Is the Median PFS Difference Meaningful for Assessing Treatment Effect?

Steven Snapinn
Amgen, One Amgen Center Drive, Thousand Oaks, CA 91320

Abstract

The magnitude of a treatment's effect on PFS is a key consideration for assessing the clinical meaningfulness of that effect. Various metrics can be used for this assessment, such as the hazard ratio or the absolute risk reduction. One metric in common use, particularly among clinicians, is the difference between groups with respect to median PFS. In this presentation I will discuss some limitations of this metric, including the fact that the difference in medians cannot be interpreted as the median causal effect of the treatment. I will also contrast the difference in medians with the absolute risk reduction and show that, for a given hazard ratio, they are in conflict with each other; based on that, I will argue that the hazard ratio is the most appropriate metric for assessing clinical meaningfulness.

Key Words: Absolute risk reduction; Clinical meaningfulness; Hazard ratio; Overall survival; Progression-free survival

1. Introduction

In clinical trials, the effects of experimental treatments are measured on a variety of scales. For a time-to-event endpoint, such as progression-free survival (PFS) or overall survival (OS), effects can be measured on both a relative scale and an absolute scale. The most common relative scale is the hazard ratio, often calculated using a Cox proportional hazards model. In oncology trials, the most common absolute scale is the difference in medians calculated from the Kaplan-Meier curves. In this paper I will argue that assessing the treatment affect using the difference in medians is seriously flawed, for three reasons: 1) The difference in medians is not an average causal effect. 2) The difference in medians is measured with poor precision. 3) For a given hazard ratio, the difference in medians is in direct contradiction to another absolute measure of effect in widespread use, the absolute risk reduction.

As examples of the use of the difference in medians, consider the Kaplan-Meier curves in Figures 1 and 2. Figure 1 represents the hypothetical results of a clinical trial in which the difference in medians is approximately 1 month, and Figure 2 represents the actual results of a published clinical trial (Van Cutsem et al 2007) in which the difference in medians is approximately 1 week. While this measure of the effect size is approximately 4 times greater in Figure 1 than in Figure 2, visual examination of the curves gives a far different impression. In Figure 1, the Kaplan-Meier curves are virtually overlapping, suggesting little, if any, treatment effect; the 1 month difference in medians is due to a barely

perceptible horizontal portion of the two curves in which one curve is slightly above 0.5 and the other is slightly below 0.5. In Figure 2, on the other hand, there appears to be a considerable treatment effect, and the hazard ratio of 0.54 seems to validate that perception; however, while the curves remain separate throughout, they come close to touching when both cross a rate of 0.5 on the y-axis.

Proponents of the use of the difference in medians must conclude, despite the visual evidence, that the treatment effect in Figure 1 is 4 times greater than that in Figure 2. If one uses the difference in medians in some cases when it is in agreement with the visual evidence, but not in cases like Figures 1 and 2, where there appears to be a contradiction, then one is basing the estimate of the treatment effect on judgment, not on a chosen metric. In the next section I will discuss how such contradictions can occur.

2. The Difference in Medians Is Not an Average Causal Effect

The goal of a randomized clinical trial is to learn about the causal effect of a treatment on a set of specified outcomes. Ideally, one would want to know the causal effect in each treated individual, in order to learn about the distribution of causal effects. In practice, however, the most one can typically learn about the distribution of causal effects is a summary statistic, such as the mean. This is due to the fact that measuring the causal effect in an individual involves an unobserved counterfactual. Specifically, for a treated patient, the causal effect is the difference between the observed outcome and the unobserved counterfactual outcome that would have been observed had the patient not been treated. While one might have a reasonable guess at the counterfactual, it is never known with certainty, and, therefore, the causal effect is never observed for an individual patient.

Consider the results in Table 1 for 10 hypothetical subjects. The table contains the results that would be obtained for these 10 subjects under two scenarios: if taking the control treatment and if taking the experimental drug. The outcome measure is a continuous variable, and smaller values represent better outcomes. This table represents an ideal situation, since one can never know both potential outcomes for any individual subject, but it serves to illustrate the problem with the difference in medians.

In this ideal situation one can calculate the causal effect of the treatment in each subject, and this is contained in the rightmost column of Table 1. Of the 10 subjects, 8 experienced a reduction of 5 units, while 2 experienced no change. The mean causal effect is, therefore, 4 units, and the median causal effect is 5 units. This represents the information that one would want to learn from a clinical trial.

Table 1

Subject #	Outcome w/ Control	Outcome w/Drug	Causal Effect
1	5	0	5
2	5	0	5
3	5	0	5
4	5	0	5
5	5	5	0
6	5	5	0
7	10	5	5
8	10	5	5
9	10	5	5
10	10	5	5

Now consider Table 2, in which the same 10 subjects have been assigned to the two treatments in a balanced way (as indicated in the column labeled “Group”). The shaded cells represent unobserved information. For example, Subject 1 is assigned to the Drug group; therefore, “Outcome w/Control” is unobserved for that subject. Since only one outcome is observed for any individual subject, the causal effect is unobserved for all subjects.

Without the information on individual causal effects, it is not possible to fully determine the distribution of causal effects. However, it is possible to learn something about that distribution through the distributions of outcomes for each of the two treatments. For example, the mean of the outcomes with the control is 7 units, and the mean of the outcomes with the drug is 3 units. Taking the difference, we obtain a mean causal effect of 4 units, in agreement with the mean calculated from the rightmost column, containing the individual causal effects.

Now consider a similar calculation of the difference in medians. The median outcome in the control group is 5 units, and the median outcome in the drug group is also 5 units. Taking the difference, therefore, one would calculate a difference in medians of 0 units, which is in disagreement with the median causal effect of 5 units calculated from the rightmost column. This example illustrates a key flaw in the calculation of the difference in medians: it is not, in general, an average causal effect.

To summarize, what we want to know is the distribution of differences between outcomes in individuals (i.e., the distribution of causal effects). However, that is not possible, since that would involve knowledge of the unobserved counterfactual. What we are able to calculate the separate distributions of outcomes with the two treatments, and from this we can calculate the difference in means. Since the difference in means is equal to the mean of the differences, this is a valid approach to estimate the mean causal effect. However, the difference in medians is not, in general, the same as the median of the differences; therefore, the difference in medians is not, in general, equal to the median causal effect.

Table 2

Subject #	Group	Outcome w/ Control	Outcome w/ Drug	Causal Effect
1	Drug	5	0	5
2	Control	5	0	5
3	Drug	5	0	5
4	Control	5	0	5
5	Drug	5	5	0
6	Control	5	5	0
7	Drug	10	5	5
8	Control	10	5	5
9	Drug	10	5	5
10	Control	10	5	5

Note the use of the words “in general” in the previous paragraph. It is true that under certain distributional assumptions the difference in medians will be equal to the median causal effect. However, without making distributional assumptions, it is not clear what relationship the difference in medians has to the distribution of causal effects, other than the following: the difference in medians must lie somewhere between the smallest possible causal effect and the largest possible causal effect.

It is this issue that can lead to the visual inconsistencies illustrated by Figures 1 and 2. In Figure 1, the difference in medians of 1 month is most likely an outlying causal effect from a distribution that is tightly packed around 0. In Figure 2, the difference in medians of 1 week is most likely at the very low end of the distribution of causal effects.

3. The Difference in Medians Is Measured With Poor Precision

Estimation of the median for a specific treatment group is straightforward: it is the point in time at which the Kaplan-Meier curve crosses 0.5 on the y-axis. Calculation of the confidence interval for median is somewhat more complicated, and involves the well-know Greenwood estimate of the variance of the Kaplan-Meier curve:

$$\widehat{\mathrm{Var}}\left[\widehat{S}(t)\right] = \widehat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

In this equation, S represents the Kaplan-Meier estimate as a function of time, and n_i and d_i represent the number of subjects at risk and the number of events, respectively, at event time t_i. Using this estimate of the variance, one can put a confidence interval around the Kaplan-Meier estimate, and the corresponding confidence bounds for the median are the points in time at which the upper and lower confidence bounds for the Kaplan-Meier curve cross 0.5 on the y-axis.

Clearly, a small study with a small number of events will result in poor precision for any measure of treatment effect. In fact, the variance of the estimate of the log hazard ratio is closely approximated by $4/d$, where d is the total number of events in the two treatment groups. For the estimate of the median, however, there are two additional factors that play a large role in determining precision. The first is sample size, most importantly the size of the risk set in the vicinity of the median. This is apparent from the Greenwood formula. In other words, the larger the fraction of subjects censored prior to reaching the median the worse the precision. The second factor is the slope of the Kaplan-Meier curves in the vicinity of the median. If the curve is relatively flat (as in the 1-month flat portion of the curves in Figure 1) then the precision will be worse than if the curve is descending rapidly (as in Figure 2). Various authors have proposed approaches for calculating a confidence interval for the difference between the estimated medians of the two treatment groups (Karrison 2007), but clearly the precision of the estimate of the difference will depend on the same factors as the precision of the estimate of the median for an individual treatment group.

For the reasons described above, there are many cases where the precision of the estimate of the hazard ratio is far better than the precision of the estimate of the difference in medians. It is therefore counterintuitive that the hazard ratio is nearly always presented along with a confidence interval, while the difference in medians is often presented without one. A hazard ratio with a very small p-value and a narrow confidence interval that does not cross unity will often be accompanied by a difference in medians that has a confidence interval extending well beyond zero, but if the confidence interval is not calculated then the researcher might not be aware of this issue.

4. The Difference in Medians Is in Direct Contradiction to the Absolute Risk Reduction

To be approved and used by clinicians, a new drug must have an effect that is not only statistically significant, but also clinically meaningful. Some researchers believe that a relative measure of the treatment effect, such as the hazard ratio, can overstate the magnitude of benefit, and therefore prefer absolute measures, such as the difference in medians, for assessing clinical meaningfulness. However, as discussed by Snapinn and Jiang (2010), it is not well understood that, for a given hazard ratio, the difference in medians is in direct contradiction to another absolute measure in common use, the absolute risk reduction.

Consider the results of two hypothetical trials in Figure 3. In each trial the survival distributions are exponential, and in both trials the hazard ratio for the treatment group (the dashed line) is 0.6 relative to the control (the solid line). The only difference between trials is that the hazard rates are relatively low in trial A and relatively high in trial B. Considering the difference in medians as the measure of treatment effect (i.e., the horizontal difference between curves at 0.5 on the y-axis), the effect size in trial A (5 years) is considerably larger than in trial B (9 months). However, when considering the absolute risk reduction at 1 year as the measure of treatment effect (i.e., the vertical difference between curves at 1 year on the x-axis), the effect size in trial B (14.9%) is considerably larger than in trial A (3.7%). In general, for a given hazard ratio, the study with lower hazard rates will have a better treatment effect when measured by the difference in medians, while the study with higher hazard rates will have a better treatment effect when measured by the absolute risk reduction.

The fact that proponents of the difference in medians and proponents of the absolute risk reduction come to opposite conclusions regarding clinical meaningfulness calls into question the value of both of these absolute measures of effect. It might be surprising to both proponents that cost-effectiveness, as measured by lifetime cost per life-year saved, can be shown to be a function solely of the cost (assumed to accrue uniformly over time) and the hazard ratio; i.e., for a given hazard ratio, cost-effectiveness is unrelated to either the difference in medians or to the absolute risk reduction. This might suggest that the hazard ratio is the most appropriate measure of clinical meaningfulness as well. But it is certain that if the difference in medians is a good measure of clinical meaningfulness, then the absolute risk difference must be an extremely poor measure, and *vice versa*.

5. Summary

In this paper I have argued that the difference in medians is a poor metric for assessing the clinical meaningfulness of a treatment's effect. Examples showing dramatic differences between the calculation of the difference in medians and the overall visual impression of the Kaplan-Meier curves can be explained by the fact that the difference in medians does not have a clear relationship to the distribution of causal effects. This issue alone should be a strong reason to avoid this metric. In addition, the difference in medians is often measured with extremely poor precision. In fact, if the confidence for the difference in medians were routinely provided, making the poor precision obvious to all, investigators might quickly abandon this metric. Finally, for a given hazard ratio, use of the difference in medians as a measure of clinical meaningfulness leads to conflicting conclusions relative to use of the absolute risk reduction. For all these reasons, relative measures of the treatment effect such as the hazard ratio are preferable to the difference in medians for assessing clinical meaningfulness.

References

Karrison, T. Comparison of Median Survival Times With Adjustment for Covariates. *Statistics in Medicine* **23**:2537-2553, 2007.

Snapinn, S., Jiang, Q. On the Clinical Meaningfulness of a Treatment's Effect on a Time-to-Event Variable. *Statistics in Medicine* **30**:2341-2348, 2011.

Van Cutsem, E., Peeters, M., Siena, S., Humblet, Y., Hendlisz, A., Neyns, B., Canon, J.-L., Van Laethem, J.-L., Maurel, J., Richardson, G., Wolf, M., Amado, R. G. Open-Label Phase III Trial of Panitumumab Plus Best Supportive Care Compared With Best Supportive Care Alone in Patients With Chemotherapy-Refractory Metastatic Colorectal Cancer. *Journal of Clinical Oncology* **25**:1654-1664, 2007.

Figure 1

A Pair of Kaplan-Meier Curves for Two Treatments With a 1-Month Difference in Medians

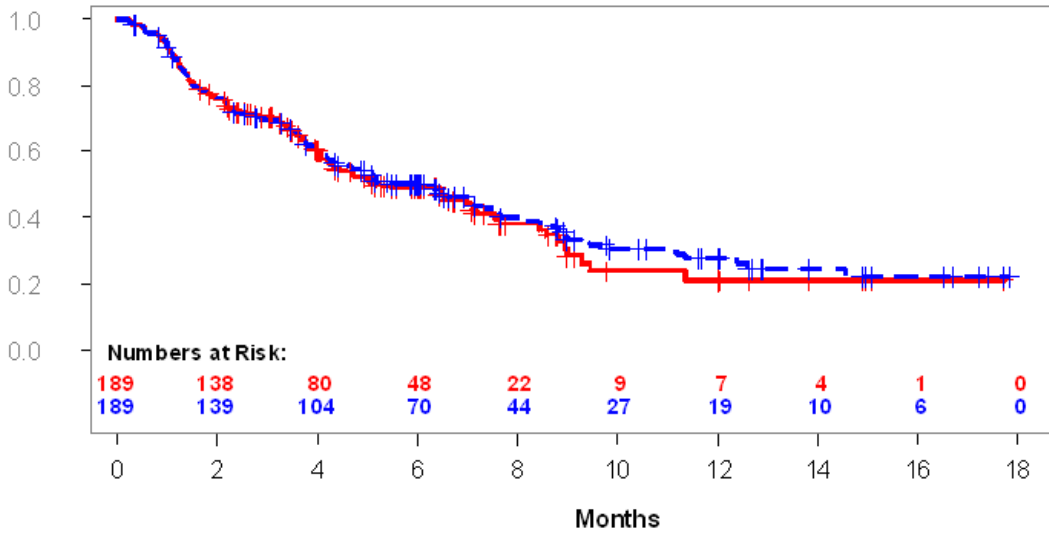
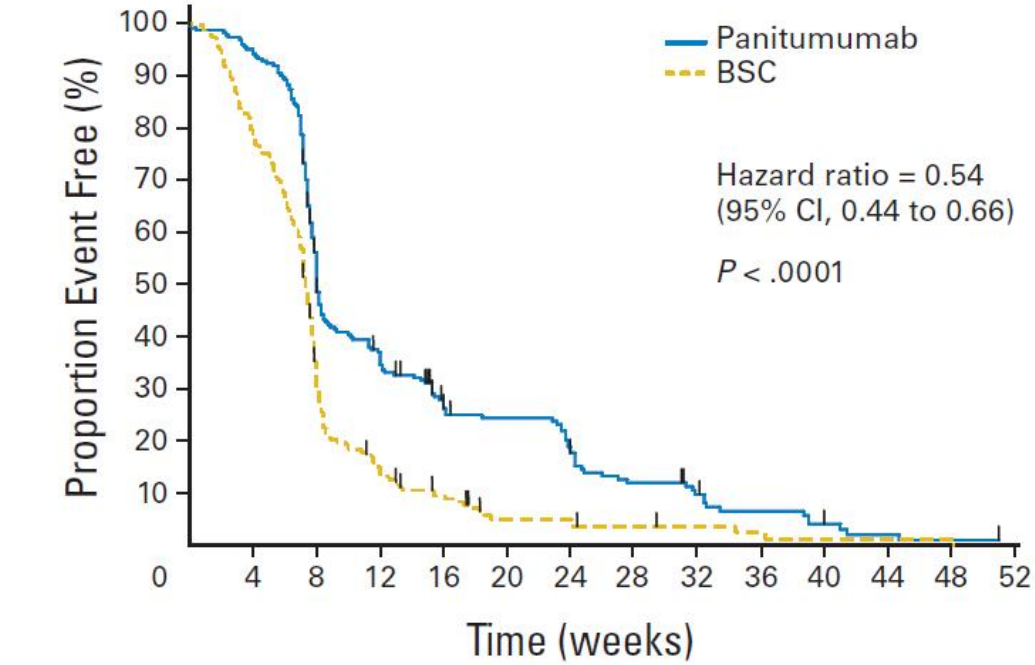


Figure 2

A Pair of Kaplan-Meier Curves for Two Treatments With a 1-Week Difference in Medians



Patients at risk

Panitumumab	231	209	118	76	49	40	31	19	13	8	5	2	1
BSC	232	175	75	31	17	7	7	4	3	2	1	1	1

Figure 3

Hypothetical Kaplan-Meier Curves for Two Clinical Trials Each With Exponential Survival Distributions and Hazard Ratio = 0.6