Interobserver and intraobserver variability of RECIST assessment in ovarian cancer

Michael Krasovitsky,1,2 Yeh Chen Lee,1,2,3 Hao-Wen Sim,2,3 Tanya Chawla,4,5 Helen Moore,6 Daniel Moses,7 Luke Baker,6 Catherine Mandel,7,8 Ania Kielar,9,10 Angus Hartery,11 Martin O’Malley,4,5 Michael Friedlander,1,2 Amit M Oza,5,12 Lisa Wang,5,12 Stephanie Lheureux,5,12 Michelle Wilson13

HIGHLIGHTS
• Evaluation of ovarian cancer using RECIST V.1.1 may be subject to interobserver and intraobserver variability.
• Reproducibility of lesion selection was high overall, but marginal for peritoneal lesions.
• Measurement of malignant lesions showed very good concordance overall, but moderate concordance for nodal lesions.

ABSTRACT
Objectives Measurement of Response Evaluation Criteria In Solid Tumors (RECIST) relies on reproducible unidimensional tumor measurements. This study assessed interobserver and intraobserver variability of target lesion selection and measurement, according to RECIST version 1.1 in patients with ovarian cancer.

Methods Eight international radiologists independently viewed 47 images demonstrating malignant lesions in patients with ovarian cancer and selected and measured lesions according to RECIST V.1.1 criteria. Thirteen images were viewed twice. Interobserver variability of selection and measurement were calculated for all images. Intraobserver variability of selection and measurement were calculated for images viewed twice. Lesions were classified according to their anatomical site as pulmonary, hepatic, pelvic mass, peritoneal, lymph nodal, or other. Lesion selection variability was assessed by calculating the reproducibility rate. Lesion measurement variability was assessed with the intra-class correlation coefficient.

Results From 47 images, 82 distinct lesions were identified. For lesion selection, the interobserver and intraobserver reproducibility rates were high, at 0.91 and 0.93, respectively. Interobserver selection reproducibility was highest (reproducibility rate 1) for pelvic mass and other lesions. Intraobserver selection reproducibility was highest (reproducibility rate 1) for pelvic mass, hepatic, nodal, and other lesions. Selection reproducibility was lowest for peritoneal lesions (intraobserver reproducibility rate 0.76 and intraobserver reproducibility rate 0.69). For lesion measurement, the overall interobserver and intraobserver intraclass correlation coefficients showed very good concordance of 0.84 and 0.94, respectively. Interobserver intraclass correlation coefficient showed very good concordance for hepatic, pulmonary, peritoneal, and other lesions, and ranged from 0.84 to 0.97, but only moderate concordance for lymph node lesions (0.58). Intraobserver intraclass correlation coefficient showed very good concordance for all lesions, ranging from 0.82 to 0.99. In total, 85% of total measurement variability resulted from interobserver measurement difference.

Conclusions Our study showed that while selection and measurement concordance were high, there was significant interobserver and intraobserver variability.

Most resulted from interobserver variability. Compared with other lesions, peritoneal lesions had the lowest selection reproducibility, and lymph node lesions had the lowest measurement concordance. These factors need consideration to improve response assessment, especially as progression free survival remains the most common endpoint in phase III trials.

INTRODUCTION
The World Health Organization (WHO) established the first internationally recognized criteria to assess the radiological burden of cancer, using a bidimensional model to provide a summation of tumor burden in 1979.1,2 These criteria were updated in 2000 and again in 2009, to reflect modern imaging, with the establishment of the Response Evaluation Criteria In Solid Tumors (RECIST version (V) 1.0 and 1.1, respectively).3,4 In comparison with the WHO criteria, RECIST used unidimensional radiographic measures, and defined criteria indicative of a complete response, partial response, stable disease, and progressive disease.4 These criteria were designed to provide an objective, uniform, and reliable method to quantify tumor burden, and used unidirectional changes in size to indicate response and have become a core component of cancer clinical trials.

Despite its centrality in clinical trials, the assumption that the unidimensional tumor measurements central to RECIST assessments can be reliably performed by different readers, or by the same reader at different times, has been challenged.3–7 Erasmus et al demonstrated that intraobserver and interobserver misclassification of unchanged lesions led to the incorrect diagnosis of progressive disease in 9.5% and 29.8% of patients respectively.5 Other studies have confirmed the presence of both interobserver and intraobserver variability in measurement for RECIST in non-gynecological cancers.6,7 Ovarian cancer is a heterogenous disease typically diagnosed at an advanced stage. It is associated with...
significant morbidity and high mortality. The clinical course is variable, but in advanced disease, it is typified by multiple cycles of treatment, high rates of recurrence, and the need for retreatment. Each phase may not be associated with radiological progression despite evidence of CA125 progression, and even in the presence of radiological abnormalities, RECIST V.1.1 criteria may not be met. It has been shown that over 50% of patients with recurrent ovarian cancer do not have RECIST defined measurable disease. Additionally, the pattern of spread of ovarian cancer is often locoregional, with a predilection for omental, peritoneal, and nodal metastases and may be associated with ascites and pleural effusions. These may not be measurable by RECIST V.1.1 criteria, even in the presence of unequivocal disease recurrence.

Primary endpoints, including response rate and progression free survival, are used routinely in phase II and III trials. Given the pivotal role played by these trials in regulatory approvals and clinical care, assessing the reliability of RECIST in ovarian cancer is warranted. To better delineate the complex relationship between ovarian cancer and RECIST, we sought to characterize the reproducibility and variability of both lesion selection and measurement, according to RECIST criteria, in ovarian cancer. We aimed to determine both the interobserver and intraobserver variability of target lesion selection and measurement among experienced radiologists.

METHODS
We performed a retrospective, imaging blinded study involving eight international radiologists from tertiary hospitals with expertise in gynecological cancers. Radiologists originated from Australia (n=3), Canada (n=3), New Zealand (n=1), and the UK (n=1). The radiologists had 3–19 years of specialist experience. The study was approved by the University Health Network Regional Ethics Board (Canada) in March 2015.

A random sample of patients with advanced ovarian cancer attending the Princess Margaret Cancer Centre between January 2005 and December 2015 was generated. Two authors, a radiologist (TC) and a medical oncologist (MW), reviewed the computed tomography (CT) scans of these patients. Key images were selected. Criteria for image selection included the presence of between one and six malignant lesions arising from the peritoneum, lymph nodes, liver, lung, pelvis, or other sites. All CT images were generated from multidetector CT scans, with routine use of contrast administration in the portal venous phase.

A total of 47 images were selected and reviewed by a second medical oncologist (SL). Following approval, these 47 images were provided to eight radiologists over two viewing sessions, 12 weeks apart. Thirty-four images were viewed on a single occasion, while 13 images were viewed on two occasions. All images were viewed using standardized EFilm lite software (Joint Department of Medical Imaging, Toronto, Ontario).

Each image had a specific task, reflective of the RECIST framework, which resulted in the selection of a total of 67 lesions per radiologist. For 29 images, radiologists were asked to select and measure a single lesion according to RECIST (task 1: 29 images × 1 lesion=29 lesions). For five images, radiologists were asked to select and measure two lesions (task 2: 5 images × 2 lesions=10 lesions). For 12 images, radiologists completed task 1 at baseline and again at 12 weeks (task 3: 12 images × 1 lesion × 2 views=24 lesions). For one image, task 2 was completed at baseline and again at 12 weeks (task 4: 1 image × 2 lesions × 2 views=4 lesions). Radiologists were blinded to clinical details and to which images would be viewed twice.

For all tasks, lesion selection was recorded, and interobserver variability of selection was calculated, and stratified by site. For all tasks, individual radiologist lesion measurement was recorded; variability of measurement was calculated among radiologists selecting the same lesion to measure. For images viewed on two occasions, interobserver and intraobserver variability of selection were assessed. For images viewed on two occasions, both interobserver and intraobserver variability of measurement were determined.

Statistical Methods
Two primary statistical methods were used to analyze variability of lesion selection and measurement. For lesion selection variability, a systematic algorithm was developed and validated by a statistician, to allow assessment of the rate of reproducibility of lesion selection. The reproducibility rate was calculated from 0 to 1, where 0 represented complete discordance of selection, and 1 represented complete concordance. To calculate the reproducibility rate, the total number of radiologists viewing an image was noted. The number of radiologists selecting each lesion within the image was assessed. The lesion selected most frequently was used to determine the reproducibility rate. For example, if eight radiologists viewed an image, and five selected lesion (a) while three selected lesion (b) to measure for RECIST, the reproducibility rate was calculated as 5/8, or 0.63. Bootstrapping methods were used to obtain confidence intervals for the reproducibility rate estimates. For each estimate, 10 000 bootstrap samples were drawn with replacement, and the reproducibility rate was recalculated for each of these bootstrap samples. The percentile method was then used to construct the 95% confidence interval (CI).

A similar calculation was performed to determine the intraobserver reproducibility rate. This calculation determined the degree of reproducibility of lesion selection by the same radiologist at two time points. For example, if eight radiologists viewed an image, and at baseline, five selected lesion (a) and three selected lesion (b), while at the second time point, five selected lesion (a), two selected lesion (b) and one, previously selecting lesion (b), now selected lesion (c), then the intraobserver reproducibility rate was calculated as 7/8, or 0.875.

To assess lesion measurement variability, the intraclass correlation coefficient was calculated. The intraclass correlation coefficient provides a numerical value between 0 and 1 that reflects the degree of observer concordance. Intraclass correlation coefficient estimates were interpreted as follows: 0.00–0.20 poor concordance; 0.21–0.40 fair concordance; 0.41–0.60 moderate concordance; 0.61–0.80 good concordance; and 0.81–1.00 very good concordance.10 Interobserver and intraobserver intraclass correlation coefficient estimates were calculated with 95% CIs.

Bland–Altman plots were generated to show the average lesion size versus percent difference in lesion size for each measurement pair, alongside 95% limits of agreement. This enabled a clinical interpretation of the high versus low intraclass correlation coefficient estimates. For instance, an intraclass correlation coefficient of
0.96 had nearly all measurement pairs within the RECIST thresholds of partial response and progressive disease; by contrast, an intraclass correlation coefficient of 0.77 had several measurement pairs beyond these thresholds (Figure 1).

Finally, a variance components analysis using a linear mixed effects model was performed to determine the relative contribution of interobserver and intraobserver variability to total measurement error. All statistical analyses were performed in R V.4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS
A total of 47 images were used, generating 67 lesion selection and measurement tasks per radiologist. Interobserver variability was assessable for all 67 tasks, with intraobserver variability assessable for the 14 tasks with images viewed on two occasions. All participating radiologists viewed all images on one occasion with seven radiologists participating in the second assessment of images. Across the eight radiologists, 82 distinct lesions were selected, with a median size of 2.7 cm (range 0.8–9.5). Table 1 summarizes the location of lesions.

**Target Lesion Selection**
Figure 2 summarizes the reproducibility rates of lesion selection. The interobserver reproducibility rate was 0.91 for all lesions, indicating that in 9% of cases, radiologists selected a different RECIST target lesion for the same image. Interobserver selection reproducibility was highest for pelvic mass lesions (reproducibility rate 1) and other lesions (reproducibility rate 1). Interobserver reproducibility was 0.88 for pulmonary lesions, 0.95 for hepatic lesions, and 0.99 for lymph node lesions. The interobserver selection reproducibility rate was lowest for peritoneal lesions (0.76).

The overall intraobserver reproducibility rate was 0.93, indicating that in 7% of cases, a single radiologist selected a different target lesion for RECIST at the second time point. Intraobserver selection reproducibility was highest for pelvic masses (reproducibility rate 1), hepatic (reproducibility rate 1), nodal (reproducibility rate 1), and other lesions (reproducibility rate 1). The intraobserver reproducibility rate was 0.86 for pulmonary lesions. Intraobserver selection reproducibility rate was lowest for peritoneal lesions (0.69).

**Target Lesion Measurement**
Figure 3 illustrates the intraclass correlation coefficients for lesion measurement. The interobserver intraclass correlation coefficient showed very good concordance at 0.84. The interobserver intraclass correlation coefficient demonstrated very good concordance for pulmonary lesions (0.90), hepatic lesions (0.97), peritoneal lesions (0.84), and other lesions (0.93). Good concordance was seen for pelvic mass lesions (0.74), with only moderate concordance for nodal lesions (0.58).

Across all domains, intraobserver concordance was higher than interobserver concordance. The overall intraobserver intraclass correlation coefficient showed very good concordance at 0.94. Intraobserver intraclass correlation coefficient showed very good concordance for pulmonary lesions (0.99), pelvic mass lesions...

(0.98), peritoneal lesions (0.92), and other lesions (0.92). As with interobserver measurement, lymph node lesions showed the lowest intraobserver measurement concordance (intraclass correlation coefficient of 0.82). It was not evaluable for hepatic lesions.

Using a mixed effects model, the relative contribution of interobserver and intraobserver measurement variability to the total variability observed was measured. Interobserver measurement difference contributed to 85% of total variability, indicating that most of the discordance of lesion size arose from the different measurements provided by different radiologists.

**DISCUSSION**

**Summary of Results**

Our study demonstrated that there was both interobserver and intraobserver variability in the selection and also the measurement of target lesions in ovarian cancer. It also highlighted that interobserver reproducibility of lesion selection in ovarian cancer was generally high, particularly for pulmonary, hepatic, nodal, and pelvic lesions. The most significant variability of selection occurred when radiologists assessed peritoneal lesions. In this setting, in 24% of cases, radiologists selected different lesions.

**Results in the Context of Published Literature**

Our findings are of importance in ovarian cancer, which classically spreads within the peritoneum. Similarly, in other malignancies with peritoneum spread, such as colorectal cancer, inappropriate selection and subsequent miscalculation of peritoneal burden has been identified as a clinical issue affecting disease response classification.

In breast cancer, a major source of variability in disease status designation according to RECIST was not the lesion measurement, but rather the variable selection of target lesion for measurement between readers. In this study, we found that when the results were dichotomized as demonstrating progression or no progression, disagreement was observed in 11 (27%) of 41 patients, indicating that different selection of target lesions led to different assessment of clinical outcome. This is clearly important when so much reliance is placed on radiological response or progression.
Our study has confirmed that interobserver measurement concordance was lowest when radiologists were asked to measure lymph node lesions. McErlean et al demonstrated similar results when they examined measurement variability of liver, lung, and nodal lesions in patients with cancer.13 The investigators found that the overall interobserver measurement agreement rate was 0.967 and 0.955 for long axis and short axis measurements, respectively; measurement consistency was highest for pulmonary lesions (interobserver agreement rates for long axis and short axis dimensions of 0.945 and 0.939, respectively) and lowest for nodal lesions (interobserver agreement rates for long axis and short axis dimensions of 0.9131 and 0.884, respectively). Our study adds to the literature by confirming that lesion location can affect measurement accuracy, with the lowest accuracy occurring in lesions arising from lymph nodes.

Intraobserver variability was also examined in our study. Intraobserver measurement concordance was very high for all lesions. Based on the intraclass correlation coefficient, lymph node lesions had the lowest measurement concordance, although this result was still consistent with very good concordance. Of note, in our study, intraobserver selection and measurement error occurred less frequently than interobserver error. This lower rate of intraobserver measurement variability compared with interobserver variability is consistent with the literature. Indeed, Muenzel et al demonstrated in their study that the median intraobserver measurement variability ranged from 4.9% to 9.6% (mean 5.9%), while the median interobserver variability ranged from 4.3% to 11.4% (mean 7.1%).14 Erasmus et al demonstrated similar results.5 In contrast with these findings, McErlean et al found minimal differences between interobserver and intraobserver variability in CT measurements, with overall intraobserver agreement rates for long axis and short axis measurements of 0.957 and 0.945, compared with interobserver agreement rates of 0.954 and 0.941.13

These findings support the view that while RECIST assessment is central to clinical trial design and assessment, it may have inherent limitations. The determination of response to therapy is usually accomplished through a combination of clinical features, biochemical changes, surgical restaging, or radiological assessment. This determination requires standardization and cross-institutional application and relies on the accurate measurement of tumor response.3,4 The current study raises questions about the reliability of this approach, particularly given the supremacy of radiological assessment in clinical trials. While this limitation may be partly addressed by blinded, independent imaging review, the fundamental uncertainty regarding how radiologists select and measure target lesions remains.

In clinical practice, assessment of disease status in ovarian cancer is more multifaceted than RECIST assessment alone. Indeed, due to the possible paucity of radiological abnormalities in patients with ovarian cancer, numerous surrogate markers of response have been investigated. These include symptom assessment,15 physical examination,16 and measurement of CA125.17–20 Despite their importance, these assessments also have a number of inherent limitations, and radiological assessment of disease remains central to the evaluation of women with ovarian cancer. It is known, for instance, that CA125 is elevated in only 69–88% of ovarian cancers,21 and that this biomarker has a sensitivity of only 62–94%, with a specificity of 91–100%, in detecting recurrent disease in ovarian cancer.22,23

To circumvent these limitations, cross sectional imaging has long been viewed as an objective marker of disease activity, a non-biased ‘snapshot’ of the ‘true’ state of malignancy, which is less prone to clinician bias. The current study adds to the literature by questioning the nature of this objective ‘truth’ and supports the arguments of authors Armato24 and Kuhl,25 that there is no singular radiological truth.

Strengths and Weaknesses
Strengths of the current study include the large number of participating radiologists and the numerous different disease sites presented in the trial images. Furthermore, the study was strengthened by a simple statistical method, which allowed for a clinically meaningful reproducibility rate to be ascertained. Few studies have thus far examined interobserver and intraobserver variability of lesion selection. Our study adds to the literature by demonstrating that intraobserver selection variability is less than interobserver variability. These results highlight the need for judicious assessment of lesion selection and ideally review by a dedicated and experienced radiologist for trial assessments.

Limitations include the low number of images demonstrating non-peritoneal and non-nodal sites of disease, as well as the lack of a comprehensive image selection committee during trial design. Additionally, only single images were provided to radiologists (rather than a complete set of CT images), and this may have limited image interpretation and produced bias in lesion selection.

Implications for Future Research
An important question this study did not address is whether the variability was impacted by the experience of the radiologist. This is an important question to consider in future studies. This study also did not address the very important group of patients with recurrent ovarian cancer who do not have RECIST defined measurable disease.8 Future solutions need to consider this group.

Additionally, it would also be relevant to determine if the number of lesions per organ impacts the variability of lesion selection. In clinical practice, the most dominant lesion would likely be the selected lesion despite having multiple other lesions. It is difficult to predict how this is impacted by the number of lesions and should be considered in future research. The small sample size of this study did not enable this to be assessed.

Future research needs to recognize the higher rates of interobserver variability, as demonstrated in this study. Utilization of dedicated oncological software or systems that allow scrapbooking of target lesions for reference, and comparison at subsequent post treatment assessment, may be means of limiting both interobserver and intraobserver variability. In the future, the role of RECIST in ovarian cancer may additionally be optimized by incorporating new technologies, such as artificial intelligence based imaging modalities or fluorodeoxyglucose–positron emission tomography–CT and by developing radiology practice standards that limit variability.

CONCLUSION
This study found that RECIST concordance of lesion selection and measurement in ovarian cancer was generally high,
particularly for parenchymal lesions. Lesion selection reproducibility for peritoneal lesions and measurement of nodal lesions were relatively variable, and thus the results could be erroneous. Intraobserver variability was lower than interobserver variability, and serial assessment by the same radiologist is advisable. Prioritization of parenchymal lesions over nodal or peritoneal lesions is important, as this may reduce variability error. Additionally, in the future, incorporating novel software systems and technologies may further mitigate some of this variability. These factors need consideration to improve response assessment, especially as progression free survival remains the most common endpoint in phase III trials.

Author affiliations
1 Medical Oncology, Prince of Wales Hospital and Royal Hospital for Women, Randwick, New South Wales, Australia
2 Faculty of Medicine and Health, University of New South Wales, Sydney, New South Wales, Australia
3 NHMRC Clinical Trials Centre, University of Sydney, Camperdown, New South Wales, Australia
4 Joint Department of Medical Imaging, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
5 University of Toronto, Toronto, Ontario, Canada
6 Department of Radiology, Auckland City Hospital, Auckland, New Zealand
7 Department of Radiology, Prince of Wales Hospital and Royal Hospital for Women, Randwick, New South Wales, Australia
8 Westmead Hospital, Westmead, New South Wales, Australia
9 Swinburne University of Technology, Hawthorn, Victoria, Australia
10 University of Ottawa, Ottawa, Ontario, Canada
11 Memorial University of Newfoundland, St John’s, Newfoundland, Canada
12 Cancer and Blood, Auckland City Hospital, Auckland, New Zealand

Presented at This has been presented at ANZGOG virtually in 2021 and in part at ASCO 2018.

Acknowledgements We acknowledge the time and effort of the radiologists in their contributions to this study, in particular Drs K Tung, H Moore, L Baker, D Moses, M O’Malley, A Kielar, C Mandel, and A Hartery.

Contributors TC, SL, MKW, LW, and AMO contributed to the design of the trial. DM, LB, HM, MO, AK, AH, and CM contributed to the study assessments. H-WS, MK, and LW contributed to the statistical analysis. All authors contributed to and reviewed the manuscript and approved for submission. MW acts as guarantor for the study.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study involves human participants and was approved by the University Health Network Regional Ethics Board (Canada) in March 2015 (15-8873-CE). The study used anonymized images from patients. No consent was necessary, as deemed by the ethics board. The study was non-interventional and did not involve collection of clinical data or follow-up.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. The data are included in the analysis in the study.

ORCID iDs Hao-Wen Sim http://orcid.org/0000-0002-2530-0523
Catherine Mandel http://orcid.org/0000-0002-3962-1301

REFERENCES