Use and Misuse of Waterfall Plots

Tiffany Shao, Lisa Wang, Arnoud J. Templeton, Raymond Jang, Francisco W. Vera-Badillo, Mairead G. McNamara, Myles Margolis, Tae Kyoung Kim, Mehrdad Sinaei, Hassan Shoushtari, Ian F. Tannock

Manuscript received April 8, 2014; revised June 11, 2014; accepted September 12, 2014.

Correspondence to: Ian Tannock, MD, PhD, DSc, Division of Medical Oncology and Hematology, Princess Margaret Cancer Centre, 610 University Avenue, Toronto, ON, M5G 2M9, Canada (e-mail: ian.tannock@uhn.ca).

Background

“Waterfall plots” are used to describe changes in tumor size observed in clinical studies. Here we assess criteria for generation of waterfall plots and the impact of measurement error in generating them.

Methods

We reviewed published waterfall plots to investigate variability in criteria used to define them. We then compared waterfall plots generated by different observers for 24 patients enrolled in a completed phase I study of solid tumors with available computed tomography (CT) scans. Tumor measurements were made independently from CT scans according to Response Evaluation Criteria in Solid Tumors 1.1 by four board-certified radiologists and four medical oncologists. Interobserver variability was quantified and compared with reference measurements reported for the phase I study. All statistical tests were two-sided.

Results

There was substantial variability in criteria used to generate published waterfall plots. In the internal study, the results were statistically significantly different between all eight readers ($P = .01$, variance = 197.1, SD = 14.0) and between the oncologists ($P = .01$, variance = 319.0, SD = 17.9), but not between the radiologists ($P = .68$, variance = 70.8, SD = 8.4). Different observers classified one to five patients as having a partial response and 12–19 patients as having stable disease. Similar variability in categorization of response was observed when these error rates were applied to published waterfall plots.

Conclusion

Waterfall plots are subject to substantial variability in criteria used to define them and are influenced by measurement errors; they should be generated by trained radiologists. Caution should be exercised when interpreting results of waterfall plots in the context of clinical trials.


Tumor response in contemporary clinical trials is usually analyzed by using Response Evaluation Criteria in Solid Tumors 1.1 (RECIST 1.1) (1). These criteria allow researchers to classify tumor response into defined categories, but report only some of the available data. Waterfall plots have become increasingly popular in depicting tumor response in modern clinical trials because of their simple and intuitive display of results for individual patients. A waterfall plot is an ordered histogram where each patient is represented by a single vertical bar that can be further subgrouped by symbols or color for visualization. It displays the magnitude of individual patients’ responses to treatment as a percent change in measurements of their tumor(s) from baseline and treats response as a continuous rather than a categorical variable. Negative and positive bars represent some degree of tumor shrinkage and growth, respectively, and the number of positive and negative bars gives a visual impression of the overall proportion of patients with tumor growth or shrinkage. For comparative trials, the position of the inflexion point (to the right of which bars start to become negative) can be a further measure of relative treatment efficacy.

Errors in measurement are associated with every test, and estimates of tumor size are subject to considerable error and bias (2,3). Such errors can multiply as two or more measurements are compared to evaluate response. Measurement error can influence the classification of patients into response categories defined by RECIST 1.1, especially in borderline cases (eg, change from partial response to stable disease).

There has been little recognition of measurement error associated with waterfall plots, despite every pair of measurements being subject to error (4–6). Bias may also occur if measurements are made with knowledge of prior treatment, generally in the direction of showing greater tumor shrinkage with new treatments. Other biases may arise from variability in methods used to generate waterfall plots, such as defining changes in tumor dimension at a fixed time vs those defining maximum shrinkage, how nonmeasurable and new lesions are
recorded and interpreted, as well as the impact of a shrinking denominator when there are missing data because of unavailable postbaseline scans.

Here we review criteria used to generate published waterfall plots. We also compare waterfall plots generated by measurements of different observers on a single data set to demonstrate the degree of variability in the results. Finally, we apply levels of measurement error determined from our internal study as well as simulations to published waterfall plots in order to assess their stability and reliability. Our hypothesis was that waterfall plots would demonstrate substantial variability, leading to differences in classification of patients among categories of tumor response.

Methods

Review of Published Waterfall Plots
We reviewed published waterfall plots to determine and compare the criteria used to generate them. Reports of clinical trials with waterfall plots published in the *Journal of Clinical Oncology* between 2008 and 2012 were identified manually. We determined the number of patients available for serial measurement and causes of drop-out for each waterfall plot, how response was measured (best response vs changes at a given time), the type of therapeutic agent, the total sample size, phase and type of study (solid tumor vs hematological tumor), whether measurable disease was required in the entry criteria, and how changes in nonmeasurable lesions were presented in patients with both measurable and nonmeasurable metastases.

Internal study
Twenty-four patients with metastatic solid tumors who participated in a phase I trial of a novel therapy (standard dose doxorubicin plus escalating doses of intravenous pantoprazole; trial identifier: NCT 01163903) at Princess Margaret Cancer Centre formed the study group. The present study and the Phase I trial were approved by the institutional Research Ethics Board at Princess Margaret Cancer Centre, and patients signed informed consent for the Phase I trial. All patients had thoracic and abdominal/pelvic CT scans performed at enrollment as well as at least one follow-up assessment. All patients included in the study had measurable disease according to RECIST 1.1 criteria (1).

CT images were made available electronically to four board-certified radiologists, each with at least 15 years experience, and four medical oncologists; the latter had completed recently their training and certification in medical oncology. Tumor measurements were performed independently using the baseline and postbaseline CT scans according to RECIST 1.1 criteria (unidimensional measurements of longest tumor diameters, short axis for lymph node measurements, up to five target lesions, and maximum of two lesions per organ) (1). The oncologists were provided with the preselected image numbers of target lesions, while the radiologists were blinded to the image numbers and asked to select target lesions. Both groups were provided with the timing of the baseline and postbaseline scans.

The follow-up measurement used to generate the waterfall plots was the lowest value of the sum of the measured lesions during treatment; the percentage change in the sum compared with baseline was plotted for each patient. Waterfall plots were thus generated for each physician and compared with the reference waterfall plot. The reference waterfall plot was obtained from analysis of results of the previous clinical trial, generated by response measurements using RECIST 1.1 criteria by two radiologists (7).

Percent changes in tumor dimension were classified using RECIST 1.1 criteria (1), where partial response (PR) is defined as a 30% reduction in sum of linear dimensions of target lesions compared with baseline and the absence of new lesions. Progressive disease (PD) is defined as a 20% increase in the sum of linear dimensions of target lesions, and at least 5mm absolute increase in the sum, or the appearance of new lesion(s). When there is neither sufficient shrinkage nor increase for PR or PD, it is defined as stable disease. Complete response (CR) is defined as disappearance of all target lesions (1). The Spearman rank correlation coefficient was also calculated between eight observers for the ordering of individual patients.

Impact of Measurement Error on Published Waterfall Plots
We applied levels of measurement error derived from our internal study to assess its effect on stability of reported results. These errors were based on the standard deviation between the trained radiologists, who participated in our internal study. We illustrate here the effect of these errors on the plotted histograms and on

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>All studies (n = 108)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of agent, No. (%)</td>
<td></td>
</tr>
<tr>
<td>Chemotherapy</td>
<td>9 (8.3%)</td>
</tr>
<tr>
<td>Targeted agent</td>
<td>78 (72.2%)</td>
</tr>
<tr>
<td>Combined treatment</td>
<td>14 (13.0%)</td>
</tr>
<tr>
<td>Other</td>
<td>7 (6.5%)</td>
</tr>
<tr>
<td>Selection of patients, No. (%)</td>
<td></td>
</tr>
<tr>
<td>Measurable disease as inclusion criteria</td>
<td>77 (71.3%)</td>
</tr>
<tr>
<td>Inclusion criteria not described</td>
<td>31 (28.7%)</td>
</tr>
<tr>
<td>Timing of assessment, No. (%)</td>
<td></td>
</tr>
<tr>
<td>Best response</td>
<td>87 (80.6%)</td>
</tr>
<tr>
<td>Fixed time</td>
<td>5 (4.6%)</td>
</tr>
<tr>
<td>Both</td>
<td>4 (3.7%)</td>
</tr>
<tr>
<td>Not described</td>
<td>12 (11.1%)</td>
</tr>
<tr>
<td>Initial sample size, median (range)</td>
<td>53 (15–463)</td>
</tr>
<tr>
<td>Final number due to shrinking</td>
<td>46 (15–375)</td>
</tr>
<tr>
<td>Number of drop-outs, median (range)</td>
<td>4 (0–119)</td>
</tr>
<tr>
<td>% of studies without drop-outs, No. (%)</td>
<td>12 (11.1%)</td>
</tr>
<tr>
<td>Causes of drop-out, No. (%)</td>
<td></td>
</tr>
<tr>
<td>Not described</td>
<td>16 (16.7%)</td>
</tr>
<tr>
<td>Not evaluable (NE) for response</td>
<td>80 (73.3%)</td>
</tr>
<tr>
<td>Reason non-evaluable (n = 80), No. (%)</td>
<td></td>
</tr>
<tr>
<td>Missing postbaseline scans, early discontinuation of therapy, death, toxicity, surgery, did not complete treatment, withdrew consent, lost to follow up, clinical deterioration, incorrect tissue diagnosis</td>
<td>54 (67.5%)</td>
</tr>
<tr>
<td>Unspecified</td>
<td>17 (21.3%)</td>
</tr>
<tr>
<td>Nonmeasurable disease</td>
<td>6 (7.5%)</td>
</tr>
<tr>
<td>New lesions</td>
<td>3 (3.8%)</td>
</tr>
</tbody>
</table>

Table 1. Characteristics of published waterfall plots
classification of patients into response categories, using two representative published waterfall plots. Simulations based on these published waterfall plots were performed to assess the effect on overall response rate. We repeatedly simulated the results of the two trials from a normal distribution 1000 times, utilizing the assessed variability from our trial for each individual.

Statistical Analysis
Variations in interobserver measurements were estimated through the SAS mixed model, the Spearman rank correlation coefficient, and the simulations. Variances and standard deviations were determined using the best measurements as outcome between all readers, as well as amongst four radiologists and four oncologists. We applied the respective measurement errors obtained from the four radiologists as +/- 16.8% to provide 95% confidence intervals around two published waterfall plots. All statistical tests were two-sided, and a P value of less than .05 was considered statistically significant.

Results
Published Waterfall Plots
We identified 108 reports of clinical trials with waterfall plots published in the Journal of Clinical Oncology in the five-year period between 2008 and 2012 (Supplementary Material, available online). The majority of the waterfall plots were generated for Phase 1 and 2 trials evaluating targeted therapy for solid tumors and used measurements of best response on and/or off treatment to generate them (Table 1). Seventy-seven of 108 trials required the presence of measurable disease as one of the entry criteria. The median number of patients included in waterfall plots was 53 (range = 15–463). The number of drop-outs ranged from 0 to 119 patients. The causes of drop-out leading to participants not evaluable (NE) for response varied widely, including unspecified, early discontinuation of therapy, missing postbaseline scans, etc. (Table 1). The majority of patients who dropped out from the studies were excluded from the waterfall plots.

Figure 1. Waterfall plots generated by each of the eight readers. Individual patients are color coded to allow comparison between readers. Each color bar represents one patient, and the dotted lines indicate a 30% reduction and 20% increase from baseline, which are the cutoff points that determine partial response and progressive disease, respectively. MO = medical oncologist; R = radiologist.
Waterfall Plots for the Internal Clinical Study

Waterfall plots generated by each of the radiologists and medical oncologists are shown in Figure 1: These plots demonstrate variability in response classification, number of patients included, and the “ordering” of individual patients.

Figure 2A shows the reference waterfall plot generated by two radiologists for the study. The waterfall plot produced using the mean of the measurements of all readers is shown in Figure 2B: The results were statistically significantly different between all eight readers ($P = .01$), with variance of 197 (SD = 14). There was no statistically significant difference between four radiologists ($P = .68$), and the variance between radiologists was 70.8 (SD = 8.4) (Figure 2C). There was a statistically significant difference amongst four oncologists ($P = .01$), and the variance between oncologists was 319 (SD = 17.9) (Figure 2D).

The Spearman rank correlation coefficient was calculated between eight observers, and it varied from 0.51 to 0.94 for the ordering of individual patients. Therefore, there were marked differences in the “ordering” of individual patients for each waterfall plot. For example, the first bar on the left of each waterfall plot in Figure 1, indicating the largest increase in the sum of tumor diameters, was represented by several different individual patients.

The number of patients classified as responders by RECIST varies from one to five among all readers. Two of four radiologists classified one instead of two patients as PR when compared with baseline. The remaining two radiologists classified the same two patients as having a PR as the reference radiologists. The number of PRs ranged from one to five for the four oncologists (Figure 1). There was also marked variability in the classification of PD as best response ranging from two to eight amongst all readers when compared with the 10 PDs indicated by the reference radiologists, but these included new lesions and unequivocal increase in nonmeasurable lesions, which were not taken into consideration by our eight test readers. The number of patients with stable disease showed similar variability, as it consists of individual patients who do not have sufficient change in response to be classified as PD or PR.

The number of patients included in the waterfall plots by the eight readers varied from 22 to 24. The reference radiologists for the trial did not identify a target lesion in one patient (this was known to the oncologists), but all four radiologists regarded the patient as having target lesion(s) that rendered him eligible for inclusion. Another patient had five baseline lesions identified, but only three of them were measured, while the other two lesions were classified as nonevaluable at follow-up by the reference radiologists and the four oncologists. However, all four radiologists were able to select various target lesions that were evaluable at follow-up for the same patient. Therefore, this patient was included in the waterfall plots only by the four radiologists.

Influence of Measurement Error on Published Waterfall Plots

We applied standard measurement errors estimated from the radiologists in the internal study to published waterfall plots, as described in the Methods section. We showed that this would lead to substantial changes in the classification of response categories in many of the published studies, and we selected two published studies to illustrate this effect (8,9). When applying the 95% confidence intervals for measuring errors, the results of all eight readers were shown to be statistically significantly different from the reference radiologists (8,9). The dotted lines in the published waterfall plots indicate 30% reduction and 20% increase from baseline, which are the cutoff points that determine partial response and progressive disease, respectively.
confidence interval of +/-16.8% among the radiologist readers to the published waterfall plots, the number of responders changed substantially, as compared with that reported. In one study with five patients reported to have PRs, applying confidence intervals (+/- 16.8%) would lead to between two and 12 patients satisfying response criteria (Figure 3A). In the second study with 12 patients reported to have PRs or CRs, applying confidence intervals (+/- 16.8%) would lead to between seven and 15 patients satisfying response criteria (Figure 3B). There was similar variability in those assigned to the SD or PD categories.

Simulations based on the two representative published waterfall plots were performed. The mean response rates were 17.9% (STD = 3%) and 45.7% (STD = 4%). The median response rates were 17.7% (range = 9.7%-29%) and 44% (range = 32%-60%). The original response rates were 16% and 48%, respectively.

**Discussion**

Our review indicated substantial variability in the reporting of published waterfall plots. Subtle differences in criteria used to generate them, a shrinking denominator as patients come off trial, and unrecognized measurement error all impact the waterfall plots and the rates of tumor response that are derived from them.

The present study evaluated interobserver variability in generating waterfall plots from a completed internal study and found considerable variability that leads to different classifications of patients.

**Figure 3.** Published waterfall plots from (A) Chan et al (7) and (B) Wong et al (8) (reproduced with permission) with mean 95% confidence intervals generated by the four radiologists in the internal study. Note that in (A), the number of partial responses (PRs) varied from published level of five to between two and 12. In (B), the number of PRs/complete responses (CRs) varied from published level of 12 to between seven and 15. There is corresponding variability in the other categories of progressive disease and stable disease. The dotted lines indicate a 30% reduction and 20% increase from baseline, which are the cutoff points that determine partial response and progressive disease, respectively. Reprinted with permission, 2012 American Society of Clinical Oncology. All rights reserved.
in a clinical trial according to categories of partial response, stable disease, and progression. This variability was less (but still present) when the waterfall plots were generated by radiologists, as compared with those generated by medical oncologists.

Previous studies have shown high interobserver variability in determining the proportion of patients with tumor response for modified Response Evaluation Criteria in Solid Tumors (RECIST) measurements (10–12). Oxnard et al. reported that increases and decreases less than 10% might be indistinguishable from variability-related changes on repeat CT scans of lung tumors (13). Also, RECIST may be suboptimal for evaluating outcomes after treatment with targeted agents (14,15), although 85% of the studies that we reviewed used waterfall plots to characterize tumor shrinkage after treatment with targeted agents alone or in combination.

In the present study, interobserver misclassification was evident because two of four radiologists classified one instead of two patients as PRs as compared with the reference waterfall plot. The number of patients thought to satisfy RECIST criteria for PR ranged from one to five for the oncologists and one to two for the radiologists (Figure 2). The misclassification of progression varied from two to eight patients for all readers when compared with 10 patients indicated by the reference radiologists for the trial, taking into consideration new lesions and unequivocal progression in nonmeasurable lesions. The classification of stable disease showed similar variability ranging from 12–19 among all readers. This raises concerns as patients can be under- or overtreated based on their misclassification in a clinical trial. The larger variance amongst oncologists compared with radiologists is almost certainly because of radiologists having more training and experience in the interpretation of CT scans. The four radiologists each had at least 15 years of experience, while the four oncologists had recently completed their training and certification. Also, some readers reported measurements for 22 patients, while others reported measurements for all 24 patients, but this was consistent among the four radiologists. Waterfall plots based on imaging should be generated by trained radiologists.

When confidence intervals on measurements derived from our internal study were applied to published waterfall plots, there was substantial uncertainty in classification of patients into response categories, even if limited to the lower rates of variability among trained radiologists (Figure 3). The mean and median of simulated response rates were similar to the original response rates. However, even though the overall response rates were similar, the range of response rates demonstrated substantial variability. Extreme cases can happen, especially when one observer tends to overestimate or underestimate because of individual bias. Also, clinical decision-making is not based on averages but on data for individual patients. Therefore, a decision for an individual patient as to whether to stay on the study or stop treatment will vary substantially, despite similar overall response rates.

There is a tendency to believe in the rigor of waterfall plots despite factors, evaluated here, that lead to their variability. Ratain et al. (16) first introduced the useful concept of a waterfall plot; however, although the waterfall plot is used descriptively in that article, waterfall plots have since been used in different ways to make treatment decisions. It is important for clinicians to recognize interobserver variability in evaluation of tumor response depicted using waterfall plots, as measurement error can influence substantially the results of clinical trials and clinical management of patients. Options may exist for patients who are not responding to current therapy, and it is important to distinguish between small but clinically reliable changes in tumor size vs measurement noise. In our opinion, trials should require a confirmation of PR and rigorous trials appropriately require a second measurement satisfying RECIST 1.1 criteria for determination of PR. Waterfall plots rely on measurements of maximum change, and using them alone to classify patients may result in overclassification of PR, as they do not require confirmation.

There are several limitations with our study. First, the sample size of the internal study of 24 patients is relatively small. Also, the radiologists were aware of the purpose of the study, and they may be more likely to be accurate with their measurements than in usual assessment. Furthermore, the medical oncologists were not blinded to target lesions, and nonmeasurable lesions or new lesions were not recorded by the eight test readers.

In conclusion, the present study has investigated interobserver variability and stability of waterfall plots using RECIST 1.1 criteria. Our findings suggest that waterfall plots demonstrate substantial variability. Caution should be exercised when interpreting results of waterfall plots in the context of clinical trials.

References
14. Choi H, Charmsangavej C, de Castro Faria S, et al. CT evaluation of the response of gastrointestinal stromal tumors after imatinib mesylate...


**Funding**

This study was funded by grant KG100252 from the Susan G Komen Foundation.

**Affiliations of authors:** Divisions of Medical Oncology and Hematology (TS, AJT, RWJ, FVB, MGM, IFT) and Biostatistics (LW), Princess Margaret Cancer Center, Department of Medicine, University of Toronto, Toronto, Canada; Joint Department of Medical Imaging, University Health Network, Toronto, Canada (MM, TKK, MS, HS).