Published in final edited form as:
Eur J Cancer. 2009 January ; 45(2): 300–310. doi:10.1016/j.ejca.2008.11.010.

A Simulation Study to Evaluate the Impact of the Number of Lesions Measured on Response Assessment

Chaya S. Moskowitz[a], **Xiaoyu Jia**[a], **Lawrence H. Schwartz**[b], and **Mithat Gönen**[a]

[a] *Department of Epidemiology and Biostastistics, Memorial Sloan-Kettering Cancer Center 1275 York Avenue, New York, NY 10065*

[b] *Department of Radiology, Memorial Sloan-Kettering Cancer Center 1275 York Avenue, New York, NY 10065*

Abstract

The objectives of this study were to evaluate whether the number of lesions that are used to measure tumor burden affects response assessment and inter-rater variability. In order to accomplish this, a simulation study was conducted. Data were generated from a mixed-effects mixture model. Parameter values to input in the model were obtained from analysis of real data. Response assessments based on ten, five, three, two, and one lesion were evaluated. There was little difference between response assessments based on five lesions and response assessments based on ten lesions. When fewer than five lesions were used to assess response, there were notable differences from the ten-lesion based response assessment. Basing response assessment on a small number of lesions tends to overestimate response rates and lead to misclassification of patients' response status. Therefore, measuring five lesions per patient appears to sufficiently capture patients' response to therapy. Measuring fewer than five lesions results in loss of information that may adversely affect clinical trial results as well as patient management.

Introduction

Response to chemotherapy is an essential part of patient care and clinical research. Responding patients are often offered prolonged treatment and non-responders are quickly switched to another treatment regimen. Phase II clinical trials using response as the primary endpoint are ubiquitous and often are the primary determinants of whether a regimen should be taken to a definitive Phase III study. Hence accurate determination of response to chemotherapy is of critical importance.

Patients who receive treatment for cancer, whether as participants in a clinical trial or simply in the course of standard therapy, usually have multiple sites of metastases, in multiple organs. It is possible that the effect of the treatment will not be identical at all sites of metastases. For example, the treatment may shrink all of the lesions but by varying degrees (Figure 1). In some

Correspondence To: Chaya S. Moskowitz, Ph.D., Department of Epidemiology and Biostatistics Memorial Sloan-Kettering Cancer Center, 307 East 63rd Street, 3rd Floor, New York, NY 10065, Phone: (646) 735-8117, Fax: (646) 735-0010, Email: moskowc1@MSKCC.ORG.

Conflict of Interest
None declared

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

instances it is even possible for certain lesions to shrink in response to treatment while others grow. When assessing response to therapy, such as with RECIST guidelines,[1] it might therefore seem necessary to measure all lesions in order to best completely evaluate whether a patient is responding to a therapy. In fact, there is some empirical evidence in the literature that the variability in tumor response measurements are substantially reduced as increasing numbers of lesions are measured.[2] Often, however, resources do not permit radiologists to evaluate every lesion, and instead a subset or selection of lesions is chosen. The original WHO criteria recommended that 5 lesions be measured.[3, 4] In the RECIST 1.0 guidelines recommendations were for measuring all lesions up to a total of ten. In patients with more than ten lesions, the choice of which lesions to measure should be based on the size of the lesion and how suitable it is for repeated measurements.

In practice, measuring up to an upper limit of ten lesions may still be difficult and require more time and effort than many radiologists are routinely able to devote. A natural question to ask is whether fewer lesions can be measured while still sufficiently capturing a patient’s response to therapy. If so, how many lesions must one measure?

A key difficulty in answering this question is that the truth is rarely, if ever, known. In order to know whether a radiologist’s repeated assessments of tumor burden accurately reflects a patient’s change in true tumor burden over the course of a therapy, after being imaged at each time point the patient would need to undergo surgery and have all of their lesions measured. Clearly this is not possible.

One potential way to address this issue is to compare the response assessment that would have been obtained had we measured fewer than ten lesions with the response assessment obtained based on the complete ten lesions. In some sense, this approach considers the response assessment based on ten lesions to be the gold standard. It must be acknowledged, however, that response assessment based on ten lesions is not necessarily “the truth.” Unmeasured lesions beyond this upper limit may change the assessment. With this caveat in mind, comparing a ten-lesion based response assessment with a response assessment based on fewer lesions would help answer the question of whether measuring less than ten lesions would substantially alter the way tumor burden is currently evaluated under the RECIST 1.0 guidelines.

Another issue to take into consideration is that response assessment is radiologist-specific. That is, each radiologist selects what he or she perceives to be the ten largest lesions and then measures these lesions to the best of their ability. Inter-rater variability in this setting, however, is not inconsequential and whether a patient is determined to have responded to treatment may in fact differ between radiologists.[5, 6] We might further question, then, whether radiologists are more likely to agree in their response assessments if they measure more lesions. It seems logical to be most comfortable with response assessments that have a high level of agreement between multiple radiologists.

In this journal, the paper entitled “Individual Patient Data Analysis to assess modifications to the RECIST criteria” evaluates the EORTC data warehouse[7] and assess change in response by decreasing the number of lesions. There was concern that this database, collated from both industrial trials and cooperative group trials may not be truly representative of total tumor burden and number of lesions. In fact the mean number of lesions in those cooperative group trials was approximately 40% lower than the industrial independently reviewed trials. Therefore, part of the rationale of this simulation study is to more precisely approximate total tumor burden.

There are several advantages to conducting a simulation study including the ability to change the parameter settings used to simulate the data and the ability to explore the results in a variety of scenarios. For these reasons, we undertook this simulation study.

Methods

The primary aim of this simulation study was to evaluate whether the number of lesions measured affects response assessment. Secondarily, we were also interested in exploring whether the number of lesions measured affects inter-rater variability.

The flow-chart in Figure 2 portrays how data were simulated. Our general approach was to generate measurements on ten lesions for each patient in hypothetical single-arm trials at two time points (before and after treatment). Based on the change in lesion size between the two time points, we determined each patient's response as per RECIST 1.0 guidelines, first using all target lesions and then using only a subset of each patient's largest lesions. By varying the number of target lesions in the subset, we could assess the effect of the number of lesions measured. In addition, at each time point we generated two lesion measurements to simulate two radiologists reading each image. In all the scenarios we studied, we repeated this process 1000 times to simulate 1000 trials. For the interested reader, the Appendix describes the simulation model and provides a detailed description of how data were generated.

Results

Description of simulated data

To better understand the results, it is helpful to have an idea of what the simulated data actually look like. To this end, we first give an example of the lesion measurements generated for individual patients.

Figure 3 shows lesion measurements from two simulated patients who responded to therapy. The first row contains a patient whose lesion measurements were generated as part of a trial with a low response rate, while the second row contains a patient who was generated as part of a trial with a high response rate. The first figure in each row corresponds to measurements made by the first radiologist, and the second figure corresponds to the repeated reading by the second radiologist. The darker shaded bar depicts the baseline measurement of a lesion while the adjacent lighter color bar depicts the follow-up measurement after therapy.

In the first plot we see a patient for whom the first radiologist measured a decrease in lesion size for all of the ten lesions. While the decreasing trend is consistent across the ten lesions, the lesions decrease in varying degrees. The second plot in the top row shows the second radiologist's measurements of the same lesions. There is a similar tendency for all the lesions to decrease in size here as well, although the measurements themselves differ somewhat from those made by the first radiologist. In contrast to the first patient, most but not all, of the ten lesions in the second patient (shown in the bottom row) are judged by the first radiologist to be shrinking. Some of the lesions regress considerably, while other lesions remain unchanged or even appear to grow slightly. Again, an identical trend is seen from the measurements made by the second radiologist, but the actual measurements made by the two radiologists differ.

Figure 4 again shows lesion measurements for individual patients except here the top row displays measurements generated for a patient with progressive disease (top row) and a patient with stable disease (bottom row). For the patient with progressive disease, the measurements made by both radiologists show an increase in lesion size for the majority of lesions with one or two lesions (depending on the radiologist) decreasing very slightly in size. For the patient with stable disease, some lesions increase while others decrease in size, but the change in measurements is relatively small in all cases.

To depict the data for an entire simulated clinical trial, in Figure 5 we present waterfall plots for a single trial with a low response rate and another trial with a high response rate. In both cases, we used a trial with $N = 50$ patients and give the measurements from only one radiologist. The waterfall plots show the percent relative change in total tumor size for each patient.

Response and progression rates

To summarize results across 1000 simulated trials, we began by estimating the response and progression rates in each trial and then averaging these rates across the 1000 trials. The response rate was calculated by combining complete and partial responders as is frequently done in Phase II studies and then dividing by the number of patients in the trial ([CR + PR]/N, where N is the number of patients in the trial). The progression rate was calculated by dividing the number of patients with progressive disease by the total sample size (PD/N). In each trial we estimated response and progression rates separately for each radiologist. Agreement between the radiologists was calculated for each trial using the proportion of overall agreement[8] based on a 2 × 2 table of either responders × non-responders or progressors × non-progressors. We report the average of this proportion of agreement across the 1000 trials. Results of these analyses are shown in Tables 2–5.

Starting with Table 2, the first row shows the average response rates for trials with 25 patients assuming all ten lesions are included in the response assessments. For both radiologists, we see an 8% average response rate across the trials. Furthermore, there is a very high level of agreement between the two radiologists when all lesions are considered. In the next row, the same patients are analyzed in each trial, but now only the five largest target lesions are included in the assessments. These results are virtually identical to when all ten lesions are used. The response rates are again 8% for both radiologists and there is a very high level of agreement between the radiologists. In the third row, only the three largest lesions are considered in the response assessment. Here we see that the average response assessments now increase to 10%. The overall proportion of agreement drops slightly from what was seen for ten and five lesions, but is still very high. In looking at the next two rows, the estimated response rates based on two lesions and one lesion continue to increase while the proportion of agreement between the radiologists continues to decrease relative to what was seen for the rates based on all ten lesions. The additional rows in Table 2 show similar results when the trials have 50 patients and 100 patients. On average we estimate the same response rates for ten lesions and five lesions with very high levels of agreement between the two radiologists. When the number of lesions considered is reduced to three or fewer lesions, we see the average response rates increase and the agreement between the radiologists decrease. In the case of three lesions, differences relative to a ten lesion based assessment are noticeable, but rather small. In the situation when only the largest lesion is considered, these differences become quite substantial.

In Table 3, the same simulation sets are analyzed except now the focus is on the progression rate. Based on the response assessment using ten lesions, the average progression rate across the different scenarios (sample sizes and radiologists) is 4%. Small differences of 1% are seen when the number of lesions evaluated is decreased to either five or three lesions. The agreement between the radiologists is very high when ten lesions are used. It is relatively similar when five lesions are used, but begins to decrease slightly when only three lesions are used. When fewer than three lesions are evaluated, the estimated progression rate increases and the agreement between the radiologists decreases even further.

Table 4 and 5 are similar to Tables 2 and 3, except here data were simulated to have a high response rate of slightly over 40%. A trend comparable to that seen in Tables 2 and 3 is found here as well. With the response assessment based on ten lesions as the reference, using five lesions yields identical response and progression rates. Using three lesions produces similar but not uniformly identical estimates. More variation between the estimates arises if only two or one lesion is considered. The agreement between the radiologists decreases as the number of lesions evaluated decreases.

Discordant patient-level response assessment

To further analyze the simulation results, we calculated the proportion of patients whose response assessment changed based on the number of lesions being assessed. For this purpose we used binary assessments, i.e. responder versus non-responder (CR+PR versus SD+PD), and then separately progressor versus non-progressor (PD versus CR+PR+SD). Within each trial we estimated the proportion of patients with discordant assessments depending on how many lesions were included in the assessment. For instance, we compared five lesions with ten lesions by counting the number of patients who were classified as responders using a ten-lesion based assessment and non-responders using a five-lesion based assessment plus the number of patients who were classified as non-responders using a ten-lesion based assessment and responders using a five-lesion based assessment. This number was divided by the number of patients in the trial and then the result was averaged over the 1000 trials in the simulation set. This computation was repeated for all pair-wise comparisons of ten, five, three, two, and one lesion. Tables 6–9 contain these results.

In Table 6 we see that when the therapy has a low response rate, for both radiologists on average less than 1% of patients are reclassified as either responders or non-responders based on whether ten or five lesions are measured. This result holds regardless of the sample size used. Depending on whether ten lesions or three lesions are measured, approximately 2% of patients have their response status change. This number increases to 6% when comparing ten lesions with two lesions, and 15% when comparing ten lesions with the single largest lesion.

Table 7 shows that on average 2% of patients are reclassified as having progressive disease or not depending on whether ten lesions or five lesions are measured. As fewer lesions are considered, the proportion of patients found to have discordant progression assessments increases.

Table 8 shows results similar to Table 6 when the therapy has a high response rate. The results are very similar to what was seen in Table 6. Similarly, Table 9 is comparable to Table 7.

Discussion

The accurate and reproducible measurement of tumor burden at baseline and follow up CT scans is of paramount importance in assessing response to therapy. The earliest use of tumor burden and imaging biomarkers mandated that more than one tumor be assessed in a patient. As imaging modalities have improved, so has the ability to detect metastatic disease and smaller changes in these metastases. Nevertheless, there remains the question of what proportion of tumor burden is necessary to assess and measure quantitatively. Actual trial data where “all lesions” are measured is not common, certainly not across multiple primary tumors and types of therapy. Therefore, we sought to evaluate this question with a simulation study whose simulation parameters are based upon actual tumor measurements.

We began by considering the scenario of ten target lesions per patient and decreasing the number of lesions measured to five, three, two or one, similar to the approach in the RECIST data warehouse analysis[7].

Across all the scenarios that were considered we consistently saw little or no difference when comparing response and progression assessments based on the five largest target lesions with these assessments based on ten target lesions. In some, but not all, situations, agreement between the radiologists decreased very slightly when comparing the five-lesion based assessment with the ten-lesion based assessment.

The differences in response assessment between ten lesions and three, two or one lesion was more prominent. It is obvious that measuring a single lesion gives misleading results. In a clinical trial of 50 patients, a typical size for Phase II studies, response rates would be overestimated by 7–14% (Tables 2 and 4). A trial of this size would often be powered to detect a difference of 15–20% in response rates[9] so this amount of overestimation can easily lead to declaring an ineffective drug promising. The amount of overestimation is 3–6% for two lesions and 1–2% for three lesions. While measuring three lesions represents an improvement over measuring two or a single lesion, the latter figures fall short of our original goal of sufficiently representing patient experience with a smaller number of lesions.

On an individual basis measuring one to three lesions continue to be less- or insufficient. When compared with 10 lesions, measuring a single lesion causes 11–16% of the patients to be misclassified. With two lesions, this misclassification drops to 5–6% and with three lesions to 2–3%. Measuring five lesions assures that at most 1% (and in many cases less than 1%) of the patients are incorrectly classified.

Another way to evaluate Tables 6–9 is to examine the misclassification rates of measuring one to three lesions as compared with measuring five lesions. Five lesions reduced misclassification rate by 2% in absolute terms when compared with three lesions and by 5–15% when compared with one or two lesions.

The tables demonstrate that measuring smaller number of lesions leads to overstated response rates and a higher proportion of misclassified patients. This is consistently seen across trials with low and high response rates as well as patients with low and high response probabilities. Measuring five lesions results in a very small loss of information, whereas measuring fewer lesions may not sufficiently capture patient response.

In all of our analyses we used the largest lesions based on our actual size, effectively assuming that the radiologist always chooses the “correct” lesions based on a size criterion. In practice, this theoretically would require measuring all the lesions first, which defies the purpose of lesion selection. Therefore our results are optimistic estimates of error introduced by measuring a smaller number of lesions. In practice we can see even higher biases in estimating the response rate and higher rates of misclassified patients as radiologists occasionally fail to include the largest lesions according to their visual inspection, or purposely may not include the largest lesions because they feel a particular lesion may not be reproducibly measurable.

There are some limitations of this simulation study. The results depend on a model which required several assumptions. The assumptions we made (such as the normality of the tumor measurements on the natural logarithm scale and the correlation structure) may deviate from the truth. Thus the simulated lesion measurements may not precisely represent true lesion measurements in clinical trials. However, it is important to keep in mind that our goal was not to perfectly model how tumors change over time, but rather to have a sensible approximation that would allow us to assess the affect of the number of lesions on trial-level summary statistics such as the response rate. It is often said that all models are wrong but some are useful.[10] Based on our analysis of actual lesion measurements, we believe that the model used in this simulation study is reasonable and useful for the stated purposes of the study.

Our model simulates lesion measurements at only two time points. This approach allows us to assess how the number of lesions affects estimates of the proportion of patients responding to a treatment and the proportion of patients having progressive disease. This approach, however, does not allow us to look or conclude about optimizing the number of lesions for time-to-event outcomes such as time to progression or progression free survival. Generating data at more than two time points is extremely complex due to the multiple possible variants in lesion growth curves in the presence of a treatment. For instance, lesions may respond to treatment and continually shrink, or they may initially respond but then begin to grow again. Alternatively, lesions may not respond to treatment and continue to grow over time. These scenarios are just several possibilities. Attempting to model this process (and in addition determining the proportion of lesions that behave in a particular fashion) is beyond the scope of this paper. It remains an important question, however, as more and more trials use progression-free survival as their primary endpoint.

A third issue to consider is that our definition of progressive disease is limited by the fact that we only considered the effect of lesions that were present at a baseline measurement. In clinical practice and in clinical trials, patients are assessed not only based upon target lesions, but non target disease and the development of new lesions. It is uncertain the impact that non target assesement, which is generally very qualitative, and new lesions, which may or may not be recognized as new metastatic disease, will have on the concordance of disease progression.

Finally we did not consider a model in which patient response is a predictor of overall survival. If, as we argued above, generating longitudinal measurements is much more complex than a snapshot in time, capturing what happens beyond progression in a simulation study is even more complicated by factors like salvage therapy, age, and co-morbid conditions and toxicities resulting from the initial treatment. Nevertheless this is also an interesting question that could be considered in the future, preferably starting with the analysis of large randomized trials.

The use of tumor measurements extends beyond RECIST [11]. Increasingly, investigators use measurements in waterfall analyses and evaluate responses, progressions and simply change in tumor size as a continuous variable.[12] The use of tumor measurements in this regard may dictate even stricter requirements for concordance and therefore the need to measure more lesions or measure lesions more accurately. Based upon the existing data and current contemporary use of tumor measurements in RECIST, it appears that decreasing the number of lesions measured to five is a satisfactory compromise between capturing “total” tumor burden and the resource commitment needed to accomplish this goal both for a single patient and for a clinical trial. This variable however will need to be continually re-evaluated with changes in therapies, in modalities used to assess these therapies, and in metrics or response criteria used to categorize the benefit of the therapy.

Acknowledgements

The authors would like to thank the members of the RECIST International Working Group for their comments during the development of the model. In particular, we are grateful to Dr. Lori Dodd for her helpful suggestions. We are also thankful to Ms. Joanne Chou for preparing the warehouse data used to obtain input values to the simulation model.

References

1. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst Feb 2;2000 92(3):205–216. [PubMed: 10655437]
2. Schwartz LH, Mazumdar M, Brown W, Smith A, Panicek DM. Variability in response assessment in solid tumors: effect of number of lesions chosen for measurement. Clin Cancer Res Oct 1;2003 9(12): 4318–4323. [PubMed: 14555501]
3. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. Cancer Jan 1;1981 47(1):207–214. [PubMed: 7459811]
4. WHO handbook for reporting results of cancer treatment. Geneva (Switzerland): World Health Organization Offset Publication No. 48; 1979.

5. Hopper KD, Kasales CJ, Van Slyke MA, Schwartz TA, TenHave TR, Jozefiak JA. Analysis of interobserver and intraobserver variability in CT tumor measurements. AJR Am J Roentgenol Oct; 1996 167(4):851–854. [PubMed: 8819370]

6. Schwartz LH, Ginsberg MS, DeCorato D, Rothenberg LN, Einstein S, Kijewski P, Panicek DM. Evaluation of tumor measurements in oncology: use of film-based and electronic techniques. J Clin Oncol May;2000 18(10):2179–2184. [PubMed: 10811683]

7. Bogaerts J, Ford R, Sargent D, Schwartz LH, Rubinstein L, Lacombe D, Eisenhauer EA, Verweij J, Therasse P. for the RECIST Working Party. Individual Patient Data Analysis to assess modifications to the RECIST criteria. Eur J Cancer. this issue

8. Fleiss, JL.; Levin, B.; Paik, MC. Statistical Methods for Rates and Proportions. New York: John Wiley & Sons, Inc; 2003.

9. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials Mar;1989 10(1): 1–10. [PubMed: 2702835]

10. Box, GEP.; Draper, NR. Empirical Model-Building and Response Surfaces. New York: John Wiley & Sons, Inc.; 1987.

11. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New Response Evaluation Criteria In Solid Tumors: Reveised RECIST Guideline (version1.1). Eur J Cancer. this issue

12. Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non small-cell lung cancer. J Natl Cancer Inst Oct 3;2007 99(19):1455–1461. [PubMed: 17895472]

13. Harville DA. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association 1977;72(358):320–338.

14. James LP, Zhao B, Moskowitz CS, Riely GJ, Miller VA, Guo P, Ginsberg MS, Kris M, Schwartz LH. Reproducibility of computed tomography (CT) measurements of lung cancer. Journal of Clinical Oncology 2008;26May 20 suppl; abstr 8002

Appendix

Model

In each simulation set, we simulated data for 1000 hypothetical Phase II studies. For each Phase II study in the same simulation set, we generated data on *N* patients with ten lesions. We specified that based on measurements from ten lesions a proportion of the patients were responders, denoted by *Z*=2, a proportion were progressors, denoted by *Z*=1, and a proportion had stable disease, denoted by *Z*=0. To generate lesion measurements, we used a mixture of three mixed effects models that allowed parameter values to differ between responders, progressors, and patients who were neither responding nor progressing:

$$\begin{aligned}\ln(y_{ijkt}) = & I(Z_i=0) \times (\delta_t^{(Z=0)} + \lambda_i^{(Z=0)} + \alpha_{j(i)}^{(Z=0)} + \varepsilon_{ijkt}^{(Z=0)}) \\ & + I(Z_i=1) \times (\delta_t^{(Z=1)} + \lambda_i^{(Z=1)} + \alpha_{j(i)}^{(Z=1)} + \varepsilon_{ijkt}^{(Z=1)}) \\ & + I(Z_i=2) \times (\delta_t^{(Z=2)} + \lambda_i^{(Z=2)} + \alpha_{j(i)}^{(Z=2)} + \varepsilon_{ijkt}^{(Z=2)}) + \beta_k\end{aligned}$$

where $i = 1,\ldots, N;\ j = 1,\ldots,10;\ k = 1,2;\ t = 0,1$. In this model, y_{ijkt} is the lesion measurement for the i^{th} patient, j^{th} lesion, at time *t*, as measured by the k^{th} radiologist. Time *t*=0 represents a baseline measurement while time *t*=1 represents a follow-up measurement. The model specifies that y_{ijkt} is a function of five components where the values of all the components except for one, $\beta_{k,}$ depend on the value of *Z* :

1. $\delta_t^{(Z=0)}$ is a fixed effect that represents the overall mean lesion measurement among responders at time *t*. Similarly, $\delta_t^{(Z=1)}$ and $\delta_t^{(Z=2)}$ represent the overall mean lesion

measurement at time *t* among patients with progression and stable disease, respectively.

2. $\lambda_i^{(Z=0)}$,$\lambda_i^{(Z=1)}$ and $\lambda_i^{(Z=2)}$ are the subject random effects for responders, progressors, and patients with stable disease, respectively. We assume $\lambda_i^{(Z=0)} \sim \text{Normal}(0,\sigma^2_{\lambda(Z=0)}),\lambda_i^{(Z=1)} \sim \text{Normal}(0,\sigma^2_{\lambda(Z=1)})$ and $\lambda_i^{(Z=2)} \sim \text{Normal}(0,\sigma^2_{\lambda(Z=2)})$.

3. $\alpha_{j(i)}^{Z=0}$,$\alpha_{j(i)}^{Z=1}$ and $\alpha_{j(i)}^{Z=2}$ are the lesion random effects which are nested within subject. We assume $\alpha_{j(i)}^{Z=0} \sim \text{Normal}(0,\sigma^2_{\alpha(Z=0)}),\alpha_{(j)(i)}^{Z=1} \sim \text{Normal}(0,\sigma^2_{\alpha(Z=1)})$ and $\alpha_{j(i)}^{Z=2} \sim \text{Normal}(0,\sigma^2_{\alpha(Z=2)})$.

4. $\varepsilon_{ijkt}^{(Z=0)}$,$\varepsilon_{ijkt}^{(Z=1)}$ and $\varepsilon_{ijkt}^{(Z=2)}$ are the random error terms with $\varepsilon_{ijkt}^{(Z=0)} \sim \text{Normal}(0,\sigma^2_{\varepsilon(Z=0)}),\varepsilon_{ijkt}^{(Z=1)} \sim \text{Normal}(0,\sigma^2_{\varepsilon(Z=1)})$, and $\varepsilon_{ijkt}^{(Z=2)} \sim \text{Normal}(0,\sigma^2_{\varepsilon(Z=2)})$.

5. β_k is the radiologist random effect. We assume that this random effect is distributed similarly among all simulated patients and that $(\beta_1, \beta_2)^T$ ~ Multivariate Normal $(0,\Sigma_\beta)$ where the components of Σ_β are σ_β^2 on the diagonal and $\sigma_{\beta 12}$ on the off-diagonal.

The natural logarithm of each lesion measurement is then determined by summing across each of these component pieces. We use the natural logarithm of the lesion measurements because in our experience with analysis of real data sets we have found that the logarithm of the measurement is more reasonably approximated by a normal distribution than are the raw, untransformed measurements.

After the lesion measurements have been generated according to this model, measurements were summed across lesions within patient for each time point and radiologist. The relative percent change from baseline to follow-up was calculated for the two radiologists. That is,

$$R_{ik}=100 \times \frac{\sum_{j=1}^{J} y_{ijk1} - \sum_{j=1}^{J} y_{ijk0}}{\sum_{j=1}^{J} y_{ijk0}}$$

is the relative percent change in tumor burden for the i^{th} patient, k^{th} reading. This yielded *N* pairs, (R_{11}, R_{12}), (R_{21}, R_{22}), …, (R_{N1}, R_{N2}), of measured changes in tumor sizes. For each patient there are two response assessments arising from the two radiologists. We were primarily interested in calculating the change within radiologist. In other words, our primary focus was on the relative change from baseline to follow-up for the first radiologist and then separately for the second radiologist, rather than studying the relative change from baseline, radiologist 1 to follow-up radiologist 2, for instance. The R_{ik} were then divided into response categories using the definitions from RECIST. A complete response (CR) was defined by $R_{ik} = -100$, while a partial response was defined as $-100 < R_{ik} \le -30$. $R_{ik} \ge 20$ denoted progressive disease (PD) and $-30 \le R_{ik} \le 20$ denoted stable disease (SD).

After response has been assessed using the measurements from ten lesions, subsets of the ten lesions were selected and analyzed for each patient. For each patient the *S* largest lesions as measured by the k^{th} radiologist at the baseline measurement were selected. That is, the lesions that are selected to be included in the subsets may differ between the two radiologists. Response assessment was calculated as above based only on the *S* lesions while ignoring the remaining lesions.

Simulations were performed in the statistical software package R (copyright The R Foundation for Statistical Computing).

Parameter values

We aimed to simulate data that would show the affect of decreasing the number of lesions measured first in Phase II studies with a substantial treatment effect and then in Phase II studies with a smaller treatment effect. In order to obtain parameter values to use in the simulation model, we analyzed existing data from actual studies on real patients.

Values for $\delta_t^{(Z=0)},\delta_t^{(Z=1)},\delta_t^{(Z=2)},\sigma^2_{\lambda(Z=0)},\sigma^2_{\lambda(Z=1)},\sigma^2_{\lambda(Z=2)},\sigma^2_{\alpha(Z=0)},\sigma^2_{\alpha(Z=1)},\sigma^2_{\alpha(Z=2)},\sigma^2_{\varepsilon(Z=0)},\sigma^2_{\varepsilon(Z=1)}$, and $\sigma^2_{\varepsilon(Z=2)}$ were obtained from several of the trials from the data warehouse that were independently reviewed ("New Response Evaluation Criteria in Solid Tumors: Revised RECIST Guideline (Version 1.1)"). We fit mixed effects models using the method of restricted maximum likelihood[13] separately to responders, progressors, and patients with stable disease for each of the eight protocols included in the data set and obtained eight sets of estimates for each of the above parameters. The PROC MIXED procedure in SAS (version 9.1 for Windows, The SAS Institute, Inc.) is used for this purpose.

We combined estimates using weighted averages of the protocol-specific estimates, first averaging over protocols with a more substantial treatment effect and then separately averaging over protocols with a smaller treatment effect. We focused on using the same progression rate in both instances, and again obtained parameter estimates by taking weighted averages over protocols with similar progression rates. Based on this analysis, we simulated data with a low progression rate and a response rate that was either moderately low or relatively high. Additional results of this analysis detailing the parameter values we used to simulate data are contained in Table 1.

Values for σ^2_β and $\sigma_{\beta 12}$ were obtained by fitting a random effects model to data collected for a separate study looking at the reproducibility of CT measurements. These data are described in detail elsewhere.[14] The parameter estimates that resulted from this analysis and were used to generate data for our simulation study are also contained in Table 1.

Within each set of simulations, we generated data for a fixed number of patients, *N*, for each hypothetical Phase II study. We used values of 25, 50, and 100 for *N* across different sets of simulations. We studied the affect on estimates of response and progression when *S*=5, 3, 2, and 1 target lesions were measured in comparison to when 10 target lesions were measured.

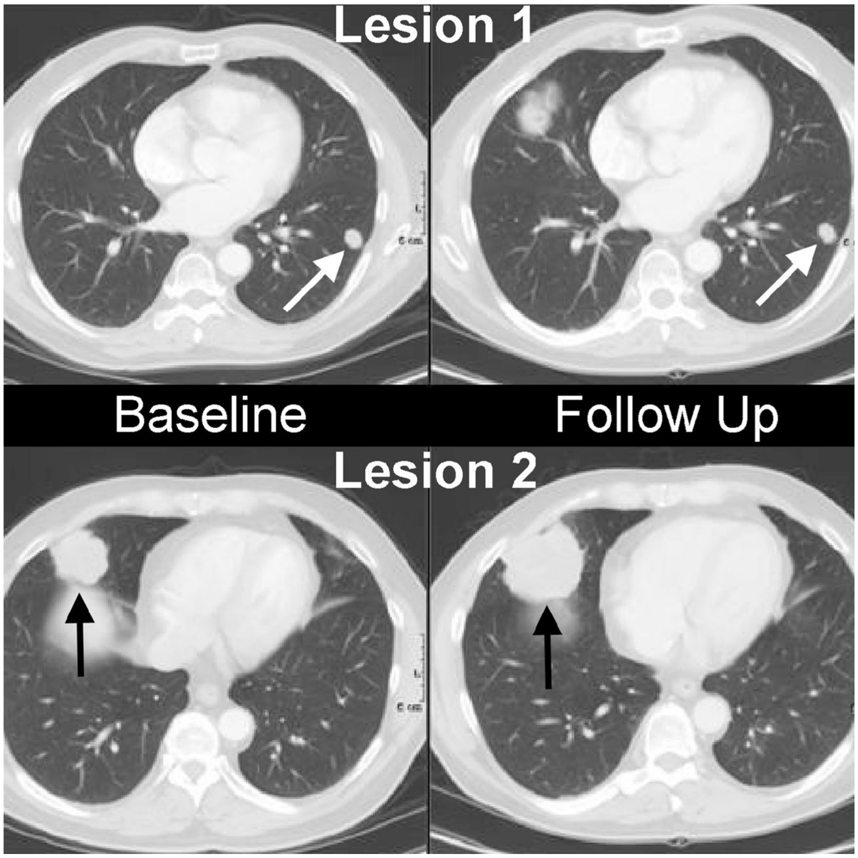

Figure 1.
Metastatic disease to the lungs. Note the smaller lesion (white arrows) has not changed in size from baseline to follow up, while the larger lesion (black arrows) has increased in size.

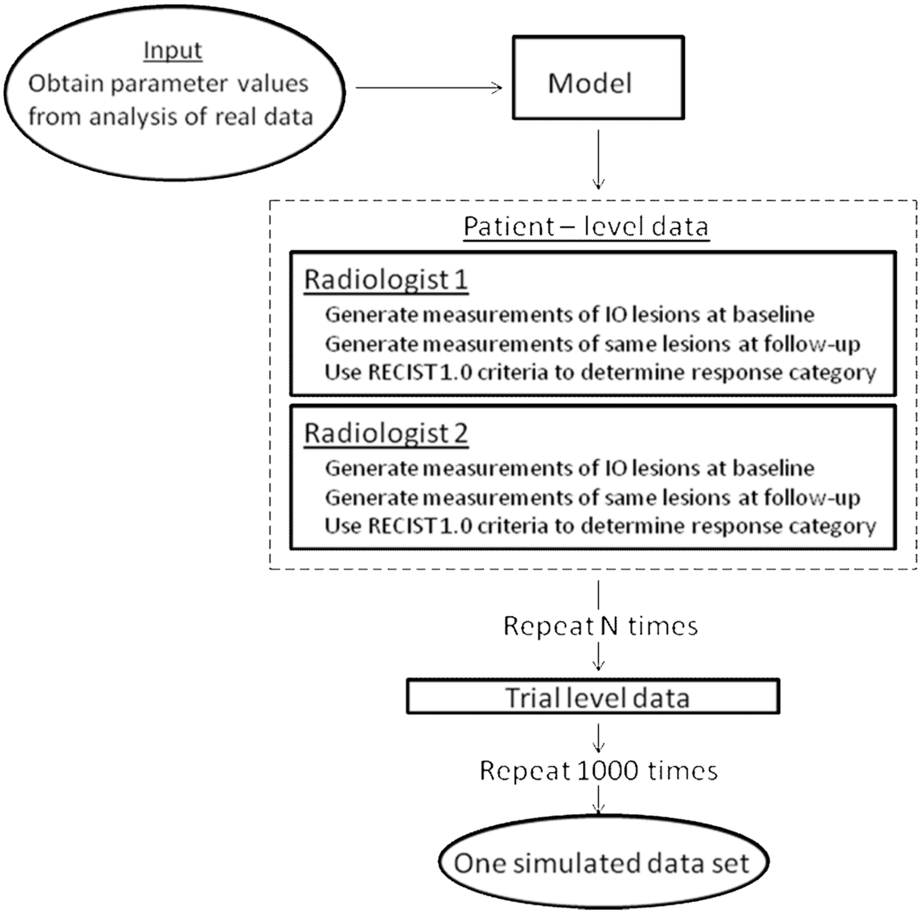
Input
Obtain parameter values from analysis of real data
Model
Patient – level data
Radiologist 1
Generate measurements of IO lesions at baseline
Generate measurements of same lesions at follow-up
Use RECIST 1.0 criteria to determine response category
Radiologist 2
Generate measurements of IO lesions at baseline
Generate measurements of same lesions at follow-up
Use RECIST 1.0 criteria to determine response category
Repeat N times
Trial level data
Repeat 1000 times
One simulated data set

Figure 2.
Flowchart of the simulation process

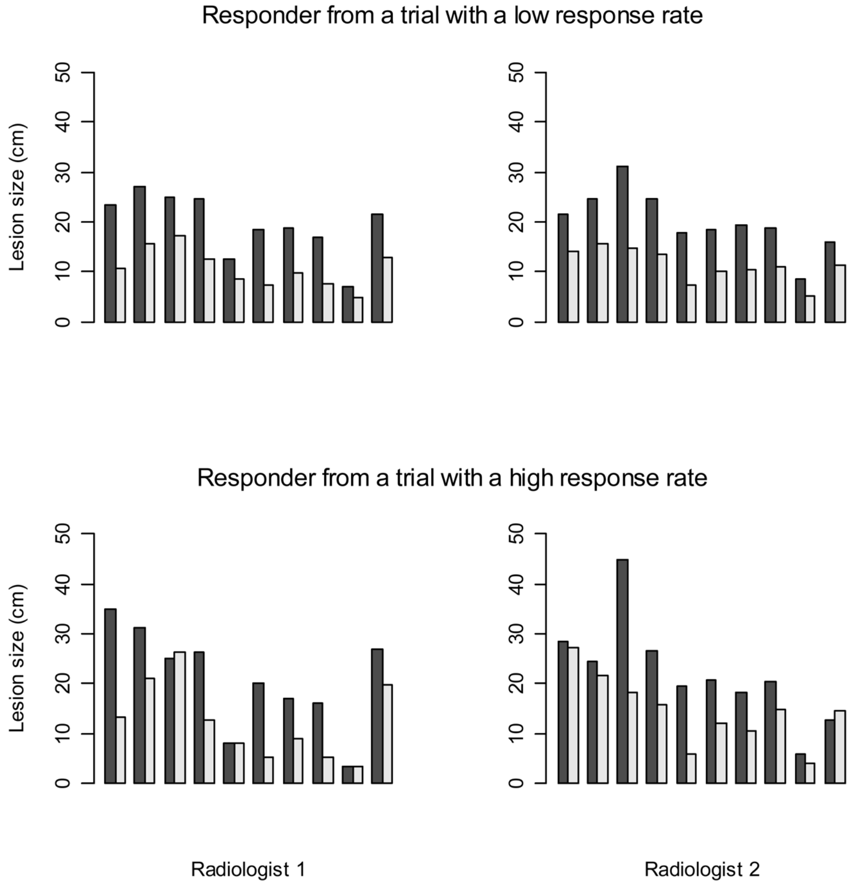

Figure 3.

An example of 10 lesion measurements at baseline (shaded bar) and follow-up (light bar) for two sample patients who were classified as responders

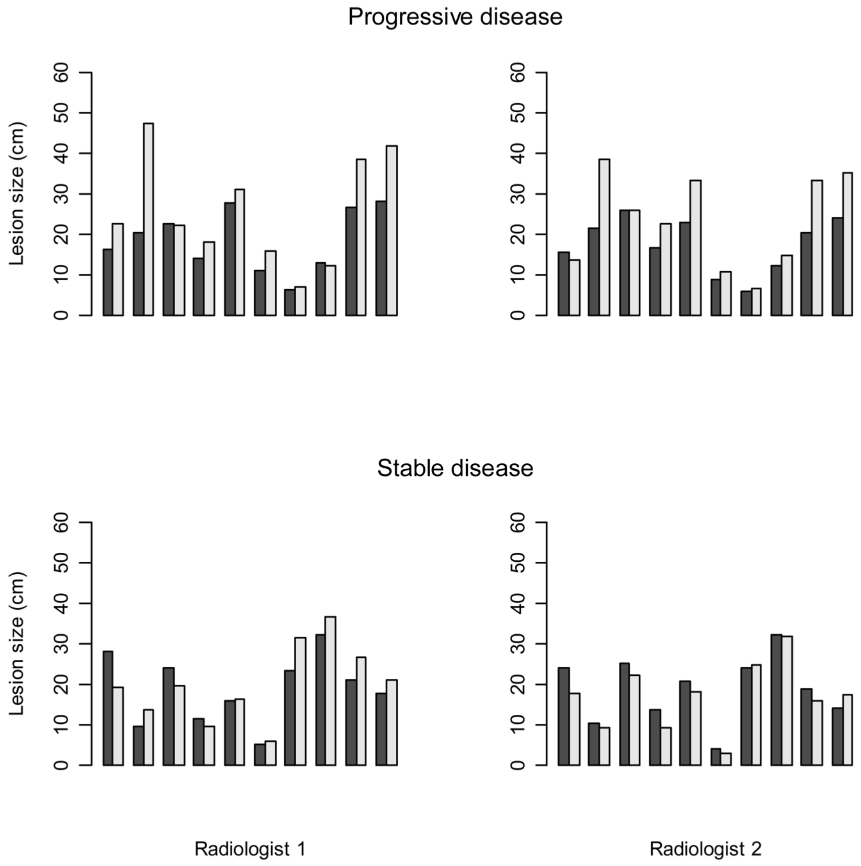

Figure 4.

An example of 10 lesion measurements at baseline (shaded bar) and follow-up (light bar) for a patient with progressive disease and a patient with stable disease.

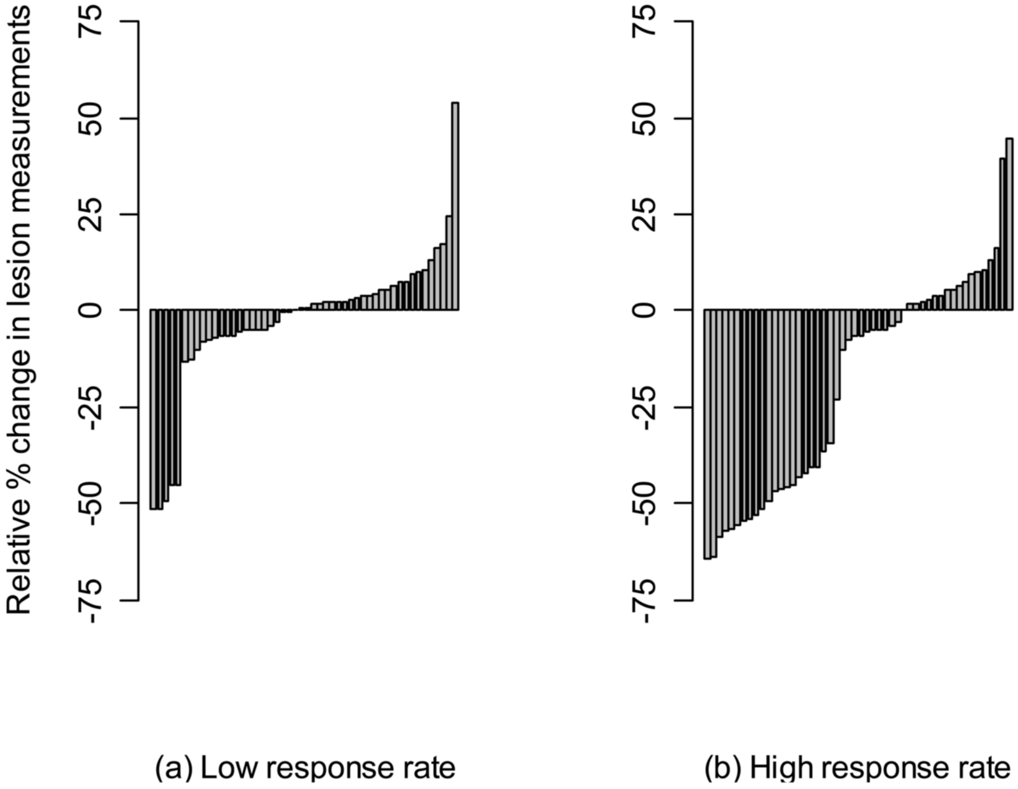

Figure 5.

Waterfall plots for a single simulated Phase II study with a low response rate and a single simulated Phase II study with a high response rate

Table 1

Parameter values used in simulation model

	δ_0	δ_1	σ_{λ}^{2}	σ_{α}^{2}	σ_{ε}^{2}	σ_{β}^{2}	$\sigma_{\beta 12}$
Responders (Z=0, low response)	3.10	2.43	0.06	0.11	0.03	0.003	0.001
Responders (Z=0, high response)	3.10	2.34	0.16	0.19	0.18	0.003	0.001
Progressors (Z=1)	3.10	3.36	0.13	0.21	0.06	0.003	0.001
Patients with stable disease (Z=2)	3.10	3.08	0.12	0.15	0.03	0.003	0.001

Table 2

Low response: average response rates across 1000 simulated trials

N	Number of lesions measured	Response rate (%)		Overall proportion of agreement
		Radiologist		
		1	2	
25	10	8%	8%	> 0.99
	5	8%	8%	> 0.99
	3	10%	10%	0.96
	2	13%	13%	0.90
	1	22%	22%	0.76
50	10	10%	10%	> 0.99
	5	10%	10%	> 0.99
	3	12%	12%	0.96
	2	16%	15%	0.90
	1	24%	24%	0.76
100	10	10%	10%	> 0.99
	5	10%	10%	0.99
	3	12%	12%	0.96
	2	15%	15%	0.90
	1	24%	24%	0.76

Table 3

Low response: average progression rates across 1000 simulated trials

N	**Number of lesions measured**	**Progression rate (%)**		**Overall proportion of agreement**
		Radiologist		
		1	**2**	
25	10	4%	4%	0.97
	5	3%	4%	0.96
	3	4%	5%	0.93
	2	6%	6%	0.90
	1	10%	10%	0.83
50	10	4%	4%	0.98
	5	4%	3%	0.96
	3	4%	4%	0.93
	2	6%	6%	0.91
	1	9%	9%	0.85
100	10	3%	4%	0.98
	5	3%	4%	0.96
	3	4%	4%	0.94
	2	6%	6%	0.91
	1	9%	9%	0.85

Table 4

High response: average response rates across 1000 simulated trials

N	**Number of lesions measured**	**Response rate (%)**		**Overall proportion of agreement**
		Radiologist		
		1	**2**	
25	10	43%	43%	0.98
	5	43%	43%	0.98
	3	44%	44%	0.95
	2	46%	46%	0.91
	1	50%	50%	0.81
50	10	43%	43%	0.98
	5	43%	43%	0.98
	3	44%	44%	0.95
	2	46%	46%	0.91
	1	50%	50%	0.81
100	10	42%	42%	0.98
	5	42%	42%	0.98
	3	43%	43%	0.95
	2	45%	45%	0.91
	1	49%	50%	0.81

Table 5

High response: average progression rates across 1000 simulated trials

N	**Number of lesions measured**	**Progression rate (%)**		**Overall proportion of agreement**
		Radiologist		
		1	**2**	
25	10	3%	3%	0.98
	5	3%	3%	0.97
	3	3%	4%	0.95
	2	4%	4%	0.94
	1	6%	7%	0.89
50	10	3%	3%	0.98
	5	3%	3%	0.97
	3	3%	3%	0.95
	2	4%	4%	0.94
	1	7%	7%	0.90
100	10	3%	3%	0.98
	5	3%	3%	0.97
	3	3%	3%	0.95
	2	4%	4%	0.93
	1	7%	7%	0.89

Table 6

Response classification in trials with a low response rate: proportion of patients classified discordantly into responders and non-responders.

Number of lesions being compared	N=25		N=50		N=100	
	Radiologist 1	Radiologist 2	Radiologist 1	Radiologist 2	Radiologist 1	Radiologist 2
10 vs 5	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
10 vs 3	0.02	0.02	0.02	0.02	0.02	0.02
10 vs 2	0.06	0.06	0.06	0.06	0.06	0.06
10 vs 1	0.15	0.15	0.15	0.15	0.15	0.15
5 vs 3	0.02	0.02	0.02	0.02	0.02	0.02
5 vs 2	0.05	0.05	0.05	0.05	0.05	0.05
5 vs1	0.15	0.15	0.15	0.14	0.15	0.14
3 vs 2	0.04	0.04	0.04	0.04	0.04	0.04
3 vs 1	0.13	0.14	0.14	0.13	0.14	0.13
2 vs 1	0.12	0.12	0.12	0.11	0.12	0.11

Table 7

Progression classification in trials with a low response rate: proportion of patients classified discordantly into progressors and non-progressors.

Number of lesions being compared	N=25		N=50		N=100	
	Radiologist 1	Radiologist 2	Radiologist 1	Radiologist 2	Radiologist 1	Radiologist 2
10 vs 5	0.02	0.02	0.02	0.02	0.02	0.02
10 vs 3	0.03	0.04	0.04	0.03	0.04	0.04
10 vs 2	0.05	0.06	0.06	0.05	0.05	0.05
10 vs 1	0.10	0.10	0.10	0.10	0.10	0.09
5 vs 3	0.03	0.03	0.03	0.03	0.03	0.03
5 vs 2	0.05	0.05	0.05	0.05	0.05	0.05
5 vs1	0.09	0.09	0.09	0.09	0.09	0.09
3 vs 2	0.04	0.04	0.04	0.04	0.04	0.04
3 vs 1	0.09	0.09	0.08	0.08	0.08	0.08
2 vs 1	0.08	0.08	0.07	0.07	0.07	0.07

Table 8

Response classification in trials with a high response rate: proportion of patients classified discordantly into responders and non-responders.

Number of lesions being compared	N=25		N=50		N=100	
	Radiologist 1	**Radiologist 2**	**Radiologist 1**	**Radiologist 2**	**Radiologist 1**	**Radiologist 2**
10 vs 5	0.01	0.01	0.01	0.01	0.01	0.01
10 vs 3	0.03	0.03	0.03	0.03	0.03	0.03
10 vs 2	0.06	0.05	0.05	0.05	0.05	0.05
10 vs 1	0.12	0.11	0.12	0.12	0.12	0.12
5 vs 3	0.02	0.02	0.02	0.02	0.02	0.02
5 vs 2	0.05	0.05	0.05	0.05	0.05	0.05
5 vs1	0.11	0.11	0.11	0.12	0.11	0.11
3 vs 2	0.04	0.03	0.04	0.04	0.04	0.04
3 vs 1	0.10	0.10	0.11	0.11	0.11	0.11
2 vs 1	0.09	0.09	0.09	0.09	0.09	0.09

Table 9

Progression classification at a single timepont in trials with a high response rate: proportion of patients classified discordantly into progressors and non-progressors

Number of lesions being compared	N=25		N=50		N=100	
	Radiologist 1	Radiologist 2	Radiologist 1	Radiologist 2	Radiologist 1	Radiologist 2
10 vs 5	0.02	0.02	0.01	0.01	0.01	0.01
10 vs 3	0.03	0.03	0.03	0.03	0.03	0.03
10 vs 2	0.04	0.04	0.04	0.04	0.04	0.04
10 vs 1	0.07	0.07	0.07	0.07	0.07	0.07
5 vs 3	0.02	0.02	0.02	0.02	0.02	0.02
5 vs 2	0.03	0.03	0.03	0.03	0.03	0.03
5 vs1	0.06	0.06	0.06	0.06	0.06	0.06
3 vs 2	0.02	0.02	0.02	0.02	0.02	0.02
3 vs 1	0.06	0.06	0.06	0.06	0.06	0.06
2 vs 1	0.05	0.05	0.05	0.05	0.05	0.05