Sources and Types of Inter-Reader Discordance in RECIST 1.1 Blinded Central Review: A Retrospective Pilot Study

William Beute, MS[1]; Veronika Matsuk, PhD[1]; Gene Kim, PharmD[1]; Shinji Yue, MS, MBA[1]; Bernard Beute, MD[1]

[1]MERIT CRO, Inc., Madison, WI, USA

4888eP

Presenting Author:
Shinji Yue, MERIT CRO, Inc.; syue@meritcro.com

MERIT
Your Clinical Endpoint Expert

BACKGROUND

- Inter-reader discordance in blinded independent central review (BICR) under RECIST 1.1 and other response criteria affects oncology trial endpoints, yet the sources and types of discordance remain poorly characterized.
- Prior work has shown that target lesion selection variability reduces response agreement (**Kuhl et al., Radiology 2019**), but no study has classified discordance into mechanistically distinct types or identified which baseline factors predict each type.

OBJECTIVES

- Classify RECIST 1.1 inter-reader discordance in BICR into mechanistically distinct types
- Identify baseline factors that predict each discordance type

METHODS

Study Design and Source Data

- Retrospective pilot analysis of independent RECIST 1.1 reads (Reader 1 versus Reader 2) from a single GI oncology trial (n = 404 participants, 1,189 visits)
- Three CDISC domains were used: Tumor Identification (TU), Tumor Results (TR), and Response (RS), restricted to RECIST 1.1 Read 1 and Read 2 evaluator fields

Discordance Type (DC) Classification

- Ordinal rank assigned per response: CR (Complete Response) = 0; NED (No Evidence of Disease) = 1; PR (Partial Response) = 2; SD (Stable Disease) = 3; PD (Progressive Disease) = 4; NN = non-CR/non-PD, tracked separately
- **DC1:** one-step ordinal difference, including the NN/PD boundary when a subject has zero target lesions
- **DC2:** PD versus non-PD, excluding NN/PD zero-target cases, which are classified as DC1
- **DC3**: CR or PR versus all other categories; may co-occur with DC1 or DC2

Baseline Predictors Evaluated

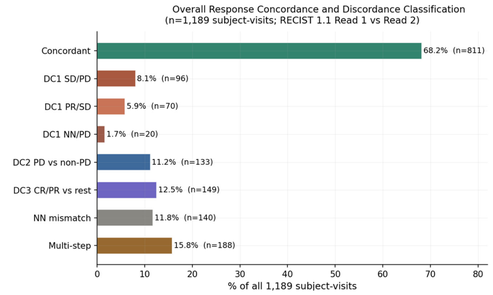
- Target lesion count difference (TU domain)
- Zero-target lesion selection discordance
- Lesion location overlap (Jaccard similarity, TULOC)
- Baseline sum-of-diameters (SOD) difference, absolute and relative (TR domain, SUMDIAM)

Statistical Analysis

- Chi-square tests (2 by 2 contingency) were used to assess binary baseline factors against each DC type
- Mann-Whitney U tests were used to assess continuous predictors (SOD difference, Jaccard similarity, target count difference) against each DC type

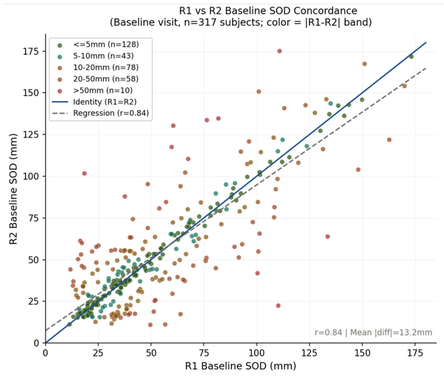
RESULTS

Figure 1. Every participant visit in which both readers recorded an overall response was classified into one of eight categories; the concordant category (green, 68.2%) serves as the reference group, representing visits in which both readers reported the same response.



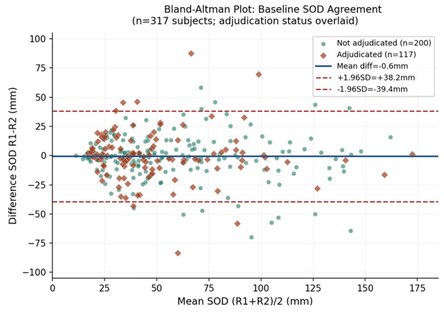
DC1 is subdivided into three sub-bars corresponding to the specific one-step boundaries observed (SD/PD, PR/SD, NN/PD); the remaining bars (multi-step, NN mismatch) represent cases outside the primary DC type framework.

Figure 2. Agreement between Reader 1 (R1) and Reader 2 (R2) in baseline sum of target lesion diameters (SOD).



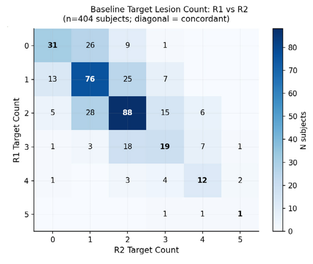
Points on the identity line (each point = 1 participant) indicate perfect agreement; point color indicates the magnitude of disagreement (dark green, less than 5 mm; red, greater than 50 mm). The regression slope of 0.87 indicates that Reader 2 tended to record slightly lower SOD values than Reader 1 at higher tumor burdens. The wide dispersion from the identity line is consistent with the mechanism underlying DC Type 3 (PR/SD discordance).

Figure 3. Agreement between Reader 1 and Reader 2 in baseline sum of target lesion diameters (SOD).



The horizontal axis represents the mean SOD of the two readers; the vertical axis represents the difference between Reader 1 and Reader 2 SOD. The mean difference (-0.6 mm) indicates negligible systematic bias, although the 95% limits of agreement (-39.4 to +38.2 mm) indicate that individual differences may be substantial.

Figure 4. Differences between readers in target lesion selection at baseline.



Each cell indicates the number of subjects with the corresponding combination of Reader 1 count (row) and Reader 2 count (column); the diagonal represents concordant pairs, with the two-lesion cell (n = 88) representing the most frequent agreement. The zero/zero cell (n = 31) represents concordant non-measurable disease. Overall, 47.5% of subjects fell off the diagonal, indicating that nearly half of subjects had a discordant target lesion count between readers.

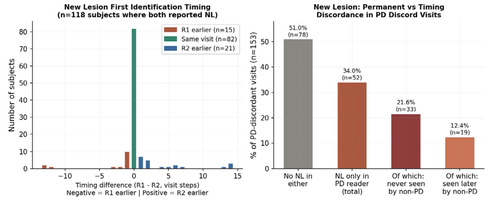
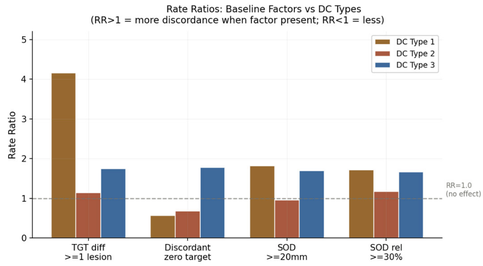
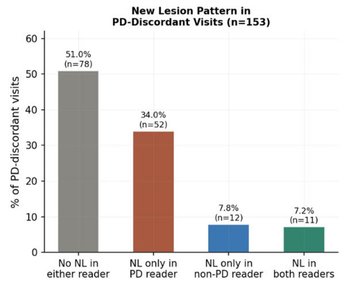


Figure 5. Specific baseline selection characteristics against each DC type tested using chi-square tests, with rate ratios calculated to quantify the strength of each relationship.



The dashed reference line at 1.0 denotes no effect. Target lesion count difference is associated with a rate ratio of 4.16 for DC1, indicating that DC1 is 4.2 times more frequent when lesion counts differ between readers, whereas the corresponding DC2 rate ratio remains near 1.0; therefore, DC2 appears to be independent of all baseline selection factors evaluated.

Figure 6. New lesion identification represents the most subjective component of RECIST assessment.

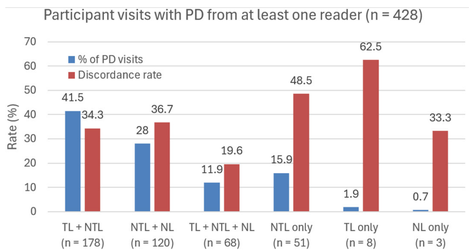


The category of no new lesion in either reader (51%) indicates that PD was attributable to target lesion growth or non-target lesion worsening. The category of a new lesion identified only by the PD-calling reader (34%) represents the most clinically actionable finding, as the other reader identified no new lesion.

Figure 7. New lesion identification represents the most subjective component of RECIST assessment.

In the left panel, the central bar (69.5%) indicates that both readers identified the lesion at the same visit; a smaller subset of cases showed a difference of five to fourteen visit steps. In the right panel, among the 52 visits in which only the PD-calling reader identified a new lesion, the other reader never identified any new lesion across the entire follow-up period in 63.5% of cases, indicating a persistent difference in perception rather than a delay in detection.

Figure 8. Visits in which at least one reader reported progressive disease (PD).



Target lesion plus non-target lesion involvement (TL+NTL), the most frequent driver (41.5% of PD visits), was associated with moderate discordance. Non-target lesion involvement alone (NTL only), although accounting for only 15.9% of PD visits, was associated with the some of the highest discordance rates (48.5%). The combination of non-target lesion worsening and new lesion appearance (NTL+NL) showed moderate discordance (36.7%).

CONCLUSIONS

- DC1 and DC3 are driven by baseline target-lesion-count differences and SOD measurement and selection mechanisms
- DC2 (PD versus non-PD) is independent of all baseline selection factors tested, consistent with a new-lesion- and non-target-driven mechanism at follow-up instead
- New lesion detection is consistently the highest driver of reader discordance, being inherently a perception task; one reader identifies a new lesion while the other does not, which is difficult to standardize with more objective, quantitative thresholds
- Zero-target lesion selection at baseline is a preventable, 100%-specific cause of the NN/PD boundary, one of the two highest-stakes disagreements observed in this cohort
- Three mechanistically distinct sources of discordance require three distinct interventions: baseline lesion-selection training for DC1 and DC3, structured new-lesion review procedures for DC2, and protocol-level adjudication triggers beyond date-of-progression rules for all discordance types

DISCLOSURES

The authors have no financial conflicts of interest to disclose.

REFERENCES:

1. Kuhl CK, et al. *Validity of RECIST Version 1.1 for Response Assessment in Metastatic Cancer: A Prospective, Multireader Study. Radiology*. 2019.