The Clinical Significance of Effect Sizes for Survival and Tumor Response Endpoints Using the Empirical Rule Effect Size

Major-Elechi BT, Sloan JA*, Novotny PJ, Sargent DJ, Grothey A, Lafky JM and Dueck AC

Abstract

Context: In planning oncology phase II and phase III clinical trials, the size of the expected effect for endpoints such as tumor response and overall survival are key parameters driving the sample size. We applied the empirical rule effect size approach, also known as the ½ standard deviation (SD) method, to define clinically significant effect sizes for overall survival and tumor response endpoints in a series of clinical trials.

Methods: The observed effect size was calculated for 12 phase II and 27 phase III completed cancer clinical trials identified by experts as being notable.

Results: The effect sizes of the phase II and phase III clinical trials ranged from -0.32 to 0.84 and 0.01 to 0.44 SDs respectively. Effect sizes for all but four of the phase II trials were less than a ½ SD. For phase III studies, the effect sizes for all but one study were below 0.4 SD and roughly 67% of them had an effect size smaller than 0.2 SDs. There were no differences across disease sites, although colorectal and breast trials did have slightly larger effect sizes.

Conclusions: Even highly noteworthy existing phase II and phase III oncology clinical trials rarely achieve the ½ SD level of clinical significance. This method allows for more ready interpretation of the clinical significance of overall survival and tumor response endpoints. It allows for cross-study comparison across different endpoints. The method also facilitates study design as it directly builds clinical significance into the study.

Keywords: Probability; Statistics in biology

Introduction

Effect size is a key parameter for estimating sample size and power as well as interpreting results for clinical and observational studies [1]. An effect size should communicate information that is useful to assessing the clinical significance of any result found in a randomized clinical trial [2]. Traditional effect size calculation of expected overall survival (OS) times for phase III clinical trials involves preliminary estimates typically from previous trials [3]. Similarly, phase II clinical trial estimates of tumor response rates (RR) are usually derived from phase I clinical trials or pilot studies [4]. The clinical meaningfulness of the effect size used in phase II and III clinical trials is most often determined by investigator opinion rather than statistical calibration [5]. The statistical significance of a trial result, which is heavily dependent on sample size, size may vary from its inferred clinical significance. Principles to judge the clinical significance of a statistically significant effect size in different clinical contexts are necessary.

In order to provide a context for discussion summarizing OS and RR, we consider an area where clinical significance has been extensively studied; that of quality of life (QOL) data. Investigations involving QOL effect size assessments typically involve an additional hurdle in that the clinical significance of changes in QOL scores has been a subject of much debate [6]. Much has been written about various approaches for assessing clinical significance for QOL endpoints including anchor-based and distributional-based approaches [7].

Among the distribution based methods, one popular approach is the so-called ERES method (Empirical Rule Effect Size) or the ½ standard deviation method [4]. Based on a combination of Chebyshev’s Theorem and Cohen’s Effect size classifications, it has been demonstrated that in a broad spectrum of applications and techniques a between-groups difference of a ½ standard deviation is clinically non-ignorable [8,9].
overall survival time, then it follows directly that $E(x) = t$, $V(x) = t^2$, $SD(x) = t$ and finally that $\frac{1}{2}SD(x) = \frac{t}{2}$. Given that $t$=median overall survival time, then $SD(x) = \frac{t}{ln2}$ and $\frac{1}{2}SD(x) = \frac{t}{2ln2}$.

Mathematical underpinnings for the $\frac{1}{2}$ standard deviation for tumor response

Assume the number successes, $x$, in a phase II clinical trial follows a binomial distribution with parameters $n$ and $p$ where $p$ is probability of success. Then it follows directly that $(x) = pV(x) = p(1-p)$, $SD(x) = \sqrt{p(1-p)}$ and finally that $\frac{1}{2}SD(x) = \frac{\sqrt{p(1-p)}}{2}$. The sample proportion, $\hat{p} = \frac{x}{n}$ can be used as an unbiased estimate of $p$.

Mathematical underpinnings for the $\frac{1}{2}$ standard deviation for hazard ratio

For phase III trials, the calibrated effect size is calculated as the difference in observed median overall survival between the two arms (i.e., observed treatment arm median $OS(\hat{m}_1)$ minus observed reference arm median $OS(\hat{m}_2)$ divided by the standard deviation of the reference arm median overall survival time computed as $m_2/ln2$. Of note, the calibrated overall survival effect size $\frac{(m_1 - m_2)}{m_2/ln2}$ can be re-written as $ln2 \frac{m_1}{m_2} - 1$ showing that the effect size is also a function of the true inverse hazard ratio.

Statistical analysis

The ERES calibration method can be applied to tumor response and overall survival endpoints. The calibrated effect size for overall survival is calculated as the difference in median overall survival divided by the standard deviation of the median overall survival time of the reference group. For the sake of emphasis in this paper, the observed difference in phase III clinical trials with three or more arms was taken as the difference between the reference group and the arm with the largest observed endpoint. In phase II trials the observed effect size was taken as the difference between the observed $p$ and the reference response rate taken from the power statement divided by the standard deviation of the reference response rate.

Sloan (2003) suggested that Cohen’s effect sizes could be utilized as a means to define a non-ignorable clinically meaningful effect size. Small, medium, and large effect sizes can be characterized by the size of a worm so as to be insignificant (small), an elephant so as to be obvious to all (large), and a duck which from various perspectives could be recognized as clinically non-ignorable (medium). The duck allegory came from the quotation of Justice Warren Berger relating to the definition of pornography stating that it may be difficult to define but people know it when they see it. Similarly a clinically meaningful effect size should be non-ignorable from various viewpoints. Norman et al. furthered this work to show that a medium effect size or $\frac{1}{2}$ SD is defensible from philosophical, clinical and even physiological perspectives.

For example, consider the following GlaxoSmithKlein phase III study (EGF104900; NCT00320385) where patients were randomized to receive single agent of lapatinib (1500 mg/daily) or a combination of lapatinib (1000 mg p.o. daily) plus trastuzumab (2 mg/kg). Women treated with monotherapy lapatinib experienced a median overall survival of 9.5 months compared with 14 months when treated with the combination (median HR: 0.74, p=0.026) [10]. This could be calibrated by saying that the effect is 0.36 standard deviations, which is a small/medium clinically meaningful effect size in this context although the p-value is statistically significant. The online only table gives further illustration for the $\frac{1}{2}$ standard deviation calibration method for overall survival and tumor response endpoints.

To demonstrate the application of the ERES method, we reviewed phase II and phase III clinical trials conducted between the years 1999 and 2012 and identified thirty-nine exemplary clinical studies derived from the North Central Cancer Treatment Group (NCCCTG), the Mayo Cancer Center, and a review conducted by Ocana and Tannock (2011) of clinical trials used by the United States Food and Drug Administration (FDA) for the approval of new molecular-targeted drugs. All studies reported median overall survival time or tumor response as a primary endpoint. The studies include the top three diagnosed cancers in the United States, as well as other cancers (7 breast cancer studies, 9 lung cancer studies, 12 gastrointestinal cancer studies and 11 other cancer studies). We further include material drawn from a recent review by the ASCO Cancer Research Committee for assessing clinically significant treatment effects11. Effect size was expressed as a multiple of the standard deviation (SD) with the following classifications: small effect size 0.2 SD, medium effect size 0.5 SD and large effect size 0.8 SD.

Results

Tables 1 and 2 present the key design parameters and observed results for the various phase II and phase III trials, respectively. For example in the first row of Table 1, results indicate for study NCCCTG 954651, a phase II trial of oral 776C85 and oral 5-FU for patients with metastatic colorectal cancer, the study was designed with an assumed null response rate of 10% and an observed response rate of 26%. The calibrated effect size for this observed response rate is equivalent to 0.53 SD, which is slightly bigger than the non-ignorable $\frac{1}{2}$ SD cut-off, which for this trial would have been 25% response rate. In Table 2, for example, the first row indicates that NCCCTG 959255, a study of different agents for anorexia and cachexia, reported an observed difference of 18 days improvement in overall survival over the control group and a median survival of 120 days (p=0.66). The calibrated effect size for this difference is 0.1 SD, a small effect size. In contrast, the WILCG study by Furuse et al., comparing concurrent vs. sequential treatment for NSCLC, reported a 3.2 month improvement in survival (p=0.04), which while statistically significant, represents a small calibrated effect size of 0.17 SD. Figure 1A presents effect sizes for the various phase II clinical trials making for a ready cross-study comparison of the magnitude of observed tumor response results. The effect sizes of the phase II clinical trials (Table 1) range from -0.32 SD (metastatic lung cancer) to 10.84 SD (metastatic breast cancer) and all but four trials have observed effects less than $\frac{1}{2}$ standard deviation. A negative effect size indicates that the tumor response was less than the assumed null hypothesis. There were no substantial differences across disease sites.

Figure 1B presents effect sizes ranging from 0.01 SD to 0.44 SD for the phase III studies (Table 2), in order of magnitude. All effect sizes are below a $\frac{1}{2}$ standard deviation threshold of a non-ignorable clinical effect with two thirds of the trials having effect sizes less than 0.2 SD. Phase III breast and colorectal cancer clinical trials had the largest effect sizes of 0.34 SD and 0.44 SD, respectively. Note, the effect sizes of all observed lung studies are less than 0.2 SD.

Ocana and Tannock performed a comprehensive review of phase III randomized clinical trials used by the FDA to approve new molecular-targeted drugs since 2000 [11,12].
<table>
<thead>
<tr>
<th>Study no.</th>
<th>Description</th>
<th>Tumor sites</th>
<th>Study type</th>
<th>Reference response rate</th>
<th>Observed response rate</th>
<th>Non-ignorable clinically significant response rate (½ SD)</th>
<th>Calibrated observed effect size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>954651 [16]</td>
<td>Phase II trial of oral 776CBS and oral 5-FU</td>
<td>GI Metastatic</td>
<td></td>
<td>10%</td>
<td>26%</td>
<td>0.25</td>
<td>0.53</td>
</tr>
<tr>
<td>983252* [17]</td>
<td>Paclitaxel, carboplatin and trastuzumab for metastatic breast cancer</td>
<td>Breast Adjuvant</td>
<td></td>
<td>40%</td>
<td>65%</td>
<td>0.64</td>
<td>0.51</td>
</tr>
<tr>
<td>983252* [17]</td>
<td>Paclitaxel, carboplatin and trastuzumab for metastatic breast cancer</td>
<td>Breast Adjuvant</td>
<td></td>
<td>40%</td>
<td>81%</td>
<td>0.64</td>
<td>0.84</td>
</tr>
<tr>
<td>N0021* [18]</td>
<td>Gemcitabine and epirubicin for the treatment of mesothelioma</td>
<td>Head and Neck</td>
<td>Other: Locally Advanced</td>
<td>10%</td>
<td>13%</td>
<td>0.25</td>
<td>0.10</td>
</tr>
<tr>
<td>N0021* [18]</td>
<td>Gemcitabine and epirubicin for the treatment of mesothelioma</td>
<td>Head and Neck</td>
<td>Other: Locally Advanced</td>
<td>10%</td>
<td>7%</td>
<td>0.25</td>
<td>-0.10</td>
</tr>
<tr>
<td>N0022 [19]</td>
<td>Oral vinorelbine for the treatment of metastatic NSCLC</td>
<td>Lung Metastatic</td>
<td></td>
<td>15%</td>
<td>3.4%</td>
<td>0.33</td>
<td>-0.32</td>
</tr>
<tr>
<td>N0044 [20]</td>
<td>Preoperative radiation and chemotherapy for locally advanced esophageal cancer</td>
<td>GI Other: Neoadjuvant</td>
<td></td>
<td>40%</td>
<td>35%</td>
<td>0.64</td>
<td>-0.10</td>
</tr>
<tr>
<td>N0087* [21]</td>
<td>Interleukin-12 and rituximab in patients with non-Hodgkin’s lymphoma</td>
<td>Lymphoma Other</td>
<td></td>
<td>50%</td>
<td>37%</td>
<td>0.75</td>
<td>-0.26</td>
</tr>
<tr>
<td>N0087* [21]</td>
<td>Interleukin-12 in combination and rituximab in patients with non-Hodgkin’s lymphoma</td>
<td>Lymphoma Other</td>
<td></td>
<td>50%</td>
<td>52%</td>
<td>0.75</td>
<td>0.04</td>
</tr>
<tr>
<td>N0149 [22]</td>
<td>Oxaliplatin and Capecitabine for adenocarcinoma</td>
<td>GI Cancer Control</td>
<td></td>
<td>15%</td>
<td>35%</td>
<td>0.33</td>
<td>0.56</td>
</tr>
<tr>
<td>N014C [23]</td>
<td>PS-341 and gemcitabine for pancreatic adenocarcinoma</td>
<td>Pancreas Metastatic</td>
<td></td>
<td>5%</td>
<td>0%</td>
<td>0.16</td>
<td>-0.23</td>
</tr>
<tr>
<td>N0242 [24]</td>
<td>Docetaxel and capecitabine for adenocarcinoma</td>
<td>GI Metastatic</td>
<td></td>
<td>20%</td>
<td>39%</td>
<td>0.40</td>
<td>0.48</td>
</tr>
</tbody>
</table>

* Randomized phase II trial.

**Table 1:** Application of ½ standard deviation calibration method for tumor response to existing phase II NCCTG clinical trials included in the sample [16-24].
The ERES method was applied to the 12 studies that listed OS as a primary outcome (3 breast studies, 3 colorectal studies, 2 lung studies and 4 other cancer trials). Figure 2 and Table 3 demonstrate that all of these studies achieve less than a ½ standard deviation and 67% (8/12) of the studies had effect sizes <0.2 SD.

Recently, the American Society for Clinical Oncology (ASCO) Cancer Research Committee derived estimates for clinically meaningful differences in outcomes for cancer clinical trials based on subjective perception of ASCO members and expert reviewers [11]. Table 4 presents the results of that effort supplemented by estimates derived via our method. Column A presents what would be a ½ standard deviation effect size which has been suggested as a non-ignorable effect [13]. The ½ SD effect sizes in Table 4 are all substantially larger than the upper estimates arrived at by expert review, with the exception of colon cancer. Hence the size of treatment effect that clinicians would consider as clinically meaningful for most cancer clinical trials is much smaller than ½ SD. Column B of Table 4 indicates that in pancreatic cancer a calibrated effect size of between 0.25 and 0.46 SD and in colon cancer a calibrated effect size between 0.35-0.87 SD could be considered clinically significant, compared to effect sizes of around 0.20 SD, for breast and lung cancer.

**Discussion**

The present article provides a calibration method for interpreting results of clinical trials within the context of similar trials in the same disease and even across disease sites. This ERES calibration method

---

**Table 2:** Application of ½ standard deviation calibration method for overall survival to existing phase III clinical trials [25-39].

<table>
<thead>
<tr>
<th>Study</th>
<th>Treatment Description</th>
<th>Tumor Type</th>
<th>OS (M)</th>
<th>SD (M)</th>
<th>Wilcoxon Test (p)</th>
<th>ERES (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTOG 9410 [31]</td>
<td>Concurrent versus sequential treatment with radiotherapy (RT) and chemotherapy (CT) for NSCLC</td>
<td>Lung</td>
<td>Metastatic</td>
<td>14.6</td>
<td>2.4</td>
<td>months</td>
</tr>
<tr>
<td>GMMA Ankara 1995 [32]</td>
<td>Concurrent versus sequential treatment with radiotherapy (RT) and chemotherapy (CT) for NSCLC</td>
<td>Lung</td>
<td>Metastatic</td>
<td>10</td>
<td>1</td>
<td>months</td>
</tr>
<tr>
<td>GLOT-GFPC NPC 9501 [33]</td>
<td>Concurrent versus sequential treatment with radiotherapy (RT) and chemotherapy (CT) for NSCLC</td>
<td>Lung</td>
<td>Metastatic</td>
<td>14.5</td>
<td>1.8</td>
<td>months</td>
</tr>
<tr>
<td>EORTC 08972-22973 [34]</td>
<td>Concurrent chemo-radiotherapy versus sequential chemo-radiotherapy for inoperable NSCLC</td>
<td>Lung</td>
<td>Metastatic</td>
<td>16.2</td>
<td>0.3</td>
<td>months</td>
</tr>
<tr>
<td>Irinotecan [35]</td>
<td>Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer</td>
<td>Colorectal/ colon</td>
<td>Metastatic</td>
<td>4.3</td>
<td>2.7</td>
<td>months</td>
</tr>
<tr>
<td>Cetuximab NCT00079066 [36]</td>
<td>Cetuximab in patients with advanced colorectal cancer</td>
<td>Colorectal/ colon</td>
<td>Metastatic</td>
<td>4.6</td>
<td>1.5</td>
<td>months</td>
</tr>
<tr>
<td>FOLFOX NCCTG N9741 [37]</td>
<td>Combinations of FU/LV, irinotecan and oxaliplatin in patients with previously untreated metastatic colorectal cancer</td>
<td>Colorectal/ colon</td>
<td>Metastatic</td>
<td>15</td>
<td>4.5</td>
<td>months</td>
</tr>
<tr>
<td>CLEOPATRA [38]</td>
<td>Combination of pertuzumab plus trastuzumab plus docetaxel as compared with placebo plus trastuzumab plus docetaxel when used as first-line treatment for HER2-positive metastatic breast cancer</td>
<td>Breast</td>
<td>Metastatic</td>
<td>12.4</td>
<td>6.1</td>
<td>months</td>
</tr>
<tr>
<td>EMILIA [39]</td>
<td>T-DM1 compared with lapatinib plus capecitabine in patients with HER2-positive advanced breast cancer previously treated with trastuzumab and a taxane</td>
<td>Breast</td>
<td>Metastatic</td>
<td>25.1</td>
<td>5.6</td>
<td>months</td>
</tr>
</tbody>
</table>
Figure 1: Calibrated effect sizes for tumor response in phase II (A) and median OS in phase III (B) NCCTG clinical trials ordered by size of effect.
Figure 2: Calibrated effect sizes for overall survival of phase III clinical trials used by the FDA for the approval of new molecular-targeted drugs since 2000 ordered by size of effect (Ocana and Tannock [12]).

Table 3: Characteristics and calibrated effect sizes for overall survival of phase III randomized clinical trials used by the United States Food and Drug Administration (FDA) for the approval of new molecular-targeted drugs during 2000-2010 [12].
has previously been used to benchmark QOL outcomes. However, the empirical data reflecting the results of recent oncology clinical trials analyzed here indicated that under the vast majority of circumstances, the calibrated effect sizes of these oncology clinical trials are less than ½ standard deviation. The ½ standard deviation criteria is not an arbitrary value but has been empirically derived based on a combination of Chebyshev’s Theorem and the pervasive “Cohen’s d” effect size. However, a limitation of the method is that it is dependent on the distribution of the survival function and also the proportional hazards assumption.

It has been previously demonstrated that the effect sizes of recent ‘positive’ randomized controlled clinical trials are smaller than those in the past and that the main predictor of a positive clinical trial is a statistically significant result [14]. Although 19 of the 27 phase III trials were statistically significant, none had effect sizes >0.5 SD and almost all (25 of 27) had effect sizes <0.2 SD. Four of the 12 phase II trials had effect sizes >0.5 SD. It is therefore imperative to develop principles that underlie the clinical significance threshold in different clinical contexts. It also opens up the question as to whether a ½ SD rule is practical for oncology trials with an OS endpoint, since it may not be achievable for survival outcomes. This is not to discourage or disparage recent and future oncology clinical trials, but hopefully put expectations into perspective. Further, the method still allows for comparative calibration across clinical trials involving survival outcomes.

A further perspective that is challenged by this work is the extensive discussion and conclusion regarding effect sizes for QOL outcomes. Often it has been noted that the size of effects seen in QOL studies is less than a ½ standard deviation and is declared a clinically insignificant result. However, if the same rubric were to be applied to tumor response and overall survival effect sizes, the vast majority of those results would also be characterized as clinically non-significant. This opens the discussion to a potential re-examination of how one can interpret clinical trial results for overall survival and tumor response outcomes versus QOL-related domains and toxicity. The ERES method can be helpful here, as described elsewhere, in the development of a quality adjusted survival metric which combines survival and toxicity/QOL into a single summary statistic [15].

**Conclusion and Implications**

The ERES method allows for a simple and mathematically consistent interpretation of the clinical significance of overall survival and tumor response studies. It allows for direct cross-study comparison across different end points and facilitates study design as it builds clinical significance into the study directly. The primary advantage of the ERES method is that it allows for a calibration of design parameters and trial results into a common metric. For example, when comparing results of two clinical trials, one might encounter an improvement of three months of overall survival that represents a small effect size in one trial and a moderate effect size in another. This is consistent with the findings of the ASCO committee that an additional month of overall survival could be a large improvement (as in metastatic pancreatic cancer) or a small improvement (as in early stage breast disease).

The next step in this line of research (in progress) is to produce real time applications that can facilitate effective communication among clinical trial development teams.

Ready-made applications for electronic devices could be employed to translate previous literature and subjective estimates of efficacy into data for constructing and comparing alternative statistical design parameters. Ultimately, the success of this approach will lie in the ability for the abstract effect size approach to be understood and readily used by individuals involved in the design and interpretation of clinical trials.

**Acknowledgements**

Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Numbers U10CA180821, U10CA180882, and UG1CA189923 to the Alliance for Clinical Trials in Oncology; CA25224 to the legacy North Central Cancer Treatment Group (NCCTG); and CA-37404, CA-35431, CA-35415, CA-35103, and CA-35269. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**References**


---

Table 4: Summary of recommended overall survival targets for meaningful clinical trial goals.

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Patient population</th>
<th>Current baseline median overall survival</th>
<th>Improvement over current overall survival that would be clinically meaningful*</th>
<th>½ Standard deviation (column A)</th>
<th>Calibrated effect size (SD) (column B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pancreatic Cancer</td>
<td>FOLFIRINOX Eligible Patients</td>
<td>10-11 months</td>
<td>4-5 months</td>
<td>7.21-7.93 months</td>
<td>0.25-0.35</td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>Gemcitabine Eligible Patients</td>
<td>6-8 months</td>
<td>3-4 months</td>
<td>4.33-5.77 months</td>
<td>0.26-0.46</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>Non-squamous cell carcinoma</td>
<td>13 months</td>
<td>3.25-4 months</td>
<td>9.38 months</td>
<td>0.17-0.21</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>Squamous cell carcinoma</td>
<td>10 months</td>
<td>2.5-3 months</td>
<td>7.21 months</td>
<td>0.17-0.21</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>Metastatic triple negative, previously untreated for metastatic disease</td>
<td>18 months</td>
<td>4.5-6 months</td>
<td>12.98 months</td>
<td>0.17-0.23</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>Disease progression on all prior therapies (or not a candidate for standard 2nd or 3rd line options)</td>
<td>4-6 months</td>
<td>3-5 months</td>
<td>2.89-4.33 months</td>
<td>0.35-0.87</td>
</tr>
</tbody>
</table>


