Published in final edited form as:

J Pediatr Surg. 2014 January ; 49(1): 154–158. doi:10.1016/j.jpedsurg.2013.09.047.

Inter-rater reliability of surgical reviews for AREN03B2: A COG renal tumor committee study[☆]

Thomas E. Hamilton^{a,*}, Douglas Barnhart^b, Kenneth Gow^c, Fernando Ferrer^d, Jessica Kandel^e, Richard Glick^f, Roshni Dasgupta^g, Arlene Naranjo^h, Ying He^h, Eric Gratiasⁱ, James Geller^j, Elizabeth Mullen^a, and Peter Ehrlich^k

^a Dana Farber Cancer Institute, Boston Children's Hospital, Boston MA

^b Primary Children's Medical Center, Salt Lake City UT

^c Seattle Children's Hospital, Seattle WA

^d Connecticut Children's Hospital, Hartford CT

^e Columbia University Medical Center, New York NY

^f Cohen Children's Medical Center of New York, New York NY

^g Cincinnati Children's Hospital, Cincinnati OH

^h Children's Oncology Group, University of Florida, Gainesville FL

ⁱ T. C. Thompson Children's Hospital, Chattanooga TN

^j Cincinnati Children's Hospital, Cincinnati Medical Center, Cincinnati OH

^k C.S. Mott Children's Hospital, University of Michigan, Ann Arbor MI

Abstract

Purpose—The Children's Oncology Group (COG) renal tumor study (AREN03B2) requires real-time central review of radiology, pathology, and the surgical procedure to determine appropriate risk-based therapy. The purpose of this study was to determine the inter-rater reliability of the surgical reviews.

Methods—Of the first 3200 enrolled AREN03B2 patients, a sample of 100 enriched for blood vessel involvement, spill, rupture, and lymph node involvement was selected for analysis. The surgical assessment was then performed independently by two blinded surgical reviewers and compared to the original assessment, which had been completed by another of the committee surgeons. Variables assessed included surgeon-determined local tumor stage, overall disease stage, type of renal procedure performed, presence of tumor rupture, occurrence of intraoperative tumor spill, blood vessel involvement, presence of peritoneal implants, and interpretation of residual

[☆]This work was supported by NIH grants U10 CA98413 (COG SDC grant) and U10 CA98543 (COG Chair's grant).

© 2014 Elsevier Inc. All rights reserved.

* Corresponding author. Department of Surgery, Boston Children's Hospital, One Boston Medical Center Place, Dowling 2 South, Boston MA 02118. Tel.: +1 617 355 3038. thomas.hamilton@childrens.harvard.edu (T.E. Hamilton).

disease. Inter-rater reliability was measured using the Fleiss' Kappa statistic two-sided hypothesis tests (Kappa, p-value).

Results—Local tumor stage correlated in all 3 reviews except in one case (Kappa = 0.9775, $p < 0.001$). Similarly, overall disease stage had excellent correlation (0.9422, $p < 0.001$). There was strong correlation for type of renal procedure (0.8357, $p < 0.001$), presence of tumor rupture (0.6858, $p < 0.001$), intraoperative tumor spill (0.6493, $p < 0.001$), and blood vessel involvement (0.6470, $p < 0.001$). Variables that had lower correlation were determination of the presence of peritoneal implants (0.2753, $p < 0.001$) and interpretation of residual disease status (0.5310, $p < 0.001$).

Conclusion—The inter-rater reliability of the surgical review is high based on the great consistency in the 3 independent review results. This analysis provides validation and establishes precedent for real-time central surgical review to determine treatment assignment in a risk-based stratagem for multimodal cancer therapy.

Keywords

Wilms Tumor; Quality assurance; Surgery; Outcomes

Multi-modality treatment for a child with Wilms Tumor (WT) is based on risk classification which includes age, tumor weight, histology, stage and molecular characteristics [1]. This requires interpretation of surgical, radiological, pathological and oncological data. Staging for WT is complicated, as both local and disease categories must be established. Briefly, Stage I tumors are completely excised. Tumor was not ruptured or biopsied prior to removal blood vessels of the renal sinus are not involved. Note: for a tumor to qualify for certain therapeutic protocols as stage I, lymph nodes must be examined microscopically and negative for disease. Stage II tumors penetrated the renal capsule but were completely excised. Tumors that extend beyond the kidney as evidenced by: penetration of the renal capsule or extensive invasion of the renal sinus; blood vessels within the nephrectomy specimen outside the renal parenchyma including those of the renal sinus, contain tumor. Note: rupture or spillage confined to the flank, including biopsy is no longer considered stage II and is now considered stage III. Residual non-hematogenous tumor present following surgery and confined to the abdomen is considered stage III. Additional stage III criteria include: positive regional lymph node metastases, penetration to the peritoneal surface or implants, gross or microscopic tumor remains postoperatively, local infiltration into vital structures, tumor spillage before or during surgery, the tumor is treated with preoperative chemotherapy before therapy regardless of type of biopsy, tumor is removed in greater than one piece (e.g. tumor thrombus in renal vein removed separately from nephrectomy specimen).

Stage IV is hematogenous metastases (lung, liver, etc) or lymph nodes outside the abdomen. Stage V is bilateral renal involvement [1]. Prior research has shown a high discordance and protocol violation rates when staging was done by an individual institution compared to a central group of experts [2,3]. Misclassification is likely to adversely impact delivery of appropriate therapy. Under staging a child can result in less therapy and an increased risk of

recurrence. Conversely, over staging can result in treatment of increased intensity with an unnecessary higher risk of both short and long-term toxicity.

Quality assurance (QA) is essential to maintain data reliability, validity, and integrity and is mandated by the National Cancer Institute for any clinical trial [4]. Most of the QA review has been performed retrospectively. However, since 2006 treatment on any therapeutic Children's Oncology Group (COG) renal tumor protocol has required enrollment in the Renal Tumor Classification, Biology, and Banking Study (AREN03B2) [5]. Risk assignment is determined by real-time central review of clinical and molecular factors of known predictive value. Central reviewers include a team of surgeons, pathologists, radiologists and oncologists. By performing the central review in real-time (data are delivered and assimilated immediately as collected, day zero is the date of surgical procedure) each individual child is assured the best risk assignment prior to the initiation of therapy. The radiological, surgical and pathology reviews are all performed within 48 h of the patient registering for the study and prior to therapy beginning. We hypothesized a high level of agreement between surgical reviewers. This study's objective was to determine the accuracy and inter-rater reliability of the surgical reviews on AREN03B2 patients.

1. Patients and methods

AREN03B2 opened in 2006 and as of May 2012 3200 patients had been enrolled. To enroll on AREN03B2 only, materials only need to be submitted by day 30. To receive an initial risk assignment by day 14, required materials including chest and abdominal imaging, pathology slides and institutional pathology reports and operative reports are requested to be submitted by day 7 after surgery. To register to the study a CT or MRI of the chest and abdomen, pathological specimen and operative note must be submitted within 14 days of the original diagnosis for central review. The radiology review and pathology review occur independently prior to the surgical review. The surgical reviewer determines the final local and disease stage and the oncologist then performs the risk assignment. There are six surgical reviewers with 4–25 years' experience with WT surgical quality reviews. The study statistician selected a blinded sample of 100 unilateral renal tumor AREN03B2 patients, enriched for patients with blood vessel involvement, spill, rupture, and lymph node involvement, for analysis. Sample "enrichment" by the statistician selected a higher level of the more difficult variables to interpret than would occur randomly to be a more stringent test of the reviewers' ability to agree. The more frequent cases of stage I disease where all disease is resected are much less challenging. Cases reviewed previously by either of the study surgeons were excluded. Two committee surgeons (each with over twelve years of experience performing surgery reviews) independently re-reviewed every patient and assigned the local and disease stage. This was compared to the initial central surgical review (Fig. 1). The key variables of interest were local and disease stage as they determined therapy. Other variables included type of renal procedure, blood vessel involvement, rupture, spill, presence of peritoneal implants, and residual disease status. Results from committee surgeons, as well as the original (different) reviewer, are summarized with contingency tables. Inter-rater reliability was determined using two methods. The Fleiss' Kappa statistic was calculated for each variable [6,7]. A Kappa score of 1.0 is perfect agreement, 0.8–0.99 almost perfect, 0.6–0.79 substantial, 0.4–0.59 moderate, 0.2–0.39 fair,

0.01–0.19 slight and <0 no agreement [8]. Two-sided hypothesis tests determined whether there was no agreement among the 3 reviewers. A p value of 0.05 was considered significant. All analyses were performed using SAS® version 9.2. P-values were not adjusted for multiple comparisons.

2. Results

In the enriched sample of 100 cases the inter-rater reliability was excellent, with almost perfect agreement. The null hypothesis was also rejected for all variables. Kappa values were almost perfect for: Type of procedure (Table 1) Kappa = 0.8357 ± 0.0394 , p-value: 0.001; local stage (Table 2) 0.9775 ± 0.037 , p-value: 0.001 and disease stage (Table 3) 0.942 ± 0.0279 , p-value: 0.001. Substantial agreement was seen for: tumor rupture 0.658 ± 0.0549 , p-value: 0.001 (Table 4); spill 0.6493 ± 0.0513 , p-value: 0.001 and tumor extension into the blood vessels 0.6470 ± 0.0547 , p-value: 0.001 (Table 5). Moderate agreement was seen with presence of residual disease with Kappa = 0.5310 ± 0.0503 , p-value: 0.001 and a fair Kappa was noted for determining the presence of peritoneal metastasis with a Kappa = 0.2753 ± 0.0493 , p-value: 0.001.

3. Discussion

Wilms tumor treatment has served as a paradigm for multi-modality cancer therapy. A large volume of data from well controlled randomized therapeutic trials has provided the basis for identification of well-defined risk groups enabling targeted therapy which aims to maximize survival and minimize toxicity. The importance of correctly staging the patient (particularly lymph node invasion) has been recognized since the 1980s [9–11]. This information resulted in recommendations to alter the staging system for NWTs-3. A prospective study by Othersen et al. [11] examined the surgeon's impression of lymph nodes and compared it to the pathological review. They found that surgeon's impression alone had only a 57% positive predictive value with a false negative rate of 31% and a false positive rate of 18%. The recommended treatment for a child (COG protocols) is based on the individual child's risk. Currently risk is stratified based on a patient's age, tumor weight, histology, local and disease stage and molecular characteristics of the tumor. This is a complex process and requires expertise to ensure proper targeted therapy. The surgeon (and the initial surgery) provides critical information for determining the local and disease stage. Proper surgical approach and procedure, performance, documentation and understanding of the findings at operation all determine local stage and therefore significantly impact therapy. Shamberger et al., in a retrospective study from NWTs-4, identified an increased risk of local recurrence when 1) surgeons failed to sample lymph nodes and 2) tumor spillage occurred. Incorporating these factors in assignment of local tumor stage was critical in the surgical protocol NWTs V that reduced local tumor recurrence and improved the survival rate of children with WT [12].

Quality assurance of a clinical trial is mandated both internally by the COG, but also externally by the National Cancer Institute. Quality assurance (QA) is essential to maintain study data reliability, validity, and homogeneity. There are several components to a successful QA program [4]. These include insightful study planning, consistent

implementation, interval assessment (with special attention to documentation of data quality and methods), and improvement. Particularly important in monitoring progress is the availability of real-time feedback to alter behavior that may negatively impact outcome.

Advances in molecular studies and clinical trials have enhanced cancer knowledge and therapy. For most cancers, staging and classification are now multifactorial, and often multidisciplinary, and therefore pose a greater chance of misclassification. For example, in a Dutch breast cancer trial study, inter-observer variation in pathological examination of breast carcinomas results in substantial differences in clinicopathological risk assessment and subsequently in adjuvant systemic treatment [8]. One method to reduce variability in risk classification is to centrally review each case by experts prior to treatment. This was done for involved-field radiotherapy (IF-RT) in a Hodgkin's disease study. An individual IFRT prescription was provided for every study patient. Central review was significant for improving correctness of stage definition, allocation to treatment groups, and extent of the IF treatment volume [13].

We have also previously shown that there is discordance in central and institutional review of pathology especially the diagnosis of focal and diffuse anaplasia [3]. To improve pathological staging on NWTs-5 central pathology review was done. NWTs-5 also included a surgical QA program however it was done several months after treatment was started. In the surgical QA analysis from NWTs 5 several key variables that impact therapy noted to be inconsistent and/or misclassified. These included; lymph node assessment, determination of rupture and spill as well as assessment of tumor extent [2]. Based on these data, AREN03B2, the first COG renal Tumor Classification, Biology and Banking Study was built around a plan for real time central risk assignment. This study provides real time review of initial imaging, pathology and surgery combined with loss of heterozygosity (LOH) studies, as well as clinical features (age, presence of underlying congenital anomalies and/or genetic predisposition syndromes) to ensure proper risk classification. All but LOH studies are completed prior to starting therapy. Initially the process included direct feedback to the surgeon if there was a protocol violation however that was discontinued in 2008.

The purpose of this study was to assess the surgical inter-rater reliability for determining local and disease stage. Inter-rater reliability is the degree of agreement among raters [14]. It gives a score of how much consensus there is in the ratings given by judges. It is useful in refining the tools employed by human judges, for example by determining if a particular scale is appropriate for measuring a particular variable. If various raters do not agree, either the scale is defective or the raters need to be re-trained. The surgical reviewers for this study consist of six dedicated pediatric surgical oncologists with between 4 and 25 years of renal tumor experience. Each underwent training prior to becoming surgical reviewer. Each reviewer performs approximately 100 cases a year.

The results of this study are encouraging and demonstrate excellent inter-rater reliability. An enriched sample was used to reflect the complexity of cases and to avoid bias towards the more frequent lower stage cases, which are often easier to interpret. Of critical importance the Kappas for the local and disease staging were almost perfect between the three reviewers. Disease stage is a key component for final risk assignment, and local stage

dictates important therapy variations. In the NWTS-5 surgical QA the area of greatest variability (and misclassification) was tumor spill or rupture. In this study, improvement was shown, with substantial agreement demonstrated between all three reviewers, Kappa = 0.6493 ± 0.0513 .

Although still significant there was less agreement for the presence of peritoneal metastasis. In this category the reviewer had three choices; “yes, no and indeterminate”. In review of the discordance noted, an explanation was found. In the practice of assigning the variable as “no” or “indeterminate” the operative report did not explicitly comment on the presence or absence of peritoneal metastasis. This variation in practice had no impact on final stage, as neither would “upstage” patient. Importantly, there was excellent consensus for the “yes” classification.

A moderate Kappa (0.5310 ± 0.0503 , p-value: 0.001) was noted between the reviewers classifying the presence of residual disease. There appears to be some variability in how the presence of nodal disease and microscopic residual was recorded. Some reviewers checked both microscopic residual and lymph node (LN) positivity when only the LN was positive while others checked the LN box only. Either practice resulted in a consistent staging of the patient as stage III. Alterations to the surgical check list, expanding and clarifying this section have occurred as a result of this audit.

A limitation of this study was the focus on the surgical review. Although the surgical review is critical for risk assignment, so are the radiological and pathological reviews. It is possible that a misclassification by one of the other central reviewers would influence the surgical review. In addition we did not compare the institutional staging to the central review staging to look at discordance rates. Finally, all three reviewers could have agreed but still be wrong. The strong statistical significance of the two-sided test makes this less likely. A future study is planned to compare institutional risk assignment to central risk assignment.

The ultimate contribution of real-time surgical review is the assignment of appropriate local and overall disease stage to determine therapy and in this, the process is excellent. Inter-rater reliability for operative findings is lower than expected and highlights potential areas for improvement. This analysis provides validation and establishes precedent for real-time central surgical review as a critical element of a multidisciplinary review utilized to determine treatment assignment in a risk-based stratagem for multimodal cancer therapy.

Discussion

Discussant: Dr. Rebecka Meyers (Salt Lake City, UT)

I loved that presentation because this is an important topic to me and having spent a lot of time doing reviews for liver tumors I am curious about the quality of the material you are reviewing. Are you reviewing operative reports, I assume?

Response: Thomas Hamilton

Exactly.

Rebecka Meyers

And how about radiographic results and I don't know if the surgeons review the pathology as well, but we are often frustrated by the quality of that data. You can't say if there are peritoneal implants if the surgeon in their report doesn't say it, and I am wondering how much of the lack of consensus on things like peritoneal implants is in the report itself and not the reviewer's.

Response: Dr. Thomas Hamilton

I think that is a very, very good point. That is exactly what we find is that if the answer was yes there is excellent agreement, like the surgeon described, there is an implant there, we stopped, we just biopsied, and we got out. But if it wasn't specifically stated that they didn't look for implants, some review may say, ah, it didn't sound like there or some may say no. That is part of us being more precise in our definitions but it's very important for us to have the surgeons do as detailed and as thorough an operative note as they can to help us to assign the appropriate risk for each child.

Discussant: Andrea Hayes-Jordan (UT MD Anderson)

I also thought that was excellent and this is also the way we want to go as far as being able to have central review, but I wondered if you could help us and the audience understand some of the definitions. The tumor spill rupture, sort of piggy-backing on Rebecka's question, was that part of the reason — even though there are only one or two that disagreed, was it the definition of tumor rupture spill that sort of tripped up a couple people and what would the other definitions that may or may not need some clarification?

Response: Thomas Hamilton

I think that is exactly what has happened. If you peel the tumor, for example, out of the vein and you think that you got it all or not and then you go look at the pathology report did they really see it all or not? Some of that is interpretation and one reviewer may say indeterminate, one may say positive, but the most examples are a biopsy before nephrectomy. Those are the type of spills that we absolutely have to catch and for all of those there is excellent agreement.

Discussant: Daniel Von Allmen (Cincinnati, OH)

Did you make any effort or has there been any effort to try to correlate what was reported on the radiology reports or in the radiology studies with what was said by the surgeon? We looked at this in neuroblastoma and there was only a 66% concordance rate between what the surgeon reported and what the radiologist showed.

Response: Dr. Thomas Hamilton

In terms of looking at the preoperative assessment?

Dr. Daniel Von Allmen

Things like peritoneal implants and these things that there is potentially poor reliability.

Thomas Hamilton

I think this has been published in a paper recently. Peter Ehrlich may be an author on that paper. Do you want to answer that?

Response: Dr. Peter Ehrlich (Ann Arbor, MI)

We did a study that has been published where the radiologist looked at their predictive power of determining both preoperative rupture and peritoneal metastases and the CAT scan. There was no correlation between that. If there is a rupture there are signs that go well but they can't predict whether there has been a rupture with any degree of accuracy and same with nodal status too.

Unidentified speaker (moderator)

Dr. Hamilton, let me ask you a general question. You've demonstrated or the COG has once again demonstrated the value of the program and it has really become a de facto standard of care. My question is to what extent do we have children with renal tumors cared for outside of that mechanism? Do you know?

Thomas Hamilton

I think it may be larger than we want but I think a very large number are involved in COG studies. There are data bases which would show us that some are not being enrolled in the COG but I think all the ones that are certainly are getting this appropriate review and receiving appropriate care.

References

1. Dome, JS.; Perlman, EJ.; Ritchey, ML., et al. Renal tumors.. In: Pizzo, PA.; Poplack, DG., editors. Principles and practice of pediatric oncology. Lippincott Williams and Wilkins; Philadelphia: 2006. p. 905-32.
2. Ehrlich PF, Ritchey ML, Hamilton TE, et al. Quality assessment for Wilms' tumor: a report from the National Wilms' Tumor Study-5. *J Pediatr Surg.* 2005; 40:208–12. [PubMed: 15868587]
3. Perlman EJ. Pediatric renal tumors: practical updates for the pathologist. *Pediatr Dev Pathol.* 2005; 8:32–338.
4. Riddick L, Simbanin C. One fish, two fish, we qc fish: controlling data quality among more than 50 organizations over a four year period. *Qual Assur.* 2001; 9(3):209–16. [PubMed: 12553084]
5. Dome, JS.; Grundy, PE.; Ehrlich, PF., et al. Renal tumor classification, biology, and banking study (AREN03B2). Childrens Oncology Group; 2013. Available from: www.childrensoncologygroup.org
6. Chen, B.; Zaebst, D.; Seel, L. A macro to calculate kappa statistics for categorizations by multiple raters. 2013. <http://www2.sas.com/proceedings/sugi30/155-30.pdf> Available from: <http://www2.sas.com/proceedings/sugi30/155-30.pdf>
7. Fleiss, JL. Statistical methods for rates and proportions. John Wiley & Sons, Inc; New York: 1981.
8. Bueno-de-Mesquita JM, Nuyten DS, Wesseling J, et al. The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment. *Ann Oncol.* 2010; 21:40–7. [PubMed: 19622588]
9. Farewell VT, D'Angio GJ, Breslow N, et al. Retrospective validation of a new staging system for Wilms' tumor. *Cancer Clinical Trials.* 1981; 4(2):167–71. [PubMed: 6265112]
10. Jereb B, Tournade MF, Lemerle J, et al. Lymph node invasion and prognosis in nephroblastoma. *Cancer.* 1980; 45(7):1632–6. [PubMed: 6245781]

11. Othersen HB, Delorimier A, Hrabovsky E, et al. Surgical evaluation of lymph node metastases in Wilms tumor. *J Pediatr Surg.* 1990; 25:330–1. [PubMed: 2156040]
12. Shamberger RC, Guthrie KA, Ritchey ML, et al. Surgery related factors and local recurrence of Wilms tumor in the National Wilms Tumor Study 4. *Ann Surg.* 1999; 229:292–7. [PubMed: 10024113]
13. Eich HT, Staar S, Grossman A. Centralized radiation oncologic review of cross-sectional imaging of Hodgkin's disease leads to significant changes in required involved field-results of a quality assurance program of the German Hodgkin Study Group. *Int J Radiat Oncol Biol Phys.* 2004; 58:1121–7. [PubMed: 15001253]
14. Vierra AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med.* 2005; 37:360–3. [PubMed: 15883903]

Type of Renal procedure	Enucleation Nephrectomy complete Nephrectomy partial Renal biopsy only Renal biopsy followed by resection Wedge resection Other
Type of Renal biopsy	Open Percutaneous
Did Major Blood Vessels outside the renal parenchyma contain tumor	Yes No Indeterminate
Did Tumor Rupture	Yes No Indeterminate
When did the tumor rupture	Preoperative Intra-operatively Indeterminate
Was Tumor cut across or spilled during surgery	Yes No Indeterminate
Were there peritoneal implants	Yes No Indeterminate
Patients disease Status	No evidence of disease Microscopic Residual Gross Residual
Local Tumor Stage	I II III
Disease Stage	I II III IV V

Fig. 1.
Surgical review unilateral renal tumor.

Table 1

Distribution of ratings in evaluating the variable 'type of renal procedure' and Fleiss' kappa.

Type of Renal Procedure	Review			Kappa (SE)	P-value
	Reviewer 1	Reviewer 2	Original		
Enucleation (Single)	0	0	0	0.8357	< 0.001
Enucleation (Multiple)	0	0	0	(0.0394)	
Complete Nephrectomy	94	94	93		
Partial Nephrectomy	1	2	3		
Renal Biopsy Only	1	1	1		
Renal Biopsy Followed by Complete Resection at Same Procedure	3	3	2		
Wedge Resection (Single)	0	0	0		
Wedge Resection (Multiple)	1	0	0		
Other (Specify)	0	0	1		

Table 2

Distribution of ratings in evaluating the variable ‘surgeon-determined local tumor stage’ and Fleiss‘ kappa.

Surgeon- Determined Local Tumor Stage	Review			Kappa (SE)	P-value
	Reviewer 1 ^a	Reviewer 2 ^a	Original ^b		
I	20	20	20	0.9775 (0.037)	<0.001
II	24	24	25		
III	47	47	46		
Unknown/NA	0	0	0		

^a 9 patients with renal cell carcinoma were not reviewed.

^b 8 patients with renal cell carcinoma were not reviewed. 1 patient without renal cell carcinoma was missing review.

Table 3

Distribution of ratings in evaluating the variable 'disease stage' and Fleiss' kappa.

Disease Stage	Review			Kappa(SE)	P-value
	Reviewer 1 ^a	Reviewer 2 ^a	Original ^b		
I	20	20	20	0.9422 (0.0279)	<0.001
II	20	19	20		
III	34	32	29		
IV	17	20	22		

^a 9 patients with renal cell carcinoma were not reviewed.

Table 4

Distribution of ratings in evaluating the variable 'tumor rupture' and Fleiss' kappa.

	Reviewer 1	Reviewer 2	Original	
Indeterminate	0	0	4	0.6858 (0.0549) < 0.001
No	82	68	72	
Yes	18	32	24	

Table 5

Distribution of ratings in evaluating the variable 'blood vessel involvement' and Fleiss' kappa.

Blood Vessel Involvement	Review			Kappa (SE)	P-value
	Reviewer 1	Reviewer 2 ^a	Original		
Indeterminate	1	0	4	0.6470 (0.0547)	< 0.001
No	93	91	89		
Yes	6	7	7		

^a2 patients were missing reviews.