HHS Public Access

Author manuscript

Int J Stat Med Res. Author manuscript; available in PMC 2019 August 08.

Published in final edited form as:

Int J Stat Med Res. 2018 ; 7(4): 137–146. doi:10.6000/1929-6029.2018.07.04.4.

Combining Survival and Toxicity Effect Sizes from Clinical Trials: NCCTG 89-20-52 (Alliance)

Brittny T. Major-Elechi[1], **Paul J. Novotny**[1], **Jasvinder A. Singh**[2], **James A. Bonner**[3], **Amylou C. Dueck**[4], **Daniel J. Sargent**[1], **Axel Grothey**[5], **Jeff A. Sloan**[1,*]

[1]Division of Biomedical Statistics and Bioinformatics, Mayo Clinic, Rochester, MN 55905, USA

[2]Birmingham Veterans Affairs Medical Center, and the Department of Medicine and Epidemiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA

[3]Department of Radiation Oncology, University of Alabama at Birmingham, Birmingham, AL 35294, USA

[4]Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ, USA

[5]Department of Medical Oncology, Mayo Clinic, Rochester, MN 55905, USA

Abstract

Background: How can a clinician and patient incorporate survival and toxicity information into a single expression of comparative treatment benefit? Sloan *et al*. recently extended the ½ standard deviation concept for judging the clinical importance of findings from clinical trials to survival and tumor response endpoints. A new method using this approach to combine survival and toxicity effect sizes from clinical trials into a quality-adjusted effect size is presented.

Methods: The quality-adjusted survival effect size (QASES) is calculated as survival effect size (ESS) minus the calibrated toxicity effect sizes (EST) (QASES=ESS-EST). This combined effect size can be weighted to adjust for the relative emphasis placed by the patient on survival and toxicity effects.

Results: As an example, consider clinical trial NCCTG 89–20-52 which randomized patients to once-daily thoracic radiotherapy (ODTRT) versus twice-daily treatment of thoracic radiotherapy (TDRT) for the treatment of lung cancer. The ODTRT vs. TDRT arms had median survival time of 22 vs. 20 months ($p=0.49$) and toxicity rate of 39% vs. 54%, ($p<0.05$). The QASES of 0.18 standard deviations translates to a quality-adjusted survival difference of 5.7 months advantage for the ODRT arm over the TDRT treatment arm (22(16.3) months), $p<0.05$). Similar results are presented for the four possible case combinations of significant/non-significant survival and toxicity benefits using completed clinical trials.

*Address correspondence to this author at the Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA; Tel: (507) 284-9985; Fax: (507) 266-2477; jsloan@mayo.edu.

Conclusions: We used a novel approach to re-analyze clinical trial data to produce a single estimate for each treatment that combines survival and toxicity data. The QASES approach is an intuitive and mathematically simple yet robust approach.

Keywords

Survival; toxicity; quality of life; effect size; quality-adjusted life years; QALY

INTRODUCTION

How can a clinician and patient incorporate survival and toxicity information into a singular expression of comparative treatment efficacy? If an anti-cancer agent, for example, has a positive effect on survival but impacts an increased symptom burden on patients, how can one integrate these disparate information sources into a scientifically supportable decision? Clinicians routinely make such integrations using experience, professional opinion and literature reviews but the process is highly individualized and ad hoc [1]. This lack of standardization leaves open the question as to what constitutes a clinically meaningful treatment effect in terms of a composite, complex entity encompassing both survival and toxicity.

Quality-adjusted life years (QALYs) represent an attempt to measure disease burden that accounts for both the quality and quantity of life lived [2]. QALY-adjusted survival analysis builds upon the Kaplan-Meier approach to survival analysis by attributing a value to each day that reflects the quality of that day. Traditionally, survival curves and toxicity tables are analyzed separately and the lack of integration between the two endpoints leads to questioning of the usefulness of QALYS in clinical settings. The issue of calibration is a major barrier to the routine use of QALYs in survival analysis. The Q-TWiST method divides overall survival (OS) into three states: toxicity (TOX), time without symptoms or toxicity before relapse (TWiST) and time from disease progression or relapse to death (REL) [3]. The overall survival time (OS) for a patient can then be written as: OS = TOX + TWiST + REL. Days with toxicity and days after relapse are considered to be of less value and are given less weight in the analyses. This method has been successfully implemented, for example, in the analysis of breast cancer trials [4] to examine the benefit of adjuvant chemotherapy treatment versus its toxic effects.

Unfortunately, issues in calibration arise in application of the QALY model to clinical data [5]. In the initial application of the Q-TWiST assignation, every toxicity incident was assumed to last for a period of three months. Upon reflection, such a severe impact of toxicity was necessary to magnify the toxicity impact on the QALY calculations so that the overall survival results could be adjusted sufficiently to observe a significant effect. Similarly, Sloan and colleagues demonstrated that the mixture of utility parameters had to be extreme to see quantitatively distinguishable results in the QALY model when applied to a clinical trial where the survival curves were on top of one another but the toxicity incidence rates were doubled from one treatment to the next [6]. In another Q-TWiST study of cyclophosphamide, epirubicin, fluorouracil versus cyclophosphamide, methotrexate, fluorouracil treatment for premenopausal women with nodepositive breast cancer, Radice

(2005) found that there was no combination of utility values that would produce a statistically significant comparison between treatments despite the fact that an intuitive inspection of the data would lead one to believe that a clinically relevant difference might well exist [7]. A recent review by Tate and Skrepnek found that only six of 284 studies that employed QALY estimates reported significant findings and highlighted a number of methodological weaknesses [8].

In a more recent paper discussing the power of QALY analyses, Sloan *et al.* demonstrated that the Q-TWiST method has may have impractical power considerations [9]. For example, they showed via a series of simulations, that a quadrupling of toxicity would be required to have 80% power to detect a statistically significant difference in survival time in many clinical trial designs. Such an imbalance of toxicity incidence across arms would cause a trial to cease long before such an effect size was observed. This invites the question as to whether the QALY approach as presently designed and implemented has practical applicability.

Across various QOL settings, 1/2 standard deviation (SD) has been shown to be clinically meaningful [10]. Sloan *et al.* recently extended the ½ standard deviation concept to clinical trials and derived a ½ standard deviation calibration method for survival and tumor response endpoints. This method expresses survival differences and tumor response rates in terms of standard deviations. We can express toxicity in terms of standard deviations as well producing a toxicity effect size. This paper extends that idea to combine the survival and toxicity effects into one quality-adjusted survival effect size.

METHODS

We have previously shown the mathematical underpinnings of the survival and tumor response effect size calculation [11] and include herein a brief synopsis in appendix 1. Similar to tumor response effect sizes, toxicity effect sizes can be calculated assuming the adverse event follows a binomial distribution. We build upon that method by providing the integration of the two effect sizes, survival and toxicity, into a single quality-adjusted effect size (Figure 1).

Combined (Calibrated) Effect Size

The combined effect size is calculated as the survival effect size minus the toxicity effect size. The subtraction is necessary to account for the difference in impact of the survival time from the toxicity rate (i.e., one is positive; one is negative in terms of QALY survival). This combined effect size can be computed for each toxicity event or for overall toxicity and also can be weighted to adjust to vary the magnitude of toxicity calibration.

$$Total\ Effect\ Size = \frac{w_1 ES_A - w_2 ES_B}{w_1 + w_2} \quad \text{where } 0 \le w_1, w_2 \le 1.$$

This calibrated effect size can be used to back-calculate a difference in overall survival that adjusts for the effects of toxicity.

$$\Delta\ OS = Calibrated\ Total\ Effect\ Size * Standard\ Deviation\ of\ the\ Reference\ Arm$$

RESULTS

To illustrate our method of calibrated effect sizes (or Quality-Adjusted Survival Effect Size (QASES)), we consider an exemplary trial in which there are differing toxicity profiles between the reference and treatment arms but no difference in overall survival. A phase III clinical trial carried out by the North Central Cancer Treatment Group, NCCTG 89–20-52, randomized patients to once-daily thoracic radiotherapy (ODTRT) versus twice-daily treatment of thoracic radiotherapy (TDRT) for the treatment of lung cancer [12]. The NCCTG is now part of the Alliance for Clinical Trials in Oncology. Table 1 gives a summary of the trial. The ODTRT arm had a median survival time of 22 months and overall toxicity profile of 39%. The TDTRT arm had a median survival time of 20 months and overall toxicity profile of 54%. The difference in overall survival was not significant (p=0.49).

The standard deviation of the overall survival endpoint of the reference group is 31.74 months, with overall survival difference of −2 months. This equates to an effect size of −0.06 months/SD. The standard deviation of toxicity of the reference group is 0.49, with toxicity difference of 0.15. This equates to an effect size of 0.3 months/SD. We can now adjust the survival effect size for this medium toxicity effect size. Given weights w_1=1 and w_2=1, the effect size is −0.18 i.e. considering toxicity events as well as the overall survival outcome the effect size is 0.18 in favor of the control arm (ODTRT).

This calibrated effect size can then be used to back calculate the difference in survival that accounts for toxicity, or in other words, a quality-adjusted survival effect size (QASES). $\Delta Median\ OS = 0.18 * \left(\frac{22}{\ln 2}\right) = -5.7\ months$. The quality-adjusted survival difference is −5.7 months. This is equivalent to a median quality-adjusted OS for the TDRT arm of 16.3 months compared to 22 months in the ODTRT arm, which would have been statistically significant had these been unadjusted mean survival times.

Table 2 further demonstrates how the combined effect sizes and back-calculated difference in survival time vary as the survival and toxicity weights, w_1 and w_2, change where w_1 + w_2=1. The weight w_1=0 is completely toxic and w_1=1 is no toxicity. In this example, the combined effect size with zero weight given to toxicity produces the least negative combined effect size −0.06 and a back calculated difference in the quality-adjusted survival time between the two arms of −1.9 (20.1 months versus 22 months). Regardless of the combination of weights chosen, the adjusted survival difference is still in favor of the ODRT arm. This is expected given a non-significant overall survival difference and an increased toxicity profile. However, the difference in overall survival benefit is now adjusted for the toxicity. We will demonstrate other cases in which a minimal significant survival benefit can be reduced given increased toxicity events as well other possible scenarios.

Examination of the 4 QASES Cases

In this section, we attempt to present exemplary results of the QASES approach via taxonomy of possible combinations of survival and toxicity results. For any given trial, the

survival data and toxicity data for a given clinical trial will be either significantly in favor of one treatment arm or statistically nonsignificant in terms of p-values. There are four possible cases that will arise in the combination of significant and non-significant survival and toxicity comparisons between treatment arms: both are non-significant, both are significant, the survival comparison is significant but the toxicity comparison is not, and the survival comparison is non-significant but the toxicity comparison is (Figure 1). The first case was exemplified in the previous section by the randomized clinical trial (RCT) with a non-significant survival comparison and synergistic toxicity comparison. In general, the first case, the treatment arm does not provide a significant increase in overall survival time and no significant decrease in overall adverse events (−/−). The second case is a RCT with a non-significant survival comparison but antagonistic toxicity comparison. In this case, the treatment arm does not provide a significant increase in overall survival but it does significantly reduce the adverse event rate (−/+). The third case is a RCT with a significant survival comparison but antagonistic toxicity comparison. In this case, the treatment arm provides a significant overall survival benefit but also a significant increase in toxicity (+/−). The fourth and final case is a RCT with a significant survival comparison and a synergistic toxicity comparison. In this case, the treatment arm significantly increases the overall survival time as compared to the control arm and there is a nonsignificant difference in the toxicity profiles or ideally a decrease in adverse events (+/+).

We apply our method of the quality-adjusted survival effect size to each of these cases using exemplary trials over the last decade. For each trial, a subset of exemplary toxicities reported from the trial results manuscript rather than a thorough examination of all toxicities reported for each trial.

Case 1. Non-significant survival benefit and a significant synergistic toxicity comparison (−/−)

(Our first example described above)

Case 2. Non-significant survival benefit but antagonistic toxicity comparison (−/+)

In this case we consider a trial in which there was a minimal, non-significant increase in median overall survival accompanied by a significant improvement in the toxicity profile. S-1 is a fourth generation oral fluoropyrimidine approved in Japan, Korea, Singapore, and China for the treatment of advanced gastric adenocarcinoma and in Japan and Korea for adjuvant therapy of gastric adenocarcinoma after a curative resection. With the hypothesis that S-1 in cisplatin/S-1 could improve overall survival, safety, and convenience compared to cisplatin/infusional fluorouacil, a non-Asian global phase III trial was initiated in March 2005 in the FLAGS trial [13].

The median overall survival time of the cisplatin/S-1 treatment group was 8.6 months compared to 7.9 months in the cisplatin/fluorouracil intravenous infusion control group (log-rank p=0.2; hazard ratio 0.92; 95% CI, 0.80–1.05). Difference in median survival time was 0.7 months. This produced an effect size of 0.06. For this calibration, the adverse event rates for neutropenia, fatigue, diarrhea and stomatitis were considered. The toxicity levels were less in the treatment group than in the control group for neutropenia (60% versus 83%),

stomatitis (6% versus 30%) and diarrhea (29% versus 38%). The fatigue toxicity levels were about equal in the treatment and control group (39% versus 39%). The combined quality-adjusted effect size increases to 0.32, 0.29 and 0.13 standard deviations once the impact of neutropenia, stomatitis and diarrhea respectively is included with weights w1 and w2=1. The quality-adjusted median survival time difference under the same weighting is larger than the overall survival difference for all adverse events except for fatigue (Table 1). Figure 2 demonstrates the range of quality-adjusted survival differences for each toxicity event with the restraint $w_1 + w_2=1$, where $w_1=1$ equates to the reported median overall survival difference. The quality-adjusted survival difference begins at 3.65 months for neutropenia, 3.33 months for stomatitis and 1.43 months for diarrhea survival weight, w_1, equal to zero, for the treatment vs. control group. The quality-adjusted median survival time advantage decreases as the weight of survival increases. If toxicity is considered an important factor in quality of life for the patient, this drug could be considered for the patient over the standard treatment although the difference in survival is minimal.

Case 3. Significant survival benefit but antagonistic toxicity comparison (+/−)

The third case is a clinical trial with significant difference in overall survival but an increased toxicity in the treatment arm.

The National Cancer Institute of Canada Clinical Trials Group (NCIC CTG) in cooperation with Australasian Gastrointestinal Tumor Group (AGITG) conducted a phase III trial compare erlotinib plus Gemcitabine with gemcitabine Alone in patients with advanced pancreatic cancer [14]. The median overall survival in the erlotinib plus Gemcitabine arm was 6.24 months compared to 5.91 months in the gemcitabine alone arm (hazard ratio 0.82; p=0.038). The difference in median overall survival is 0.33 months, which is a difference of 10 days and translates into an effect size of 0.04 standard deviations. The adverse events that were considered for the calibration are diarrhea, fatigue, stomatitis, and composite measure of any toxicity of grade 3 or 4. The percentages were greater in the treatment group for each event: diarrhea (56% versus 41%), fatigue (89% versus 86%), stomatitis (23% versus 14%) and any toxicity (grade 3/4) (62% versus 57%).

The quality-adjusted effect sizes with weights $w_1=1$ and $w_2=1$ reduced to −0.13 standard deviations for diarrhea, −0.02 standard deviations for fatigue, −0.11 standard deviations for stomatitis and −0.03 standard deviations for any grade 3 or 4 toxicity (Table 1). These quality-adjusted effect sizes correspond to a reduction in the median overall survival difference from 0.33 months to quality-adjusted survival differences of −1.13 months for diarrhea, −0.20 months for fatigue, −0.94 months for stomatitis and −0.27 months for any grade 3 or 4 toxicities. A negative difference in quality-adjusted survival time or effect size represents a benefit in the control arm compared to the treatment arm. A secondary weighting scheme of $w_1 + w_2=1$ was applied to the quality-adjusted Effect Size in 0.2 increments. The results are shown in Figure 2. A survival weight equal to one, $w_1=1$, corresponds to a toxicity weight equal to zero and an adjusted median survival difference equal to the original difference in median overall survival. The adjusted median survival difference is negative for the three adverse events beginning at $w_1=0$ and require a large weight (greater than 60%) on survival to produce a positive survival difference. A positive

difference in survival occurs at about w_1=0.7 for fatigue and any grade 3/4 toxicity and at w_1=0.9 for stomatitis and diarrhea.

Case 4. Significant survival benefit and synergistic toxicity comparison (+/+)

The Clinical Evaluation of Pertuzumab and Trastuzumab (CLEOPATRA) study was designed to assess the efficacy and safety of pertuzumab plus trastuzumab plus docetaxel, as compared with placebo plus trastuzumab plus docetaxel, as first-line treatment for patients with HER2-positive metastatic breast cancer [15]. The median overall survival in the control group is 18.5 versus 12.4 months in the control group (hazard ratio for death 0.62; $p<0.001$). This corresponds to a difference of 6.1 months in overall survival between the two arms and an effect size of 0.34 standard deviations. The adverse events that were considered were diarrhea, neutropenia and fatigue. The percentages were greater in the treatment group for diarrhea (66.8% versus 46.3%), and marginally larger for neutropenia (52.8% versus 49.6%) and fatigue (37.6% versus 36.8%). The quality-adjusted effect sizes with weights w_1=1 and w_2=1 reduced to −0.04 standard deviations for diarrhea, 0.14 standard deviations for neutropenia and 0.16 standard deviations for fatigue (Table 1). These quality-adjusted effect sizes correspond to a reduction in the median overall survival difference from 6.1 months to −0.63 months for diarrhea, 2.48 months for neutropenia, and 2.9 months for fatigue. A secondary weighting scheme of $w_1 + w_2$=1 was applied to the Quality-adjusted Effect Size in 0.2 increments. The results are shown in Figure 2. A survival weight equal to one, w_1=1, corresponds to a toxicity weight equal to zero and an adjusted median survival difference equal to the original difference in median overall survival. The adjusted median survival difference is negative for the three adverse events beginning at w_1=0 and a positive difference in survival occurs at about w_1=0.1 for neutropenia and fatigue i.e. a 90% weight on toxicity is required to produce negative difference in the quality-adjusted survival. However, the weight on survival must be greater than 0.5 for a positive survival difference in quality-adjusted survival when considering diarrhea as an adverse event.

DISCUSSION

The impact of experimental treatments on the human experience is not restricted to a simple prolongation of life (or lack thereof), but rather is a combination of quantity and quality of life. It is difficult however to distill the impact on both quantity and quality of life into a single summary statistic. The QASES method demonstrated in this manuscript attempts to overcome this challenge by transforming survival and toxicity comparisons into a single metric by expressing results in terms of a quality-adjusted survival effect size described in terms of standard deviations. This method allows for calibration of individual adverse events and survival effect sizes as well as overall toxicity combined with survival. The QASES metric has the potential to facilitate more effective communication between patients and clinicians concerning what is best for the patients in terms of an overall patient-relevant outcome, by combining both survival and quality of life. The method allows for patients to incorporate their personal views on the relative weighting they ascribe to quantity and quality of life with reference to treatment choices, and using this approach, physicians can provide information to the patients that they need to make these assessments.

One potential future application of this method in the clinic would be to develop this method further, obtain patient ratings of toxicities/harms against benefits, provide estimates for an average patient, comparing various treatment regimens by this single statistic. We envision that with simplified graphic displays on touchscreens developed using existing data from various treatments, where patient can input their values/preferences regarding the importance of various harms/benefits to them, it can produce an individualized statistic/s for patients to help them make personalized decisions. Such decisions will be based on more explicit information and will likely align better with their preferences, values and goals. Health care providers can provide information to the patients that they need to make these assessments.

One advantage of the QASES method is the simplicity/generalizability of the approach, which only requires summary statistics for survival and toxicity results for a given clinical trial. While all of the examples in this paper are in the realm of cancer, the methods apply equally as well to other diseases. Further work is underway to facilitate the presentation of the QASES results in real time in a clinical setting via computerized decision-making applications.

As with any statistic, there are limitations with the QASES method. For example, in unscrupulous hands, the method could be misapplied by only including toxicities that favor a given treatment. Further, in the case where a treatment imparts a non-significant overall survival effect but an improved toxicity profile, the resultant increase in quality-adjusted survival is not actually going to result in living longer, but better, and it will be important to make that distinction. This, in patient terms, may be described as a “fuller (more satisfying) life” rather than a “longer life”. Similarly, since this method is based on group effect sizes, it cannot precisely reflect each individual patient’s specific likelihood of success given certain treatment options. The QASES method should not be used without considering all aspects of care and well-being such as economic, social, and regulatory variables.

The groundbreaking work of Gelber and colleagues pioneered the combination of survival and toxicity data and involves much more complicated modeling and estimation procedures [16]. The QASES method is not intended to replace or disrespect that exemplary body of work. In fact, the QASES method was developed as a result of findings obtained by applying the original Gelber QALY model to the first example study in this manuscript [2]. That work led us to uncover parametric calibrations that would be necessary to exist to optimally apply the Gelber model [9]. The QASES is a simple metric that is intended as a potential alternative to the more complicated Gelber model without having to delve into strong distributional assumptions. Both techniques aspire to the larger objective of effectively calibrating, interpreting and communicating results of clinical trials to both clinicians and patients.

The QASES method is also intended to compliment clinical judgment, not replace it. In response to ASCO’s recently published estimates for clinically meaningful benchmarks for treatment efficacy [17], we applied the QASES method to compare the clinical subjective with the statistical objective estimates of treatment efficacy [9]. Our results indicated that the clinician benchmarks were expressing the belief that even statistically small effect sizes have

a place in defining success for new cancer treatments. We further identified that the estimates for colorectal cancer were notably larger than for other diseases. It is in this complimentary manner that the QASES method can be used to calibrate, compare and interpret various approaches estimate the effect of anti-cancer treatments.

Acknowledgments

FUNDING SUPPORT

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Numbers U10CA180882 (to the Alliance Statistics and Data Center), UG1CA189823 and U10CA180821 (to the Alliance for Clinical Trials in Oncology), U10CA025224, U10CA037404, U10CA035431, U10CA035415, U10CA035103, and U10CA035269. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Dr. Singh is also supported by the resources and the use of facilities at the VA Medical Center at Birmingham, Alabama, USA.

FINANCIAL CONFLICT

JAS has received research grants from Takeda and Savient and consultant fees from Savient, Takeda, Regeneron, Merz, Iroko, Bioiberica, Crealta/Horizon and Allergan pharmaceuticals, WebMD, UBM LLC and the American College of Rheumatology. JAS serves as the principal investigator for an investigator-initiated study funded by Horizon pharmaceuticals through a grant to DINORA, Inc., a 501 (c)(3) entity. JAS is a member of the executive of OMERACT, an organization that develops outcome measures in rheumatology and receives arms-length funding from 36 companies; a member of the American College of Rheumatology's (ACR) Annual Meeting Planning Committee (AMPC); Chair of the ACR Meet-the-Professor, Workshop and Study Group Subcommittee; and a member of the Veterans Affairs Rheumatology Field Advisory Committee. JAS is the editor and the Director of the UAB Cochrane Musculoskeletal Group Satellite Center on Network Meta-analysis. Other authors declare no conflicts.

APPENDIX 1

Mathematical Underpinnings for the Survival Endpoints Effect Size

Distribution

Assume the survival time, x, follows an exponential distribution of $f(x) = \frac{1}{t}e^{\frac{x}{t}}$ where $x \geq 1$ and t=mean overall survival time. Then it follows directly that $E(x) = t$, $Var(x) = t^2, Sd(x) = t$ and finally that $\frac{1}{2}Sd(x) = \frac{t}{2}$. Given that t=median overall survival time, then $Sd(x) = \frac{t}{\ln 2}$ *and* $\frac{1}{2}Sd(x) = \frac{t}{2\ln 2}$.

Effect Size

The calibrated effect size is the difference in overall survival between the two arms (i.e. treatment arm OS – reference arm OS) divided by the standard deviation of the reference arm survival time.

Example

Consider the following illustration. 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). The difference in survival times between the arms can be calibrated to 0.36 standard deviations ((14.5–9.5)/(9.5/ln2)). This is a small/medium

effect size according to the Cohen guidelines of small (0.2), medium (0.5), and large (0.8) effect sizes (Cohen, 1988).

Mathematical Underpinnings for the Toxicity Effect Size

Distribution

Assume the amount of toxicity in a clinical trial follows a Binomial distribution with parameters n and p where p is probability of a toxic event. Then it follows directly that $E(x) = p,\ Var(x) = p(1-p), Sd(x) = \sqrt{p(1-p)}$ and finally that $\frac{1}{2}Sd(x) = \frac{1}{2}\sqrt{p(1-p)}$. The sample proportion, $\hat{p} = \frac{x}{n}$ can be used as an unbiased estimate of p.

Effect Size

The calibrated effect size for the adverse events is the difference in toxicity between the two arms (i.e. toxicity in treatment- toxicity in control) divided by the standard deviation of the reference group.

Additional examples are provided in S1 for survival and S2 for toxicity below.

S1:

½ Standard Deviation of Median Survival Time

SD Difference for Median Survival Analysis	Median survival = 6 months	Median survival = 1 year
SD	8.7 months	17.3 months
½ SD	4.3 months	8.7 months
¼ SD	2.2 months	4.3 months
1/5 SD	1.7 months	3.5 months

S2:

½ Standard Deviation of Toxicity

SD Difference for Adverse Event Rates	Adverse Event Rate = 50%	Adverse Event Rate = 25%
SD	8.7 months	17.3 months
½ SD	4.3 months	8.7 months
¼ SD	2.2 months	4.3 months
1/5 SD	1.7 months	3.5 months

REFERENCES

[1]. Braun MS, Seymour MT. Balancing the efficacy and toxicity of chemotherapy in colorectal cancer. Ther Adv Med Oncol 2011; 3(1): 43–52. 10.1177/1758834010388342 [PubMed: 21789155]

[2]. Gelber RD, Gelman RS, Goldhirsch A. A quality-of-life-oriented endpoint for comparing therapies. Biometrics 1989; 45(3): 781–95. 10.2307/2531683 [PubMed: 2790121]

[3]. Sloan JA, Sargent DJ, Lindman J, et al. A new graphic for quality adjusted life years (Q-TWiST) survival analysis: the Q-TWiST plot. Qual Life Res 2002; 11(1): 37–45. 10.1023/A:1014401516011 [PubMed: 12003054]

[4]. Cole BF, Gelber RD, Gelber S, et al. A quality-adjusted survival (Q-TWiST) model for evaluating treatments for advanced stage cancer. Journal of Biopharmaceutical Statistics 2004; 14(1): 111–24. 10.1081/BIP-120028509 [PubMed: 15027503]

[5]. Revicki DA, Feeny D, Hunt TL, et al. Analyzing oncology clinical trial data using the Q-TWiST method: clinical importance and sources for health state preference data. Qual Life Res 2006; 15(3): 411–23. 10.1007/s11136-005-1579-7 [PubMed: 16547779]

[6]. Sloan JA, Bonner JA, Hillman SL, et al. A quality-adjusted reanalysis of a Phase III trial comparing once-daily thoracic radiation vs. twice-daily thoracic radiation in patients with limited-stage small-cell lung cancer (1). Int J Radiat Oncol Biol Phys 2002; 52(2): 371–81. 10.1016/S0360-3016(01)01819-3 [PubMed: 11872282]

[7]. Radice D, Redaelli A. Q-TWiST analysis of cyclophosphamide, epirubicin, fluorouracil versus cyclophosphamide, methotrexate, fluorouracil treatment for premenopausal women with node-positive breast cancer. Pharmaco Economics 2005; 23(1): 69–75. 10.2165/00019053-200523010-00006

[8]. Tate WR, Skrepnek GH. Quality-adjusted time without symptoms or toxicity (Q-TWiST): patient-reported outcome or mathematical model? A systematic review in cancer. Psychooncology 2014.

[9]. Sloan JA, Sargent DJ, Novotny PJ, et al. Calibration of quality-adjusted life years for oncology clinical trials. J Pain Symptom Manage 2014; 47(6): 1091–99 e3. [PubMed: 24246787]

[10]. Jeff A Sloan DV-C, Kamath Celia C, Sargent Daniel J, Novotny Paul, Atherton Pamela, Allmer C, Fridley BL, Frost Marlene, Loprinzi CL. Detecting worms, ducks, and elephants: A simple approach for defining clinically relevant effects in quality-of-life measures. Journal of Cancer Integrative Medicine 2003; 1(1): 41–47.

[11]. Sloan JA, Major B. Combining survival and toxicity effect sizes from clinical trials into an interpretable,quality-adjusted survival effect size estimate of treatment efficacy [abstract]. Poster presentation. (2014 6 11) J Clin Oncol 2014; 32(15_suppl): 6630.

[12]. Bonner J, Sloan J, Shanahan T. Phase III comparison of twice-daily compared with once-daily thoracic radiotherapy in limited stage small-cell lung carcinoma. Journal of Clinical Oncology 1999; 17: 2681–91. 10.1200/JCO.1999.17.9.2681 [PubMed: 10561342]

[13]. Ajani JA, Rodriguez W, Bodoky G, et al. Multicenter Phase III Comparison of Cisplatin/S-1 With Cisplatin/Infusional Fluorouracil in Advanced Gastric or Gastroesophageal Adenocarcinoma Study: The FLAGS Trial. Journal of Clinical Oncology 2010; 28(9): 1547–53. 10.1200/JCO.2009.25.4706 [PubMed: 20159816]

[14]. Moore MJ, Goldstein D, Hamm J, et al. Erlotinib Plus Gemcitabine Compared With Gemcitabine Alone in Patients With Advanced Pancreatic Cancer: A Phase III Trial of the National Cancer Institute of Canada Clinical Trials Group. [16] Journal of Clinical Oncology 2007; 25(15): 1960–66. 10.1200/JCO.2006.07.9525

[15]. Baselga J, Cortes J, Kim S-B, et al. Pertuzumab plus [17] Trasruzumab plus Docetaxel for Metastatic Breast Cancer. New England Journal of Medicine 2012; 366: 109–19. 10.1056/NEJMoa1113216 [PubMed: 22149875]

[16]. Bowling A Research Methods in Health: Investigating Health and Health Services. Buckingham, UK: Open University Press 1997.

[17]. Ellis LM, Bernstein DS, Voest EE, et al. American Society of Clinical Oncology perspective: Raising the bar for clinical trials by defining clinically meaningful outcomes. J Clin Oncol 2014; 32(12): 1277–80. 10.1200/JCO.2013.53.8009 [PubMed: 24638016]

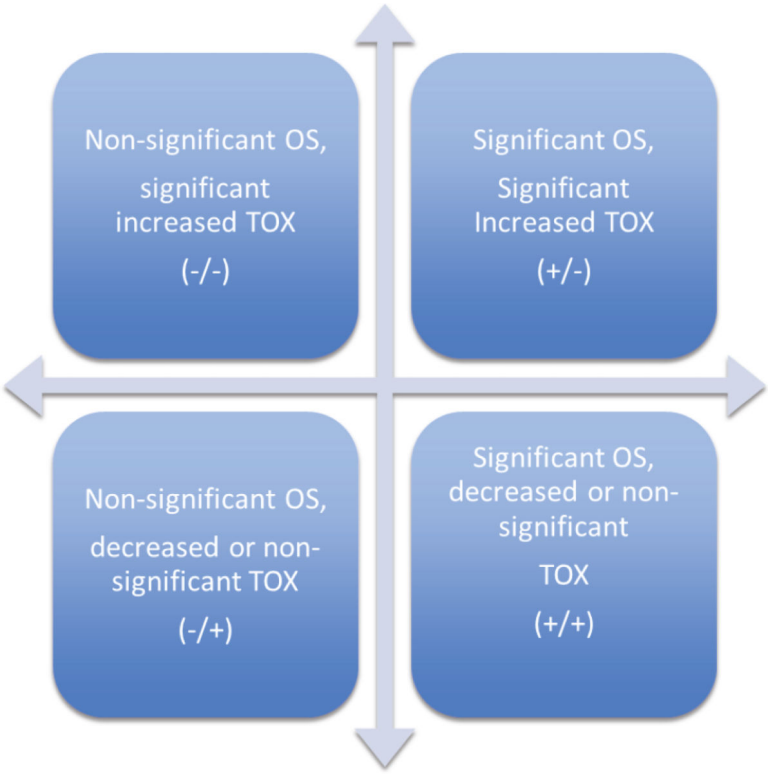

Figure 1:
Illustration of four possible combinations of overall comparative effectiveness of any two treatments for a given accounting for major benefit, survival (yes, no) and major harm, toxicity (yes, no).

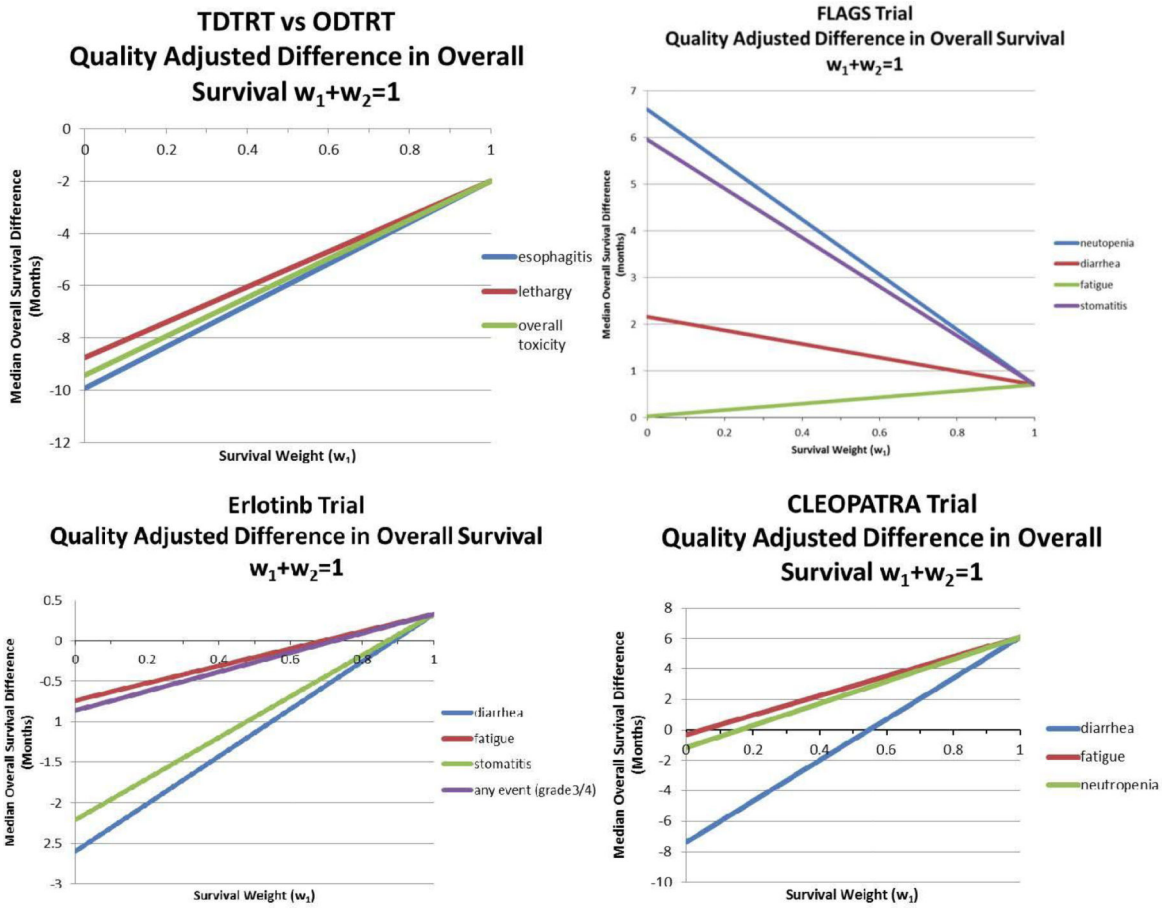

Figure 2:

Y-axis represents median survival difference between comparator treatments for the four possible case examples, indicating significant differences in survival (yes, no) and toxicity (yes, no). X-axis shows the relative weighting of survival (0–1) to toxicity (0–1). Each line with in each panel represents different QALY estimates for specific toxicity outcomes (e.g. nausea, fatigue etc.). Each line starts on the left weighing survival at 0 (w1=0) and toxicity at 1 (w2=1); at the right, weighing survival at 1 (w1=1) and toxicity at 0 (w2=0).

Table 1:

Case Examples for each of the Four Possible Combinations of Overall Comparative Effectiveness of Any Two Treatments for a Given Accounting for Major Benefit, Survival (Yes, No) and Major Harm, Toxicity (Yes, No)

Endpoint	Treatment	Control	Difference	Sd of Control	Effect Size (months/SD)	Quality-adjusted Effect Size
Case 1 example: NCCTG 89-20-52: TDRT vs. ODTRT [12]						
Median Overall Survival (months)	20	22	−2	31.74	−0.06	
Toxicity (proportion of people)						
Overall Toxicity (nonhematologic)	54%	39%	15%	49%	0.30	−0.18
Case 2 example FLAGS trial: S-1/cisplatin vs. cisplatin/infusional fluorouracil [13]						
Median Overall Survival (months)	8.6	7.9	0.7	11.40	0.06	
Toxicity (proportion of people)						
Neutropenia	61%	83%	−22%	38%	−0.58	0.32
Fatigue	39%	39%	−0.1%	49%	−0.002	0.03
Diarrhea	29%	38%	−9%	49%	−0.19	0.13
Stomatitis	6%	30%	−24%	46%	−0.52	0.29
Case 3 example: Erlotinib + Gemcitabine vs. Gemcitabine [14]						
Median Overall Survival (months)	6.24	5.91	0.33	8.53	0.04	
Toxicity (proportion of people)	56%	41%	15%	49%	0.30	−0.13
Fatigue	89%	86%	3%	35%	0.09	−0.02
Stomatitis	23%	14%	9%	35%	0.26	−0.11
Any Toxicity (grade 3/4)	62%	57%	5%	50%	0.10	−0.03
Case 4 example: CLEOPATRA Trial: Pertuzumab + Trastuzumab + Docetaxel vs. Trastuzumab + Docetaxel + Placebo [15]						
Median Overall Survival (months)	18.5	12.4	6.1	17.9	0.34	
Toxicity (proportion of people)	67%	46%	21%	50%	0.41	−0.04
Neutropenia	53%	50%	3%	50%	0.06	0.14
Fatigue	38%	37%	1%	48%	0.02	0.16

Sd, standard deviation.

Table 2:

Weighted Combination of the Two Effect Sizes (w1+w2=1) Related to Case Example 1 $\frac{w_1 ES_A - w_2 ES_B}{w_1 + w_2}$ = weighted combined ES = w_1 * (−0.06) - (1−w_1)*(0.30)

W_1	Combined Effect Size	Back Calculated Median Difference
0.0	−0.30	−9.42
0.1	−0.28	−8.68
0.2	−0.25	−7.93
0.3	−0.23	−7.19
0.4	−0.20	−6.45
0.5	−0.18	−5.71
0.6	−0.16	−4.97
0.7	−0.13	−4.23
0.8	−0.11	−3.48
0.9	−0.08	−2.74
1.0	−0.06	−2.00