# RECIST 1.1: An Analysis of the Classification Stability of a Tumor Measurement Scale in Oncology Clinical Trials

**Author:** Pat Callahan

**University:** Ludwig-Maximilians-Universität München
**Department:** Department of Medicine
**Institute:** Institute for Biometry and Epidemiology
**Program of Study:** MSc Epidemiology

---

**Cooperating Institution:** Boehringer Ingelheim AG
**Division:** Global Biostatistics and Data Sciences
**Thesis Advisor:** Dr. Cornelia Ursula-Kunz

**Reviewer:** Dr. rer. biol. hum. Dennis Freuer M.Sc.

München, Deutschland

July 31st, 2025

# Table of contents

# Abstract

**Background**: Accurate and unbiased assessment of tumor response to treatment is essential in cancer clinical trials. The Response Evaluation Criteria in Solid Tumors (RECIST) is the standard for evaluating tumor response, and assessments are commonly performed both by local site investigators and by blinded independent central reviewers. While RECIST has been widely adopted, questions remain about its reliability and the consistency of tumor response classification across different raters. Site investigators in particular may exhibit bias in tumor measurement or identification compared to blinded central reviewers. Previous studies have either individually examined inter-rater reliability or compared site investigator and central reviewer assessments, but gaps remain in terms of meta-analyzing reliability, assessing site investigator and central reviewer discrepancies, and in understanding the impact of RECIST's threshold definitions on classification stability.

**Purpose**: This study aims to address these gaps through three main analyses. First, we conduct a general meta-analysis of inter-rater reliability for RECIST to establish an overall estimate of agreement. Second, we analyze discrepancies between site investigator and central reviewer assessments of tumor response, building on existing literature and providing new data from three cancer clinical trials. Third, we perform a sensitivity analysis to evaluate how changes in the disease response and progression thresholds affect the classification of tumor response. Together, these analyses provide a comprehensive assessment of RECIST reliability and its implications for clinical trial outcomes.

**Methods**: First, systematic review of the literature was conducted to assess the inter-rater reliability (IRR) of RECIST in terms of Cohen's and Fleiss' kappa statistics. Second, a retrospective analysis of three cancer clinical trials was performed, comparing site investigator and central reviewer assessments of tumor response. Specifically, trial endpoints of time to progression, time to response, and duration of response were analyzed for differences in hazard ratios between site investigators and central reviewers. Differences in hazard ratios between raters within studies were scrutinized, and these differences were synthesized into meta-analyses. Traditional null-hypothesis significance testing as well as equivalence testing on a hazard ratio range of $[0.80, 1.25]$ were performed to assess the significance of differences (or equivalence) in hazard ratios. Differences between raters in objective response rate were also analyzed using Cochran's Q and McNemar's tests. Third, a sensitivity analysis was conducted to evaluate how changes in RECIST's disease response and progression thresholds affect the classification of tumor response. This involved simulating different threshold definitions and assessing their impact on the aforementioned IRR and trial endpoint analyses.

**Results**: The meta-analysis of inter-rater reliability for RECIST revealed a substantial level of agreement across studies, with a pooled kappa coefficient of 0.66. Of the included studies, 4 of them were directly from clinical trials, and exhibited clustering around a lower kappa value of ~0.45. The analysis of discrepancies between site investigator and central reviewer assessments showed that while there were some significant differences in hazard ratios within individual studies, no systematic differences were observed. Moreover, we were able to demonstrate equivalence between site investigators and central reviewers in the time to progression and time to response outcomes. The sensitivity analyses demonstrated that changes in RECIST's threshold definitions do not significantly affect the differences between site investigators and central reviewers with regards to tumor response classification.

**Conclusion**: IRR analyses show substantial overall agreement between raters, but this might be lower in the real-world setting of clinical trials. Follow-up analyses of trial endpoints, however, do not point to systematic differences between site investigators and central reviewers, but individual studies still may show differences. These findings are consistent with previous literature, which has shown that RECIST is overall a reliable tool for assessing tumor response, but also highlight that individual studies likely benefit from the inclusion of blinded central reviewers. The sensitivity analyses further confirm that RECIST's arbitrary threshold definitions do not significantly impact classification stability, suggesting that RECIST remains a robust tool for evaluating tumor response in clinical trials.

# Notice

**Boehringer Ingelheim wishes to thank the following companies who have contributed data to TransCelerate's DataCelerate platform, which facilitates the sharing of control arm data between participating member companies: [AbbVie, Amgen, Astellas, AstraZeneca, Boehringer Ingelheim, Bristol-Myers Squibb, EMD Serono, GlaxoSmithKline, Genentech, Janssen, Lilly, Novartis, Novo Nordisk, Pfizer, Sanofi, Shionogi, and UCB Biosciences.] Neither TransCelerate Biopharma nor any of the above-referenced companies have contributed to, approved or are in any way responsible for the author's research results.**

# Acknowledgements

I am grateful to LMU Munich's Department of Medicine IBE and the support staff and professors therein who have helped enable my success in this program and in this thesis, and who have provided a supportive and enriching academic environment. I extend my sincere appreciation to my thesis supervisor, Dr. Cornelia Ursula-Kunz, for her guidance and support throughout this thesis process (and for always challenging me to write more). As well, my many thanks to Boehringer Ingelheim for their support of this project and the opportunity to work with them on this thesis.

I owe tremendous gratitude to my previous mentor, Prof. Katherine Rankin, who has been a great source of inspiration and support throughout my academic career, and without whose mentorship I would not have developed the skills and knowledge necessary to complete this thesis or my master's degree.

I wish to thank the friends who have supported me along the way, and particularly those who have been on this journey with me in the Epidemiology program. I am especially thankful to my roommates in Munich who have been an unending source of laughter, support, and companionship during my time here, making the experience all the more enjoyable and memorable, and who have been instrumental in helping me to feel at home in a new city and culture.

And finally, I am forever grateful to my mom and dad for their support in all my endeavors, and for supporting me in finding many opportunities to learn and grow and to repeatedly find new homes thousands of miles away from *home*. My heartfelt thanks to my partner for her support, patience, and especially for all the delicious homecooked meals that kept me healthy and sane over the last several weeks. I also want to thank my siblings for their encouragement and support over the last several years. And of course, my gratitude extends to all of our family pets: past, present and future who have brought and bring joy, companionship, and love to my life.

# Listings

# List of Figures

# List of Tables

# List of Abbreviations

| Abbreviation | Expansion |
|---|---|
| BICR | Blinded Independent Central Review |
| CDISC | Clinical Data Interchange Standards Consortium |
| CI | Confidence Interval |
| CR | Complete Response |
| DCR | Disease Control Rate |
| DOR | Duration of Response |
| EMA | European Medicines Agency |
| FDA | Food and Drug Administration |
| HR | Hazard Ratio |
| HTD | Historical Trial Data |
| IRR | Inter-Rater Reliability |
| LME | Linear Mixed Effects |
| NCT | National Clinical Trial |
| NHST | Null Hypothesis Significance Testing |
| ORR | Objective Response Rate |
| OS | Overall Survival |
| PD | Progressive Disease |
| PFS | Progression-Free Survival |
| PR | Partial Response |
| PSOC | Placebo and Standard of Care |
| RCT | Randomized Controlled Trial |
| RECIST | Response Evaluation Criteria in Solid Tumors |
| RS | Disease Response Domain |
| SD | Stable Disease |
| SDTM | Study Data Tabulation Model |

| Abbreviation | Expansion |
|---|---|
| SLD | Sum of Longest Diameters |
| TOST | Two One-Sided Tests |
| TR | Tumor Results Domain |
| TTP | Time to Progression |
| TTR | Time to Response |
| TU | Tumor Domain |
| WHO | World Health Organization |
| df | degrees of freedom |

Table 1.: Table of Abbreviations

# 1. Introduction

This thesis examines the reliability of the Response Evaluation Criteria in Solid Tumors 1.1 (RECIST) (1), a widely used tool for assessing tumor response in oncology clinical trials. Reliable assessment of tumor response is crucial for evaluating treatment efficacy, yet concerns exist regarding the consistency of RECIST interpretations across different raters. Beginning with a general overview of cancer biology, global cancer epidemiology, and the challenges of measuring treatment outcomes in oncological studies, this Introduction establishes the context for understanding the importance of standardized response criteria. It further explores the role of clinical trials in drug development, regulatory requirements for tumor response assessment, and the operational challenges of implementing RECIST in multi-center trials. Following this, the express purpose of this thesis is articulated.

The subsequent Methods chapter details the research approach used to address the reliability questions, including a systematic literature review on inter-rater reliability of RECIST, retrospective analysis of discrepancies between site investigator and central reviewer tumor assessments using RECIST in several clinical trials, and sensitivity analyses examining how threshold definitions affect classification differences in the context of these trials. Following the presentation of the statistical methodology, the Results section reports findings from the meta-analysis of inter-rater reliability, quantifies site vs. central reviewer assessment discrepancies, and evaluates how varying threshold definitions impact response categorization.

The Discussion chapter contextualizes these findings within existing literature and explores their implications for clinical trials and research methodology while also highlighting some of the limitations of this thesis. Finally, the Conclusion synthesizes the key insights, readdresses the primary research questions, and offers recommendations for enhancing the reliability of tumor response assessment in future oncology trials.

## 1.1. A Brief Primer on Cancer

Before diving into the specifics of tumor response assessment and the reliability of RECIST, it is necessary to review several aspects of cancer as a disease and the value in developing treatments against cancers. This section thus provides an elementary overview of cancer biology, its global burden, and the challenges associated with treatment and response assessment.

### 1.1.1. Cancer 101

Cancer is fundamentally a disease of dysregulated cellular behavior, rooted in alterations at the molecular and cellular levels (2). Under normal physiological conditions, cells adhere to regulatory pathways that govern their growth, division, differentiation, and death. In contrast, cancer arises when these regulatory systems are disrupted, leading to unrestrained cell proliferation and, ultimately, tumor formation (2). The transformation from a normal to a malignant cell is typically driven by the accumulation of genetic and epigenetic mutations that interfere with key cellular processes (2,3). These mutations often activate oncogenes which promote cell division, and they can inactivate tumor suppressor genes which ordinarily function to restrain growth or induce apoptosis[1] in damaged cells (3,4). These cells thus possess the ability to proliferate indefinitely, resist apoptosis, and ignore signals that would normally inhibit growth.

Metastasis, the process by which cancer cells spread beyond the primary tumor site, represents one of the most clinically significant and challenging aspects of cancer (5,6). Once metastasized, tumors are often less responsive to localized treatments such as surgery or radiation, necessitating systemic therapies that are typically less targeted and more toxic (6). Metastasis also correlates strongly with poorer prognosis, and its presence at diagnosis is a key determinant of clinical outcomes (5). Complicating the diagnosis and treatment of cancer is its remarkable heterogeneity (6). Cancer is not a singular disease but rather a collection of disorders characterized by diverse genetic profiles and clinical behaviors (5,6). Even within a single type of cancer, significant variation can exist between patients and even within the different tumors of a single patient (5,7).

It is also important to distinguish between two broad categories of cancer: solid tumors (i.e. carcinomas and sarcomas) and hematological malignancies (2). Solid tumors arise

[1] Apoptosis is a form of programmed cell death that occurs in multicellular organisms, allowing for the removal of damaged or unwanted cells.

in tissues and organs such as the breast, colon, lungs, or in connective tissue such as muscle and bone although this latter grouping, so-called, sarcomas, are far less common (2). In either case though, the cancer from these solid tumors tends to form localized masses (2). In contrast, hematological malignancies, such as leukemia and lymphoma, originate in the blood or bone marrow and typically disseminate early in the disease process (2). Given their distinct biological behaviors, diagnostic criteria, and therapeutic strategies, this thesis will focus exclusively on cancers that present as solid tumors. This focus will allow for a more coherent investigation into the cellular and molecular dynamics specific to solid tumor biology and their implications for diagnosis, treatment, and patient outcomes.

### 1.1.2. Cancer in Context: Global Burden

While the biological bases of cancer are generally well understood, the disease continues to present a substantial global public health challenge. Cancer represents one of the leading causes of morbidity and mortality worldwide (8), with millions of new cases diagnosed annually. Current estimates indicate approximately 20 million incident cases of cancer per year globally alongside an estimated 9 million annual deaths (9). Of these, the 5 most common types include lung, breast, colorectal, prostate, and stomach cancers, which together account for nearly half of all new cases, and generally all present as solid tumors (9).

These figures, though significant, represent only point estimates in an evolving epidemiological landscape. Global trends in cancer incidence and mortality have shown concerning increases over recent decades, with projections suggesting this burden may rise by up to 75% to 35 million incident cases by 2050 (9). However, these statistics exhibit considerable variation across cancer types, geographic regions, and healthcare access levels (9). Improved diagnostic capabilities and expanded access to healthcare systems might contribute to higher reported incidence rates in some regions (10), though the overall trend toward increasing cancer burden appears consistent (9).

Given its current and future profound impact on global health and an economic burden in the trillions (11), cancer remains a critical priority for healthcare systems, researchers, and policymakers worldwide with nearly 4 billion USD in funding in 2024 from the US government, UK government, and European Commission alone (12). The United States' National Institutes of Health dominates this figure with over 3 billion USD allocated to cancer research per year (12), accounting for approximately 10% of the total NIH budget (12,13). The scale of these investments underscores the urgency of developing more effective approaches to cancer diagnosis, treatment, and monitoring including improved methods for assessing treatment response in clinical settings.

### 1.1.3. Treatment Challenges

Despite remarkable advances in cancer research and substantial ongoing investment (14), cancer remains one of medicine's most formidable challenges due to its inherent complexity and heterogeneity (14,15). Each tumor possesses a unique genomic profile with distinct mutations and cellular characteristics that can vary both between patients with the same cancer type and even within different regions of a single tumor (3,15). This molecular diversity means treatments effective for one patient may fail in another, even if their clinical presentations are similar (15).

The available treatment options for cancer are diverse to meet the diversity of tumor characteristics, and can generally be divided into the so-called "pillars" of cancer treatment. Different authors appear to define different categories and quantities of pillars (16,17), but surgery, radiation therapy, chemotherapy, and immunotherapy, appear to be agreed-upon pillars with targeted therapy, hormone therapy, and cell therapy also being included at times (17). Regardless of the classification system used, each of these treatment modalities has its own strengths and limitations; the choice of treatment often depends on the specific characteristics of the tumor, its stage, and the patient's overall health (17). Surgical interventions, for example, can potentially be curative for localized disease, but cannot address metastases (18). Radiation therapy risks damaging adjacent healthy tissues (18), while chemotherapy often lacks specificity, causing substantial global toxicity (16). Without going into detail of the other modalities, it suffices to say that each different treatment comes with its own set of positives and negatives that must be weighed in the context of the cancer being treated and the needs of the individual patient receiving treatment.

Given the careful balancing act required to provide treatments that provide therapeutic benefit without excessive downsides, a central concern in cancer treatment is thus determining the effectiveness of a given therapy. This is particularly challenging in oncology, where treatment responses can be complex and multifaceted (19,20). Unlike many other medical conditions, cancer treatments often do not yield immediate or straightforward outcomes such as psuedoprogression, wherein tumor sizes on imaging might swell as a direct result of the treatment rather than due to true progression (21). Moreover, tumors may shrink, stabilize, or even grow despite treatment, and these changes can occur at different rates depending on the individual patient and the specific therapy used (4,21). Current evaluation methods, particularly imaging-based assessments of tumor burden, introduce additional variables related to technique and interpretation, emphasizing the need for standardized and validated criteria like RECIST while acknowledging their inherent limitations.

## 1.2. Measuring Tumor Response in Solid Tumors: RECIST

As the previous section highlighted, determining the effectiveness of cancer therapies presents unique challenges due to the complex and variable nature of tumor responses. Given these challenges, standardized assessment tools become essential for consistent evaluation across different clinical settings. While the discussions of RECIST criteria and the broader challenges of measuring tumor response in clinical trials are inherently intertwined, we begin by examining RECIST itself as a framework, assuming the reader has at least a passing familiarity with clinical trials, and we will address specific challenges of clinical trials later in this Introduction.

This section explores the historical context that necessitated the development of standardized response criteria, followed by a detailed overview of RECIST's technical specifications and algorithmic structure. The discussion then extends to the practical implementation of RECIST in clinical settings, focusing particularly on inter-rater reliability challenges, discrepancies between site investigator and central reviewer assessments, and the complexities of applying these criteria consistently in clinical trials. Understanding both the technical aspects of the RECIST algorithm and its real-world reliability challenges is essential for interpreting the findings presented in this thesis and for appreciating their broader implications for clinical trial methodology, regulatory decision-making, and ultimately, patient care in oncology.

### 1.2.1. Historical Context: The Need for Standardization

The development of standardized tumor response assessment criteria emerged largely over the course of the early and mid 20th century in large part due to inconsistent and often dangers evaluation methods (22). One of the earliest examples of attempts to standardize the general process of a clinical trial in the modern sense (i.e., with a proper control group, clearly defined evaluation procedures, and objective endpoints) was published in 1960 by Zubrod et al. (22,23). However, it wasn't until 1979 that the WHO introduced the first internationally recognized tumor response assessment standard, coinciding roughly with the wider availability of computed tomography and magnetic resonance imaging (24,25). Prior to this standardization, assessment of tumor response in oncology trials was highly variable and often relied on subjective evaluations by local investigators (22), leading to inconsistencies in how treatment effects were measured and reported across different institutions and studies.

The establishment of the World Health Organization (WHO) tumor assessment criteria in 1979 represented a pivotal advancement in oncology research. The WHO guidelines introduced a bidimensional approach to measuring and assessing tumor burden, wherein "the sum of the products of the two longest diameters in the perpendicular dimensions of all tumors" was calculated (26) and four key response categories were defined: Complete Response (CR), Partial Response (PR: 50% reduction), Stable Disease (SD), and Progressive Disease (PD: 25% increase or new lesions). Widely adopted and validated across numerous tumor types, the WHO criteria provided a consistent framework that enabled meaningful comparisons of treatment efficacy across trials for the first time (24).

Defining these cut-offs and categories is thus the measurement basis for calculating key trial endpoints such as Objective Response Rate (ORR), Progression-Free Survival (PFS), and Overall Survival (OS). These endpoints are critical for evaluating the effectiveness of new therapies and for regulatory approval processes, as they provide standardized measures of treatment benefit that can be compared across different studies and patient populations. Of course, there is an implicit assumption in these WHO criteria that tumor burden accurately reflects disease status and is predictive of outcomes; we discuss this topic in detail in Section 1.3.2. Despite this significant step forward, the WHO system was not without limitations. Its approach to measuring tumor response allowed for tracking of an indeterminate number of solid tumors, was vulnerable to manual error, and imposed a substantial time burden on clinicians (26,27). The practical demands of measuring multiple diameters per lesion made assessments time-consuming, while variability in how different observers selected and measured lesions further undermined reproducibility (27).

### 1.2.2. Improving on WHO Criteria: RECIST

In response to the limitations of the WHO criteria, RECIST was developed as a simplified and standardized framework for tumor response assessment with the first version, RECIST 1.0, introduced in 2000 (27). The most fundamental innovation of RECIST over the WHO guidelines was the adoption of unidimensional measurements, where only the longest diameter of target lesions would be measured, rather than the more complex bidimensional approach required by WHO criteria (27). This change alone substantially reduced computational complexity and measurement error, making the assessment process more efficient and reproducible across different investigators and trial sites.

RECIST 1.0 also introduced a more structured approach to lesion selection and categorization. The criteria allowed for up to 10 target lesions (with a maximum of 5 per organ), which would be measured and tracked throughout treatment (27). Additionally, the framework formalized the concept of non-target lesions (lesions not directly measured but still assessed qualitatively for response or progression) and established the identification of new lesions as an absolute marker of disease progression (27). The calculation process was dramatically simplified by introducing the sum of longest diameters (SLD) as the primary metric, eliminating the need for the complex product calculations required by WHO criteria (27). RECIST 1.0 also established new threshold values for calculating progression (>20% increase) and response (>30% decrease) in target tumor SLDs (27). This new framework was rapidly adopted across the oncology community due to its practicality, efficiency, and capacity to enhance consistency in multi-center trials.

Building on nearly a decade of implementation experience, RECIST 1.1 was introduced in 2009 to address certain limitations identified in the original criteria (1). This update further streamlined the assessment process by reducing the maximum number of target lesions from 10 to 5 (with no more than 2 per organ), which research had shown was sufficient for accurate response assessment while further reducing measurement burden (1). RECIST 1.1 also provided more detailed guidance on lesion selection, emphasizing the importance of choosing measurable lesions and excluding non-measurable abnormalities from target designation (1). RECIST 1.1 also established specific criteria for lymph node assessment, specifying minimum size requirements and measurement approaches (1). Additionally, it introduced an absolute minimum size threshold of 5mm for target lesions that could not be accurately measured, and implemented a similar absolute increase requirement to determine progression of disease (1). These refinements, alongside enhanced guidance on imaging techniques and assessment timing, further improved the standardization and reliability of tumor response evaluation in clinical trials.

Having traced the evolution of RECIST from its origins to the current 1.1 version, we can now examine its technical framework in detail. The historical development of these criteria contextualizes why certain specifications were adopted and how they address previously identified limitations in tumor assessment methodology. The following section provides a comprehensive overview of RECIST 1.1's operational structure. This detailed understanding is essential for interpreting the reliability analyses presented later in this thesis, as the technical nuances of RECIST-based lesion classification and response categorization are of direct relevance to the consistency of assessments across different raters.

### 1.2.3. Technical Specifications of RECIST 1.1

As alluded to, lesions are classified into three distinct categories in RECIST 1.1, each with specific roles in determining treatment response. These categories are target lesions, non-target lesions, and new lesions. Target lesions and non-target lesions (if any) are evaluated at baseline and subsequently measured at each imaging assessment, while new lesions are identified based on their appearance during the course of treatment. The classification of lesions into these categories is critical for the overall response assessment, as it determines how changes in tumor burden are interpreted and categorized. The RECIST 1.1 framework provides clear definitions and measurement guidelines for each lesion type, ensuring that assessments are both standardized and reproducible across different clinical settings. The paragraphs below provide a look at each category and the algorithmic approach used to assess treatment response based on these classifications is presented in Table 1.1.

**Target lesions** form the quantitative foundation of the RECIST assessment. These lesions must be measurable in at least one dimension and meet specific size criteria: a minimum of 10 mm in the longest diameter for non-lymph nodes and 15 mm for lymph nodes (where the longest perpendicular diameter is used). RECIST 1.1 limits the selection to a maximum of five target lesions with no more than two per organ. These constraints ensure focus on the most clearly measurable disease manifestations while maintaining a manageable assessment burden. For these target lesions, the SLD is calculated at each imaging assessment, and a Target Response is determined based on specific threshold changes: CR requires disappearance of all target lesions; PR is defined as at least a 30% decrease in SLD compared to baseline; PD occurs with at least a 20% increase in SLD compared to the smallest SLD recorded since treatment initiation (i.e. since the point of nadir[2]); and SD applies when neither PR nor PD criteria are met.

**Non-target lesions** complement the quantitative assessment of target lesions. These are abnormalities that, while identified at baseline, are not selected for measurement due to their small size, irregular shape, or poor delineation. Although not measured directly, non-target lesions are still evaluated qualitatively for response or progression. Their assessment contributes significantly to the overall response determination, particularly when "unequivocal progression" is observed, which can indicate PD regardless of target lesion measurements.

**New lesions** constitute the third category and are defined as any lesions that were not present at baseline but appear during treatment. The identification of new lesions serves as an absolute marker of disease progression, overriding any positive changes observed in target or non-target lesions. This categorical classification reflects the biological significance of disease spread to new sites, which fundamentally indicates treatment failure regardless of responses elsewhere.

The overall response assessment in RECIST 1.1 integrates findings from all three lesion categories according to a predetermined algorithm. Table 1.1 summarizes the possible combinations of target, non-target, and new lesion assessments and their corresponding overall response classifications.

| Target Lesions | Non-Target Lesions | New Lesions | Overall Response |
|---|---|---|---|
| CR | CR | No | CR |
| CR | Non-CR/non-PD | No | PR |
| CR | Not evaluated | No | PR |
| PR | Non-PD or not all evaluated | No | PR |
| SD | Non-PD or not all evaluated | No | SD |
| Not all evaluated | Non-PD | No | NE |
| PD | Any | Yes or No | PD |
| Any | PD | Yes or No | PD |
| Any | Any | Yes | PD |

Table 1.1.: Summary of RECIST 1.1 Response Categories and Combinations

[2] Nadir is the lowest point reached by a variable, in this case the SLD of target lesions, during treatment. It is used as a reference point for determining progression.

It is worth emphasizing certain critical aspects of the RECIST Overall Response calculation that have particular relevance for the reliability analyses presented in this thesis. The presence of new lesions or unequivocal progression in non-target lesions serves as an absolute indicator of disease progression, regardless of favorable changes observed in target lesions. This means that even if target lesions demonstrate CR, PR, or SD, the detection of new lesions or significant worsening of non-target lesions will override these positive findings, resulting in an overall classification of PD. This hierarchical decision structure reflects the biological understanding that cancer spread to new sites fundamentally represents treatment failure, regardless of responses in previously identified disease locations.

An important aspect of RECIST evaluations to keep in mind is that the Overall Response can take any of the values CR, PR, or SD up until the point at which a patient is classified as PD at which point they would generally be removed from study participation. This is of practical importance for endpoints that include positive responses such as ORR, time to response (TTR), and duration of response (DoR) as these endpoints are calculated using only patients who achieved CR or PR at any point during treatment. Moreover, because nadir can be a shifting point, the SLD used to determine progression can change over the course of treatment, meaning that a patient who was previously classified as PR or SD could later be reclassified as PD if their SLD increases by 20% from the lowest point recorded since treatment initiation. This dynamic nature of RECIST assessments means that RECIST ratings are only quasi-ordinal in nature, as the numeric basis for the ratings can change over time. Figure 1.1 below illustrates four possible patterns of RECIST 1.1 Target Lesion assessments over time, demonstrating how classification is a moving target for individuals that change based on the SLD of target lesions.

Each row of the four rows in this figure can be thought of as the trajectory of a patient's SLD over the course of a baseline visit and 5 follow-up visits, with the y-axis normalized to 100% of the baseline SLD. Greyed out regions indicate timepoints not yet observed i.e. the rows should be read left to right. The red horizontal lines indicate the thresholds for progression (20% increase from nadir), while the green horizontal lines indicate the thresholds for response (30% decrease from baseline). In the first row, the patient neither achieves a response nor experiences progression, remaining classified as SD throughout the treatment course. In the second row, the patient's SLD decreases slightly, lowering the nadir value and thus the threshold for progression in the process. However, their disease course begins to worsen, and they are ultimately classified as PD at the final visit as they have crossed the red line indicating a 20% increase in SLD from the nadir. The third row illustrates a patient who achieves a PR at their third follow-up visit as they cross under the green line, and who simply continues to show response over the following visits. The fourth row shows a patient who achieves a PR at their second follow-up visit, but whose disease worsens over time, ultimately being classified as PD at the final visit. This illustrates how RECIST classifications can change over time based on the SLD of target lesions and the appearance of new lesions.

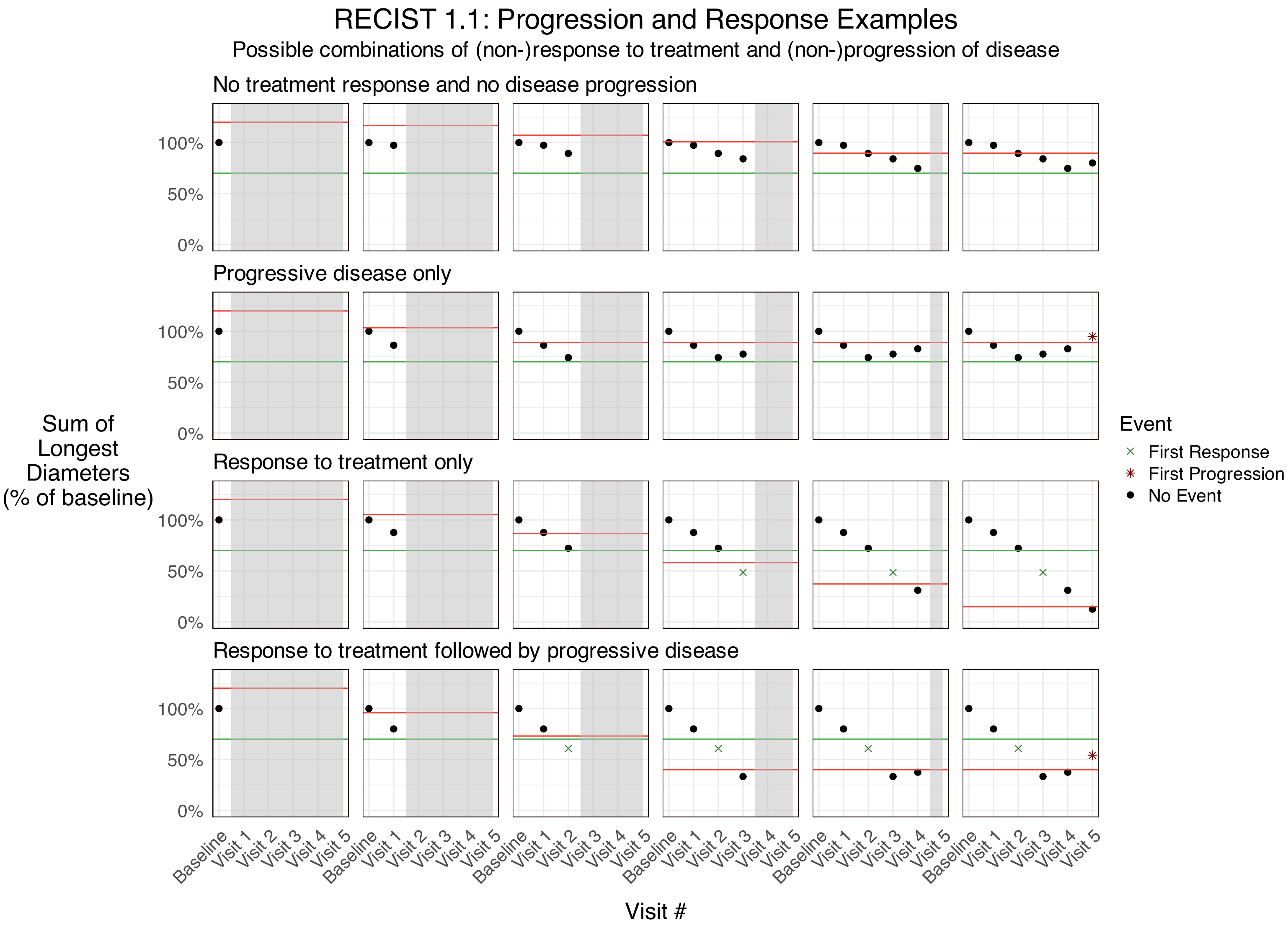


Figure 1.1.: Example of RECIST 1.1 Response Assessments Over Time

While the overview provided here covers the core elements of RECIST 1.1, it necessarily omits some nuances addressed in the complete guidelines such as specific criteria for lymph nodes, handling of non-measurable lesions, and special considerations for certain imaging modalities. Additional technical specifications in the original publication (1) include detailed criteria for lymph node assessment, guidance for handling non-measurable or non-evaluable lesions, and special considerations for particular imaging modalities. Nevertheless, the framework described above establishes the essential foundation required for understanding RECIST's application in clinical trials and, more specifically, for interpreting the reliability analyses that form the central focus of this thesis.

## 1.3. Towards Treatment: Clinical Trials

With some background on cancer and knowledge of the RECIST criteria in mind, we can now turn to the specific context of clinical trials and the role of tumor response assessment in evaluating cancer treatments. This section provides an overview of the importance of clinical trials, how success is measured in those trials, and the regulatory requirements for and difficulties in tumor response assessment.

### 1.3.1. Clinical Trials as the Gateway to Therapies

Clinical trials serve as the critical gateway through which new cancer treatments must pass before reaching patients. These systematic studies provide the controlled environment necessary to evaluate safety, efficacy, and optimal dosing of novel therapies (22). Regulatory bodies including the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have established rigorous frameworks governing the conduct of these trials, emphasizing not only scientific validity but also ethical considerations and patient safety (22,28). These regulatory requirements underscore the fundamental role of clinical trials as the sole legitimate pathway for new cancer therapies to gain approval and enter standard clinical practice.

The development of cancer treatments represents a substantial investment of time, expertise, and financial resources. A new therapeutic agent typically undergoes many years of laboratory research, preclinical testing, and multiple phases of clinical trials before receiving regulatory approval (29). This extensive development timeline makes the accuracy and reliability of trial endpoints particularly crucial; inefficiencies or errors in measuring treatment effects can significantly delay or even derail the development of potentially beneficial therapies. Phase II and Phase III trials are especially critical in this process, as they provide the primary evidence for efficacy and safety that supports regulatory decision-making (30). These later-phase trials rely heavily on standardized assessment criteria like RECIST to reliably measure treatment responses consistently across diverse clinical settings.

Given that regulatory approval hinges on demonstrating meaningful clinical benefit, measurement tools must be both precise and reliable to distinguish genuine therapeutic effects from artifacts or natural disease variation (21,31). Standardized frameworks like RECIST provide this essential foundation by enabling consistent assessment across different investigators, trial sites, and patient populations while facilitating regulatory review through a common evaluation language.

### 1.3.2. Measuring Treatment Success & Surrogate Endpoints

The primary objective of any cancer treatment is to improve patient outcomes, with overall survival (i.e. the length of time a patient lives after starting treatment) representing the most definitive measure of therapeutic benefit (32,33). Despite its unequivocal clinical relevance, overall survival presents significant methodological challenges as a primary endpoint in clinical trials. Studies using overall survival as their primary endpoint often require extended follow-up periods as they necessarily aim to capture the long-term effects of treatment, which substantially increases trial duration, cost, and complexity (32). Such extended timelines are frequently incompatible with the urgent need to bring effective therapies to patients and the economic realities of drug development (32). Furthermore, the impact of subsequent treatments after disease progression can confound the interpretation of overall survival data, potentially obscuring the true effect of the investigational therapy.

Given these limitations, surrogate endpoints have become essential tools in oncology research, providing earlier signals of efficacy that can accelerate therapeutic development and regulatory decision-making (34). These surrogate measures serve as proxies for clinical benefit that can be assessed more rapidly than overall survival. Common surrogate endpoints include outcomes reliant on measurements like tumor response and patient-reported quality-of-life (34,35). Importantly, several RECIST-based metrics have emerged as particularly valuable surrogate endpoints in oncology trials. Progression-Free Survival

(PFS) measures the time patients live without disease worsening; Objective Response Rate (ORR) quantifies the proportion of patients achieving complete or partial tumor shrinkage; Disease Control Rate (DCR) captures those with complete response, partial response, or stable disease; Time to Progression (TTP) documents the interval from treatment initiation to disease advancement; and Duration of Response (DoR) records how long responses are maintained before progression or death (34). These endpoints are often employed as primary or secondary endpoints in solid tumor clinical trials depending on what patient outcomes the trial hopes to optimize (33).

While surrogate endpoints facilitate more efficient drug development, their use requires careful interpretation (33). The correlation between surrogate endpoints and meaningful clinical outcomes like overall survival varies across cancer types, treatment modalities, and patient populations (33). A treatment may demonstrate impressive tumor shrinkage yet fail to extend survival or improve quality of life. Such a disconnect underscores the need for cautious assessment of surrogate measures. Nevertheless, regulatory agencies including the FDA and EMA generally accept well-validated surrogate endpoints as components of the evidence package supporting approval, particularly in Phase II and Phase III trials. These later-stage trials, involving hundreds to thousands of participants, rely heavily on standardized assessment frameworks to ensure consistent measurement and interpretation of key endpoints.

Tumor response as measured by RECIST has been validated as a key indicator of treatment efficacy and predictor of survival in clinical practice. This validation provides the foundation for RECIST's central role in defining and measuring the surrogate endpoints described above. However, the reliability of these RECIST-derived endpoints depends critically on accurate and consistent tumor assessment. Measurement challenges such as technical issues with imaging, observer-related variability, and patient-related factors can compromise endpoint validity. Inaccurate response assessment can lead to misclassification of treatment efficacy, inappropriate treatment decisions, and ultimately impact patient outcomes. Given that regulatory approval and clinical adoption of new therapies often hinge on these surrogate measures, ensuring their reliability through standardized assessment criteria becomes a matter of paramount importance.

### 1.3.3. RECIST as the Regulatory Standard

Regulatory authorities worldwide have embraced RECIST 1.1 as a preferred methodology for tumor response assessment in clinical trials. Both the FDA and EMA explicitly recognize RECIST in their guidance documents, with the FDA publishing non-binding recommendations that specifically reference RECIST as an established approach for evaluating tumor response in solid tumors (36,37). While these agencies do not *mandate* RECIST's use, they strongly encourage standardized and validated methodologies that ensure consistency across trial sites and facilitate regulatory review (36). RECIST has become the de facto standard precisely because it provides a framework that satisfies regulatory expectations for objective, reproducible, and clinically meaningful endpoints (38).

The widespread regulatory acceptance of RECIST stems from multiple factors. First, standardized criteria like RECIST enable efficient regulatory review by providing a common framework for comparing results across different studies and therapeutic agents (38,39). Second, RECIST facilitates international harmonization, ensuring consistency across multinational trials and regulatory submissions to different authorities (1). Finally, regulatory agencies expect robust quality assurance in clinical trials, and the FDA helped mold RECIST as it was consulted during development (1).

While RECIST enjoys broad regulatory support and validation, important questions persist regarding its systematic validity in certain contexts. The framework's reliance on unidimensional measurements may not fully capture complex response patterns (e.g. volumetric changes), particularly with novel therapies that may induce changes in tumor density or vascularity without necessarily affecting the tumor size. Furthermore, the arbitrary nature of the threshold values used to define response categories, namely the 30% reduction for partial response and 20% increase for progressive disease, lacks explicit biological rationale (31). These inherent limitations underscore the need for ongoing critical evaluation of RECIST's performance across different treatment modalities and cancer types, even as it remains the regulatory standard for response assessment in oncology trials.

### 1.3.4. Operational Challenges of RECIST in Clinical Trials

Despite RECIST's standardized framework, significant challenges persist in its practical implementation across clinical trials (39). Even when adhering to RECIST guidelines, rater variability remains a substantial concern, with multiple factors contributing to inconsistent assessments (40). These challenges can be broadly categorized into technical, observer-related, and lesion-specific factors. Technical factors include variations in imaging protocols, slice thickness, and contrast timing, which can significantly affect how tumors appear and are measured (1). Observer-related factors encompass differences in experience level, training background, and specialty expertise among evaluators (40,41). Lesion characteristics such as size, location, morphology, and enhancement patterns further complicate consistent assessment. Additionally, measurement methodology differences including the use of manual versus automated tools and variations in software systems can introduce another layer of potential inconsistency in tumor response evaluation (42). A particularly significant source of variability lies in the selection of target lesions at baseline (43,44). Since RECIST allows for the designation of up to five target lesions from potentially numerous eligible lesions, different evaluators may select different subsets of tumors for measurement. This initial divergence cascades throughout the assessment process, as subsequent measurements and response classifications are directly tied to these baseline selections. These challenges highlight the need for standardized imaging protocols, consistent training programs for investigators, and ongoing quality control processes to ensure assessment reliability.

In the context of clinical trials, a critical distinction exists between assessments performed by site investigators (also called “local” or “enrolling” investigators) and those conducted through blinded independent central review (BICR) (45). Site investigators are typically the clinicians directly involved in patient care and treatment decisions, while BICR involves independent radiologists or oncologists who review imaging data without knowledge of treatment allocation. (Of note, a single study may have many site investigators, each of whom may assess the same patient independently; we will generally refer to them in the singular form “site investigator” in our Methods and Results section for simplicity.) This separation is intended to minimize bias and ensure objective evaluation of treatment response (39).

Beaumont et al. (46) examined this distinction in a Phase II study, finding notable differences in the determination of progressive disease between site investigators and central reviewers. These differences were not merely random variations but reflected systematic disparities in how the two groups applied RECIST criteria. Ford et al. had previously observed that BICR workflow processes are “specifically intended to produce greater consistency in image interpretation” compared to site investigator assessments, typically employing two independent BICR raters with an adjudicator to resolve discrepancies which may produce better results than site investigators (45). Ford's comprehensive analysis identified multiple sources of variability between site and central assessments, including differences in training protocols, potential treatment bias among site investigators who may have clinical knowledge of patients, varying experience with different tumor types, and disparate interpretations of RECIST guidelines (see Table 1 from Ford for a detailed overview of these variability sources) (45).

While BICR is generally considered the gold standard for response assessment due to its structured approach and reduced potential for bias, it is not always feasible due to substantial cost and time constraints (47). Consequently, some trials rely exclusively or partly on site investigator assessments, which might introduce variability and bias into response evaluations (48). This disparity is particularly pronounced in target lesion selection, where systematic differences may emerge between specialized radiologists (i.e. central reviewers) and clinicians with less imaging experience (i.e. site investigators) (45). Furthermore, site investigators may be influenced by treatment knowledge in open-label studies, potentially affecting their assessments in ways that blinded central reviewers would not experience (48).

The extent of such differences between site and central reviewers has been empirically investigated in several key studies. Zhang et al. (48) conducted a meta-analysis of 76 phase III randomized clinical trials (RCTs) of anticancer agents for solid tumors that included assessments from both site investigators and central reviewers. While their analyses found no systematic differences between site and central reviewers across the four trial endpoints (ORR, DRR, PFS, and TTP) they examined, they observed that statistically inconsistent inferences could be made in nearly a quarter of the trials depending on which assessment source was used. This suggests that although differences may not be systematic across the entire landscape of oncology trials, they can nevertheless have significant implications for specific studies and endpoints. Corroborating these findings, Jacobs et al. (49) examined 24 phase II and III RCTs of anticancer agents for solid tumors specifically in breast cancer. Their investigation reached essentially the same conclusion as Zhang et al., though with the notable limitation that Jacobs' work focused exclusively on PFS as the endpoint of interest, leaving questions about other important RECIST-derived endpoints unaddressed.

## 1.4. Research Gaps Restated and Aims of the Thesis

The preceding sections have established that while RECIST 1.1 provides a standardized framework vital for tumor response assessment in clinical trials, significant reliability challenges persist in its implementation. Despite this widespread regulatory acceptance and empirical reliability analyses done by the likes of Zhang et al. and Jacobs et al., some questions persist about the consistency of RECIST interpretations across different raters, particularly between site investigators and central reviewers (48,49).

Several specific knowledge gaps regarding RECIST reliability can thus be seen in the literature covered in this introduction. First, despite numerous individual studies examining inter-rater reliability in RECIST assessments, no comprehensive meta-analysis has synthesized these findings to establish baseline expectations for agreement levels across raters in any context. Second, while some studies have noted differences between site investigator and central reviewer assessments, systematic analyses of how these differences affect various trial endpoints remain limited in terms of the endpoints they have analyzed including the absence of TTR and DoR. Third, the impact of RECIST's arbitrary threshold values, namely the 30% reduction defining partial response and 20% increase indicating progression, on inter-rater reliability has not been comprehensively evaluated. These thresholds, though widely accepted, lack explicit biological rationale and may contribute to assessment variability when measurements fall near these cutoff points.

Addressing these knowledge gaps has significant clinical relevance for oncology trials. Improved understanding of RECIST reliability can enhance the accuracy of tumor response assessment, potentially reducing measurement error and increasing confidence in trial outcomes. Additionally, quantifying the extent and patterns of disagreement between site investigators and central reviewers could inform more efficient trial designs and monitoring practices. Finally, evaluating the impact of threshold values on response classification may guide future refinements to RECIST criteria, particularly as novel treatment modalities with atypical response patterns become more prevalent in oncology.

This thesis therefore aims to systematically address these gaps through three primary research objectives. First, we will conduct a comprehensive meta-analysis of existing inter-rater reliability studies using RECIST criteria to establish baseline agreement expectations. Second, we will analyze several trial endpoints to compare site investigator assessments with central review outcomes, examining whether systematic differences exist in response classification. As a corollary to these analyses, one would expect that site investigators and central reviewers will not differ at all in an ideal scenario, and we thus also conduct equivalence testing to formally assess this hypothesis on the same endpoints. Finally, we will evaluate how varying the threshold values in RECIST criteria affects inter-rater reliability and classification stability, potentially identifying cutpoints that maximize (dis-)agreement between raters.

With this background established, we proceed to the Methods chapter, which outlines the research design and methodology employed to address these questions.

# 2. Methods and Materials

This section is broadly divided into two parts: the first part describes the data sources and how the data for this study are sourced, while the second part details the statistical analyses performed on the data. The methods and materials described here are designed to ensure transparency, reproducibility, and rigor in the research process, adhering to best practices in meta-analyses, and in scientific reporting more broadly.

## 2.1. Data Sources

Data for this study are sourced from two primary domains: a systematic review of the literature and proprietary data extracted from clinical trials. The literature review focuses on peer-reviewed studies reporting inter-rater reliability metrics for RECIST 1.1 assessments, while the clinical trial data are obtained from three specific trials that utilize RECIST 1.1 criteria for evaluating treatment responses made available through TransCelerate BioPharma Inc.'s DataCelerate™ platform (50,51). The trials are selected based on their use of RECIST 1.1 criteria, the availability of assessments from both site investigators and central reviewers, and the inclusion of data on target lesions, non-target lesions, new lesions, and SLD measurements. Details for both data sources are provided below.

### 2.1.1. Literature Search

A systematic search of the literature is undertaken to identify peer-reviewed studies that report inter-rater reliability metrics for assessments conducted using RECIST 1.1. The search is designed to comprehensively capture studies that quantitatively evaluate the degree of agreement between multiple independent raters, employing established statistical measures of inter-rater reliability.

#### 2.1.1.1. Search Strategy & Study Eligibility

The primary search was conducted in the PubMed/MEDLINE database using a combination of Medical Subject Headings (MeSH) and free-text terms (52). The search strategy encompasses two core conceptual domains: employment of RECIST criteria in the study and inter-rater reliability. Searches were conducted with Essie syntax to optimize sensitivity to identifying relevant studies (53). The search terms included combinations of expressions including "RECIST inter-rater reliability", "(RECIST OR Response Evaluation Criteria in Solid Tumors)", and "(Cohen's kappa OR Fleiss' kappa OR kappa)"

In addition to the primary search, targeted supplementary searches were carried out to further enhance the comprehensiveness of the review and capture potentially relevant studies not indexed in PubMed. A supplementary search of Google Scholar was performed using the keywords "RECIST interrater reliability", and the first 50 results retrieved were screened for eligibility, with attention to studies published in non-MEDLINE indexed journals.

Eligibility criteria were defined a priori to ensure the systematic and unbiased inclusion of relevant studies. To be included, studies must be published in peer-reviewed journals, report quantitative inter-rater reliability statistics for RECIST 1.1 assessments (either Cohen's $\kappa$ or Fleiss' $\kappa$), and involve two or more independent qualified raters (i.e. medical doctors) evaluating identical imaging datasets. Only studies that provide sufficient methodological detail to enable data extraction and critical appraisal were considered. Furthermore, only articles published in English and after the introduction of RECIST 1.1 in 2009 are eligible (1).

Studies were excluded if they employ the outdated RECIST 1.0 criteria, if they employ variants of RECIST such as mRECIST (54) and iRECIST (55), if they report only descriptive rather than statistical measures of agreement, and if they only provide continuous measures of inter-rater reliability such as intra-class correlation coefficients. In addition, studies lacking sufficient detail regarding statistical methodology or reliability calculation procedures, studies not reporting inter-rater reliability metrics, and duplicated studies are all excluded.

#### 2.1.1.2. Study Screening and Selection Process

To ensure transparency and reproducibility in the identification, screening, and selection of studies, we follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (56). The study selection process is documented using a PRISMA flow diagram, which details the number of records identified, screened, assessed for eligibility, and included in the final analysis, as well as the reasons for exclusion at each stage. This approach provides a standardized and rigorous framework for reporting systematic reviews and meta-analyses.

All searches were conducted between April 1 and April 2, 2025. The study selection process was carried out in two stages to minimize bias and ensure comprehensive identification of eligible studies. In the first stage, titles and abstracts of all retrieved records were independently screened by PC to determine potential eligibility. Records that clearly meet inclusion criteria (i.e. because $\kappa$ is reported in the article abstract) or whose eligibility cannot be determined based on the abstract alone were retained for full-text evaluation.

In the second stage, the full texts of potentially eligible studies were assessed in detail against the predefined inclusion and exclusion criteria. Studies that meet all inclusion criteria were incorporated into the final dataset. In addition to database searching, the reference lists of all included studies were manually reviewed to identify any additional relevant articles not captured by the primary or supplementary search strategies. The entire study selection process was documented and reported in Figure 2.1 in accordance with the PRISMA guidelines to ensure transparency and reproducibility. The diagram was produced using the {PRISMA}[1] R Package (57).

#### 2.1.1.3. Data Extraction

For data extraction, we used a standardized form developed specifically for the purposes of this review. Extracted study characteristics include bibliographic information (author, year, and database), number of raters, the type of inter-rater reliability statistic reported (Cohen's $\kappa$ or Fleiss' $\kappa$), the estimate for the reported statistic, and the corresponding standard error or confidence interval. The context in which the study is conducted (e.g. clinical trial, retrospective study, etc.) was also collected, and summarized as either being a Clinical Trial or Other Study as the only major distinction of relevance for this paper is whether

[1] It is customary in the R ecosystem to denote the name of a package by enclosing it in curly braces e.g. {packageName}. We will follow this convention throughout the thesis.

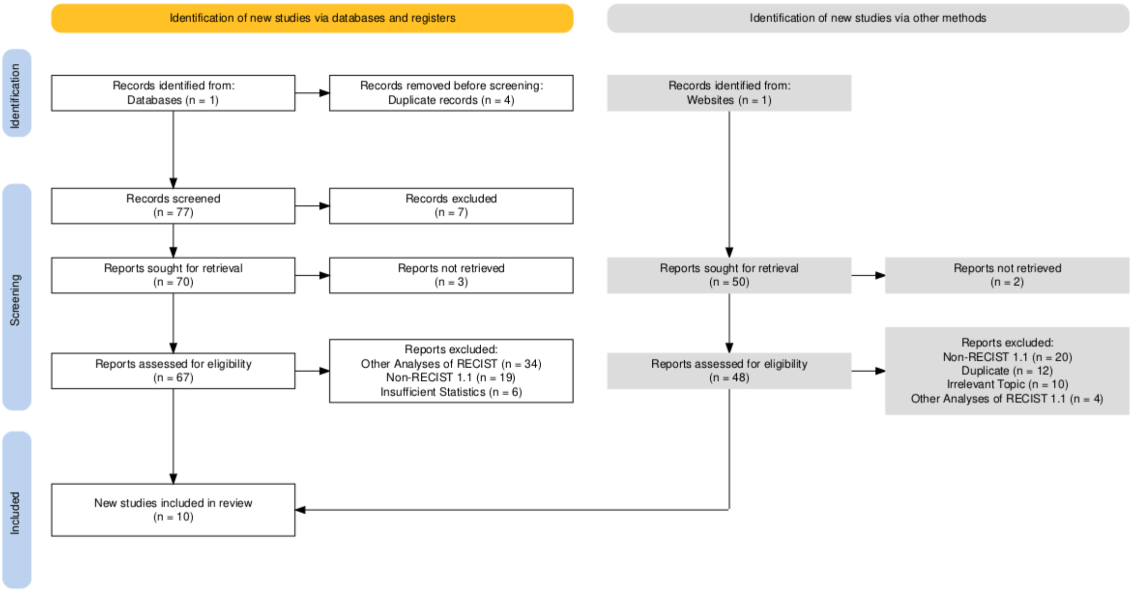

Identification of new studies via databases and registers
Identification of new studies via other methods
Identification
Screening
Included
Records identified from:
Databases (n = 1)
Records removed before screening:
Duplicate records (n = 4)
Records identified from:
Websites (n = 1)
Records screened
(n = 77)
Records excluded
(n = 7)
Reports sought for retrieval
(n = 70)
Reports not retrieved
(n = 3)
Reports sought for retrieval
(n = 50)
Reports not retrieved
(n = 2)
Reports assessed for eligibility
(n = 67)
Reports excluded:
Other Analyses of RECIST (n = 34)
Non-RECIST 1.1 (n = 19)
Insufficient Statistics (n = 6)
Reports assessed for eligibility
(n = 48)
Reports excluded:
Non-RECIST 1.1 (n = 20)
Duplicate (n = 12)
Irrelevant Topic (n = 10)
Other Analyses of RECIST 1.1 (n = 4)
New studies included in review
(n = 10)

Figure 2.1.: PRISMA flow diagram of study selection for inclusion in the meta-analysis.

or not the study data is derived from a clinical trial. The term “Other Study” is chosen to describe the remaining studies as they are done in a mix of prospective (40,58,59) and retrospective (41,44,60–64) designs across 1 to 3 sites, but such study design elements are not otherwise of significance to this review.

When studies fail to directly report sufficient details of their inter-rater reliability statistics, namely the standard error of the estimate, these values are derived from the reported 95% confidence intervals using standard statistical formulas. Studies that neither report confidence intervals nor standard errors are necessarily excluded from the meta-analyses.

### 2.1.2. Trial Data

Potential clinical trials were identified by querying the ClinicalTrials.gov API v2.0.3 “studies” endpoint (65), specifically targeting the `query.outc` field, which allows searching for both primary and secondary outcome measures. This field was queried using Essie syntax (66) for the string: “RECIST OR response evaluation criteria in solid tumors OR sum of longest diameters OR SLD”, to identify studies that directly or indirectly referenced RECIST criteria in their primary or secondary outcomes. We did not restrict the search criteria to strictly “RECIST” or “response evaluation criteria in solid tumors” because it was noted in preliminary searches that some studies only referenced the terms “sum of longest diameters” and “SLD” while implicitly using the RECIST rating criteria.

Studies identified via querying the ClinicalTrials.gov API were then cross-referenced with those available through TransCelerate BioPharma Inc.'s DataCelerate™ platform. TransCelerate is a non-profit organization promoting collaboration across biopharmaceutical companies, with approximately 20 major pharmaceutical partner companies (67). The DataCelerate™ platform in particular aims to facilitate the sharing of high-quality historical trial data (HTD) across companies in a manner compliant with U.S. Code of Federal Regulations privacy laws as well as adhering to other globally relevant data privacy laws (51); in other terms, it provides a secure, standardized environment ensuring regulatory compliance and protects patient confidentiality while enabling collaborative research and secondary analyses across the pharmaceutical industry. Of note, the data available for this project via DataCelerate™ is exclusively placebo group/standard of care (PSOC) data, also commonly referred to as "control group" data (51).

The subset of studies identified by cross-referencing results from ClinicalTrials.gov against DataCelerate™ undergo manual review to confirm the use of RECIST 1.1 criteria, the availability of RECIST assessments from both site investigators and central reviewers, and the inclusion of data on target lesions, non-target lesions, new lesions, and SLD measurements. Three trials meet all inclusion criteria: NCT02395172, NCT03434379, and NCT03631706. Two of these trials investigate treatments for non-small cell lung cancer, while the third focuses on hepatocellular carcinoma. Detailed trial characteristics are summarized in Table 2.1.

| Trial ID | NCT02395172 | NCT03434379 | NCT03631706 |
|---|---|---|---|
| **Cancer** | Non-Small Cell Lung Cancer | Hepatocellular Carcinoma | Non-Small Cell Lung Cancer |
| **Label** | Open | Open | Open |
| **Phase** | Phase 3 | Phase 3 | Phase 3 |
| **Control Group Size** | n=329 | n=128 | n=146 |
| **Control Group Treatment** | Docetaxel (cytotoxic) | Sorafenib (targeted cytostatic) | Pembrolizumab (immunotherapy) |
| **Population Notes** | NSLCC patients what have progressed disease | Advanced or metastatic HCC patients | NSLCC patients |
| **Locations** | 260 international sites | 119 international sites | 119 international sites |

Table 2.1.: Characteristics of Included Clinical Trials

The therapeutic agents in these trials represent three distinct mechanisms of action: cytotoxic treatments that induce cell death primarily through chemotherapy (68), cytostatic agents that inhibit cell division (68), and immunotherapies that activate targeted immune responses (69). This distinction is methodologically significant as these different modalities produce distinct response patterns that may affect RECIST assessment validity. For example, immunotherapies can cause pseudoprogression due to immune cell infiltration before actual tumor response occurs, potentially leading to misclassification under standard RECIST criteria (21).

For each clinical trial, data are structured in accordance with the Study Data Tabulation Model (SDTM) as defined by the Clinical Data Interchange Standards Consortium (CDISC) (70). The data standards developed by CDISC ensure consistency and regulatory compliance in the submission of clinical trial data, and also helps facilitate interoperability and data analyses within and across institutions analyzing trial data. Specifically, the tumor (TU), tumor results (TR), and disease response (RS) domains are employed to capture and organize data related to tumor assessments and treatment responses.

The TU domain is used to uniquely identify tumors and lesions, specifically their classification as target, non-target, or new lesions, as per RECIST 1.1 guidelines (71). The TR domain records quantitative and qualitative measurements of these identified tumors, such as size and progression status, which are essential for calculating the SLD and assessing changes over time (71). The RS domain captures the overall response evaluations (e.g., complete response, partial response, stable disease, or progressive disease) based on the RECIST criteria (72), and is largely derived from the TU and TR domains. These domains work in concert to provide a comprehensive and traceable representation of tumor burden through monitoring of individual tumors and the SLD of those tumors, as well as treatment efficacy by way of deriving outcomes according to the RECIST criteria (73).

Data extracted and/or derived from the TU, TR, and RS SDTM domains include SLD measurements and RECIST 1.1 assessments for overall response, target lesions, non-target lesions, and new lesions. These data are available at scheduled imaging time points as defined by each study protocol. At each of these time points, imaging assessments are independently performed by three raters: one on-site investigator and two central reviewers.

## 2.2. Statistical Analyses

The statistical analyses are divided into three major sections reflecting the three main lines of enquiry of this thesis. First, the procedures for the IRR meta-analyses are covered including a review of Cohen's and Fleiss' $\kappa$, the mathematical bases for transforming the IRR data, and details of how the meta-analyses are conducted and interpreted. Second, we present a multi-part analysis of differences between site investigators and central reviewers for the trial endpoints ORR, TTP, TTR, and DoR. Finally, this chapter concludes by describing the sensitivity analysis which aims to elucidate whether differences between site investigators and central reviewers in trial outcome measures has a dependency on the disease progression and response thresholds defined by RECIST. A helpful framing question to keep in mind while reading the methods and results for the sensitivity analyses is as follows: "if we change the response and progression thresholds, how much more or less do the site investigators and central reviewers agree with one another?"

### 2.2.1. Inter-Rater Reliability Meta-Analysis

A meta-analysis is a statistical technique that combines the results of multiple independent studies to produce a single summary estimate of an effect or association (74). By synthesizing data across studies, meta-analyses increase statistical power, improve estimates of effect size, and help identify patterns or sources of heterogeneity that may not be apparent in individual studies (74). This approach is especially valuable in fields where individual studies may be small or yield conflicting results, as it provides a more comprehensive and objective assessment of the evidence. Meta-analyses typically employ either fixed-effect or random-effects models, with the former assuming that all studies estimate the same underlying effect size, while the latter allows for variability between studies in terms of populations, methodologies, and other factors (74). The choice of model depends on the research question, the degree of heterogeneity among studies, and the assumptions about the data.

As such we conducted a series of random-effects meta-analyses focusing on inter-rater reliability (IRR) statistics in order to evaluate the consistency of RECIST 1.1 assessments across different raters of the same data. Specifically, we analyzed Cohen's $\kappa$, Fleiss' $\kappa$, and a combined set of both measures, with all analyses restricted to the Overall RECIST 1.1

Outcome, which was generally the only available endpoint across the studies identified in the literature search.

#### 2.2.1.1. IRR Measures Review: Cohen's and Fleiss' $\kappa$

Cohen's kappa ($\kappa$) is a statistical measure assessing the degree of agreement between *two* raters/observers who each classify the same items. Unlike a simple percentage agreement measure, Cohen's $\kappa$ accounts for agreement between raters that occurs by chance alone which provides a more robust measurement of rater agreement (75). The calculation of Cohen's $\kappa$ is given by (75):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.1}$$

where

$$p_o = \sum_{i=1}^{k} p_{ii} \tag{2.2}$$

is the observed proportion of agreement and $p_{ii}$ denotes the proportion of items for which both raters independently assigned category $i$, and $k$ is the total number of categories. Likewise,

$$p_e = \sum_{i=1}^{k} p_{iA} \cdot p_{iB} \tag{2.3}$$

is the expected proportion of agreement by chance, with $p_{iA}$ is the proportion of items that rater A assigns to category $i$, $p_{iB}$ is the proportion of items that rater B assigns to category $i$, and $k$ is the number of categories.

Interpretation of the corresponding values is relatively straightforward as $\kappa$ is mathematically bounded on the interval $[-1, 1]$ with positive values indicating rater agreement and negative values indicating disagreement (76). However, observed values are typically assumed to be on the range $(0, 1]$ as this is the range in which raters show partial agreement up until perfect agreement at $\kappa = 1$ (76). Approximate qualitative interpretations of values on this range have been proposed by Landis et al. (77), and are reproduced in Table 2.2 below.

| Kappa | Interpretation |
|---|---|
| <0.00 | Poor agreement |
| 0.00-0.20 | Slight agreement |
| 0.21-0.40 | Fair agreement |
| 0.41-0.60 | Moderate agreement |
| 0.61-0.80 | Substantial agreement |
| 0.81-1.00 | Almost perfect agreement |

Table 2.2.: Interpretation of Cohen's Kappa Values according to Landis et al. (1977)

While Table 2.2 provides a useful framework for interpreting Cohen's $\kappa$ values, it is important to note that these benchmarks are arbitrary and may not apply universally across all contexts. The interpretation of $\kappa$ values can vary depending on the specific application, the nature of the data, and the clinical or research context in which the ratings are made. Therefore, while these benchmarks provide a general guideline and serve as the starting point for discussions around the IRR in this paper, they must be applied with caution and in conjunction with other contextual factors.

As with other statistical measures, the interpretation of Cohen's $\kappa$ should likewise be informed by the variability of the data (i.e. confidence intervals and standard errors) and the specific clinical or research context in which it is applied. In some cases, reviewing the marginal distributions of the data may provide additional insights into the nature of the source of (dis-)agreement between raters.

An important limitation of Cohen's $\kappa$ is its sensitivity to the strictness of category definitions, particularly when categories are conceptually similar or adjacent in clinical meaning. In the context of RECIST 1.1, for example, a complete response (CR) and a partial response (PR) are treated as distinct categories, even though both represent favorable treatment outcomes, as will be explored in our analyses of objective response rate. Because $\kappa$ penalizes all disagreements equally, however, it does not account for the practical or clinical proximity of such categories. As a result, disagreements between raters that involve adjacent or similar categories can disproportionately lower the $\kappa$ value, even if the distinction is of limited clinical impact. This limitation should be considered when interpreting $\kappa$ values in settings where the boundaries between categories may be somewhat arbitrary or where minor disagreements do not necessarily reflect meaningful differences in clinical judgment.

In addition to advocating for caution in interpreting Cohen's $\kappa$ values, it is also important to note that the statistic is only applicable for assessing agreement between *two* raters. When more than two raters are involved, Fleiss' $\kappa$ is typically employed as an extension of Cohen's $\kappa$ that allows for the assessment of agreement among multiple raters simultaneously (78). Fleiss' $\kappa$ is calculated using a similar formula to Cohen's $\kappa$ that calculates an average agreement between raters while adjusting for expected agreement, and the yielded values are also on the range, $[-1, 1]$ (78).

#### 2.2.1.2. Compiling Inter-Rater Reliability Estimates

Estimates of Cohen's and Fleiss' $\kappa$ statistics and their corresponding standard errors were extracted directly from the included studies when available. When standard errors were not reported, they were derived from the reported confidence intervals using standard statistical formulas assuming a normal distribution and a 95% confidence level. Specifically, the standard error (SE) was calculated as:

$$SE = \frac{CI_{upper} - CI_{lower}}{2 \cdot Z_{0.975}} \tag{2.4}$$

where $CI_{upper}$ and $CI_{lower}$ are the upper and lower bounds of the confidence interval, respectively, and $Z_{0.975}$ is the critical value from the standard normal distribution corresponding to a 95% confidence level (approximately 1.96).

To supplement the literature-based estimates, Fleiss' $\kappa$ was also calculated for the RECIST Overall outcome using data from the control arm data of the three clinical trials described previously. A standard R implementation for calculating the standard error ($SE$) of Fleiss' $\kappa$ could not be readily located[2], and recent research has shown variance calculations for Fleiss' $\kappa$ that are derived from the variance calculates for Cohen's $\kappa$ (78) may be inappropriate for constructing confidence intervals (79). Rather than seek out a package implementing a potentially incorrect calculation of the $SE$ of Fleiss' $\kappa$, $SE$ values for these trial-based estimates were obtained via non-parametric bootstrapping with 10,000 resamples. All IRR calculations for the clinical trial data were performed using the {irr} package in R (80).

[2] The R package {rel} apparently contained a function for calculating the standard error around Fleiss' $\kappa$, but the package has since been removed from CRAN.

#### 2.2.1.3. Logit Transformation of Kappa Values

A log transformation is frequently applied to effect size estimates in meta-analyses to (81), particularly for measures such as odds ratios, risk ratios, and hazard ratios. The primary rationale is that these effect sizes are inherently positive and often exhibit right-skewed distributions. Applying the natural logarithm (ln) to these values thus serves several purposes: it stabilizes the variance, normalizes the distribution, and converts multiplicative relationships into additive ones (81), which simplifies statistical modeling and interpretation. On the log scale, confidence intervals are more likely to be symmetric, and the assumption of normality underlying many meta-analytic methods is better satisfied.

However, log transformation is not universally appropriate. It cannot be applied to effect sizes that can take negative values or are bounded (such as correlation coefficients or $\kappa$ statistics), as the transformation may yield values outside the theoretical range or introduce interpretational difficulties. Additionally, back-transforming results to the original scale can sometimes complicate interpretation, especially for non-statistical audiences. In the case of Cohen's and Fleiss' $\kappa$, which are both bounded on the range $[-1, 1]$, a special case of the log-transformation, the logit, was applied to obtain many of the same benefits seen in a log transformation, including stabilizing variance and ensuring that confidence intervals remained within the theoretical bounds of the statistic. The transformation was defined as:

$$\text{logit}(\kappa) = \log\left(\frac{\kappa + 1}{1 - \kappa}\right) \tag{2.5}$$

The logit transformation is particularly well-suited for $\kappa$ statistics because direct application of a standard logarithmic transformation is not possible for values that can be negative or zero, nor does it preserve the bounded nature of the statistic (82). The logit function maps the interval $(-1, 1)$ onto the entire real line, removing the bounds and allowing for the use of statistical methods that assume unbounded, normally distributed variables. Perhaps most importantly in the context of performing meta-analyses, this transformation ensures that confidence intervals, when back-transformed, remain within the theoretical limits of the $\kappa$ statistic. As a result, the logit transformation enables more accurate and interpretable meta-analytic inference for bounded agreement measures.

To estimate the standard error of the logit-transformed kappa, the delta method was used (83). In simple terms, the delta method provides a way to estimate the standard error of a transformed statistic by using the standard error of the original statistic and the slope (derivative) of the transformation (84). It works by approximating how much uncertainty in the original value "carries over" after applying a mathematical function, such as the logit. This allows us to calculate confidence intervals for transformed values, even when the transformation is non-linear (84). Our approach is similar to Carpentier et al.(82) who applied such a transformation to Cohen's $\kappa$ data, but they assumed bounding of $\kappa$ from 0 to 1. We followed the same approach, but with the more general bounding of $[-1, 1]$ to account for the fact that $\kappa$ can take negative values, and our steps were informed by Muche et al. (84).

If $\hat{\theta}$ is an estimator of a parameter $\theta$ and $g(\cdot)$ is a differentiable function, then:

$$\mathrm{Var}(g(\hat{\theta})) \approx (g'(\theta))^2 \cdot \mathrm{Var}(\hat{\theta}) \tag{2.6}$$

In our case, we are ultimately interested in the standard error of the logit-transformed $\kappa$ estimate, which can be obtained by simply taking the square root of the variance:

$$\mathrm{SE}[g(\hat{\theta})] \approx |g'(\theta)| \cdot \mathrm{SE}(\hat{\theta}) \tag{2.7}$$

The derivative of the logit transformation is given by:

$$\begin{aligned} g'(\kappa) &= \frac{d}{d\kappa} \log\left(\frac{\kappa+1}{1-\kappa}\right) \\ &= \frac{d}{d\kappa}\left[\log(\kappa+1) - \log(1-\kappa)\right] \\ &= \frac{1}{\kappa+1} + \frac{1}{1-\kappa} \\ &= \frac{(1-\kappa)+(\kappa+1)}{(\kappa+1)(1-\kappa)} \\ &= \frac{2}{1-\kappa^2} \end{aligned} \tag{2.8}$$

Substituting this into the equation for the standard error of the logit-transformed $\kappa$ gives us:

$$\mathrm{SE}[g(\hat{\theta})] = \frac{2}{1-\kappa^2} \cdot \mathrm{SE}(\hat{\theta}) \tag{2.9}$$

And replacing $\mathrm{SE}(\hat{\theta})$ with $SE_\kappa$, the standard error of the original $\kappa$ estimate, we obtain the final formula for the standard error of the logit-transformed $\kappa$:

$$SE_{\mathrm{logit}(\kappa)} = \frac{2 \cdot SE_\kappa}{1 - \kappa^2} \tag{2.10}$$

where $SE_\kappa$ is the standard error of the original $\kappa$ estimate and $\kappa$ is the original $\kappa$ estimate. This is the transformation applied to all $\kappa$ estimates prior to conducting the meta-analyses, and the back-transformation is applied to the pooled estimates and confidence intervals after fitting the random-effects model.

#### 2.2.1.4. Meta-Analysis Procedure

For each set of $\kappa$ estimates (Cohen's, Fleiss' and the combined set), meta-analyses were performed on the logit-transformed values to ensure that pooled estimates and confidence intervals remained within the theoretical bounds of the statistic. The logit transformation was applied to each study's $\kappa$ value using Equation 2.5 and its standard error using Equation 2.10, as described above. Meta-analyses were conducted using a random-effects model, which accounts for both within-study and between-study variability. This approach is appropriate given the expected heterogeneity in study populations, imaging protocols, and assessment procedures across the included studies. The estimate of the average effect from a random-effects analysis will be more conservative than that of a fixed-effect analysis, but provides a more appropriate, generalizable result given the heterogeneity of the included studies.

The random-effects model was implemented using the `rma()` function from the {metafor} package in R (85), specifying the Restricted Maximum Likelihood (REML) estimator to estimate the between-study variance ($\tau^2$). Each study's contribution to the pooled estimate was weighted by the inverse of its total variance (the sum of within-study and between-study variance components). After fitting the model, pooled logit-transformed estimates and their confidence intervals were back-transformed to the original $\kappa$ scale using the inverse logit function:

$$\kappa = \frac{e^x - 1}{e^x + 1} \tag{2.11}$$

Heterogeneity among studies was assessed using the $I^2$ statistic and Cochran's $Q$ test, both of which are standard outputs of the {metafor} package. The $I^2$ value quantifies the amount of variation in estimates that is attributable to between-study differences, with higher values indicating greater heterogeneity–that is, higher heterogeneity may be indicative of different effects or methodological differences between studies (86). Forest plots were generated to visually summarize individual study estimates, their confidence intervals, and the overall pooled effect, all presented on the original $\kappa$ scale.

Publication bias is assessed using both visual and statistical methods. First, we generate funnel plots for each meta-analysis, plotting the logit-transformed $\kappa$ estimates from individual studies against their corresponding standard errors. In the absence of publication bias, these plots are expected to display a symmetrical, inverted funnel shape, as studies with larger standard errors (typically smaller studies) should scatter widely at the bottom of the plot, while those with smaller standard errors (larger studies) cluster more narrowly at the top (87). Asymmetry in the funnel plot may suggest the presence of publication bias, such as the selective publication of studies with larger or more significant effect sizes, or other small-study effects (87). However, it is important to note that funnel plot asymmetry can also arise from sources unrelated to publication bias, including true heterogeneity between studies, methodological differences, or simple chance (86,87).

To formally test for funnel plot asymmetry, we applied Egger's regression test using the. Egger's test is a statistical method that evaluates whether there is a systematic relationship between study effect sizes and their standard errors, which would be indicative of small-study effects or publication bias (88). Specifically, Egger's test regresses the standardized effect sizes (effect size divided by its standard error) on the inverse of the standard error. A significant intercept in this regression suggests that smaller studies tend to report larger (or smaller) effect sizes than would be expected by chance, consistent with the presence of publication bias. The test provides a p-value for the null hypothesis that the intercept is zero (i.e., no asymmetry). While Egger's test increases the objectivity of publication bias assessment, it also has limitations: it may be underpowered when the number of studies is small (typically fewer than 10) (86), and it can be influenced by between-study heterogeneity or outliers. Therefore, results from Egger's test should be interpreted in conjunction with visual inspection of the funnel plot and consideration of the broader context of the included studies. Funnel plots and Egger's test results are generated using the {metafor} package in R (85).

### 2.2.2. Comparing Site Investigators vs. Central Reviewers

While the preceding section examined general agreement on RECIST in a variety of clinical trial and controlled study environments, the following section presents a comprehensive, multi-part analysis comparing RECIST 1.1 assessments made by site investigators and central reviewers in the three included clinical trials. Given the complexity and breadth of the data, the analyses are organized into a series of focused components, each addressing a distinct aspect of rater agreement and its implications for clinical trial outcomes.

First, we assess the general agreement between raters regarding the RECIST Overall outcome. In opposition to the Fleiss' $\kappa$ values used previously as an average across **all** raters, pairwise Cohen's $\kappa$ coefficients were calculated to check whether particular pairs of raters exhibited higher or lower agreement with one another. Following this, we aimed to begin quantifying to what degree raters disagree with each other by conducting linear mixed effects models treating the RECIST outcome as an ordinal variable, with rater as a fixed effect predictor and random intercepts specified for each subject to account for repeated measures. To then explore pairwise differences between raters, we conducted post-hoc comparisons of estimated marginal means (EMMs) between the raters. These analyses provide a preliminary assessment of rater-specific differences in RECIST outcome assignment, and help identify whether systematic biases exist in the classification of treatment response between site investigators and central reviewers although they do not explore in what ways the raters disagree with each other.

Second, we examine the ORR, a key binary endpoint in oncology trials, to determine whether systematic differences exist in the classification of treatment response between site investigators and central reviewers. This analysis provides further insight into potential discrepancies in response categorization, and is more nuanced than the preceding analyses as it provides a direct comparison of the proportion of patients classified as responders (complete or partial response) by each rater group. Moreover, these analyses of ORR begin to reveal whether differences identified in step one materially affect trial outcomes or patient-level endpoints. To assess overall differences in ORR detection among raters, we used Cochran's Q test as an omnibus test for matched binary data, followed by post-hoc pairwise McNemar's tests with within-study Bonferroni p-value adjustments to identify specific differences between raters within each study.

Third, we extend our investigation to time-to-event outcomes, including Time to Response (TTR), Time to Progression (TTP), and Duration of Response (DoR). By leveraging survival analysis techniques, we assess not only whether raters agree on the occurrence of key clinical events, but also whether they differ in the timing of detecting these events. This approach allows for a more nuanced evaluation of how rater variability may influence the interpretation of trial efficacy. Moreover, such modeling is of particular importance given that RECIST 1.1 assessments are often used to derive these time-to-event outcomes, and discrepancies in rater evaluations could lead to significant differences in the estimated timing of responses or progression in the context of clinical trials and patient care.

While our analyses of ORR and TTP build upon the methodological framework established by Zhang et al. (48), our examination of TTR and DoR represents a novel contribution to the literature. These latter time-to-event outcomes have not been previously investigated in the context of RECIST 1.1 inter-rater reliability, despite their clinical significance in oncology trials. By including these additional endpoints, we provide a more comprehensive assessment of how rater discrepancies may impact the full spectrum of clinically relevant outcome measures derived from RECIST evaluations.

Finally, we synthesize the results of these comparisons across trials using meta-analytic methods. This meta-analysis is designed to quantify the overall magnitude and direction of rater differences, while accounting for between-study heterogeneity. Importantly, our meta-analytic approach incorporates both traditional null hypothesis significance testing (NHST) as well as formal equivalence testing (using the Two One-Sided Tests, or TOST, procedure). This dual approach enables us to distinguish between statistically significant differences, statistical equivalence, and inconclusive findings, providing a rigorous and interpretable summary of rater agreement on time-to-event outcomes.

Together, these analyses offer a detailed and methodologically robust assessment of the impact of rater discrepancies on key clinical trial endpoints, with direct implications for the design, conduct, and interpretation of RECIST-based studies.

#### 2.2.2.1. Preliminary Agreement Analyses Between Site and Central Reviewers

To explore potential differences in RECIST 1.1 assessments between site investigators and central reviewers, a series of analyses were conducted using data from the three clinical trials described previously. These analyses focused exclusively on the Overall RECIST 1.1 outcome and do not include data from the literature review.

The first component of this analysis focuses on the general agreement between raters regarding the RECIST Overall outcome. Pairwise inter-rater reliability (IRR) was calculated using Cohen's $\kappa$ coefficient for each combination of raters: site investigator vs. central reviewer 1, site investigator vs. central reviewer 2, and central reviewer 1 vs. central reviewer 2. This approach quantifies the degree of agreement between each pair of raters beyond what would be expected by chance, providing a robust measure of inter-rater reliability. All IRR analyses were conducted using R (version 4.3.2) and a mix of the {irr} package (80) and the {psych} package (89).

IRR serves as a very general indicator of overall agreement, but, as discussed previously, it does not provide insight into the specific nature of rater disagreements. To address this limitation, we conducted a more detailed analysis of RECIST outcome assignment using linear mixed effects models. These models treated the RECIST outcome as an ordinal variable, with rater as a fixed effect and random intercepts for each subject to account for repeated measures. Non-evaluable timepoints were removed, and the data was ordered from worst outcome to best outcome as follows: progressive disease (PD), Non-CR/Non-PD, stable disease (SD), partial response (PR), and complete response (CR). Post-hoc pairwise comparisons of estimated marginal means (EMMs) between raters were performed using the {emmeans} package, with Tukey adjustment for multiple comparisons. This approach allows us to explore systematic differences in RECIST outcome assignment between raters while placing soeme level of emphasis on the ordinal nature of the outcome measures.

A linear mixed effects model is particularly well-suited for this analysis because it accounts for the hierarchical structure of the data, where multiple ratings are nested within individual subjects. By including random intercepts for each subject, the model appropriately handles the correlation between repeated measurements from the same individual, thereby avoiding inflated type I error rates that can arise from treating observations as independent. The model is specified as follows:

Let $y_{ij}$ denote the outcome for subject $i$ as rated by rater $j$, and let the categorical variable $\text{rater}_{ij}$ indicate which rater provided the assessment.

$$y_{ij} = \beta_0 + \beta_1 \cdot I(\text{rater}_{ij} = \text{CR1}) + \beta_2 \cdot I(\text{rater}_{ij} = \text{CR2}) + b_{0i} + \varepsilon_{ij} \tag{2.12}$$

where:

- $\beta_0$ is the fixed intercept (mean outcome for ratings by the Site Investigator),
- $\beta_1$ is the fixed effect for Central Reviewer 1 (difference from Site Investigator),
- $\beta_2$ is the fixed effect for Central Reviewer 2 (difference from Site Investigator),
- $b_{0i} \sim \mathcal{N}(0, \sigma_b^2)$ is the subject-specific random intercept,
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the residual error term,
- $I(\cdot)$ is an indicator function equal to 1 if the condition is true and 0 otherwise.

Post-hoc pairwise comparisons between raters are performed on the mixed model fixed-effect estimates using the {emmeans} package, applying Tukey adjustment for multiple comparisons within each study. This approach provides a preliminary assessment of the presence of rater-specific differences in RECIST outcome assignment, though it should be considered a crude estimation of differences given the quasi-ordinal nature of the outcome means the actual value of differences may not be interpreted in a numeric, linear fashion; more nuanced analyses of key RECIST metrics like disease progression and response are presented in subsequent sections to address this shortcoming.

#### 2.2.2.2. Objective Response Rate (ORR)

To assess whether the presence of discrepancies in RECIST evaluations affects the classification of treatment response, we compare the Objective Response Rate (ORR) between site investigators and central reviewers for each trial. ORR is defined as the proportion of patients achieving either a RECIST-rated complete or partial overall response at any point during the trial.

To test for differences in ORR detection among raters, we use Cochran's Q test as an omnibus test, which is specifically designed for matched/paired binary data (90). Cochran's Q test evaluates whether there are significant differences in proportions across three or more related groups by testing the null hypothesis that all raters classify patients as responders at the same rate (90,91). This is particularly useful for identifying overall disagreement in binary outcomes (such as response vs. no response) when the same subjects are assessed by multiple raters.

We further conduct post-hoc pairwise McNemar's tests to identify specific differences between raters, applying Bonferroni adjustments for multiple comparisons within each study. McNemar's test is used to compare paired proportions between two related groups, focusing on discordant pairs (i.e., cases where the two raters disagree) (92,93). It is especially useful for detecting systematic differences in binary classifications between two raters, such as whether one rater is more likely than another to classify a patient as a responder (92). The two-way contingency tables for these pairwise comparisons are provided in the appendix for reference. This approach provides a robust framework for detecting both overall and pairwise differences in response classification between rater groups.

#### 2.2.2.3. Time to Event Analyses

While ORR captures *whether* a response occurred, it does not reflect the timing of that response which is an essential consideration in evaluating treatment efficacy and informing clinical decision-making. To address this, we extended our analysis to include several time-to-event outcomes, allowing us to assess not only whether raters agreed on the occurrence of an event, but also whether they differed in how quickly they identified it. The following time-to-event outcomes were analyzed:

- Time to Progression (TTP): the time from baseline evaluation to disease progression
- Time to Response (TTR): the time from baseline evaluation to the first recorded response
- Duration of Response (DoR): the time from initial response to subsequent progression

Survival analysis is selected for its ability to incorporate multiple raters simultaneously, account for censoring, which is common in clinical trial data, and take advantage of the fact that evaluation time points are aligned across raters in the trial datasets. These features make survival models particularly well-suited for comparing the timing of key clinical events between site investigators and central reviewers.

##### 2.2.2.3.1. Cox Proportional Hazards Models

To compare the timing of clinical events between site investigators and central reviewers, we employed Cox proportional hazards models, a semi-parametric approach that estimates the hazard ratio (HR) for event occurrence between groups while adjusting for censoring (94). This method enabled us to quantify whether one rater group systematically identified responses or progressiocox proportional hazard modeln earlier or later than another.

Although Cox models are typically used for time-to-event data, they can also be applied to compare the timing of discrete events such as treatment responses or disease progression. In this context, the hazard function represents the instantaneous risk of an event occurring at a given time point, conditional on survival up to that time. By modeling the hazard function separately for site investigators and central reviewers, we could directly compare the relative timing of events between these two groups.

Three key assumptions underlying the validity of the Cox proportional hazard model are as follows:

1. **Proportional Hazards Assumption**: The hazard ratios are constant over time, meaning the relative risk of an event occurring in one group compared to another does not change as time progresses (95,96).
2. **Independence of Survival Times**: The survival times of individuals are independent, meaning that the event timing for one individual does not influence that of another (95,97).
3. **Non-informative Censoring**: Censoring is non-informative, meaning that the reasons for censoring (e.g., loss to follow-up) are unrelated to the likelihood of the event occurring (97).

To address assumptions regarding proportional hazards, we conduct several diagnostic checks. We first create Kaplan-Meier survival curves for all TTP, TTR, and DoR for all three trials in order to visualize the time-to-event distributions for each rater group. These curves are then evaluated using Schoenfeld residuals and log(-log) survival curves, which help confirm whether the hazard ratios remain constant over time (95).

Independence of survival times was assumed based on the design of the trials, where each patient's event timing was independent of others. However, because we restructured the data such that patients could be included in a data set multiple times (e.g. when multiple raters determined the event of interest had occurred), we adjusted our analysis to account for possible intra-cluster correlation. Specifically, we used the `coxph()` function from the {survival} package in R (98), specifying the `cluster` argument as the subject ID, effectively treating each study participant as a cluster. Specifying a cluster variable in this way allows the model to account for the correlation of survival times within subjects. This adjustment is crucial for obtaining valid standard errors and confidence intervals in the presence of multiple raters assessing the same subjects, but will not affect the hazard ratio estimates.

#### 2.2.2.4. Estimation and Comparison of Hazard Ratios

The log-rank test is a statistical test used to compare the survival distributions of two or more groups, and is a standard method for checking for differences in time-to-event outcomes (99,100). In our analysis, the log-rank test was initially employed as a standard exploratory step to evaluate overall survival curve differences between site investigators and central reviewers, and the associated p-values for determining if *a* difference is persent are included in the Kaplan-Meier curves. However, while the log-rank test is a widely used method for comparing survival curves, it has limitations in terms of interpretability. Specifically, it does not provide estimates of the magnitude of differences in event timing, and, by extension, it does not allow for direct comparisons of hazard rates between different groups (99). As such, we opted to focus on post-hoc comparisons of the Cox model hazard ratios instead.

Hazard ratios (HRs) were estimated using the `coxph()` function from the {survival} package in R (98). The models were specified with the time-to-event outcome as the response variable, the rater group (site investigator, central reviewer one, central reviewer two) as a predictor, and subject ID as a cluster variable to account for intra-cluster correlation. The proportional hazards assumption was checked using Schoenfeld residuals and log(-log) survival curves, as described previously. The models were fitted separately for each time-to-event outcome (TTP, TTR, DoR) across the three trials. The Site Investigator group was selected as the reference group in all models, and, as such, the resulting hazard ratios (HRs) can be interpreted relative to this group; hazard ratio values greater than 1 indicate that the central reviewer(s) identified events later than site investigators, and values less than 1 indicate that central reviewer(s) identified events earlier.

The fitted Cox models are then passed to the `emmeans()` function from the {emmeans} package in R (101) to estimate the pairwise marginal means between rater groups (i.e. the differences in hazard ratios). This approach has the benefit of automatically employing the logged hazard ratios which can be mathematically summed and subtracted, and later back-transformed to the original hazard ratio scale[3]. The calculated differences in logged hazard ratios thus represent differences in the average hazard of an event occurring at any given time point between rater groups. In line with the interpretation of hazard ratios, a

[3] These calculations are mathematically equivalent to calculating the ratio in hazard ratios between rater groups; we simply chose to calculate differences in logged hazard ratios to facilitate the use of the {emmeans} package and to ensure that the resulting confidence intervals were symmetric around zero (no difference).

positive difference indicates that the central reviewer(s) identified events later than the site investigators, while a negative difference indicates that the central reviewer(s) identified events earlier.

This approach allowed for more interpretable and direct comparisons of event timing between rater groups, providing estimates of the average hazard over time. By leveraging the differences in hazard ratios, we were able to extract clinically meaningful contrasts such as whether one rater group consistently identified responses or progression earlier than the other while also maintaining an analysis structure parallel to analyses seen in clinical trials and regulatory submissions. The estimated marginal means (EMMs) of hazard ratios were then used to summarize the differences in event timing between site investigators and central reviewers across the three trials in a meta-analysis.

#### 2.2.2.5. Meta-Analysis of Rater Differences

To assess whether systematic differences in event timing between site investigators and central reviewers were consistent across trials, we performed a random-effects meta-analysis of the differences in estimated marginal means (EMMs) of the logged hazard ratios (HRs) for key time-to-event outcomes. This approach enabled us to synthesize evidence from all three trials, accounting for both within- and between-study variability, and to evaluate the generalizability of rater-related discrepancies.

The meta-analysis was conducted on the EMM-based differences in logged HRs, which quantify the average difference in event timing (e.g., TTP, TTR, or DoR) between rater groups. By focusing on the logged HRs, we ensured that the effect sizes were on a scale suitable for meta-analytic pooling and that the resulting confidence intervals could be interpreted symmetrically around zero (no difference).

A random-effects model was chosen to reflect the expectation that the true effect size may vary across studies due to differences in trial design, patient populations, or assessment procedures. Each study's effect size was weighted by the inverse of its variance, ensuring that more precise estimates contributed more to the pooled result. Between-study heterogeneity was quantified using the $I^2$ statistic and Cochran's $Q$ test, providing insight into the consistency of rater differences across trials.

Forest plots were generated to visually summarize the individual and pooled estimates, along with their 95% confidence intervals. After meta-analysis, the pooled difference in logged HRs and its confidence interval were exponentiated to return to the HR scale, facilitating clinical interpretation. A pooled HR greater than 1 indicates that central reviewers, on average, identified events later than site investigators, while a pooled HR less than 1 suggests earlier identification by central reviewers.

This meta-analytic framework provided a robust summary of rater-related timing discrepancies, contextualizing the magnitude and direction of these effects across diverse clinical settings. By synthesizing results from multiple trials, we were able to draw more generalizable conclusions about the impact of rater group on key clinical trial endpoints.

#### 2.2.2.6. Equivalence Testing

Recognizing that a lack of statistical significance in traditional hypothesis testing does not imply equivalence, we conducted formal equivalence tests to assess whether the observed differences in hazard ratios (HRs) between site investigators and central reviewers were small enough to be considered clinically negligible. Specifically, we applied the Two One-Sided Tests (TOST) procedure, a widely accepted method for testing statistical equivalence (102,103).

The TOST procedure is designed to formally test whether an observed effect falls within a pre-specified equivalence margin—an interval of values considered so close to the null value (e.g., $HR = 1$) that any difference is deemed not clinically meaningful. Lacking clear guidance on what should constitute equivalence between raters, we defined an equivalence margin of $[0.80, 1.25]$ a priori, which is consistent with thresholds used in bioequivalence studies (104), where a 20% deviation in either direction is often considered acceptable for treatment comparisons or measurement agreement (104).

The TOST procedure involves the following steps (103):

1. Define the equivalence interval: Specify the range of effect sizes (here, $HR \in [0.80, 1.25]$) that are considered practically equivalent to no difference.
2. Conduct two one-sided hypothesis tests:
   - $H_{01} : HR \leq 0.80$ (the effect is meaningfully smaller than the lower bound)
   - $H_{02} : HR \geq 1.25$ (the effect is meaningfully larger than the upper bound)
3. Interpretation: If both null hypotheses are rejected at the chosen significance level ($\alpha = 0.05$), the observed HR is considered statistically equivalent to 1.0, indicating no meaningful difference between rater groups.

To align with the TOST framework, 90% confidence intervals for the hazard ratios were used, as this corresponds to the 1–2 rule (i.e., $1 - 2 \times 0.05 = 0.90$) for equivalence testing (103). This approach provides a more rigorous and interpretable framework for concluding equivalence than simply failing to reject the null hypothesis in a traditional test. It allows us to distinguish between true similarity and statistical inconclusiveness, which is particularly important in regulatory and clinical decision-making contexts. TOST analyses were performed using the `TOSTone()` function from the R package {TOSTER} (105).

The major value of equivalence testing lies in its ability to distinguish between *no evidence of a difference* and *evidence of no meaningful difference*. Traditional NHST can only reject or fail to reject the null hypothesis of no effect; a non-significant result may simply reflect insufficient power, not true equivalence (103). In contrast, the TOST procedure allows us to make a positive statement about similarity: if the observed effect and its confidence interval fall entirely within the equivalence bounds, we can conclude that any difference is too small to be of concern with the important caveat that that the equivalence bounds should be established with guidance from subject-matter experts like clinicians and statisticians (103).

It is important to note, as highlighted in the comment above, that it is technically possible—though uncommon—to both reject the null hypothesis of no difference (the NHST) *and* establish equivalence (TOST) simultaneously. This comes across contradictory at the surface-level, but it can occur in situations with very wide equivalence margins, small sample sizes, or particular data artifacts, but more often reflects a true underlying effect that is statistically detectable yet still within the range considered clinically unimportant.

By incorporating both NHST and equivalence testing, our analysis provides a more nuanced and rigorous assessment of rater differences, allowing us to distinguish between statistically significant differences, true equivalence, and inconclusive findings. This dual approach is particularly valuable in regulatory and clinical decision-making contexts, where the distinction between “not different” and “equivalent” has important practical implications.

### 2.2.3. Sensitivity Analyses of RECIST 1.1 Thresholds

A central concern in RECIST-based clinical trials is the potential for systematic differences in tumor response assessments between site investigators and central reviewers. While central review is often implemented to enhance consistency and reduce bias, it remains unclear to what extent observed discrepancies are driven by subjective interpretation of imaging versus the rigid application of fixed response thresholds. RECIST 1.1 defines specific percentage changes in tumor burden to classify treatment response and disease progression (e.g., a partial response [PR] is defined as a 30% decrease in the sum of diameters of target lesions from baseline, while progressive disease [PD] is defined as a 20% increase from the nadir, with an absolute increase of at least 5 mm). Although these thresholds are widely accepted, they are ultimately arbitrary. When tumor measurements fall near these cutoffs, even small variations in measurement or interpretation can lead to discordant classifications between raters.

#### 2.2.3.1. Methodological Approach

To investigate whether the fixed nature of these thresholds contributes to inter-rater discrepancies, we conducted a high-resolution sensitivity analysis. This analysis systematically evaluated how varying the RECIST-defined cutoffs for PR and PD affects both inter-rater agreement and key clinical outcome estimates.

We generated a series of modified datasets by replacing the standard 30% (PR) and 20% (PD) thresholds with all possible combinations of values ranging from 0% to 100% inclusive, in 1% increments. This resulted in 10,201 unique threshold pairs (101 × 101). For each threshold combination, tumor response categories were reclassified for every patient and rater based on the new cutoffs. Overall response was then recalculated using standard RECIST 1.1 logic, incorporating the modified target lesion responses along with the original non-target lesion and new lesion data.

To quantify the impact of these threshold changes, we computed Fleiss' $\kappa$ for both target lesion response (based solely on the modified percent change thresholds) and overall response (based on full RECIST logic). This allowed us to assess how agreement between raters varied across the threshold space.

In parallel, we recalculated key clinical outcomes ORR, TTR, TTP, and DoR for each rater using the modified classifications. For ORR, we recalculated the proportion of patients classified as responders (CR or PR) for each rater at each threshold pair, and reperformed Cochran's Q tests.

For the time-to-event outcomes of TTR, TTP, and DoR, we recalculated event times based on modified outcome classifications corresponding to different RECIST threshold definitions. These recalculated times were used to fit Cox proportional hazards models, as previously described, to estimate hazard ratios for each reviewer group at each threshold pair. Following a general rule of thumb for Cox regression (106), we filtered out data sets with fewer than $(k-1) * 10$ cases, where $k$ is the number of raters in the study. In other terms, if fewer than 20 events of interest (TTR, TTP, or DoR) were observed across all 3 raters in a data set, then the Cox models at those thresholds were ignored. This is done to ensure that a reasonable number of events are present in the data set to allow for meaningful comparisons between raters (106).

To assess how these classification threshold changes influenced the relative risk associated with reviewer assessments, we conducted a difference-in-differences (DiD) analysis comparing hazard ratios between site investigators and central reviewers.

In all models, the site investigator group served as the reference, meaning their hazard ratio remained constant across thresholds, while the central reviewer's hazard ratio varied according to the applied RECIST definitions for progression and response. The DiD was calculated as the difference in central reviewer hazard ratios between the original and modified thresholds, effectively isolating the impact of threshold changes on estimated risk. We defined the primary contrast of interest as the difference in hazard ratios between the site investigator and the central reviewer at each threshold. Specifically, we calculated:

$$\Delta_{\text{RECIST}} = HR_{\text{Site}} - HR_{\text{Central (Original)}}, \quad \Delta_{\text{Sensitivity}} = HR_{\text{Site}} - HR_{\text{Central (New)}} \tag{2.13}$$

where $\Delta_{\text{RECIST}}$ represents the difference in hazard ratios at the original RECIST criteria (20% progression, 30% response) and $\Delta_{\text{Sensitivity}}$ represents the difference in hazard ratios at the new combination of disease progression and response thresholds. Then, the difference-in-differences was computed as:

$$\begin{aligned}\Delta_{\text{RECIST}} - \Delta_{\text{Sensitivity}} &= (0 - HR_{\text{Central Reviewer Original}}) - (0 - HR_{\text{Central Reviewer New}}) \\ &= HR_{\text{Central Reviewer New}} - HR_{\text{Central Reviewer Original}}\end{aligned} \tag{2.14}$$

Because the site investigator's hazard ratio remained fixed, this expression simplifies to the change in the central reviewer's hazard ratio across thresholds. A positive DiD value indicates that the central reviewer's hazard ratio increased under the new threshold. Conversely, a negative DiD suggests that the central reviewer's estimated hazard decreased under the new classification rule. It should also be noted that we refer here to a single central reviewer because these time-to-event analyses employed the averaged hazard ratios across the two central reviewers as this simplified the analysis and interpretation of results. The averaging was done to reduce the number of heatmaps generated and because our analyses here do not focus on differences between central reviewers as individuals, but rather on differences between site investigators and central reviewers as a group.

#### 2.2.3.2. Visualization of Results

In order to reduce the complexity of the results and facilitate interpretation, we generated heatmaps that display the difference in agreement and outcome estimates between the values found at each alternative threshold pair and those found at the actual RECIST 1.1 cut-off criteria (30% for PR and 20% for PD). The heatmaps were constructed such that each cell represents the difference in either Fleiss' $\kappa$ or DiD of the hazard ratio between the modified threshold pair and the standard RECIST thresholds. Positive values indicate that the modified thresholds resulted in higher agreement or more favorable outcomes compared to the standard RECIST definitions, while negative values indicate lower agreement or less favorable outcomes. In the case of hazard ratios, the values of the central reviewers were first averaged before calculating the difference from the site investigator values, allowing for a direct comparison of the relative timing of events between site investigators and central reviewers. This was done primarily to reduce the number of heatmaps generated, and because our analyses do not focus on the differences between the central reviewers as individuals, but rather on the differences between site investigators and central reviewers as a group.

These heatmaps provide a visual summary of how sensitive inter-rater agreement and clinical outcome estimates are to the choice of PR and PD thresholds, and help identify regions of the threshold space where agreement is maximized or where outcome estimates diverge from those obtained using the standard RECIST definitions. They were generated using the {ggplot2} package in R (107).

This sensitivity analysis provides a systematic framework for evaluating the robustness of RECIST-based classifications and clinical outcomes to the choice of response thresholds. By exploring a wide range of possible cutoffs and visualizing the resulting changes in agreement and outcome estimates, this approach helps to clarify the extent to which observed rater discrepancies may be attributable to the arbitrariness of the RECIST criteria, and informs the interpretation and design of future RECIST-based studies.

# 3. Results

The results of the preceding analyses are presented in this chapter, again divided into three major sections. The first section covers the results of the IRR analyses, including a summary of the studies selected for inclusion and an overview of the meta-analysis results including forest and funnel plots. The second section is a deep dive into the degree of concordance we observe between site investigators and central reviewers both within individual studies and averaged across studies. Regarding the latter, we specifically present both traditional null hypothesis tests and equivalence tests of the meta-analysis results. Finally, we close out the results with qualitative interpretations of the sensitivity analyses, aiming to identify and summarize commonalities and differences across the three trials.

## 3.1. Inter-Rater Agreement is Substantial

### 3.1.1. Summary of Selected Studies

A total of ten studies are ultimately identified in the literature that meet the inclusion criteria for this meta-analysis. Of these, four studies report inter-rater reliability using Cohen's $\kappa$, while six studies use Fleiss' $\kappa$ as their primary measure. All but one of the included studies are conducted in non-clinical trial contexts, comprising a mix of retrospective and prospective serial enrollment studies, most often at a single site. Across these studies, the raters are typically radiologists or oncologists with substantial experience in tumor measurement, ensuring a high level of expertise in the assessment process. As mentioned in the Methods section (Section 2.1.2), this data is further supplemented by three clinical trials wherein we calculate Fleiss' $\kappa$ for the site investigators and central reviewers. These trials are identified by their NCT numbers in the meta-analysis results. A complete list of the studies included in the meta-analysis, along with their key characteristics and inter-rater reliability measures, is provided in Table 3.1.

This table contains columns for the Author, the type of $\kappa$ measure (Cohen's or Fleiss') used, the year of study publication, a general remark about the context of the study, the number of raters, comments about the type of raters employed, and the number of patients and scans available for each study. Although not directly included in the table, the total number of scans is generally a multiple of ~2 or ~3 indicating that 2 or 3 scans were evaluated per individual, with some variation due to study attrition. The table also includes footnotes to highlight important details about the studies, such as whether the study was conducted across multiple sites, whether lesions were selected by consensus among raters, and whether the study explicitly examined differences in baseline tumor selection by raters. These footnotes provide additional context for interpreting the inter-rater reliability measures and should be considered when drawing conclusions from the meta-analysis results.

| Author | κ Measure | Year | Context | No. Raters | Type of Rater(s) | No. Patients | Total Scans |
|---|---|---|---|---|---|---|---|
| [§]Aghighi et al. (60) | Cohen's | 2016 | Local Study | 2 | Radiologists | 74 | 148 |
| El Homsi et al. (41) | Fleiss' | 2024 | Local Study | 7 | Radiologists | 159 | 318 |
| Felsch et al. (61) | Cohen's | 2017 | Clinical Trial | 2 | Site Investigators, Central Reviewer | 170 | 484 |
| Ghobrial et al. (58) | Cohen's | 2017 | Local Study | 2 | Radiologist, Oncologist | 28 | 56 |
| Ghosn et al (62) | Fleiss' | 2021 | Local Study | 3 | Radiologists | 37 | 74 |
| Karmakar et al. (63) | Cohen's | 2019 | Local Study | 2 | Radiologists | 61 | 82 |
| Kuhl et al (40) | Fleiss' | 2019 | Local Study | 3 | Radiologists | 316 | 932 |
| [¶]Oubel et al. (44) | Fleiss' | 2015 | Local Study | 3 | Two oncologists, One radiologist | 11 | 33 |
| Tovoli et al. (64) | Fleiss' | 2018 | Local Study | 3 | Radiologists | 77 | 154 |
| [†]Zimmermann et al. (59) | Fleiss' | 2021 | Local Study | 3 | Radiologists | 42 | 84 |

Table 3.1.: Summary of studies included in the meta-analysis of inter-rater reliability (IRR) measures for RECIST 1.1.

[§] This study was conducted across multiple sites, with scans obtained from three different locations.

[¶] In this study, the selection of lesions at baseline was performed by consensus among the raters. This approach is likely to inflate both the estimated inter-rater reliability and the precision of the results; therefore, these findings should be interpreted with caution.

† This study explicitly examined differences that arose when raters selected *different* baseline tumors.

For the studies utilizing Fleiss' $\kappa$, nearly all involved three raters, with the exception of one study that included seven radiologists. The primary aim of this outlier study was to investigate how the number of years of experience among raters influenced inter-rater reliability. This focus on rater experience can add an important dimension to the understanding of variability in tumor measurement, but was not the primary focus of this meta-analysis and should be kept in mind when interpreting the results.

It is also important to highlight the study by Felsch et al. (61), which provided three distinct sets of ratings: one from the site investigators and two from the central reviewers. This study calculated Cohen's $\kappa$ for both the comparison between central reviewer 1 and central reviewer 2, as well as for the consensus central reviewer versus the site investigator. For the purposes of this meta-analysis, only the consensus central reviewer versus site investigator comparison was included, as it was deemed the most relevant. Unfortunately, Fleiss' $\kappa$ was not calculated for this study, limiting direct comparison with the other included studies.

### 3.1.2. Overall Meta-Analysis Results

The overall meta-analysis of IRR for RECIST 1.1 yields a pooled $\kappa$ estimate of 0.66 (95% CI: 0.56 to 0.73), indicating substantial agreement among raters across studies as seen in Figure 3.1. This forest plot illustrates the individual $\kappa$ estimates for each study, along with their confidence intervals, and the overall pooled estimate. The pooled estimate and associated 95% confidence interval, represented by the black diamond at the bottom, suggests that, on average, raters are able to agree on tumor measurements and classifications with a substantial level of reliability.

However, the analysis also reveals considerable heterogeneity, with a Q statistic of 250.82 (df = 12, $p < 0.0001$), an $I^2$ value of 95.96%, and a $tau^2$ of 0.23. These values suggest that nearly all of the observed variability in $\kappa$ estimates is attributable to real differences between studies rather than chance alone. Such high heterogeneity is not uncommon in meta-analyses of rater agreement, especially when studies differ in design, populations, or rater experience, and it underscores the importance of interpreting the pooled estimate with caution.

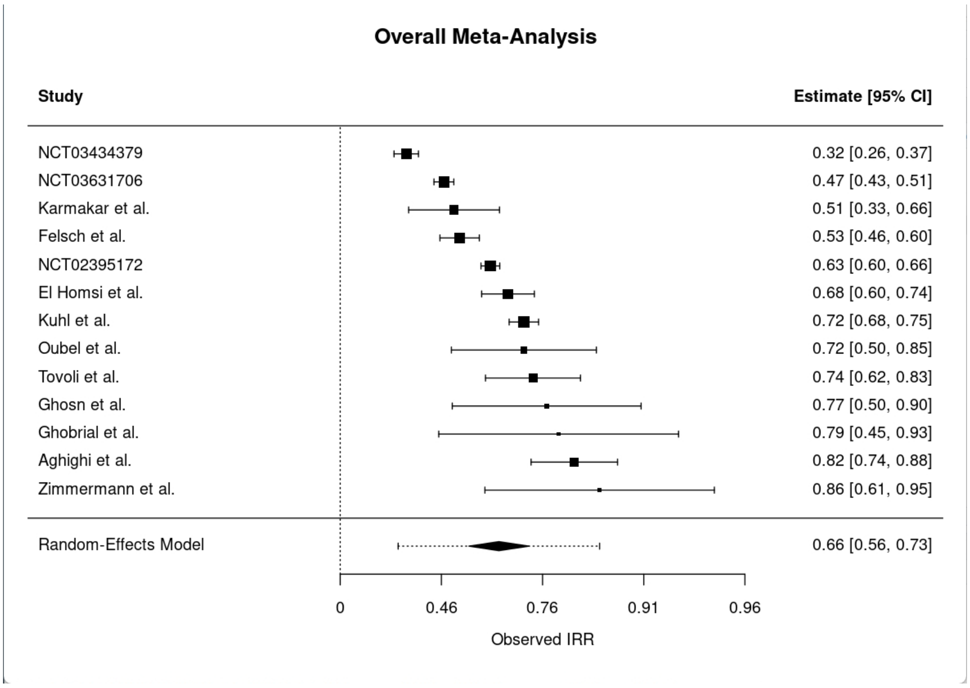


Figure 3.1.: Forest plot of inter-rater reliability (IRR) meta-analysis

For completeness, separate meta-analyses are also conducted for studies reporting Fleiss' $\kappa$ and Cohen's $\kappa$ individually. The pooled Fleiss' $\kappa$ is 0.65 (95% CI: 0.54 to 0.74; Q = 224.23, df = 8, p < 0.0001; $I^2$ = 96.80%; tau$^2$ = 0.23), and the pooled Cohen's $\kappa$ is 0.67 (95% CI: 0.46 to 0.81; Q = 25.22, df = 3, p < 0.0001; $I^2$ = 88.48%; tau$^2$ = 0.34). These supplemental analyses, along with their corresponding figures, are provided in the appendix. The results for both measures are consistent with the overall finding of substantial inter-rater agreement, though the high heterogeneity persists within each subgroup and is consistent with several of the footnotes in Table 3.1. Meta-analysis plots and funnel plots for Cohen's $\kappa$ and Fleiss' $\kappa$ are available in the appendix at Figure A.2 and Figure A.1, respectively.

To assess the potential for publication bias, funnel plots are generated for each inter-rater reliability measure included in the meta-analysis as seen in Figure 3.2. This funnel plot displays the individual study estimates of IRR on the x-axis against their standard errors on the y-axis. The studies are represented by points, with the overall pooled estimate indicated by a vertical line. The funnel plot is expected to be symmetric around the pooled estimate if there is no publication bias or small-study effects.

Visual inspection of these plots suggests the presence of asymmetry, raising concerns about apparent publication bias. To formally test for this, Egger's test for funnel plot asymmetry is performed using the `regtest()` function from the metafor package. The results of

Egger's test are statistically significant (p = 0.002), providing further evidence that publication bias or small-study effects may be present in the included studies. These findings should be considered when interpreting the pooled estimates, as they indicate that the observed effect sizes may be influenced by factors such as selective reporting or differences in study size.

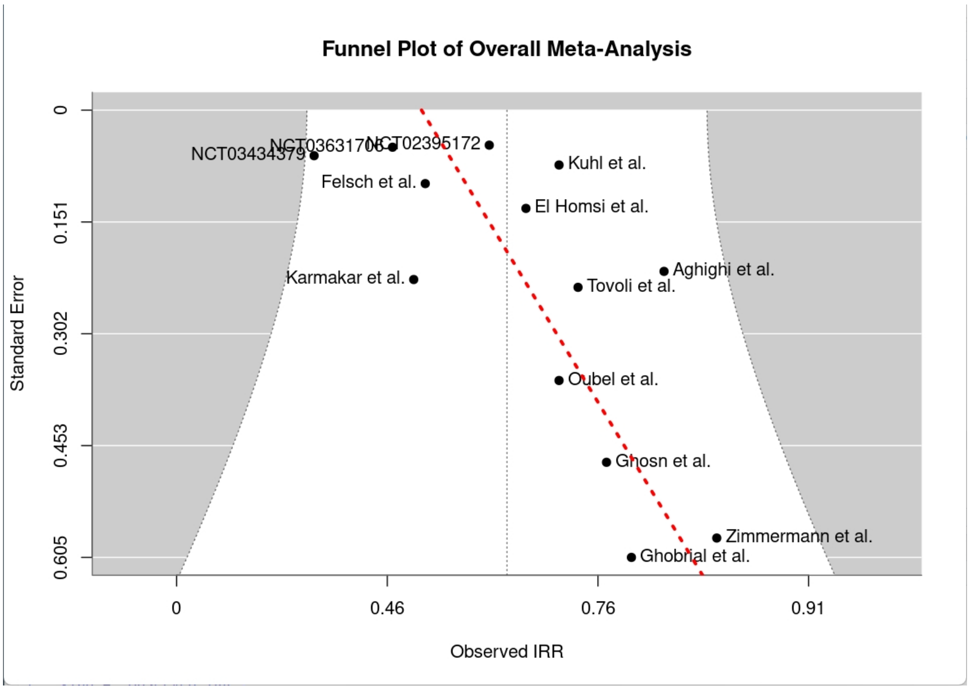


Figure 3.2.: Funnel plot of inter-rater reliability (IRR) meta-analysis

However, the asymmetry observed in the funnel plot should not be interpreted solely as evidence of publication bias. It is likely influenced by the small sample size of the clinical trial data and the fact that all clinical trials were conducted by the same research group. Additionally, Egger's test is generally recommended for meta-analyses with at least 10 studies, and our analysis includes just slightly over that minimum recommendation at 13. This data limitation should be considered when interpreting the results. Another possible explanation for the observed asymmetry is that the clinical trials may simply have different inter-rater reliability compared to the non-clinical trial studies, potentially reflecting a different underlying "population" of raters than the radiologists used in almost all other studies. This clustering of clinical trials in the funnel plot suggests a potential limitation of the meta-analysis, but does not necessarily indicate publication bias.

More likely, the results reflect small study effects, a common phenomenon in meta-analyses where smaller studies tend to report larger effect sizes than larger studies. Methodological differences may also contribute, as seen in Oubel et al. (44), where raters selected tumors at baseline by consensus rather than independently. This approach can overestimate inter-rater reliability, as consensus selection increases the likelihood of agreement among raters.

An additional observation from the meta-analysis is that the four clinical trials included in the dataset appear to cluster together in terms of effect size, with the funnel plot suggesting these studies are distributed around a different mean (visually estimated at ~0.45) compared to the mean calculated for all studies pooled together. This pattern may indicate that inter-rater reliability is somewhat lower in clinical trial settings than in non-clinical trial contexts. However, this trend is not formally tested for statistical significance due to the small sample size of the clinical trial subgroup. As such, this observation should be interpreted cautiously and considered a potential area for further investigation.

Despite high heterogeneity in the studies and potential small sample bias, the overall findings of the meta-analysis suggest that inter-rater reliability for RECIST 1.1 is moderate to substantial, with a pooled $\kappa$ estimate of 0.66. Even in the context of clinical trials, where the pooled $\kappa$ was slightly lower at approximately 0.45, there still appears to be a reasonable level of agreement among raters from a statistical perspective. However, this level of agreement may not be sufficient to ensure consistent and reliable tumor measurements across different raters within a clinical care or trial context, where precise measurements are crucial for treatment decisions and trial outcomes.

The results of our follow-up analyses, particularly the evaluation of differences in hazard ratios and important time-to-event outcomes, thus investigate the practical implications of these findings in clinical trial settings.

## 3.2. Concordance in Clinical Trial Outcomes

### 3.2.1. Site Investigators and Central Reviewers Agreement Appears Study-Dependent

To assess inter-rater reliability (IRR), we examine pairwise Cohen's $\kappa$ estimates across the studies, as summarized in Table 3.2. This table provides a detailed overview of the pairwise Cohen's $\kappa$ estimates, with each column representing a study and each row representing a pairwise comparison between raters. The estimates are presented alongside their 95% confidence intervals and the number of images jointly assessed, providing a clear picture of the level of agreement between the raters in each study. The table also includes the number of cases assessed by each rater, which is important for interpreting the reliability estimates.

In two of the three studies (NCT03434379 and NCT03631706), the highest agreement is observed between the two site investigators, whereas in NCT02395172, this pair has the lowest agreement suggesting potentially high variability or inconsistency in which raters disagree with one another. Moreover, the range of Cohen's $\kappa$ values across all studies is broad, from 0.286 to 0.803, further indicating substantial variability in IRR. This variability is consistent with the high heterogeneity observed in the meta-analysis, suggesting that agreement between raters can differ considerably depending on the study context and rater pairings. The pairwise Cohen's $\kappa$ estimates provide a useful summary of the level of agreement between raters, but the follow-up analyses using linear mixed effects (LME) models with estimated marginal means (emmeans) provide a more nuanced understanding of the differences between raters.

It is worth explicitly noting that the value in adding these estimated differences on top of the pairwise Cohen's $\kappa$ estimates is that they allow us to at least partially account for the quasi-ordinal nature of the RECIST scale. That is, Cohen's $\kappa$ penalizes **all differences** in ratings equally, regardless of whether the differences are of meaning. For example, depending on the context, one might consider stable disease, partial response, and complete response all as sufficient outcomes, but the IRR analysis will necessarily penalize any differences in outcomes by rater. By contrast, using a linear mixed effects model with the RECIST outcomes set to an ordinal scale allowed us to assign a crude numeric value to outcomes that reflect the ordered nature of the scale. This additional layer of analysis can help to clarify the nature of the differences between raters and provide insights into how these differences may impact clinical decision-making or trial outcomes.

| Comparison | NCT02395172 | NCT03434379 | NCT03631706 |
|---|---|---|---|
| Central Reviewer 1 vs Central Reviewer 2 | $\kappa$ = 0.485 [0.437, 0.534], n=946 | $\kappa$ = 0.352 [0.284, 0.419], n=466 | $\kappa$ = 0.513 [0.465, 0.56], n=822 |
| Site Investigator vs Central Reviewer 2 | $\kappa$ = 0.677 [0.636, 0.719], n=967 | $\kappa$ = 0.307 [0.233, 0.382], n=425 | $\kappa$ = 0.424 [0.374, 0.474], n=788 |
| Site Investigator vs Central Reviewer 1 | $\kappa$ = 0.803 [0.769, 0.836], n=962 | $\kappa$ = 0.286 [0.211, 0.361], n=424 | $\kappa$ = 0.49 [0.441, 0.539], n=787 |

Table 3.2.: Pairwise Cohen's $\kappa$ estimates across studies

Note: While all raters were intended to evaluate the same number of cases, in practice the central reviewers assessed a more complete set of cases than the site investigators. As a result, the central reviewers have more cases represented in the figure.

Full model results for the LMEs are provided in the appendix (Equation A.1, Equation A.2, and Equation A.3), as these are not the primary focus of this section and the point estimates of these models should not be interpreted directly because the RECIST scale can only be considered quasi-ordinal. The main emphasis here, instead, is on the pairwise comparisons of raters using estimated marginal means. As discussed in the Methods (Section 2.2.2.1), the absolute values of these estimates should not be over-interpreted due to the quasi-ordinal nature of the scale; rather, the focus should be on whether the differences are large enough to suggest meaningful disagreement or concordance between raters.

We present here the results of the pairwise comparisons from one of the studies, NCT03631706, in Table 3.3 to demonstrate patterns of disagreement that can occur. The complete set of results for all three studies can be find at Table A.1. This table summarizes the pairwise contrasts of raters from the linear mixed effects model, including the estimated differences in RECIST outcomes, standard errors, degrees of freedom, t-ratios, and p-values. The p-values can be understood as whether a given pairwise comparison is statistically significant, with the Bonferroni method applied to adjust for multiple comparisons. The stars in the p-value column indicate significance levels at $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

In the example of NCT03631706, there is a significant difference between the central reviewers as well as a significant difference between the site investigator and central reviewer 1, but no difference between the site investigator and central reviewer 2. Meanwhile, studies NCT02395172 and NCT03434379 show no statistically significant differences between any of the raters, suggesting that the raters are in agreement on the RECIST outcomes for these studies.

| Contrast | Estimate | SE | df | t.ratio | p.value |
|---|---|---|---|---|---|
| Site Inv. - Reader 1 | 0.228 | 0.041 | 2507.766 | 5.630 | $0^{***}$ |
| Site Inv. - Reader 2 | 0.064 | 0.041 | 2507.790 | 1.586 | 0.2519 |
| Reader 1 - Reader 2 | -0.164 | 0.042 | 2501.910 | -3.865 | $3e\text{-}04^{***}$ |

Table 3.3.: Pairwise contrasts of raters from the linear mixed effects model in NCT03631706

These results might seem counterintuitive when comparing them to the IRR analyses as, for example, the Cohen's $\kappa$ estimates seen in study NCT03434379 might lead one to believe there are substantial differences between the raters. However, the LME analyses suggest that these differences are not statistically significant when accounting for the quasi-ordinal nature of the data, indicating that the raters are generally in agreement on the RECIST outcomes for this study. This discrepancy highlights the importance of considering both pairwise comparisons and more complex modeling approaches when evaluating inter-rater reliability in clinical trials. By employing a combination of methods, we can gain a more comprehensive understanding of how raters interpret and apply the RECIST criteria, ultimately informing efforts to improve consistency and accuracy in tumor response assessment.

### 3.2.2. Differences Between Raters Extend to Objective Response Rate

While overall agreement between raters is generally acceptable when measuring basic IRR and no differences between raters are identified in two of the three studies when using a mixed modeling strategy, there are notable exceptions in agreement when it comes to evaluating ORR. The omnibus Cochran's Q test for each study are available in Table 3.4; this table summarizes the results of Cochran's Q test for ORR analyses across the three studies. The table includes the study name, Cochran's Q statistic, p-value, and degrees of freedom (df) for each study. The p-values are presented in a format that indicates statistical significance, with stars denoting levels of significance at $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

The results of the Cochran's Q test indicate that there are statistically significant differences in ORR among the raters for studies NCT03434379 and NCT03631706, with p-values less than 0.05 in both cases. This suggests that at least one rater in each of these studies classifies responses differently from the others. To further explore these differences, pairwise McNemar's post-hoc tests are conducted using the `mcnemar.test()` function in R, which is appropriate for paired nominal data.

| Study | Cochran's Q | p-value | df |
|---|---|---|---|
| NCT02395172 | 0.05 | 0.973 | 2 |
| NCT03434379 | 6.33 | $0.042^{*}$ | 2 |
| NCT03631706 | 12.07 | $0.002^{**}$ | 2 |

Table 3.4.: Cochran's Q test results for objective response rate analyses.

The post-hoc results of the pairwise McNemar's tests, which provides a chi-squared statistic on 1 degree of freedom, were summarized as p-values in Table 3.5. This table presents the results of the pairwise McNemar comparisons between raters for each study, summarized simply as the p-values of the pairwise comparisons. The p-values are presented in a format that indicates statistical significance after Bonferroni adjustment, with stars denoting levels of significance at $^*p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. The results indicate that, in NCT03434379, the site investigator disagreed significantly with central reviewer 2 (as indicated by a significant p-value), while in NCT03631706, significant disagreement was observed between the site investigators and both central reviewers. The direction and magnitude of differences between raters can vary across studies, however.

| | Central Reviewer 1 | Site Investigator | Site Investigator |
|---|---|---|---|
| Study | vs. Central Reviewer 2 | vs. Central Reviewer 1 | vs. Central Reviewer 2 |
| NCT02395172 | 1 | 1 | 1 |
| NCT03434379 | 1 | 0.634 | $0.048^*$ |
| NCT03631706 | 1 | $0.018^*$ | $0.018^*$ |

Table 3.5.: Pairwise post-hoc McNemar's test results for ORR analyses: p-values.

To illustrate how these differences appear in the data itself, contingency table examples for significant and non-significant pairwise comparisons are provided in Table 3.6 and Table 3.7, respectively. That is, the contingency table in Table 3.6 shows how ORR data looks in the event of significant disagreement between raters, namely that the difference between off-diagonal cells is pronounced. On the contrary, the contingency table in Table 3.7 illustrates a lack of significant disagreement, with balanced off-diagonal cell values. (The full set of contingency tables for all studies can be found in the appendix at Table A.2, Table A.3, and Table A.4).

| | Site: Response | Site: No Response |
|---|---|---|
| Central : Response | 55 | 16 |
| Central: No Response | 3 | 72 |

Table 3.6.: Contingency table for Central Reviewer 2 and Site Investigator in NCT03631706.

| | Site: Response | Site: No Response |
|---|---|---|
| Central: Response | 268 | 5 |
| Central: No Response | 5 | 50 |

Table 3.7.: Contingency table for Central Reviewer 2 and Site Investigator in NCT02395172.

Of particular note in these contingency tables is that the overall number of disagreements between site investigator and central reviewers is relatively low, with most cases showing agreement on whether a response is present or not. This suggests that while there are some differences in how responses are classified, the raters overall accuracy is generally quite high. Based on the contingency table in Table 3.6, one might be tempted to conclude that site investigators are more hesitant to classify someone as having a partial or complete treatment response as an error in this direction could lead to a patient not receiving a treatment that could be beneficial. However, this interpretation should be approached with caution, as the differences in ORR are not necessarily indicative of a systematic bias in one direction or another. In fact, the contingency tables in study NCT03434379 (see Table A.3) show the opposite pattern, with the central reviewers being more likely to classify patients as having a response than the site investigators.

Overall, the results of this sub-analysis further suggest that, when disagreements do occur, they are typically between the site investigator and central reviewers, with the two central reviewers generally in agreement with one another. However, the presence of significant differences between raters in detection of ORR is study-dependent, and the direction of any differences can vary across studies. Moreover, the overall number of disagreements is relatively low, suggesting that while there may be some variability in how responses are classified, the raters overall accuracy is generally quite high. This finding underscores the importance of considering the context and specific characteristics of each study when interpreting inter-rater reliability and agreement on treatment response.

### 3.2.3. Other Key Trial Outcomes Show No Differences Between Raters

While the ORR analyses reveal some significant differences between raters, it is worth noting that ORR is typically only employed as a secondary outcome and is a more rudimentary measure of treatment response than the more complex time-to-event outcomes.

The latter are often more clinically relevant and can provide a more nuanced understanding of treatment effects. In this section, we present the results of time-to-event analyses, which are conducted for three key outcomes: TTP, TTR, and DoR. These analyses are performed using both Kaplan-Meier survival curves and comparison of the hazard ratios derived from Cox proportional hazards models, with the aim of assessing whether there are any significant differences between site investigators and central reviewers in terms of these important clinical endpoints.

Kaplan-Meier curves provide a visual representation of the time-to-event data and are particularly valuable for comparing event rates between different groups—in this case, between assessments made by site investigators versus central reviewers. In these curves, the y-axis represents the probability of being event-free (i.e., survival probability), while the x-axis represents time. For TTP analyses, the event of interest is disease progression, and the curves show the probability of remaining progression-free over time. Similarly, for TTR, the curves display the probability of remaining response-free (i.e., not yet having achieved a response), and for DoR, they show the probability of maintaining a response without progression. When interpreting these curves visually, curves that are higher indicate better outcomes for TTP and DoR (longer time to progression or longer duration of response), while for TTR, a lower curve is preferable (faster time to response). If the curves for different raters overlap substantially, this suggests strong agreement in their assessments. Conversely, separation between curves indicates systematic differences in how raters are evaluating events, with the degree of separation reflecting the magnitude of disagreement. Statistical significance of these visual differences can be assessed using log-rank tests, which evaluate whether the observed separation between curves could have occurred by chance.

Kaplan-Meier survival curves for TTP, TTR, and DoR in study NCT03434379 are shown in Figure 3.3, Figure 3.4, and Figure 3.5, respectively as examples of the time-to-event analyses. Complete results for the three studies including all Kaplain-Meier curves, log(-log(Survival)) plots, and Schoenfeld residuals plots are provided in the appendix at Section A.2.3.1, Section A.2.3.2, and Section A.2.3.3.

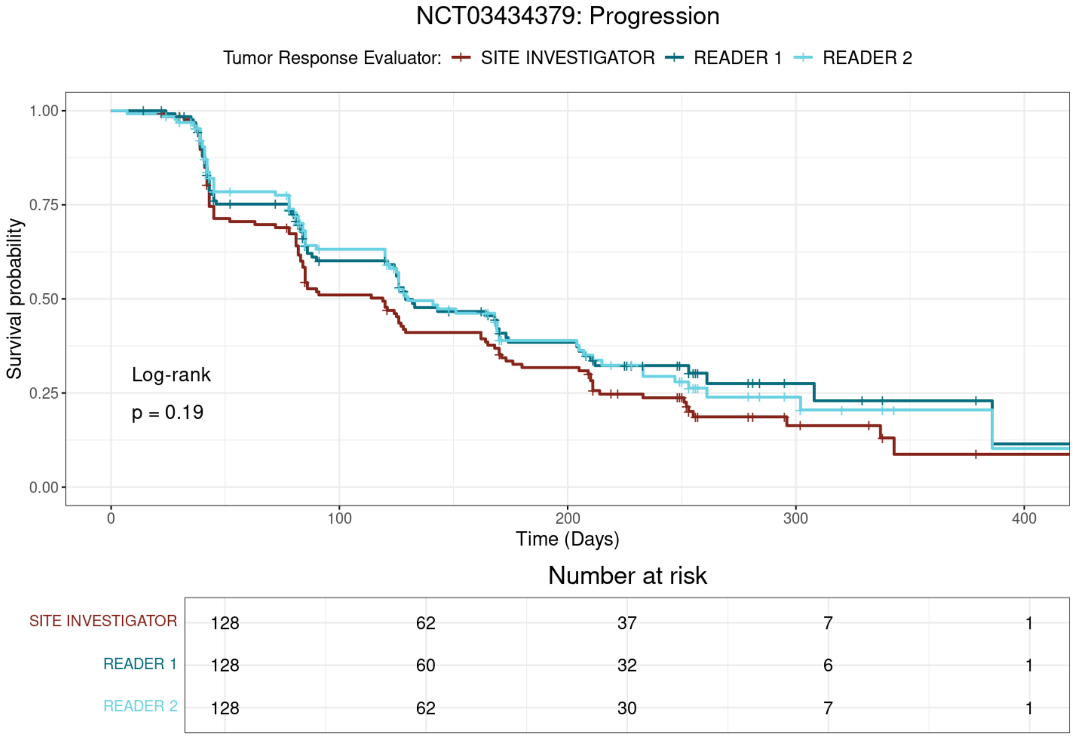


Figure 3.3.: Kaplan-Meier survival curves for TTP in NCT03434379

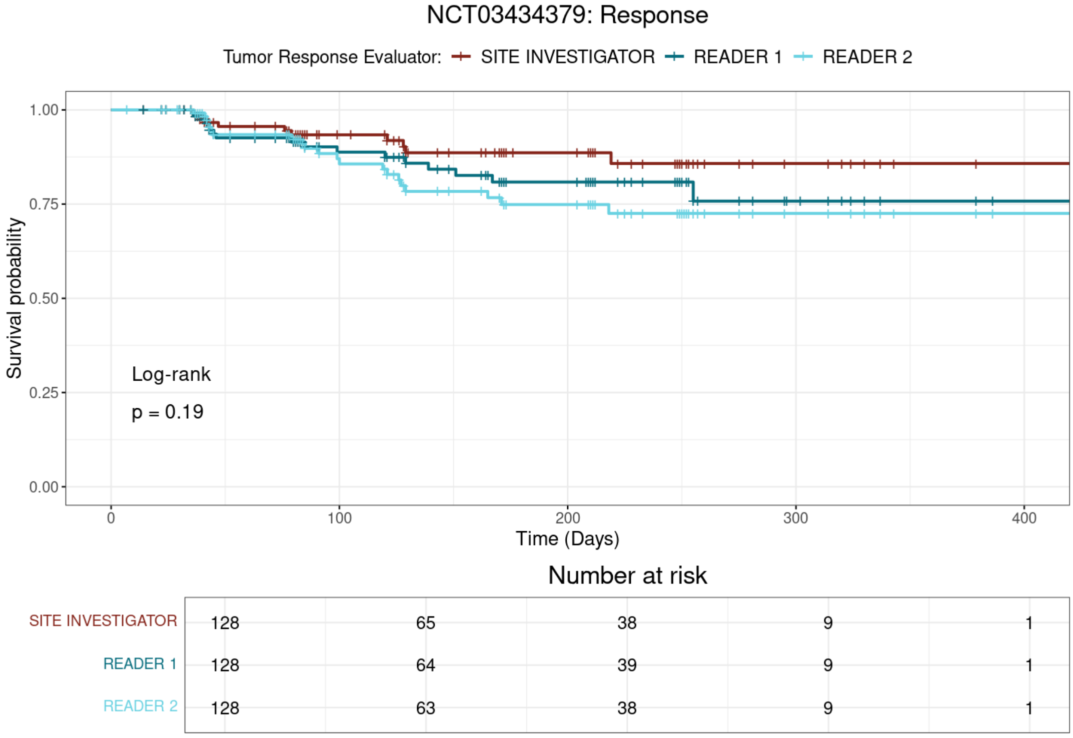


Figure 3.4.: Kaplan-Meier survival curves for TTR in NCT03434379

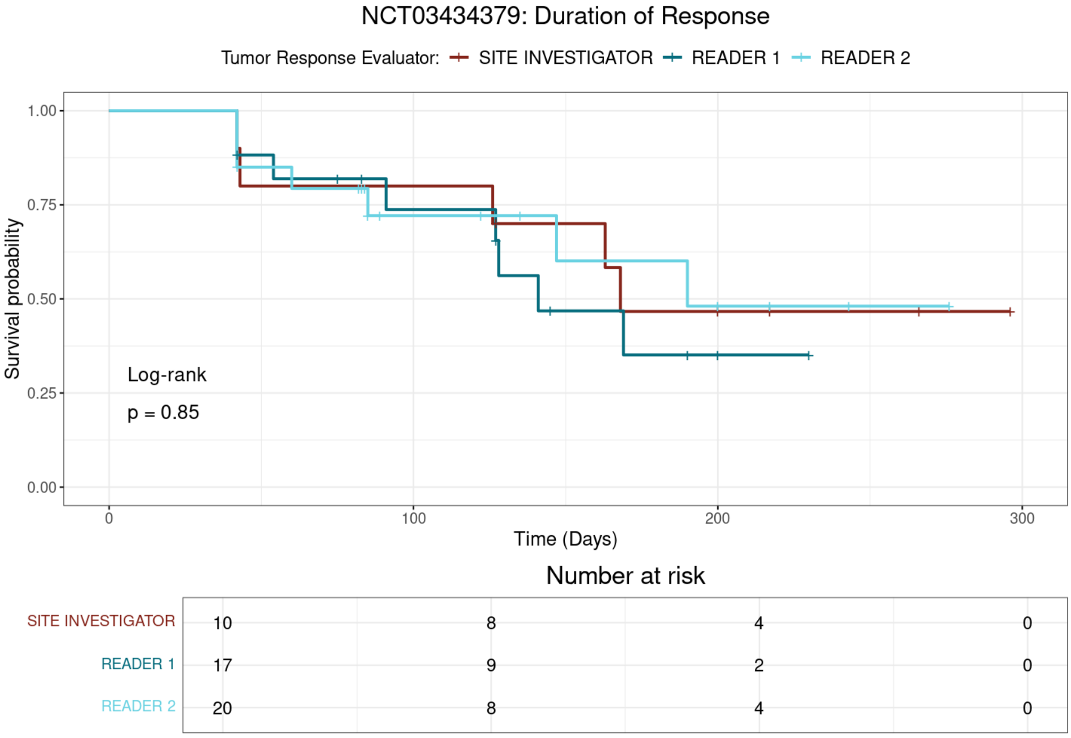


Figure 3.5.: Kaplan-Meier survival curves for DoR in NCT03434379

Notably, in study NCT03434379, where significant differences in ORR have been observed in the preceding analyses, the plot TTP (Figure 3.3) suggests that the site investigator may be quicker to assess progression than the central reviewer, as the site investigator's curve appears to be consistently below that of the central reviewers. Likewise, the site investigator might be slower to identify and classify response to treatment compared to the central reviewer, as indicated by the site investigator's curve being consistently above that of the central reviewers in the TTR plot (Figure 3.4). However, none of the other studies show such clear differences in time-to-event outcomes between raters, and the overall results suggest that, while there may be some variability in how raters assess time-to-event outcomes, these differences do not appear to be systematic or clinically meaningful.

To more accurately quantify differences in time-to-event outcomes, we employ Cox proportional hazards models to estimate hazard ratios (HRs) for each outcome. We verify the proportional hazards assumption through visual inspection of log(-log(Survival)) plots, which show approximately parallel curves between groups, and Schoenfeld residual plots, which confirm no systematic trends or deviations over time across all outcomes. These diagnostic checks support the validity of our Cox models for subsequent hazard ratio analyses. (The complete set of diagnostic plots for the Cox models, including log(-log(Survival)) plots and Schoenfeld residuals plots, are provided in the appendix at Section A.2.3.)

To formally assess differences in time-to-event outcomes, we calculate the differences in hazard ratios between site investigators and central reviewers for all studies and outcomes. As an aside on interpretation of the values: all differences are calculated as Site Investigator − Central Reviewer[1], meaning that a positive difference indicates that the site investigator has a greater hazard ratio and thus a shorter time to event occurrence, while a negative difference indicates that the central reviewer has a greater hazard ratio and thus a shorter time to event occurrence. We avoid interpreting the differences in subjective terms of e.g. "good" or "bad" because the interpretation of these differences is context-dependent and can vary based on the specific clinical scenario, and instead focus on both statistical significance and whether the differences indicate a faster or slower time to event occurrence.

Across all comparisons, no significant differences are found except for a single pairing in study NCT03434379, where the site investigator has a significantly lower hazard ratio for TTR compared to central reviewer 1. This marked difference between raters can be noted in Figure 3.4, which shows the KM survival curves for TTR in NCT03434379. In this case, the site investigator's curve is consistently and clearly above that of central reviewer 2, indicating a longer time to response for the site investigator ($p = 0.035$). This is in contrast to the log-rank test results, which do not detect the presence of a statistically significant difference between any groups.

And again, visual inspection of the TTP plot in Figure 3.3 for NCT03434379 suggests that the site investigator may be quicker to assess progression than the central reviewers, as the site investigator's curve appears to be consistently below that of the central reviewers. This difference is quantified in the calculation of hazard ratios, wherein the site investigator has a greater hazard ratio than the central reviewers. This difference trends towards significance when comparing the site investigator with central reviewer 1 ($p = 0.096$), but as can be seen in Table 3.8, none of the differences in hazard ratios are statistically significant across the studies for TTP.

[1] Cox modeling is done with the raters set as a factor (i.e. ordinal) variable in R with Site Investigator as the reference value. Thus, its estimate is always 0, the other raters' HRs are relative to this value, and the direction of the difference indicates whether the site investigator has a greater or lesser hazard ratio than the central reviewers.

| study | Contrast | estimate | SE | p.value |
|---|---|---:|---:|---|
| NCT02395172 | Site Inv. - Reader 1 | 0.109 | 0.060 | 0.166 |
| NCT02395172 | Site Inv. - Reader 2 | -0.065 | 0.047 | 0.354 |
| NCT03434379 | Site Inv. - Reader 1 | 0.242 | 0.117 | 0.096· |
| NCT03434379 | Site Inv. - Reader 2 | 0.223 | 0.111 | 0.112 |
| NCT03631706 | Site Inv. - Reader 1 | -0.117 | 0.109 | 0.532 |
| NCT03631706 | Site Inv. - Reader 2 | 0.069 | 0.100 | 0.773 |

Table 3.8.: Differences in Hazard Ratios between site investigators and central reviewers for time to progression (TTP)

| study | Contrast | estimate | SE | p.value |
|---|---|---:|---:|---|
| NCT02395172 | Site Inv. - Reader 1 | 0.000 | 0.106 | 1 |
| NCT02395172 | Site Inv. - Reader 2 | 0.006 | 0.060 | 0.995 |
| NCT03434379 | Site Inv. - Reader 1 | -0.441 | 0.317 | 0.346 |
| NCT03434379 | Site Inv. - Reader 2 | -0.667 | 0.269 | 0.035* |
| NCT03631706 | Site Inv. - Reader 1 | 0.112 | 0.092 | 0.44 |
| NCT03631706 | Site Inv. - Reader 2 | 0.069 | 0.089 | 0.719 |

Table 3.9.: Differences in Hazard Ratios between site investigators and central reviewers for time to response (TTR)

| study | Contrast | estimate | SE | p.value |
|---|---|---:|---:|---|
| NCT02395172 | Site Inv. - Reader 1 | -0.068 | 0.187 | 0.931 |
| NCT02395172 | Site Inv. - Reader 2 | 0.151 | 0.176 | 0.666 |
| NCT03434379 | Site Inv. - Reader 1 | -0.275 | 0.520 | 0.858 |
| NCT03434379 | Site Inv. - Reader 2 | -0.036 | 0.523 | 0.997 |
| NCT03631706 | Site Inv. - Reader 1 | -0.068 | 0.237 | 0.955 |
| NCT03631706 | Site Inv. - Reader 2 | 0.332 | 0.263 | 0.417 |

Table 3.10.: Differences in Hazard Ratios between site investigators and central reviewers for duration of response (DoR)

More generally, only one statistically significant difference in hazard ratios is found across all studies and outcomes, which is for TTR in NCT03434379. This suggests that, while there may be some variability in how raters assess time-to-event outcomes, these differences do not generally appear to be statistically or clinically relevant for an individual study. The results of the hazard ratio analyses are summarized in Table 3.8, Table 3.9, and Table 3.10 for TTP, TTR, and DoR, respectively.

Although few if any individual differences are found between the site investigators and central reviewers in terms of time to event outcomes, it remains possible that small, systematic differences could be detected with a larger sample size of studies and raters. To address this, meta-analyses of the time-to-event outcomes are conducted, explored in the next section. These meta-analyses aim to determine whether any systematic differences between site investigators and central reviewers could be identified across multiple studies, thereby providing a more robust assessment of inter-rater reliability for RECIST 1.1 in clinical trial settings.

### 3.2.4. Meta-Analyses Confirm Lack of Differences Between Raters

The meta-analyses of time-to-event outcomes (TTP, TTR, and DoR) reveal no statistically significant differences between site investigators and central reviewers across studies. The forest plots in Figure 3.6, Figure 3.7, and Figure 3.8 illustrate the pooled hazard ratios for each outcome, with the central vertical line indicating no difference ($\Delta HR = 0$). The squares represent the point estimates of the hazard ratios for each study, with the size of the square proportional to the weight of the study in the meta-analysis. The horizontal lines extending from each square represent the 95% confidence intervals for the hazard ratios, and the diamond at the bottom represents the pooled estimate across all studies. A diamond that does not cross the vertical line indicates a statistically significant difference, while a diamond that crosses the line indicates no significant difference.

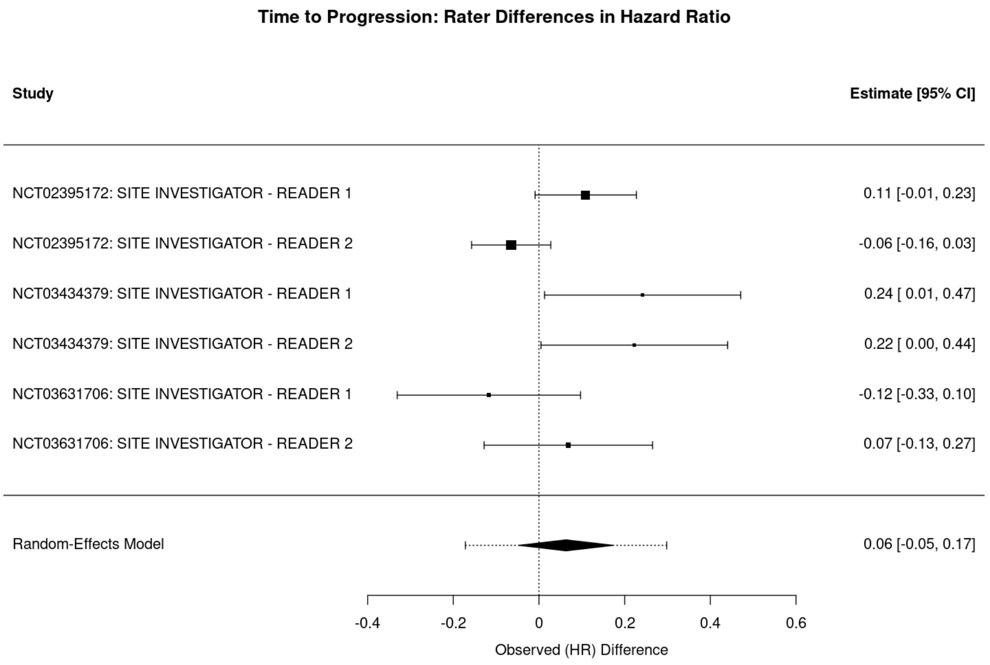


Figure 3.6.: Forest plot of hazard ratios for TTP from the meta-analysis of time to event outcomes

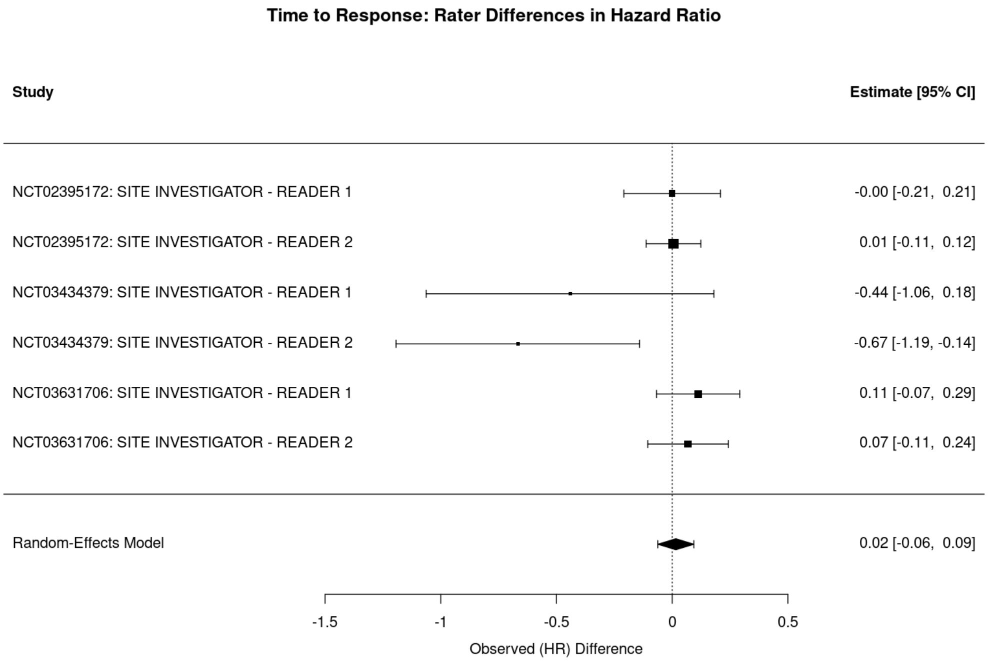


Figure 3.7.: Forest plot of hazard ratios for TTR from the meta-analysis of time to event outcomes

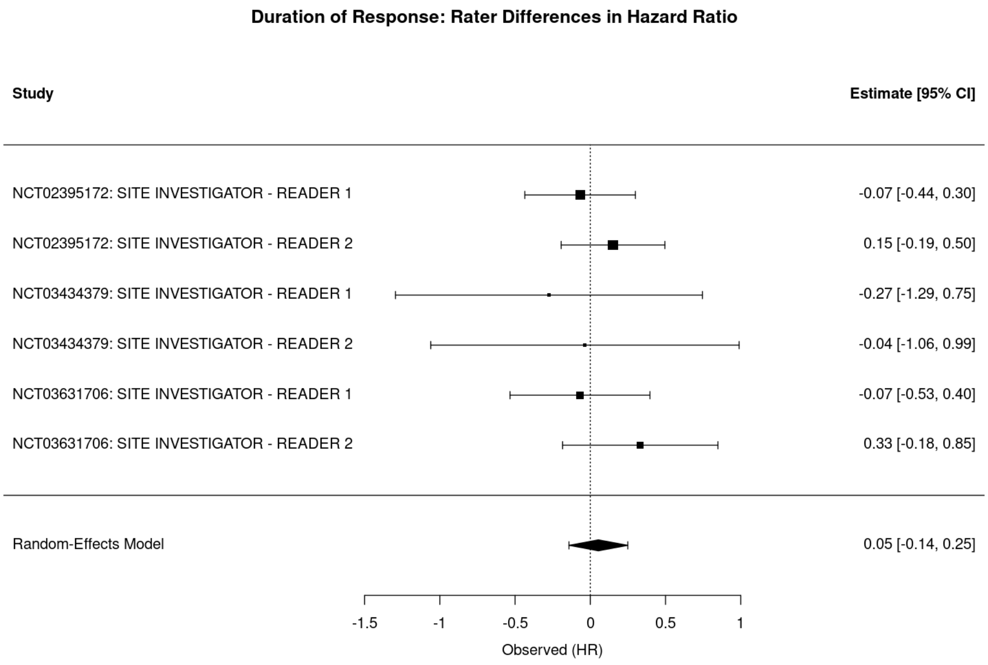


Figure 3.8.: Forest plot of hazard ratios for DoR from the meta-analysis of time to event outcomes

For each outcome, the pooled difference in hazard ratios is small and the 95% confidence intervals include zero, indicating a lack of systematic bias between rater groups. Specifically, for TTP, the difference in hazard ratios is 0.06 (95% CI: -0.05 to 0.17, p = 0.262, Figure 3.6), with moderate heterogeneity ($I^2$ = 63.4%, $^2$ = 0.0112, heterogeneity p = 0.0155, df = 5). For TTR, the difference is 0.02 (95% CI: -0.06 to 0.09, p = 0.691, Figure 3.7), with negligible heterogeneity ($I^2$ = 0.01%, $^2$ = 0.0000, heterogeneity p = 0.0739, df = 5). For DoR, the difference is 0.05 (95% CI: -0.14 to 0.25, p = 0.590, Figure 3.8), with no observed heterogeneity ($I^2$ = 0.0%, $^2$ = 0.0000, heterogeneity p = 0.771, df = 5).

Interpretation of the heterogeneity statistics further supports these findings. The $I^2$ statistic quantifies the percentage of total variation across studies attributable to heterogeneity rather than chance. For TTP, an $I^2$ of 63.4% indicates moderate heterogeneity, suggesting that some observed differences in hazard ratios may reflect real differences between studies. In contrast, $I^2$ values near zero for TTR and DoR indicate that almost all variability is due to sampling error rather than true heterogeneity. The $^2$ statistic represents the estimated between-study variance in true effect sizes, with higher values (as in TTP) indicating more variability in underlying effects, and values near zero (as in TTR and DoR) indicating little to no between-study variance. The heterogeneity test p-value, derived from Cochran's Q test, assesses whether observed variability in effect sizes exceeds what would be expected by chance. A significant p-value (e.g., p = 0.0155 for TTP) suggests real heterogeneity, while non-significant values (as in TTR and DoR) indicate no evidence of excess heterogeneity.

Across all outcomes, positive values indicate that site investigators tend to report slightly higher hazard ratios than central reviewers, suggesting more rapid assessments of progression, response, or shorter durations of response. However, as previously noted, none of these differences reach statistical significance i.e. the 95% confidence intervals for all outcomes include zero.

Although none of the meta-analyses reach significance, it is still worthwhile noting that not all data points are fully independent, as each study contributes two sets of central reviewer–site investigator pairs. This could potentially double-weight some trials and narrow confidence intervals. Nevertheless, the consistent lack of significant differences across outcomes supports the robustness of these findings. In larger meta-analyses or those with more diverse reviewer groups, the non-independence of central reviewers who likely share training, software, and protocols should be considered when interpreting results.

### 3.2.5. Equivalence Between Raters Established for TTR and TTP

To formally assess whether the outcomes measured by site investigators and central reviewers could be considered equivalent, we apply the TOST procedure to the differences in hazard ratios for TTP, TTR, and DoR. The equivalence bounds are set a priori as hazard ratios (HRs) of 0.8 to 1.25, which correspond to -0.223 to 0.223 on the log scale. These bounds represent the range of differences that would be considered not clinically meaningful, and thus, if the 90% confidence interval for the difference in log hazard ratios falls entirely within this interval, equivalence can be concluded.

The TOST plots in Figure 3.9, Figure 3.10, and Figure 3.11 summarize the results of the TOST procedure for TTP, TTR, and DoR, respectively. These plots provide a visual framework for interpreting equivalence testing results. In each plot, the dashed vertical lines represent the equivalence bounds of -0.223 to 0.223 on the log scale (corresponding to HRs of 0.8 to 1.25). The area between these lines defines the region of practical equivalence. If the 90% confidence interval for the difference in means (represented by the horizontal line with the point estimate as a dot) lies entirely within this region, equivalence is supported. If the confidence interval crosses either bound, however, equivalence cannot be established.

Our analyses show distinct patterns across the three outcomes. For TTP and TTR, the confidence intervals are fully contained within the equivalence bounds, providing strong evidence of practical equivalence between site investigators and central reviewers for these outcomes. In contrast, the DoR interval marginally exceeds the upper equivalence bound, suggesting that while the difference is likely small, we cannot formally conclude equivalence for this outcome with the available data.

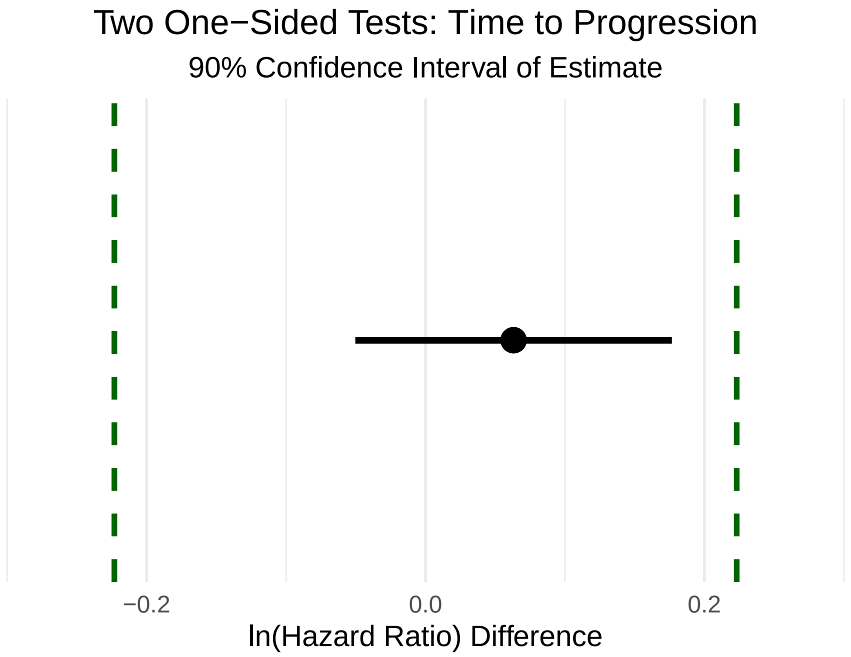


Figure 3.9.: Equivalence testing for time to progression (TTP) between site investigators and central reviewers using the TOST method.

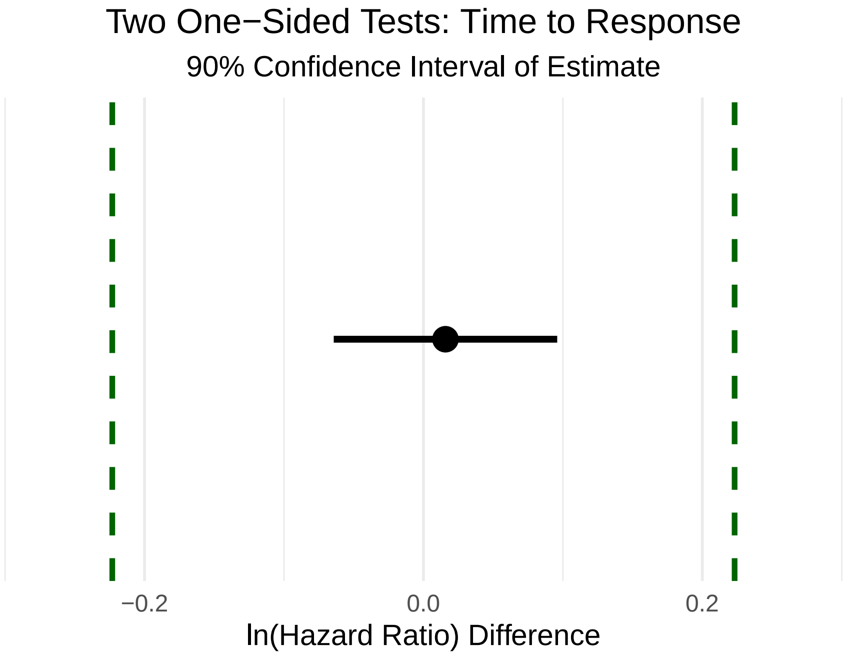


Figure 3.10.: Equivalence testing for time to response (TTR) between site investigators and central reviewers using the TOST method.

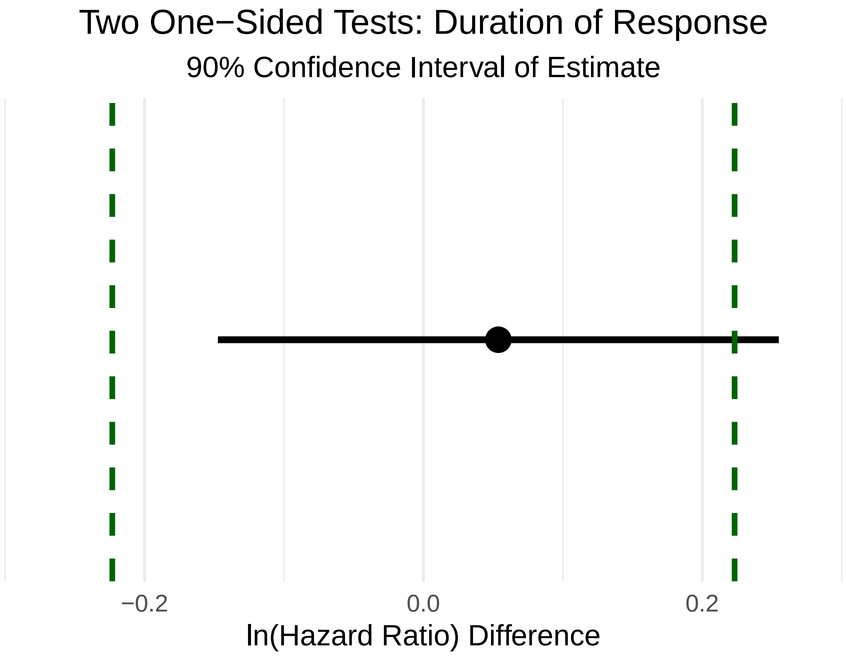


Figure 3.11.: Equivalence testing for duration of response (DoR) between site investigators and central reviewers using the TOST method.

The quantitative TOST analyses indicate that both TTP and TTR are statistically equivalent between site investigators and central reviewers. For TTP, the 90% confidence interval for the difference in log hazard ratios is [-0.05, 0.177], with an equivalence test result of t(5) = -2.840 and a p-value of 0.0181. For TTR, the 90% confidence interval is

[-0.064, 0.096], with t(5) = -5.214 and a p-value of 0.00171. In both cases, the confidence intervals fall entirely within the equivalence bounds, supporting the conclusion that any differences between site investigators and central reviewers are not clinically meaningful for these outcomes. In contrast, for DoR, the 90% confidence interval is [-0.147, 0.255] with t(5) = -1.697 and a p-value of 0.0752. Here, the upper bound of the confidence interval slightly exceeds the equivalence margin, and thus formal equivalence cannot be established. However, the interval is close to the bounds, suggesting that the two groups likely have similar durations of response, but the available data do not provide sufficient certainty to draw a definitive conclusion.

It is important to note several limitations when interpreting these results. First, the number of studies and events included in the DoR analysis is limited, resulting in wider confidence intervals and reduced statistical power to detect equivalence. This limitation is common in meta-analyses of clinical trial endpoints, especially for outcomes that are less frequently observed or reported. Second, the choice of equivalence bounds, while based on conventional thresholds for clinical relevance, remains somewhat arbitrary and may not capture all perspectives on what constitutes a meaningful difference. Finally, the TOST procedure assumes that the data are sufficiently powered to detect equivalence; in cases where sample sizes are small, failure to demonstrate equivalence may reflect limited data rather than true differences between groups.

Overall, these results provide strong evidence that site investigators and central reviewers yield equivalent results for TTP and TTR, and likely similar results for DoR, though the latter cannot be confirmed with high confidence. These findings support the robustness of RECIST-based time-to-event endpoints to rater differences in the context of clinical trials, but also highlight the need for larger studies to more definitively assess equivalence for less common outcomes such as DoR.

## 3.3. RECIST Thresholds are Robust

This section presents a series of sensitivity analyses designed to evaluate how key trial outcomes respond to changes in the RECIST thresholds for disease progression and partial response. For each analysis, we systematically recalculate RECIST-based outcomes using the available SLD data, varying the progression and response thresholds across a plausible range from [0, 100]%. The results are visualized as heatmaps, which provide an intuitive summary of how inter-rater reliability, objective response rate, and time-to-event outcomes shift as the thresholds are altered.

In these heatmaps, the axes represent the progression and partial response thresholds, while the color scale indicates either the deviation from the original estimate (for IRR and time-to-event outcomes) or the p-value from Cochran's Q test (for ORR). Rather than focusing on statistical significance at each point, these visualizations are intended to highlight the overall sensitivity or robustness of the outcomes to threshold selection. Regions of stability, sensitivity, or missing data can be readily identified, allowing for a nuanced interpretation of how RECIST criteria perform under different assumptions. Full details for interpreting each heatmap are provided in the relevant sub-sections below.

It is also important to note that, in the cases of ORR and TTR, changes in the threshold values for *progression* do not affect the corresponding heatmaps, and similarly, changes in the threshold values for *partial response* do not affect the heatmaps for TTP. Despite this, all heatmaps were plotted with both dimensions to provide a consistent visual scale for comparison across outcomes. This approach allows readers to directly compare the sensitivity of different endpoints to threshold selection, even when one axis does not influence a particular outcome.

### 3.3.1. IRR Shows Stability Across Thresholds

The IRR sensitivity analyses, visualized as heatmaps, illustrate how inter-rater reliability responds to changes in the RECIST thresholds for disease progression and partial response. For both the Target Lesion and Overall Outcome, the heatmaps display values relative to the original IRR calculated at the standard RECIST criteria (20% progression, 30% response), with intersecting black lines marking these reference thresholds. Negative values indicate a decrease in IRR compared to the original, while positive values indicate an increase.

For the Target Lesion outcome, the heatmaps (see Figure 3.12) show a clear trend: as either threshold approaches 100%, $\kappa$ values decrease. This is expected, as very few cases are classified as having a response or progression at such extreme thresholds, reducing the opportunity for agreement between raters. Additionally, in clinical trials, patients are typically removed from the study once progression is observed, making high progression thresholds less relevant and further limiting available data in these regions. Aside from this expected decline at the extremes, the heatmaps for the Target Lesion outcome do not display systematic patterns across studies, suggesting that IRR is generally stable across a wide range of plausible threshold values.

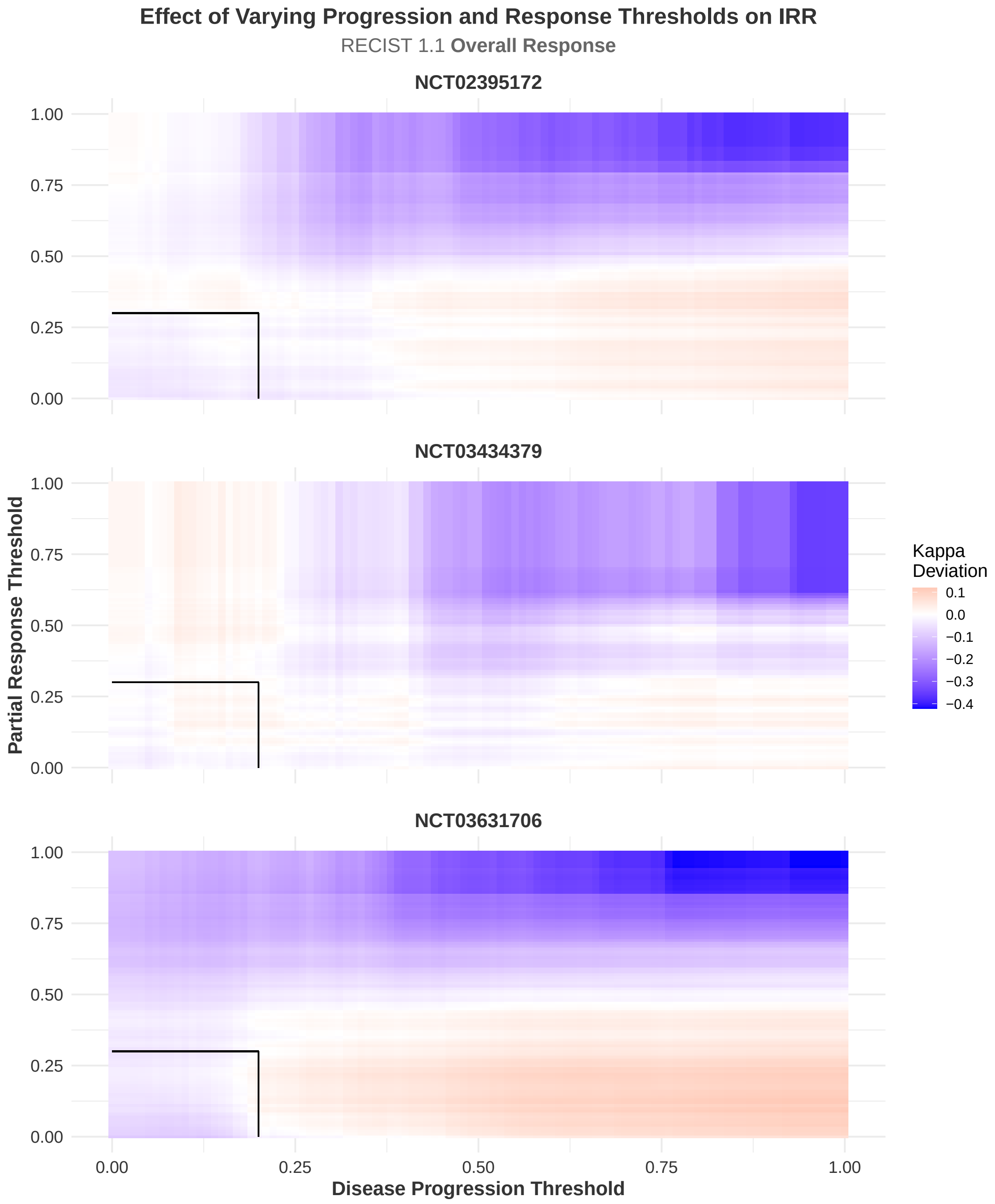


Figure 3.12.: Heatmap of change in inter-rater reliability (IRR) across RECIST thresholds, target lesions

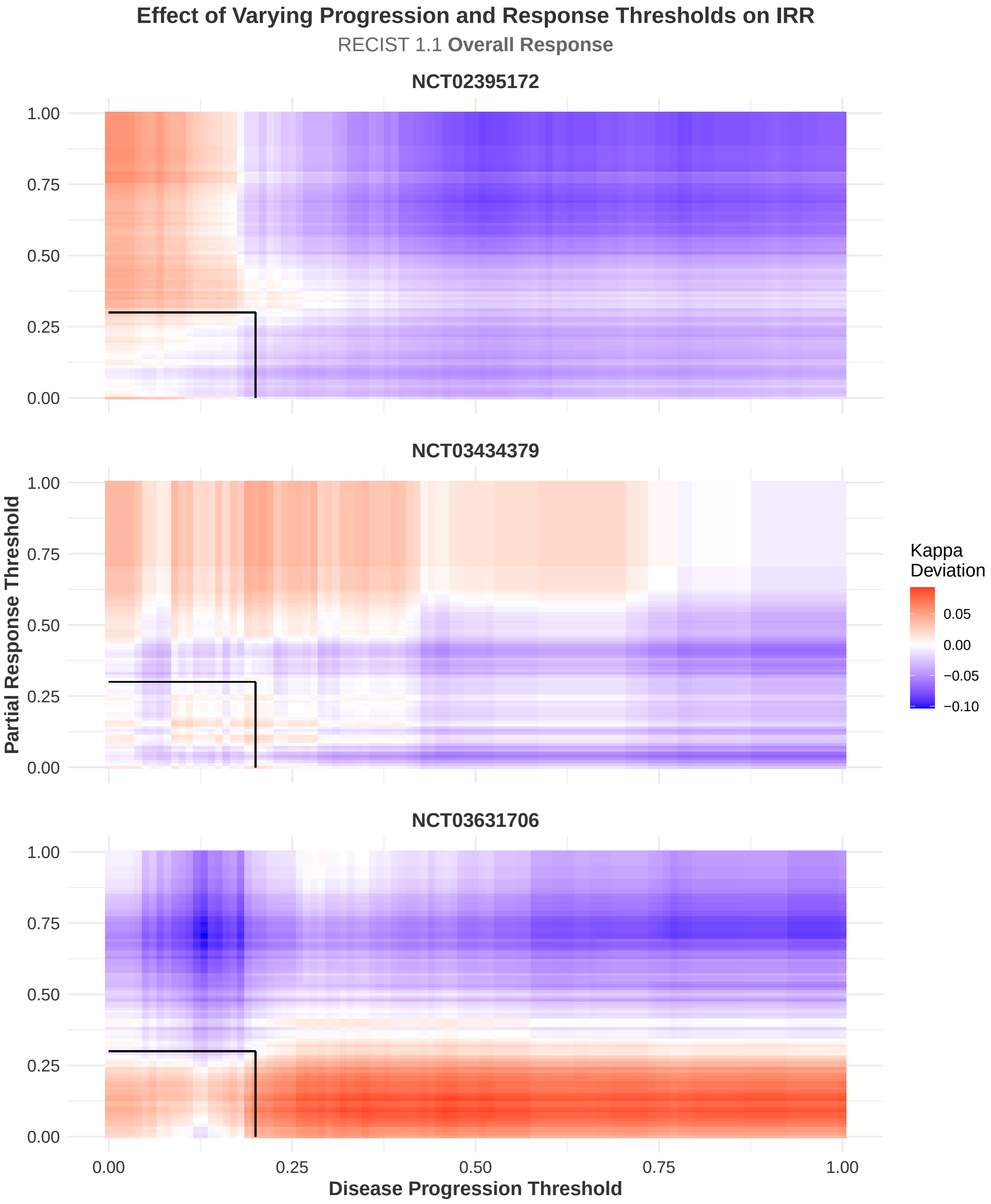


Figure 3.13.: Heatmap of change in inter-rater reliability (IRR) across RECIST thresholds, overall response

The Overall Outcome is a composite measure that incorporates information from the Target Lesion, Non-Target Lesion, and New Lesion outcomes, providing a more comprehensive assessment of tumor response. The heatmaps for the Overall Outcome (Figure 3.13) display a similar lack of systematic trends or abrupt changes in IRR as thresholds are varied. Notably, the deviation in $\kappa$ for the heatmaps of target lesions ranges from -0.422 to 0.119, while the deviation in $\kappa$ for th heatmaps of overall response ranges from -0.103 to 0.093.

This indicates that IRR can be somewhat sensitive to changes in the RECIST thresholds used for calculating the Target outcome, particularly at extreme values, but the IRR for Overall response is well-stabilized by the information gain from non-target and new lesions.

The much smaller range of $\kappa$ deviations in the overall outcome heatmaps suggests that the RECIST criteria are robust to changes in the thresholds for disease progression and response, and that the overall IRR is not significantly affected by these changes. This robustness supports the reliability of RECIST when applied in a comprehensive manner across different studies and raters.

### 3.3.2. ORR Unaffected by RECIST Thresholds

The interpretation of ORR heatmaps differs from the IRR analyses in that we focus on p-values from Cochran's Q tests rather than differences in effect estimates. This approach is selected because comparing differences in p-values across thresholds has limited interpretive value without constant reference to the conventional alpha threshold of 0.05. Instead, our analysis aims to identify regions where statistically significant differences might emerge either within or across studies as thresholds change.

While more precise statistical analyses could be performed by selecting specific clinically relevant thresholds for progression and partial response, then conducting targeted pairwise comparisons between raters at these points, such an approach would require a priori hypotheses about optimal thresholds. Without such predefined hypotheses, we instead utilized heatmaps as a visualization tool to comprehensively examine the sensitivity of ORR measurements to changes in RECIST thresholds and their subsequent effect on inter-rater differences.

For all outcomes studied in this analysis, seen in Figure 3.14, we observed no systematic changes in Cochran's Q p-values across the range of RECIST thresholds. For example, in study NCT02395172, the p-values within about ± 10% of the original response threshold remained consistently above the conventional alpha threshold of 0.05, indicating no significant differences between raters. However, nearly the opposite pattern was observed in NCT03631706, where p-values were consistently below or near 0.05 within ± 10% of the original response threshold. Above all, given the relatively small number of participants who reached a disease response at the original threshold, it is likely reasonable to conclude that most of the variation seen within plots here is due to sampling error rather than systematic differences between raters.

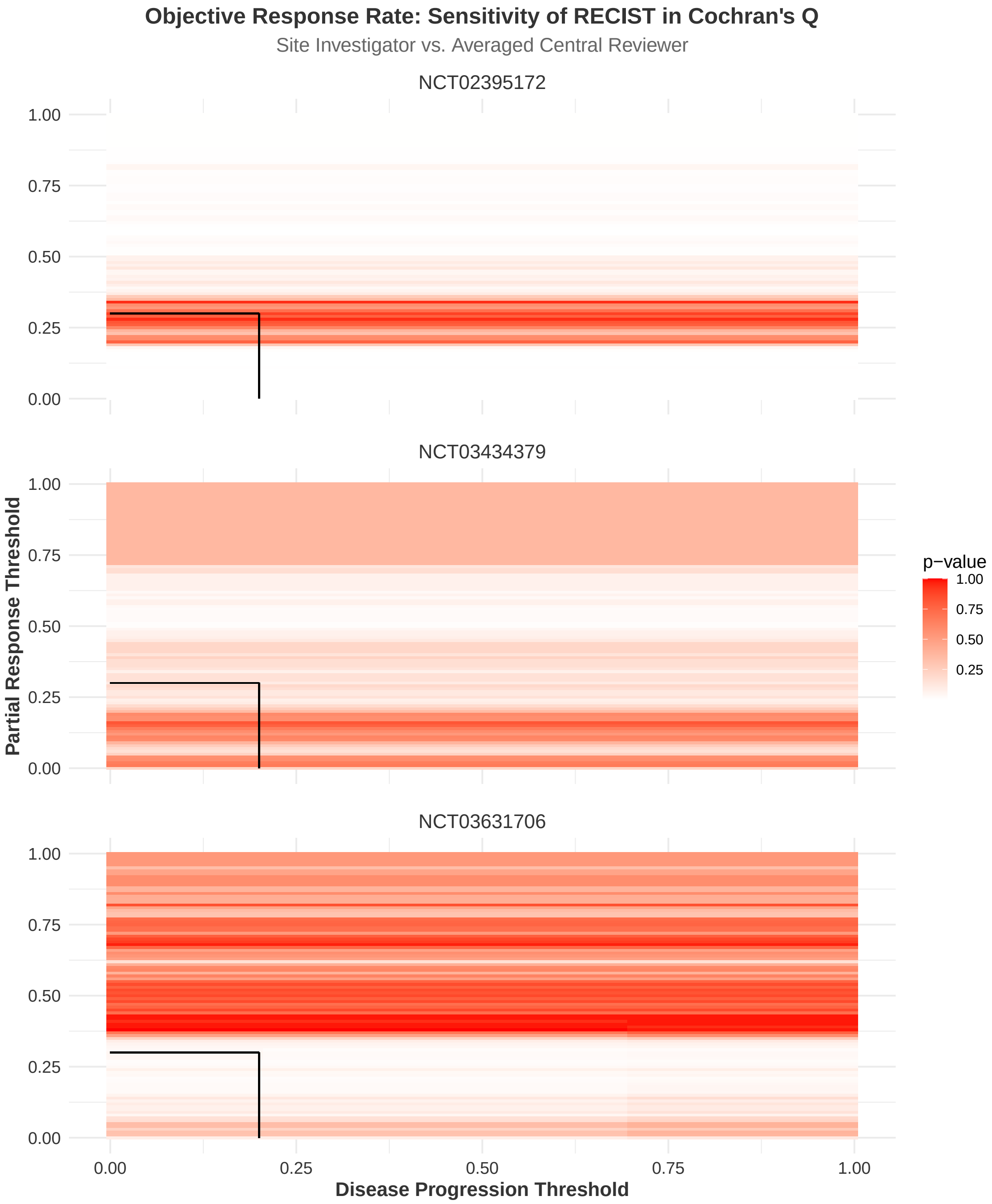


Figure 3.14.: Heatmap of Change in Differences between Raters for Objective Response Rate (ORR) across RECIST thresholds

### 3.3.3. Time-to-Event Outcomes Show no Patterns Across Changing Thresholds

To evaluate how classification threshold changes influenced the estimated risk associated with reviewer assessments across the time-to-event outcomes, we conducted a difference-in-differences (DiD) analysis comparing hazard ratios between site investigators and central reviewers. In all Cox proportional hazards models, the site investigator served as the reference group, and thus their hazard ratio remained constant across thresholds. The central reviewer's hazard ratio, by contrast, varied depending on changes in the RECIST thresholds for disease progression and response.

As noted in the Methods section in Equation 2.14, the calculations for the DiD analysis can be simplified to the following equation:

$$\Delta_{\text{RECIST}} - \Delta_{\text{Sensitivity}} = HR_{\text{Central Reviewer New}} - HR_{\text{Central Reviewer Original}}$$

where $\Delta_{\text{RECIST}}$ represents the difference in hazard ratios between the site investigator and central reviewer at the original RECIST thresholds, and $\Delta_{\text{Sensitivity}}$ represents the difference in hazard ratios at any given new threshold. The DiD value thus represents the change in the central reviewer's hazard ratio relative to the site investigator's hazard ratio as the RECIST thresholds are varied.

A positive DiD value indicates that the central reviewer's hazard ratio increased under the new threshold. Conversely, a negative DiD suggests that the central reviewer's estimated hazard decreased under the new classification rule. Additionally, it should again be noted that we refer here to a single central reviewer because we have averaged the hazard ratios of the two central reviewers *within* each study. This averaging was done to simplify the analysis and interpretation, as the two central reviewers generally agreed closely in their hazard ratio estimates and the display and interpretation of such results would have been unwieldy.

The TTP heatplot (Figure 3.15) shows generally decreasing values of DiD both when the disease progression threshold is increased and decreased with the exception of study NCT02395172. In this study, the DiD values slightly increased as the progression threshold was raised. The general overall decreases in DiD values indicate that the central reviewers showed, on average, a decrease in their estimated hazard ratios relative to the site investigators as the disease progression threshold was increased (i.e. the discrepancy in

HRs between the site investigator and central reviewer hazard ratios decreased). However, changes in any direction are not generally noticeable at all within ∼ ±5% of the original progression threshold, and large decreases in DiD values are only observed at the extremes of the progression threshold range. This suggests that the RECIST thresholds for disease progression are robust to changes in the progression threshold, and that the differences between raters would not be expected to change significantly with small changes in the progression threshold.

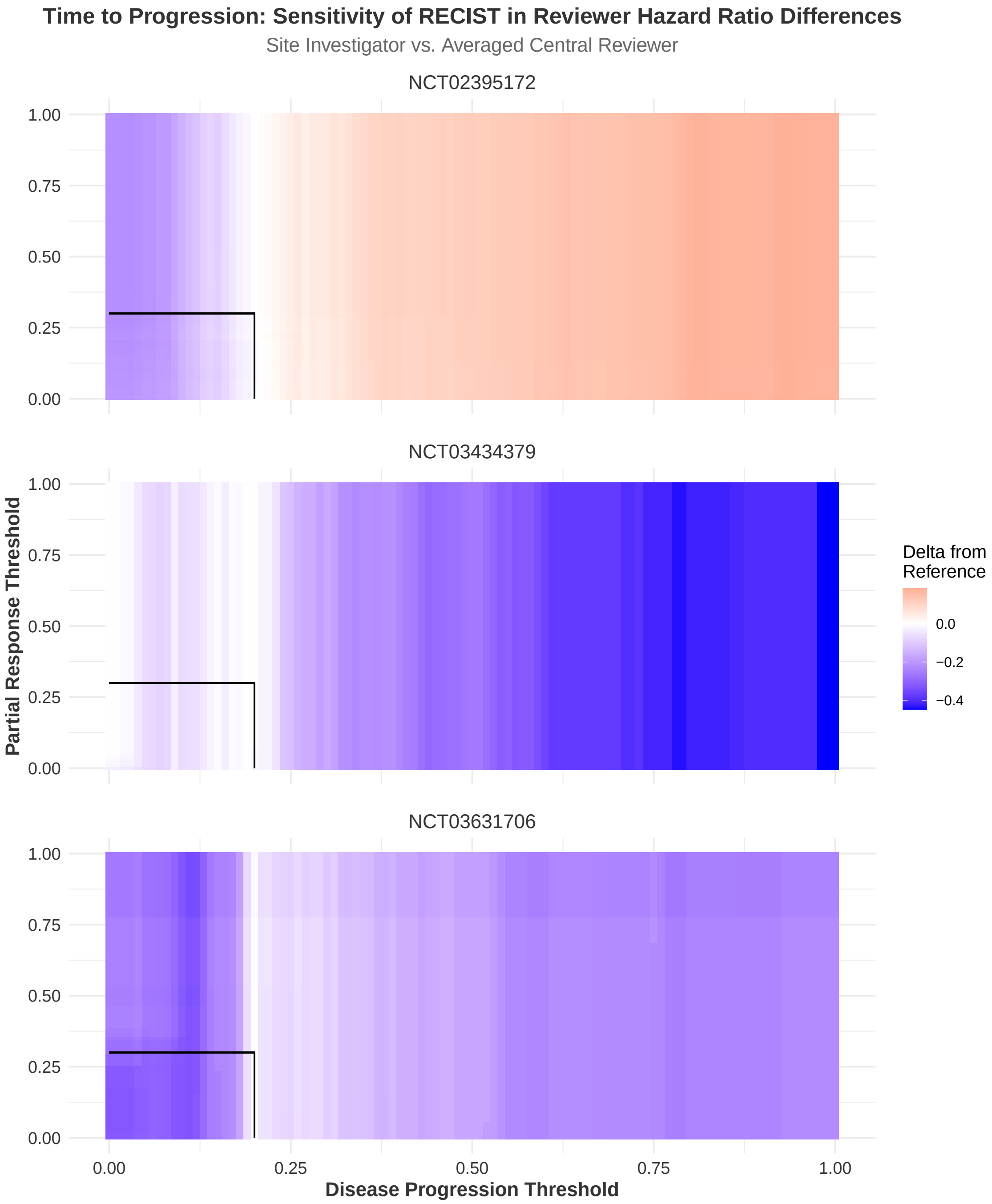


Figure 3.15.: Heatmap of Change in Differences between Raters for Time to Progression (TTP) across RECIST thresholds

Regarding the TTR and DoR sensitivity analyses, it is important to first note that the TTR and DoR heatmaps contain extensive regions of missing data points. This occurs due to the filtering procedure outlined in the Methods section (Section 2.2.3.1), where data sets with fewer than $(k-1)*10$ cases were excluded, with $k$ representing the number of raters in the study. This filtering process ensures that a reasonable number of events are present in each data set to allow for meaningful comparisons between raters. While heatmaps without such filtering are available in the appendix (Figure A.33, Figure A.34), they are not included in the main text as they do not provide additional insights beyond what is already presented here. Furthermore, some of the observed DiD values in the unfiltered maps are extremely large (exceeding an HR of 15 in some cases), making them difficult to interpret in a meaningful clinical context.

With these considerations in mind, the results of the TTR sensitivity analyses in (Figure 3.16) generally mirror those of the TTP sensitivity analyses, with the notable exception that the DiD values for TTR tend to be larger in magnitude. This increased range of DiD values likely stems from the fact that TTR events are generally less frequent than TTP events, making the DiD values more sensitive to changes in the RECIST thresholds. This interpretation is supported by the observation that up to 50% of the data points in the TTR heatmaps are missing due to the filtering procedure, while the TTP heatmaps maintain complete data coverage because a sufficient number of progression events were observed at all threshold combinations.

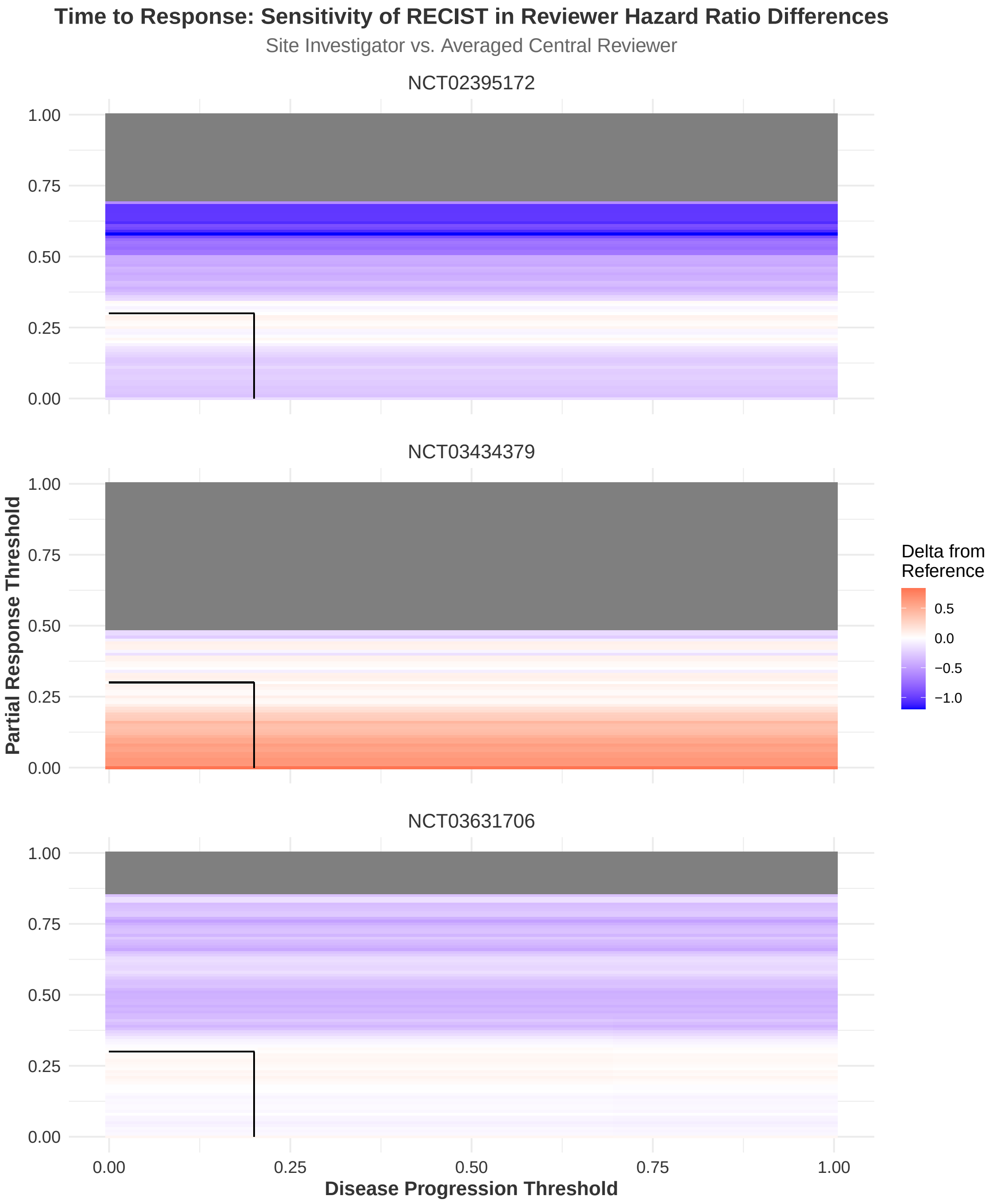


Figure 3.16.: Heatmap of Change in Differences between Raters for Time to Response (TTR) across RECIST thresholds

Interpretation of the DoR heatmaps (Figure 3.17) follows a pattern almost identical to that of the TTR heatmaps, with both displaying similar ranges of DiD values and similar patterns across the RECIST thresholds. The DoR heatmaps also exhibit a large number of missing data points due to the filtering procedure, but with an important distinction: missing data points are visible across both dimensions of the heatmap. This occurs because the DoR outcome requires both an identified response event and a subsequent progression event to calculate a duration of response. Consequently, at certain thresholds, response may be observed without progression, leading to an apparent step-down pattern from left to right in the heatmaps. This pattern is particularly pronounced in study NCT03631706.

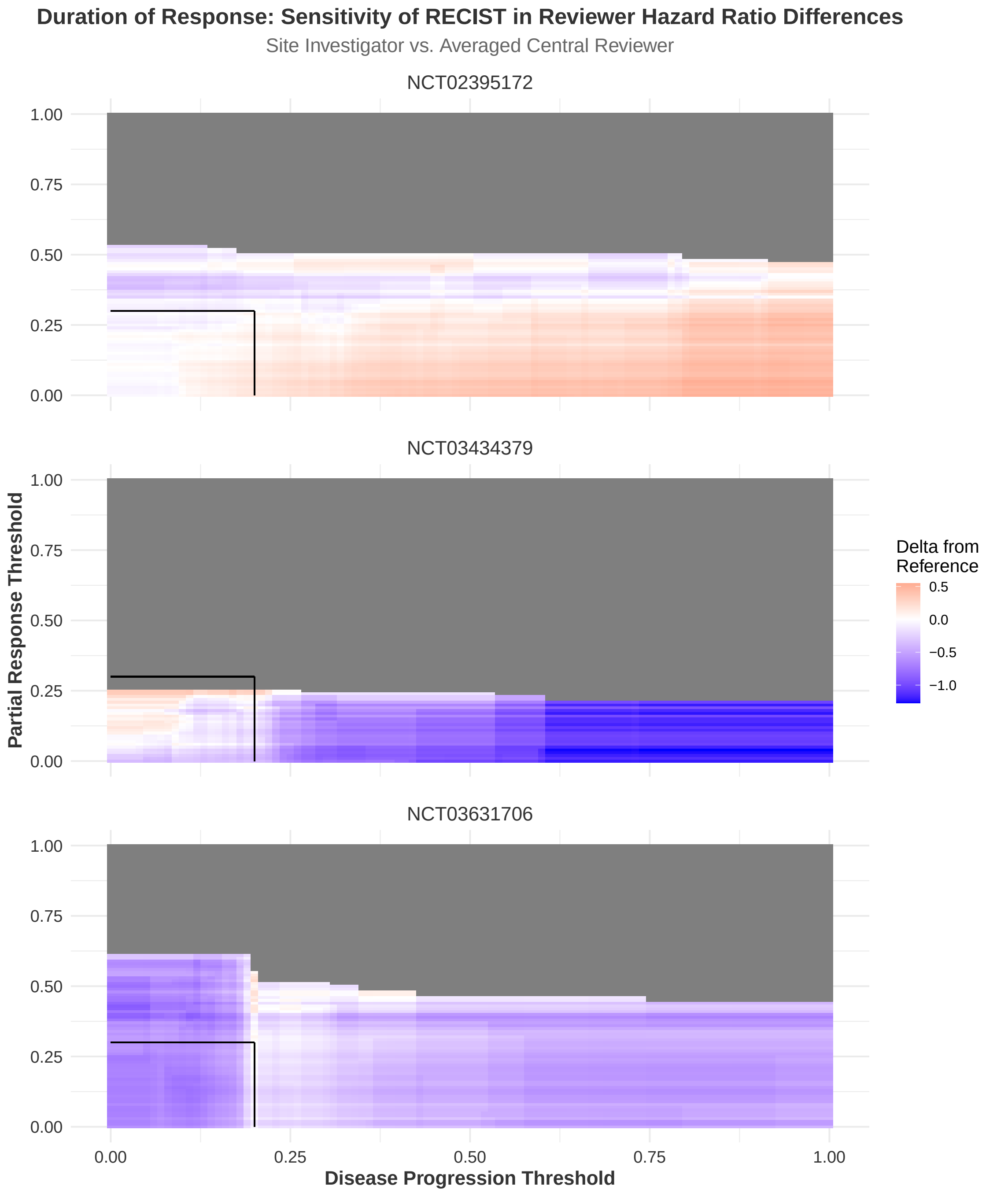


Figure 3.17.: Heatmap of Change in Differences between Raters for Duration of Response (DoR) across RECIST thresholds

Collectively, the time-to-event sensitivity analyses provide strong evidence that the RECIST thresholds for disease progression and response are robust to changes in the thresholds, with differences between raters remaining largely unaffected by threshold adjustments. These findings align with the results from both the IRR and ORR analyses, which similarly suggest that the RECIST criteria demonstrate general robustness to threshold changes and that differences between raters are neither systematic nor clinically meaningful. The heatmaps offer a clear visual representation of these findings, facilitating identification of regions where the RECIST thresholds exhibit stability or sensitivity to changes in the progression and response parameters.

In general, the DiD analyses reveal minimal systematic differences between site investigators and central reviewers across the range of RECIST thresholds examined, with exceptions occurring only at extreme threshold values. This pattern reinforces the conclusion that the RECIST criteria are robust to threshold adjustments and that differences between raters remain clinically insignificant across a reasonable range of threshold values. Overall, these sensitivity analyses provide compelling evidence for the reliability of RECIST-based time-to-event endpoints in clinical trials, supporting confidence in their continued use for assessing treatment response.

# 4. Discussion

Maintaining parallel structure with the previous chapters, this section discusses the results of the analyses in three parts that are mostly self-contained: the IRR meta-analysis, the site investigator and central reviewer analyses, and the sensitivity analyses. Each part discusses the results of the analyses and their interpretation. Overall limitations of the study are addressed in a separate subsection.

## 4.1. Results interpretation

### 4.1.1. IRR Meta-analysis

The results of our meta-analysis demonstrate that the inter-rater reliability (IRR) of the RECIST 1.1 tumor measurement scale can be considered substantial based on Landis' interpretation (77), with a pooled $\kappa$ of 0.66, Cohen's $\kappa$ of 0.67, and a pooled Fleiss' $\kappa$ of 0.65. These values indicate raters generally agree on the classification of tumor measurements, supporting the reliability of RECIST 1.1 as a standardized assessment tool across different evaluators.

However, our analysis also revealed potential limitations in the available evidence. Egger's test for funnel plot asymmetry yielded a p-value less than 0.05, suggesting the presence of publication bias in the studies included in our meta-analysis. This finding indicates that some relevant studies might have been excluded from our analysis, potentially affecting the comprehensiveness of our results and warranting caution in their interpretation. Another, potentially more likely explanation for the observed funnel plot asymmetry is that there is simply a dearth of studies analyzing IRR in the context of RECIST 1.1, particularly those that report IRR values for both site investigators and central reviewers as sponsors of clinical trials generally would not have a clear incentive to calculate or report these estimates. Such a lack of studies may lead to an overrepresentation of studies with higher

IRR values, as these are more likely to be published, while studies with lower IRR values may be underrepresented or unpublished.

Of particular note, we observed a distinct clustering pattern where data from clinical trials centered around a lower mean $\kappa$ value compared to data from non-clinical trials. This pattern suggests that the IRR of the RECIST 1.1 tumor measurement scale may be lower in clinical trial settings than in other contexts which is an important finding that raises questions about the contextual reliability of RECIST 1.1. This difference could just be an artifact of serial disagreements between raters within the same patients, but it is also likely we are simply looking at a different data context considering the composition of raters: non-clinical trial studies generally involved exclusively radiologists, while clinical trials generally have two radiologists as the central reviewers and a mix of specialists as the site investigators. This diversity in professional backgrounds and training could explain the observed variability in IRR values and highlights the need for further research to confirm these preliminary findings.

### 4.1.2. Site Investigator and Central Reviewer Analyses

Our pairwise IRR analyses reveal considerable variability in agreement between raters within and across studies, irrespective of whether they are site investigators or central reviewers. The wide range of pairwise Cohen's $\kappa$ values (0.286 to 0.803) across studies indicates some degree of inconsistency in how different raters interpret and apply RECIST criteria. However, Cohen's $\kappa$ penalizes all disagreements equally even if differences have only small clinical relevance, which can be the case for RECIST, depending on the trial outcome being used (e.g. TTP uses only progression as an indicator whereas ORR uses partial and complete response information). To account for the quasi-ordinal nature of the data, we conduct follow-up analyses using linear mixed effects models, which assign numerical values to outcomes based on their clinical favorability. These analyses identified significant differences between site investigator and one central reviewer, as well as between two central reviewers themselves, but only in one of the three studies examined. The other two studies show no significant differences between site investigators and central reviewers, suggesting quantifiable rater disagreements are the exception rather than the norm.

Overall, these preliminary analyses indicate an absence differences between site investigators and central reviewers regarding objective response rates. The variations that do exist appear attributable to random variation rather than systematic bias with the RECIST criteria, and a high degree of variability could be expected after examining the contingency tables for response rates; the observed quantity of responses to treatment were generally quite low. However, it is important to note that such within-study differences were also observed by Zhang et al. (48) in their examination of other clinical trial endpoints. Their findings alongside our analyses underscore the importance of considering potential rater discrepancies within individual trials, as they can lead to meaningful differences in outcome assessments even when systematic biases are not present across multiple studies (48).

Extending our analysis to time-to-event outcomes and hazard ratios further strengthens these conclusions. Across individual studies, we identify only a single instance of significant difference in hazard ratios between site investigators and central reviewers, with no consistent directional pattern of differences across studies. Pooling the data from all three studies, we find no statistically significant differences in hazard ratios for time-to-progression (TTP), time-to-response (TTR), or duration of response (DoR) between site investigators and central reviewers. This finding suggests that, despite the observed variability in pairwise IRR analyses, the overall agreement between rater groups remains robust when considering time-to-event outcomes. However, the absence of differences based on NHST does not necessarily imply equivalence between the two rater groups, as it is possible that differences exist but are not statistically significant due to limited sample sizes or other factors. To address this limitation and provide a more rigorous assessment, we employed formal equivalence testing using the TOST procedure, which yields particularly informative results.

Our equivalence analyses demonstrated statistical equivalence between site investigators and central reviewers for both TTP and TTR on a hazard ratio range of $[0.80, 1.25]$. However, equivalence could not be definitively established for DoR, though the data suggested potential similarity. The inability to confirm equivalence for DoR likely stems from limited sample sizes and wider confidence intervals rather than true clinical differences, as the point estimates were similar but the data lacked sufficient power to establish formal equivalence within our pre-defined margins. It is also worthwhile noting that the equivalence bounds we established could be interpreted as relatively liberal, which may have contributed to being about to establish equivalence for TTP and TTR.

Regardless, these findings again align with the work of Zhang et al. on rater agreement in tumor measurement classification although their approach lacked formal equivalence testing. Likewise, recent work by Jacobs et al. (49) focusing specifically on breast cancer trials similarly concluded that site investigators and central reviewers demonstrated good agreement in tumor measurement classification. While their methodological approach also paralleled ours in using meta-analysis to compare outcomes between rater groups, their study was more limited in scope, focusing exclusively on progression-free survival without conducting equivalence testing. Our use of formal equivalence testing and evaluation of multiple additional endpoints provides a broader understanding of RECIST reliability across different clinical contexts and allows for a more thorough assessment of IRR within the RECIST 1.1 framework.

### 4.1.3. Sensitivity Analyses

Our sensitivity analyses of Target Outcomes revealed important insights into the robustness of RECIST 1.1 criteria. While we observed some decreases in IRR when varying the assessment thresholds, these reductions only manifested at extreme threshold values that would rarely be encountered in clinical practice due to their clinical extremeness. Across all three clinical trials, we found no consistent patterns in the sensitivity analyses of Target Outcomes, which supports the conclusion that the IRR of the RECIST 1.1 tumor measurement scale remains stable across reasonable variations in the thresholds used to define target outcomes. This stability reinforces confidence in the reliability of RECIST 1.1 as a standardized assessment framework for oncology trials.

Perhaps the most significant finding from our IRR sensitivity analyses was the critical contribution of Non-Target Lesion and New Lesion measurements to the overall assessment reliability. The Overall Response classifications demonstrated markedly greater stability throughout our sensitivity analyses compared to Target Lesion measurements alone. This enhanced stability can be attributed to the additional contextual information provided by Non-Target Lesion and New Lesion measurements, which effectively compensate for potential variability in Target Lesion assessments. This finding underscores the importance of comprehensive tumor assessment in clinical trials and validates the RECIST 1.1 approach of integrating multiple types of lesion measurements into the final response determination.

With respect to trial endpoints, specifically ORR, TTP, TTR, and DoR, our sensitivity analyses further confirmed the robustness of the RECIST 1.1 framework. Across a range of alternative threshold values for disease progression and response, we observed that the RECIST 1.1 measurement scale maintained consistent performance characteristics. Notably, modifications to these thresholds produced no discernible impact on the degree of agreement (or disagreement) between site investigators and central reviewers across any of the three clinical trials examined. This consistency strongly suggests that RECIST 1.1 functions as a stable tool for tumor classification and that the currently established threshold values are appropriate for clinical trial applications. The framework's resilience to threshold adjustments indicates that variations in measurement technique or interpretation within reasonable bounds are unlikely to substantially affect trial outcomes, lending further credibility to RECIST 1.1 as a reliable standard for oncology research.

## 4.2. Study Limitations

### 4.2.1. Data Availability and Quality

Our meta-analysis faces several important limitations related to the quantity and characteristics of available data. The analysis includes only 14 studies, just above the recommended minimum of 10 studies for meta-analytic approaches, which constrains the generalizability of our findings. This limitation is compounded by the considerable heterogeneity among the included studies in terms of study design, rater populations, and cancer contexts, potentially affecting the validity of our pooled estimates. The relatively small sample size also prevents us from conducting meaningful subgroup analyses to explore the influence of potentially important confounding variables, such as cancer type which is a factor that could be particularly relevant given that imaging techniques and tumor growth patterns vary substantially across different malignancies.

For our more in-depth analyses of site investigator and central reviewer agreement, we are further limited by access to only three clinical trials. This restricted sample size inevitably limits the generalizability of our findings regarding RECIST reliability in clinical trial settings. A more fundamental limitation of the trial data is the standard imaging schedule, typically performed at 4-6 week intervals. This relatively sparse temporal sampling limits our ability to characterize tumor growth and decay patterns with precision, potentially obscuring subtle differences in assessment timing between site investigators and central reviewers. Consequently, our time-to-event analyses may lack sufficient sensitivity to detect all meaningful differences between rater groups. The use of a tumor growth model that accounts for different growth and decay patterns could have provided a more nuanced understanding of tumor dynamics, but such models are inherently complex and difficult to develop due to the non-linear nature of tumor growth.

Arguably one of the largest limitations of the clinical trial data is that only control group data were available for our analyses. This limitation restricts our ability to draw conclusions about the IRR of RECIST 1.1 in the context of active treatment, where the dynamics of tumor response may differ significantly from those observed in control groups. Future research should aim to include both control and treatment arms to provide a more comprehensive understanding of RECIST 1.1's reliability across different clinical scenarios.

An additional constraint is that our analyses focused exclusively on the RECIST 1.1 criteria, which may not be the optimal measurement scale for all tumor types. This focus limits the applicability of our findings to alternative response criteria such as iRECIST (for immunotherapy) or mRECIST (for hepatocellular carcinoma), which are increasingly used in specific therapeutic contexts. Our analytical approach could also have been expanded to include endpoints not addressed in this study. For instance, a meta-analysis similar to that conducted by Zhang et al. (48) could have been performed specifically for PFS and DCR, potentially providing additional insights into the IRR of RECIST 1.1 assessments. This represents a valuable direction for future research that could complement and extend our current findings.

### 4.2.2. Analytical Approaches

Our methodological approach to measuring IRR had several inherent limitations. While Cohen's $\kappa$ and Fleiss' $\kappa$ are widely accepted metrics for assessing agreement between raters, they do not account for potentially information such as the similarity of different levels of the outcome measure or the experience level of the raters. Additionally, for some studies, a continuous measure of IRR such as the intraclass correlation coefficient might have provided more nuanced insights into agreement patterns. However, we prioritized methodological consistency across studies, which necessitated using categorical measures of agreement that could be applied uniformly across the heterogeneous literature.

A further significant methodological challenge in our survival analyses was the use of Cox regression modeling, which required us to address violations of the independence of observations assumption. Although we implemented statistical corrections by specifying clustering of individuals, this approach, which did enable us to compare hazard ratios between raters, remains methodologically debatable. Similarly, our analyses of objective response rates might have benefited from logistic regression modeling, which would have permitted the estimation of odds ratios and corresponding confidence intervals, potentially offering a more clinically interpretable metric of agreement. However, our primary focus was on detecting the presence of differences rather than precisely quantifying their magnitude, which our chosen approach adequately accomplished.

For our analyses of ordinal RECIST data, we acknowledge that more sophisticated approaches such as ordinal logistic mixed effects models or proportional odds models might have better accounted for the inherent quasi-ordinality of response classifications. However, we deliberately employed basic linear mixed effects models as a pragmatic means to detect differences between rater groups without overcomplicating the analytical framework. This decision was justified by our subsequent time-to-event analyses, which provided more clinically relevant outcome measures. As noted earlier, tumor growth modeling that accounts for non-linear growth and decay patterns could have provided deeper insights, but the development of such models remains challenging due to the biological complexity and inter-individual variability of tumor dynamics.

A further limitation specific to our sensitivity analyses stems from the constraints inherent in clinical trial data, where information collection is bounded by patients' actual clinical i.e. data is only available up until their participation in the study is discontinued. This fundamental constraint prevented us from observing the full spectrum of possible tumor growth and decay patterns as measured by SLD, limiting our ability to comprehensively characterize the IRR of the RECIST 1.1 tumor measurement scale across all theoretical threshold values. Particularly relevant to our threshold-modifying approach was the inability to observe how patients who progressed under the standard 20% threshold might have behaved at higher progression thresholds. Such patients may have required several additional weeks or months to reach the more extreme thresholds we examined in our sensitivity analyses. While a tumor growth modeling approach could theoretically address this limitation by simulating disease trajectories beyond observed timepoints, the considerable heterogeneity in tumor behavior—characterized by non-linear growth patterns and highly variable individual responses—renders such models exceptionally difficult to develop and validate. This limitation underscores the inherent challenge in fully exploring the theoretical boundaries of measurement criteria within the constraints of real-world clinical data.

With this breakdown of the overall results and limitations of our study, we can now turn to the implications of our findings for the future of tumor response assessment in clinical trials. The next section revisits the broader context of tumor response assessment and discusses how our results can inform future research and practice in this area, as well as potential avenues for further investigation.

# 5. Conclusion

The validity and reliability of clinical trial results in oncology fundamentally depend on accurate tumor measurement and consistent assessment of disease response and progression. Without standardized evaluation criteria, trial outcomes would be subject to substantial variability, potentially leading to erroneous conclusions about treatment efficacy. RECIST was developed precisely to address this need, providing a standardized framework for tumor assessment that has been widely adopted in clinical trials globally. Since its initial introduction in 2000 and subsequent refinement to RECIST 1.1 in 2009, this framework has become the cornerstone for evaluating therapeutic responses in solid tumor oncology trials (1).

Despite RECIST 1.1's widespread adoption and critical importance in drug development, there has been surprisingly limited systematic research examining its fundamental reliability. Particularly notable is the absence of comprehensive meta-analyses synthesizing inter-rater reliability (IRR) data across multiple studies and contexts. This knowledge gap is significant because consistent measurement and interpretation between different raters, whether at the same institution or across different trial sites, is essential for ensuring that reported treatment effects reflect genuine biological responses rather than measurement inconsistencies or subjective interpretations.

That is not to say that no work has been done in this domain; indeed, previous research has made valuable contributions to our understanding of RECIST reliability, particularly within the context of clinical trials. For example, Zhang et al. (48) and, more recently, Jacobs et al. (49) conducted meta-analyses investigating differences between site investigators and central reviewers in several key clinical trial outcome measures. These studies provide important insights into potential discrepancies in tumor assessments, with an absolutely critical conclusion by Zhang et al. that "statistically inconsistent inferences could be made in many trials" depending on whether the site investigator or central reviewer assessments are used (48). However, these studies are limited because the range of trial outcomes analyzed is constrained and neither study offers an analysis of equivalence between rater groups.

An additional critical gap in the literature concerns the empirical impact of RECIST's defined threshold values. The 30% decrease threshold for partial response and 20% increase threshold for progressive disease were established based on expert consensus rather than extensive empirical validation. To date, no comprehensive research has investigated whether these specific thresholds might systematically influence agreement between raters or whether alternative thresholds might yield more consistent assessments. This question is not merely academic as even small changes in these threshold values could potentially affect trial outcomes and subsequent treatment approvals.

This thesis addresses these important gaps through three complementary analytical approaches. First, it presents a comprehensive meta-analysis of IRR studies to establish an overall estimate of RECIST 1.1's reliability across diverse contexts. Second, it performs a detailed comparison of site investigator versus central reviewer assessments across multiple clinical trial endpoints, including several not previously examined in the literature. Finally, it conducts novel sensitivity analyses exploring how variations in the disease response and progression threshold values might affect the classification consistency of tumor measurements between raters. Together, these analyses provide a more comprehensive evaluation of RECIST 1.1's reliability than has previously been available, with important implications for the interpretation of clinical trial results and potential refinements to the RECIST framework.

Our meta-analysis reveals that the IRR of the RECIST 1.1 criteria demonstrates substantial agreement between raters, with a pooled Cohen's and Fleiss' $\kappa$ estimate of 0.66. According to the widely accepted Landis and Koch interpretive scale (77), this value indicates substantial agreement, suggesting that RECIST 1.1 provides reasonably consistent tumor assessments across different evaluators. However, this level of agreement, while encouraging, still leaves room for improvement in measurement consistency especially when considering how clearly defined the RECIST criteria are. Of particular concern is a notable pattern in our data: the four clinical trial studies included in our meta-analysis consistently show lower IRR values compared to studies conducted in non-clinical trial settings. This finding raises important questions about whether RECIST 1.1 reliability may be context-dependent, with potentially lower consistency in the high-stakes environment of clinical trials where diverse specialists (oncologists and radiologists) may be evaluating the same images with different training backgrounds. With more data on RECIST reliability in clinical trial contexts, this observation could be more deeply investigated to determine whether specific factors contribute to the observed variability such as compounding discrepancies due to serial assessments of the same patients.

From a methodological perspective, we also contribute to the statistical foundations of $\kappa$ analysis by developing a more mathematically rigorous approach for scaling $\kappa$ values to logit values using the delta method. This methodological advancement extends previous work that assumed $\kappa$ values were bounded to the interval [0, 1], clarifying that the full theoretical range of $\kappa$ is [-1, 1]. While in most practical RECIST applications we expect positive agreement values, our method provides a more statistically sound framework for meta-analytic comparisons that may encounter unusual cases of systematic disagreement.

To further examine RECIST reliability exclusively within clinical trial contexts, we conducted detailed comparisons between site investigator and central reviewer response assessments. Although there are detectable differences in ORR determinations in some cases, the direction and magnitude of these differences vary inconsistently across studies and may be an artefact of a small sample of response to treatment within the trials. More tellingly, an analysis of pairwise IRR between site investigators and central reviewers reveals a broad range of $\kappa$ values, indicating variability in agreement. Notably, in two of the three clinical trials we examined, the pairwise $\kappa$ values between site investigators and central reviewers are lower than those between different central reviewers. This pattern suggests a systematic tendency for greater consistency among specialized central reviewers than between site investigators and central review teams although the sample size is small and the results should be interpreted with caution. This finding highlights the potential value of central review processes in enhancing measurement reliability, particularly in complex clinical trial settings where multiple raters may interpret tumor images differently.

To determine whether observed differences in tumor classification have meaningful clinical implications, we modeled time-to-event outcomes using Cox regression models. Our analysis of TTP, TTR, and DoR yields particularly informative results. No significant differences in hazard ratios are detected between site investigators and central reviewers across these critical endpoints. More definitively, using two one-sided tests (TOST) for equivalence, we demonstrate statistical equivalence between site investigators and central reviewers for both TTP and TTR outcomes. Although equivalence could not be established for DoR, the point estimates are similar, suggesting that the inability to confirm equivalence likely stems from limited sample sizes and wider confidence intervals rather than potential clinical differences.

These time-to-event findings align with previous research by Zhang et al. (48) showing general concordance between rater groups for primary trial endpoints. However, our results also confirm an important nuance: while average agreement across multiple trials appears to be consistent, individual trials can exhibit meaningful discrepancies between site and central assessments. Such trial-specific variations, though statistically expected in some proportion of studies, can have profound implications for the interpretation of results and the determination of treatment efficacy (48). This observation lends support to the continued practice of BICR in clinical trials to ensure data quality.

Our sensitivity analysis of the RECIST 1.1 threshold values confirms the stability of current standards as the established disease response (30% decrease) and progression (20% increase) thresholds appeared to be robust across testing. That is, when we vary these thresholds within clinically reasonable ranges, we find no substantial effects on classification consistency or inter-rater agreement, and we only observed large deviations in agreement between raters at the extremes of classification thresholds. This indicates that the consensus-derived RECIST thresholds are appropriate and stable across different clinical contexts and rater groups. However, this analysis is limited by available data, particularly for extreme threshold values, because this clinical trial data only includes patients up until the point of disease progression, and therefore we cannot assess the full range of RECIST classifications. Future research could explore the effects of more extreme threshold values on classification consistency, particularly in trials involving treatments that may not produce the expected tumor shrinkage or growth patterns, and this could potentially be accomplished through simulation studies or additional empirical data collection.

The collective evidence from our three analytical approaches supports several important conclusions about RECIST 1.1 reliability. First, RECIST 1.1 demonstrates substantial overall reliability as reflected in good $\kappa$ values, though with room for improvement, particularly in clinical trial settings where we observed consistently lower agreement. Second, differences between site investigators and central reviewers, while present, generally do not translate into clinically meaningful differences in trial endpoints, with the important caveat that individual trials may still exhibit consequential discrepancies. Third, the current RECIST threshold values appear empirically justified, showing stability across various analytical conditions.

It is important to recognize that this study hardly acknowledges a broader contextual issue regarding the type of treatment received and the relevance of RECIST criteria to different treatment methodologies; RECIST was originally developed when chemotherapy (cytotoxic treatments) and surgical interventions represented the primary available options, and the expected response pattern (i.e. apoptosis and clearing of dead cells) as a result of chemotherapy meant that measuring tumor growth or decline could be assumed as an adequate proxy for disease response (2,68,69). However, the contemporary therapeutic landscape has evolved dramatically, introducing treatment modalities that may not necessarily produce the tumor shrinkage that RECIST was designed to detect (21). This study alone examines trials involving cytotoxic, cytostatic, and immunotherapy treatment modalities (Table 2.1), representing only three of the apparent seven "pillars of cancer therapy," which also include surgery, radiotherapy, hormonal therapy, and cell therapy (69). Cytostatic agents, for example, work by halting tumor growth rather than causing regression, while immunotherapies may initially cause tumor swelling due to immune cell infiltration before any reduction occurs (69). Therefore, one might reasonably expect differential utility of RECIST depending on the treatment type, with potentially decreasing utility as cytostatic and immunotherapies become more prevalent in clinical practice, given that these treatments may achieve therapeutic benefit without the measurable tumor shrinkage that RECIST criteria prioritize. Future studies should explore how RECIST 1.1 performs across these diverse treatment modalities, particularly in the context of immunotherapies and other novel approaches that may not conform to traditional tumor response patterns.

Furthermore, RECIST may soon face competition from emerging technological approaches, including software that can automatically segment tumors (42,108) and three-dimensional methods that might better assess overall tumor burden (109). In effect, the employment of these advanced techniques could potentially replace the need for human raters in some contexts, thereby enhancing measurement consistency and reducing variability. However, these innovations also raise important questions about how to integrate new technologies with existing frameworks like RECIST 1.1, particularly regarding the interpretation of results and the establishment of new standards for tumor assessment. While we intentionally focused our analysis on the reliability of current RECIST 1.1 criteria to avoid diluting the primary aims of this thesis, these broader considerations will require careful attention in future research as the tumor assessment landscape continues to evolve.

Looking forward, our findings suggest several promising directions for improving RECIST implementation. A particularly valuable approach would be to explore collaborative methodologies for tumor identification at baseline, as demonstrated by Oubel et al. (44), involving enhanced coordination between site investigators and central reviewers. This strategy could directly address the observed trend of lower inter-rater reliability in clinical trial settings by establishing shared understanding of target lesions from the outset of treatment. By developing standardized protocols for collaborative baseline assessments, the oncology research community could potentially enhance measurement consistency throughout the treatment course, further strengthening the reliability of this critical evaluation framework.

Such methodological innovations would complement RECIST's solid foundation while addressing the specific contexts where our research identified opportunities for improvement. As cancer therapeutics continue to advance and diversify, maintaining and enhancing the reliability of response assessment tools like RECIST 1.1 remains essential for generating valid and generalizable clinical trial results. And despite the introduction of competing assessment frameworks, RECIST 1.1 remains the gold standard for tumor response evaluation in oncology trials, with its rigorous methodology and established reliability continuing to underpin the development of new cancer treatments.

# References

1. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer [Internet]. 2009 Jan 1;45(2):228–47. Available from: https://www.ejcancer.com/article/S0959-8049(08)00873-3/fulltext

2. Cooper GM. The Development and Causes of Cancer. In Sinauer Associates; 2000. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9963/

3. Wainwright EN, Scaffidi P. Epigenetics and cancer stem cells: Unleashing, hijacking, and restricting cellular plasticity. Trends in Cancer [Internet]. 2017 May;3(5):372–86. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5506260/

4. Rodriguez-Brenes IA, Komarova NL, Wodarz D. Tumor growth dynamics: Insights into evolutionary processes. Trends in Ecology & Evolution [Internet]. 2013 Oct 1;28(10):597–604. Available from: https://www.sciencedirect.com/science/article/pii/S0169534713001420

5. Gerstberger S, Jiang Q, Ganesh K. Metastasis. Cell [Internet]. 2023 Apr 13;186(8):1564–79. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10511214/

6. Fares J, Fares MY, Khachfe HH, Salhab HA, Fares Y. Molecular principles of metastasis: a hallmark of cancer revisited. Signal Transduction and Targeted Therapy [Internet]. 2020 Mar 12;5(1):28. Available from: https://www.nature.com/articles/s41392-020-0134-x

7. Gavish A, Tyler M, Simkin D, Kovarsky D, Castro LNG, Halder D, et al. The transcriptional hallmarks of intra-tumor heterogeneity across a thousand tumors.

8. Curtin SC, Tejada-Vera B, Bastian BA. Deaths: Leading Causes for 2020. National vital statistics reports. 2023 Dec 1;72(13):1–115.

9. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians [Internet]. 2024;74(3):229–63. Available from: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834

10. Ma Z, Richardson LC. Cancer screening prevalence and associated factors among US adults. Preventing Chronic Disease [Internet]. 2022 Apr 21;19:E22. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9044902/

11. Chen S, Cao Z, Prettner K, Kuhn M, Yang J, Jiao L, et al. Estimates and projections of the global economic cost of 29 cancers in 204 countries and territories from 2020 to 2050. JAMA Oncology [Internet]. 2023 Apr 1;9(4):465–72. Available from: https://doi.org/10.1001/jamaoncol.2022.7826

12. Are key funders of cancer research slowing down their spending? [Internet]. 2025. Available from: https://www.nature.com/nature-index/news/funding-cancer-research-grant-trends-investment

13. Gaind N. How the NIH dominates the world's health research — in charts. Nature [Internet]. 2025 Mar 10;639(8055):554–5. Available from: https://www.nature.com/articles/d41586-025-00754-4

14. Liu B, Zhou H, Tan L, Siu KTH, Guan XY. Exploring treatment options in cancer: tumor treatment strategies. Signal Transduction and Targeted Therapy [Internet]. 2024 Jul 17;9(1):175. Available from: https://www.nature.com/articles/s41392-024-01856-7

15. Chakraborty S, Rahman T. The difficulties in cancer treatment. ecancermedicalscience [Internet]. 2012 Nov 14;6:ed16. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4024849/

16. Siamof CM, Goel S, Cai W. Moving beyond the pillars of cancer treatment: Perspectives from nanotechnology. Frontiers in Chemistry [Internet]. 2020 Nov 10;8:598100. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7683771/

17. Bailly C, Thuru X, Quesnel B. Combined cytotoxic chemotherapy and immunotherapy of cancer: Modern times. NAR Cancer [Internet]. 2020 Feb 17;2(1):zcaa002. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8209987/

18. Sonkin D, Thomas A, Teicher BA. Cancer treatments: Past, present, and future. Cancer Genetics [Internet]. 2024 Aug 1;286-287:18–24. Available from: https://www.sciencedirect.com/science/article/pii/S2210776224000243

19. Lim ZF, Ma PC. Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. Journal of Hematology & Oncology [Internet]. 2019 Dec 9;12(1):134. Available from: https://doi.org/10.1186/s13045-019-0818-2

20. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nature Reviews Clinical Oncology [Internet]. 2018 Feb;15(2):81–94. Available from: https://www.nature.com/articles/nrclinonc.2017.166

21. Ma Y, Wang Q, Dong Q, Zhan L, Zhang J. How to differentiate pseudoprogression from true progression in cancer patients treated with immunotherapy. American Journal of Cancer Research [Internet]. 2019 Aug 1;9(8):1546–53. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6726978/

22. Junod S. FDA and clinical drug trials: A short history. FDLI Update [Internet]. 2008;2008:55. Available from: https://heinonline.org/HOL/Page?handle=hein.journals/fdliup2008&id=123&div=&collection=

23. Zubrod CG, Schneiderman M, Frei E, Brindley C, Lennard Gold G, Shnider B, et al. Appraisal of methods for the study of chemotherapy of cancer in man: Comparative therapeutic trial of nitrogen mustard and triethylene thiophosphoramide. Journal of Chronic Diseases [Internet]. 1960 Jan 1;11(1):7–33. Available from: https://www.sciencedirect.com/science/article/pii/0021968160901375

24. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. Cancer [Internet]. 1981;47(1):207–14. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-0142%2819810101%2947%3A1%3C207%3A%3AAID-CNCR2820470134%3E3.0.CO%3B2-6

25. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors. Journal of nuclear medicine : official publication, Society of Nuclear Medicine [Internet]. 2009 May;50(Suppl 1):122S. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2755245/

26. Choi JH, Ahn MJ, Rhim HC, Kim JW, Lee GH, Lee YY, et al. Comparison of WHO and RECIST Criteria for Response in Metastatic Colorectal Carcinoma. Cancer Research and Treatment : Official Journal of Korean Cancer Association [Internet]. 2005 Oct 31;37(5):290. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2785927/

27. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. JNCI: Journal of the National Cancer Institute [Internet]. 2000 Feb 2;92(3):205–16. Available from: https://doi.org/10.1093/jnci/92.3.205

28. Ahmed R. Ensuring Quality Medicine: A Comprehensive Overview of EMA and DGDA's History, Structure, and Functions. RADINKA JOURNAL OF HEALTH SCIENCE [Internet]. 2024 Dec 31;2(2):254–66. Available from: https://rjupublisher.com/ojs/index.php/RJHS/article/view/362

29. Brown DG, Wobst HJ, Kapoor A, Kenna LA, Southall NT. Clinical development times for innovative drugs. Nature reviews Drug discovery [Internet]. 2022 Nov;21(11):793–4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9869766/

30. Mahan VL. Clinical Trial Phases. International Journal of Clinical Medicine [Internet]. 2014 Dec 4;05(21):1374. Available from: http://www.scirp.org/journal/PaperInformation.aspx?PaperID=52733&#abstract

31. Zhao B, Lee SM, Lee HJ, Tan Y, Qi J, Persigehl T, et al. Variability in assessing treatment response: Metastatic colorectal cancer as a paradigm. Clinical cancer research : an official journal of the American Association for Cancer Research [Internet]. 2014 Jul 1;20(13):3560–8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4337392/

32. Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. Statistics in medicine [Internet]. 2012 Nov 10;31(25):2973–84. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3551627/

33. Delgado A, Guddati AK. Clinical endpoints in oncology - a primer. American Journal of Cancer Research [Internet]. 2021 Apr 15;11(4):1121–31. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8085844/

34. Kemp R, Prasad V. Surrogate endpoints in oncology: When are they acceptable for regulatory and clinical decisions, and are they currently overused? BMC Medicine [Internet]. 2017 Jul 21;15:134. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5520356/

35. Zettler M, Basch E, Nabhan C. Surrogate end points and patient-reported outcomes for novel oncology drugs approved between 2011 and 2017. JAMA Oncology [Internet]. 2019 Sep;5(9):1358–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6613294/

36. Food, Drug Administration C for DE and. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics [Internet]. 2018. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics

37. European Medicines Agency. Guideline on the clinical evaluation of anticancer medicinal products. 2023 Nov 18; Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-evaluation-anticancer-medicinal-products-revision-6_en.pdf

38. Fournier L, Geus-Oei LF de, Regge D, Oprea-Lager DE, D'Anastasi M, Bidaut L, et al. Twenty years on: RECIST as a biomarker of response in solid tumours an EORTC imaging group – ESOI joint paper. Frontiers in Oncology [Internet]. 2022 Jan 10;11:800547. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8784734/

39. Ruchalski K, Braschi-Amirfarzan M, Douek M, Sai V, Gutierrez A, Dewan R, et al. A primer on RECIST 1.1 for oncologic imaging in clinical drug trials. Radiology: Imaging Cancer [Internet]. 2021 May 14;3(3):e210008. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8183261/

40. Kuhl CK, Alparslan Y, Schmoee J, Sequeira B, Keulers A, Brümmendorf TH, et al. Validity of RECIST Version 1.1 for Response Assessment in Metastatic Cancer: A Prospective, Multireader Study. Radiology. 2019 Feb;290(2):349–56.

41. El Homsi M, Bou Ayache J, Fernandes MC, Horvat N, Kim TH, LaGratta M, et al. Comparison of abbreviated and complete MRI protocols for treatment response assessment of colorectal liver metastases. European Radiology [Internet]. 2024 Dec 10; Available from: https://doi.org/10.1007/s00330-024-11277-3

42. Baidya Kayal E, Kandasamy D, Yadav R, Bakhshi S, Sharma R, Mehndiratta A. Automatic segmentation and RECIST score evaluation in osteosarcoma using diffusion MRI: A computer aided system process. European Journal of Radiology. 2020 Dec;133:109359.

43. Abramson RG, McGhee CR, Lakomkin N, Arteaga CL. Pitfalls in RECIST data extraction for clinical trials: Beyond the basics. Academic radiology [Internet]. 2015 Jun;22(6):779–86. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4429002/

44. Oubel E, Bonnard E, Sueoka-Aragane N, Kobayashi N, Charbonnier C, Yamamichi J, et al. Volume-based response evaluation with consensual lesion selection: A pilot study by using cloud solutions and comparison to RECIST 1.1. Academic Radiology [Internet]. 2015 Feb 1;22(2):217–25. Available from: https://www.sciencedirect.com/science/article/pii/S1076633214003729

45. Ford R, Schwartz L, Dancey J, Dodd LE, Eisenhauer EA, Gwyther S, et al. Lessons learned from independent central review. European Journal of Cancer [Internet]. 2009 Jan 1;45(2):268–74. Available from: https://www.ejcancer.com/article/S0959-8049(08)00879-4/fulltext

46. Beaumont H, Evans TL, Klifa C, Guermazi A, Hong SR, Chadjaa M, et al. Discrepancies of assessments in a RECIST 1.1 phase II clinical trial – association between adjudication rate and variability in images and tumors selection. Cancer Imaging [Internet]. 2018 Dec 11;18:50. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6288919/

47. Dodd LE, Korn EL, Freidlin B, Jaffe CC, Rubinstein LV, Dancey J, et al. Blinded independent central review of progression-free survival in phase III clinical trials: important design element or unnecessary expense? Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2008 Aug 1;26(22):3791–6.

48. Zhang J, Zhang Y, Tang S, Jiang L, He Q, Hamblin LT, et al. Systematic bias between blinded independent central review and local assessment: Literature review and analyses of 76 phase III randomised controlled trials in 45 688 patients with advanced solid tumour. BMJ Open [Internet]. 2018 Sep 10;8(9):e017240. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6144327/

49. Jacobs F, Molinelli C, Martins-Branco D, Marta GN, Salmon M, Ameye L, et al. Progression-free survival assessment by local investigators versus blinded independent central review in randomized clinical trials in metastatic breast cancer: A systematic review and meta-analysis. European Journal of Cancer [Internet]. 2024 Jan 1;197:113478. Available from: https://www.sciencedirect.com/science/article/pii/S0959804923007803

50. Gill D. Re-inventing clinical trials through TransCelerate. Nature Reviews Drug Discovery [Internet]. 2014 Nov;13(11):787–8. Available from: https://www.nature.com/articles/nrd4437

51. Yin PT, Desmond J, Day J. Sharing Historical Trial Data to Accelerate Clinical Development. Clinical Pharmacology & Therapeutics [Internet]. 2019 Sep 20;106(6):1177–8. Available from: http://dx.doi.org/10.1002/cpt.1608

52. Baumann N. How to use the medical subject headings (MeSH). International Journal of Clinical Practice [Internet]. 2016;70(2):171–4. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/ijcp.12767

53. Ide NC, Loane RF, Demner-Fushman D. Essie: A concept-based search engine for structured biomedical text. Journal of the American Medical Informatics Association : JAMIA [Internet]. 2007;14(3):253–63. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244877/

54. Lencioni R, Llovet J. Modified RECIST (mRECIST) Assessment for Hepatocellular Carcinoma. Seminars in Liver Disease [Internet]. 2010 Feb;30(01):052–60. Available from: http://www.thieme-connect.de/DOI/DOI?10.1055/s-0030-1247132

55. Seymour L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekar S, et al. iRECIST: Guidelines for response criteria for use in trials testing immunotherapeutics. The Lancet Oncology [Internet]. 2017 Mar;18(3):e143–52. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5648544/

56. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ [Internet]. 2021 Mar 29;n71. Available from: http://dx.doi.org/10.1136/bmj.n71

57. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. *PRISMA2020*: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Systematic Reviews [Internet]. 2022 Mar 27;18(2). Available from: http://dx.doi.org/10.1002/cl2.1230

58. Ghobrial FEI, Eldin MS, Razek AAKA, Atwan NI, Shamaa SSA. Computed tomography assessment of hepatic metastases of breast cancer with revised response evaluation criteria in solid tumors (RECIST) criteria (version 1.1): Inter-observer agreement. Polish Journal of Radiology [Internet]. 2017 Oct 20;82:593–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5894063/

59. Zimmermann M, Kuhl CK, Engelke H, Bettermann G, Keil S. CT-based whole-body tumor volumetry versus RECIST 1.1: Feasibility and implications for inter-reader variability. European Journal of Radiology [Internet]. 2021 Feb 1;135:109514. Available from: https://www.sciencedirect.com/science/article/pii/S0720048X2030704X

60. Aghighi M, Boe J, Rosenberg J, Von Eyben R, Gawande RS, Petit P, et al. Three-dimensional radiologic assessment of chemotherapy response in ewing sarcoma can be used to predict clinical outcome. Radiology [Internet]. 2016 Sep;280(3):905–15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5006736/

61. Felsch M, Zaim S, Dicken V, Lehmacher W, Scheuring UJ. Comparison of central and local serial CT assessments of metastatic renal cell carcinoma patients in a clinical phase IIB study. Acta Radiologica [Internet]. 2017 Feb 1;58(2):249–55. Available from: https://doi.org/10.1177/0284185116642634

62. Ghosn M, Derbel H, Kharrat R, Oubaya N, Mulé S, Chalaye J, et al. Prediction of overall survival in patients with hepatocellular carcinoma treated with y-90 radioembolization by imaging response criteria. Diagnostic and Interventional Imaging [Internet]. 2021 Jan 1;102(1):35–44. Available from: https://www.sciencedirect.com/science/article/pii/S2211568420302217

63. Karmakar A, Kumtakar A, Sehgal H, Kumar S, Kalyanpur A. Interobserver Variation in Response Evaluation Criteria in Solid Tumors 1.1. Academic Radiology. 2019 Apr;26(4):489–501.

64. Tovoli F, Renzulli M, Negrini G, Brocchi S, Ferrarini A, Andreone A, et al. Inter-operator variability and source of errors in tumour response assessment for hepatocellular carcinoma treated with sorafenib. European Radiology [Internet]. 2018 Sep 1;28(9):3611–20. Available from: https://doi.org/10.1007/s00330-018-5393-3

65. ClinicalTrials.gov API | ClinicalTrials.gov v2.0.3 [Internet]. 2025. Available from: https://clinicaltrials.gov/data-api/api

66. Ide NC, Loane RF, Demner-Fushman D. Essie: A concept-based search engine for structured biomedical text. Journal of the American Medical Informatics Association : JAMIA [Internet]. 2007;14(3):253–63. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244877/

67. Gill D. Re-inventing clinical trials through TransCelerate. Nature Reviews Drug Discovery [Internet]. 2014 Nov;13(11):787–8. Available from: https://www.nature.com/articles/nrd4437

68. Anttila JV, Shubin M, Cairns J, Borse F, Guo Q, Mononen T, et al. Contrasting the impact of cytotoxic and cytostatic drug therapies on tumour progression. PLoS Computational Biology [Internet]. 2019 Nov 18;15(11):e1007493. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6886869/

69. Bailly C, Thuru X, Quesnel B. Combined cytotoxic chemotherapy and immunotherapy of cancer: Modern times. NAR Cancer [Internet]. 2020 Feb 17;2(1):zcaa002. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8209987/

70. SDTM [Internet]. 2024. Available from: https://www.cdisc.org/standards/foundational/sdtm

71. Tumor/Lesion Identification & Results | CDISC [Internet]. Available from: https://www.cdisc.org/kb/ecrf/tumorlesion-identification-results

72. Oncology Disease Response (RS) Supplements | CDISC [Internet]. Available from: https://www.cdisc.org/kb/articles/cdisc-published/oncology-disease-response-rs-supplements

73. Wolff F, Gering V. The oncology specific domains TU, TS and RS: What to know as a statistical analyst. 2018;

74. Cochrane handbook for systematic reviews of interventions version 6.4 (updated august 2023). Cochrane, 2023 [Internet]. 2023. Available from: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-13#section-13-3-5-6

75. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement [Internet]. 1960 Apr;20(1):37–46. Available from: http://dx.doi.org/10.1177/001316446002000104

76. McHugh ML. Interrater reliability: The kappa statistic. Biochemia Medica [Internet]. 2012 Oct 15;22(3):276–82. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

77. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics [Internet]. 1977;33(1):159–74. Available from: https://www.jstor.org/stable/2529310

78. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76(5):378–82.

79. Gwet KL. Large-sample variance of fleiss generalized kappa. Educational and Psychological Measurement [Internet]. 2021 Aug;81(4):781–90. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8243202/

80. Irr: Various coefficients of interrater reliability and agreement [Internet]. The R Foundation; 2005. Available from: http://dx.doi.org/10.32614/CRAN.package.irr

81. FENG C, WANG H, LU N, CHEN T, HE H, LU Y, et al. Log-transformation and its implications for data analysis. Shanghai Archives of Psychiatry [Internet]. 2014 Apr;26(2):105–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/

82. Carpentier M, Combescure C, Merlini L, Perneger TV. Kappa statistic to measure agreement beyond chance in free-response assessments. BMC Medical Research Methodology [Internet]. 2017 Apr 19;17(1):62. Available from: https://doi.org/10.1186/s12874-017-0340-6

83. Oehlert GW. A note on the delta method. The American Statistician [Internet]. 1992 Feb 1;46(1):27–9. Available from: https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475842

84. Muche R. Applied survival analysis: Regression modeling of time to event data.: DW hosmer, jr., s lemeshow. New york: John wiley, 1999, pp.386, US$89.95. ISBN: 0-471-15410-5. International Journal of Epidemiology [Internet]. 2001 Apr 1;30(2):408–9. Available from: https://doi.org/10.1093/ije/30.2.408

85. Viechtbauer W. Metafor: Meta-analysis package for r [Internet]. The R Foundation; 2009. Available from: http://dx.doi.org/10.32614/CRAN.package.metafor

86. Cochrane handbook for systematic reviews of interventions version 6.4 (updated august 2023). Cochrane, 2023 [Internet]. 2023. Available from: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-13#section-13-3-5-6

87. Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. Journal of Clinical Epidemiology [Internet]. 2001 Oct 1;54(10):1046–55. Available from: https://www.jclinepi.com/article/S0895-4356(01)00377-8/fulltext

88. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ [Internet]. 1997 Sep 13;315(7109):629–34. Available from: https://www.bmj.com/content/315/7109/629

89. Revelle W. Psych: Procedures for psychological, psychometric, and personality research [Internet]. The R Foundation; 2007. Available from: http://dx.doi.org/10.32614/CRAN.package.psych

90. Cochran WG. The comparison of percentages in matched samples. Biometrika [Internet]. 1950;37(3/4):256–66. Available from: https://www.jstor.org/stable/2332378

91. West SL, Gartlehner G, Mansfield AJ, Poole C, Tant E, Lenfestey N, et al. Table 7, Summary of common statistical approaches to test for heterogeneity [Internet]. 2010. Available from: https://www.ncbi.nlm.nih.gov/books/NBK53317/table/ch3.t2/

92. Sundjaja JH, Shrestha R, Krishan K. McNemar And Mann-Whitney U Tests. In Treasure Island (FL): StatPearls Publishing; 2025. Available from: http://www.ncbi.nlm.nih.gov/books/NBK560699/

93. McNEMAR Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947 Jun;12(2):153–7.

94. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society Series B (Methodological) [Internet]. 1972;34(2):187–220. Available from: https://www.jstor.org/stable/2985181

95. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit. British Journal of Cancer [Internet]. 2003 Aug;89(4):605–11. Available from: https://www.nature.com/articles/6601120

96. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: Multivariate data analysis – an introduction to concepts and methods. British Journal of Cancer [Internet]. 2003 Aug 4;89(3):431–6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394368/

97. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer [Internet]. 2003 Jul;89(2):232–8. Available from: https://www.nature.com/articles/6601118

98. Therneau TM. Survival: Survival analysis [Internet]. The R Foundation; 2001. Available from: http://dx.doi.org/10.32614/CRAN.package.survival

99. Bland JM, Altman DG. The logrank test. BMJ : British Medical Journal [Internet]. 2004 May 1;328(7447):1073. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403858/

100. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports. 1966 Mar;50(3):163–70.

101. Lenth RV. Emmeans: Estimated marginal means, aka least-squares means [Internet]. The R Foundation; 2017. Available from: http://dx.doi.org/10.32614/CRAN.package.emmeans

102. Phillips KF. Power of the two one-sided tests procedure in bioequivalence. Journal of Pharmacokinetics and Biopharmaceutics [Internet]. 1990 Apr 1;18(2):137–44. Available from: https://doi.org/10.1007/BF01063556

103. Lakens D. Equivalence tests. Social Psychological and Personality Science [Internet]. 2017 May;8(4):355–62. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5502906/

104. Treadwell JR. Methods Project 1: Existing Guidance for Individual Trials. In Agency for Healthcare Research; Quality (US); 2012. Available from: https://www.ncbi.nlm.nih.gov/books/NBK98984/

105. Lakens D, Caldwell A. TOSTER: Two one-sided tests (TOST) equivalence testing [Internet]. The R Foundation; 2016. Available from: http://dx.doi.org/10.32614/CRAN.package.TOSTER

106. Ogundimu EO. Adequate sample size for developing prediction models is not simply related to events per variable. Journal of Clinical Epidemiology. 2016;

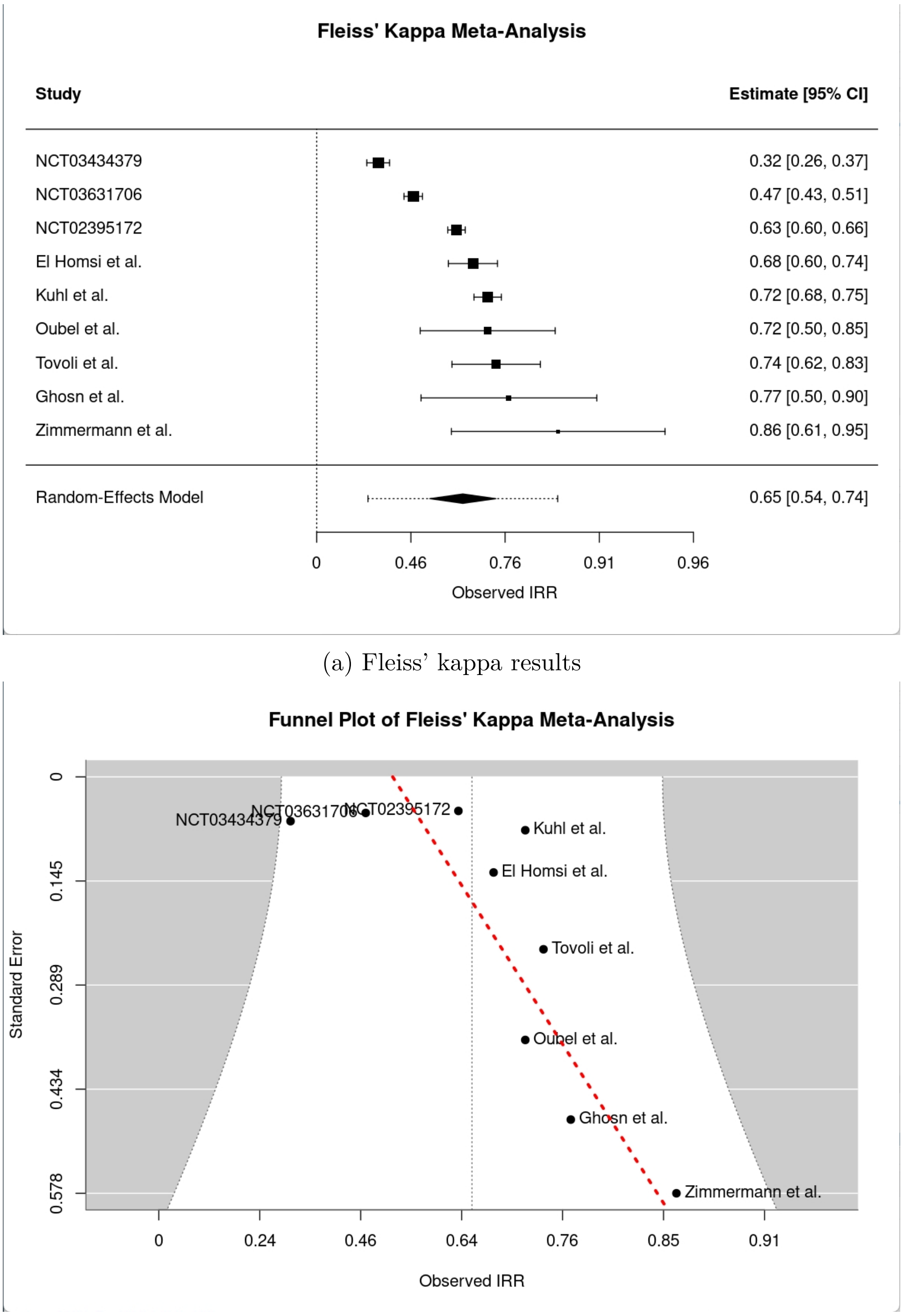
107. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, et al. ggplot2: Create elegant data visualisations using the grammar of graphics [Internet]. The R Foundation; 2007. Available from: http://dx.doi.org/10.32614/CRAN.package.ggplot2

108. Wulff AM, Fabel M, Freitag-Wolf S, Tepper M, Knabe HM, Schäfer JP, et al. Volumetric response classification in metastatic solid tumors on MSCT: Initial results in a whole-body setting. European Journal of Radiology [Internet]. 2013 Oct 1;82(10):e567–73. Available from: https://www.ejradiology.com/article/S0720-048X(13)00284-2/fulltext

109. Mozley PD, Bendtsen C, Zhao B, Schwartz LH, Thorn M, Rong Y, et al. Measurement of tumor volumes improves RECIST-based response assessments in advanced lung cancer. Translational Oncology [Internet]. 2012 Feb 1;5(1):19–25. Available from: https://www.sciencedirect.com/science/article/pii/S1936523312800568

# A. Figures

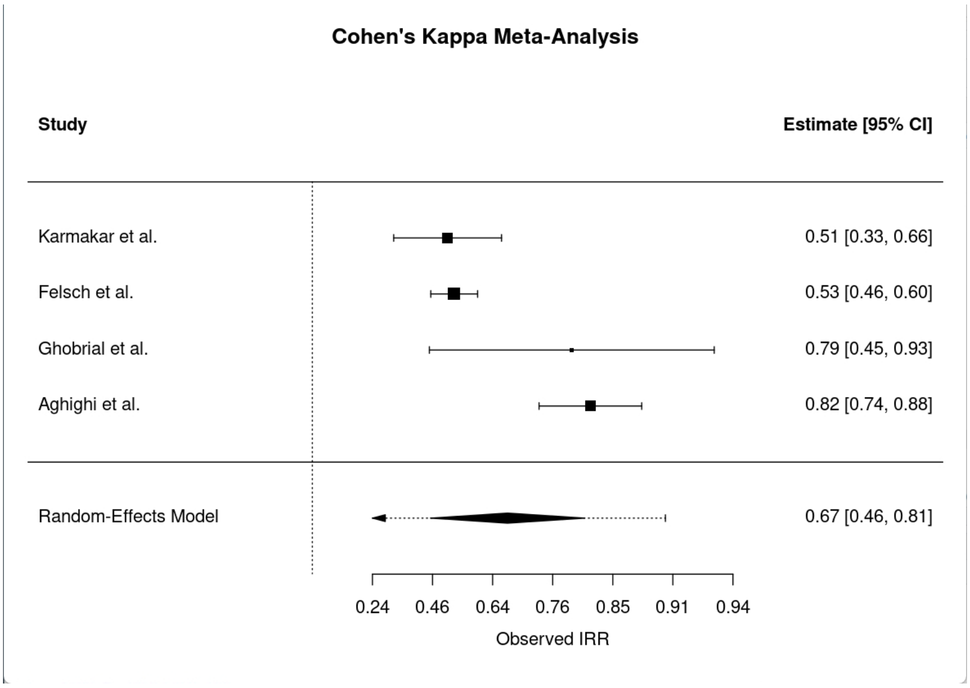
## A.1. IRR Analyses



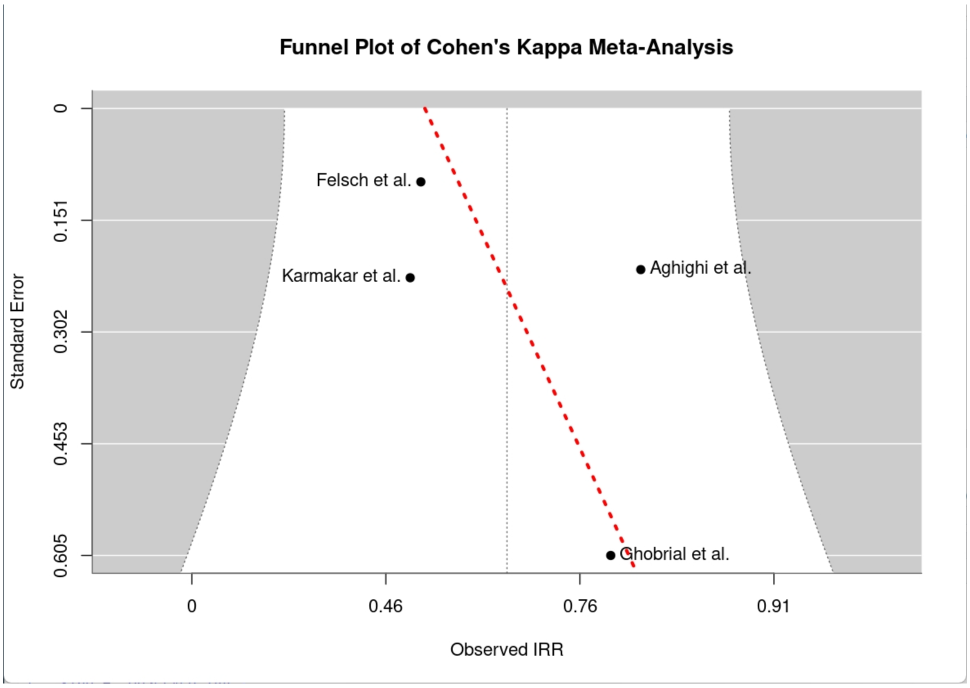
(a) Fleiss' kappa results

(b) Fleiss' kappa funnel plot

Figure A.1.: Fleiss' kappa results and funnel plot for the inter-rater reliability (IRR) meta-analysis.



(a) Cohen's kappa results



(b) Cohen's kappa funnel plot

Figure A.2.: Cohen's kappa results and funnel plot for the inter-rater reliability (IRR) meta-analysis.

## A.2. Trial Results

### A.2.1. Overall Agreement

#### A.2.1.1. Estimated Marginal Means (EMMs)

| Study | Contrast | Estimate | SE | df | t.ratio | p.value |
|---|---|---:|---:|---:|---:|---|
| NCT02395172 | Site Inv. - Reader 1 | -0.0180 | 0.0372 | 2613.316 | -0.4836 | 0.8791 |
| NCT02395172 | Site Inv. - Reader 2 | 0.0371 | 0.0372 | 2617.056 | 0.9983 | 0.578 |
| NCT02395172 | Reader 1 - Reader 2 | 0.0551 | 0.0373 | 2626.392 | 1.4773 | 0.302 |
| NCT03434379 | Site Inv. - Reader 1 | 0.0519 | 0.0621 | 1262.717 | 0.8352 | 0.6812 |
| NCT03434379 | Site Inv. - Reader 2 | 0.0579 | 0.0622 | 1262.846 | 0.9302 | 0.6213 |
| NCT03434379 | Reader 1 - Reader 2 | 0.0060 | 0.0614 | 1259.941 | 0.0972 | 0.9948 |
| NCT03631706 | Site Inv. - Reader 1 | 0.2284 | 0.0406 | 2507.766 | 5.6305 | 0*** |
| NCT03631706 | Site Inv. - Reader 2 | 0.0644 | 0.0406 | 2507.789 | 1.5859 | 0.2519 |
| NCT03631706 | Reader 1 - Reader 2 | -0.1640 | 0.0424 | 2501.910 | -3.8645 | 3e-04*** |

Table A.1.: Estimated marginal means (EMMs) between all raters for the objective response rate (ORR) analyses

#### A.2.1.2. LME Equations

$$\begin{aligned}\text{RSST}\hat{\text{R}}\text{ESC}_i &\sim N\left(2.51_{\alpha_{j[i]}} + 0.02_{\beta_1}(\text{RSEVALID}_{\text{READER 1}}) - 0.04_{\beta_2}(\text{RSEVALID}_{\text{READER 2}}), \sigma^2\right)\\ \alpha_j &\sim N\left(0, 0.73\right), \text{ for USUBJID j = 1,...,J}\end{aligned} \tag{A.1}$$

Equation: Estimated marginal means for the objective response rate (ORR) for study NCT02395172.

$$\begin{aligned}\text{RSST}\hat{\text{R}}\text{ESC}_i &\sim N\left(2.34_{\alpha_{j[i]}} - 0.05_{\beta_1}(\text{RSEVALID}_{\text{READER 1}}) - 0.06_{\beta_2}(\text{RSEVALID}_{\text{READER 2}}), \sigma^2\right)\\ \alpha_j &\sim N\left(0, 0.54\right), \text{ for USUBJID j = 1,...,J}\end{aligned} \tag{A.2}$$

Equation: Estimated marginal means for the objective response rate (ORR) for study NCT03434379.

$$\begin{aligned}\text{RSST}\hat{\text{R}}\text{ESC}_i &\sim N\left(2.98_{\alpha_{j[i]}} - 0.23_{\beta_1}(\text{RSEVALID}_{\text{READER 1}}) - 0.06_{\beta_2}(\text{RSEVALID}_{\text{READER 2}}), \sigma^2\right)\\ \alpha_j &\sim N\left(0, 0.89\right), \text{ for USUBJID j = 1,...,J}\end{aligned} \tag{A.3}$$

Equation: Estimated marginal means for the objective response rate (ORR) for study NCT03631706.

## A.2.2. Objective Response Rate Results

### A.2.2.1. McNemar's Tests

```
$reader1_reader2
        READER 2
READER 1 FALSE TRUE
   FALSE   253   18
   TRUE     19   37

$reader1_site
        SITE INVESTIGATOR
READER 1 FALSE TRUE
   FALSE   259   13
   TRUE     14   42

$reader2_site
        SITE INVESTIGATOR
READER 2 FALSE TRUE
   FALSE   268    5
   TRUE      5   50
```

Table A.2.: Contingency tables for the McNemar's tests study NCT02395172.

```
$reader1_reader2

        READER 2

READER 1 FALSE TRUE

   FALSE   100   11

   TRUE      7   10


$reader1_site

        SITE INVESTIGATOR

READER 1 FALSE TRUE

   FALSE   106    5

   TRUE     11    6


$reader2_site

        SITE INVESTIGATOR

READER 2 FALSE TRUE

   FALSE   105    2

   TRUE     12    9
```

Table A.3.: Contingency tables for the McNemar's tests study NCT03434379.

```
$reader1_reader2
        READER 2
READER 1 FALSE TRUE
   FALSE    62    9
   TRUE      9   66

$reader1_site
        SITE INVESTIGATOR
READER 1 FALSE TRUE
   FALSE    55   16
   TRUE      3   72

$reader2_site
        SITE INVESTIGATOR
READER 2 FALSE TRUE
   FALSE    55   16
   TRUE      3   72
```

Table A.4.: Contingency tables for the McNemar's tests study NCT03631706.

### A.2.3. Survival Analyses

### A.2.3.1. Time to Progression (TTP) Analyses

#### A.2.3.1.1. NCT02395172

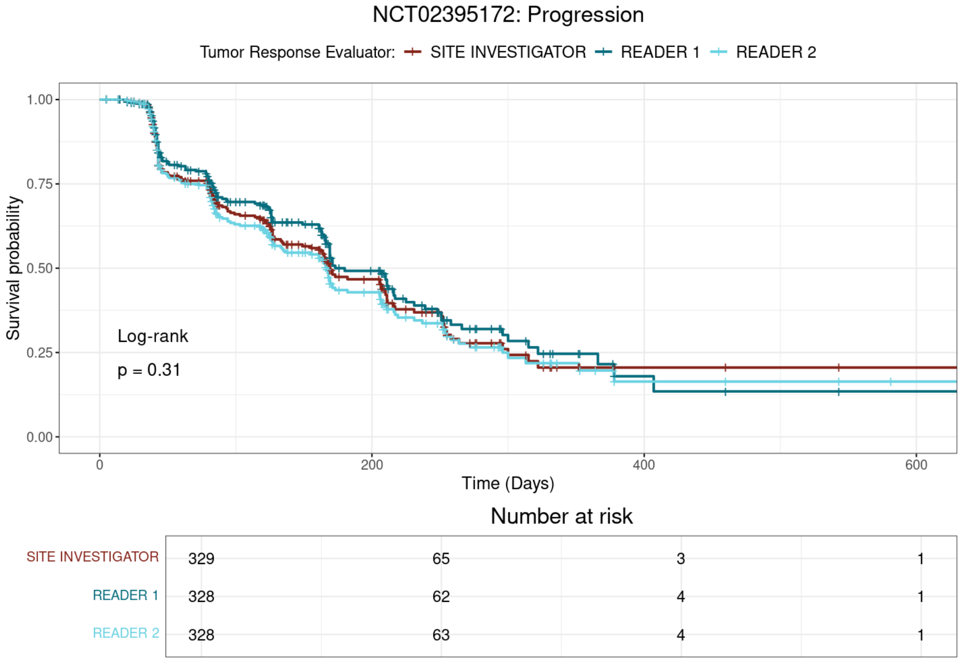


Figure A.3.: Kaplan-Meier survival plot for time to progression (TTP) for study NCT02395172.

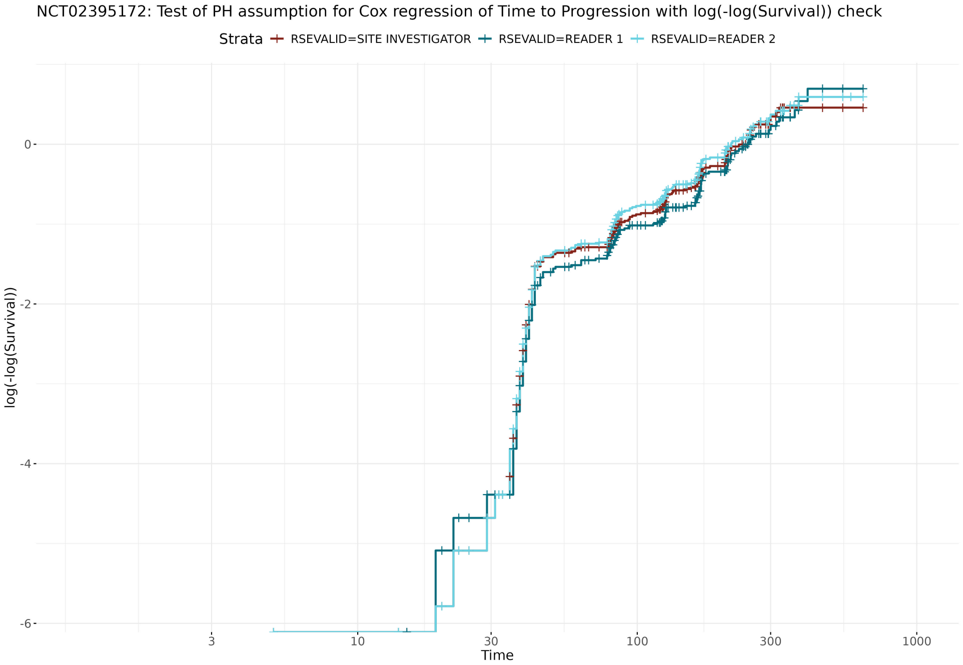


Figure A.4.: Log-minus-log survival plot for time to progression (TTP) for study NCT02395172.

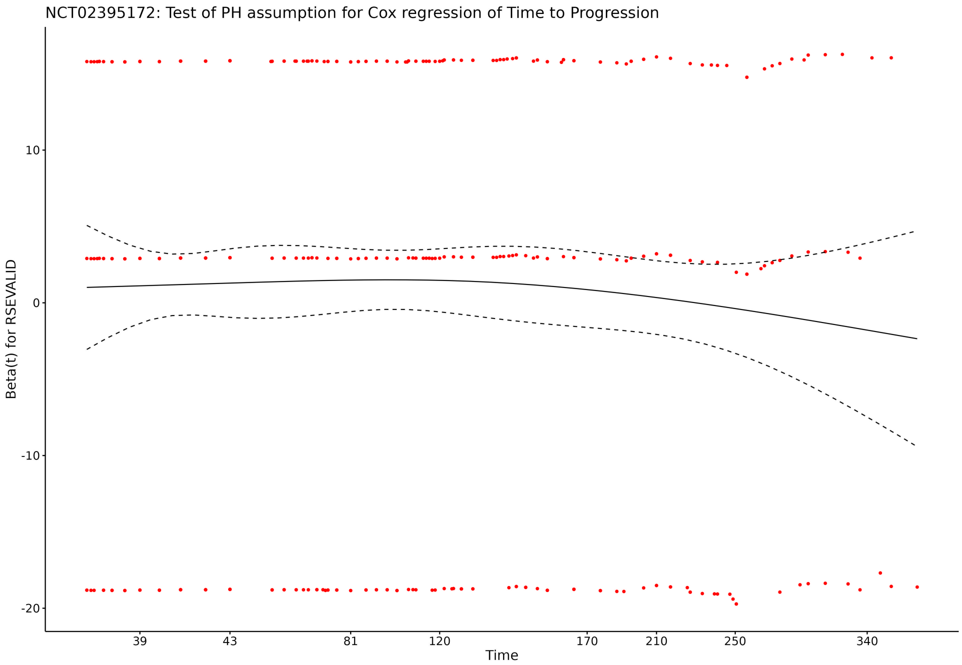
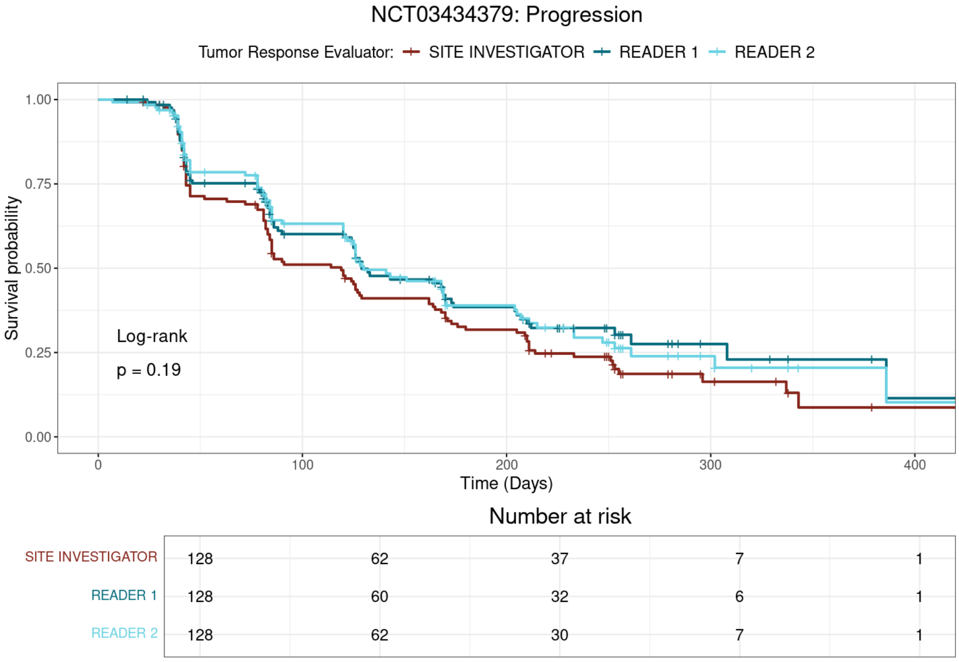


Figure A.5.: Schoenfeld residuals for the Cox proportional hazards model for time to progression (TTP) for study NCT02395172.

### A.2.3.1.2. NCT03434379



| | | | | | |
|---|---|---|---|---|---|
| SITE INVESTIGATOR | 128 | 62 | 37 | 7 | 1 |
| READER 1 | 128 | 60 | 32 | 6 | 1 |
| READER 2 | 128 | 62 | 30 | 7 | 1 |

Figure A.6.: Kaplan-Meier survival plot for time to progression (TTP) for study NCT03434379.

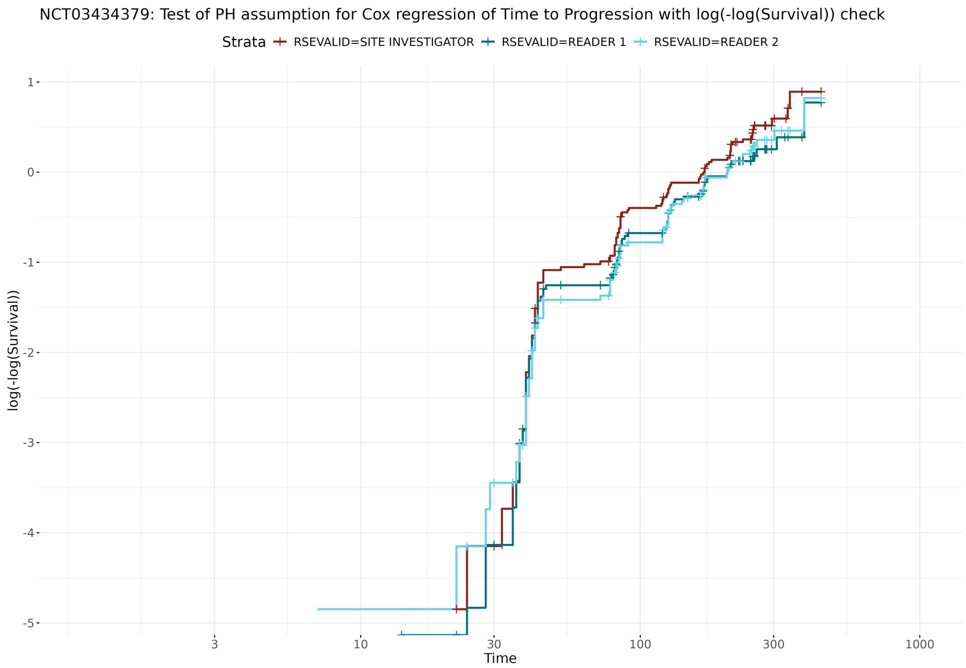


Figure A.7.: Log-minus-log survival plot for time to progression (TTP) for study NCT03434379.

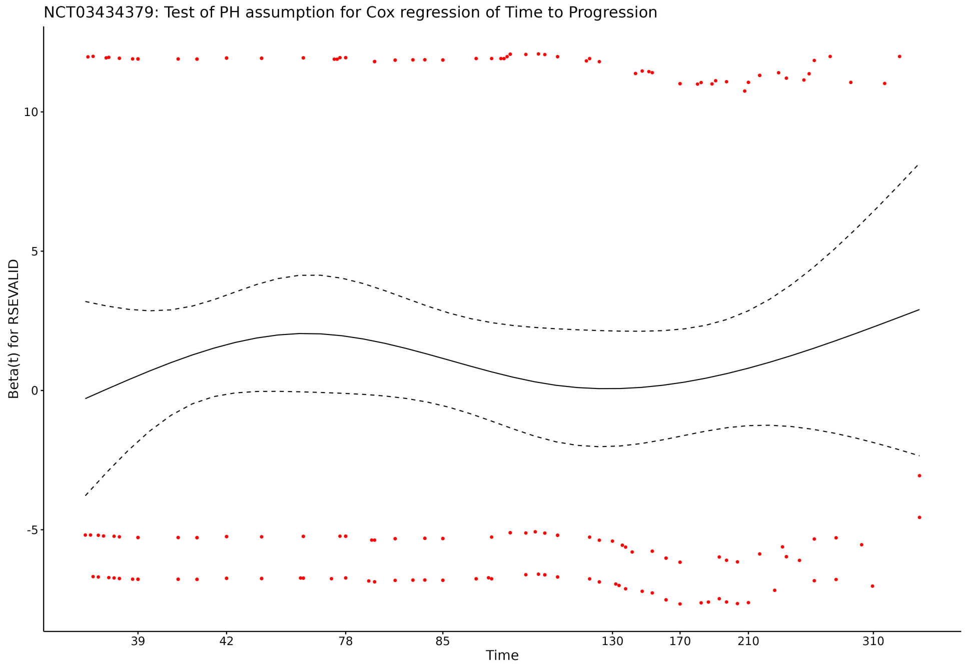


Figure A.8.: Schoenfeld residuals for the Cox proportional hazards model for time to progression (TTP) for study NCT03434379.

### A.2.3.1.3. NCT03631706

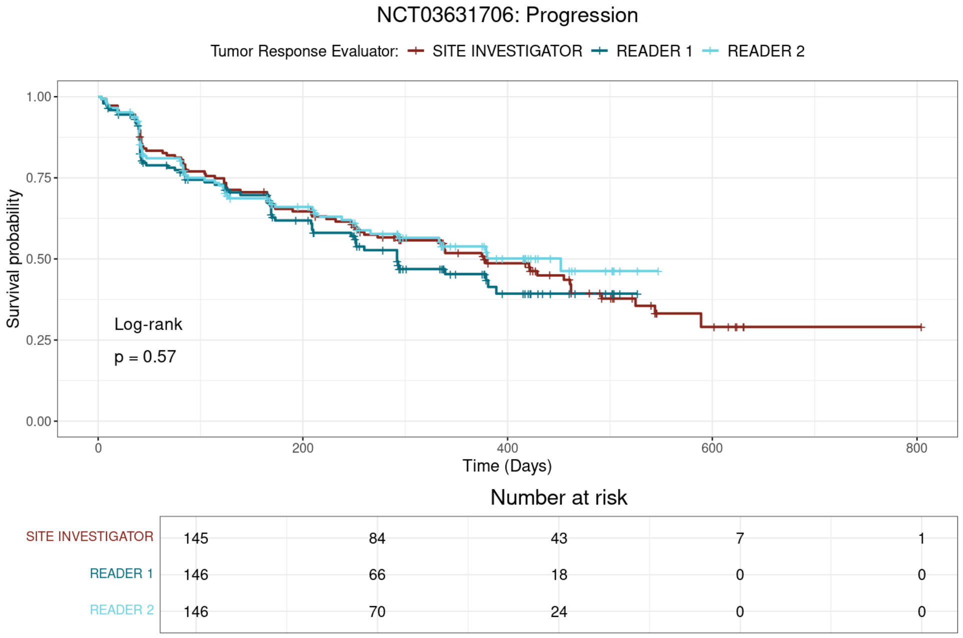


Figure A.9.: Kaplan-Meier survival plot for time to progression (TTP) for study NCT03631706.

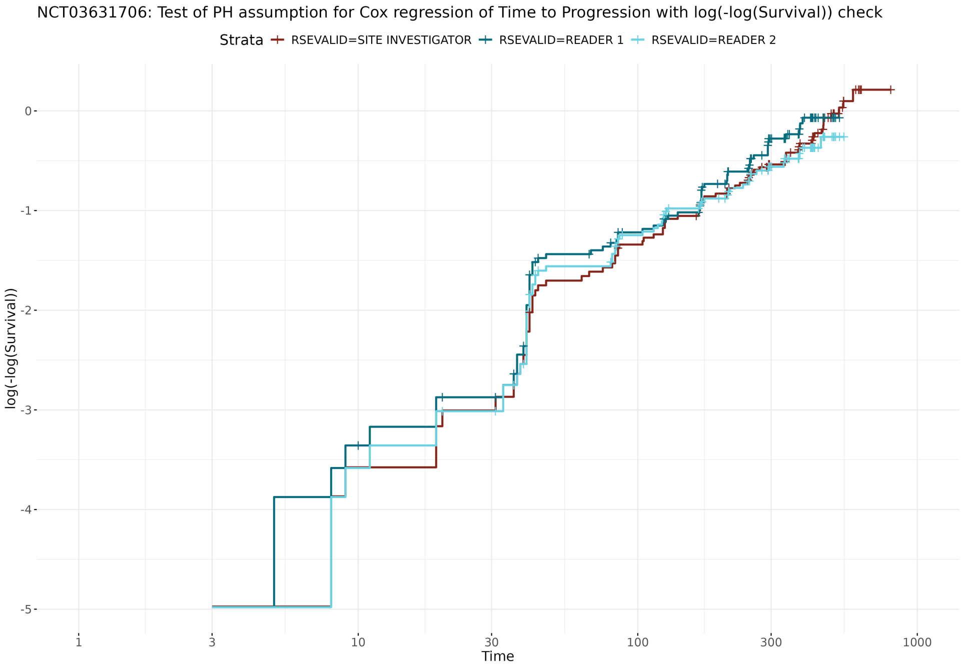


Figure A.10.: Log-minus-log survival plot for time to progression (TTP) for study NCT03631706.

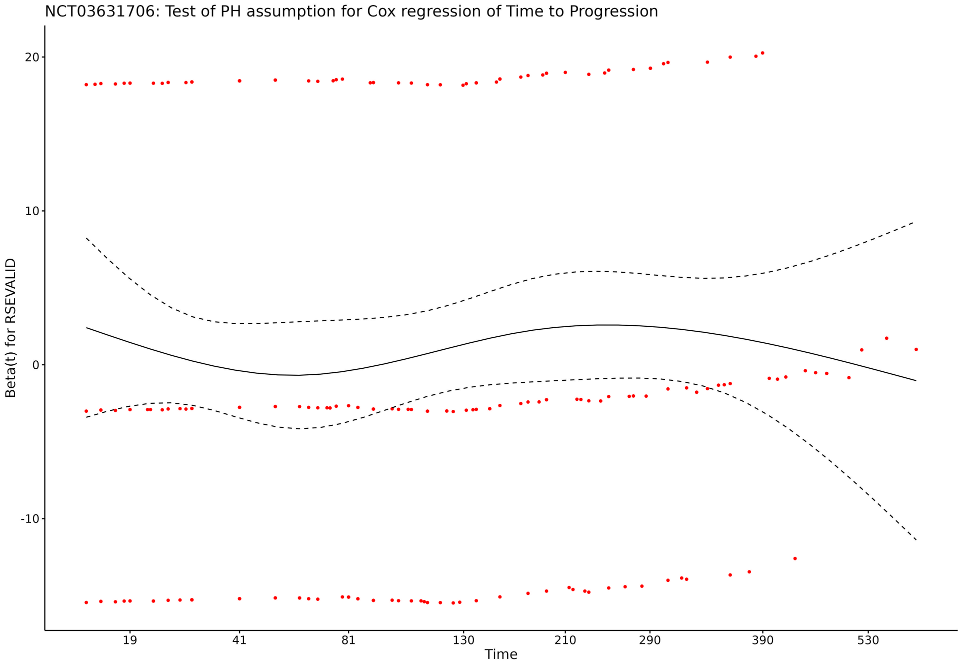
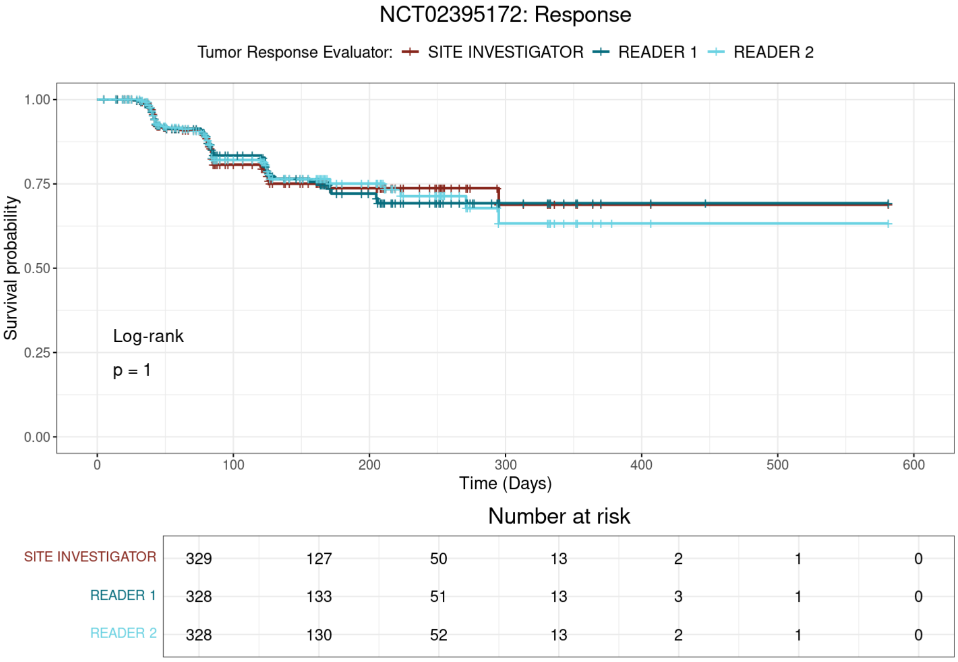


Figure A.11.: Schoenfeld residuals for the Cox proportional hazards model for time to progression (TTP) for study NCT03631706.

### A.2.3.2. Time to Response (TTR) Analyses

#### A.2.3.2.1. NCT02395172



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| SITE INVESTIGATOR | 329 | 127 | 50 | 13 | 2 | 1 | 0 |
| READER 1 | 328 | 133 | 51 | 13 | 3 | 1 | 0 |
| READER 2 | 328 | 130 | 52 | 13 | 2 | 1 | 0 |

Figure A.12.: Kaplan-Meier survival plot for time to response (TTR) for study NCT02395172.

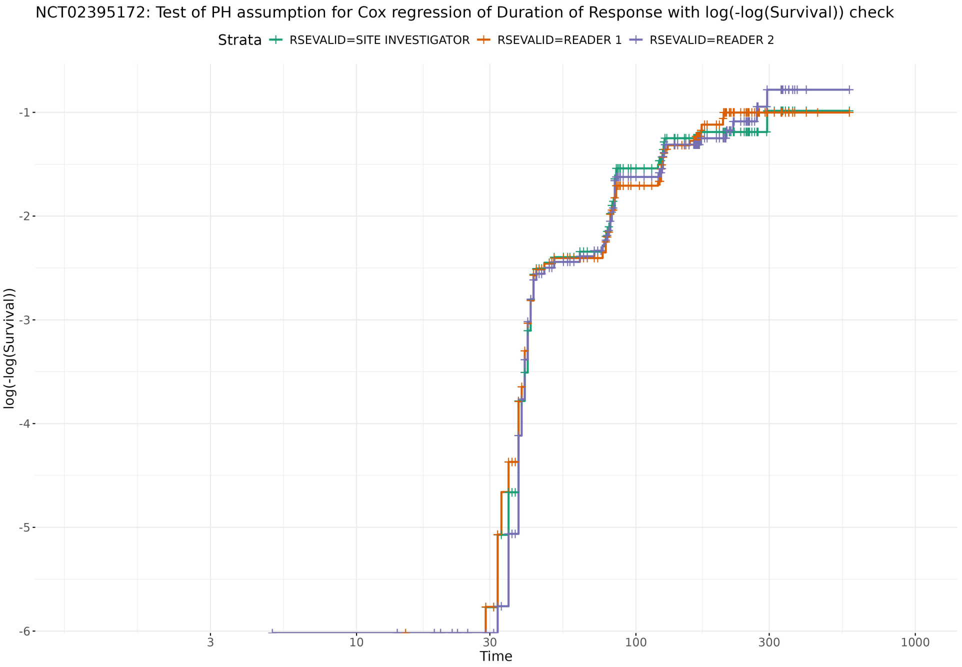


Figure A.13.: Log-minus-log survival plot for time to response (TTR) for study NCT02395172.

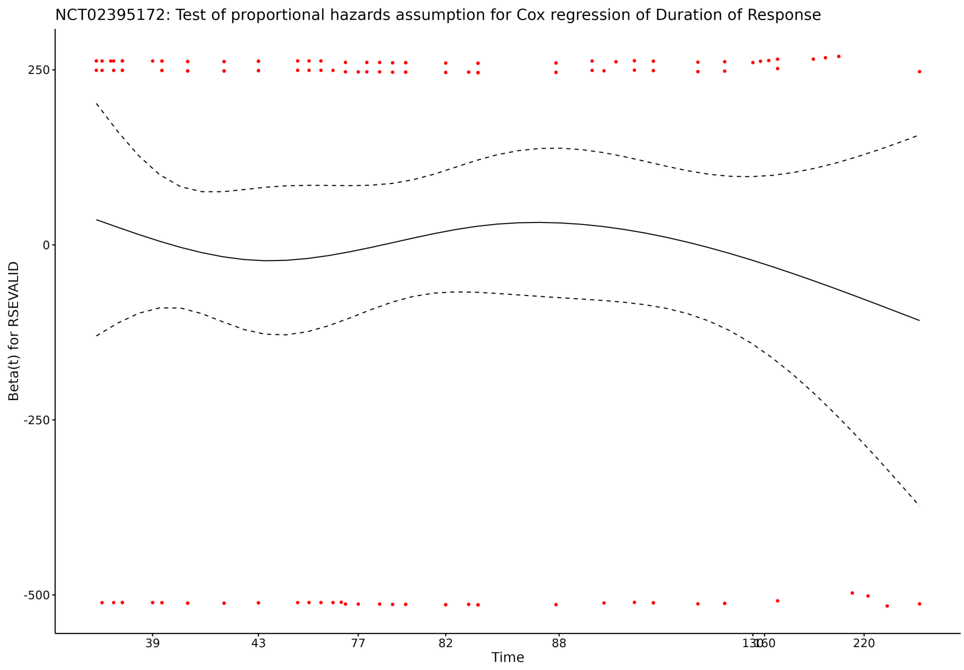


Figure A.14.: Schoenfeld residuals for the Cox proportional hazards model for time to response (TTR) for study NCT02395172.

#### A.2.3.2.2. NCT03434379

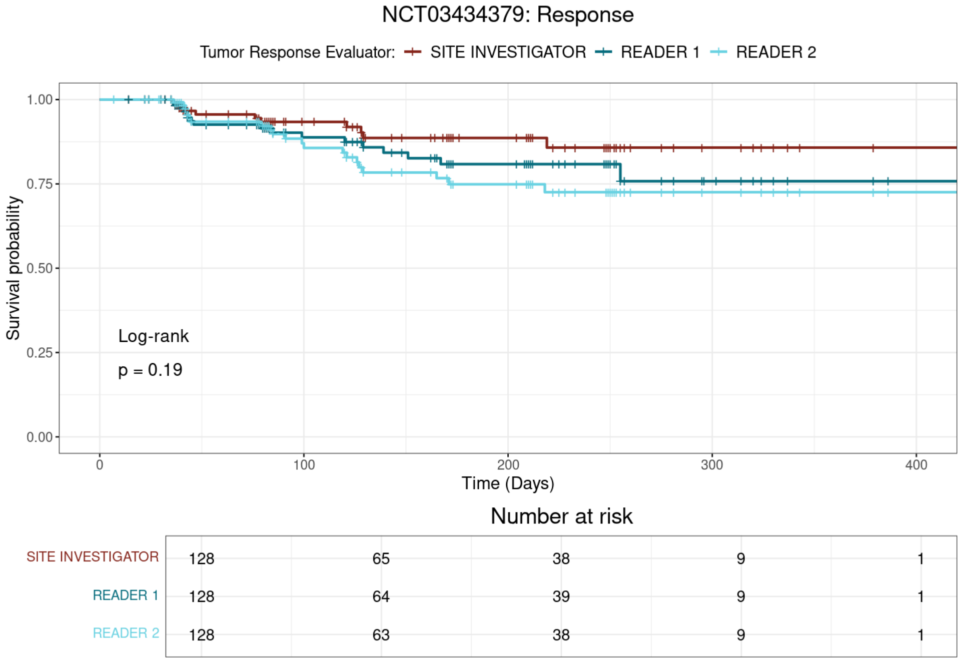


Figure A.15.: Kaplan-Meier survival plot for time to response (TTR) for study NCT03434379.

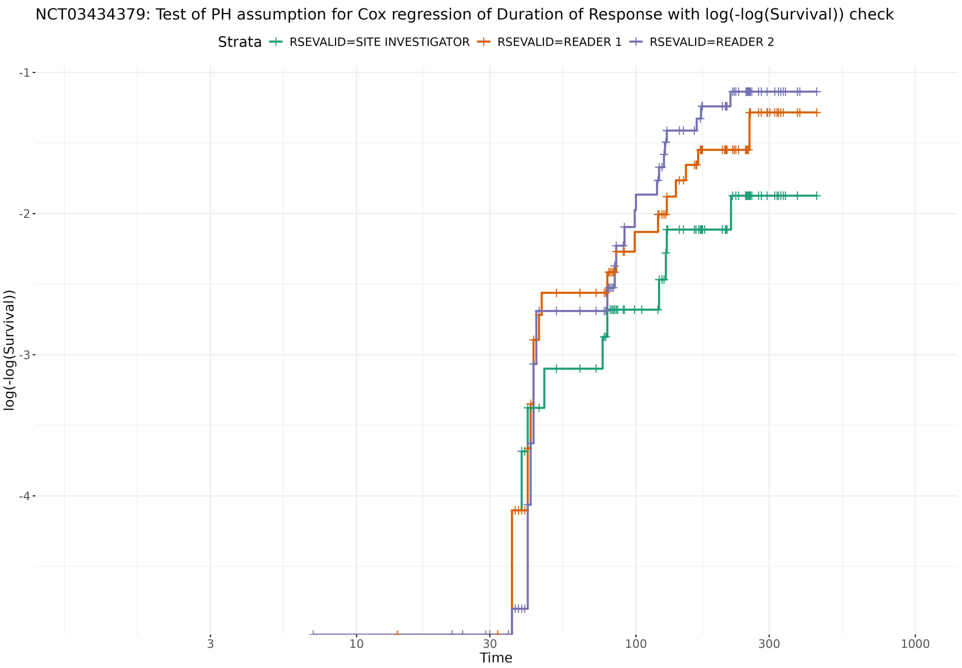


Figure A.16.: Log-minus-log survival plot for time to response (TTR) for study NCT03434379.

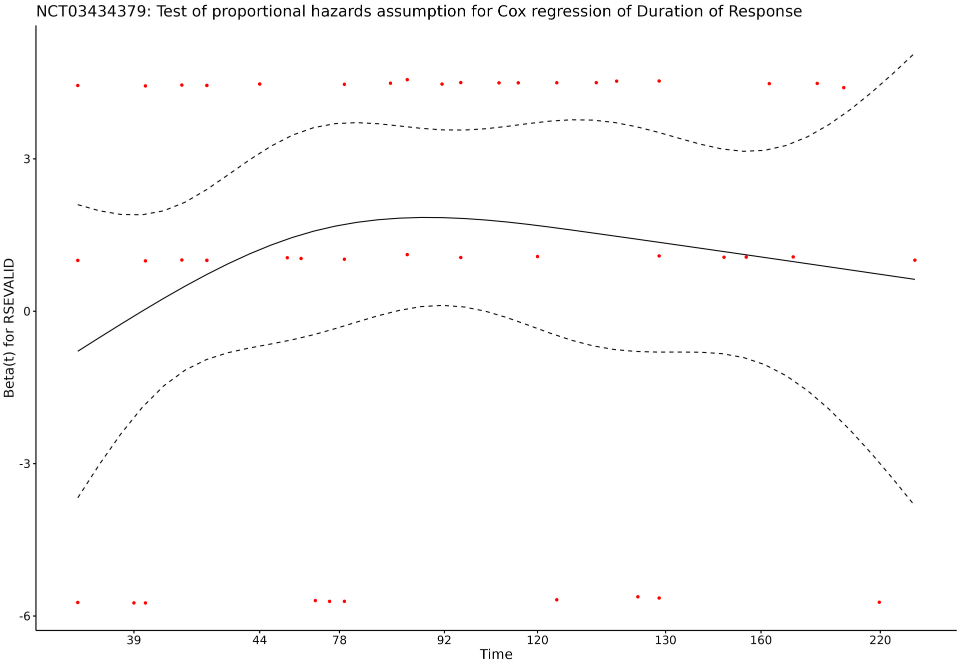


Figure A.17.: Schoenfeld residuals for the Cox proportional hazards model for time to response (TTR) for study NCT03434379.

### A.2.3.2.3. NCT03631706

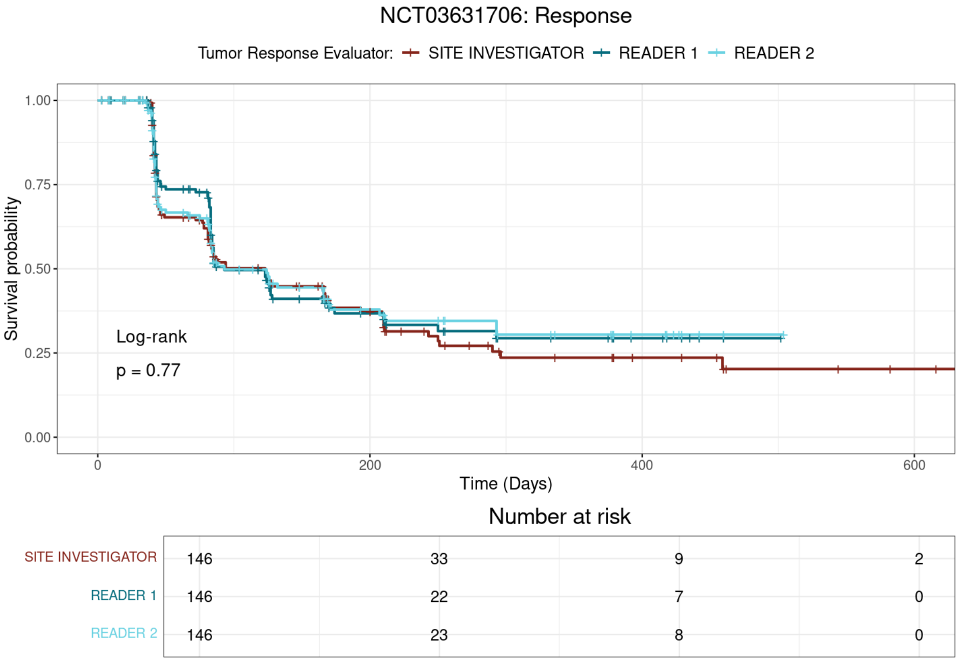


Figure A.18.: Kaplan-Meier survival plot for time to response (TTR) for study NCT03631706.

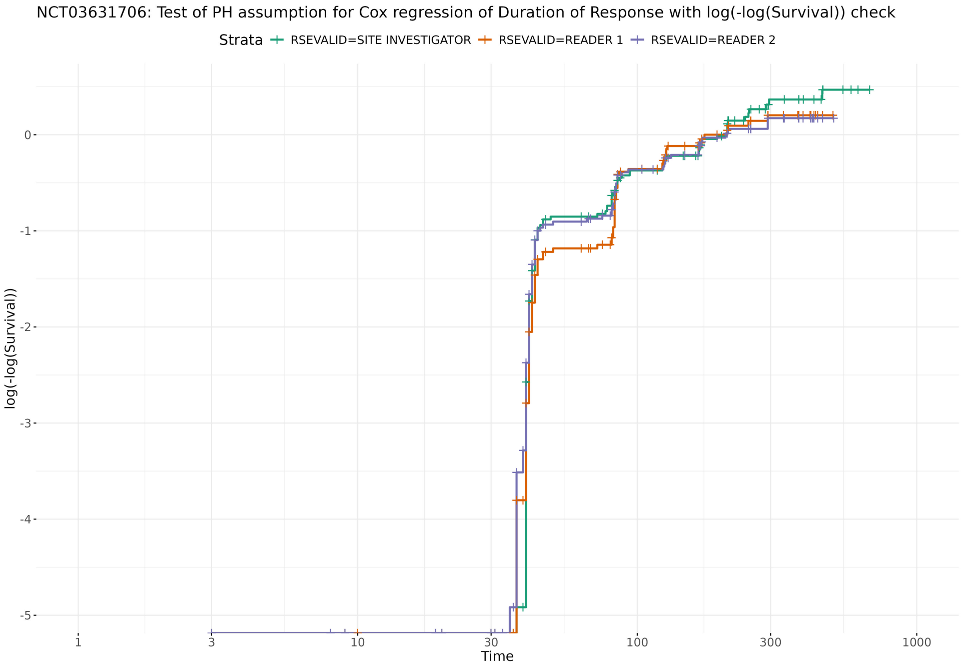


Figure A.19.: Log-minus-log survival plot for time to response (TTR) for study NCT03631706.

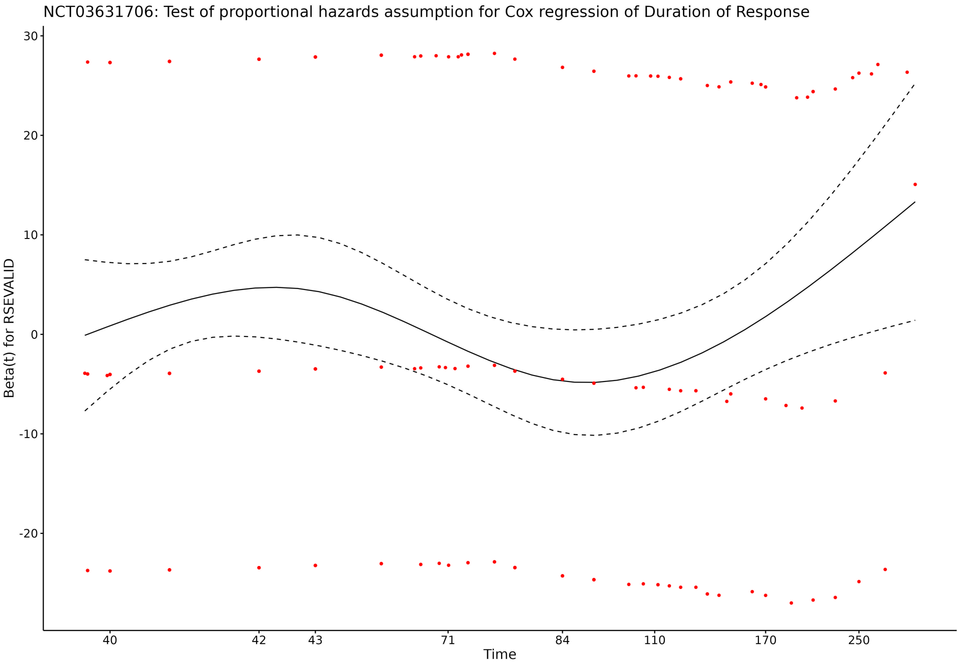


Figure A.20.: Schoenfeld residuals for the Cox proportional hazards model for time to response (TTR) for study NCT03631706.

### A.2.3.3. Duration of Response (DoR) Analyses

#### A.2.3.3.1. NCT02395172

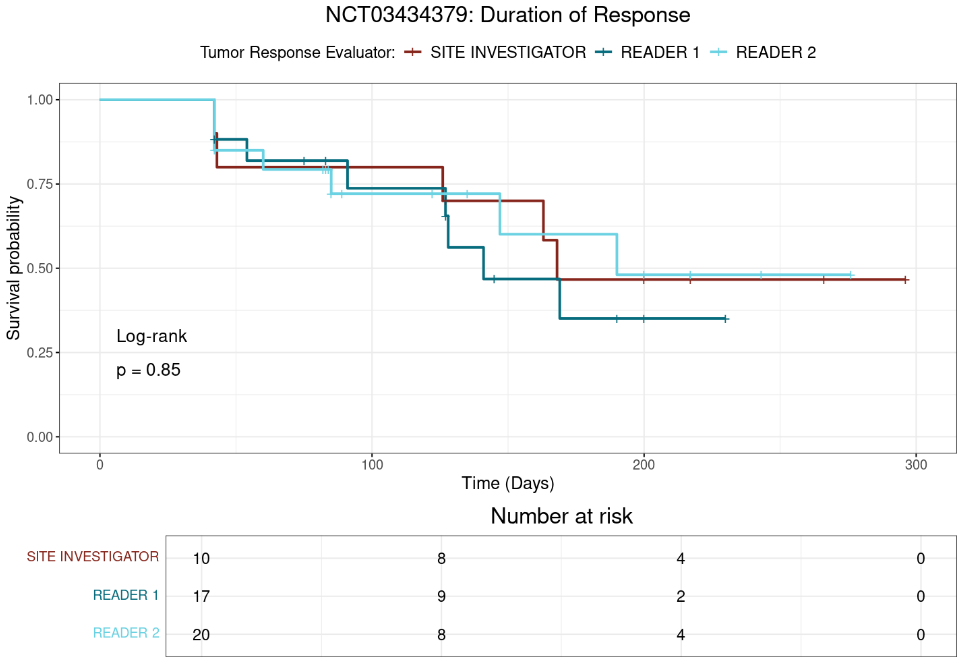


Figure A.21.: Kaplan-Meier survival plot for duration of response (DoR) for study NCT02395172.

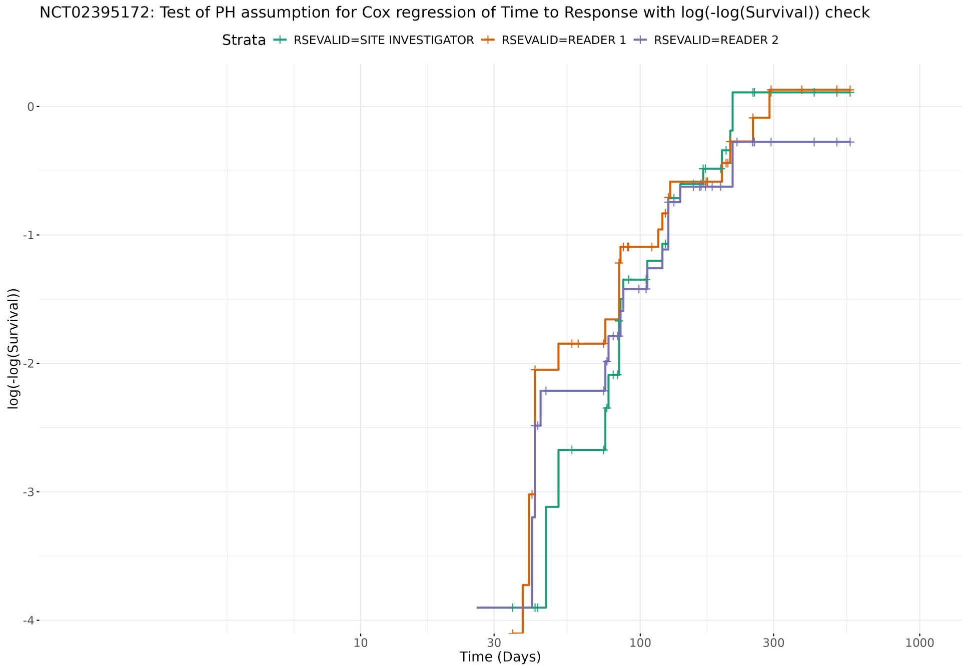


Figure A.22.: Log-minus-log survival plot for duration of response (DoR) for study NCT02395172.

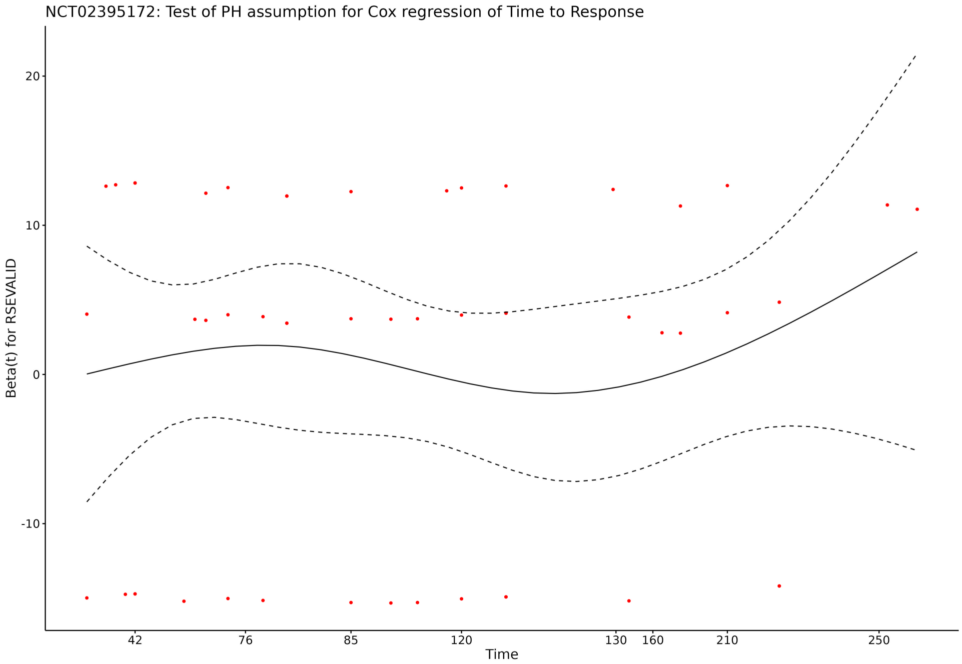
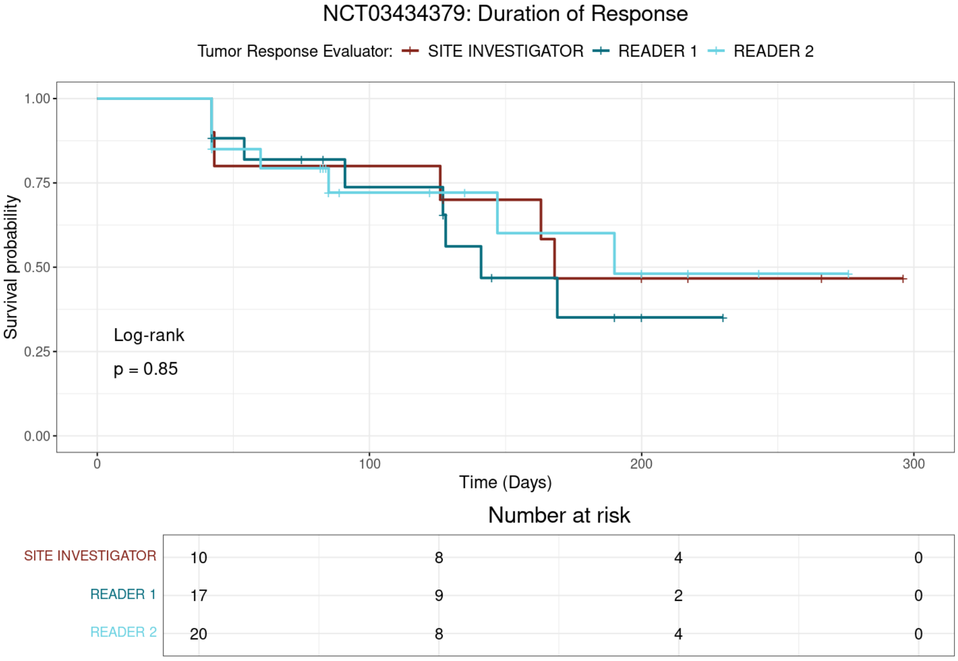


Figure A.23.: Schoenfeld residuals for the Cox proportional hazards model for duration of response (DoR) for study NCT02395172.

### A.2.3.3.2. NCT03434379



| | | | | |
|---|---|---|---|---|
| SITE INVESTIGATOR | 10 | 8 | 4 | 0 |
| READER 1 | 17 | 9 | 2 | 0 |
| READER 2 | 20 | 8 | 4 | 0 |

Figure A.24.: Kaplan-Meier survival plot for duration of response (DoR) for study NCT03434379.

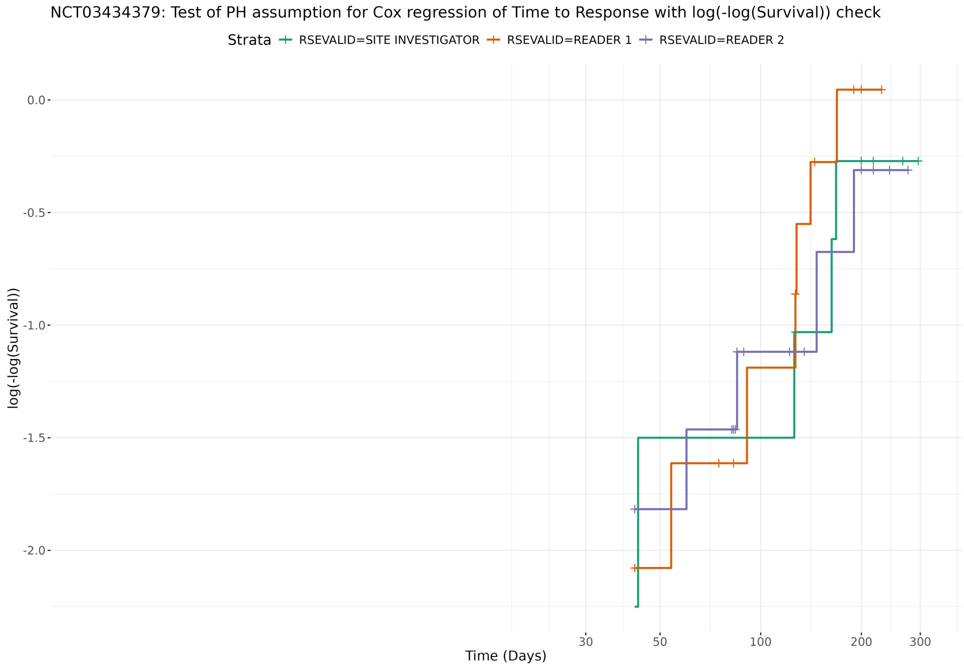


Figure A.25.: Log-minus-log survival plot for duration of response (DoR) for study NCT03434379.

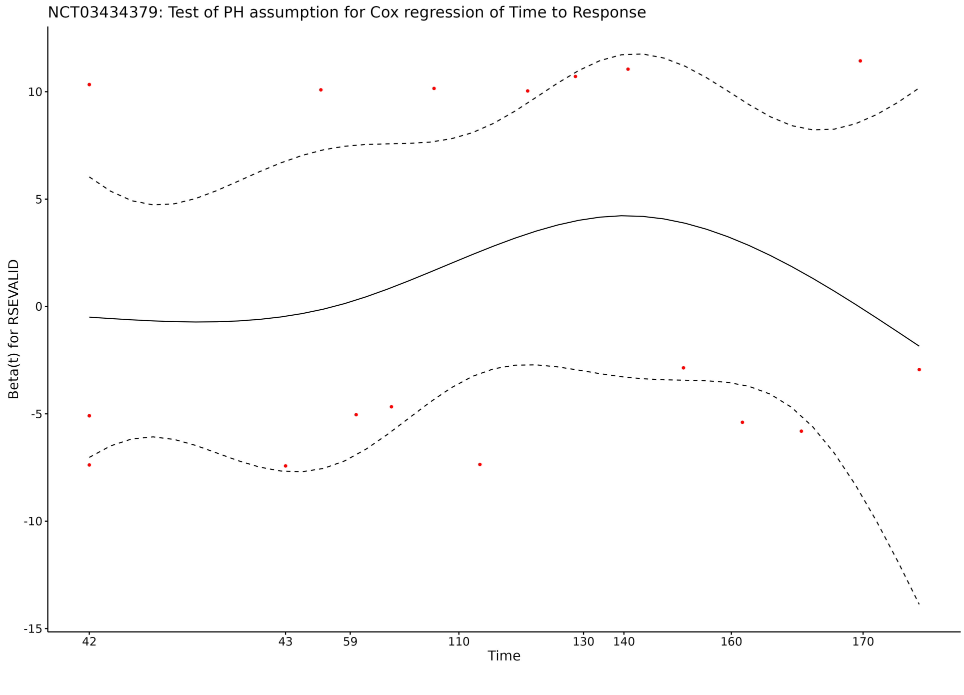
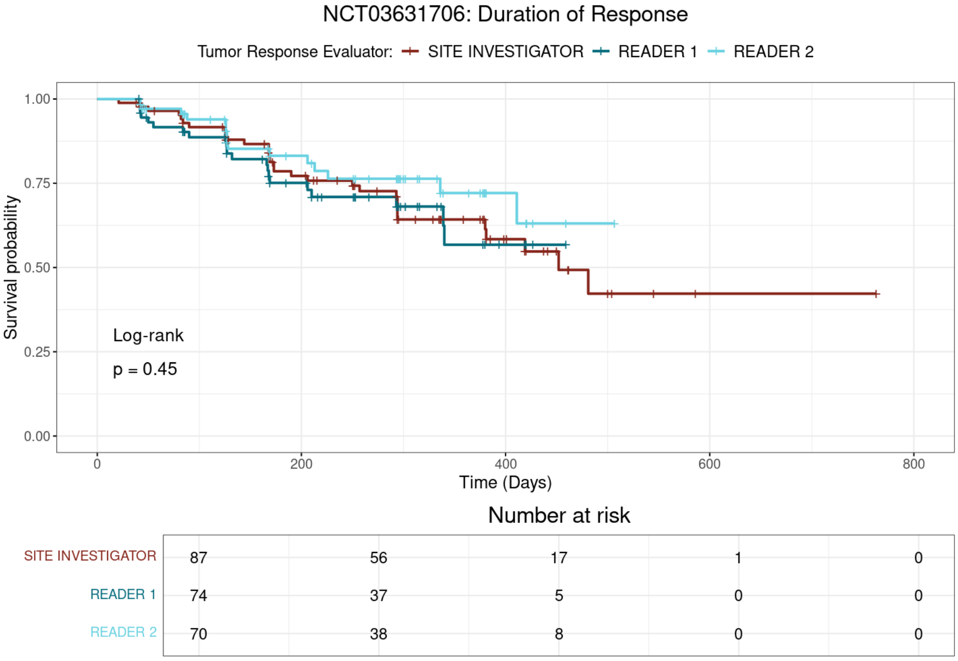


Figure A.26.: Schoenfeld residuals for the Cox proportional hazards model for duration of response (DoR) for study NCT03434379.

### A.2.3.3.3. NCT03631706



| | | | | | |
|---|---|---|---|---|---|
| SITE INVESTIGATOR | 87 | 56 | 17 | 1 | 0 |
| READER 1 | 74 | 37 | 5 | 0 | 0 |
| READER 2 | 70 | 38 | 8 | 0 | 0 |

Figure A.27.: Kaplan-Meier survival plot for duration of response (DoR) for study NCT03631706.

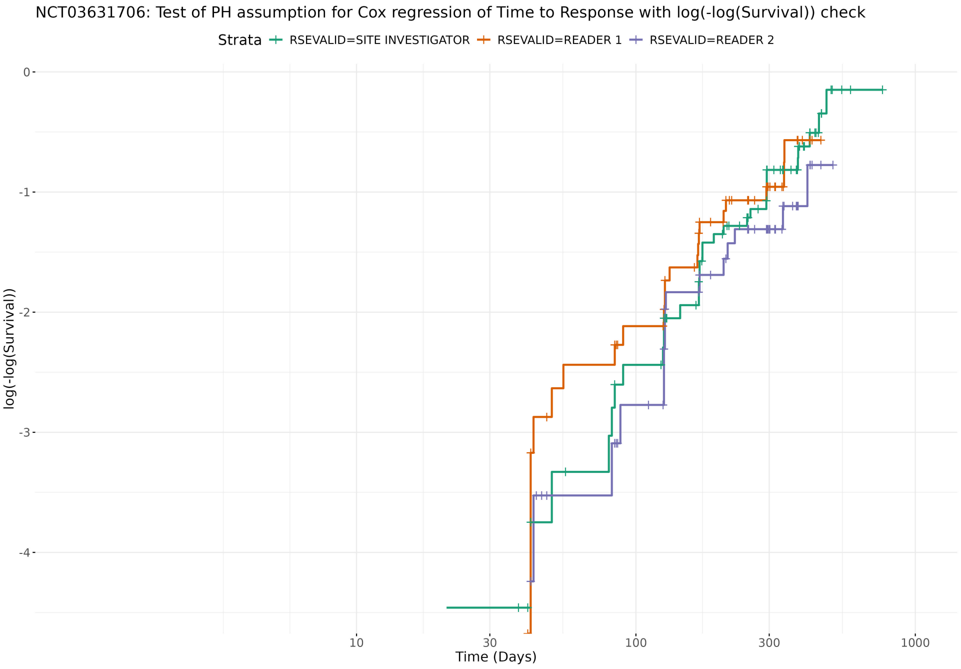


Figure A.28.: Log-minus-log survival plot for duration of response (DoR) for study NCT03631706.

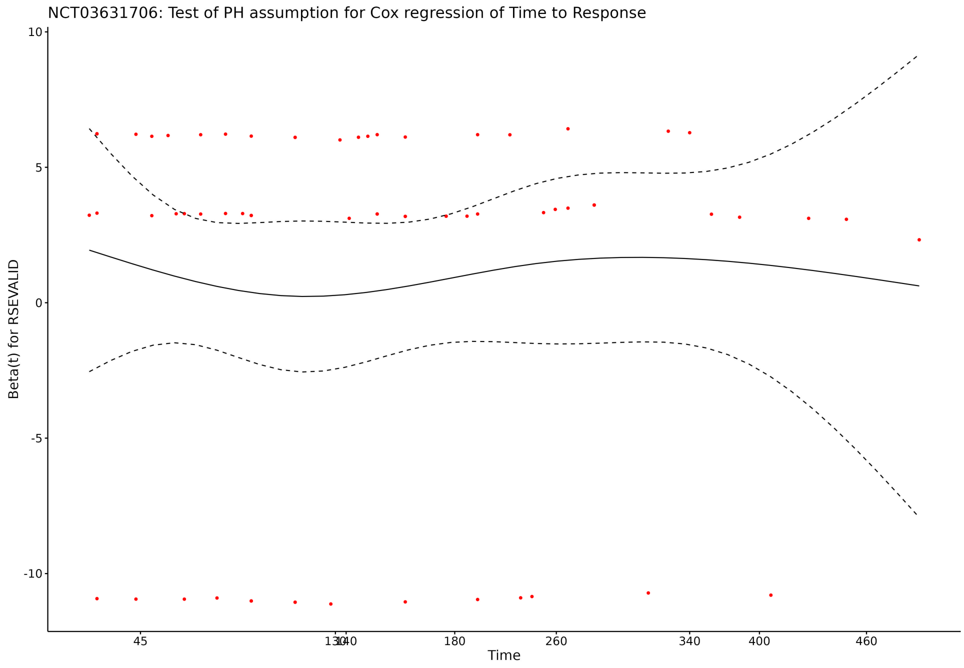


Figure A.29.: Schoenfeld residuals for the Cox proportional hazards model for duration of response (DoR) for study NCT03631706.

## A.2.4. Time-to-Event Meta-Analyses

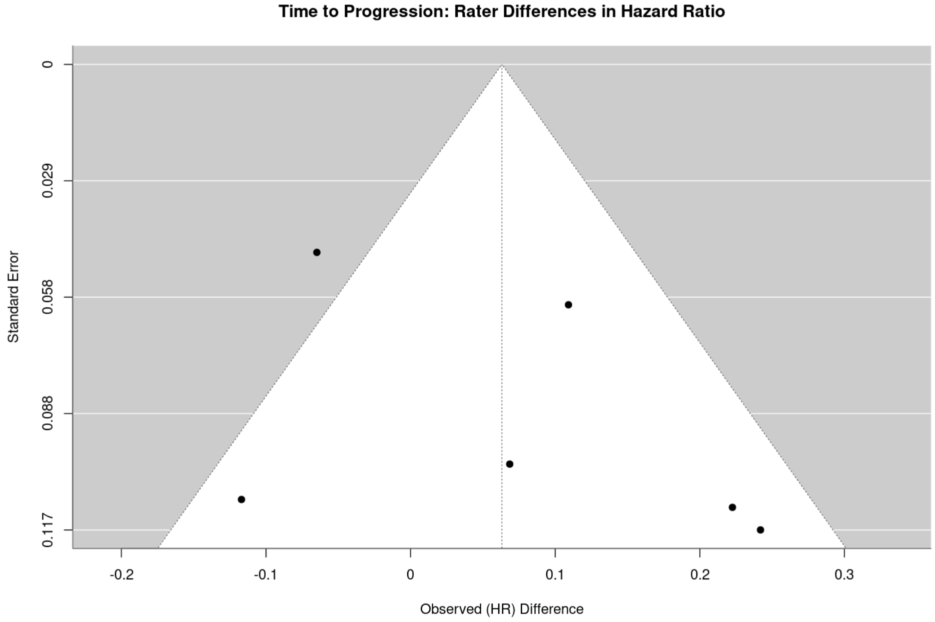


Figure A.30.: Forest plot of hazard ratios for TTP from the meta-analysis of time to event outcomes

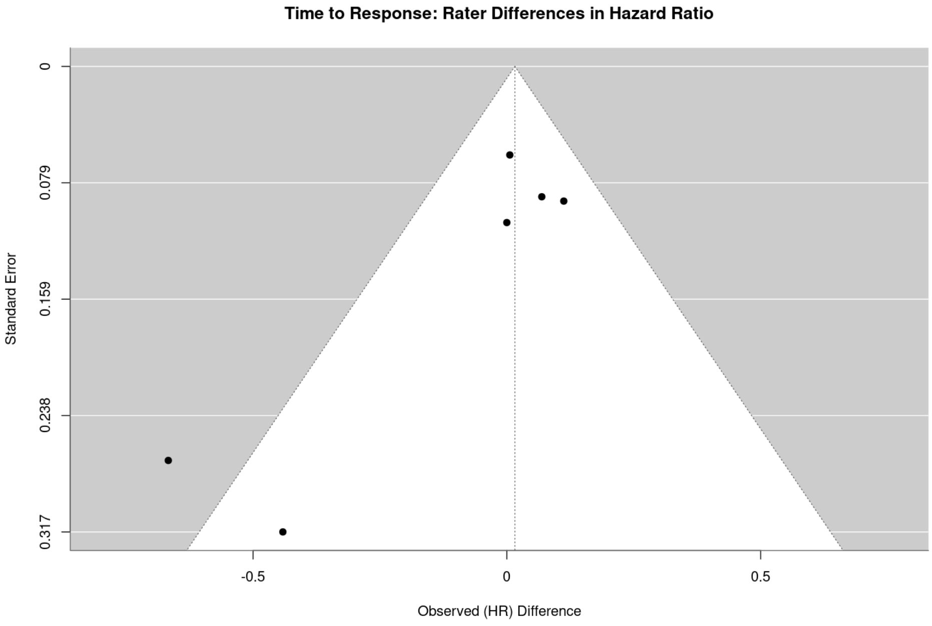


Figure A.31.: Funnel plot of hazard ratios for TTR from the meta-analysis of time to event outcomes

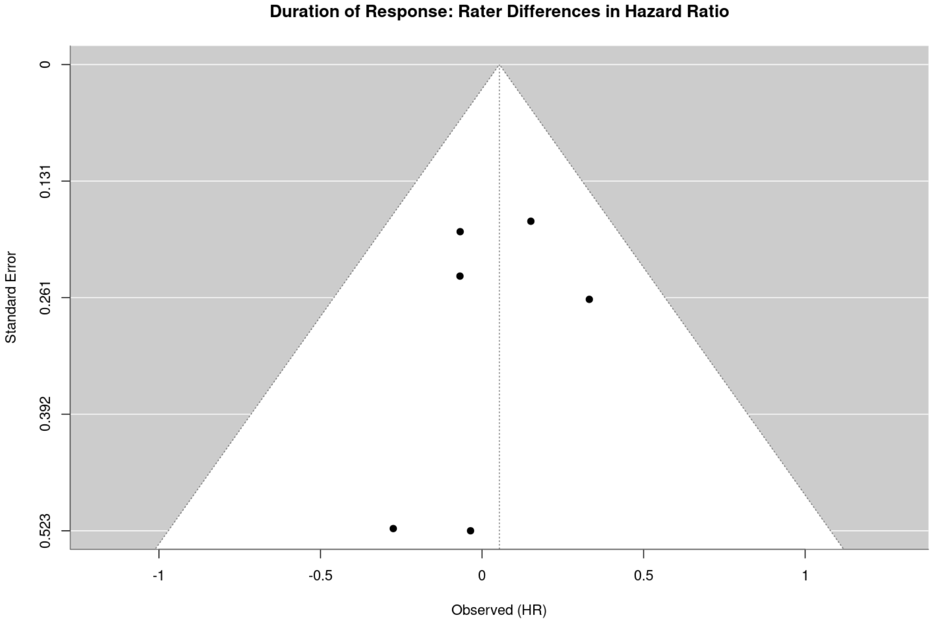


Figure A.32.: Funnel plot of hazard ratios for DOR from the meta-analysis of time to event outcomes

## A.3. Sensitivity Analyses

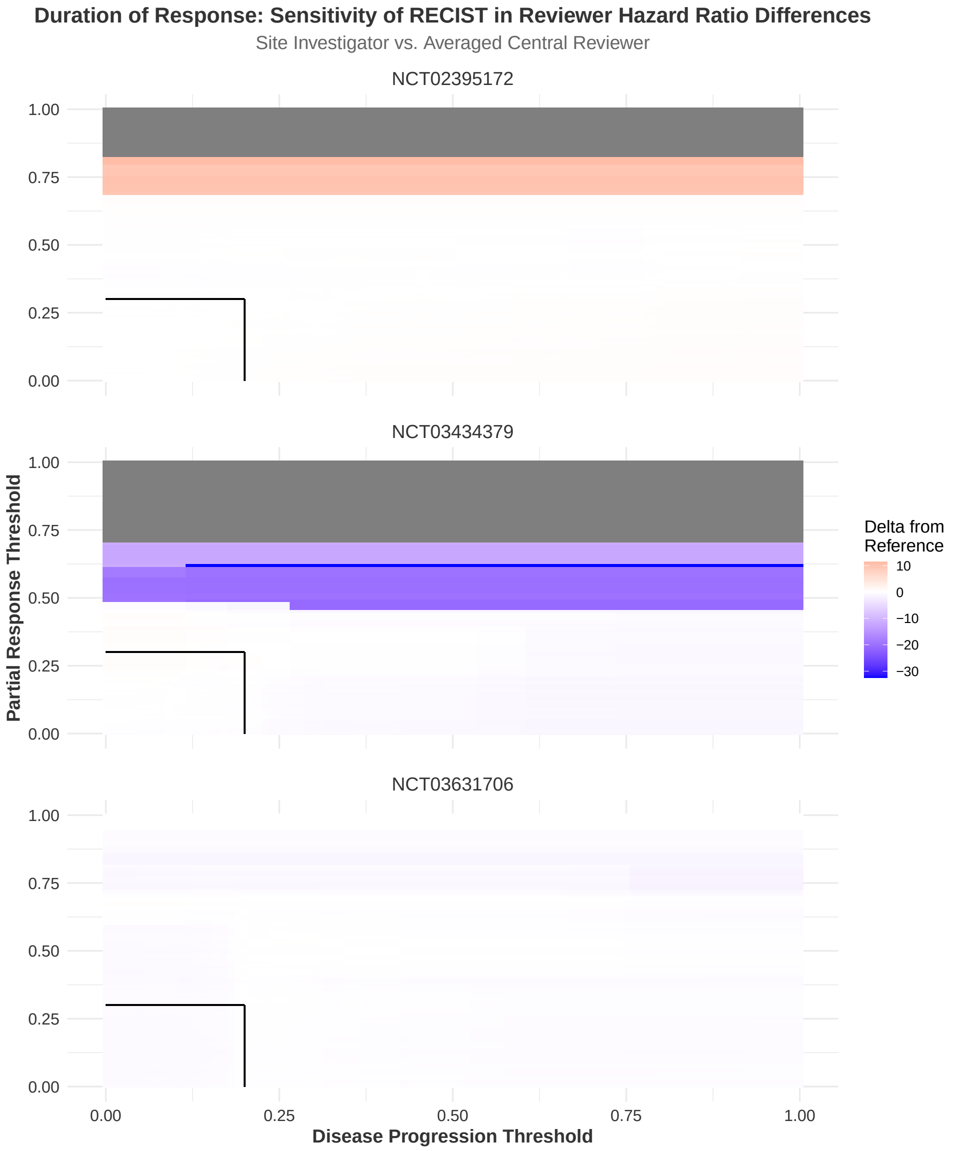


Figure A.33.: Heatmap of Change in Differences between Raters for DoR across RECIST thresholds, unfiltered data

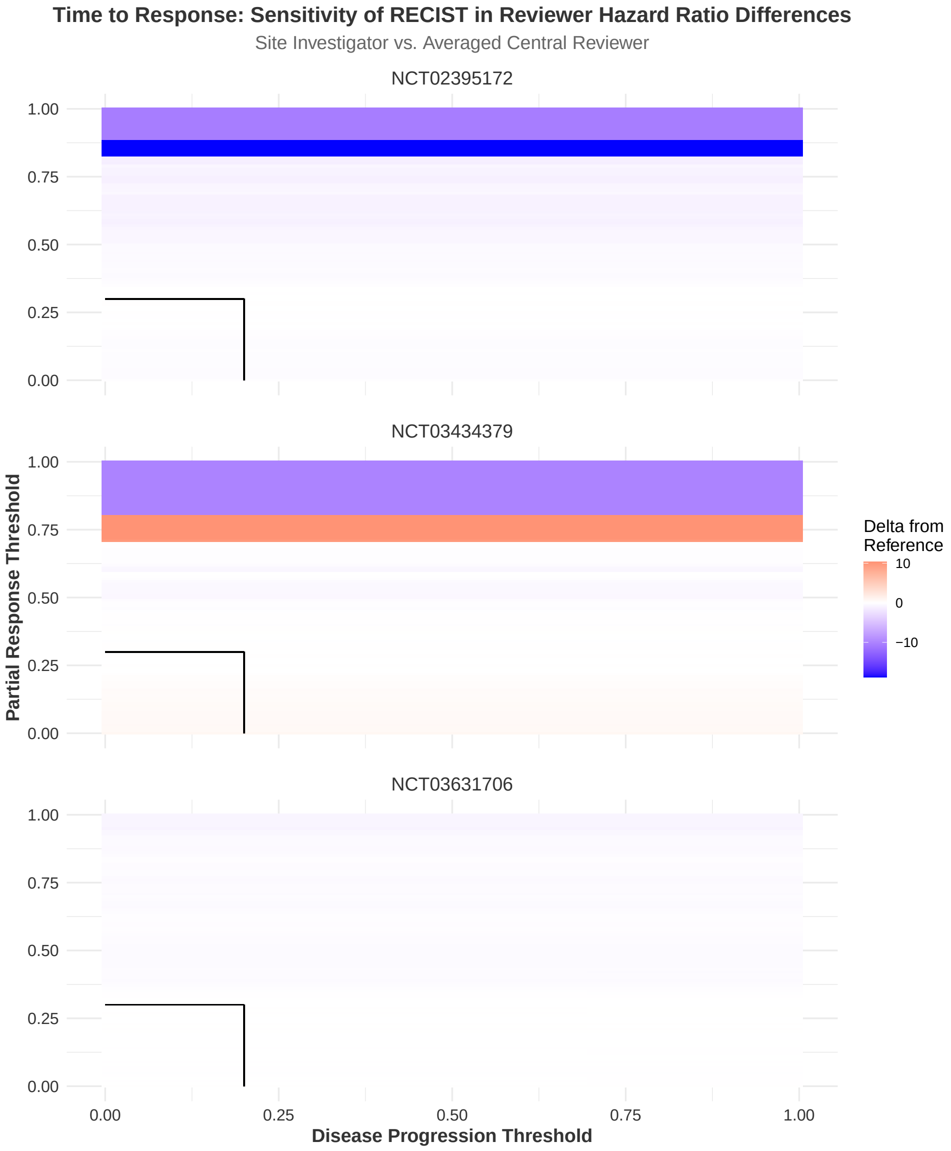


Figure A.34.: Heatmap of Change in Differences between Raters for TTR across RECIST thresholds, unfiltered data

# B. Code

Due to the length of the code required to reproduce the analyses in this thesis, the code is not included in the PDF output of this document. Instead, the code is available in the following locations:

1. In the USB drives that were provided with the thesis.
2. In the online version of the thesis:

    a. The readable version is available at neuroshepherd.github.io/masters-thesis

    b. And underlying code for this site is thus available at github.com/neuroshepherd/masters-thesis

The code in this appendix is intended to be for proof of work. As such, many of the file paths in the appendix are relative to "missing" directories or have been redacted for privacy reasons, and original, patient-level data files are not included here. Estimates from simulations and some other components that are not dependent on the original data files are included in the `data` directory of the online repository.

# C. Curriculum Vitae

```
Error in normalizePath(path, mustWork = TRUE) :
  path[1]="../cv/Patrick Callahan CV July 2025.pdf": No such file or directory
```

```
Error in knitr::include_graphics("../cv/cv_page_1.pdf") :
  Cannot find the file(s): "../cv/cv_page_1.pdf"
```

```
Error in knitr::include_graphics("../cv/cv_page_2.pdf") :
  Cannot find the file(s): "../cv/cv_page_2.pdf"
```

# D. Contributions

| Effort | Individual Contribution | Contributions by others (who, what, how much) |
|---|---|---|
| Conception | 20% | Original questions/hypotheses around the reliability and sensitivity of the RECIST scale were posited by Dr. Ursula-Kunz. PC followed up with more specific questions and analyses. |
| Field Work | 100% | Data for meta-analyses were collected by PC. Data from clinical trials was already available through a partner company, Transcelerate™. |
| Data Evaluation | 100% | All analyses were planned and performed by PC as well as the production of tables and figures. |
| Manuscript | 95% | Dr. Ursula-Kunz provided feedback on sections of the thesis. Mostafa Nasr also provided feedback on drafts. |

Elements of skeleton code for this thesis were generated using ChatGPT 4.1 and Claude Sonnet 3.7. The text of this thesis was likewise edited with the assistance of these AI tools, which provided suggestions and improvements to writing style throughout the writing process. All analyses and interpretations performed in this thesis, however, are my own.

I have written this thesis on my own without any other resources than those specified. The electronic and printed versions of this thesis are identical in terms of content, but formatting may differ. I have not submitted this thesis nor any part of it in any other examination procedure.

Printed Name

Signature

Date and Location