (2001) How accurate is helical CT volumetric assessment in renal tumors? *Eur Radiol* 11:1435–1438
14. Punnen S, Haider MA, Lockwood G, et al. (2006) Variability in size measurement of renal masses smaller than 4 cm on computerized tomography. *J Urol* 176:2386–2390
15. Rosenkrantz AB, Mussi TC, Somberg MB, Taneja SS, Babb JS (2012) Comparison of CT-based methodologies for detection of growth of solid renal masses on active surveillance. *AJR* 199:373–378
16. Pierorazio PM, Hyams ES, Mullens JK, Allaf ME (2012) Active surveillance of small renal masses. *Rev Urol* 14:13–19
17. Rendon RA, Stanietzky N, Panzarella T, et al. (2000) The natural history of small renal masses. *J Urol* 164:1143–1147
18. Kunkle DA, Chen DY, Greenberg RE, et al. (2009) Metastatic progression of enhancing renal masses under active surveillance is associated with rapid interval growth of the primary tumor. *J Urol* 179:1089
19. Prasad SR, Jhaveri KS, Saini S, et al. (2002) CT tumor measurement for therapeutic response assessment: comparison of unidimensional, bidimensional, and volumetric techniques—initial observations. *Radiology* 225:416–419
20. Lederle FA, Wilson SE, Johnson GR, et al. (1995) Variability in measurement of abdominal aortic aneurysms. Abdominal Aortic Aneurysm Detection and Management Veterans Administration Cooperative Study Group. *J Vasc Surg* 21:945–952
21. Singh K, Jacobsen BK, Solberg S, et al. (2003) Intra- and interobserver variability in the measurements of abdominal aortic and common iliac artery diameter with computed tomography. The Tromsø Study. *Eur J Vasc Endovasc Surg* 25:399–407
22. Cayne NS, Veith FJ, Lipsitz EC, et al. (2004) Variability in maximal aortic aneurysm measurements on CT scan: significance and methods to minimize. *J Vasc Surg* 39:811–815
23. Diehm N, Kickuth R, Gahl B, et al. (2007) Intraobserver and interobserver variability of 64-row computed tomography abdominal aortic aneurysm neck measurements. *J Vasc Surg* 45:263–268
24. Hopper KD, Kasales CJ, Slyke MA, et al. (1996) Analysis of interobserver and intraobserver variability in CT tumor measurements. *AJR* 167:851–854
25. Cohan RH, Sherman LS, Korobkin M, Bass JC, Francis IR (1995) Renal masses: assessment of corticomedullary-phase and nephrographic-phase CT scans. *Radiology* 196:445–451
26. Yuh BI, Cohan RH (1999) Different phases of renal enhancement: role in detecting and characterizing renal masses during helical CT. *AJR* 173:747–755