Research Article

Statistics in Medicine

Received 8 August 2010, Accepted 7 March 2011 Published online 26 April 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4256

On the clinical meaningfulness of a treatment's effect on a time-to-event variable

Steven Snapinn*† and Qi Jiang

The standard analysis of a time-to-event variable often involves the calculation of a hazard ratio based on a survival model such as Cox regression; however, many people consider such relative measures of effect to be poor expressions of clinical meaningfulness. Two absolute measures of effect are often used to assess clinical meaningfulness: (1) many disease areas frequently use the absolute difference in event rates (or its inverse, the number-needed-to-treat) and (2) oncology frequently uses the difference between the median survival times in the two groups. While both of these measures appear reasonable, they directly contradict each other. This paper describes the basic mathematics leading to the two measures and shows examples. The contradiction described here raises questions about the concept of clinical meaningfulness. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: time-to-event; hazard ratio; event rates; median survival times; clinical meaningfulness

Introduction

To be approved and used by clinicians, a new drug must not only be effective, but the magnitude of its effect must be clinically meaningful. Unfortunately, the concept of clinical meaningfulness (also referred to as clinical significance) has not been clearly defined. This is true in general, but there is particular lack of clarity in the setting of a time-to-event variable, such as overall survival. In fact, two fundamental approaches to assessing clinical meaningfulness are used in this setting, and it is not widely appreciated that these two approaches directly contradict each other. This paper discusses the concept of clinical meaningfulness in the context of a time-to-event variable and the inherent contradiction in the standard approaches.

A time-to-event variable is used as the primary efficacy outcome in many areas of clinical research. For example, oncology treatments are often evaluated on the basis of time to death, or time to progression or death (known as 'progression-free survival'); cardiovascular treatments are often evaluated on the basis of time to a major cardiovascular event; and osteoporosis drugs are often evaluated on the basis of time to fracture. The typical analysis of this type of variable involves a statistical model such as the Cox proportional hazards model, and the calculation of a hazard ratio.

The first step in assessing the efficacy of the treatment is to determine whether the treatment and control are significantly different (i.e. whether there is a sufficient evidence to conclude that the treatment has any effect at all). While some people may question the value of hypothesis testing (although not many would go as far as Rozeboom, as reported by Thompson [1]: '*Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students...*'), there is a little debate that hypothesis testing in and of itself is not sufficient. Nearly everyone would agree that the data must provide evidence that the magnitude of the treatment's effect is clinically meaningful.

Amgen Inc., One Amgen Center Drive, 24-2-C, Thousand Oaks, CA 91320, U.S.A.
**Correspondence to: Steven Snapinn, Amgen Inc., One Amgen Center Drive, 24-2-C, Thousand Oaks, CA 91320, U.S.A.*
†*E-mail: ssnapinn@amgen.com*

Unfortunately, the term 'clinical meaningfulness' (or 'clinical significance') has no standard definition, although some authors have attempted to define it. For example, Greenstein [2] wrote '*clinical significance denotes a change that may alter how a clinician will treat a patient, and this value judgment varies depending on the situation*.' Kraemer *et al.* [3] specifically addressed clinical significance in the context of treatments for psychiatric conditions: '*The clinical significance of a treatment is based on external standards provided by clinicians, patients, and/or researchers. Unfortunately, to date there is little consensus about the criteria for these efficacy standards*.'

Many people strongly assert that measures of treatment effect like the one typical in a time-to-event trial, the hazard ratio, are inappropriate measures of clinical meaningfulness. For example, Kraemer *et al.* [3] wrote: '*…relative risk reduction measures … can give an inflated impression of the size of the effect, thus exaggerating clinical significance*' Replogle and Johnson [4] had similar comments: '[*These measures of effect*] *reflect relative disparities and are perhaps less useful in clinical practice than measures of absolute benefit or harm*.'

The absolute measures most frequently recommended are the absolute risk reduction (ARR) and the number-needed-to-treat (NNT). The NNT is simply the inverse of the ARR; hence, while clinical meaningfulness increases as ARR increases and as NNT decreases, the two measures are mathematically equivalent. According to Leung [5], '*NNT has been highlighted as a meaningful measure of clinical significance*.' Kraemer *et al.* [3] wrote '*For this reason, the preferred effect size at this time would tend to be … NNT or risk difference….*' Replogle and Johnson [4] wrote '*Absolute risk reduction accounts for the baseline control group event rate and is a more realistic quantification of treatment effect than relative measures. Number needed to treat (NNT) estimates the therapeutic effort needed to prevent one additional adverse event. NNT incorporates both relative risk reduction and the event rate without treatment…. Thus, NNT has more-obvious implications for clinical decision making than risk estimates expressed in relative terms*.'

Few authors have focused specifically on time-to-event outcomes. One paper that did, by Hildebrandt *et al.* [6], found that many authors apply NNT inappropriately; however, the clear consensus among many authors is that relative measures such as the hazard ratio are inappropriate to assess clinical meaningfulness, whereas absolute measures such as ARR or NNT are preferred. One instructive example in the literature is a report by Ridker *et al.* [7] of the JUPITER trial, a lipid-lowering trial that showed a highly significant benefit on cardiovascular endpoints. Since the trial specifically targeted a low-risk population, the event rates in both treatment arms were low, and, consequently, the ARR was low. As if to highlight the perceived small benefit, the authors present Kaplan-Meier curves that cover the entire scale from 0 to 100%, even though the curves themselves take up less than 10% of the area. In the paper, the authors comment that '*…our trial evaluated the use of rosuvastatin for the prevention of first cardiovascular events; therefore, the absolute event rates are lower than would be expected among patients with a history of vascular disease, a fact that should be taken into account in considering whether the use of statin therapy … would be cost effective if applied widely*.' (Note that we will take up the issue of cost-effectiveness later in this paper.)

While most authors focus on ARR and NNT as measures of absolute treatment benefit, the median time to event is widely used, particularly in trials of oncology treatments. When trials proceed long enough so that more than 50% of the patients experience the event of interest, the median time to event can be calculated. The difference between groups in the median time-to-event is another measure of an absolute treatment benefit. For example, Fleming *et al.* [8] wrote '*…a statistically significant 1 month improvement in PFS would provide much less persuasive evidence of a survival benefit than a statistically significant and clinically important 3- to 4-month improvement in PFS in a patient with 6- to 8-month median OS*.'

Interestingly, these two fundamentally different approaches to assessing the clinical meaningfulness of a treatment's effect, the ARR or NNT and the median difference, directly contradict to each other, and we are unaware of any publication that points this out. Here we describe this contradiction, first mathematically and then conceptually. We then discuss the implications of this contradiction, related both to the concept of clinical meaningfulness and to the concept of cost-effectiveness.

Opposing views of clinical meaningfulness: Mathematical results

Assume that a time-to-event variable has an exponential distribution with hazard rate γ in patients taking the standard treatment (or placebo). Further suppose that an experimental treatment reduces the event rate, with hazard ratio $0 < \lambda < 1$.

Table I. Event-free rates and event rates corresponding to cross-over time (time at which ARR begins to decrease).

λ	Event-free rates (standard, experimental)	Event rates (standard, experimental)
0.5	0.250, 0.500	0.750, 0.500
0.667	0.296, 0.444	0.704, 0.556
0.8	0.328, 0.410	0.672, 0.590
lim $\lambda \to 1$	0.368	0.632

Median event times

The median survival times (note that we use 'survival times' as a general term to refer to 'event-free times,' whether or not the outcome is survival) in the two groups are given by:

$$\text{standard treatment: } \frac{\ln(2)}{\gamma}$$

$$\text{experimental treatment: } \frac{\ln(2)}{\gamma\lambda}.$$

A measure of treatment benefit is the difference between these medians, or $(1-\lambda)\ln(2)/\gamma\ \lambda$. We are interested in how the hazard rate, γ, influences this benefit. Holding λ constant, the rate of change in the treatment benefit, as γ changes, is $-(1-\lambda)\ln(2)/\gamma^2\lambda$. This is negative regardless of the values of γ and λ. Therefore, as the hazard rate increases, the magnitude of the treatment benefit, as measured by the difference in medians, decreases.

Absolute event rates

As a function of time, t, the event-free rates are

$$\text{standard treatment: } e^{-\gamma t}$$

$$\text{experimental treatment: } e^{-\gamma\lambda t}$$

Therefore, the benefit of the experimental treatment, as a function of time, is the difference between these rates, or $e^{-\gamma\lambda t} - e^{-\gamma t}$. Holding λ and t constant, the rate of change in the treatment benefit, as γ changes, is $t(e^{-\gamma t} - \lambda e^{-\gamma\lambda t})$. This is positive for smaller values of t and negative for larger values of t; the time of cross-over from a positive to a negative slope, denoted $t*$, satisfies $e^{-\gamma t*} = \lambda e^{-\gamma\lambda t*}$, or the time at which the event-free rate in the control group is equal to the product of the event-free rate in the experimental group and the hazard ratio.

Solving for $t*$ we get $t* = \ln(\lambda)/\gamma(\lambda-1)$, which corresponds to the time at which the event-free rates in the two treatment groups are

$$\text{control treatment: } \lambda^{1/(1-\lambda)}$$

$$\text{experimental treatment: } \lambda^{\lambda/(1-\lambda)}.$$

As illustrated in Table I, regardless of the value of λ, the event-free rate of the control treatment at time $t*$ is always less than e^{-1} while the event-free rate of the experimental treatment is always greater than e^{-1}; both rates approach e^{-1} (or 0.368) as $\lambda \to 1$.

In other words, at any time at which less than 63% of the control group has experienced the event (which is the typical region of interest when assessing the effect of a new treatment), as the hazard rate increases the magnitude of the treatment benefit, as measured by the ARR, also increases.

These results are illustrated in Table II, which shows how ARR, NNT and the median difference change as the hazard rate changes, for fixed values of the hazard ratio.

Opposing views of clinical meaningfulness: hypothetical example

Note the contrast between the results for the two measures of treatment benefit: As the hazard rate increases, the treatment benefit as measured by the ARR at any fixed point in time *increases*

Table II. Effect of the hazard rate on ARR, NNT and the median difference for fixed hazard ratio at time $t=1$.

		Event-free rates				Medians		
Hazard ratio	Hazard rate	Control	Treatment	ARR	NNT	Control	Treatment	Difference
0.9	0.1	0.905	0.914	0.009	110.0	6.931	7.702	0.770
	0.3	0.741	0.763	0.023	44.3	2.310	2.567	0.257
	0.5	0.607	0.638	0.031	32.2	1.386	1.540	0.154
	0.7	0.497	0.533	0.036	27.8	0.990	1.100	0.110
	0.9	0.407	0.445	0.038	26.1	0.770	0.856	0.086
0.7	0.1	0.905	0.932	0.028	36.3	6.931	9.902	2.971
	0.3	0.741	0.811	0.070	14.3	2.310	3.301	0.990
	0.5	0.607	0.705	0.098	10.2	1.386	1.980	0.594
	0.7	0.497	0.613	0.116	8.6	0.990	1.415	0.424
	0.9	0.407	0.533	0.126	7.9	0.770	1.100	0.330
0.5	0.1	0.905	0.951	0.046	21.6	6.931	13.863	6.931
	0.3	0.741	0.861	0.120	8.3	2.310	4.621	2.310
	0.5	0.607	0.779	0.172	5.8	1.386	2.773	1.386
	0.7	0.497	0.705	0.208	4.8	0.990	1.980	0.990
	0.9	0.407	0.638	0.231	4.3	0.770	1.540	0.770

(assuming that the control event rate is relatively low), whereas the treatment benefit as measured by the difference in medians *decreases*. This leads to an inconsistency in conclusions drawn using the two measures. For any given hazard ratio, when the hazard rate is low the treatment benefit as measured by the ARR (or, equivalently, the NTT) is low, and the treatment benefit as measured by the difference in medians is high; when the hazard rate is relatively high the treatment benefit as measured by the ARR is high, and the treatment benefit as measured by the difference in medians is low.

Two opposing conclusions cannot simultaneously be correct, which raises important questions: Is there a 'correct' measure of treatment benefit, and, if so, which one? Since both measures of clinical meaningfulness appear reasonable (a large absolute benefit seems like a good thing, as does a large increase in median survival time), what is the meaning of 'clinical meaningfulness'?

Consider the following example, based on an assumption of exponential survival distributions. Suppose a cardiovascular clinical trial (labeled Trial A) found a statistically significant reduction in the rate of a clinical event from 9.5% over one year in the control group to 5.8% in the experimental group, with a hazard ratio of 0.6. One could imagine some concern that the magnitude of the treatment effect, an ARR of 3.7% or an NNT of approximately 27 patients, is not large enough to be clinically meaningful. Now suppose that a second trial (labeled Trial B) found a statistically significant reduction in the rate of the same clinical endpoint from 45.1% over one year in the control group to 30.2% in the experimental group, also with a hazard ratio of 0.6. In this case, the ARR is 14.9%, or an NNT of approximately 7. Physicians would be more likely to believe that the experimental treatment in Trial B provided a clinically meaningful benefit than the treatment in Trial A. If, however, one were to measure treatment effect by the difference in median time-to-event, the conclusion would be the opposite: In Trial A the corresponding median times are 6.9 years in the control group and 11.6 years in the experimental group, for an increase of nearly 5 years, whereas in Trial B the corresponding median times are 1.16 years in the control group and 1.93 years in the experimental group, for an increase of only about 9 months. There would almost certainly be more concern about the clinical meaningfulness of a 9-month increase in median survival than a 5-year increase. (Note that under the notation above, these examples used $t=1$, $\lambda=0.6$, and $\gamma=0.1$ for Trial A and $\gamma=0.6$ for Trial B.)

This discordance is illustrated in Figure 1, which shows the hypothetical results described above. The top panel simply shows the exponential survival curves for the two treatment arms in each of the two trials; the x-axis is time in years, and the y-axis is the proportion of patients who have experienced the event. The middle panel truncates the curves at 1 year, a time at which the event rates are typical of those obtained in trials of new cardiovascular treatments. The absolute treatment effects for the two

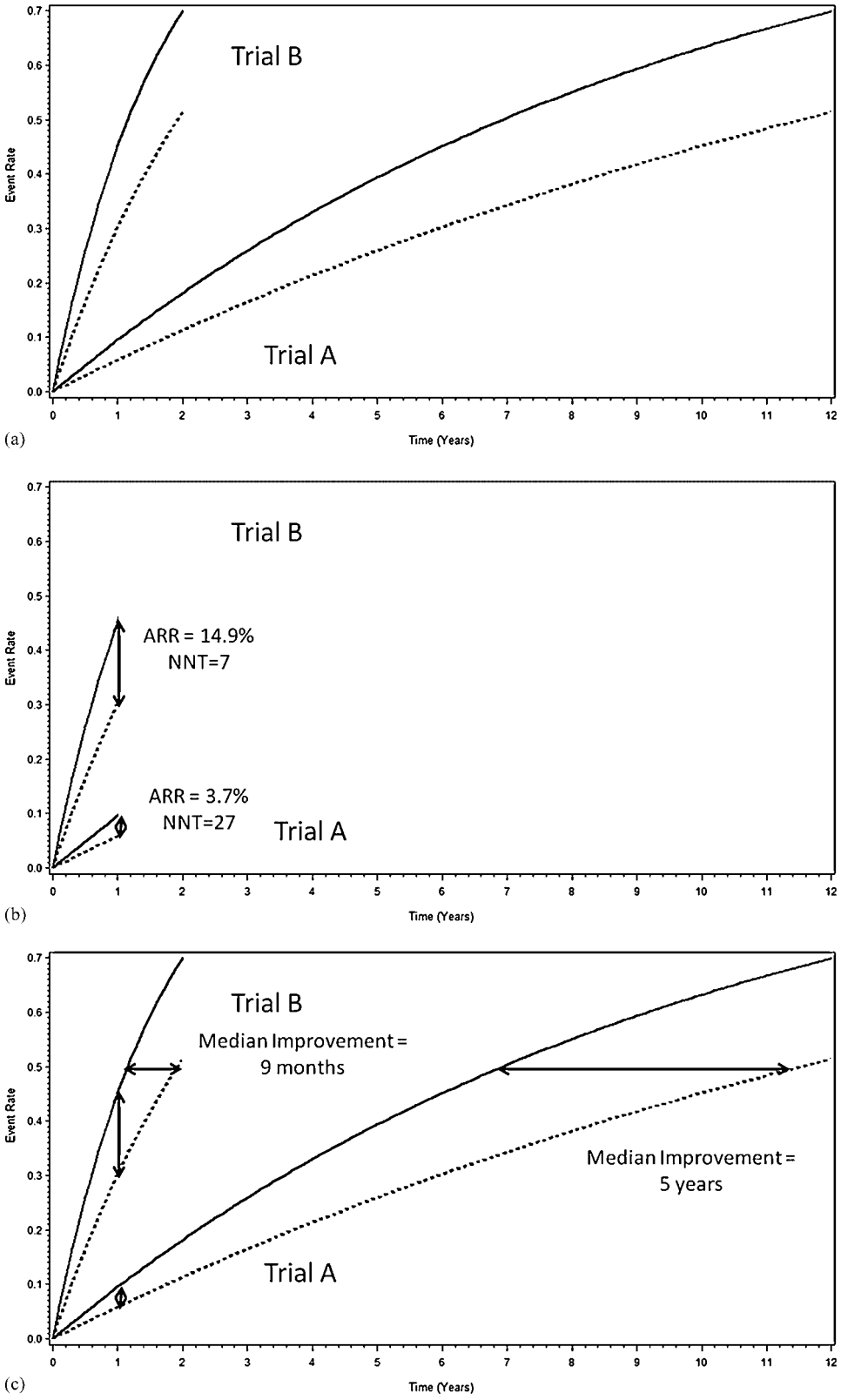

Figure 1. Kaplan–Meier curves for two hypothetical trials, both with exponential survival distributions and a hazard ratio of 0.6. In both trials the control group is the solid line and the experimental group is the dashed line. Trial A has relatively low hazard rates and Trial B has relatively high hazard rates.

trials are illustrated by the vertical separation between the curves at 1 year; this separation is clearly larger for trial B, in which the hazard rates are higher. The bottom panel again shows the same curves, but now extended beyond the time at which 50% of the patients in both treatment arms have experienced the event; this type of curve is more typical of oncology trials. In this case, the horizontal separation between these treatments at the 50% point on the y-axis, which represents the median treatment benefit, is clearly larger for trial A, in which the hazard rates are lower.

As shown mathematically in the previous section, this inconsistency is not specific to any example, but will happen in every case when the survival distributions are exponential and the hazard ratio is the same in both trials.

Cost effectiveness

Perhaps the primary reason that clinicians and regulators require that a new drug has a clinically meaningful effect is that all drugs have risks and costs, and the benefit of the drug must outweigh those risks and costs. If there were a hypothetical drug that we knew with absolute certainty had no risks and no costs, it is difficult to imagine a clinical benefit too small to allow that drug to be used. Therefore, rather than focusing solely on clinical meaningfulness, it seems reasonable to consider cost-effectiveness as well.

Consider the same model described earlier, in which survival distributions are exponential, with hazard rate γ in patients taking the standard treatment, and hazard ratio λ. Now assume that the treatment has a cost associated with it that accrues at the constant rate of c per unit time. We use the term 'cost' loosely to include both financial costs and side effects. For example, the cost might include \$2/day to purchase a daily pill; it might include the nausea one feels every time one takes the pill; or it might include a composite measure of the two.

We define cost-effectiveness to be the ratio of the lifetime cost of the treatment to the life-years saved by the treatment (or additional life-years free of the event, for nonfatal events). Under the model defined above, the expected life-years saved is the difference between life expectancy with the treatment, or $1/(\gamma\lambda)$, and life expectancy without it, or $1/\gamma$. Therefore,

$$\text{Expected Life-Years Saved} = 1/(\gamma\lambda) - 1/\gamma = (1-\lambda)/(\gamma\lambda).$$

The expected lifetime cost of the treatment is simply the product of the cost per unit time and the life expectancy with the treatment, or

$$\text{Expected Lifetime Cost} = c/(\gamma\lambda).$$

Cost-effectiveness is the ratio of these factors, or

$$\text{Cost-Effectiveness} = \{c/(\gamma\lambda)\}/\{(1-\lambda)/(\gamma\lambda)\} = c/(1-\lambda).$$

This equation for cost-effectiveness, $c/(1-\lambda)$, is remarkable for its simplicity, but even more remarkable for the fact that, though it involves the hazard ratio, it does not involve the hazard rate! Consider the consequence of that fact: For a given hazard ratio, neither the ARR, NNT nor the median difference has any impact on this measure of cost-effectiveness. (Of course, this formulation assumes that cost is constant between the populations being compared. This is almost certainly a reasonable assumption with respect to financial cost, but it may not always be true with respect to risks.) In the example described above, despite the difference in ARR or NNT and in median survival between trials A and B, the cost-effectiveness is identical.

This result seems to fly in the face of common wisdom regarding clinical meaningfulness and cost-effectiveness, but it is less surprising when one considers the contradiction between ARR or NNT and the median difference. A person who is convinced that the cost-effectiveness must be greater when ARR is high and NNT is low (i.e. in a high-risk population) must also be convinced that a treatment that has a smaller impact on the median difference is the more cost-effective. On the other hand, a person who is convinced that cost-effectiveness must be greater when the median difference is larger (i.e. in a low-risk population) must also be convinced that a treatment with a smaller ARR and a larger NNT is the more cost-effective. The equation $c/(1-\lambda)$ shows that neither position is correct.

Extension to the Weibull distribution

The Weibull distribution is a two-parameter generalization of the exponential distribution. The shape parameter, k, determines whether the hazard rate is constant over time ($k=1$, in which case the Weibull distribution reduces to the exponential distribution), decreases over time ($0<k<1$) or increases over time ($k>1$), and a scale parameter, s, that determines the magnitude of the hazard rate for a given value of k. Note that if the survival distributions for two treatment groups are Weibull with the same shape parameter, the hazard ratio, λ, is equal to the ratio of the scale parameters to the power k.

While the results in this paper were derived assuming exponential survival distributions, they also apply to the Weibull distribution when the two groups have the same shape parameter. Specifically: (1) for a given hazard ratio, the ARR increases as the hazard rates increase as long as the event rates are relatively low. The event-free rates corresponding to the change-over time do not depend on k; i.e. they are $\lambda^{1/(1-\lambda)}$ in the control group and $\lambda^{\lambda/(1-\lambda)}$ in the treatment group, just as in the exponential case. (2) For a given hazard ratio, the difference in medians decreases as the hazard rates increase. (3) Cost-effectiveness depends on cost, the hazard ratio, and the shape parameter, but not on the scale parameter, as follows: cost-effectiveness$=c/(1-\lambda^{1/k})$. Alternatively, cost-effectiveness$=c/(1-r)$, where r is the ratio of the scale parameters in the two groups.

Discussion

In this paper we have mathematically derived two results that have important implications for understanding the concepts of clinical meaningfulness and cost-effectiveness in the context of a time-to-event variable: (1) that two commonly used criteria to assess clinical meaningfulness, the ARR (or, equivalently, the NNT) and the median difference, directly contradict each other; and (2) that cost-effectiveness depends on the cost and the hazard ratio (and, in the Weibull case, the shape parameter), but not the overall risk of the population (i.e. the hazard rate). The implications for the evaluation of treatment effects can be important, since they imply that all measures of absolute treatment effect are of little value compared with a relative measure, the hazard ratio.

It is true that proponents of absolute measures have cogent arguments for their choices. One can easily imagine a proponent of the ARR or NNT arguing that it is illogical to expose patients to costs and risks in an attempt to prevent an event that they are unlikely to experience in the near future. Similarly, one can easily imagine a proponent of the median difference arguing that it is illogical to expose patients to costs and risks in an attempt to postpone an event only briefly. The problem is not that these arguments are defective; on the contrary, both make good sense. However, it is important for each of these proponents to recognize that if their argument is correct, then the other proponent's argument must be incorrect.

One important limitation of our results is that we derived them under an assumption of exponential survival distributions (i.e. constant hazard rates) and of constant costs over time. However, the results extend to the Weibull distribution (assuming that the shape parameter is the same in the two treatment groups), which is quite a flexible distribution, allowing for both increasing and decreasing hazard rates over time. Another limitation is that we have not considered the issue of competing risks. For example, consider a treatment to prevent fracture; our model assumes that all patients will eventually suffer a fracture, when, in fact, some will die without having had one. Despite these limitations, our results show that the conventional wisdom that relative measures of treatment effect are inappropriate for assessing clinical meaningfulness is simply not valid.

It is interesting to consider whether these results can be extended to other variables, say, a binary variable such as cured or not cured. Some authors have criticized the use of NNT specifically in the setting of a time-to-event variable (e.g. Kristiansen and Gyrd-Hansen [9]), but it is not clear why an absolute measure of benefit would be of little value for a time-to-event variable, but great value for a binary variable. This and other questions are left for future evaluations.

Acknowledgements

The authors thank Janet Wittes and Reynold Spector for their careful review of this manuscript, and their extremely helpful advice.

References

1. Thompson B. Statistical significance and effect size reporting: portrait of a possible future. *Research in the Schools* 1998; **5**:33–38.
2. Greenstein G. Clinical versus statistical significance as they relate to the efficacy of periodontal therapy. *Journal of the American Dental Association* 2003; **134**:583–591.
3. Kraemer HC, Morgan GA, Leech NL, Gliner JA, Vaske JJ, Harmon RJ. Measures of clinical significance. *Journal of the American Academy of Childhood Adolescent Psychiatry* 2003; **42**:1524–1529.
4. Replogle WH, Johnson WD. Interpretation of absolute measures of disease risk in comparative research. *Family Medicine* 2007; **39**:432–435.
5. Leung WC. Balancing statistical and clinical significance in evaluating treatment effects. *Postgraduate Medical Journal* 2001; **77**:201–204.
6. Hildebrandt M, Vervölgyi E, Bender R. Calculation of NNTs in RCTs with time-to-event outcomes: a literature review. *BMC Medical Research Methodology* 2009; **9**:21.
7. Ridker PM, Danielson E, Fonseca FAH, Genest J, Gotto Jr AM, Kastelein JJP, Koenig W, Libby P, Lorenzatti AJ, MacFadyen JG, Nordestgaard BG, Shepherd J, Willerson JT, Glynn RJ, for the JUPITER Study Group. Rosuvastatin to prevent cardiovascular events in men and women with elevated C-reactive protein. *The New England Journal of Medicine* 2008; **359**:2195–2207.
8. Fleming TR, Rothmann MD, Lu HL. Issues in using progression-free survival when evaluating oncology products. *Journal of Clinical Oncology* 2009; **27**:2874–2880.
9. Kristiansen IS, Gyrd-Hansen D. Cost-effectiveness analysis based on the number-needed-to-treat: common sense or non-sense? *Health Economics* 2004; **13**:9–19.