Journal of Pediatric Surgery 49 (2014) 154–158

Contents lists available at ScienceDirect

Journal of Pediatric Surgery

journal homepage: www.elsevier.com/locate/jpedsurg

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 Gratias [i], James Geller [j], Elizabeth Mullen [a], 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] Cinncinnati 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

ARTICLE INFO

Article history:
Received 23 September 2013
Accepted 30 September 2013

Key words:
Wilms Tumor
Quality assurance
Surgery
Outcomes

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 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.

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

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

* 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.
E-mail address: thomas.hamilton@childrens.harvard.edu (T.E. Hamilton).

0022-3468/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jpedsurg.2013.09.047

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

ARENO3B2 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