

# Analysis of the Cause of Discordance Between Two Radiologists in the Assessment of Radiographic Response and Progression for Subjects Enrolled in Breast Cancer Clinical Trials Employing Blinded Independent Central Review



Kristin Borradaile MS<sup>1</sup>, Robert Ford MD<sup>1</sup>, Michael O'Neal MD<sup>1</sup>, Kevin Byrne MD<sup>1</sup>

<sup>1</sup> RadPharm Imaging Core Lab (Princeton, NJ)

## Background Information

The Food & Drug Administration (FDA) advocates blinded independent central review (BICR) of radiographic exams for registrational oncology studies when the primary endpoint is based on tumor measurements such as progression-free survival, time to progression or objective response rate. Current FDA guidance recommends multiple independent readers evaluate each subject during BICR. One consequence of this is the potential for discordance between readers on the outcome of the subjects. Discordance between readers is adjudicated by a third reader to determine the final outcome. There are no published metrics regarding the cause of discordance between BICR radiologists.

## Methods

BICR data was blinded and pooled to identify cases in which the two primary readers were discordant in outcome (overall best response, best response date, or progression date). Radiographs from 459 subjects were reviewed to determine the cause of discordance and whether it resulted from a justifiable interpretation difference where neither reader was incorrect in their assessment, or if it resulted from an assessment error by one reader. We acknowledge there may have been bias introduced into this process, as the interpretations were judged by radiologists from the same facility as the original reviewers.

## Results

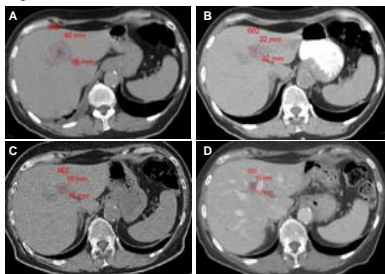
In 37% of cases (168/459), discordance resulted from a difference in lesion selection between readers. In 137 cases, the difference was justifiable, but in 31, the choice of lesions by one reader was not thought to represent the overall extent of disease at baseline. In 30% of cases (139/459), discordance resulted from a difference in the perception of new lesions. In 88 cases, the difference was justifiable, but in 51, one reader's assessment was judged to be incorrect during review. In 13% of cases (58/459), discordance resulted from a difference in the qualitative assessment of progressive disease (PD) based on non-target (NT) disease. In 39 cases, the difference was justifiable, but in 19, one reader's assessment was deemed incorrect during review. In 11% of cases (52/459), image quality issues resulted in a justifiable interpretation difference between readers. In 9% of cases (41/459), discordance resulted from a difference in lesion measurements between readers. In 37 cases, the difference was justifiable, but in 4, one reader's measurements were deemed incorrect during review. In 1 case, discordance resulted from a lack of clinical information during the review. Examples of justifiable discordance are provided.

Figure 1

READER 1	TARGET LESIONS	MEDIASTINAL ADENOPATHY (LEFT)
		LIVER MASS
		LIVER MASS
	NON-TARGET LESIONS	LIVER MASS
READER 2	TARGET LESIONS	LUNG MASS (LEFT)
		LUNG MASS (RIGHT)
		LIVER MASS
	NON-TARGET LESIONS	LIVER MASS (BILATERAL)
		BONE MASS
		MEDIASTINAL ADENOPATHY (BILATERAL)
		LUNG MASS (BILATERAL)
		LIVER MASS

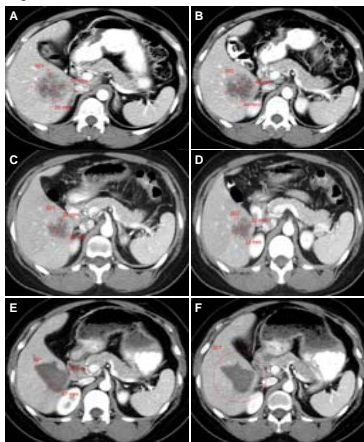
In this example, there was discordance in the date of response between reviewers (R1 & R2) due to a justifiable difference in lesion selection. R2 chose more lung disease and R1 identified hilar adenopathy. Because the thoracic lesions (classified as target disease by R2) responded at a more rapid rate than the liver lesions, R2 confirmed a partial response (PR) one time point (TP) before R1. The best response and progression date between reviewers were concordant but a justifiable difference in number and classification of lesions resulted in a response date discordance by one time point.

Figure 2



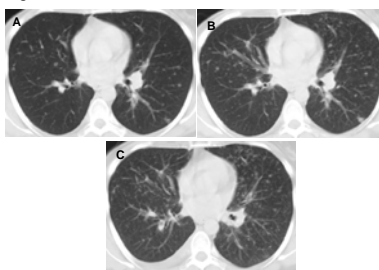
In this example, there was a justifiable discordance in response between readers based on inconsistent scan technique between assessment points. Both readers identified the same target lesion (002) in the liver, however inconsistent contrast administration made comparison between assessment points difficult. At baseline (A), the liver was imaged long after contrast was administered, resulting in delayed scans. At TP2 (B), the liver was imaged without contrast, making comparison with the baseline exam difficult. At TP3 (C), the liver was imaged temporally similar to the baseline, however the images were reconstructed with a different filter, further complicating the assessment. TP4 (D) was performed correctly with the liver imaged in the portal venous phase. At TP3 (C), R2 tried to follow the lesion, while R1 did not believe reliable measurements could be made and assessed the lesion as unevaluable. At TP4 (D), both reviewers measured the lesion and R2 confirmed PR. R1 was unable to confirm PR due to the prior unevaluable time point, and therefore, assigned a best response of stable disease (SD).

Figure 3



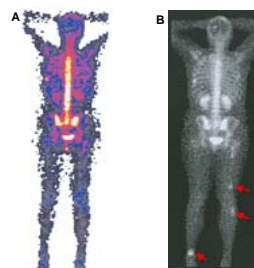
This example represents a justifiable interpretation difference in lesion measurements as well as an interpretation difference resulting from a lack of clinical information. Both R1 & R2 identified the same liver lesion as part of the baseline target disease (A-B, assessed as lesion 001 by R1 and 002 by R2). R1 & R2 measured the lesion on different image slices at baseline and subsequent time points (TP3 shown in C-D). At TP4, the lesion appeared to have increased in size (E-F). R1 measured the lesion which resulted in progression of target disease. R2 indicated the lesion had changed in appearance, decreased in density, and local treatment with radiofrequency ablation was suspected. The lesion was therefore assessed as unevaluable by R2. This represents a justifiable interpretation difference given the lack of clinical information.

Figure 4



In this example, there was discordance in progression date resulting from a justifiable difference in the perception of non-target disease progression (NT-PD). R1 assessed NT-PD at TP2 based on an increase in miliary disease (B). R2, however, did not believe that progression was unequivocal until TP3 (C). This resulted in discordance of progression date by one time point.

Figure 5



In this example, there was discordance in progression date based on a justifiable difference in the perception of new lesions resulting from image quality issues. At baseline (A), a poor quality color paper bone scan was received. At TP2 (B), a grayscale bone scan with slightly better resolution was received. At TP2, R2 identified increased uptake on the bone scan in three areas where CT correlation was not possible and assessed PD. R1 did not assess PD because in their judgment, an adequate comparison with baseline could not be made. This represents a justifiable perception difference as comparison between time points was difficult due to poor quality scans and changes in scan technique.

## Discussion

**Lesion Selection:** When radiologists function as independent reviewers, there is the potential for them to identify different target and non-target lesions that each considers representative of the subject's overall extent of disease. There is also the potential for the reviewers to identify the same lesions but classify them differently between target and non-target lesions. Because lesions do not respond or progress at the same rate, differences between reviewers with regard to response or progression of lesions can be observed.

**Perception Differences:** Due to the partially subjective nature of radiographic assessments, the presence of image artifacts, benign intercurrent diseases and occasional image quality issues, some radiologists will be more confident that a new radiographic finding truly represents a new metastatic lesion (and therefore unequivocal PD) earlier than others. Similarly, the threshold for assessing NT-PD may differ between reviewers. This is unavoidable and is in part the rationale for the two reviewer paradigm.

**Lesion Measurements:** There is a component of inter-reader measurement variability which can contribute to outcome differences between radiologists. The potential for measurement variability has many dependencies: lesion size, margination, conspicuity, phase of contrast enhancement, measurement technique (manual, semi-automated, fully automated), etc.

**Image Quality:** Given the global nature of large oncology studies and the technical equipment available in certain countries, independent reviewers often have to evaluate images of sub-optimal quality. The threshold for determining when an imaging exam is of inadequate diagnostic quality may differ slightly from one reviewer to the next. Due to the potential for reviewers to identify different lesions at baseline, missing or poor quality image data may impact one reviewer's assessment while not impacting the assessment made by the other reviewer. Similarly, missing or incomplete clinical data may impact the assessment made by one reviewer but not the other.

## Conclusion

Some factors that cause discordance are process-driven or due to justifiable interpretation differences, as was observed in 77% of cases in this review (354/459). Examples include lesion selection, image quality issues, inter-reader measurement variability, perception of new lesions and the assessment of NT-PD. Separate from these factors are reader assessment errors, such as those identified in 23% of cases (105/459). This compares favorably to reports in the literature regarding radiologist error rates. The BICR process of adjudication by a third reader identifies and mitigates these justifiable discordances and interpretive errors.