Correlation of surgeon radiology assessment with laparoscopic disease site scoring in patients with advanced ovarian cancer


HIGHLIGHTS
- Surgeon radiology review did not correlate highly with diagnostic laparoscopy findings.
- Correlation between radiology review and laparoscopy did not vary by rank or experience.
- Surgeon radiology review best assessed stomach involvement and was worst for liver involvement.

ABSTRACT
Background Radiographic triage measures in patients with new advanced ovarian cancer have yielded inconsistent results.

Objective To determine the correlation between surgeon radiology assessment and laparoscopic scoring by disease sites in patients with newly diagnosed advanced stage ovarian cancer.

Methods Fourteen gynecologic oncology surgeons from a single institution performed a blinded review of preoperative contrast-enhanced CT imaging from patients with advanced stage ovarian cancer. Each of the patients had also undergone laparoscopic scoring assessment, between April 2013 and December 2017, to determine primary resectability using the validated Fagotti scoring method, and assigned a predictive index value score. Surgeons were asked to provide expected predictive index value scores based on their blinded review of the antecedent CT imaging. Linear mixed models were conducted to calculate the correlation between radiologic and laparoscopic score for surgeons individually, and as a group. Once the model was fit, the inter-class correlation and 95% CI were calculated.

Results Radiology review was performed on 20 patients with advanced stage ovarian cancer who underwent laparoscopic scoring assessment. Surgeon faculty rank included assistant professor (n=5), associate professor (p=4), and professor (n=5). The kappa inter-rater agreement was −0.017 (95% CI −0.023 to −0.005), indicating low inter-rater agreement between radiology review and actual laparoscopic score. The inter-class correlation in this model was 0.06 (0.02–0.21), indicating that surgeons do not score the same across all the images. When using a clinical cut-off point for the predictive index value of 0.8, the probability of agreement between radiology and actual laparoscopic score was 0.56 (95% CI 0.49 to 0.73). Examination of disease site sub-scales showed that the probability of agreement was as follows: peritoneum 0.57 (95% CI 0.51 to 0.62), diaphragm 0.54 (95% CI 0.48 to 0.60), mesentry 0.51 (95% CI 0.45 to 0.57), omentum 0.61 (95% CI 0.55 to 0.67), bowel 0.54 (95% CI 0.44 to 0.64), stomach 0.71 (95% CI 0.65 to 0.76), and liver 0.36 (95% CI 0.31 to 0.42). The number of laparoscopic scoring cases, tumor reductive surgery cases, or faculty rank was not significantly associated with overall or sub-scale agreement.

Conclusions Surgeon radiology review did not correlate highly with actual laparoscopic scoring assessment findings in patients with advanced stage ovarian cancer. Our study highlights the limited accuracy of surgeon radiographic assessment to determine resectability.

INTRODUCTION
The approaches by which gynecologic oncology surgeons triage patients with advanced ovarian cancer to either primary surgery or neoadjuvant chemotherapy are diverse, inconsistent, and poorly reproducible. Pre-operative CT, serum CA-125, clinical examination, and patient factors are among the most commonly used non-invasive modalities to triage patients but have been met with variable success and it has been difficult to standardize across surgical practices. Accurate assessment of tumor burden and pattern of spread at initial diagnosis is paramount in order to determine ability to achieve optimal tumor cytoreduction and avoid a futile laparotomy.

A laparoscopic-based scoring assessment to determine primary resectability in patients with advanced stage ovarian cancer has been previously reported in the literature. Laparoscopic scoring has demonstrated value in achieving an overall high positive predictive value for sub-optimal primary tumor cytoreduction, thereby reducing rates of futile laparotomy at cytoreductive surgery. We have previously reported that incorporating the laparoscopic scoring algorithm validated by Fagotti et al into our standard triage of patients with newly diagnosed advanced stage ovarian cancer improved our complete gross resection (R0) rates at both primary and interval resection.
tumor reductive surgeries. Those patients undergoing R0 resection at primary surgery experienced the greatest survival benefit, thus, selecting patients who would achieve the most benefit from an aggressive surgical approach.\(^\text{14}\)

Some criticism of laparoscopic assessment of patients with new advanced ovarian cancer has been the need for an additional surgical procedure to obtain laparoscopic data, which could lead to delay in primary therapy, or complication risks from the laparoscopic procedure. Thus, a question is raised as to whether a thorough and systematic surgeon review of pre-operative imaging can replace laparoscopic scoring assessment. The objective of this study was to determine the correlation between surgeon radiology assessment and laparoscopic scoring by disease sites in patients with newly diagnosed advanced stage ovarian cancer.

**METHODS**

This retrospective imaging review protocol was approved by the University of Texas MD Anderson Cancer Center institutional review board (PA 18-0259). Fourteen gynecologic oncology surgeons from a single institution performed a blinded review of contrast-enhanced CT imaging from 20 patients with advanced stage ovarian cancer. All patients had previously undergone laparoscopic scoring assessment to determine primary resectability at tumor reductive surgery.

### Table 1

<table>
<thead>
<tr>
<th>Patient characteristics</th>
<th>n=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at diagnosis (years), median (range)</td>
<td>65.5 (36–80)</td>
</tr>
<tr>
<td>BMI (kg/m(^2)), median (range)</td>
<td>28.4 (21.1–41)</td>
</tr>
<tr>
<td>ECOG performance status</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>8 (40%)</td>
</tr>
<tr>
<td>1</td>
<td>9 (45%)</td>
</tr>
<tr>
<td>2</td>
<td>1 (5%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>2 (10%)</td>
</tr>
<tr>
<td>Baseline CA-125 (U/mL), median (range)</td>
<td>675 (61–12 472)</td>
</tr>
<tr>
<td>Stage at diagnosis</td>
<td></td>
</tr>
<tr>
<td>IIIC</td>
<td>17 (85%)</td>
</tr>
<tr>
<td>IVA</td>
<td>2 (10%)</td>
</tr>
<tr>
<td>IVB</td>
<td>1 (5%)</td>
</tr>
<tr>
<td>Laparoscopic assessment score (median)</td>
<td>9 (0–14)</td>
</tr>
<tr>
<td>Type of TRS</td>
<td></td>
</tr>
<tr>
<td>Primary</td>
<td>5 (26%)</td>
</tr>
<tr>
<td>Interval</td>
<td>14 (74%)</td>
</tr>
<tr>
<td>Unknown(^*)</td>
<td>1</td>
</tr>
<tr>
<td>Remaining tumor size at TRS</td>
<td></td>
</tr>
<tr>
<td>R0</td>
<td>15 (79%)</td>
</tr>
<tr>
<td>&lt;1 cm</td>
<td>1 (5%)</td>
</tr>
<tr>
<td>≥1 cm</td>
<td>3 (16%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1</td>
</tr>
</tbody>
</table>

**RESULTS**

Radiology review was performed on 20 patients with advanced stage ovarian cancer who underwent laparoscopic scoring assessment. Clinical and demographic data from patients are presented in Table 1. Median age was 65.5 years (range 36–80), median body mass index was 28.4 kg/m\(^2\) (range 21.1–41), and median baseline CA-125 level was 675 U/mL (range 61–12 472). The majority of patients had stage IIIC disease (85%) and median laparoscopic predictive index value score was 9 (range 0–14). Surgeon faculty academic rank included assistant professor (n=5), associate professor (n=4), and professor (n=5). Median surgeon experience during the study period with laparoscopic scoring surgery was 13 cases (range 1–28) and tumor reductive surgery was 22.5 cases using the Fagotti validated scoring algorithm\(^6\) (Online supplemental table 1) between April 2013 and December 2017. These 20 patients were randomly selected from our institutional ovarian cancer database (PA 16–1010) with predictive index value scores ranging from 0 to 14 (Online supplemental table 2). The patients with predictive index value scores <8 were offered primary surgery and those with scores ≥8 received neoadjuvant chemotherapy. Surgeons recorded predictive index value scores based on their blinded review of CT imaging and reports in the anatomic locations evaluated as a part of the Fagotti validated scoring system. Patients were excluded if CT imaging was non-contrast-enhanced or was performed outside of MD Anderson to ensure the quality of the imaging was consistent across all patient subjects.

Summary statistics of the radiologic and laparoscopic scores were calculated by the surgeon. Box plots were created to depict the scores graphically. The percent agreement (yes/no based on clinical cut-off point of 8) was calculated by surgeon and across all surgeons. The kappa inter-rater agreement statistic and corresponding 95% CI were calculated for the radiology scores. Linear mixed models were conducted to correlate the correlation between radiology and laparoscopic score for each surgeon and as a group. We regressed laparoscopic score on radiology score. A random intercept was included and our subject effect was the clinician. Once the model was fit, the inter-class correlation and 95% CI were calculated. We also fit a model with the image as our subject effect to calculate within-image correlations. We then categorized each score as agreement versus non-agreement based on the pre-defined cut-off point of 8 for treatment decision-making. Similar generalized linear mixed models were then conducted with agreement (yes/no) as outcome and clinical and demographic variables of interest as covariates to assess if certain covariates are associated with scoring agreement. Our sample size was chosen to give adequate precision for our estimates of inter-class correlation. When the sample size is 280, a two-sided 95% CI computed using the large-sample normal approximation for an inter-class correlation based on 14 raters will extend about 0.037 (0.042) from the observed inter-class correlation when the expected intra-class correlation is 0.7 (0.3).

All statistical analyses were performed using Stata/MP v15.0 (College Station, Texas, USA). Study data were collected and managed using REDCap (Research Electronic Data Capture) electronic data capture tools hosted at MD Anderson.\(^{15}\)
(range 2–48). Following laparoscopic scoring assessment, five patients (26%) underwent primary surgery and 15 patients (74%) received neoadjuvant chemotherapy. Of those who received neoadjuvant chemotherapy, 14 patients did undergo interval tumor reduc-
tive surgery and one patient was lost to follow-up following the
laparoscopic scoring assessment. The majority of patients under-
went optimal tumor cytoreduction to no gross residual disease (R0, 79%), compared with optimal <1 cm (5%) and sub-optimal ≥1 cm (16%).

Summary statistics for each surgeon’s scores based on agree-
ment of the laparoscopic assessment predictive index value score of
<8 or ≥8 are listed in Online supplemental table 3). Average agree-
ment among all surgeons for all cases was 55.7%. Agreement by
experience level was 51% for assistant professor, 70% for associate
professor, and 56% for professor level faculty surgeons. Figures 1
and 2 depict agreement and summary scores. The kappa inter-rater
agreement was −0.017 (95% CI −0.023 to −0.005), indicating low
inter-rater agreement between surgeon radiology review and actual
laparoscopic score (Figure 2). To account for the correlations, within
images and within surgeons, linear mixed models were conducted
to calculate the inter-class correlation. When we treated image as
the subject effect, our inter-class correlation was 0.22 (95% CI 0.10
to 0.40). An inter-class correlation of 0 would imply total agreement
between surgeons about the images, thus our inter-class correla-
tion indicates mild agreement between surgeons on scoring each
image. When treating each surgeon as the subject effect, the inter-
class correlation in this model was 0.06 (95% CI 0.02 to 0.21).
An inter-class correlation of 1 would indicate the surgeon tended
to score every image the same (ie, particularly high or low all the
time). Thus our inter-class correlation in this model indicated that
surgeons do not score the same across all the images.

Our next agreement analyzed was defined based on the clin-
ical cut-off predictive index value score of 8. Then generalized
linear mixed models were conducted to assess whether surgeon
academic rank, number of laparoscopic scoring procedures
performed by each surgeon, and number of tumor reductive surgery
cases performed by each surgeon were associated with agreement
(Table 2). When using a clinical cut-off predictive index value of 8,
the probability of agreement between radiology and actual laparo-
scopic score was 0.56 (95% CI 0.49 to 0.73). The number of laparo-
scopic scoring cases, tumor reductive surgery cases, and surgeon
title was not significantly associated with agreement.

A similar analysis was performed to assess agreement within
each disease site sub-scale (Table 3) within the laparoscopic
scoring algorithm. Online supplemental figure 1 includes a scatter
plot of radiology and laparoscopic scores by each disease site
sub-scale. Online supplemental table 4a includes the calculations
of probability of agreement from the general linear models. When
evaluating the disease site sub-scales, the probability of agreement
was as follows: peritoneum 0.57 (95% CI 0.51 to 0.62), diaphragm
0.54 (95% CI 0.48 to 0.60), mesentery 0.51 (95% CI 0.45 to 0.57),
omentum 0.61 (95% CI 0.55 to 0.67), bowel 0.54 (95% CI 0.44 to
0.64), stomach 0.71 (95% CI 0.65 to 0.76), and liver 0.36 (95% CI
0.31 to 0.42). The number of laparoscopic scoring cases, tumor
reductive surgery cases, or surgeon academic rank was not signifi-
cantly associated with overall or sub-scale agreement. Online
supplemental table 4b includes the kappa statistics for the inter-
rater agreement on the radiology sub-scale scores.

An analysis by surgeon was performed comparing the surgeon
radiology score and projected triage based on this score (primary
surgery or neoadjuvant chemotherapy) with the actual laparoscopic
score and actual laparoscopic triage results (Online supplemental

Figure 1  Scatter plot of laparoscopy compared with surgeon radiology scores. Red dots=agreement between laparoscopy score and surgeon radiology score; blue dot=no agreement between laparoscopy score and surgeon radiology score; Lap score=actual predictive index value score at the time of laparoscopy; Rad score=predictive index value score by faculty surgeon review of CT imaging.
Median percentage agreement in triage results (primary surgery or neoadjuvant chemotherapy) was 57.5% and median projected futile laparotomy rate was 42.5%. Two patients did receive an actual laparoscopic predictive index value score of 6, but based on disease burden assessment at the discretion of the surgeon, were triaged to neoadjuvant chemotherapy.

DISCUSSION

Surgeon radiology review did not correlate well with actual laparoscopic scoring assessment findings in patients with advanced stage ovarian cancer, regardless of surgeon experience level. By disease site sub-scale, the best agreement was noted when evaluating stomach involvement, and the worst with liver involvement. Our study highlights the limited accuracy of surgeon radiographic assessment to determine resectability.

Previous studies have shown poor correlation between CT radiographic predictors with findings at the time of tumor reductive surgery and ability to predict sub-optimal resection.¹ This study, and others evaluating structured radiology reporting of disease burden by anatomic location by experienced radiologists combined with patient clinical factors, have not been reproducible to predict resection results at the time of tumor cytoreduction in advanced ovarian cancer.²⁻⁴ Implementation of laparoscopy-based scoring systems has overcome the limitations of pre-operative radiology assessment tools,¹⁶‑¹⁷ which has led to reduction in unnecessary laparotomy procedures in unresectable cases.⁷⁻⁹ However, widespread implementation of laparoscopic assessment for primary resectability in patients with new advanced ovarian cancer cases been met with resistance and controversy.

Our study adds to the limited literature comparing laparoscopic scoring assessment findings with detailed radiographic review data.
Original research

Table 3  Agreement of scores by disease site sub-scale

<table>
<thead>
<tr>
<th>Lap score</th>
<th>Radiology score</th>
<th>N</th>
<th>%</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peritoneal carcinomatosis</td>
<td>0</td>
<td>54</td>
<td>51.92</td>
<td>71</td>
<td>40.57</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>50</td>
<td>48.08</td>
<td>104</td>
<td>59.43</td>
</tr>
<tr>
<td>Diaphragmatic disease</td>
<td>0</td>
<td>69</td>
<td>42.33</td>
<td>29</td>
<td>28.16</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>94</td>
<td>57.67</td>
<td>74</td>
<td>71.84</td>
</tr>
<tr>
<td>Mesenteric disease</td>
<td>0</td>
<td>57</td>
<td>41.01</td>
<td>55</td>
<td>39.01</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>82</td>
<td>58.99</td>
<td>86</td>
<td>60.99</td>
</tr>
<tr>
<td>Omental site of disease</td>
<td>0</td>
<td>69</td>
<td>64.49</td>
<td>71</td>
<td>41.04</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>38</td>
<td>35.51</td>
<td>102</td>
<td>58.96</td>
</tr>
<tr>
<td>Bowel infiltration</td>
<td>0</td>
<td>24</td>
<td>16.00</td>
<td>4</td>
<td>3.08</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>126</td>
<td>84.00</td>
<td>126</td>
<td>96.92</td>
</tr>
<tr>
<td>Stomach infiltration</td>
<td>0</td>
<td>178</td>
<td>75.11</td>
<td>18</td>
<td>62.07</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>59</td>
<td>24.89</td>
<td>11</td>
<td>37.93</td>
</tr>
<tr>
<td>Liver surface</td>
<td>0</td>
<td>67</td>
<td>28.63</td>
<td>3</td>
<td>9.38</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>167</td>
<td>71.37</td>
<td>29</td>
<td>90.63</td>
</tr>
</tbody>
</table>

by surgeons. There have been recent reports from Ahmed et al comparing the accuracy of CT and laparoscopy in predicting the peritoneal carcinomatosis index score. CT and laparoscopy sensitivity were 94.9%, 98.3%, specificity 86.7%, 80.4%, positive predictive value 97.9%, 96.8%, negative predictive value 72.2%, 88.8%, and accuracy 93.8%, 95.7%, respectively. However, CT diagnostic performance was less accurate than laparoscopy in pelvic and small intestinal regions. The authors concluded that both CT and laparoscopy seem to be effective tools for assessment of peritoneal carcinomatosis using the peritoneal carcinomatosis index score.18 19 No studies have compared CT diagnostic performance with laparoscopic scoring using the validated Fagotti scoring algorithm for surgical resectability. There are inherent differences in the peritoneal carcinomatosis index, which analyzes tumor size by anatomic location.20 Many previous studies have used the peritoneal carcinomatosis index to correlate surgical resectability and its impact on survival in advanced ovarian cancer.21–24 Although the peritoneal carcinomatosis index has been shown to be a validated radiologic assessment tool in advanced cancers with peritoneal spread, studies have suggested its poor usefulness as a triage test to reliably identify patients with advanced ovarian cancer who are likely to have complete cytoreductive surgery.24 A recent study suggested that selected peritoneal carcinomatosis index regions, such as the small intestine with adjacent mesentery and the hepato-duodenal ligament, are more predictive of complete resection and survival than the entire peritoneal carcinomatosis index.24 This is similar to evaluating the disease distribution in distinct anatomic areas by the Fagotti laparoscopic assessment model. However, there has been no prospective validated correlation between the peritoneal carcinomatosis index and laparoscopic assessment in advanced ovarian cancer.

The goal of any triage modality in a patient with newly diagnosed advanced ovarian cancer should be the reproducibility of the method and generalizability to practicing gynecologic oncology surgeons. Thus, the utility of structured radiology reports by specialized gynecologic oncology radiologists may not be practical in a setting that does not have access to this resource. The ability for a gynecologic oncology surgeon to interpret high-quality, contrast-enhanced CT imaging in order to make their own decisions is important. Our study, which was conducted in large tertiary cancer center, showed that our surgeons’ radiology review did not correlate highly with actual laparoscopic scoring assessment findings in patients with advanced stage ovarian cancer, and thus could have altered surgical decision-making if radiology review alone had been used in these cases. Our data support the high likelihood of futile laparotomy or of a patient being inappropriately explored if radiology review alone had been used.

The strengths of our correlative study are its blinded nature and its evaluation of a large number of surgeons in a tertiary cancer center with a high volume of patients with advanced ovarian cancer. Additional strengths include evaluating surgeons of different experience and rank in their ability to predict laparoscopic score based on CT review, which leads to generalizability of the study findings to gynecologic oncology surgeons in the community. We also chose patients with varying degrees of tumor burden in order to evaluate the correlation of CT review with laparoscopy in all disease burden types. Our study is one of the first to evaluate the correlation of CT review with laparoscopy findings by disease site sub-scale based on the Fagotti algorithm. The weaknesses of the study include the small number of cases selected to review, which could lead to inherent selection bias and possible inaccuracy of the results if more cases were evaluated prospectively. We also elected not to compare the performance of a radiologist interpreting CT imaging with laparoscopic findings as laparoscopic assessment is based on surgical resectability, which is difficult to determine from a radiology perspective. Based on our results, we have included a pre-laparoscopy surgeon radiology review and scoring to our quality improvement process at our institution to provide additional prospective cases to report in the future.

In conclusion, our results suggest that the validated laparoscopic scoring algorithm should be considered the ‘gold standard’ assessment tool for primary resectability in patients with new advanced ovarian cancer. Further prospective studies are needed to determine the utility of novel imaging modalities which may enhance surgeon ability to predict surgical resectability, eliminating the need for laparoscopic scoring assessment.

Author affiliations
1Department of Gynecologic Oncology and Reproductive Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
2Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
3Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Washington, Seattle, Washington, USA

REFERENCES


