Statistical Issues in Oncologic Clinical Drug Development

PhD Thesis by Published Works

Kevin J Carroll BSc, MSc, CStat, CSci, Honorary Senior Lecturer

October 2013

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there from must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Contents

14. Carroll KJ. Back to basics: explaining sample size in outcome trials, are statisticians doing a thorough job? Pharmaceutical Statistics, 2009; 8: 333–345. 236

15. Ellis S, Carroll KJ and Pemberton K. Analysis of Duration of Response in Oncology Trials. Contemporary Clinical Trials, 2008; 29: 456–465 250

Introduction

Statistical science is central to the process of pharmaceutical clinical drug development. New medicines have to be evaluated in a series of clinical trials in humans over many years for which the experimental design and data analysis is critical, requiring expert statistical input and knowledge. Traditionally trials are classified in relation to the phase of development: Phase I denotes the first trials in human volunteers and patients looking at parameters such as drug distribution, metabolism and excretion; Phase II denotes clinical 'proof-of-concept' and dose ranging trials where the aim is to establish probable efficacy of the new drug; Phase III denotes the confirmatory or 'pivotal' testing phase where the hope is to establish efficacy in controlled trials, where patients are randomized to the experimental drug or control, often but not always placebo, typically in a double-blinded manner (where neither the physician or the patient is aware of treatment assignment). It is Phase III trials that form the basis of the drug manufacturer's application for licensure of the new drug to the prevailing regulatory authorities; and Phase IV denotes post licensure trials to characterise further the safety of the drug, explore new medical uses and to provide any additional information as required by the licensing authorities. As such, it is clear that drug development is heavily regulated worldwide. This regulation includes complex statistical guidance governing trial design and analysis, further necessitating expert statistical input in the drug development process.

Over the past decade, drug development has become increasingly challenging. At an estimated cost of $0.8-2b, developing a drug has never been more expensive (Chuang-Stein et al 2011). Yet, despite increased expenditure, overall pharmaceutical research and development productivity remains low. Failure rates of 80% in Phase II and 50% in Phase III have been reported (Arrowsmith 2011a-b). Two thirds of Phase III failures are reported as due to not demonstrating a positive treatment effect reflecting poorly on the quality of Phase II design and decision making (Arrowsmith 2011b). And with high profile drug use withdrawals over recent years such as Vioxx, Avandia, Xigirs and Acomplia, regulators are demanding

ever more data to support drug approval, post-approval safety assessment and reimbursement (EMA 2009a and 2011, MHRA 2004 and 2010, FDA 2004a, Lilly 2011).

This difficult environment has stimulated clinical researchers and statisticians to consider alternative approaches and flexible designs, particularly in early development, resulting in a jump in the associated literature. Oncologic drug development has been the focus of much of this attention; medical need is high and improvements in molecular diagnosis and the associated potential for patient selection offer the opportunity to deliver targeted drugs to specific patient populations with breakthrough improvements in effectiveness together with a compelling benefit:risk balance. Statistical developments have included the application of flexible designs such as PII/PIII 'seamless' designs, novel patient selection strategies as well as analytical methods to evaluate the value of surrogate endpoints and approaches to cope with issues in survival analysis such as 'crossover' from control to active treatment upon disease progression (Williams et al 2002, Rimawi and Hilsenbeck 2012). At the same time there has been a surge in the literature regarding the use of Bayesian methodologies in drug development, particularly in relation to adaptive designs and in the area of health technology assessment with approaches such as network meta-analyses (Lumley 2002, Sutton et al 2008, Lu et al 2004, Caldwell et al 2005, Edwards et al 2009, Jones et al 2011).

Despite these advances, many statistical issues persist in the clinical development of oncologic, and other, therapeutics. Several of the key issues have been discussed by Carroll in a series of peer reviewed publications over past years. These publications are provided in the Appendix. The aim of this covering chapter is to collate and critique these publications, demonstrating how they form a coherent, related body of statistical work and, in so doing, provide recommendations for the future clinical development of oncologic medicines. These issues and challenges include the following:

1.1 The use of biomarkers and surrogate endpoints to accelerate oncologic drug development

With the advent of ever more sophisticated proteomic, genomic and genetic technologies, efforts to gain a more in-depth biologic understanding of disease, particularly in oncology, is leading to the discovery of multiple new biomarkers that may reflect underlying disease processes. Such biomarkers are frequently considered as vital patient selection tools that will help to identify those most likely to benefit from a new drug, and, in so doing, will reduce costs, lessen risk and shorten development times. Herceptin (trastuzumab), Gleevec (imatinib mesylate) and Iressa (gefitinib), are often cited biomarker-led development successes others are encouraged to emulate (EMA 2009b, FDA 2010a and 2013a). It is further hoped that biomarkers can be used as surrogate endpoints in the regulatory drug approval process and therefore provide a substitute for clinical outcomes, accelerating the availability of new medicines. However, many statistical issues remain such that biomarker strategies may not in all cases deliver the advantages hoped for. Supporting publications are:

1. Newling D, **Carroll K** and Morris T. Is prostate-specific antigen progression a surrogate for objective clinical progression in early prostate cancer? *Journal of Clinical Oncology*, ASCO Annual Meeting Proceedings (Post-Meeting Edition). Vol 22, No 14S (July 15 Supplement), 2004: 4652.
2. Collette L, Burzykowski T, **Carroll K**, Newling D, Morris T and Schroder F. Is prostate-specific antigen a surrogate for survival in advanced prostate cancer? *Journal of Clinical Oncology*, 2004 ASCO Annual Meeting Proceedings (Post-Meeting Edition). Vol 22, No 14S (July 15 Supplement), 2004: 4551.
3. Collette L, Burzykowski T, **Carroll KJ** et al. Is prostate-specific antigen a valid surrogate end point for survival in hormonally treated patients with metastatic prostate cancer? Joint Research of the European Organisation for Research and Treatment of Cancer, the Limburgs Universitair Centrum, and AstraZeneca Pharmaceuticals. 2005, *Journal of Clinical Oncology*; 23:6139–6148.
4. Buyse M, Burzykowski T, **Carroll K** et al. Progression-Free Survival Is a Surrogate for Survival in Advanced Colorectal Cancer. *Journal of Clinical Oncology*, 2007; 25:5218-5224.

5. **Carroll KJ**. Biomarkers in drug development: friend or foe? A personal reflection gained working within oncology**.** *Pharmaceutical Statistics,* 2007; 6(4): 253–260.

1.2 Randomized Phase II designs: the use of progression free survival as an endpoint and decision making from Phase II to Phase III.

Historically oncology Phase II studies have been relatively small, single arm response rate driven studies (Simon 1989). This reflected that, in the past, treatments for disease were invariably cytotoxics designed to kill cells so that tumor shrinkage was viewed as a direct measure of drug effectiveness. However, newer molecularly targeted treatments, such as the EGFR, VEGF and mTOR inhibitors as well as new biologic monoclonal anti-bodies such as bevacuzimab, panitumumab and ipilimumab, have a different, cytostatic mode of action meaning tumor shrinkage is no longer considered an appropriate endpoint (Korn 2001 et al, Stadler and Ratain 2000, Eskens et al 2000, Simon et al 2001). Rather, the period of time alive and free from growth and spread of disease, typically referred to as progression free survival (PFS), is a more appropriate endpoint to test these modern cytostatics and should be subject to fewer Type II errors than would use of standard tumor shrinkage. The use of PFS in Phase II oncology has given rise to larger and longer randomised, controlled trials which have served to raise a number of important statistical issues. These issues relate mainly to the perceived prohibitive size of Phase II when using a PFS endpoint and the potential for bias when using progression as an endpoint, and further give rise to the related question as to how best to use more informative randomised Phase II data to better predict the chance of success in subsequent Phase III trials. Supporting publications are:

6. Stone A, Wheeler C, **Carroll K** and Barge A. Optimizing randomized phase II trials assessing tumor progression. *Contemporary Clinical Trials,* 2007; 28(2):146-52.
7. **Carroll KJ**. Analysis of progression-free survival in oncology trials: some common statistical issues*. Pharmaceutical Statistics,* 2007; Vol 6(2): 99-113.
8. **Carroll KJ.** Decision Making from Phase II to Phase III and the Probability of Success: Reassured by 'Assurance'? *Journal of Biopharmaceutical Statistics,* 2013; Vol 23(5):1188-1200.

1.3 Non-inferiority trial design and analysis.

Several regulatory guidelines have been developed over past years to govern the various facets of active control, non-inferiority (AC, NI) trial design and analysis (FDA 2010b; EMA 2000 and 2005). Despite this, much statistical discussion and debate remains regarding the true nature of 'NI' assessment and the associated feasibility of 'NI' trial design given the traditionally conservative approaches employed by regulators, in particular the FDA. Issues of assay sensitivity and constancy are well known, but more fundamental is the question of what, primarily, is the statistical goal of an AC, NI trial and whether the hurdles imposed by the regulators in NI assessment represent an arbitrarily higher standard for drug approval based merely on trial design alone. There are also illogicalities associated with standard 'fixed' margin and 'percent preservation of effect' approaches to NI assessment which need to be highlighted in addition to the relative efficiency of these two approaches as compared to the third most commonly used approach, the so-called 'synthesis' method. These matters relating to AC, NI trial design and analysis are of relevance across all therapeutic areas, though particularly in oncology where gold standard randomised placebo control trials are often seen as unethical, necessitating the use of AC trials. Supporting publications are:

9. **Carroll K** and Milsted R. Barriers to clinical development in oncology: The impact of new thinking around non-inferiority. *Journal of Clinical Oncology*, ASCO Annual Meeting Proceedings (Post-Meeting Edition). Vol 22, No 14S (July 15 Supplement), 2004: 6082.
10. **Carroll K**, Milsted B and Lewis JA. Design and analysis of non-inferiority mortality trials in oncology. Letter to the Editor. *Statistics in Medicine* 2004, Vol 23(17): 2771-2774.
11. **Carroll KJ**. Active-controlled, non-inferiority trials in oncology: arbitrary limits, infeasible sample sizes and uninformative data analysis. Is there another way? *Pharmaceutical Statistics*, 2006; Vol 5(4): 283-293.
12. **Carroll KJ**. Statistical issues and controversies in active-controlled, 'non-inferiority' trials. Statistics in Biopharmaceutical Research, 2013. *Statistics in Biopharmaceutical Research* 2013; 5:3, 229-238.

1.4 The use of parametric methods in the analysis of oncology clinical trial data.

With the publication of Cox's partial likelihood approach in 1972, Cox proportional hazards regression and the associated log rank test have become the mainstay of statistical testing for time to event data in oncology and beyond (Cox 1972). Given its widespread use and acceptance across the clinical research spectrum, statisticians and researchers seldom challenge its dominance or consider other parametric alternatives. Yet, parametric methods, such as Weibull modelling, can offer greater flexibility and breadth of insight whilst still delivering core results very similar to their non-parametric counterpart. For Weibull modelling, this is the case even when the underlying distribution of survival times is known not to be truly Weibull. Further, parametric approaches can be helpful in tackling common problems such as the analysis of duration of response to provide an unbiased, unconditional evaluation based on all randomized patients rather than a non-randomized comparison in the subset of patients who responded, which is the typical approach in oncology trials despite this being in contradiction of the European Medicines Agency (EMA) anti-cancer guideline (EMA 2012a).

Supporting publications are:

13. **Carroll KJ**. On the use and utility of the Weibull model in the analysis of survival data. *Controlled Clinical Trials* 2003; 24: 682–701.
14. **Carroll KJ**. Back to basics: explaining sample size in outcome trials, are statisticians doing a thorough job? *Pharmaceutical Statistics,* 2009; 8: 333–345.
15. Ellis S, **Carroll KJ** and Pemberton K. Analysis of Duration of Response in Oncology Trials. *Contemporary Clinical Trials*, 2008; 29: 456–465.

The remainder of this Summary Chapter is structured as follows; Section 2 discusses publications relating to surrogate endpoint and biomarkers; Section 3 discusses publications regarding issues with PFS as an endpoint in Phase II trials and the use of such data in predicting the chance of success in Phase III; Section 4 discusses papers relating to active-control, 'non-inferiority' trials; and Section 5 briefly discusses papers on parametric models in the analysis of time to event data.

Section 6 then closes with a collation of the recommendations made for future oncology trial design.

2 The use of biomarkers and surrogate endpoints to accelerate oncologic drug development

2.1 Background

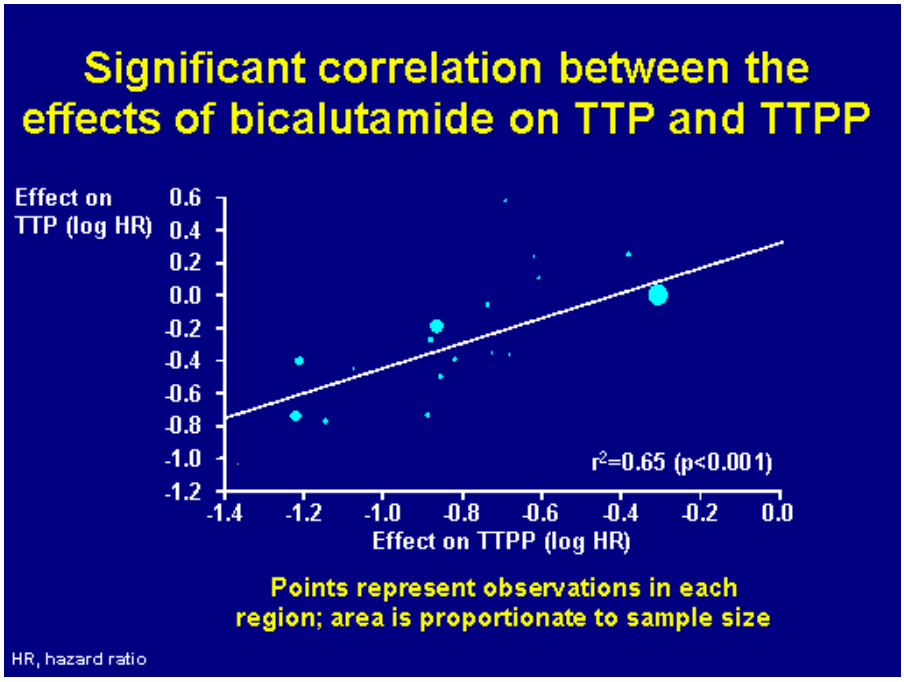
In oncology the gold standard clinical endpoint remains overall mortality. This can typically result in long term trials, especially in adjuvant settings (i.e. where treatment is given as an adjunct to therapy of curative intent) and slow growing malignances such as prostate cancer. As newer more effective second-line treatments become available, the utility of survival as means to assess drug effects in the first-line setting has become increasingly difficult across a range of malignancies, leading to an increasing focus on alternative biomarker and clinical endpoints such as progression-free survival to play the role of surrogates. Statistical evaluation of potential biomarkers and surrogate endpoints has been an important consideration for drug development over recent years. Prentice (1989) and Freedman et al (1992) were the first to lay down formal theoretical criteria to assess endpoint surrogacy. Subsequently, Fleming and DeMets (1996) wrote an important article regarding the use of assumed surrogate endpoints in the drug approval process and the high risk of drawing inappropriate conclusions. While the criteria forwarded by Prentice offer a technically complete description of the properties required to fulfil surrogacy, these criteria can prove difficult to apply in practice. For example, the requirement to prove the null (i.e. that the treatment effect on the true endpoint was reduced to zero conditional upon the presence of the surrogate) is, strictly speaking, unachievable. More recently Molenberghs et al (2001, 2002, 2004, 2010), Buyse (2009) and Buyse et al (1998, 2000, 2010) have taken a different approach, offering pragmatic meta-analytic methods to quantify the relationship between treatment effects on some purported surrogate and on the clinical endpoint itself across a set of relevant, randomized trials. This method yields measures of association between endpoints within the patient and, most importantly, between treatment effects across trials and, therefore, allows estimation of the 'surrogate threshold effect', i.e. the amount by which the surrogate would need to be impacted by treatment to result in a meaningful effect on the clinical endpoint itself (Burzykowski et al 2006).

2.2 Publication 1:

Newling D, <u>Carroll K</u> and Morris T (2004). Is prostate-specific antigen progression a surrogate for objective clinical progression in early prostate cancer?

In prostate cancer, or in healthy men deemed at risk of developing prostate cancer, prostate specific antigen (PSA) is a well-accepted and extensively used plasma biomarker of disease status. Both clinicians and patients have used PSA for more than 20 years to monitor disease and even small movements in PSA level over relatively short time periods can result in the application of treatments such as local radiotherapy, surgery or systemic treatment with medical castration or oral anti-androgen therapy. However, despite its widespread use, the utility of PSA as a surrogate endpoint for clinical outcome was not formally assessed until 2004. Access to clinical trial data in over 8,000 prostate cancer patients from the bicalutamide (an oral, once daily, non-steroidal anti-androgen) Early Prostate Cancer Programme allowed, for the first time, thorough statistical evaluation of PSA as a potential surrogate for longer term objective, radiologic progressive disease (ORPD) on CT or MRI scan (Wirth et al 2008). Using the Buyse and Molenberghs approach, this work was published in abstract form by Newling, Carroll and Morris at the American Society of Clinical Oncology (ASCO) 2004. A summary slide from the presentation is reproduced below. By examining the data in distinct regions, it was determined that while the R^2 between treatment effects on PSA and ORPD was either 0.52 ($p<0.001$) or 0.65 ($p<0.001$) (the R^2 value being dependent on the inclusion of a potentially influential data point), nevertheless a large treatment effect on PSA, i.e. a hazard ratio (HR) in the region of 0.50, was required to predict a modest effect on ORPD, i.e., a HR in the region 0.80 to 0.85.

Publication 1. Newling, Carroll, Morris (2004)



2.3 Publications 2 and 3:

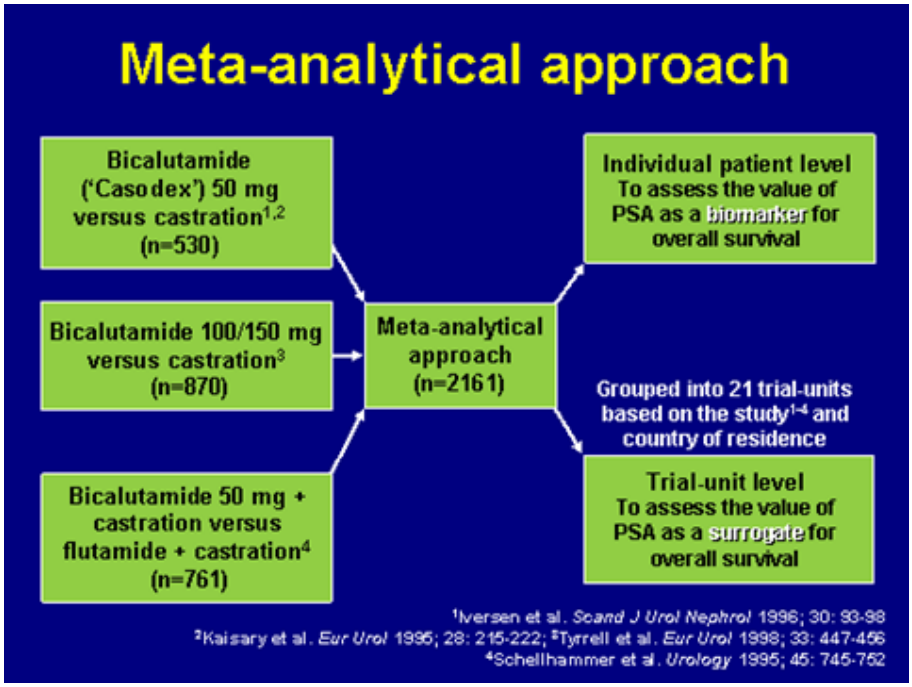
Collette L, Burzykowski T, Carroll K et al (2004). Is prostate-specific antigen a surrogate for survival in advanced prostate cancer?

Collette L, Burzykowski T, Carroll KJ et al (2005). Is prostate-specific antigen a valid surrogate end point for survival in hormonally treated patients with metastatic prostate cancer?

This work led to research and collaboration with the European Organisation for Research and Treatment of Cancer (EORTC) and the Limburgs Universitair Centrum Belgium regarding the utility of PSA progression as a surrogate endpoint for overall mortality in advanced prostate cancer. Findings were published by Collette, Burzykowski, Carroll et al initially in abstract form at the 2004 ASCO meeting and subsequently in full in the *Journal of clinical Oncology* in 2005. Data from over 2000 patients across 5 randomized clinical trials were analyzed to produce estimates of patient-level and trial-level association. Results showed that while, at the patient-level, PSA tracked reasonably well with overall survival suggesting some utility in the day-day clinical management of patients, it performed poorly at the trial-level

in terms of a true surrogate endpoint for the effect of bicalutamide treatment on mortality in advanced prostate cancer (PCa) patients. Table 5 and Fig 4B are from the 2005 publication are reproduced below, along with a summary slide from the 2004 ASCO Annual Meeting Proceedings:

Publication 2. Collette, Burzykowski, Carroll (2004)



Publication 3. Collette, Burzykowski, Carroll (2005): Figure 4B

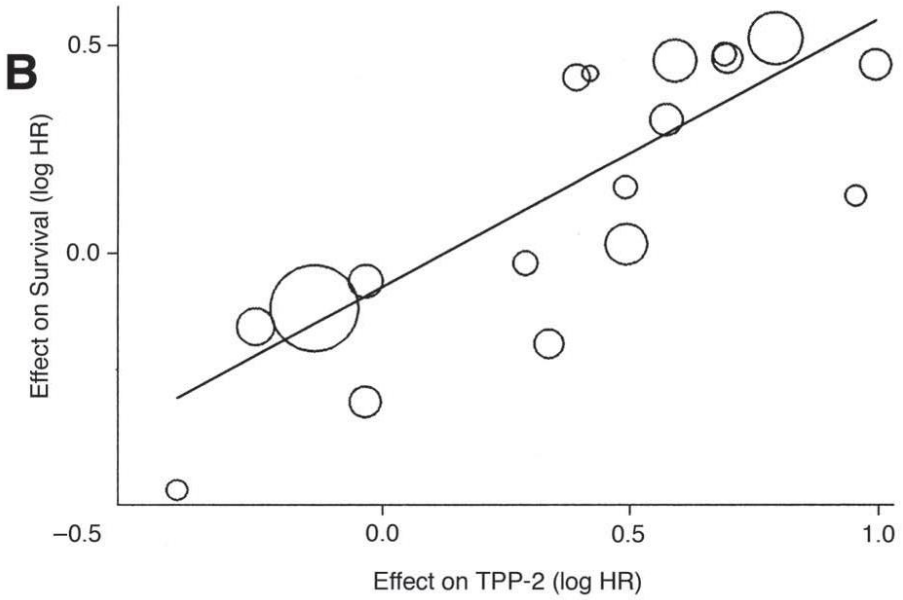


Fig 4. The treatment effects on time to prostate-specific antigen (PSA) progression (TTPP) and on overall survival. The circles represent the observations in the trial units, and their size is proportionate to the trial-unit sample size. The line represents the prediction from an estimated (weighted) regression line. (A) TTPP-1: $R^2_{trial} = 0.21$. (B) TTPP-2: $R^2_{trial} = 0.66$. HR, hazard ratio.

TTPP-1: PSA value above normal (4 ng/mL), representing a first increase ≥20% above the nadir. TTPP-2: PSA value >2.5 times the normal range (10 ng/mL), representing a first increase ≥50% above the moving average (based on three consecutive measurements) nadir. This increase had to be either the last observed value or be sustained for at least 4 weeks.

Publication 3. Collette, Burzykowski, Carroll (2005): Table 5

Table 5. Summary of the Results

PSA End Point	Patient-Level Association Between PSA and Survival		Trial-Level Association Between PSA and Survival		
		SE	R^2_{trial}	SE	95% CI
PSA response (decline by ≥ 50% from baseline)	$\theta_{patient} = 1.94$	0.33	0.08	0.14	0 to 0.49
PSA normalization (≤ 4 ng/mL)	$\theta_{patient} = 4.90$	0.52	0.41	0.18	0.05 to 0.72
TTPP-1	$\tau_{patient} = 0.52$	0.004	0.21	0.17	0 to 0.56
TTPP-2	$\tau_{patient} = 0.61$	0.02	0.66	0.13	0.30 to 0.85
Longitudinal PSA measurements	$R^2_{patient} > 0.9$ at all times > 7 mo	—	0.68	0.12	Undetermined

Abbreviations: PSA, prostate-specific antigen; TTPP, time to PSA progression; $\theta_{patient}$, survival odds ratio; $\tau_{patient}$, concordance coefficient between time to PSA progression and duration of survival.

TTPP-1: PSA value above normal (4 ng/mL), representing a first increase ≥20% above the nadir. TTPP-2: PSA value >2.5 times the normal range (10 ng/mL), representing a first increase ≥50% above the moving average (based on three consecutive measurements) nadir. This increase had to be either the last observed value or be sustained for at least 4 weeks.

As can be seen, R^2_{Trial} was weak apart from TTPP-2 with a value of 0.66. Longitudinal measures also appear to have a better R^2_{Trial}, however the incompleteness of PSA measures over time make this measure somewhat unreliable.

These data, both in early and advanced disease settings, were presented at the US Food and Drug Administrations' Public Workshop on Clinical Trial Endpoints in prostate cancer drug development, June 21-22, 2004. It was concluded that, for clinical trials in early prostate cancer, some form of PSA based progression could be included as part of a composite disease progression endpoint based primarily on objective radiologic and imaging evidence; and, further, that such a composite endpoint could be used in support of regulatory approval.

Despite this successful contribution to the debate regarding PSA surrogacy in prostate cancer, the approach taken was narrow and, with additional data and analyses, could have been more informative. For example, research was limited in terms of the number of trials involving the anti-androgen bicalutamide such that the unit of analysis using the Molenberghs approach became region within trial. The extent to which this may have influenced the results is hard to judge given region (in terms of country) was pre-determined. The inclusion of more trial data would have served to strengthen results but no further trial data were available. Data on other prostate cancer drugs in the same (anti-androgen) class would have provided both greater confidence and more generalizable results; however, the analytic approach at the time required individual patient data which was not available to the authors. Improved meta-regression techniques available today may have made incorporation of published summary level data on other drugs more feasible.

2.4 Publication 4:

Buyse M, Burzykowski T, Carroll K et al (2007). Progression-Free Survival Is a Surrogate for Survival in Advanced Colorectal Cancer

Building further in this area, a novel presentation was made to the FDA's 2004 Advisory Committee Meeting panel regarding trial endpoints in colorectal cancer (CRCa) research in which the utility of PFS as a surrogate for overall mortality in first line disease was assessed based upon data from over 1200 patients in 3 randomised clinical trials comparing the novel chemotherapy treatment raltitrexed with standard flouracil (FU) therapy (FDA 2004b, 2004c). The analysis showed that approximately half of the treatment effect on survival was explained by the effect of treatment on PFS and that, given a PFS increase of 50%, survival would be expected to increase by 29% with 95% CI (13%, 48%). These data helped to support a positive vote from the Committee to allow the use of PFS as a primary endpoint in first-line CRCa trials. This work was subsequently followed by collaboration with Buyse and Burzykowski resulting in a JCO publication in 2007. In this work 10 first-line CRCa trials comparing FU + leucovorin to either FU alone or raltitrexed were used to characterize the relationship between PFS and overall survival and 3 separate trials, comparing FU + leucovorin with or without irinotecan or oxaliplatin treatment, were used for validation purposes. Results were very encouraging for the use of PFS as a surrogate for overall survival (OS). At the patient level, PFS and OS were highly correlated with R^2 coefficient 0.82 and, at the trial level; and treatment effects were also well correlated with coefficient 0.74, 95% CI (0.44, 1.03) after exclusion of one influential trial. The data also estimated the surrogate threshold effect on PFS to be hazard ratio of 0.77 to predict a non-zero beneficial effect on overall survival. Finally validation using 3 separate trials showed a reasonable alignment between observed and predicted treatment effects on OS. Figure 4 and Table 2 from the publication are reproduced below.

Publication 4. Buyse, Burzykowski, Carroll (2007): Figure 4

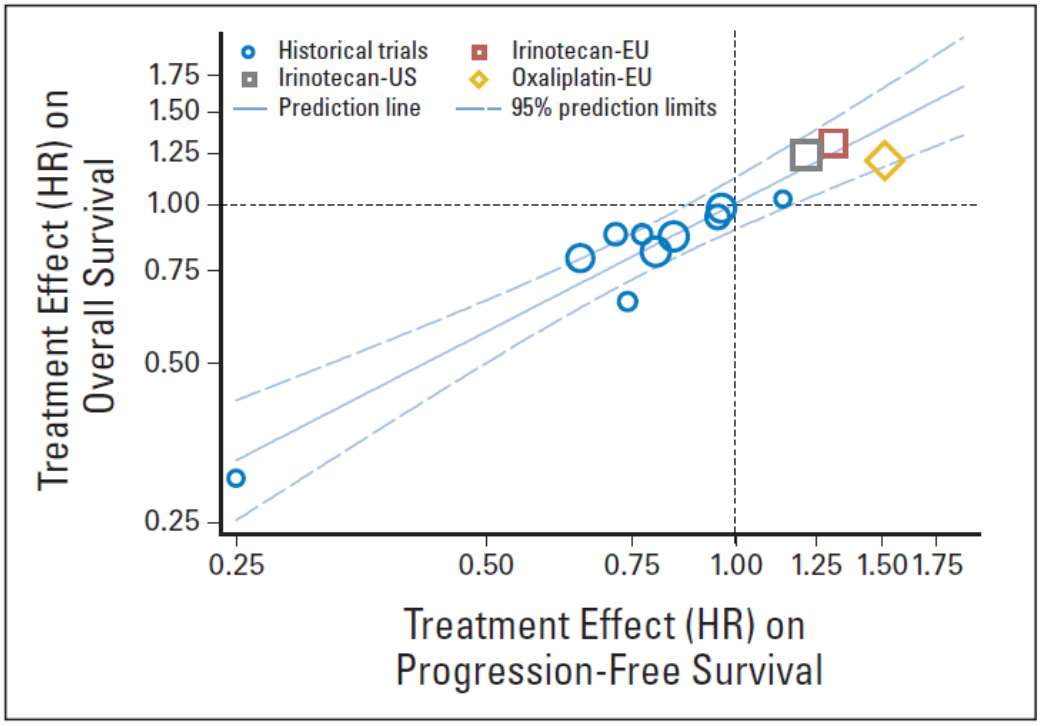


Fig 4. Correlation between treatment effects on progression-free and on overall survival in historical trials (circles), in irinotecan trials (squares), and in oxaliplatin trial (diamond). A logarithmic scale is used for both axes. Symbol size is proportional to the number of patients. HR, hazard ratio; EU, European Union.

Publication 4. Buyse, Burzykowski, Carroll (2007): Table 2

Table 2. Observed Versus Predicted OS HRs

Trial	Observed OS HR	95% CI	Predicted OS HR*	95% Predicted Interval	% FU + LV + New Drug: Total Receiving Second-Line Treatment	FU + LV + New Drug: Receiving New Drug	FU + LV + New Drug: Crossed Over to New Drug	FU + LV + New Drug: Other Second-Line Treatment	% FU + LV: Total Receiving Second-Line Treatment	FU + LV: Receiving New Drug	FU + LV: Crossed Over to New Drug	FU + LV: Other Second-Line Treatment
Irinotecan-EU[4]	1.31	1.02 to 1.67	1.25	1.00 to 1.55	39	0	16	23	58	31	13	14
Irinotecan-US[3]	1.24	1.00 to 1.53	1.17	0.96 to 1.43	52	0	5†	47	70	56	5†	9
Oxaliplatin-EU[5]	1.21	0.94 to 1.55	1.40	1.12 to 1.75	58	0	30	28	61	28	20‡	23

NOTE. HRs are HR of FU + LV *v* FU + LV + new drug.
Abbreviations: OS, overall survival; HR, hazard ratio; FU, fluouracil; LV, leucovorin; EU, European Union.
*Prediction based on observed progression-free survival HR.
†Reported as < 5%.
‡10% of patients received both oxaliplatin and irinotecan in second-line treatment and are counted only once in the overall 61%.

This body of work evaluating potential surrogate endpoints in PCa and CRCa sits within a broader framework of similar research using meta-analytic methodology to examine intermediate endpoints, predominantly in oncology. Examples include Sargent et al (2005) for 3 year disease-free survival as a surrogate for 5 years overall survival in adjuvant colorectal cancer; Tang et al (2007) for endpoints in colorectal cancer; Burzykowski et al (2008) for endpoints in advanced breast cancer; Saad et al (2010) for endpoints in breast cancer and colorectal cancer; Buyse et al (2011) for endpoints in leukaemia and Laporte et al (2013) for endpoints in non-small cell lung cancer (NSCLC). Taken together, this research has had a positive impact on how non-survival intermediate clinical endpoints are viewed and utilised in oncologic research, particularly by regulatory authorities who, in some instances (such as 1st line treatment in CRCa and NSCLC), have changed policy regarding the endpoints and evidence required to support drug approval. That said, the underlying Buyse and Molenberghs methodology relies upon multiple trials having been conducted in a given area, all of which must have collected data on both the intermediate and true clinical outcome. This clearly means that any new potential biomarker will take many years and many millions of dollars to assess and, hence, this approach does not readily serve to accelerate the drug development process. Further in all of the preceding referenced work, no account was taken of competing risks associated with long term follow-up of the surrogate and true clinical outcome; for example, patients can stop treatment due to an adverse event prior to meeting the surrogate and/or the true clinical outcome, or additional anti-cancer treatments can be given by the treating physician if considered medically indicated. In either case these intervening events represent competing risks since they can result in curtailed patient follow-up such that the surrogate and/or the true clinical outcome are censored and unobserved. The extent to which this issue may have affected published results and conclusions is unknown. The solution is not obvious, though pragmatically it would seem appropriate to mandate (by informed consent) full intent-to-treat (ITT) follow-up of all patients for overall mortality post attainment of the surrogate endpoint as a condition of acceptance of the surrogate in a regulatory context.

2.5 Publication 5:

Carroll KJ (2007). Biomarkers in drug development: friend or foe? A personal reflection gained working within oncology.

Following on from specific evaluation of surrogate endpoints in PCa and CRCa, the statistical and strategic usefulness of biomarkers and surrogate endpoints in oncologic drug development was reviewed more broadly in this paper. Fundamentally, the key goal researchers have in using a biomarker is to follow a targeted development strategy; to identify 'responsive' patients. However, this is an over simplification. Rather the goal should be to use the biomarker to identify patients 'more likely to benefit' from a given intervention. And as later reiterated by both Chakravarty et al (2011) and Fleming and Powers (2012), of primary interest are *predictive* biomarkers, i.e biomarkers that are 'effect modifiers', associated with a differential treatment effect, as opposed to biomarkers that are merely prognostic for the disease. In the paper, two critical underlying assumptions are identified that fundamentally determine usefulness of biomarker driven drug development strategies: (i) that biomarker positive patients experience a treatment effect while biomarker negative patients do not, and (ii) the biomarker diagnostic that determines whether a patient is 'positive' or 'negative' is perfect, with 100% sensitivity and specificity. Examination of these assumptions reveals that with relatively modest departures from the ideal, the strategic advantages of a targeted development are quickly eroded. Tables I-III from the paper are reproduced below.

Publication 5. Carroll (2007): Tables I to III

Table I. True median survival for new and control anti-cancer drugs.

	Median survival on control (months)	Median survival on new (months)	Treatment effect HR* new:control
Biomarker positive (25%)	6	12	0.50
Biomarker negative (75%)	6	6	1.00
All patients	6	7.5	0.80

*HR = hazard ratio.

Table II. The impact of an imperfect test.

Sensitivity (%), specificity (%)	PPV* (%)	Median survival on control (months)	Median survival on new (months)	HR new:control	Number required to enter	Number required to screen
100, 100	100	6	12	0.50	117	468
95, 75	56	6	9.4	0.64	260	613
75, 95	83	6	11	0.55	149	663
75, 75	50	6	9	0.68	317	845

*PPV = positive predictive value.

Table III. The impact of a small, non-zero effect in biomarker 'negative' patients.

	Median survival on control (months)	Median survival on new (months)	Treatment effect HR* new:control	Number of patients required (screened)
Biomarker positive (25%)	6	12	0.50	117 (468)
Biomarker negative (75%)	6	7.5	0.80*	—
All patients	6	8.7	0.69	384 (384)

*Effect size in 'negative' patients = 1/3 effect size in 'positive' patients.

Table I shows median survival times for drug and control in biomarker positive patients (25% of the overall trial population), biomarker negative patients and overall. These data feed into Table II which displays the number needed to screen (for the biomarker) and the number of patients required to be entered into the trial for the biomarker positive, negative and overall populations. A perfect test requires only 117 biomarker positive patients to be entered and 468 to be screened. However, as the specificity and sensitivity of the test falls, the numbers needed to enter and to screen quickly rise. Table III shows the impact of there being a small treatment effect in biomarker negative patients. With a treatment effect one third of that in biomarker positive patients, a trial in the overall population requires fewer patients to be entered than are required to screen in a biomarker positive trial.

With respect to utilizing biomarkers as true surrogate endpoints to support regulatory approval, it was argued expectations are, in general, too high. Longstanding clinical endpoints such as PFS are only just being accepted statistically as endpoints for drug approval on the basis of cumulative data from multiple, well-controlled randomised trials. Set against this background it was postulated as unlikely that new biomarkers would be quickly accepted as substitutes for clinical outcomes in drug evaluation. As described by Carroll et al (2008), perhaps the best that can be hoped for might be to utilise new routes for approval such as conditional approval (EU) and existing routes such as accelerated approval (FDA) on the basis of a biomarker endpoint, with a commitment to conduct further trials post-approval to confirm clinical benefit.

While the preceding publications served to identify issues in the use of surrogate endpoints and biomarkers in oncologic development, they did not offer alternative trial design solutions that might make biomarker driven developments more statistically appealing. Over past years, several important and related papers have appeared which attempt to do this. Zhao et al (2010) considers the design and analysis of trials with a biomarker 'sensitive' sub-population defined at

randomization, addressing issues relating to multiplicity, enrichment and optimality in terms of the best balance of power in the overall and biomarker defined subpopulations. In earlier work, Freidlin and Simon (2005) described an adaptive 'signature design' where biomarker or 'classifier' is derived in the first stage of a two stage study. The classifier is then applied to the second stage patients such that treatment effect is tested in the overall population as well as in the biomarker determined subgroup of second stage patients. Wang et al. (2007) proposed another adaptive design that enables the sponsor to restrict the enrolment of non-sensitive patients in the second stage. They also describe a 'prospective-retrospective' design where evaluation of a sensitive subgroup is pre-specified but testing for biomarker status can occur after randomisation or even after clinical outcomes have been observed. Jenkins et al (2010) considers a Phase II/III seamless design that allows, via a weighted combination test, for the possible identification of a sensitive sub-population at Phase II based on PFS and describes the strategies for Phase III analysis of OS while ensuring control of the overall Type I error rate. More recently Quan et al (2012) considers a very similar approach for adaptive, post-randomization identification of sensitive sub-population.

As an alternative to the designs offered by Jenkins and Quan, it is possible to derive a design in which there is a prospectively defined biomarker subgroup of interest and an interim analysis that allows certain decisions to be made regarding the future conduct of the trial. For example, suppose a trial is sized with 1-β power to test the hypothesis $\mathrm{H}_0\colon \theta = 0$ vs $\mathrm{H}_1\colon \theta \neq 0$ with 1-sided Type I error α where θ represents the effect of treatment. If T is a sufficient statistic for θ with θ =Δ assumed under the alternative and $\mathrm{T} \sim N(\Delta, V_1)$, then $\Delta^2 = V\left(\mathrm{z}_\alpha + \mathrm{z}_\beta\right)^2$ if we employ the simplifying assumption that $V_0 = V_1 = V$ where V_0 is the variance of T under the null. Suppose further there is a biomarker defined fraction s ($0 \leq s \leq 1$) of the trial population in whom an enhanced treatment effect is hypothesized. Let $\theta_S (> \theta)$ denote the true treatment effect in this subpopulation with sufficient statistic $\mathrm{T_S} \sim N(\Delta_S, V_S)$ such that $V_S^{-1} = I_S = s \cdot I$ where $I = V^{-1}$. With an

interim analysis planned at information time f, the relationship between overall population and the subset at both the interim and final analyses is as follows:

	Interim	final
Subpopln	$I \cdot s \cdot f$	$I \cdot s$
Not Subpopln	$I \cdot (1-s) \cdot f$	$I \cdot (1-s)$
Overall	$I \cdot f$	I

If $\mathrm{T_f}$ and $\mathrm{T_{Sf}}$ represent the test statistics for the overall and subpopulation at the interim, then $(\mathrm{T_{Sf}T_f}, \mathrm{T_S}, \mathrm{T})' \sim$ MVN with expectation $\underline{\Delta}' = (\Delta_\mathrm{S}, \Delta_\mathrm{S}, \Delta, \Delta)'$, correlation matrix

$$\underline{\rho} = \begin{pmatrix} 1 & \sqrt{\mathrm{s}} & \sqrt{\mathrm{f}} & \sqrt{\mathrm{sf}} \\ & 1 & \sqrt{\mathrm{sf}} & \sqrt{\mathrm{f}} \\ & & 1 & \sqrt{\mathrm{s}} \\ & & & 1 \end{pmatrix}$$ and variance $\left(\mathrm{V_{sf}}, \mathrm{V_f}, \mathrm{V_s}, \mathrm{V}\right)'$

At the interim we might then have the following decision rule (or some other variant of interest) to determine how the trial should proceed:

- If $\mathrm{T_{Sf}} \leq \mathrm{c}_1$ or $\mathrm{T_f} \leq \mathrm{k}_1$ or both, stop for efficacy
- If $\mathrm{c}_1 \leq \mathrm{T_{Sf}} \leq \mathrm{c}_2$ and $\mathrm{k}_1 \leq \mathrm{T_f} \leq \mathrm{k}_2$ then continue recruiting all patients and test both overall and subpopulation for efficacy in final analysis
- If $\mathrm{c}_1 \leq \mathrm{T_{Sf}} \leq \mathrm{c}_2$ and $\mathrm{T_f} \geq \mathrm{k}_2$ then continue recruiting only patients in the subpopulation and test for efficacy in final analysis
- If $\mathrm{k}_1 \leq \mathrm{T_f} \leq \mathrm{k}_2$ and $\mathrm{T_{Sf}} \geq \mathrm{c}_2$ then continue recruiting all patients and test only the overall population and test for efficacy in final analysis
- If $\mathrm{T_{Sf}} \geq \mathrm{c}_2$ and $\mathrm{T_f} \geq \mathrm{k}_2$ stop for futility

From this it is straightforward to calculate the overall Type I error given s, c_j and k_j, j=1 to 3:

$$\Pr\left(T_{Sf} \le c_1 | \underline{\theta} = \underline{0}\right) + \Pr(T_f \le k_1 | \underline{\theta} = \underline{0}) +$$
$$\Pr\left(T_{Sf} \le c_1, T_f \le k_1 | \underline{\theta} = \underline{0}\right) +$$
$$\Pr(c_1 \le T_{Sf} \le c_2, k_1 \le T_f \le k_2, T_S \le c_3 | \underline{\theta} = \underline{0}) +$$
$$\Pr(c_1 \le T_{Sf} \le c_2, k_1 \le T_f \le k_2, T \le k_3 \left| \underline{\theta} = \underline{0}\right) +$$
$$\Pr(c_1 \le T_{Sf} \le c_2, T_i \ge k_2, T_S \le c_3 \left| \underline{\theta} = \underline{0}\right) +$$
$$\Pr(k_1 \le T_f \le k_2, T_{Sf} \ge c_2, T \le k_3 | \underline{\theta} = \underline{0})$$

And, similarly, the overall Type II error:

$$\Pr\left(T_{Sf} \ge c_2 , T_f \ge k_2 \left| \underline{\theta} = \underline{\Delta}\right.\right) +$$
$$\Pr(c_1 \le T_{Sf} \le c_2, k_1 \le T_f \le k_2, T_S \ge c_3 | \underline{\theta} = \underline{\Delta}) +$$
$$\Pr(c_1 \le T_{Sf} \le c_2, k_1 \le T_f \le k_2, T \ge k_3 \left| \underline{\theta} = \underline{\Delta}\right.) +$$
$$\Pr(c_1 \le T_{Sf} \le c_2, T_f \ge k_2, T_S \ge c_3 \left| \underline{\theta} = \underline{\Delta}\right.) +$$
$$\Pr(k_1 \le T_f \le k_2, T_{Sf} \ge c_2, T \ge k_3 | \underline{\theta} = \underline{\Delta})$$

Finally, an important review of the regulatory Issues associated specifically with the use of biomarkers in oncology is offered by Chakravarty et al (2011). The observation made is that most biomarker driven analyses submitted to FDA tend to be retrospective and, to strengthen the credibility of such analyses, it is suggested that biomarker related hypotheses should have a pre-determined scientific (biological) basis, the associated assay should be well characterized with good analytical performance and there should be a predefined plan for analysis relating to the biomarker, including alpha control. Further, he raises a key concern regarding the typically low fraction of patients for whom biomarker status is attained in clinical trials leading to a potentially biased convenience sample. These, and other, issues are then illustrated through the example of a panitumumab in CRCa and the retrospective analysis of patients with KRAS mutations for which only a small fraction of patients were tested in the main pivotal trial.

2.6 Recommendations

Based upon the preceding publications and discussion, the following recommendations for oncology trial design and analysis can be made:

- Trials that employ a surrogate endpoint as primary should mandate full ITT follow-up of all patients post attainment of the surrogate for overall mortality. Only in this way can the true benefit of the intervention be assessed and the value of the surrogate is assessed.
- In biomarker driven trial design, routine and naive statistical assumptions regarding (i) the precise dichotomous determination of biomarker 'negative' and 'positive' patients and (ii) the complete absence of treatment effect in 'negative' patients should be abandoned. Rather design options should be offered for a range of assumptions that allow a non-zero effect in 'negative' patients and accommodate a less than perfect assay for biomarker measurement.
- For biomarker driven developments, flexible designs should be routinely considered. In particular the designs described by Zhao (2010), Jenkins (2011) and Wang (2007) offer three feasible opportunities to identify biomarker defined patient subpopulations that achieve enhanced treatment benefit whilst controlling the overall Type I error. These designs in particular should be evaluated when considering a biomarker driven oncology Phase III development strategy.

3 Randomized Phase II designs: the use of progression free survival as an endpoint and decision making from Phase II to Phase III

3.1 Background

With newer targeted small molecules and humanised antibody treatments designed to have a cytostatic effect on the tumor, the traditional single arm, open Phase II oncology design and routine use of response rate as the Phase II endpoint are increasingly no longer appropriate tools to test and screen clinically new anti-cancer treatments for potential effectiveness. Rather, arguments have been offered in favour of double blind randomised Phase II with PFS as the endpoint.

Some of the first authors to raise these issues were Stadler and Ratain (2000), Korn et al (2001), Eskens et al 2000 and Simon et al (2001). However, concerns were subsequently raised by the FDA regarding the use of progression as an endpoint (Williams et al 2002). In addition to the question as to whether improvements in PFS *per se* represent a clinical benefit to the patient, a methodological concern was that, unlike death, the exact time a patient progresses is not known (the data being interval censored between clinic visits). The concern therefore is this might lead to a downward bias when estimating the hazard ratio between two treatments, prompting suggestions that very frequent clinic visits might be required to more accurately determine the time progression occurred. A further related concern is asymmetric follow-up of treatment arms (perhaps due to treatments having different administration schedules as can be common in oncology, or worsening of disease prompting more frequent medical assessments on one treatment relative to the other) and the scope for this to introduce bias. And, finally, there is concern on behalf of sponsors and regulators alike relating to the likely prohibitive size of an event driven randomised Phase II with PFS and the primary endpoint, leading to suggestions to use a short term endpoint such as the percentage of patients with progressive disease at some early time point, e.g. at 8 weeks (often referred to as the '*rate of progressive disease*' (PD) at x weeks), in preference to PFS over the full trial follow-up period.

3.2 Publication 6:

Stone A, Wheeler C, Carroll K and Barge A (2007) Optimizing randomized phase II trials assessing tumor progression.

A selection of these issues were addressed by Stone, Wheeler, Carroll and Barge (2007). Firstly, it was shown analytically that the use of a PD-type endpoint at some early fixed point in follow-up was highly inefficient. Table 1 from the publication is reproduced below:

Publication 6. Stone, Wheeler, Carroll (2007): Table 1

Table 1
Number of patients and duration of trial to detect a hazard ratio of 0.67 with 80% power and a one-sided significance level of 20%

Endpoint[a]	No. of patients[b]	Duration[c] (months)
PFS	100[d]	11
PD rate 10% vs. 6.8%	852	43
PD rate 20% vs. 13.8%	414	22
PD rate 30% vs. 21.2%	278	16
PD rate 50% vs. 37.0%	164	12

[a] For PD rate — comparison that results in a hazard ratio of 0.67.
[b] Total number of patients for a two-arm trial.
[c] Assumes patients are recruited uniformly at 20/month; events follow an exponential distribution with medians of 4 and 6 months, respectively.
[d] Sixty-nine events are required, a greater number of patients would reduce the duration for this endpoint.

This table describes a two arm trial requiring 69 events to test the hypothesis that the true hazard ratio between experimental treatment and control is 0.67 with 80% power and a 1-sided alpha level of 20%. Exponential PFS times are assumed with medians of 6 and 4 months on experimental and control respectively. It can be seen that a PD endpoint requires substantially more patients than a PFS endpoint to deliver a given power. This is entirely as expected since early PD endpoints are associated with drastically fewer PFS events than an analysis based upon all PFS events accruing over the full trial follow-up period. Assuming medians of m_C on control and $h^{-1} \cdot m_C$ on experimental where h is the hazard ratio, the probability of a PFS event in a trial with a planned median follow-up of T months is given by $\tilde{\pi} = 2(\pi_C^{-1} + \{1 - (1 - \pi_C)^h\}^{-1})^{-1}$ where $\pi_C = 1 - 0.5^{T/m_C}$. Therefore the increase in the total sample size required when using a PD endpoint at a control rate of p_C^{-1} is given by $\tilde{\pi}/\tilde{p}$ where $\tilde{p} = 2(p_C^{-1} + \{1 - (1 - p_C)^h\}^{-1})^{-1}$. For example, if $m_C = 4$ months, $h = 0.67$ and T=8.5 months, then $\tilde{\pi} = 0.69$ so that, with a target of 69 events for 80% power and a 1-sided α level of 0.20, $n = 69/0.69 = 100$. Comparing to an analysis at a PD rate on control of 0.10, $\tilde{p} = 0.08$ and, therefore, $\tilde{\pi}/\tilde{p} = 8.54$. Hence, and as shown in Table 1 above, the PD rate analysis requires approximately 8.5 times more patients that a PFS analysis.

To handle differential follow-up two approaches were suggested, the first used an 'event count' approach where the proportion of events (regardless of the clinic visit at which they were detected) would be analysed using a complementary log-log link. This approach was shown to be reasonable with little loss of power providing (i) proportionality of the hazard ratio over time and (ii) the data were not overly mature at the time of the analysis. The second approach briefly discussed was a 'grouped' or interval censored analysis (the performance of which was evaluated in detail in the subsequent publication Carroll 2007(b)). With regards to the frequency of clinic visits to assess progression status, it was shown that, contrary to feedback commonly attained from FDA, very frequent visits added little in terms of statistical efficiency. Table 2 from the paper is reproduced below:

Publication 6. Stone, Wheeler, Carroll (2007): Table 2

Table 2
Simulated power for a trial designed to have 80% power to detect a hazard ratio of 0.67 at a one-sided significance level of 20%, for various scanning frequencies

Visit frequency	Cox proportional hazards			Grouped analysis		
	Power[a]	Hazard ratio[b]	Follow-up[c] (weeks)	Power[a]	Hazard ratio[b]	Follow-up[c] (weeks)
a) Comparison of treatments with medians of 4 and 6 months						
Constant	79.8%	0.67	50	NA	NA	NA
2 weeks	79.6%	0.67	51	79.9%	0.66	51
1 month	79.0%	0.67	52	79.8%	0.66	52
2 months	78.1%	0.67	54	80.1%	0.66	54
4 months	74.5%	0.69	59	79.4%	0.66	59
b) Comparison of treatments with medians of 8 and 12 months						
Constant	79.6%	0.67	86	NA	NA	NA
2 weeks	79.6%	0.67	87	80.0%	0.66	86
1 month	79.2%	0.67	88	80.0%	0.66	87
2 months	78.8%	0.67	91	80.0%	0.66	90
4 months	77.5%	0.68	95	79.7%	0.66	94
6 months	76.0%	0.69	100	79.6%	0.66	98
8 months	73.4%	0.70	104	79.5%	0.66	103

Each row is the result of 5000 simulations, each with 50 observations per treatment arm and waiting for 69 events to occur, assuming an exponential distribution with the stated medians.
NA, not applicable.
[a] Proportion of simulations with a one-sided p-value <0.2 in favor of the more effective therapy.
[b] Calculated as the geometric mean of the hazard ratios estimated from each dataset.
[c] Calculated as the time from start of recruitment to the 69th event observed, patients recruited over 26 weeks.

The simulations in Table 2 show how, with a Cox analysis based upon PFS time assigned to the clinic visit at which it was first detected, power falls and the treatment effect estimate attenuates as the time interval between visits lengthens whereas, with a grouped or interval censored analysis, power is maintained and the treatment effect is estimated without bias.

The last issue addressed in the paper relates to the practice of censoring PFS time on receipt of additional anti-cancer therapy. Given that the decision to provide additional therapy to the patient is commonly related to worsening health status and/or toxicity associated with randomised treatment, censoring is clearly informative and hence statistically problematic. To address this issue, an ITT approach was recommended whereby all patients would be followed for radiographic assessment of disease through the planned duration of the trial regardless of the introduction of additional anti-cancer therapy.

3.3 Publication 7:

Carroll KJ. Analysis of progression-free survival in oncology trials: some common statistical issues (2007).

This paper further explores the issue that progression times cannot be determined exactly in a clinical trial and the problematic regulatory advice to censor for progression on receipt of additional anti-cancer treatment. Assuming exponential progression times, it is shown that, even when clinic visit schedules are the same for drug and control treatments, the common practice of taking the time of progression to be the date of the clinic visit at which it was detected results in a downwardly biased estimate of the underlying hazard ratio, θ (experimental:control), that consequently erodes power. If $\hat{\theta}$ denotes the estimated hazard ratio, then:

$$\hat{\theta} = \frac{\bar{T}_C}{\bar{T}_E} \qquad E\left[\hat{\theta}\right] = \frac{V_C\left(1-e^{-\lambda_E V_E}\right)}{V_E\left(1-e^{-\lambda_C V_C}\right)} \neq \frac{\lambda_E}{\lambda_C}$$

where, for x = C, E, λ_x represents the true and event rate on treatment x, $\bar{T}_x = \sum_{i=1}^{n_x} T_{ix}/d_x$ represents the reciprocal of the estimated event rate, n_x and d_x are the total number of patients and events respectively, T_{ix} is the time to event or censoring for patient on treatment x, and V_x represents the time interval between scheduled clinic visits. The proof of this result by maximum likelihood is provided in Appendix A of the paper and illustrated in Table I, reproduced below:

Publication 7. Carroll (2007): Table I

Table I. Bias and loss of power associated with assigning time of progression to the scheduled clinic visit at which it was detected.

Hazard ratio, θ	Median PFS on E (months)	Median PFS on C (months)	Interval between clinic visits, V, (months)	Expected HR[a], $\hat{\theta}$	Log rank power[b] (%)	Relative increase in d to compensate for loss in power[c]
0.667	6	4	0.5	0.677	87.8	1.07
			1	0.686	85.4	1.16
			2	0.705	80.0	1.34
			4	0.740	67.2	1.81
0.75	8	6	0.5	0.755	88.5	1.05
			1	0.761	86.9	1.11
			2	0.771	83.3	1.23
			4	0.792	75.0	1.51
0.80	12	9.6	0.5	0.803	89.1	1.03
			1	0.806	88.1	1.07
			2	0.811	85.9	1.14
			4	0.822	81.1	1.30

Here the expected value of the hazard ratio, $E[\hat{\theta}]$, is shown alongside the resulting reduced power of a log rank test (bearing in mind the trial was originally sized for 90% power and a 2.5% 1-sided α level). And, in the last column, the increase in the number of events required to maintain 90% power is provided.

Given the power loss associated with the estimate $\hat{\theta}$, it was further shown that to retain at least 100(1-γ)% power in a study sized for 100(1-β)% power, $\gamma > \beta$, the interval between clinic visits should be no larger than:

$$V' = (median\ PFS\ on\ control) \times \frac{2}{\ln(2)} \frac{1}{\theta} \left(\frac{\theta^k - \theta}{1 - \theta^k} \right)$$

where θ is the true hazard ratio and $k = \left| \frac{z_\alpha + z_\gamma}{z_\alpha + z_\beta} \right|$. Note since $\gamma > \beta$, then $k < 1$, and since θ is defined as the hazard ratio for experimental to control, then $\theta < 1$. Therefore the term $\frac{\theta^k - \theta}{1 - \theta^k}$ in the expression V' is always positive. The full derivation of V' is provided in Appendix B of the paper and illustrated in Table II, reproduced below:

Publication 7. Carroll (2007): Table II

Table II. Maximum inter-visit interval length to maintain at least 80% power[a] for varying θ and median PFS values.

Hazard ratio, θ	$\frac{2}{\ln(2)}\frac{1}{\theta}\left(\frac{\theta^k-\theta}{1-\theta^k}\right)$	Median PFS on C (months)	Visits at least every V' months
0.8	0.5058	4	2.0
		6	3.0
		9	4.6
		12	6.1
0.75	0.5219	4	2.1
		6	3.1
		9	4.7
		12	6.3
0.667	0.5520	4	2.2
		6	3.3
		9	5.0
		12	6.6
0.50	0.6315	4	2.5
		6	3.8
		9	5.7
		12	7.6

[a]Assuming trial originally powered at 90% ($\beta=0.1$), 2.5% 1-sided α level to detect a HR size θ.

For example, to retain at least 80% power in a trial sized for 90% power, the interval between clinic visits should be no longer than around 50-60% of the expected median time to progression on control.

In addition to bias and loss of power when clinic visit schedules are the same for both treatment and control, it was shown that asymmetric visit schedules were associated with bias and inflated Type I error. These findings are illustrated in Table III of the paper reproduced below.

Publication 7. Carroll (2007): Table III

Table III. Inflation in Type I error resulting from asymmetric visit scheduling in a trial with 508 events (sized to detect an assumed hazard ratio of 0.75, 90% power, 2.5% 1 sided α).

Median PFS on E and C (HR = 1)	Interval between visits on C (months)	Interval between visits on E (months)	Expected HR[a], $\hat{\theta}$	Type I error[b] (1-sided)
4	0.5	1	0.959	0.069
	1	2	0.920	0.152
6	1	1.5	0.972	0.050
	1	2	0.945	0.092
	2	3	0.946	0.090
9	1	1.5	0.981	0.040
	2	3	0.963	0.062
	3	4	0.964	0.061
12	1	2	0.973	0.050
	2	3	0.972	0.050
	3	4	0.972	0.050
	4	6	0.946	0.090

Here the expected value of the hazard ratio, $E[\hat{\theta}]$, is displayed for a range of asymmetric visit schedules when the true HR is unity. $E[\hat{\theta}]$ is seen to be biased in favor of the treatment with the longer time between scheduled clinic visits and the associated Type I error is consequently inflated.

In light of these findings, an alternative estimate of treatment effect, $\breve{\theta}$, was suggested assuming a common time interval between scheduled clinic visits $V = V_E = V_C$:

$$\breve{\theta} = \frac{ln\left(1 - \frac{V}{\bar{T}_E}\right)}{ln\left(1 - \frac{V}{\bar{T}_C}\right)}$$

In Table IV of the paper reproduced below, $\breve{\theta}$ was shown in simulations to be unbiased with essentially no loss of statistical efficiency:

Publication 7. Carroll (2007): Table IV

Table IV. Hazard ratio estimates resulting from 1000 simulations of a trial with 200 patients (100 per arm) in which all patients achieve an event.

Hazard Ratio, θ	Median PFS on E (months)	Median PFS on C (months)	Interval between clinic visits, V, (months)	Expected value of $\hat{\theta}$[a]	$\hat{\theta}$[b]	$\breve{\theta}$[c]	SE $\ln(\breve{\theta})$[d]
0.667	6	4	0.5	0.677	0.679	0.672	0.1438
			1	0.686	0.681	0.669	0.1418
			2	0.705	0.690	0.664	0.1388
			4	0.740	0.718	0.668	0.1428
0.75	8	6	0.5	0.755	0.751	0.746	0.1483
			1	0.761	0.759	0.750	0.1438
			2	0.771	0.764	0.748	0.1376
			4	0.792	0.782	0.751	0.1433
0.80	12	9.6	0.5	0.803	0.804	0.802	0.1425
			1	0.806	0.808	0.803	0.1440
			2	0.811	0.811	0.802	0.1377
			4	0.822	0.817	0.799	0.1474

In this Table, $E[\hat{\theta}]$ is shown alongside the geometric mean $(\hat{\theta}^b)$ of simulated hazard ratio estimates where the time of progression was assigned to the clinic visits at which it was detected. As anticipated a close match is observed. Also displayed is geometric mean $(\breve{\theta}^c)$ and associated standard error (on the log scale) of simulated hazard ratio estimates based on the estimator $\breve{\theta}$. Relative to $\hat{\theta}$, the alternative $\breve{\theta}$ is seen to be unbiased with a log scale standard error very close to that associated with the original powering of the trial (i.e. with 200 events, the standard error of the hazard ratio estimate based on a log rank test is expected to be $\sqrt{4/200} = 0.1414$).

This new estimate $\breve{\theta}$ and its SE were shown to represent an interval-censored analysis which, in turn, was shown to be closely related to an analysis of the proportion of events observed over the follow-up period using a complementary log-log link.

With respect to the FDA recommendation to censor PFS on the receipt of additional anti-cancer therapy, the serious bias and grossly inappropriate conclusions that can result from such informative censoring were highlighted. Akin to the accepted norm for mortality, full ITT follow-up for progression assessment was recommended regardless of receipt of additional anti-cancer therapies.

Further, as encapsulated in guidance from both FDA and CHMP at the time, with regards to the common request from regulatory authorities for an independent centralised review (ICR) of radiographic data in patients deemed to have progressed by the trial investigator, associated key difficulties were highlighted. These included (i) the handling of patients where the investigator and independent review disagree and radiological follow-up has ceased and (ii) common practice to review only data in patients who have progressed. The challenge with (i) is the absence of investigator follow-up is most likely to be informative, resulting in such patients being closer on average to progressing than patients where neither the investigator nor independent reviewer assigned progression. And (ii) will always

lead to a less precise estimate of the treatment effect since the number of progression events can only go down. It was argued that a more satisfactory approach would be to take (in addition to ICR of patients deemed to have progressed by the investigator) a random sample of non-progressing patients to estimate the fraction of patients without progression reclassified as progressive by independent review. The overall number of progression events could then be estimated under independent review and treatment groups compared accordingly.

While addressing many PFS related issues, the paper did not evaluate other lifetime distributions such as the Log Normal, Weibull or Gamma family and the impact this might have had on the conclusions drawn with regards to determining of the timing of progression and the value of interval-censored analysis. Also, the empiric non-parametric approach offered by Turnbull (1976) for interval censored data was not examined. In brief, this technique proceeds as follows:

- Data are $(L_i, U_i]$, for i=1,…,n patients with $U_i=\infty$ meaning the i^{th} patient is right censored at L_i.
- $t_0, t_1,\ldots, t_m$ = set of time-points that includes all the points L_i and U_i, i=1,…,n.
- For each patient define $a_{ik}=1$ if $(t_{k-1}, t_k]$, k=1,…,m, is contained in the interval $(L_i, U_i]$, and 0 otherwise.
- Let $S(t_k)$ be an initial estimate of the survivor function. Update $S(t_k)$ as follows:
    - Calculate $p_k = S(t_{k-1}) - S(t_k)$, k=1,…,m
    - Estimate the number of events which occurred at t_k by $d_k = \sum_{i=1}^{n} \frac{a_{ik}p_k}{\sum_{j=1}^{m} a_{ik}p_j}$

      and number at risk by $r_k = \sum_{j=k}^{m} d_j$
    - Use d_k and r_k to provide an updated product-limit estimate of the survivor function. Repeat until convergence.
- d_k and r_k can then be used to construct a log-rank test.

In addition to not evaluating Turnbull, the non-parametric generalization of the log-rank test to interval-censored analysis offered by Sun J et al (2005) was also not explored, primarily because there was no validated software available and no way

of providing an estimate of the hazard ratio was described. More recently Zhang and Sun J (2010) and Sun X et al (2013) offer reviews of methods for interval-censored analysis. In particular, Sun X et al (2013) suggests two variance estimates for the generalised log-rank test which, together with the associated Fisher's score, could potentially be used to provide a Pike estimate for the hazard ratio.

In respect of the recommendation to implement ITT follow-up and analysis of PFS, while contrary to the FDA anti-cancer guideline (2007), this is more consistent with the EMA guideline (2012a, 2012b) which leans toward an ITT approach. This recommendation was recently supported by Denne et al (2013) who reports a re-analysis of 28 Phase III trials with a PFS endpoint using both ITT and non-ITT approaches. This re-evaluation showed a tendency for non-ITT analysis to result in upwardly biased estimates of treatment effect. Rothmann et al (2013) performed a re-elevation of 14 'add-on' oncology trials (where drug and control are provided as an adjunct to continuing background standard care) submitted to FDA looking for evidence whether censoring PFS on the provision of further anti-cancer treatment was informative. He concluded that there was evidence, but only for the active arm and not the control arm. However, this conclusion seems strange. While blinding of the 14 trials is not stated, being 'add-on' in design, drug and control treatments are typically double-blind and, further, the introduction of anti-cancer therapy would seem most plausibly to be an event related to the worsening of disease independent of the randomised therapy received.

Also not reviewed was the possibility of estimating informatively censored PFS times using methods such as (i) Inverse Probability of Censoring Weighting (IPCW) as described by Robins and Finkelstein (2000) and Rimawi and Hilsenbeck (2012) or (ii) Rank Preserving Structural Failure Time (RPSFT) analysis as described by Robins and Tsiatis (1991) and Branson and Whitehead (2002). While application of such methods might provide some theoretical estimate of the treatment effect on PFS in the absence of informative censoring, the fundamental issue is that such an estimate has no practical meaning or value to the patient or prescribing physician. This is because this estimate is entirely academic and can never be realised in

practice as patients *will* fail and *will* receive additional anti-cancer treatments. In line with Korn et al (2011) and Fleming et al (2009), the only meaningful estimate of treatment effect is based on an ITT analysis as this reflects the true value of treatment in the routine medical care of the patient. However, to provide an ITT analysis of PFS will likely require a fundamental change in the philosophy of trial conduct with an alteration to the patient informed consent. This change (i) would be explicit that patients are consenting to follow-up for the full, planned duration of the trial and not follow-up to the point randomised treatment stops or additional treatment is added and (ii) would need to be unequivocally clear to the trial investigator that they *must* continue to monitor the patient and collect radiographic assessments of disease for the entire, protocol defined trial follow-up period regardless of the introduction of additional anti-cancer treatment or dropout due to toxicity or adverse event.

Finally, several of the chief recommendations made in the 2007 paper were echoed and supported in subsequent publications by Amit et al (2011) and Stone et al (2011), who reported findings from a cross industry working group on PFS assessment in oncology trials. Based on a reanalysis of multiple industry trials and simulation studies, these authors conclude that a random sample audit approach to ICR of progression data was sufficient; that an ITT philosophy should be followed for PFS and that an interval-censored analysis of PFS data should be a default sensitivity analysis in oncology trials. Senior European regulators Pignatti, Hemmings and Jonsson (2011) support recommendations regarding an ITT approach and interval censored analysis, however they expressed caution regarding ICR by random sample and highlighted the limitations of meta-analyses and simulations to substantiate important new recommendations regarding PFS assessment and trial design in oncology.

3.4 Publication 8:

Carroll KJ (2013). Decision Making from Phase II to Phase III and the Probability of Success: Reassured by 'Assurance'?

Closely related to the issue of Phase II oncology design and choice of primary endpoint is how best to use the resulting data in decision making to inform drug developers regarding the likelihood that much larger and more expensive Phase III development will deliver a 'successful' outcome, usually meaning pivotal trials will deliver positive results for the primary efficacy endpoint with $p<0.025$ 1-sided. Given the high failure rate observed in Phase III across industry over the past decade, this area has become increasingly important. 'Assurance', or expected power, was described by O'Hagan et al (2005) and Stallard et al (2005) as a means of using Phase II outcomes, or other prior information regarding the assumed treatment effect, to predict the chance of success in proposed Phase III trials. This technique is gaining favour within industry as evidenced by recent papers from Chuang-Stein (2006), Chuang-Stein et al (2011) and Kirby et al (2012). Su (2010) looks at a variation of assurance where by the observed Phase II data and some assumed prior for the true treatment effect are combined to form a hybrid prior which is then used as the basis of calculations relating to probability of success in Phase III. More recently, Nikolakopoulos et al (2013) takes a similar approach in combining Phase II biomarker data with an assumed prior for the predictive value of the biomarker versus the true clinical outcome and using this in the context of an assurance calculation for Phase III.

In practice, however, the use of assurance in oncology, and elsewhere, often leads to confusion amongst non-statisticians and decision makers due to some of its seemingly strange and counter-intuitive properties. *"Decision Making from Phase II to Phase III and the Probability of Success: Reassured by 'Assurance' "* examines the properties of assurance when the prior for Phase III is defined by the preceding Phase II data. In this case, Phase III assurance is given by:

$$pr(x > c) = 1 - \Phi\left(\frac{c - m}{\sqrt{s^2 + \sigma^2}}\right)$$

where x represents the treatment effect estimate arising in the planned Phase III trial which is sized to detect a true treatment effect θ with Type I and Type II errors, α and β so that the anticipated variance of x is given by $\sigma^2 = \theta^2/(z_\alpha + z_\beta)^2$ where $z_u = \Phi^{-1}(1-u)$ and $\Phi^{-1}(.)$ represents the inverse standard Normal distribution function; and where $c = z_\alpha \sigma$ is the critical value of the test such that if $x > c$ the null is rejected in favour of the alternative; and where the Phase II treatment effect estimate is m, with variance s^2.

In particular, it was shown that, maximally, the probability of success (PoS) by assurance = 1- {1-sided Phase II p-value} when the intended Phase III was large relative to the completed Phase II. Hence, if p=0.2 1-sided in Phase II, assurance cannot exceed 80% even if Phase III has a million patients and a conventional power >99.999%. And when integrated over prior Phase II data, it was shown that the outcomes of two independent Phase III trials are Bivariate Normal with correlation $\rho(x,y) = \frac{s^2}{s^2+\sigma^2}$ where x and y represent the treatment effect estimates from two identical Phase III's each with variance σ^2, and s^2 represents the variance of the treatment effect estimate from Phase II. The proof of this result is provided in Appendix B of the paper. For example, and as discussed in the paper, if there were two Phase III oncology trials of identical size and design, each requiring 508 events to provide 90% power to detect a true PFS hazard ratio of 0.75 and a preceding Phase II trial that provided a PFS hazard ratio estimate of 0.75 on 70 events, then the assurance for each of the Phase III trials individually would be 67%; but since Phase IIIs outcomes are correlated by assurance with $\rho = 0.88$ then the probability of two successful Phase III trials is not simply $0.67^2 = 0.449$, but rather 0.60.

Table I of the paper displays some key observations regarding assurance and is reproduced below:

Publication 8. Carroll (2013): Table 1

Table 1 Summary of key features of assurance

		Phase III	
		Small	Large
Phase II	Small	$s^2 \to \infty, s^2 \gg \sigma^2$ Assurance = Phase II one-sided p-value (= 50%)	$\sigma^2 \to 0, s^2 \gg \sigma^2$ Assurance = Phase II one-sided p-value
	Large	$s^2 \to 0, s^2 \ll \sigma^2$ Assurance = Power of Phase III trial at $\theta_{\text{TRUE}} = \text{m}$	$s^2 > 0, \sigma^2 > 0$ with $s^2 = k\sigma^2$ Assurance = $\Phi^{-1}\left(\frac{z_\beta}{\sqrt{k+1}}\right)$ if Phase III sized in region of $\theta_{\text{TRUE}} = \text{m}$

where β is Type II error used in the sizing for Phase III and z_β is the corresponding standard Normal deviate

It was also argued that while statistical methods can serve to assist good decision making, more basic, fundamental considerations in terms of Phase II design and analysis were the key to effective decision making in drug development. These include: (i) ensuring Phase II is well-controlled, randomised and double-blinded where possible. (ii) Performing two Phase II trials since positive outcomes from two Phase II's of moderate size are generally more reassuring than a single larger Phase II, especially in disease settings with softer trial endpoints such as central nervous system (CNS) disorders like depression and schizophrenia which often use subjective rating scales completed by the physician and not the patient. If a single Phase II is designed with $\Delta^2 = s^2(z_\alpha + z_\beta)$, where Δ is the hypothesized treatment effect of interest and s^2 the variance of an associated unbiased estimator for Δ, then as compared to two Phase IIs each with $\Delta^2 = 2s^2(z_\alpha + z_\gamma)$, then $\gamma = 1 - \Phi\{(z_\alpha + z_\beta)/\sqrt{2} - z_\alpha\}$. Therefore, if α=0.05 and 1-β=0.90 (0.80), 1-γ^2=0.887 (0.793) so that two Phase IIs with N/2 patients provide similar power to a single Phase II with N patients. (iii) Predefining the Phase III Go/No Go decision rule for 'success' and, importantly, stick to it. (iv) Minimizing the potential for serious bias by avoiding multiple interim analyses, particularly when the study is open. (v) If the primary fails the Go/No Go, not looking for 'signals' elsewhere in an attempt to salvage by means of extensive (post-hoc) subgroup analyses (vi) Ensuring senior leaders and decision makers are talented, well experienced drug developers with a proven track record. (viii) Including an experienced, technically expert statistician in the heart of the decision making process.

This work on decision making is limited in scope to the concept of assurance and, further, to the definition of success in Phase III being p<0.025 1-sided for the primary endpoint. Other definitions might include achieving p<0.025 and an observed difference between treatments greater than some predefined amount. However, such variants on the definition of success would not alter the fundamental properties of assurance and the need to use the concept with caution in late stage drug development. Other statistical approaches and strategies to decision making have been discussed in the literature by Senn (1996) and Julious

and Swank (2005) in terms of a sponsors overall development portfolio. Also not addressed in this work is the interesting question to what extent a small positive oncology Phase II trial might over-estimate the true underlying treatment effect. The observation that many positive Phase II trials have been followed by negative Phase III's (Arrowsmith 2011a and 2011b) suggests there may be some over estimation of effect in Phase II. Analytically the question may be framed as follows: if x represents the treatment effect estimate from Phase II with variance s_x^2 and expectation μ, then we are interested in $E(x|x \geq c)$ where $c = \mu + s_x\Phi^{-1}(1-p_x)$ and $\Phi^{-1}(.)$ is the standard Normal cumulative distribution inverse and p_x is the one-sided Phase II p-value. Then

$$E(x|x \geq c) = \int_{x=c}^{\infty} \frac{xf(x)}{p_x}dx = \mu + s_x \int_{z=(c-\mu_x)s_x^{-1}}^{\infty} \frac{z\phi(z)}{p_x}dz \ > \ \mu$$

Hence, if proceeding to Phase III is contingent upon achieving a 1-sided Phase II p-value $\leq p_x$ then positive Phase IIs will tend to overestimate the true treatment effect μ. For example if $\mu = 0$, $\sigma_x^2 = 1$ and $p_x = 0.1$, then $E(x|x \geq c) = 1.754$; similarly, if $\mu = 5$ and $\sigma_x^2 = 16$, then $E(x|x \geq c) = 5 + 4 \times 1.754 = 12.01$. By focusing only on positive Phase II trials, the true treatment effect can be overestimated.

3.5 Recommendations:

- The routine practice of assigning the time of progression to the clinic visit at which it was first detected results in a downwardly biased estimate of the hazard ratio and, thus, reduces power. If clinic visit schedules are not closely matching between treatments, this bias is increased and Type I error increased.
- To ameliorate these issues:
    - Differential follow-up should be avoided
    - Interval censored analysis should be conducted as per Sun to avoid bias and maintain power.

    - A Turnbull estimate of the CDF should be provided as standard
    - Under proportionality, an analysis of the number of PFS events over the trial period using complementary log–log link will provide for an unbiased comparison between treatments with reasonable power so long as no more than around 75% of patients have had a PFS event.
    - If a traditional approach to analyse progression at the visit where it was detected using a log-rank test, then number of events should be increased accordingly to offset the loss in power.
- Very frequent clinic visits are not statistically necessary to provide an accurate estimate of the treatment effect.
- Censoring on additional anti-cancer treatments should be abandoned and an ITT approach to PFS trial design and analysis employed, commensurate with the long established approach for overall survival.
- Full ICR is often unnecessary; a random sampling approach that draws from progressed and non-progressed patients is preferable.
- The increasing use of assurance to estimate the probability of success in Phase III based on Phase II data is problematic and often confusing to the non-statistician and, therefore, should be used with caution.

4 Issues with active-controlled, 'non-inferiority' designs

Active-controlled, 'non-inferiority' (AC NI) trials are a key feature of many drug developments, not only in oncology but across the spectrum of therapeutic areas. Such trials commonly take place when there are issues of feasibility or ethics preventing the conduct of a placebo controlled trial. In such circumstances the experimental drug is compared to an active control that has previously been shown to be safe and effective. Blackwelder (1982) was amongst the first to argue that non-rejection of the null was not a valid basis for concluding equivalence, and rather recommended a one-sided test that the control is more effective than drug by some pre specified amount or 'fixed margin'. By rejecting the null in favour of the alternative allowed a conclusion that drug is non-inferior (or more correctly not substantially inferior) to control. Definition of the margin is often difficult and not infrequently controversial (Hung et al. 2005; Lange and Freitag 2005). The aim typically is to set the margin to rule out a 'minimally clinically meaningful difference', but there is considerable subjectivity and variability in doing so. Due to concerns regarding 'constancy', i.e. the effectiveness of control being identical in the current AC NI trial and in previous historical trials that defined the effect of control, and other issues such as the relevance of the historical trial data in terms of items such as patient population, current medical practice, length of treatment etc., regulators such as FDA have typically employed conservative approaches to the determination of the NI. The most common approach has been the so-called '95-95' rule whereby the lower 1-sided 97.5% confidence limit for the effect of the control vs placebo based on historical data is chosen as the margin; or, more often, 50% of this lower limit is set as the margin to allow demonstration that at least 50% of control effect has been retained by the new drug (Holmgren 1999, FDA 1999). Clearly use of the lower confidence limit or 50% thereof, introduces a considerable conservatism into the analysis, reducing power and increasing the risk effective new drugs will be mistakenly judged ineffective.

A key problem with the fixed margin approach is that despite the margin being informed by historical data on control, the variability inherent in these data is not factored into the NI analysis itself. A simple alternative is to use the 'synthesis' approach that explicitly combines the historical estimate of control vs placebo, $\hat{\theta}_{cp}$ with variance V_{cp}, and the estimate of control vs drug from the current AC trial, $\hat{\theta}_{ec}$ with variance V_{ec}, treating both as random variables. Under constancy, the true effect of experimental to placebo, θ_{ep}, is then estimated indirectly as $\hat{\theta}_{cp} + \hat{\theta}_{ec}$ with variance $V_{cp} + V_{ec}$ and unconditional Type I and Type II error probabilities can be accurately determined. Many authors (Rothmann et al. 2003; Rothmann 2005; Hauck and Anderson 1999; Hung et al. 2003; Hasselblad and Kong 2001; Holmgren 1999; Simon 1999; Wang et al. 2002; Wang and Hung 2003; Snapinn 2004) recommend the synthesis approach to test for preservation of a fixed, pre-specified non-zero fraction (usually 50% or 75%) of the control effect.

4.1 Publications 9 and 10:

Carroll K and Milsted R (2004). Barriers to clinical development in oncology: The impact of new thinking around non-inferiority.

Carroll K, Milsted B and Lewis JA (2004). Letter to the Editor: Design and analysis of non-inferiority mortality trials in oncology.

In 2003 Rothmann of FDA published an important paper that highlighted the conservatism of the fixed margin approach when using 50% of the lower 1-sided 97.5% confidence limit. Rothmann altered this to derive the lower (1-ζ)100% CL for control vs placebo that, when 50% was taken to define the NI limit and assuming constancy, would result in a synthesis based test for NI of exactly 2.5% 1-sided (Rothmann et al 2003). Nevertheless, this fundamental approach continued to impose a heavy burden on researchers, often requiring impractical and infeasible trials and therefore threatening to inhibit the development of new medicines.

This issue was highlighted by Carroll and Milsted in abstract from at the 2004 American Society of Clinical Oncology meeting, and further by Carroll, Milsted and Lewis by Letter to the Editor in *Statistics in Medicine* in the same year in which it

was shown that even with a strong historical control effect with p<0.001, extremely challenging AC trial sizes would result. An alternative approach was illustrated that did not require pre-specification of a specific margin or percent effect retention. Rather, having obtained the AC trial data, the comparison to historical control could be readily displayed on a continuum in relation to the likely fraction of the control effect retained, from zero to 100 percent. In such an 'effect retention likelihood' plot, a zero effect retained allows estimation of the degree of benefit over placebo, albeit through indirect measures. This would then be more in line with the efficacy standards required by both US and EU law, both of which call for substantial evidence of efficacy (versus placebo) to be established, with no requirement on relative efficacy with respect to existing agents.

4.2 Publication 11:

Carroll KJ (2006). Active-controlled, non-inferiority trials in oncology: arbitrary limits, infeasible sample sizes and uninformative data analysis. Is there another way?

Issues surrounding the use of NI trials in the regulatory assessment of new anticancer medicines were thrown into sharp relief in the July 2004 FDA Advisory Committee discussion of pemetrexed as second line treatment for non-small cell lung cancer (NSCLC). At the meeting FDA cited concerns regarding the limited amount of data used to support (FDA's own) approval of the control drug. Critically, Dr. Richard Pazdur, Director of the Division of Oncology Drug Products, FDA opened the meeting stating that "A certain proportion of the control effect . . . should be preserved to demonstrate non-inferiority" thus introducing a double regulatory standard for the assessment of drug effectiveness based solely on trial design (FDA 2004d). FDA went on to conclude that pemetrexed could not be proven to be non-inferior to docetaxel since the upper 95% CL for the hazard ratio of 1.20 exceeded the protocol pre-defined non-inferiority limit of 1.11. This judgement was despite pemetrexed demonstrating 78% retention of the docetaxel effect. Contrary to FDAs view, Advisory Committee members voted in favour of approval of the drug on the basis that it appeared to have 'similar' survival to docetaxel with

a more favourable side effect profile. And in their positive review and subsequent approval, EMA judged that 'although non-inferiority was not formally established, the data submitted are robust enough to conclude that a clinically significant inferiority of pemetrexed to docetaxel in terms of efficacy in this population is unlikely' (EMA 2006).

Building upon earlier publications, this case catalyzed the 2006 publication of *'arbitrary limits, infeasible sample sizes and uninformative data analysis'*. In the absence at that time of any formal guidance from FDA regarding the trial design and analysis of AC, NI trials, this paper took the opportunity to highlight many of the associated statistical issues and regulatory precedent. The infeasibility of NI trials designed using FDA's 95-95 rule was highlighted and illustrated in Table 1, as reproduced below.

Publication 11. Carroll (2006): Table I

Table I. The number of deaths required to prove a new treatment retains $\frac{1}{2}$ of the effect of control treatment (HR = 0.50) using Rothmann *et al* methodology (true HR new:control is unity, 90% power, α 2.5% one-sided).

(Historical) *p*-value for control vs placebo	Upper 95% CI for HR(*T*:*C*) must be less than	Approx No. of deaths required to prove 50% effect retention
0.049	1.004	3 000 000
0.02	1.12	3082
0.01	1.18	1563
0.001	1.27	735
0.0001	1.31	572
0.00001	1.33	505
0.000001	1.35	459
$<<0.000001$[a]	1.41	350

[a] Equivalent to the control effect being known with (virtually) complete certainty.

Further, the fundamental purpose of an AC was discussed and argued to be to provide evidence that a new drug would have been better than placebo if placebo could have been included; with a secondary *supportive* purpose being to estimate (indirectly) the size of the effect of the new drug relative to control. It was then shown that both of these goals could be simultaneously addressed using an effect retention likelihood approach. In contrast to previous publications, the paper gave a clearer understanding of how this concept was constructed as follows: Given the fundamental hypothesis to be tested is $H_0: \theta_{ep} = 0$ vs $H_1: \theta_{ep} > 0$, due to concerns regarding constancy, the alternative is often replaced with $H_1: \theta_{ep} \geq f \times \theta_{cp}$ to allow demonstration the experimental drug retains some non-zero fraction f, $0 \leq f \leq 1$, of the historical control effect. Then, using the notation previously described regarding the synthesis method, the test statistic then becomes $z^* = \frac{\hat{\theta}_{ec} + (1-f)\hat{\theta}_{cp}}{\sqrt{\hat{V}_{ec} + (1-f)^2 \hat{V}_{cp}}}$.. Plotting the inverse cumulative Normal function $\Phi^{-1}(z^*)$ against f provides the 'effect retention likelihood' plot. This approach is illustrated in Figure 1 of the paper, reproduced below:

Publication 11. Carroll (2006): Figure 1

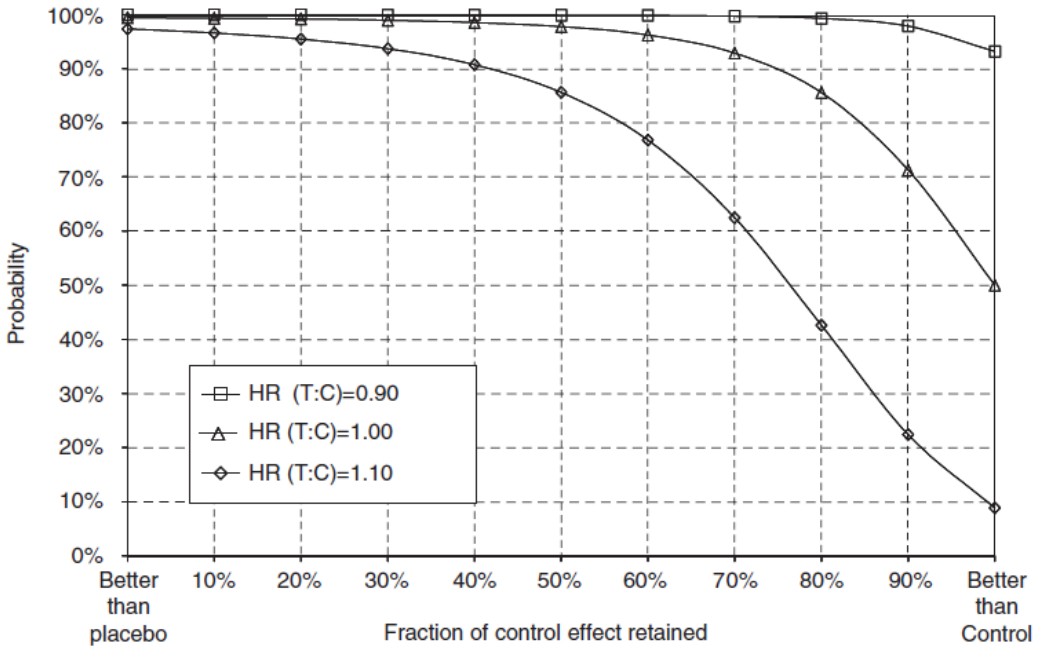


Figure 1. Effect retention likelihood plot: historical control effect, HR=1.5, p=0.005; Active-control trial design with 800 events, 90% power, 2.5% one-sided a level.

The figure shows that, with a historical control:placebo hazard ratio of 1.5, p=0.005, if an AC trial is conducted comparing experimental to control with 800 events and a hazard ratio treatment:control (HR (T:C)) of unity is observed, there is (i) a 99.4% probability that experimental would have beaten placebo (zero effect retention) and (ii) (for example) a 97.8% chance that 50% of the control effect has been retained by experimental.

In line with Senn (2005), it was argued that while of value in trial design, the predefined NI fixed limit is of little value in analysing and interpreting NI trials. Rather, the judgement as to what is and is not an unacceptable loss of effectiveness of the control treatment should lie with the 'consumer', that is with physicians and their patients or the regulatory authority acting on the patients behalf. And, crucially, it was further highlighted that the typical FDA requirement for at least 50% effect retention represented an arbitrarily higher burden of evidence for drug approval based solely upon trial design.

4.3 Publication 12:

Carroll KJ (2013). Statistical issues and controversies in active-controlled, 'non-inferiority' trials

Following release of EMA's regulatory guidance on NI trial design and analysis (EMA 2000, EMA 2005), which in turn followed ICH E9 (1998) and E10 (2000), FDA released their draft 'Guidance for Industry Non-Inferiority Clinical Trials' (FDA 2010b). The key issues offered in this guidance, which remains draft, together with many other pertinent matters relating to AC, NI trial design and analysis were the subject the 2013 paper by Carroll *"Statistical issues and controversies in active-controlled, 'non-inferiority' trials."*

In their draft guidance, FDA introduced not one, but two NI margins, 'M1' and 'M2'. M1 was defined as "the entire effect of the active control" estimated from historical data so that meeting this margin was identified to be directly equivalent to the first objective an active-control 'NI' trial, being to establish indirectly that the

new drug would have beaten placebo if placebo could have been included. FDA defined their M2 margin as 'the largest clinically acceptable' loss of effectiveness of drug relative to control, being broadly equivalent to some fraction of the control effect having been preserved by drug (often times 50%). Meeting only M1 would generally be insufficient and, therefore, these newly defined margins served to perpetuate the arbitrarily higher standard of evidence required by FDA for establishing effectiveness via an AC, NI trial. Also, in defining M1 as a fixed margin FDA continued to not take into account the uncertainty associated with the historical data on control, and in defining M2 FDA failed to address the inherent subjectivity, vagaries and idiosyncrasies associated with the clinically driven determination of this second hurdle. Rather, M1 and M2 are considered simply as fixed, constant margins and the absence of a common, consistent definition for M2 is not addressed. Treating these margins as fixed is in certain contrast to ICH guidance which calls for a justification of the margin that must take into account the historical data and its uncertainty. To ignore the underlying uncertainty in the historical data giving rise to the M1 margin in particular, being used to demonstrate indirect effectiveness vs placebo, is arguably improper statistically. Use of the synthesis method is more satisfying in this regard as the uncertainty in the historical control effect estimate is directly accounted for.

In line with Peterson, Carroll, Chuang-Stein et al (2010), this paper argued that regulatory standards based on fixed margins or some percent preservation of the historical control effect were arbitrary and lacking scientific justification, and their use introduced logical inconsistencies in the decision making process such that effective treatments could be denied approval. Owing to these illogicalities, it was argued that the qualities of an experimental drug in order to allow it to be judged effective should be independent of the trial design evaluating that drug. This would then imply that if preservation of effect is not required when a new treatment is evaluated in a placebo-controlled trial, it should not be required when the treatment is evaluated in an active-controlled trial. Hence, it was argued that after accounting for any methodological concerns with the trials used to establish efficacy, a common standard of evidence should apply regardless of trial design

which, therefore, should not be greater than the standard used for previous approvals for the same indication. Further, it was argued that the synthesis method for combining results of an active-controlled trial with results of historical trials could, with appropriate consideration of constancy, adequately characterises of the efficacy of a new treatment relative to placebo and would avoid the illogicalities highlighted with fixed margin and preservation of effect approaches (Peterson, Carroll, Chuang-Stein et al (2010))

Further, the performance of the three main methods for assessing effectiveness via an AC, NI trial were compared. Hypotheses were laid out for the synthesis, preservation of effect and 'fixed margin' approaches and associated formulae derived that specified the amount of information required to test these hypotheses with 1-sided Type I error α and a given, common power, 1-β. In this comparison the 'fixed margin' was defined as a given fraction, $0 \leq \eta \leq 1$, of the lower 1-sided 97.5% CL for the control effect, consistent FDA's 95-95 rule. Key observations were thus made possible under the usual assumption that the true difference between drug and control was in the null:

1. The synthesis method is always more efficient than the preservation of effect method for testing indirectly the hypothesis a new drug is efficacious.
2. To determine effectiveness of experimental drug with 1-β power, both synthesis and preservation of effect methods require that the historical data provides an estimate of the control effect with $p < 2\Phi^{-1}\{-(z_\alpha + z_\beta)\}$, meaning a historical control effect with a z-value of >3.24 or p<0.0012 is required to achieve 90% power to test indirectly the hypothesis that a new drug is efficacious.
3. The maximum power achievable in respect of this hypothesis, via either the synthesis or preservation of effect method, is $\Phi^{-1}\{|z_{CP}|-z_\alpha\}$, where $z_{CP}$the z-value for the historical control effect estimate.
4. The same issues and power cap do not apply to the 'fixed' margin approach where any level of power can theoretically be achieved providing $\hat{\theta}_{cp} - z_\alpha\sqrt{V_{cp}} > 0$, i.e. the estimated control effect is significant with p<0.05.

5. For the common choice f =η = 0.5, the 'fixed' approach is less efficient than the preservation of effect method when the control effect has significance $p < 2\Phi^{-1}\left\{-0.5z_\alpha\left\{\left(1+z_\beta/z_\alpha\right)^2+1\right\}\right\}$, or p < 0.00025 when 90% power is desired.
6. Otherwise, the 'fixed' approach is more efficient.

The observation that both the preservation of effect and synthesis methods require a historical control effect estimate with p ≤0.0012 is of interest since this may suggest there is little need to further discount historical data (Snapinn 2004). And in contrast to the 'fixed margin' approach, the presence of a power cap for the preservation of effect and synthesis methods is important as it illustrates that an AC, NI approach is futile unless data on the historical control is relatively strong (or unless there was very good reason to believe the new drug would be marginally better in efficacy than control (Fleming 2008)). This has implications for the amount of evidence required to grant licensure for the first drug in a new indication, especially in oncology. If, for example, in some late stage malignant disease with no currently proven therapies, the FDA or EMA were to approve a drug that improved survival relative to best supportive care with p<0.01, then it would subsequently be impossible for other researchers or a pharmaceutical manufacturer to prove indirectly the effectiveness of any new drug via an AC, NI trial with 90% power.

To illustrate these observations an example is offered of a control drug with hazard ratio versus placebo of 0.667 95% CI (0.524, 0.849), p=0.0010 for overall survival, representing a 50% increase in the event rate for placebo relative to control based on 264 events. The 'fixed' NI limit would typically be = $0.849^{0.5}$ = 0.921. It is shown that, under constancy, an active control trial would require 6270 events (≈24x more than the 264 events characterising the historical effect of control) to deliver 90% power to test, indirectly, the effectiveness of drug. In contrast, the preservation of effect method with f = 0.5 would require 35,073 events (≈130x more than for historical control) while the synthesis method (3) would require 8768 events (≈33x more than for historical control). The sensitivity of the power calculation to the strength of

historical evidence characterising the control effect estimate was further illustrated in Table 1 of the paper reproduced below:

Publication 12. Carroll 2013(b): Table 1

Table 1. Comparison of sample size requirements in an AC trial to test indirectly at the 1-sided 2.5% α level the hypothesis that a new drug is efficacious

Historical data								Number of events required in AC trial			
Histrcl HR for C:P[1]	No. of Events	V_{CP}[2]	95% CI	2-sided p-value	Assumed true HR for E:C[3]	f or η	'Fixed' NI limit	Synthesis	'Fixed'	Preservation	Max power for efficacy[5]
0.667	100	0.0400	(0.451, 0.987)	0.04289	1	0.5	0.994	--[4]	1,000,128	--	53%
0.667	200	0.0200	(0.506, 0.880)	0.00419	1	0.5	0.938	--	10,297	--	82%
0.667	264	0.0152	(0.524, 0.849)	0.00100	1	0.5	0.921	8,768	6,273	35,073	91%
0.667	327	0.0122	(0.537, 0.828)	0.00025	1	0.5	0.910	1,185	4,747	4,740	96%
0.667	500	0.0080	(0.560, 0.795)	0.00001	1	0.5	0.892	526	3,188	2,103	99%
0.667	1,000	0.0040	(0.589, 0.755)	<0.00001	1	0.5	0.869	345	2,129	1,378	100%

1: Historical estimate of the control effect, C=control, P=placebo

2: V_{CP}=variance of the historical control estimate

3: Hypothesized effect for drug vs control effect, E=drug

4: Variability of historical control effect estimate too great to achieve 90% power to test indirectly the hypothesis that a new drug is efficacious

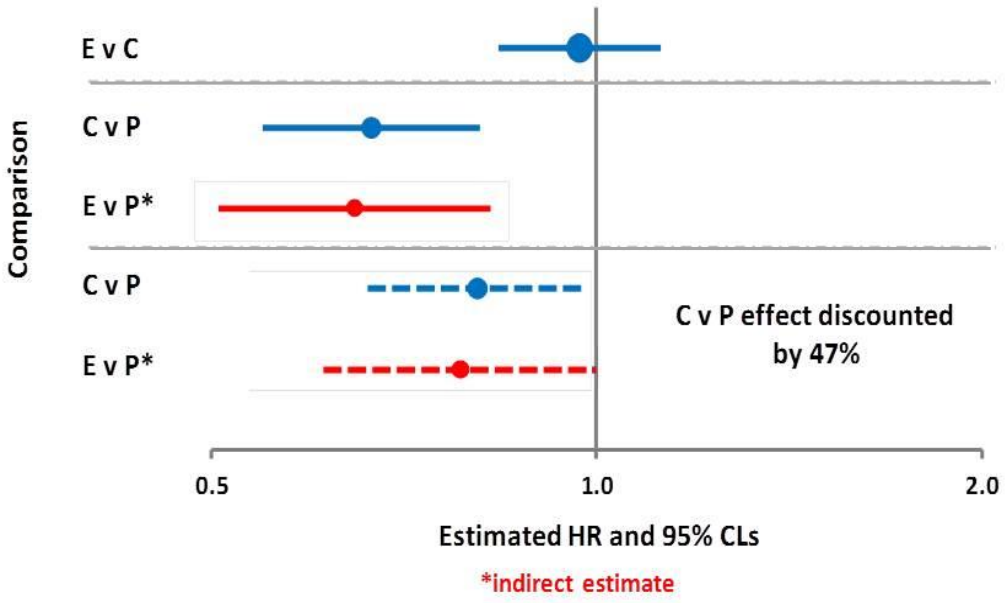
5: Maximum power achievable to test indirectly the hypothesis that a new drug is efficacious versus placebo

Other issues addressed in the paper included per-protocol (PP) vs intent-to-treat (ITT) analyses within the context of an AC, NI trial. In line with ICH E9 (1998) which states that *"…it is especially important to minimise the incidence of violations of the entry criteria, non-compliance, withdrawals, losses to follow-up, missing data and other deviations from the protocol, and also to minimise their impact on the subsequent analyses.",* it was argued that the common practice of applying PP analyses to an AC, NI trial was not ideal. Rather than remove violators and deviators from the analysis, a more appropriate strategy is to execute AC trials to rigorous and exacting standards, to minimize protocol non-adherence and ensure full ITT follow-up of all randomised patients so the trial evidence generated is of the highest completeness and quality (Fleming 2008). In this way regulators and the scientific community can rely upon the data and what it shows. This goes to the heart of 'assay sensitivity' in terms of ensuring the AC trial is of the highest possible scientific standard, regardless of whether the objective is superiority or 'NI'.

In terms of design, it was argued the AC, NI trial should be powered on the basis of proving indirectly that the new drug is effective vs placebo and then the results analysed using the synthesis method. Given concerns regarding constancy, inclusion of analyses that discount the historical control data can, in some circumstances, be helpful and informative. In a manner similar to that proposed by Rothmann (2003), an approach was illustrated whereby the maximum degree of discounting of the historical control effect was calculated that would yet still provide an indirect estimate of drug effectiveness with $p<0.025$ 1-sided. This, coupled with appropriate data display, including a likelihood effect retention plot, would provide a clear and transparent display of the relevant data. Figures 2 and 3 from the paper are reproduced below:

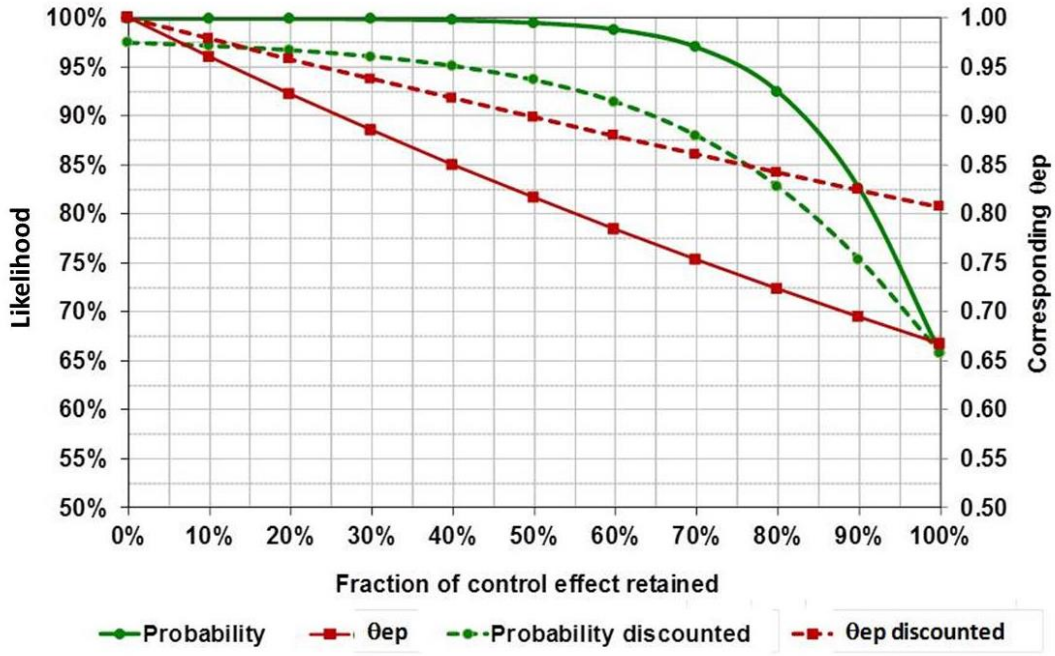
Publication 12. Carroll 2013(b): Figure 2

Figure 2. Active control, 'NI' analysis



Publication 12. Carroll 2013(b): Figure 3

Figure 3. Effect retention likelihood plot for CvP HR=0.67 (0.55, 0.81) and EvC HR=0.97 (0.84, 1.12)



Graphical display of the data in this fashion is more informative than the usual analysis associated with active-control, 'NI' trials. Employing a Bayesian interpretation of the data, we can discern there is a 99.98% chance that drug is effective versus putative placebo; and a 99.5% chance that 50% of the effect of control has been retained; and a 65.8% chance that 100% of the effect of control has been retained, i.e. that drug is superior to control. Similarly, if the historical control effect is discounted by as much as 43%, there is a 97.5% chance that drug is effective versus putative placebo; and a 93.7% chance that 50% of the effect of control has been retained. Discounting further would not allow rejection of $\mathrm{H}_0\colon \theta_{\mathrm{ep}} = 0$. Hence, by examining the data in terms of the effect retention likelihood, it seems clear that, in this example, the AC trial data provides confidence that drug is efficacious and, further, there is a high chance that drug retains a large fraction of the control effect.

The overall aim of the preceding publications on AC, NI issues was to highlight the practical infeasibility, subjectivity and illogicality of FDA's longstanding approach to NI assessment and, in so doing, offer a more reasonable statistical alternative for the design and analysis of AC, NI trials. However, the fundamental problem with AC, 'NI' trials as a vehicle to establish drug effectiveness is the understandable mistrust of indirect comparisons, and this issue is only addressed briefly in the preceding work that advocates the synthesis method. The constancy assumption is inherently unverifiable in any NI context and for any NI method that does not include a placebo arm in the AC trial. In terms of the latter, several authors have discussed designs involving experimental (E), control (C) and placebo (P) (Munzel 2009, Mielke et al 2008, Hasler et al 2008, Koch and Röhmel2004). In such a design inference regarding the amount, ω, of the control effect retained by experimental is straightforward: the contrast $\hat{C}(1-\omega) - \hat{E} + \omega\hat{P} > 0$ would provide a direct test of the hypothesis that experimental retains 100ω% of the control effect. However, this kind of design is seldom possible in oncology, particularly in late stage disease. Also not examined in detail is the method of discounting described by Snapinn (2004). And while this offers a quantitative approach to address constancy concerns, it is inherently

subjective introducing arbitrarily defined discounting parameters into an already complex construct.

It is of interest to consider closely related matters within the context of health economic assessment of medicines. In this field there has been an upsurge in published papers describing the use of Network Meta-Analyses (NMA) and, to a lesser extent, Comparative Effectiveness (CE) assessment. Jones (2011) gives an excellent summary of the statistical methodology underpinning NMA. Broadly, NMA brings together the outcomes of trials involving a set of treatments but, unlike a conventional meta-analysis, these trials do not each have to contain the treatments to be compared. Rather, trials are combined in a fashion akin to a balanced incomplete block design but in a Bayesian framework with priors employed for individual drug effects. Interestingly there is essentially no critique or criticism in the literature of NMAs regarding the issue of constancy and other basic issues such as the comparability of trial populations, trial conduct, follow-up and analysis. Yet the outcomes of such analyses are used to make critical health care decisions. It seems scientifically incongruous to have such an array of scientific regulatory guidance relating to making indirect inference regarding the efficacy of single drug on the basis of an AC,NI trial, yet no scientific guidance regarding the arguably much more troubling matter of combining multiple trials with incomplete (if any) representation of a set of treatments for which the issues of constancy and combinability based on design, population, conduct and contemporaneousness must be considered more acute.

An additional matter not addressed in the preceding publications relates to the statistical meta-analysis of historical trial data on control. Typically, regulators such as FDA tend to use random effects meta-analysis to combine historical data for use as reference in an NI analysis. A case in point is warfarin which is provided as an example in FDA's guidance document (FDA 2010b). These data are reproduced below:

Placebo-controlled warfarin trial data

Study	warfarin		placebo			
	Events (E_w)	Patient-years exposure (T_w)	Events (E_p)	Patient-years exposure (T_p)	Log RR[1]	Variance[2] log RR
AFASAK I	9	413	21	398	-0.8843	0.1538
BAATAF	3	487	13	435	-1.5793	0.4059
AEFT	21	507	54	405	-1.1691	0.0617
CAFA	7	237	11	241	-0.4352	0.2254
SPAF I	8	260	20	244	-0.9798	0.1671
SPINAF	9	489	24	483	-0.9932	0.1487

1: RR = $\frac{E_w}{T_w}\frac{T_p}{E_p}$. 2: Variance log RR = $\frac{T_w - E_w}{T_w E_w} + \frac{T_p - E_p}{T_p E_p}$, as per FDA formulae.

A DerSimonian and Laird (1986) random effects meta-analysis provides the following results:

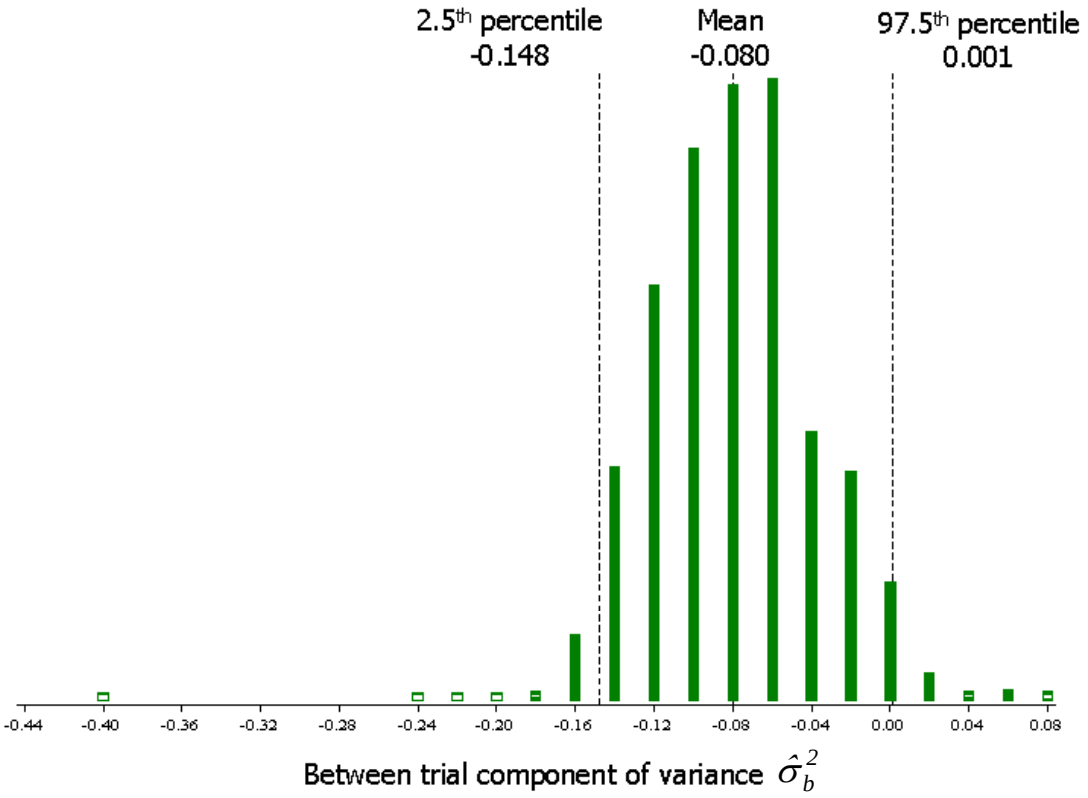
Parameter	Estimate
Mean within trial variance, $\overline{\sigma}_i^2$	0.1938
Between trial component variance, σ_b^2	-0.0681
According to DerSimonian and Laird, when σ_b^2 is estimated to be negative, then the meta-analysis proceeds with σ_b^2 set at zero. This then gives the following results, matching those obtained by FDA:	
Relative risk (RR)	0.3615
SE log RR	0.1537

In order to calculate the 95% confidence interval (CI) for the RR, and thus derive the usual 95-95 'fixed' limit, FDA employs a t-distribution on 6 degrees of freedom according to Follmann and Proschan (1999). However, a t-distribution is arguably inappropriate in this instance as the between trial component of variance, σ_b^2, is zero. Application of a t-distribution when a z-distribution applies can easily give rise to confusing and counter-intuitive results when combining historical trial data. For example, suppose two trials, each with 250 events (and, thus, a SE ln(RR) of

$\sqrt{4/250}$), give similar RRs of 0.75. Both trials are thus statistically significant with 95% CI (0.59, 0.96), p=0.0229. The combined estimate, either by random or fixed effects analysis, is therefore 0.75 on 500 events (SE ln(RR) of $\sqrt{4/500}$). Application of the z-distribution gives a 95% CI of (0.63, 0.89), p=0.0013, whereas application of the t-distribution gives a 95% CI of (0.24, 2.34), p=0.19. If there were 3 rather than 2 trials, the z-distribution gives 95% CI of (0.65, 0.87), p=0.00008 and the t-distribution gives 95% CI of (0.55 1,03), p=0.0588. The loss of significance despite the doubling or tripling of events occurs because it is assumed variability in the overall estimate is driven by the number of trials, either 2 or 3, and not the number of events within a trial. This makes little sense when the estimated between trial component of variance is zero.

It might be argued that the estimated between trial variance component of -0.0681 is a point estimate and the true value of σ_b^2 could be larger and positive. However, this is not supported by bootstrap resampling analyses (Westfall and Young 1993, Westfall et al 1993) . The empirical distribution for σ_b^2 based on all possible re-samples is shown in the following figure. The 95% CI (based on the 2.5th and 97.5th percentiles) for σ_b^2 is (–0.148, 0.001) and truncating for negative estimates gives an interval of [0, 0.001). It can be safely concluded, therefore, that the between trial component of variability for the warfarin placebo controlled trials is zero.

Empirical bootstrap distribution of the between trial component of variance resulting from a DerSimonian and Laird random effects meta-analysis



The table below displays 95% CI estimates for the RR, warfarin:placebo, together with the corresponding NI limit, assuming firstly a conservative t-distribution and then, secondly, a more appropriate z-distribution. Also provided for comparison are the results of a bootstrap analysis on the RR estimate resulting from a DerSimonian and Laird random-effects meta-analysis.

Method	RR estimate	SE log RR	Assumed asymptotic distribution for RR estimate	95% CI for RR	NI-margin
Random effects analysis	0.361	0.154	t-distribution	0.248 to 0.527	1.38
			z-distribution	0.267 to 0.489	1.43
Bootstrap analysis*	0.366	0.101	None	0.311 to 0.462	1.47

*** CI based on the 2.5th and 97.5th percentiles of the empirical bootstrap distribution for the log RR. RR estimate is the geometric mean of resample estimates of the log RR; median is 0.362.**

Employing a t-distribution to calculate the CI for the RR results in a conservative estimate for the NI limit of 1.38. The use of the z-distribution provides an NI limit of 1.43, which is more in line with the distribution free empirical bootstrap estimate of 1.47. While the difference in these limits may seem modest, they have a large impact on the sample size for an AC, NI trial. Using an NI limit of 1.38 requires 405 events (13,500 patients for a 3% annual event rate) for 90% power as compared to 329 events (10,950 patients) and 283 events (9,440 patients) for NI limits of 1.43 and 1.47 respectively. The use of the latter NI limits therefore result in sample size savings of 20% and 30% respectively.

4.4 Recommendations

- A single regulatory standard for the assessment of drug effectiveness should apply regardless of trial design.
- AC, NI trials should be sized to determine efficacy vs putative placebo. The use of the punitively conservative 95-95 rule and 50% effect preservation should be abandoned. Both approaches are subject to serious illogicalities which render them unsuitable for licensing decisions.
- Effectiveness should be assessed via the synthesis method. This provides a test of size 2.5% 1-sided under constancy and is statistically more efficient than the 'fixed' margin and preservation of effect approaches.
- Constancy should be addressed either through discounting or use of effect retention likelihood methodology which allows the percentage of control effect retained to be assessed on a continuum from 0% (= likelihood superior to placebo) to 100% = (likelihood superior to control).
- The arbitrarily determined additional M2 margin should be abandoned. The extent of drug effectiveness beyond beating placebo alone can be assessed via the effect retention likelihood.
- To instill confidence in the reliability of an AC, NI assessment, careful a-priori examination of the historical trial data and their relevance to the current AC, NI design and setting is essential.

- Meta-analysis of historical trial data should not employ a Follmann and Proschan (1999) adjustment by default.
- In AC, NI analyses, patient exclusions and non-randomised PP analyses should be avoided. Stringent standards for trial conduct and execution should be established at the outset to ensure close adherence to the protocol and deliver full patient follow-up.

5 The use of parametric methods in the analysis of oncology clinical trial data

In oncology trials, the log-rank test and related Cox regression analysis are the established tools for time to event endpoint analysis such as PFS and overall survival. Their prominence comes as a result of extensive use in clinical research over the past four decades. A key appeal is that there is no need to assume a given form for the underlying survivor function in order to make inferences about relative event rates. Yet, in direct parallel with other areas of clinical statistics such as the routine and widespread use of mixed models for longitudinal data or the use of Poisson and Negative Binomial modelling in the analysis of exacerbations in chronic obstructive pulmonary disease, if the distribution of survival times can be well approximated, parametric failure-time modelling can be powerful and informative, allowing a wider set of inferences to be made than either the log-rank or Cox approaches can offer.

5.1 Publications 13 and 14:

Carroll KJ (2003). On the use and utility of the Weibull model in the analysis of survival data.

Carroll KJ (2009). Back to basics: explaining sample size in outcome trials, are statisticians doing a thorough job?

The use and potential benefits of modelling time to event data were reviewed and explored in this paper. The simple two parameter Weibull $f(t) = \alpha\lambda t^{\alpha-1}e^{-\lambda t^{\alpha}}$ with $t > 0$, $\alpha > 0$, $\lambda > 0$, where λ represents the event rate and α the scale, was explored since it is both an accelerated failure-time model (AFT) and a proportional hazards model, being the only member of the AFT family to possess both properties. Analysis via the Weibull distribution therefore allows simultaneous description of treatment effects in terms of (i) the usual hazard ratio and (ii) the relative increase (or decrease) in the time to an event, referred to in the paper to as

the “event time ratio” (ETR) to illustrate the close parallel with the better known hazard (or event) rate ratio. A quantification of the treatment effect in terms of increased time is often desired by clinicians and investigators who can find the concept of the HR difficult to understand.

In the paper it was shown analytically that Weibull and Cox based analyses provide asymptotically unbiased, equally efficient estimates of the hazard ratio regardless of whether proportionality holds. Model fitting and parameterisation using the software procedure PROC LIFEREG in SAS was explained where the scale parameter $\alpha = \sigma^{-1}$ and log event rate $\ln(\lambda) = -\left(\mu - \underline{\beta}'\underline{x}\right)$ (SAS® 2011). Further, the important relationship between the ratio of percentiles of the Weibull distribution and the hazard ratio (HR) under proportionality was provided, being $\frac{t_{pA}}{t_{pB}} = HR^{-\alpha^{-1}}$ where t_{px} represents time to reach the p^{th} percentile, $x = A, B$. These basic features of a Weibull analysis were illustrated by example in Tables 2 and 3 as reproduced below:

Publication 13. Carroll 2003: Table 2

Table 2. Results of Weibull and Cox analyses

	Weibull: modeling event time ratio				Cox: modeling log hazard ratio		
	$\hat{\mu}$	$SE^a\hat{\mu}$	$\hat{\sigma}$	$SE^a\hat{\sigma}$			
	8.977	0.158	0.7275	0.0307			
	$\hat{\beta}$	$SE^a\hat{\beta}$	t	$-\hat{\beta}/\hat{\sigma}$	$\hat{\gamma}$	$SE\hat{\gamma}$	t
Randomized treatment[b]	0.4022	0.0706	5.70	−0.5529	−0.5544	0.0947	−5.85
Log PSA[c] at diagnosis	−0.2005	0.0352	−5.70	0.2756	0.2772	0.0471	5.89
Disease stage	0.3802	0.0746	5.10	−0.5226	−0.5265	0.1002	−5.25
Radiotherapy	−0.3184	0.0987	−3.23	0.4377	0.4382	0.1347	3.25
Observation	−0.6184	0.0837	−7.39	0.8500	0.8548	0.1096	7.80
Moderately differentiated	−0.1456	0.0891	−1.63	0.2001	0.1977	0.1222	1.62
Poorly differentiated	−0.3973	0.0937	−4.24	0.5461	0.5500	0.1275	4.31

[a] Standard error.
[b] Covariance between scale and treatment parameters was estimated to be 0.00047305.
[c] Prostate-specific antigen.

Publication 13. Carroll 2003: Table 3

Table 3. Estimated hazard (HR) and event time ratios (ETR) for active relative to placebo

Cox			Weibull					
HR	SE[a]	95% CI[b]	HR	SE	95% CI	ETR	SE	95% CI
0.574	0.0947	0.477, 0.692	0.575	0.0947	0.477, 0.693	1.495	0.0706	1.302, 1.717

[a] Standard error.
[b] Confidence interval.

A direct means of testing for departures from proportionality was derived and, in cases where proportionality did not hold, it was described how treatments could still be compared statistically via the ratio of integrated hazards, $\int_0^T h_i(u)du/T$ for $i = A, B$, over some time period 0-T (where typically T would represent the trial follow-up period). This would provide a measure of the average event rate on treatment A relative to the same for treatment B over the period (0-T]. For the Weibull this ratio would be $\frac{\lambda_A}{\lambda_B} T^{\alpha_A - \alpha_B}$. The approximate standard error of this quantity can be derived using the delta method if the parameters λ_x and α_x, $x = A, B$, together with their standard errors are estimated separately for treatments A and B using appropriate software (such as PROC LIFEREG in SAS® 2011).

An important use of the Weibull in predicting data maturation was described and illustrated by example. Formulae were derived that allowed the percent of additional events to be estimated by simulation over the period (0,T+S] given current data over (0-T]. Finally, the issue of departures from the Weibull distribution and potential impact on inference was addressed by means of further simulation studies where data from a range of non-Weibull settings (LogNormal, Gamma and piece-wise exponential) were simulated and then analysed assuming a Weibull model and also by Cox regression. A very close match was observed in results from Weibull and Cox analyses, underscoring the robustness of inference on the hazard ratio to departures from the Weibull. However, for the estimation of percentiles and the ETR, a reasonable match to Weibull distribution is required. Weibull and Cox analyses of data from a piecewise exponential distribution is displayed in Table 4 of the paper, reproduced below:

Publication 13. Carroll 2003: Table 4

Table 4. Simulation of piecewise exponential: analysis by Cox and by Weibull

			Cox analysis			Weibull analysis					
λ_1[a]	μ_A/μ_B[b]	$\tilde{\mu}_A/\tilde{\mu}_B$[c]	*HR*[d]	*SE*[e] *ln HR*	*t*	HR	SE ln HR	*t*	ETR[f]	SE In ETR	*t*
0.01	1.25	1.13	0.834	0.1199	−1.51	0.826	0.1185	−1.62	1.099	0.0585	1.62
0.01	1.50	1.26	0.716	0.1270	−2.64	0.702	0.1251	−2.83	1.191	0.0612	2.86
0.10	1.25	1.10	0.872	0.1142	−1.19	0.874	0.1073	−1.25	1.115	0.0868	1.25
0.10	1.50	1.21	0.783	0.1195	−2.05	0.784	0.1123	−2.16	1.221	0.0920	2.17
1	1.25	1.00	0.995	0.1096	−0.05	0.982	0.1341	−0.33	1.033	0.1568	0.14
1	1.50	1.00	0.987	0.1127	−0.12	0.967	0.1407	−0.25	1.043	0.1658	0.25

[a] Common event rate over first 3 months.
[b] Ratio of mean times to event; μ_A = *6 months throughout.*
[c] Ratio of median times to event.
[d] Hazard ratio.
[e] Standard error.
[f] Event time ratio.

In the '*Back to basics*' paper the use of parametric models in sample size estimation for time to event endpoints was described. The standard sizing for a time to event trial relies on accruing a given number of events, E, with the total number of patients to be recruited then determined by $E\pi^{-1}$ where π is the probability of an event over the trial follow-up period. Typically patient entry times, r, are assumed to be Uniformly distributed $U(0,R)$, and times to event exponential with parameter λ leading to $\pi = 1 - \frac{e^{-\lambda F}(1-e^{-\lambda R})}{\lambda R}$ where F is the minimum follow-up period. However these assumptions are often unrealistic (Williford 1987). It was shown in the paper that, more generally, the probability of an event over the trial period R+F is given by:

$$\int_{r=0}^{R}\int_{t=r}^{R+F} f(t|r)f(r)dtdr = 1 - E_r\left[e^{-\lambda(R+F-r)^{\alpha}}\right] \approx 1 - e^{-\lambda(R+F-E[r])^{\alpha}}$$

with the approximation holding if λ is small (meaning, typically, that the accrual period R < median time to event) and where r denotes patient entry time as before and t denotes the time to event following a Weibull distribution with parameters α and λ. If entry times are uniform then $1 - e^{-\lambda(R+F-E[r])^{\alpha}} = 1 - e^{-\lambda(0.5R+F)^{\alpha}}$. Extension to non-uniform patient entry times, where $f(r) = \frac{\gamma e^{-\gamma r}}{1-e^{-\gamma R}}$ and $E[r] = \frac{1}{\gamma} - \frac{Re^{-\gamma R}}{1-e^{-\gamma A}}$, is straightforward.

5.2 Publication 15:

Ellis S, Carroll KJ and Pemberton K (2008). Analysis of Duration of Response in Oncology Trials.

The duration of response (DoR) in the subset of responding patients is commonly evaluated in oncology trials. However, a formal statistical comparison between drug and control for DoR is clearly biased since the groups being compared are defined by the post-treatment outcome of response rather than by randomisation and, therefore are, by definition, non-comparable. The EMA anti-cancer guideline (EMA 2012a) states no formal statistical comparison between treatment groups should be made on the basis of DoR in the subset of responding patients. This

paper built on earlier work by Temkin (1978) and Begg and Larson (1982) who considered the probability of being in response function (PBRF) as an alternative means of describing the difference between treatments in relation to response. It was shown in the paper that using the stochastic formulation of Begg and Larson and assuming exponentially distributed transition times, the area under the PBRF curve (if defined to infinity), would provide an estimate of the expected DoR (EDoR) in all patients and not only those that responded. It was further shown that a novel, alternative formulation using a mixture distribution approach not only arrived at the same result as Begg and Larson when exponential times were assumed, but was also more flexible, allowing other, better fitting parametric failure-time distributions (such as the Weibull) to be employed. This new approach was illustrated using data from the gefitinib INTACT trial (Herbst et al 2004) and a proposal for transparent presentation of analysis results offered. Figure 3 of the paper, being the empiric PRBF for the gefitinib example, is reproduced below along with Table 3, which provides the associated EDoR analysis results, and Figure 4b which displays the Weibull model fit to the duration of response times in responding patients:

Publication 15. Ellis, Carroll, Pemberton 2008: Figure 3

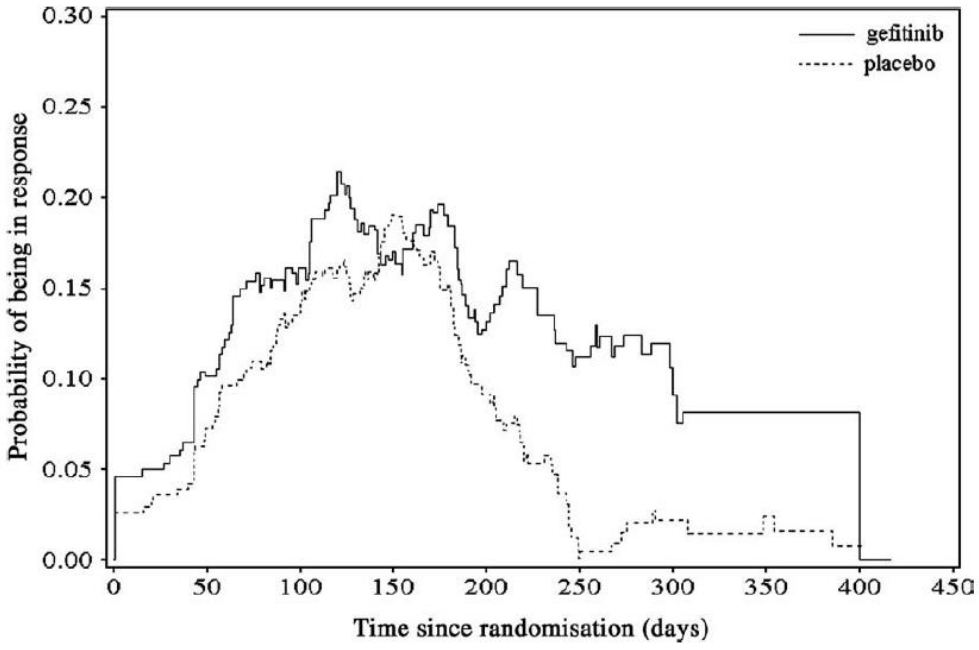


Fig. 3. Probability of being in response as a function of follow-up time: gefitinib 500 mg vs. placebo, INTACT 2.

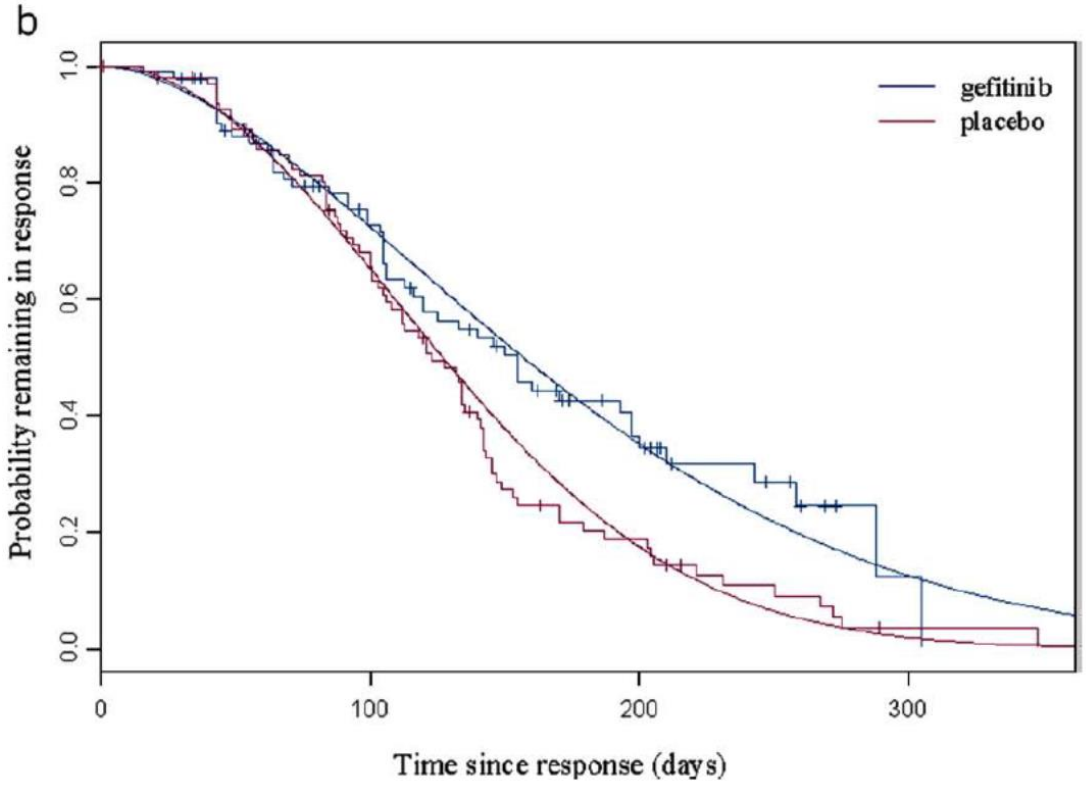
Publication 15. Ellis, Carroll, Pemberton 2008: Table 3

Table 3
Gefitinib vs. placebo, INTACT 2. Comparison of treatments for Expected Duration of Response using exponential, Weibull and log Normal densities

	Exponential		Weibull		Log Normal	
	Gefitinib N=347	Placebo N=345	Gefitinib N=347	Placebo N=345	Gefitinib N=347	Placebo N=345
Response rate, % [1]	30.6%	29.9%	30.6%	29.9%	30.6%	29.9%
Mean DoR[a] [2]	221.6	148.8	173.7	134.7	202.6	139.5
SE[b] DoR	0.137	0.115	0.083	0.057	0.131	0.074
EDoR[c] [1]x[2]	67.7	44.4	53.1	40.2	61.9	41.7
Ratio of EDoR and 95% CI[d]	1.524		1.320		1.486	
	(1.003 to 2.313)		(0.977 to 1.783)		(1.025 to 2.155)	
	P=0.048		P=0.07		P=0.04	

[a]DoR = Duration of response in responding patients, days.
[b]SE = standard error.
[c]EDoR = Expected duration of response, days.
[d]CI = Confidence interval.

Publication 15. Ellis, Carroll, Pemberton 2008: Figure 4b



When an evaluation of duration of response is desired it was argued that (i) the PBRF should be used to descriptively display data (ii) response rates should be provided for the duration of response in responding patients, including the associated Kaplan–Meier curves (without any formal comparison or p-value attached) and (iii) the expected duration of response and associated statistical comparison be provided; and then these measures (i)-(iii) be laid out as per the examples provided in Tables 2 and 3 of the paper. This would then ensure data were presented and displayed transparently, allowing statisticians and non-statisticians to most readily appreciate the relative difference between treatments.

In these papers, extensions and other uses of parametric modelling of time to event data were briefly discussed. For example, Wei and Glidden (1997) suggest AFT models may be useful for multivariate failure-time data and Keiding et al (1997) have suggested AFT approaches may be useful in random effects survival analyses (or frailty modelling), emphasizing intuitive interpretation of the Weibull. Sposto (2002) examines parametric cure models, concluding that they are at least as good

as Cox-based approaches and are to be preferred when proportionality fails to hold, allowing simultaneous assessment of covariate effects on both the proportion cured and the failure rate among those not cured.

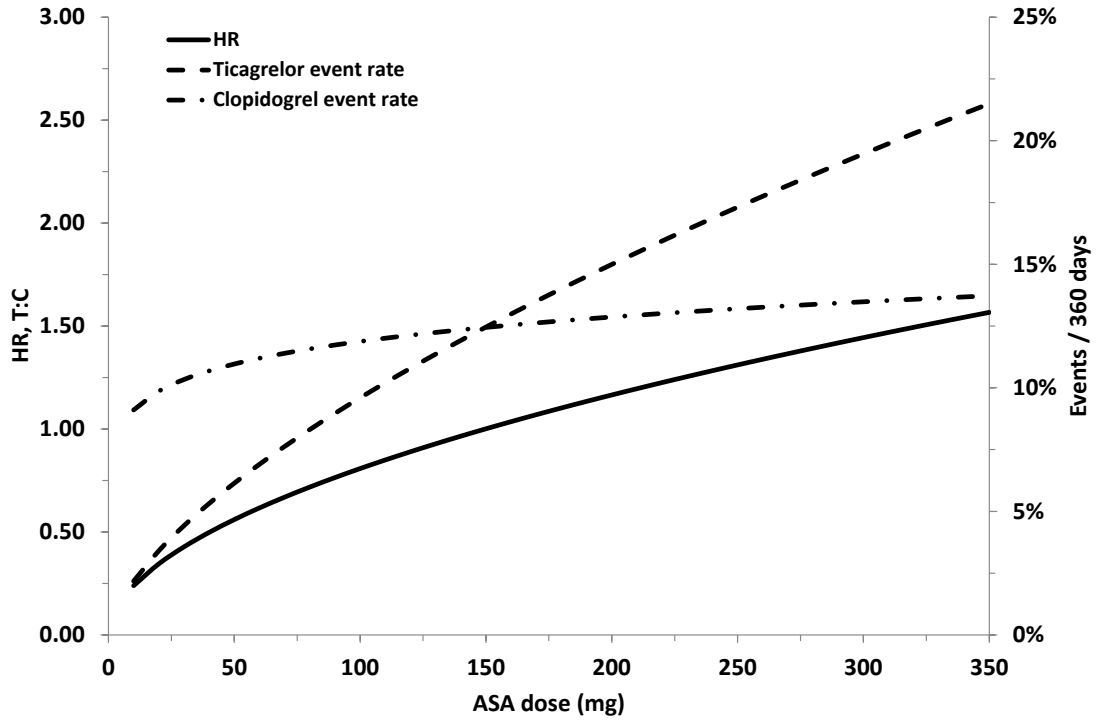
Not addressed in either paper, however, were other important matters pertaining to the routine use of Weibull and other parametric models for failure time data. For example, the use of diagnostics to assess model fit and adequacy was not addressed. Issues relating to power and sample size determination associated with a Weibull analysis were not covered, being an area where further research would be beneficial. And while Cox and Weibull analyses were found to give similar estimates of the hazard ratio irrespective of the underlying time to event distribution, the extent to which Weibull misspecification would impact the estimate of percentiles and, in particular, the median time to event was not addressed. Further research and guidance in this area would be most helpful in assessing the risk of bias when making inferences based on an assumed Weibull (or any other AFT) distribution.

Despite the benefits and versatility of Weibull modeling of time to event data, surprisingly few applications in late stage clinical drug development have been documented. One recent and important example is provided by the comparison of ticagrelor and clopidogrel in the treatment of acute coronary syndromes (Wallentin et al 2009). In the large pivotal phase III registration trial, 'PLATO', a highly significant qualitative interaction was observed between randomized treatment and concomitant aspirin dose for the primary endpoint of time to first of CV death, non-fatal myocardial infarction and non-fatal stroke. While the relationship between the hazard ratio and aspirin dosage was easily estimated via Cox modeling, US and EU regulatory authorities were particularly interested in the relationship between the event rate and aspirin dose on each arm and, in turn, how this related to the hazard ratio. Weibull modeling was therefore used. For each of ticagrelor and clopidogrel separately, the time to event was modeled as in SAS® using PROC LIFEREG with terms for intercept and log(aspirin dose) as a continuous covariate and a Weibull link function (SAS 2011). The resulting parameter estimates were

thus $\hat{\mu}$, $\hat{\beta}$ and $\hat{\sigma}$ for the intercept, regression co-efficient for log(aspirin dose) and scale respectively, with corresponding variance and covariance estimates $\hat{V}_x$ and $\hat{C}_{x,y}$ with $x, y = \hat{\mu}, \hat{\beta}, \hat{\sigma}$ and $x \neq y$.

As highlighted previously, since SAS® parameterises with $\alpha = \sigma^{-1}$ and $\ln(\lambda) = -\left(\mu - \underline{\beta}'\underline{x}\right)$, then $H_T = -(\mu + \beta \log(D)) + \log(T)/\sigma$ where D = log(aspirin dose) and H_T denotes the log integrated hazard over (0, T]. Hence, $\hat{H}_T = -\left(\hat{\mu} + \hat{\beta}\log(D)\right) + \log(T)/\hat{\sigma}$ with $\widehat{Var}(\hat{H}_T) = \left[\frac{\partial H_T}{\partial \mu}\right]^2_{\hat{\mu}} \hat{V}_\mu + \left[\frac{\partial H_T}{\partial \beta}\right]^2_{\hat{\beta}} \hat{V}_\beta + \left[\frac{\partial H_T}{\partial \sigma}\right]^2_{\hat{\sigma}} \hat{V}_\sigma + 2\left[\frac{\partial H_T}{\partial \mu}\right]\left[\frac{\partial H_T}{\partial \beta}\right] \hat{C}_{\mu,\beta} + 2\left[\frac{\partial H_T}{\partial \mu}\right]\left[\frac{\partial H_T}{\partial \sigma}\right] \hat{C}_{\mu,\sigma} + 2\left[\frac{\partial H_T}{\partial \beta}\right]\left[\frac{\partial H_T}{\partial \sigma}\right] \hat{C}_{\beta,\sigma}$. The log hazard ratio can then be defined as $\hat{H}_{T_{tica}} - \hat{H}_{T_{clop}}$ with variance estimate $\widehat{Var}_{tica}(\hat{H}_T) + \widehat{Var}_{clop}(\hat{H}_T)$.

The results of the analysis are shown in the figure below:



This analysis therefore revealed that the observed interaction with aspirin dosage was mainly due to an interaction with ticagrelor treatment, with the 360 day event rate rising from around 5% up to 25% as the aspirin dose rose, whereas a much

flatter relationship is seen for clopidiogrel. These results were stated to be of importance in FDA's review and eventual approval of the drug, and led to a black box warning that ticagrelor should not be co-administered with aspirin doses exceeding 100mg per day (FDA 2011a-c).

5.3 Recommendations

- In oncology trials, Weibull modelling of time to event data should be routinely performed in support of the more standard Cox analysis. The same covariates and strata should be used for both analyses.
- Treatment effects should be reported in terms of both the hazard ratio and the ETR since the latter provides a clinically useful direct measure of the relative improvement in the time to an event.
- Concerns regarding an exact distributional match to the Weibull distribution should not be overstated given a Weibull-based analysis provides very similar results to a Cox based analysis regardless of the true underlying distribution of the time to event.
- With respect to predicting data maturation, Weibull modelling should be used as standard in preference to the often used and very simplistic assumption of a constant event rate over time.
- Treatments should not be formally compared on the basis of DoR in responding patients. Rather, an EDoR analysis should be performed and a comparison made on the basis of all randomised patients. Results of such an analysis should be presented clearly and transparently as recommended in the associated paper.

6 Collated Recommendations for Oncologic Clinical Trial Design and Analysis

On the use of biomarkers and surrogate endpoints to accelerate drug development:

- Trials that employ a surrogate endpoint as primary should mandate full ITT follow-up of all patients post attainment of the surrogate for overall mortality. Only in this way can the true benefit of the intervention be assessed and the value of the surrogate is assessed.
- In biomarker driven trial design, routine and naive statistical assumptions regarding (i) the precise dichotomous determination of biomarker 'negative' and 'positive' patients and (ii) the complete absence of treatment effect in 'negative' patients should be abandoned. Rather design options should be offered for a range of assumptions that allow a non-zero effect in 'negative' patients and accommodate a less than perfect assay for biomarker measurement.
- For biomarker driven developments, flexible designs should be routinely considered. In particular the designs described by Zhao (2010), Jenkins (2011) and Wang (2007) offer three feasible opportunities to identify biomarker defined patient subpopulations that achieve enhanced treatment benefit whilst controlling the overall Type I error. These designs in particular should be evaluated when considering a biomarker driven oncology Phase III development strategy.

Randomized Phase II designs: On the use of progression free survival as an endpoint and decision making from Phase II to Phase III

- The routine practice of assigning the time of progression to the clinic visit at which it was first detected results in a downwardly biased estimate of the hazard ratio and, thus, reduces power. If clinic visit schedules are not

closely matching between treatments, this bias is increased and Type I error increased.

- To ameliorate these issues:
    - Differential follow-up should be avoided
    - Interval censored analysis should be conducted as per Sun to avoid bias and maintain power.
    - A Turnbull estimate of the CDF should be provided as standard
    - Under proportionality, an analysis of the number of PFS events over the trial period using complementary log–log link will provide for an unbiased comparison between treatments with l reasonable power so long as no more than around 75% of patients have had a PFS event.
    - If a traditional approach to analyse progression at the visit where it was detected using a log-rank test, then number of events should be increased accordingly to offset the loss in power.
- Very frequent clinic visits are not statistically necessary to provide an accurate estimate of the treatment effect.
- Censoring on additional anti-cancer treatments should be abandoned and an ITT approach to PFS trial design and analysis employed, commensurate with the long established approach for overall survival.
- Full ICR is often unnecessary; a random sampling approach that draws from progressed and non-progressed patients is preferable.
- The increasing use of assurance to estimate the probability of success in Phase III based on Phase II data is problematic and often confusing to the non-statistician and, therefore, should be used with caution.

Issues with active-controlled, 'non-inferiority' designs:

- A single regulatory standard for the assessment of drug effectiveness should apply regardless of trial design.

- AC, NI trials should be sized to deter-mine efficacy vs putative placebo. The use of the punitively conservative 95-95 rule and 50% effect preservation should be abandoned. Both approaches are subject to serious illogicalities which render them unsuitable for licensing decisions.
- Effectiveness should be assessed via the synthesis method. This provides a test of size 2.5% 1-sided under constancy and is statistically more efficient than the 'fixed' margin and preservation of effect approaches.
- Constancy should be addressed either through discounting or use of effect retention likelihood methodology which allows the percentage of control effect retained to be assessed on a continuum from 0% (= likelihood superior to placebo) to 100% = (likelihood superior to control).
- The arbitrarily determined additional M2 margin should be abandoned. The extent of drug effectiveness beyond beating placebo alone can be assessed via the effect retention likelihood.
- To instill confidence in the reliability of an AC, NI assessment, careful a-priori examination of the historical trial data and their relevance to the current AC, NI design and setting is essential.
- Meta-analysis of historical trial data should not employ a Follmann and Proschan (1999) adjustment by default.
- In AC, NI analyses, patient exclusions and non-randomised PP analyses should be avoided. Stringent standards for trial conduct and execution should be established at the outset to ensure close adherence to the protocol and deliver full patient follow-up.
- In AC, NI analyses, patient exclusions and non-randomised PP analyses should be avoided. Stringent standards for trial conduct and execution should be established at the outset to ensure close adherence to the protocol and deliver full patient follow-up.

On the use of parametric methods in the analysis of oncology clinical trial data:

- In oncology trials, Weibull modelling of time to event data should be routinely performed in support of the more standard Cox analysis. The same covariates and strata should be used for both analyses.
- Treatment effects should be reported in terms of both the hazard ratio and the ETR since the latter provides a clinically useful direct measure of the relative improvement in the time to an event.
- Concerns regarding an exact distributional match to the Weibull distribution should not be overstated given a Weibull-based analysis provides very similar results to a Cox based analysis regardless of the true underlying distribution of the time to event.
- With respect to predicting data maturation, Weibull modelling should be used as standard in preference to the often used and very simplistic assumption of a constant event rate over time.
- Treatments should not be formally compared on the basis of DoR in responding patients. Rather, an EDoR analysis should be performed and a comparison made on the basis of all randomised patients. Results of such an analysis should be presented clearly and transparently as recommended in the associated paper.

References

Amit O, Mannino, Stone AM, Bushnell W et al (2011). Blinded independent central review of progression in cancer clinical trials: Results from a meta-analysis and recommendations from a PhRMA working group. European Journal of Cancer; 47:1772–1778.

Arrowsmith J. (2011a). Phase II failures: 2008–2010. Nature Reviews Drug Discovery; 10: 328–329.

Arrowsmith J. (2011b). Phase III and submission failures: 2007–2010. Nature Reviews Drug Discovery; 10:87–88.

Begg BB, Larson M (1982). A study of the use of the Probability-of-being-response function as a summary of Tumour response data. Biometrics;38:59–66.

Blackwelder WC (1982). Proving the Null Hypothesis. Controlled Clinical Trials; 3:345–353.

Branson M, Whitehead J (2002). Estimating a treatment effect in survival studies in which patients switch treatment. Statist. Med.; 21:2449-2463.

Burzykowski T, Buyse M (2006). Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharm Stat; 5(3):173-86.

Burzykowski T, Buyse M, Piccart-Gebhart MJ et al (2008). Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. J Clin Oncol; 26(12):1987-92.

Buyse M, Molenberghs G (1998). Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics; 54:1014–1029.

Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics; 1(1):49-67.

Buyse M (2009). Use of meta-analysis for the validation of surrogate endpoints and biomarkers in cancer trials. Cancer J; 15(5):421-5.

Buyse M, Sargent DJ, Grothey A, Matheson A, de Gramont A (2010). Biomarkers and surrogate end points--the challenge of statistical validation. Nat Rev Clin Oncol; 7(6):309-17.

Buyse M, Michiels S, Squifflet P et al (2011). Leukemia-free survival as a surrogate end point for overall survival in the evaluation of maintenance therapy for patients with acute myeloid leukemia in complete remission. Haematologica; 96(8):1106-1112.

Caldwell DM, Ades AE, Higgins PT (2005). Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ; 331:897-900.

Carroll K, Chaudri Ross H, Evans D et al (2008). Conditional approval: discussion points from the PSI conditional approval expert group. Pharmaceutical Statistics; 7: 263–269.

Chakravarty G, Rothmann M, Sridhara R (2011). Regulatory Issues in use of Biomarkers in Oncology Trials. Statistics in Biopharmaceutical Research; 3(4):569-576.

Chuang-Stein, C (2006). Sample size and the probability of a successful trial. Pharmaceutical Statistics; 5:30–309.

Chuang-Stein C, Kirby, S, French, F et al (2011). A quantitative approach for making go/no go decisions in drug development. Drug Information Journal; 45:187–202.

Cox DR (1972). Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B;34:187–220.

Denne JS, Stone AM, Bailey-Iacona R, Chen T-T (2013). Missing Data and Censoring in the Analysis of Progression-Free Survival in Oncology Clinical Trials, Journal of Biopharmaceutical Statistics; 23(5):951-970.

DerSimonian R, Laird N (1986). Meta-Analysis in Clinical Trials. Controlled Clinical Trials; 7:177-188.

Edwards SJ, Clarke MJ, Wordsworth S, Borrill J (2009). Indirect comparisons of treatments based on systematic reviews of randomised controlled trials. Int J Clin Pract; 63(6):841–854.

EMA (2000). Committee for Proprietary Medicinal Products for Human Use (CPMP). Points to Consider on Switching Between Superiority and Non-Inferiority. EMEA London. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003658.pdf [last accessed October 2013].

EMA (2005). Committee for Medicinal Products for Human Use (CHMP). Guideline on the Choice of the Non-inferiority Margin. EMEA London. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf [last accessed October 2013].

EMA (2006). European Public Assessment Report, Alimta: Initial Marketing Authorisation Scientific Discussion. EMEA, London. Available at http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Scientific_Discussion/human/000564/WC500025606.pdf [last accessed October 2013].

EMA (2009a). Acomplia: Withdrawal of the marketing authorisation in the European Union. EMA, London. Available at

http://www.ema.europa.eu/docs/en_GB/document_library/Public_statement/2009/11/WC500012189.pdf [last accessed October 2013].

EMA (2009b). European Assessment Report Iressa. EMA, London. Available at http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/001016/WC500036361.pdf [last accessed October 2013].

EMA (2011). Xigris (drotrecogin alfa (activated)) to be withdrawn due to lack of efficacy. EMA, London. Available at http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2011/10/news_detail_001373.jsp&mid=WC0b01ac058004d5c1 [last accessed October 2013].

EMA (2012a). CHMP Guideline on the evaluation of anticancer medicinal products in man. EMA, London. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/01/WC500137128.pdf [last accessed October 2013].

EMA (2012b). Appendix 1 to the CHMP guideline on the evaluation of anticancer medicinal products in man. EMA, London. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2011/12/WC500119965.pdf [last accessed October 2013].

Eskens FA, Verweij J (2000). Clinical studies in the development of new anticancer agents exhibiting growth inhibition in models: facing the challenge of a proper study design. Crit Rev Oncol Hematol;34:83–8.

FDA (1999). CBER FDA Memorandum, 1999. Summary of CBER Considerations on Selected Aspects of Active Controlled Trial Design and Analysis for the Evaluation of Thrombolytics in Acute MI.

FDA (2004a). Public Health Advisory: Safety of Vioxx. FDA, September 2004. http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm106274.htm [last accessed October 2013].

FDA (2004b). Oncologic Drugs Advisory Committee: Colorectal cancer endpoint discussion. 4th May 2004. Transcript page 312. Available at http://www.fda.gov/ohrms/dockets/ac/04/transcripts/4037T2.DOC [last accessed October 2013].

FDA (2004c). Oncologic Drugs Advisory Committee: Colorectal cancer endpoint discussion. 4th May 2004. Available at

http://www.fda.gov/ohrms/dockets/ac/04/slides/4037s2.htm [last accessed October 2013].

FDA (2004d). Oncologic Drug Products Advisory Committee meeting. 27 July 2004. Transcript, page 12. Available at: http://www.fda.gov/ohrms/dockets/ac/04/transcripts/2004-4060T1.htm [last accessed October 2013].

FDA (2007). Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics, 2007. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071590.pdf [last accessed October 2013].

FDA (2010a). New and Generic Drug Approvals. HERCEPTIN (trastuzumab) Intravenous Infusion product labeling, October 2010. Available at http://www.accessdata.fda.gov/drugsatfda_docs/label/2010/103792s5256lbl.pdf [last accessed October 2013].

FDA (2010b). Guidance for Industry Non-Inferiority Clinical Trials. FDA, 2010. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf [last accessed October 2013].

FDA (2011a). Center for Drug Evaluation and Research. Brilinta, Product Label. Food and Drug Administration 2011. Available at http://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/022433Orig1s000Lbl.pdf [last accessed October 2013].

FDA (2011b). Center for Drug Evaluation and Research. Brilinta, Statistical Review. Food and Drug Administration. Available at http://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/022433Orig1s000StatR.pdf [last accessed October 2013].

FDA (2011c). Center for Drug Evaluation and Research. Brilinta, Summary Review. Food and Drug Administration 2011. Available at http://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/022433Orig1s000SumR.pdf [last accessed October 2013].

FDA (2013a). New and Generic Drug Approvals. GLEEVEC (imatinib mesylate) tablets for oral use product labeling, February 2013. Available at http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/021588s038lbl.pdf {last accessed October 2013].

Fleming TR, DeMets DL (1996). Surrogate end points in clinical trials: are we being misled? Annals of Internal Medicine; 125(7):605–613.

Fleming TR (2008). Current issues in non-inferiority trials. Statistics in Medicine; 27(3):317–332.

Fleming TR, Rothmann MD, Lu LH (2009). Issues in Using Progression-Free Survival When Evaluating Oncology Products. J Clin Oncol; 27:2874-2880.

Fleming TR, Powers JH (2012). Biomarkers and surrogate endpoints in clinical trials. Statist in Med; 31:2973–2984.

Follmann DA, Proschan MA (1999). Valid Inference in Random Effects Meta-Analysis. Biometrics; 55(3):732–737.

Freedman LS, Graubard BI, Schatzkin A (1992). Statistical validation of intermediate endpoints for chronic diseases. Statistics in Medicine; 11:167–178.

Freidlin B, Simon R (2005). Adaptive Signature Design: An Adaptive Clinical Trial Design for Generating and Prospectively Testing A Gene Expression Signature for Sensitive Patients. Clin Cancer Res;11(21) :7872-7878.

Hasler M, Vonk R, Hothorn LA (2008). Assessing non-inferiority of a new treatment in a three-arm trial in the presence of heteroscedasticity. Statist. Med.; 27:490–503.

Hasselblad V, Kong DF (2001). Statistical Methods for Comparison to Placebo in Active-Control Trials. Drug Information Journal; 35:435–449.

Hauck WW, Anderson S (1999). Some Issues in the Design and Analysis of Equivalence Trials. Drug Information Journal; 33:109–118.

Herbst RS, Giaccone G, Schiller JS, et al (2004). Gefitinib in combination with paclitaxel and carboplatin in advanced non-small-cell lung cancer: a phase III trial- INTACT 2. J Clin Oncol; 22:785–94.

Holmgren EB (1999). Establishing Equivalence by Showing That a Specified Percentage of the Effect of the Active Control Over Placebo Is Maintained. Journal of Biopharmaceutical Statistics; 9:651–659.

Hung J, Wang S-J, Tsong Y, Lawrence J, O'Neil R (2003). Some fundamental issues with non-inferiority testing in active controlled trials. Statistics in Medicine; 22:213–225.

Hung J, Wang S-J, O'Neil R (2005). A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. Biometrical Journal; 47(1):28–36.

ICH E9 (1998). Note for Guidance on Statistical Principles for Clinical Trials. ICH Technical Coordination, EMEA: London, 1998. Available at

http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf last accessed October 2013].

ICH E10 (2000). Note for Guidance on Choice of Control Group for Clinical Trials. ICH Technical Coordination, EMEA: London, 2000. Available at http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E10/Step4/E10_Guideline.pdf [last accessed October 2013].

Jenkins M, Stone A, Jennison C (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. Pharm. Stat; 10:347–356.

Jones B, Roger J, Lane P et al (2011). On behalf of PSI Health Technology Special Interest Group, Evidence Synthesis sub-team. Statistical approaches for conducting network meta-analysis in drug development. Pharmaceut. Statist.; 10:523–531.

Julious JA, Swank DJ (2005). Moving statistics beyond the individual clinical trial: applying decision science to optimize a clinical development plan. Pharmaceut. Statist.; 4:37–46.

Keiding N, Andersen PK, Klein JP (1997). The role of frailty models and accelerated failure-time models in describing heterogeneity due to omitted covariates. Statist. Med.;16:215–224

Kirby S, Burke J, Chuang-Stein C, Sin C (2012). Discounting phase 2 results when planning phase 3 clinical trials. Pharmaceutical Statistics; 11:373–385.

Koch A, Röhmel J (2004). Hypothesis Testing in the "Gold Standard" Design for Proving the Efficacy of an Experimental Treatment Relative to Placebo and a Reference. Journal of Biopharmaceutical Statistics; 14(2):315-325.

Korn EL, Arbuck SG, Pluda JM et al (2001). Clinical trial designs for cytostatic agents: are new approaches needed? J. Clin. Oncol.;19:265–72.

Korn EL, Freidlin B, Abrams JS (2011). Overall Survival As the Outcome for Randomized Clinical Trials With Effective Subsequent Therapies. J. Clin. Oncol.; 29:2439-2442.

Lange S, Freitag G (2005). Special Invited Papers Section: Therapeutic Equivalence—Clinical Issues and Statistical Methodology in Noninferiority Trials: Choice of Delta: Requirements and Reality—Results of a Systematic Review. Biometrical Journal; 47:12–27.

Laporte S, Squifflet P, Baroux N et al (2013). Prediction of survival benefits from progression-free survival benefits in advanced non-small-cell lung cancer: evidence from a meta-analysis of 2334 patients from 5 randomised trials. BMJ Open 2013;3:e001802.

Lilly (2011). Lilly Announces Withdrawal of Xigris Following Recent Clinical Trial Results. Eli Lilly and Company, October 2011. Available at http://newsroom.lilly.com/releasedetail.cfm?releaseid=617602 [last accessed October 2013].

Lu G, Ades AE (2004). Combination of direct and indirect evidence in mixed treatment comparisons. Statist. Med.;23:3105-3124.

Lumley T (2002). Network meta-analysis for indirect treatment comparisons. Statist. Med.;21:2313-2324.

MHRA (2004). Immediate withdrawal of rofecoxib. MHRA, London, September 2004. Available at http://www.mhra.gov.uk/Safetyinformation/Safetywarningsalertsandrecalls/Safetywarningsandmessagesformedicines/CON1004263 [last accessed October 2013].

MHRA (2010). Rosiglitazone (Avandia, Avandamet): Withdrawal from clinical use. MHRA, London, September 2010. Available at http://www.mhra.gov.uk/home/groups/pl-p/documents/websiteresources/con094122.pdf [last accessed October 2013].

Mielke M, Munk A, Schacht A (2008). The assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints. Statist. Med.; 27:5093–5110.

Molenberghs G, Geys H, Buyse M (2001). Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. Statistics in Medicine; 20:3023–3038.

Molenberghs G, Buyse M, Geys H, et al (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. Controlled Clinical Trials; 23:607–625.

Molenberghs G, Burzykowski T, Alonso A, Buyse M (2004). A perspective on surrogate endpoints in controlled clinical trials. Stat Methods Med Res; 13(3):177-206.

Molenberghs G, Burzykowski T, Alonso A, Assam P, Tilahun A, Buyse M (2010). A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials. Stat Methods Med Res; 19(3):205-36.

Munzel U (2009). Nonparametric non-inferiority analyses in the three-arm design with active control and placebo. Statistics in Medicine; 28(29):3643-56.

Nikolakopoulos S, van der Wal WM, Roes KCB (2013). An Analytical Approach to Assess the Predictive Value of Biomarkers in Phase II Decision Making. Journal of Biopharmaceutical Statistics; 23(5):1106-1123.

O'Hagan A, Stevens JW, Campbell MJ (2005). Assurance in clinical trial design. Pharmaceutical Statistics; 4:18–201.

Peterson P. Carroll K, Chuang-Stein C et al (2010). PISC Expert Team White Paper: Toward a Consistent Standard of Evidence When Evaluating the Efficacy of an Experimental Treatment From a Randomized, Active-Controlled Trial, Statistics in Biopharmaceutical Research; 2(4):522-531.

Pignatti F, Hemmings R, Jonsson B (2011). Is it time to abandon complete blinded independent central radiological evaluation of progression in registration trials? European Journal of Cancer; 47:1759–1762.

Prentice RL (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. Statistics in Medicine; 8:431–440.

Quan H, Zhou D, Mancini P et al (2012). Adaptive Patient Population Selection Design in Clinical Trials. Statistics in Biopharmaceutical Research; 4(1):86-99.

Rimawi M, Hilsenbeck SG (2012). Making sense of clinical trial data: is inverse probability of censoring weighted analysis the answer to crossover bias? J Clin Oncol; 30(4):453-8

Robins J, Tsiatis A (1991). Correcting for non-compliance in randomised trials using rank preserving structure failure time model. Communications in Statistics - Theory and Methods; 20:2069-2631.

Robins JM, Finkelstein DM (2000). Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics; 56:779-788.

Rothmann M (2005). Type I Error Probabilities Based on Design Stage Strategies With Applications to Noninferiority Trials. Journal of Biopharmaceutical Statistics; 15:109–127.

Rothmann M, Li N, Chen G, Chi GYH, Temple R, Tsou H-H (2003). Design and analysis of non-inferiority mortality trials in oncology. Statistics in Medicine; 22:239–264.

Rothmann M, Koti K, Lee KY et al (2013). Evaluating and Adjusting for Premature Censoring of Progression-Free Survival, Journal of Biopharmaceutical Statistics; 23(5):1091-1105.

Saad ED, Katz A, Hoff PM, Buyse M (2010). Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. Ann Oncol; 21(1):7-12.

Sargent DJ, Wieand HS, Haller DG et al (2005). Disease-Free Survival Versus Overall Survival As a Primary End Point for Adjuvant Colon Cancer Studies: Individual Patient Data From 20,898 Patients on 18 Randomized Trials. J Clin Oncol; 23:8664-8670.

SAS Institute Inc. 2011. SAS/STAT® 9.3 User's Guide. Cary, NC, USA: SAS Institute Inc.

Senn S (1996). Some statistical issues in project prioritization in the pharmaceutical industry. Statist. Med.; 15:2689-2702.

Senn S (2005). 'Equivalence is different' – some comments on therapeutic equivalence. Biometrical Journal; 47(1):104–107.

Simon R (1989). Optimal Two-Stage Designs for Phase II Clinical Trials. Controlled Clinical Trials; 10: 1-10.

Simon R (1999). Bayesian design and analysis of active control clinical trials. Biometrics; 55:484–487.

Simon R, Steinberg SM, Hamilton M, et al (2001). Clinical trial designs for the early clinical development of therapeutic cancer vaccines. J Clin Oncol;19:1848–54.

Snapinn S (2004). Alternatives for Discounting in the Analysis of Noninferiority Trials. Journal of Biopharmaceutical Statistics; 14:263–273.

Sposto R (2002). Cure model analysis in cancer: an application to data from the Children's Cancer Group. Statist. Med.; 21:293–312.

Stadler WM, Ratain MJ (2000). Development of target-based antineoplastic agents. Invest New Drugs;18:7–16.

Stallard N, Whitehead J, Cleal S (2005). Decision-making in a phase II clinical trial: a new approach combining Bayesian and frequentist concepts. Pharmaceut. Statist.; 4:119–128.

Stone AM, Bushnell W, Denne, J et al (2011). Research outcomes and recommendations for the assessment of progression in cancer clinical trials from a PhRMA working group. European Journal of Cancer; 47:1763–1771.

Su Z (2010). Assessing the success probability of a phase III clinical trial based on phase II data. Contemporary Clinical Trials; 31:620–623.

Sun J, Zhao Q, Zhao X (2005). Generalized Log-Rank Tests for Interval-Censored Failure Time Data. Scand J Statist; 32:49–57.

Sun X, Li X, Chen C, Song Y (2013). A Review of Statistical Issues with Progression-Free Survival as an Interval-Censored Time-to-Event Endpoint. Journal of Biopharmaceutical Statistics; 23(5):986-1003.

Sutton A, Ades AE, Cooper N et al (2008). Use of indirect and mixed treatment comparisons for technology assessments. Pharmacoeconomics;26:753-767.

Tang PA, Bentzen SM, Chen EX, Siu LL (2007). Surrogate end points for median overall survival in metastatic colorectal cancer: literature-based analysis from 39 randomized controlled trials of first-line chemotherapy. J Clin Oncol.;25(29):4562-8.

Temkin NR (1978). An analysis for transient states with application to tumour shrinkage. Biometrics 1978;34:571–80.

Turnbull BW (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. J R Stat Soc Ser B; 38:290-295.

Wallentin, L. et al (2009). Ticagrelor Versus Clopidogrel in Patients With Acute Coronary Syndromes. New England Journal of Medicine; 361:1045–1057.

Wang S-J, Hung J, Tsong Y (2002). Utility and pitfalls of some statistical methods in active controlled clinical trials. Controlled Clinical Trials; 23:15–28.

Wang S-J, Hung J (2003). TACT method for noninferiority testing in active controlled trial. Statistics in Medicine; 22:227–238.

Wang S-J, O'Neill RT, Hung HM (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat; 6(3):227-44.

Wei LJ, Glidden DV (1997). An overview of statistical methods for multiple failure-time data in clinical trials. Statist Med; 16:833–839.

Westfall P, Young S (1993). Resampling based multiple testing: examples and methods for p-value adjustment. New York: Wiley; 1993.

Westfall P, Young S, Wright P (1993). On adjusting p-value for multiplicity. Biometrics, 1993; 49(3): 941-945.

Williams G, He K, Chen G, Chi G, Pazdur R (2002). Operational bias in assessing time to progression (TTP). Proc Am Assoc Cancer Res 2002 (abst 975).

Williford WO et al (1987). The 'Constant Intake Rate' assumption in interim recruitment goal methodology for multicenter clinical trials. Journal of Chronic Diseases; 40:297–307.

Wirth MP, Hakenberg OW, Froehner M (2008). Adjuvant hormonal treatment - the bicalutamide early prostate cancer program. Front Radiat Ther Oncol; 41:39-48.

Zhang X, Sun J (2010). Regression analysis of clustered interval-censored failure time data with informative cluster size. Computational Statistics & Data Analysis; 54(7): 1817–1823.

Zhao YD, Dmitrienko A, Tamura R (2010). Design and Analysis Considerations in Clinical Trials With a Sensitive Subpopulation. Statistics in Biopharmaceutical Research; 2(1):72-83.

Appendix: Supporting Publications

1. Newling D, **Carroll K** and Morris T. Is prostate-specific antigen progression a surrogate for objective clinical progression in early prostate cancer? 2004, Journal of Clinical Oncology, ASCO Annual Meeting Proceedings (Post-Meeting Edition). Vol 22, No 14S (July 15 Supplement), 2004: 4652. Available at http://meeting.ascopubs.org/cgi/content/abstract/22/14_suppl/4652 [last accessed October 2013]

Journal of Clinical Oncology, 2004 ASCO Annual Meeting Proceedings (Post-Meeting Edition).
Vol 22, No 14S (July 15 Supplement), 2004: 4652
© 2004 American Society of Clinical Oncology

Abstract

Is prostate-specific antigen progression a surrogate for objective clinical progression in early prostate cancer?

D. Newling, K. Carroll and T. Morris

AstraZeneca, Macclesfield, United Kingdom

4652

Background: It is a common misconception that a correlation between endpoints is enough to demonstrate surrogacy. To show true surrogacy, the effect of an intervention on an intermediate endpoint relative to a control treatment needs to reliably predict the effect of the intervention on the clinical outcome of interest. Valid surrogate endpoints are needed to accelerate availability of information about new therapies for early prostate cancer where clinical progression and survival times are prolonged. The usefulness of prostate-specific antigen (PSA) progression as a surrogate for objective clinical progression has therefore been assessed in the bicalutamide ('Casodex') Early Prostate Cancer (EPC) Program, the world's largest prostate cancer treatment program. **Methods:** Individual data from all 8113 patients across 21 countries in the EPC Program were examined. Time to PSA progression and time to objective clinical progression were the endpoints evaluated. The effect of treatment on time to PSA progression was compared with the effect of treatment on time to objective clinical progression. **Results:** Analyses suggest that time to PSA progression is a modest surrogate endpoint for the effect of a hormonal intervention on objectively confirmed progression in patients with early prostate cancer (r^2 = 0.52–0.65, $p<0.001$). An intervention that produces around a 50% reduction in the risk of PSA progression would likely result in around a 20% reduction in the risk of objective clinical progression. **Conclusions:** The effect of treatment on PSA progression is moderately predictive for the effect of hormonal treatment on objective clinical progression. A large effect on PSA progression predicts for a smaller, but nonetheless clinically important effect on objective clinical progression. These data suggest that a large positive treatment effect on time to PSA progression is reasonably likely to reflect a clinically important delay in objective clinical progression, making PSA progression a valid endpoint for the evaluation of hormonal medicines in early prostate cancer.

Is prostate-specific antigen a surrogate for objective clinical progression in early prostate cancer?

Newling D, Carroll K, Morris T

AstraZeneca
Macclesfield, Cheshire, UK

Introduction

- Surrogate endpoints may aid the development of prostate cancer therapies
- The biomarker prostate-specific antigen (PSA) is a promising potential surrogate for prostate cancer progression
- Surrogacy requires that the treatment effect on PSA can predict the treatment effect on objective clinical progression

Publication 1: Newling, Carroll, Morris 2004.

Objective

To determine whether PSA progression may be a surrogate endpoint for clinical disease progression in patients with early non-metastatic prostate cancer, using data from 8113 patients in the bicalutamide ('Casodex') 150 mg Early Prostate Cancer (EPC) program

Publication 1: Newling, Carroll, Morris 2004.

Publication 1: Newling, Carroll, Morris 2004.

The bicalutamide 150 mg EPC program

- **Three geographically distinct trials conducted across 21 countries (Trials 23, 24 and 25)**
- **Examining bicalutamide 150 mg/day (n=4052) or placebo (n=4061) in addition to standard care**
- **Endpoints**
  - overall survival
  - time to objectively confirmed disease progression (progression-free survival)[1]
  - time to PSA progression[2]

fined as [1]time from randomization to earliest occurrence of objective progression or death from any use without progression; [2]time between randomization and earliest occurrence of PSA doubling from seline, objective progression, or death from any cause in the absence of progression

Publication 1: Newling, Carroll, Morris 2004.

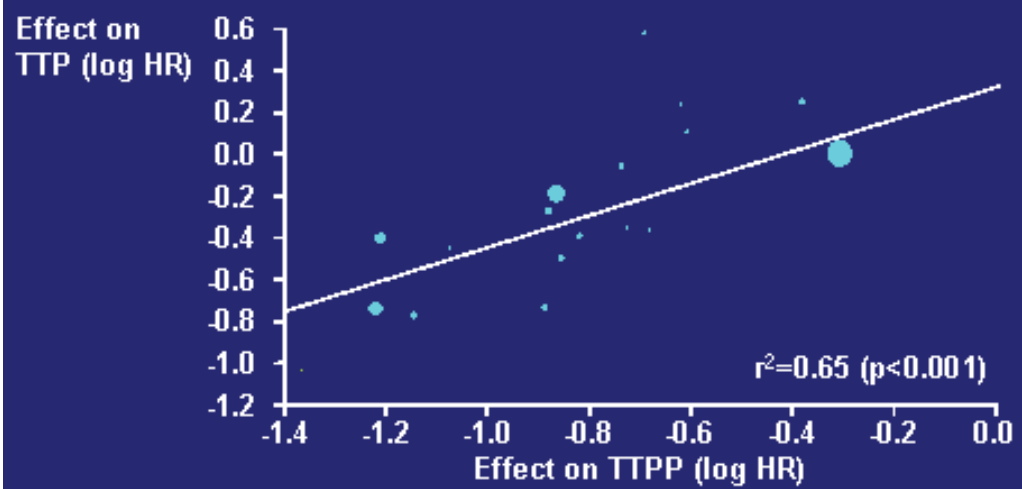
Methods

- Clinical endpoint: time to objectively confirmed disease progression (TTP)
- Surrogate endpoint: time to PSA progression (TTPP)
- Previously accepted meta-analytic methodology[1] for the assessment of intermediate endpoints and potential surrogates used
- Relative treatment effects on TTP and TTPP estimated by region
- Control analysis performed which excluded data from largest region (Trial 23, conducted in USA and Canada)

[1]Buyse & Molenberghs. *Biometrics* 1998; 54: 1014-29 / Buyse et al. *Biostatistics* 2000; 1: 49-67

Publication 1: Newling, Carroll, Morris 2004.

Significant correlation between the effects of bicalutamide on TTP and TTPP



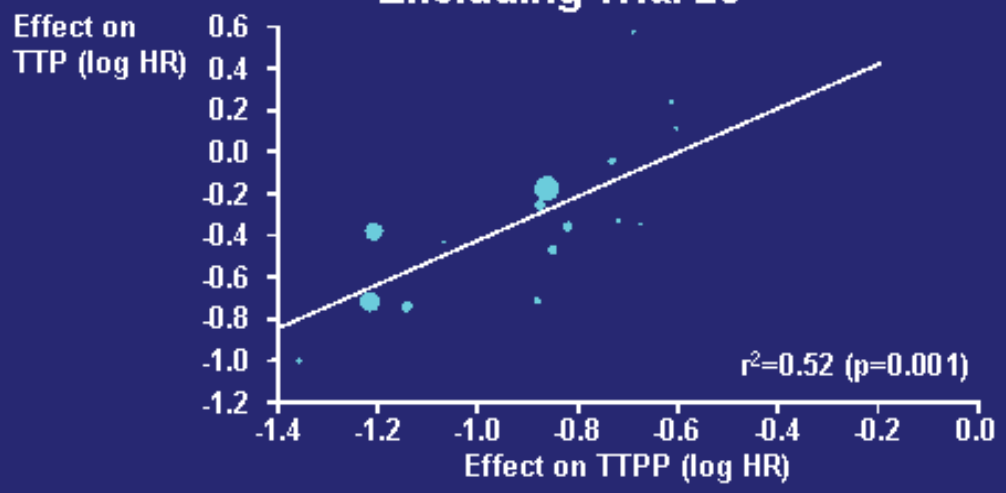
Points represent observations in each region; area is proportionate to sample size

HR, hazard ratio

Publication 1: Newling, Carroll, Morris 2004.

Significant correlation between the effects of bicalutamide on TTP and TTPP

Excluding Trial 23

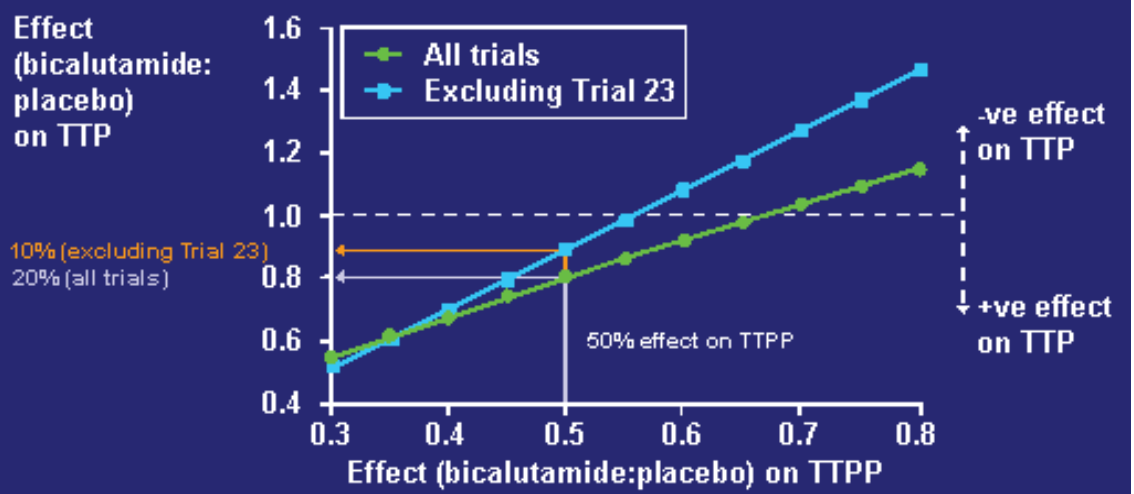


Points represent observations in each region; area is proportionate to sample size

HR, hazard ratio

Publication 1: Newling, Carroll, Morris 2004.

Prediction of treatment effect on TTP from effect observed on TTPP



50% reduction in risk of PSA progression results in a 10-20% reduction in risk of objective clinical progression

Conclusions

In early prostate cancer

- The effect of hormonal treatment on the surrogate endpoint of PSA progression is moderately predictive for the effect on objective clinical progression
- A large positive effect on time to PSA progression is reasonably likely to reflect a clinically important delay in objective clinical progression

Publication 1: Newling, Carroll, Morris 2004.

2. Collette L, Burzykowski T, **Carroll K**, Newling D, Morris T and Schroder F. Is prostate-specific antigen a surrogate for survival in advanced prostate cancer? Journal of Clinical Oncology, 2004 ASCO Annual Meeting Proceedings (Post-Meeting Edition). Vol 22, No 14S (July 15 Supplement), 2004: 4551. Available at http://meeting.ascopubs.org/cgi/content/abstract/22/14_suppl/4551 [last accessed October 2013]

Journal of Clinical Oncology, 2004 ASCO Annual Meeting Proceedings (Post-Meeting Edition).
Vol 22, No 14S (July 15 Supplement), 2004: 4551
© 2004 American Society of Clinical Oncology

Abstract

Is prostate-specific antigen a surrogate for survival in advanced prostate cancer?

L. Collette, T. Burzykowski, K. Carroll, D. Newling, T. Morris and F. Schroder

EORTC Data Center, Brussels, Belgium; Limburgs Universitair Centrum, Diepenbeek, Belgium; AstraZeneca, Macclesfield, United Kingdom; Erasmus Medical Centrum, Rotterdam, Netherlands

4551

Background: Surrogate endpoints are needed to shorten the duration of Phase III clinical trials in advanced prostate cancer. Prostate-specific antigen (PSA) is the most studied biomarker in prostate cancer. This study attempts to validate PSA endpoints as surrogates for overall survival (OS). **Methods:** Individual data from 2161 patients with advanced prostate cancer treated in studies comparing bicalutamide ('Casodex'), either as monotherapy or in combination with an LHRHa, with castration were used in a meta-analytical approach to surrogate endpoint validation. PSA response, several definitions of time to PSA progression and longitudinal PSA measurements were considered. **Results:** The analyses confirmed the known association between the PSA endpoints and OS at the individual patient level (biomarker association). However, when comparing patients treated with bicalutamide-based treatment or castration, the effect of hormonal intervention on the PSA endpoint did not predict the effect on OS with a high degree of precision. The association between intervention on any PSA endpoint and OS, measured by the determination coefficient R2 (ranging from 0.10–0.66 for PSA progression, to 0.69 for the whole PSA profile) was generally low. **Conclusions:** It is a common misconception that a correlation at the individual level between PSA and OS is enough to demonstrate surrogacy. To demonstrate true surrogacy, a high correlation between the treatment effect on the surrogate and the treatment effect on the true endpoint needs to be established across groups of patients treated with two alternative interventions. The level of association observed in this study between the treatment effect on PSA endpoints and that observed on OS was in general low, showing that in Phase III clinical trials of hormonal treatments in advanced prostate cancer, treatment effects on OS cannot be predicted from observed treatment effects on PSA endpoints with a high degree of precision. This study indicates that PSA is unlikely to be a useful surrogate for OS in advanced prostate cancer. 'Casodex' is a trademark of the AstraZeneca group of companies.

Publication 2: Collette, Burzykowski, Carroll 2004

Is prostate-specific antigen a surrogate for survival in advanced prostate cancer?

Collette L,[1] Burzykowski T,[2] Carroll K,[3] Newling D,[3] Morris T,[3] Schröder F[4]

[1]EORTC Data Center, Brussels, Belgium
[2]Limburgs Universitair Centrum, Diepenbeek, Belgium
[3]AstraZeneca, Macclesfield, UK
[4]Erasmus Medical Centrum, Rotterdam, The Netherlands

Publication 2: Collette, Burzykowski, Carroll 2004

Aims

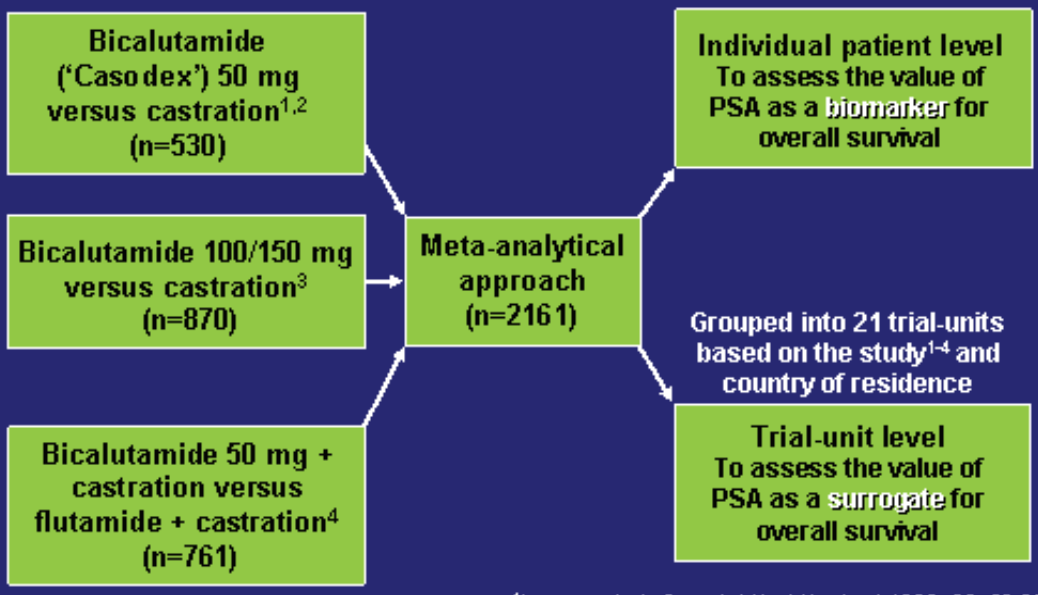
- Use of a surrogate endpoint in clinical trials would speed the development of new prostate cancer therapies
- In this study, various prostate-specific antigen (PSA) endpoints were assessed as surrogates for overall survival in advanced (metastatic) prostate cancer

Biomarker:
an intermediate outcome that is correlated with the true clinical outcome at the individual patient level

Surrogate:
a biomarker that is intended to serve as a substitute for the true endpoint in comparative treatment trials. It should allow prediction of the effect of a therapeutic intervention on the true endpoint with sufficient precision

Publication 2: Collette, Burzykowski, Carroll 2004

Meta-analytical approach



[1]Iversen et al. *Scand J Urol Nephrol* 1996; 30: 93-98
[2]Kaisary et al. *Eur Urol* 1995; 28: 215-222; [3]Tyrrell et al. *Eur Urol* 1998; 33: 447-456
[4]Schellhammer et al. *Urology* 1995; 45: 745-752

Publication 2: Collette, Burzykowski, Carroll 2004

Potential surrogate endpoints for overall survival

Potential surrogate endpoint	Definition
PSA response	A PSA decline from baseline level (≥20 ng/mL) ≥50% at 2 subsequent observations ≥4 weeks apart
Time to PSA progression-1	Time to a ≥20% increase above the nadir and which exceeded the upper normal limit (ie >4 ng/mL)
Time to PSA progression-2	Time to an increase ≥50% above the moving average (based on 3 consecutive measurements) nadir and which exceeded 2.5 times the upper normal limit (ie a level >10 ng/mL). This increase had to be either the last observed value or be sustained for ≥4 weeks
Longitudinal PSA profile	The complete series of PSA measurements in each patient

Publication 2: Collette, Burzykowski, Carroll 2004

Statistical methods

- The relative treatment effects on the survival endpoint (log survival ratio) and on the various PSA endpoints were estimated using the meta-analytical validation approach[1]
- The squared correlation between these treatment effects (R^2_{trial}) was estimated from the slope of the regression line

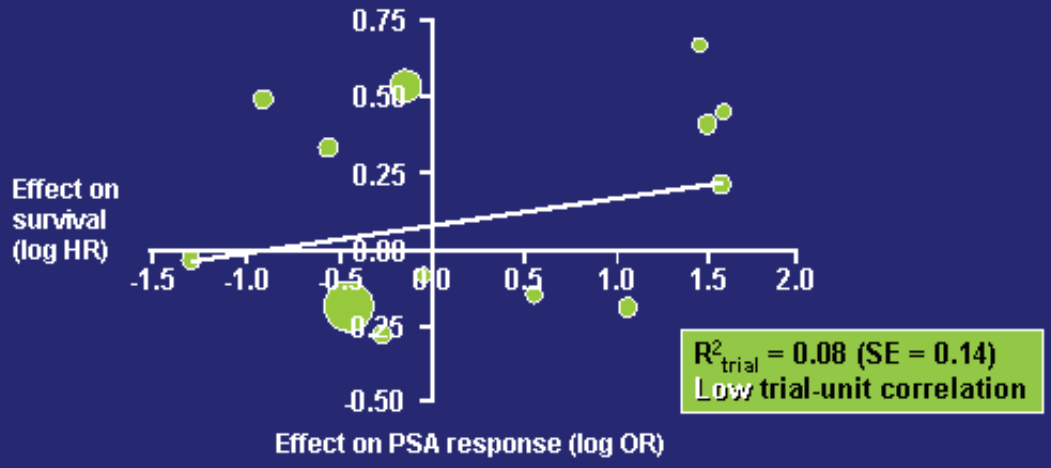
R^2_{trial} close to 1 = proof of surrogacy
(ie a precise prediction of the treatment effect on survival from the treatment effect on the PSA endpoint)

[1]Buyse et al. *Biostatistics* 2000; 1: 49-68

Estimated treatment effects on the PSA response

Trial-unit level

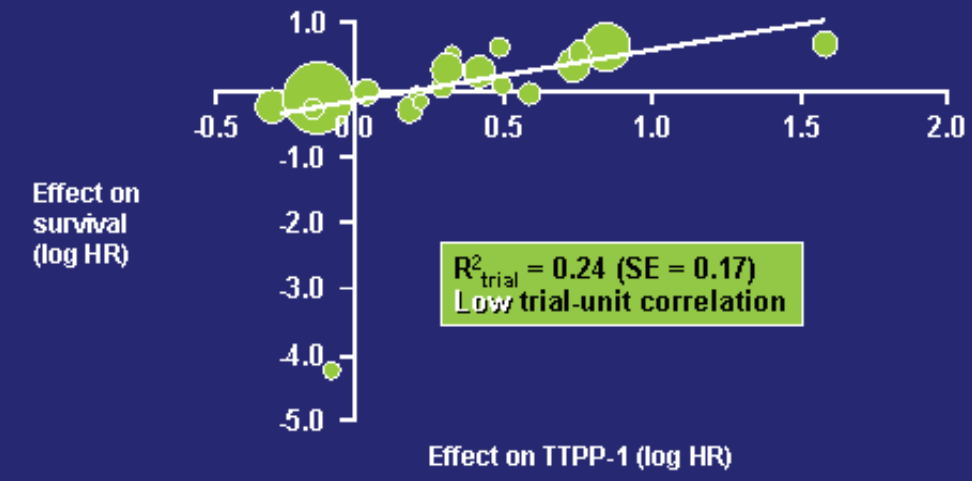
Publication 2: Collette, Burzykowski, Carroll 2004



Survival HR was analyzed for 13 trial-units (n=1606)
Each circle represents an individual trial-unit and their size is proportionate to the sample size
The line represents the prediction from an estimated (weighted) regression line
HR, hazard ratio; OR, odds ratio; PSA, prostate-specific antigen; SE, standard error

Estimated treatment effects on survival against TTPP-1

Trial-unit level

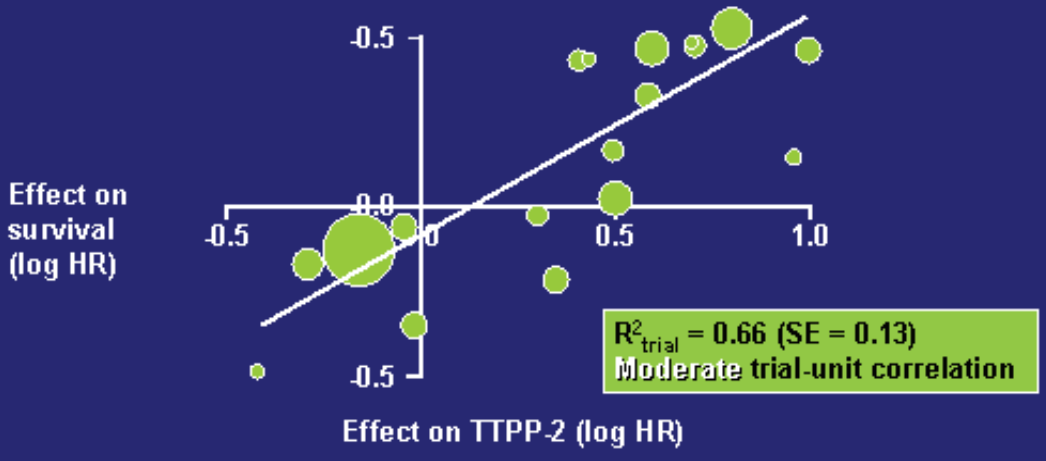


TTPP-1 was analyzed for 19 trial-units (n=2070)
Each circle represents an individual trial-unit and their size is proportionate to the sample size
The line represents the prediction from an estimated (weighted) regression line
HR, hazard ratio; SE, standard error; TTPP-1, time to prostate-specific antigen progression-1

Publication 2: Collette, Burzykowski, Carroll 2004

Estimated treatment effects on survival against TTPP-2

Trial-unit level



TTPP-2 was analyzed for 18 trial-units (n=2043)
Each circle represents an individual trial-unit and their size is proportionate to the sample size
The line represents the prediction from an estimated (weighted) regression line
HR, hazard ratio; SE, standard error; TTPP-2, time to prostate-specific antigen progression-2

Publication 2: Collette, Burzykowski, Carroll 2004

Correlation between longitudinal PSA and overall survival

Trial-unit level

Publication 2: Collette, Burzykowski, Carroll 2004

- The mean profiles of log-transformed PSA measurements for groups of patients with similar observation times showed a quadratic curvature
- PSA profiles were therefore modeled as a function of time and the square root of time
- At the trial-unit level:

R^2_{trial} = 0.68 (SE = 0.12)
Moderate trial-unit correlation

Association between PSA endpoints and survival*

Publication 2: Collette, Burzykowski, Carroll 2004

	Individual patient correlation	Trial-unit correlation
PSA response	High	Low
Time to PSA progression-1	Moderate	Low
Time to PSA progression-2	Moderate	Moderate
Longitudinal PSA	High	Moderate

True surrogacy = a high correlation between the treatment effect on the surrogate and the treatment effect on the true endpoint (overall survival), which needs to be established across groups of patients treated with the new and standard interventions

*Median 3.25 years' follow-up

Conclusions

- At the individual patient level, the analyses confirm the known association between PSA endpoints and overall survival, and thus the value of PSA as a biomarker
- At the trial-unit level, the association between PSA-based endpoints and overall survival was generally low to moderate
- Overall survival cannot therefore be reliably predicted across groups of patients on the basis of PSA
- Analysis using prostate cancer survival as the true endpoint led to similar findings

PSA is unlikely to be a valid surrogate for overall survival for use in Phase III clinical trials of hormonal treatments in advanced prostate cancer

Publication 2: Collette, Burzykowski, Carroll 2004

3. Collette L, Burzykowski T, **Carroll KJ** et al. Is prostate-specific antigen a valid surrogate end point for survival in hormonally treated patients with metastatic prostate cancer? Joint Research of the European Organisation for Research and Treatment of Cancer, the Limburgs Universitair Centrum, and AstraZeneca Pharmaceuticals. 2005, Journal of Clinical Oncology; 23:6139–6148.

VOLUME 23 · NUMBER 25 · SEPTEMBER 1 2005

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Is Prostate-Specific Antigen a Valid Surrogate End Point for Survival in Hormonally Treated Patients With Metastatic Prostate Cancer? Joint Research of the European Organisation for Research and Treatment of Cancer, the Limburgs Universitair Centrum, and AstraZeneca Pharmaceuticals

Laurence Collette, Tomasz Burzykowski, Kevin J. Carroll, Don Newling, Tom Morris, and Fritz H. Schröder

From the European Organisation for Research and Treatment of Cancer Data Center, Brussels; Limburgs Universitair Centrum, Diepenbeek, Belgium; AstraZeneca Pharmaceuticals, Macclesfield, United Kingdom; and Erasmus Medical Centrum, Rotterdam, the Netherlands.

Submitted August 27, 2004; accepted May 11, 2005.

Authors' disclosures of potential conflicts of interest are found at the end of this article.

Address reprint requests to Laurence Collette, MSc, European Organization for Research and Treatment of Cancer Data Center, Avenue Emmanuel Mounier 83/11, B-1200 Brussels, Belgium; e-mail: laurence.collette@eortc.be.

© 2005 by American Society of Clinical Oncology

0732-183X/05/2325-6139/$20.00

DOI: 10.1200/JCO.2005.08.156

ABSTRACT

Purpose
The long duration of phase III clinical trials of overall survival (OS) slows down the treatment-development process. It could be shortened by using surrogate end points. Prostate-specific antigen (PSA) is the most studied biomarker in prostate cancer (PCa). This study attempts to validate PSA end points as surrogates for OS in advanced PCa.

Patients and Methods
Individual data from 2,161 advanced PCa patients treated in studies comparing bicalutamide to castration were used in a meta-analytic approach to surrogate end-point validation. PSA response, PSA normalization, time to PSA progression, and longitudinal PSA measurements were considered.

Results
The known association between PSA and OS at the individual patient level was confirmed. The association between the effect of intervention on any PSA end point and on OS was generally low (determination coefficient, < 0.69).

Conclusion
It is a common misconception that high correlation between biomarkers and true end point justify the use of the former as surrogates. To statistically validate surrogate end points, a high correlation between the treatment effects on the surrogate and true end point needs to be established across groups of patients treated with two alternative interventions. The levels of association observed in this study indicate that the effect of hormonal treatment on OS cannot be predicted with a high degree of precision from observed treatment effects on PSA end points, and thus statistical validity is unproven. In practice, non-null treatment effects on OS can be predicted only from precisely estimated large effects on time to PSA progression (TTPP; hazard ratio, < 0.50).

J Clin Oncol 23:6139-6148. © 2005 by American Society of Clinical Oncology

INTRODUCTION

Phase III cancer clinical trials that evaluate the clinical benefit of new treatment options often require large patient numbers and long follow-up. Recent advances in the understanding of the biologic mechanisms of disease development have resulted in the emergence of a large number of potentially effective new agents. There is also increasing public pressure for promising new drugs to receive marketing approval as rapidly as

Information downloaded from jco.ascopubs.org and provided by at BOEHRINGER on June 30, 2013 from 148.188.1.60
Copyright © 2005 American Society of Clinical Oncology. All rights reserved.

possible, in particular for life-threatening diseases such as cancer. For these reasons, there is an urgent need to find ways of shortening the duration of cancer clinical trials. The duration of phase III trials results from the use of long-term clinical end points (clinical progression and survival). Therefore, to replace this end point (the "true" end point) by another ("surrogate") end point that could be measured earlier, more conveniently, or more frequently and would adequately reflect the benefit of new treatments on the clinical end point(s) seems to be an attractive solution.

"Biomarkers" (ie, physical signs or laboratory measurements that occur in association with a pathological process or that have putative diagnostic and/or prognostic utility[1]) are generally regarded as the best candidate surrogate end points. A biomarker is an intermediate outcome that is correlated with the true clinical outcome for an individual patient. It may be useful for diagnostic or prognostic information on a particular patient. It is a common misconception that established biomarkers necessarily make valid surrogate end points. To this aim, it is required that "the effect of treatment on a surrogate end point must be reasonably likely to predict clinical benefit."[2] Thus, "surrogacy" is a concept that relates to groups of patients. To demonstrate surrogacy, a strong association between the treatment effects on the surrogate and on the true end point needs to be established across groups of patients treated with the new and standard interventions.

The validation of a candidate surrogate end point is not straightforward. Until recently, the statistical methods developed for this purpose used the data from a single trial.[3-5] These methods suffer from numerous drawbacks: some of them are too stringent to be of practical value, whereas others are based on nontestable assumptions.[6,7] To overcome these limitations, a new methodology, known as the "meta-analytic" validation approach, was developed recently.[8-10] This method uses large databases from multiple randomized clinical trials and aims at measuring directly the association between the treatment effects on the surrogate and the true end point.

In the field of prostate cancer (PCa), prostate-specific antigen (PSA) has probably been the most studied biomarker. It has been investigated as a potential surrogate end point across disease stages,[11-14] and in hormone-refractory patients in particular.[15-18] In a recent article, Buyse et al[19] considered several PSA-based end points in androgen-independent patients treated with liarozole (an imidazole-like compound that causes elevation of retinoic acid, postulated to have antitumor activity), cyproterone acetate, or flutamide. They showed that despite a strong association at the individual patient level, none of the end points qualified as a surrogate for overall survival (OS). In early PCa, Newling et al[20] found a modest correlation between the effect of Casodex on time to PSA progression (TTPP) and on objectively confirmed progression. In primary metastatic PCa, several studies demonstrated some level of association between a post-therapy fall in PSA or a PSA relapse on treatment and long-term survival prognosis.[21-25] However, this merely qualifies PSA as a biomarker. In trial NCI-INT-105, treatment differences in post-therapy PSA levels did not translate into survival differences.[26] Thus, whether PSA is a valid surrogate for survival in hormonally treated PCa remains an open question. This question is of importance, because the use of PSA could shorten the time to the end point from between several months in advanced disease[27] to several years in early disease.[28]

The objective of the present research is to assess PSA-based end points as surrogates for OS in hormone-naïve metastatic PCa using the meta-analytic approach. The data from > 2,000 patients treated with bicalutamide (Casodex) that were made available by AstraZeneca Pharmaceuticals were used for this purpose.

PATIENTS AND METHODS

Individual data from three large international randomized trials of AstraZeneca's Casodex Development Program were used (301/302,[29,30] 306/307,[31] and US trial 0001[32,33]; Table 1). In studies 301/302 and 306/307, Casodex monotherapy (50 and 150 mg/day, respectively) was compared to medical or surgical castration. In the US trial, Casodex (50 mg/day) in combination with goserelin or leuprolide acetate was compared to the combination of flutamides (750 mg/day) and castration in a 2 × 2 factorial design. All patients were newly diagnosed with metastatic PCa. Four hundred eighty patients with T3-4 M0 disease and elevated PSA from trial 306/307 were excluded. Survival was an end point in all studies, although time to treatment failure (Table 1) was the primary end point in most. PSA was monitored at months 1, 2 (except US trial), and 3 and then every 3 months until month 18 (trial 301/302) or death (other trials). For the analysis, the PSA test date was assumed to be the visit date.

End Points

We considered OS calculated from randomization to the date of death or last visit as the true end point. PCa-specific survival was defined similarly but with deaths unrelated to PCa or treatment censored at the last visit. PSA response, PSA normalization, TTPP, and the complete series of PSA measurements ("PSA profile") were successively assessed as potential surrogate end points for OS.

Patients who had a baseline PSA level at least five times above the normal range (> 20 ng/mL) were included in the analyses of PSA response and PSA normalization. Patients qualified for PSA response if their PSA declined by at least 50% from baseline level at two subsequent observations at least 4 weeks apart. Patients in whom the decline reached a value below or equal to normal (4 ng/mL) qualified for PSA normalization.[25]

Two definitions of TTPP were assessed: (1) For TTPP-1, PSA progression was defined as a PSA value above normal (4 ng/mL), representing a first increase ≥ 20% above the nadir[25] (eg, with a PSA nadir of 2 ng/mL, a minimum increase to 4 ng/mL [100% increase] is required, whereas with a PSA nadir of 3.5 ng/mL, a 20% increase to 4.2 ng/mL is enough). (2) For TTPP-2, PSA progression was defined as a PSA value > 2.5 times the normal range (10 ng/mL), representing a first increase ≥ 50% above the moving average (based on three

Information downloaded from jco.ascopubs.org and provided by at BOEHRINGER on June 30, 2013 from 148.188.1.60
Copyright © 2005 American Society of Clinical Oncology. All rights reserved.

Table 1. Trials Used in the Analysis

Trial	
301/302[28,29]	
Patients	Stage D2, fit for orchidectomy; ECOG performance status 0-2; no prior systemic therapy for prostate cancer, no previous radiotherapy to the prostate within 3 months of entry
Treatments	Bicalutamide (50 mg/d) *v* castration (orchidectomy in trial 301, orchidectomy or goserelin 3.6 mg monthly injection in trial 302)
Design	Open two-arm randomization
Objective	To compare bicalutamide to castration in a pooled analysis
Efficacy end points	Time to treatment failure (objective progression, change of treatment, death as a result of any cause)*; overall survival
Results	Bicalutamide (50 mg/d) demonstrated significantly worse time to progression and survival in trial 301; the trend was not significant in trial 302; by pooled analysis, both end points were significantly worse with bicalutamide than with castration
306/307[30]	
Patients	Metastatic (M1) or locally advanced with PSA five-fold in excess of the upper normal limit (T3-4 M0); only the M1 patients were included in the presently reported analyses; fit for orchidectomy; ECOG performance status 0-2; no prior systemic therapy for prostate cancer, no previous radiotherapy to the prostate within 3 months of entry
Treatments	Bicalutamide (100 or 150 mg/d) or castration (medical or surgical at the patient's discretion)
Design	Initially 2 (Casodex 100 mg):2 (Casodex 150 mg):1 (castration) then changed to 2:1 randomization between Casodex 150 mg and castration
Objective	To demonstrate noninferiority of Casodex 150 mg in comparison to castration by excluding a risk increase of 25%
Efficacy end points	Time to treatment failure (addition of systemic therapy or withdrawal from therapy, objective progression, or death)*; overall survival; objective response
Results (in M1)	Significant differences in favor of castration were found for time to treatment failure (HR, 1.43; 95% CI, 1.20 to 1.71 in favor of castration) and overall survival (HR, 1.30; 95% CI, 1.04 to 1.64)
US trial[31,32]	
Patients	Stage D2 only; ECOG performance status 0-2; no prior systemic therapy for prostate cancer
Treatments	Bicalutamide (50 mg/d) *v* flutamide (250 mg tid) in combination with goserelin acetate (3.6-mg monthly injection) or leuprolide acetate (7.5-mg monthly injection)
Design	2 × 2 factorial design, blinding for the LHRH-A randomization
Objective	To demonstrate noninferiority of bicalutamide + LHRH-A relative flutamide + LHRH-A by excluding a relative-risk increase of 25%
Efficacy end points	Time to treatment failure (addition of systemic therapy or withdrawal from therapy, objective progression, or death)*; overall survival
Results	Noninferior time to treatment failure (HR, 0.93 in favor of bicalutamide; 95% CI, 0.79 to 1.10) Noninferior overall survival (HR, 0.87 in favor of bicalutamide; 95% CI, 0.72 to 1.05)

Abbreviations: ECOG, Eastern Cooperative Oncology Group; PSA, prostate-specific antigen; HR, hazard ratio; LHRH-A, luteinizing hormone-releasing hormone agonist.
*A rising PSA was not considered a sign of progression in any of the studies.

consecutive measurements) nadir. This increase had to be either the last observed value or be sustained for at least 4 weeks[19] (eg, with a nadir of 2 ng/mL at three consecutive occasions, a 500% increase to 10 ng/mL is needed to reach the end point, whereas after a nadir of 7 ng/mL, a 50% increase to 10.5 ng/mL is enough).

Patients who died or are alive without PSA progression were censored at the time of death or last visit, respectively.

Statistical Methods

The meta-analytic approach to surrogate end-point validation has been detailed extensively elsewhere.[6,9,34-36] Thus, we shall only summarize the key features. The method is rooted in the concept that a valid surrogate end point must enable one to predict with sufficient precision the treatment effect on the true clinical end point (OS) from the observed treatment effect on the surrogate (PSA-based) end point. Unlike traditional validation methods such as the Prentice criteria,[3] this new methodology does not require that any of those effects be statistically significant. Indeed, when data from several trials are available, the method consists of simultaneously estimating the relative treatment effects on the survival end point and on the PSA end point (log odds ratio of PSA response or normalization, log hazard ratio [HR] of PSA progression, treatment effect on the longitudinal PSA measurements) in each trial. A model that estimates the association between the treatment effects on the true end point and the corresponding effects on the PSA end points (PSA response,[34] TTPP,[35] or longitudinal PSA measurements[36]) in a way similar to standard linear regression (although mathematically more sophisticated) is then adjusted. As in linear regression, the strength of the association is measured by the squared correlation coefficient that we shall denote R^2_{trial}. This coefficient also indicates the precision with which the treatment effect on the survival end point can be predicted from the observed treatment effect on the surrogate. The maximal possible value of R^2_{trial} is 1, which indicates a perfect prediction. In practice, observing $R^2_{trial} = 1$ is not possible, and one rather seeks a value close to 1, which indicates a strong association between the treatment effects and thus a relatively precise prediction.[9,35] Additionally, the model quantifies the association between the PSA-based end point and the survival end point at the individual patient level. Parameters quantifying the strength of the association at this level will be denoted by the subscript "patient." They can be regarded as measures of validity of the PSA end point as a biomarker for predicting duration of survival.

Only three trials were available, which is too few to allow a precise estimation of R^2_{trial}. Therefore, the patients were grouped

Information downloaded from jco.ascopubs.org and provided by at BOEHRINGER on June 30, 2013 from 148.188.1.60
Copyright © 2005 American Society of Clinical Oncology. All rights reserved.

by the trial they entered and their country of residence, as done by Buyse et al.[19] These groups will be henceforth referred to as "trial units."

RESULTS

After excluding nonmetastatic patients and those with no baseline or follow-up PSA measurements, the individual data from 2,161 patients classified into 21 trial units were available for the analysis (Table 2). Their baseline and treatment characteristics are listed in Table 3. More than half of the patients presented with six or more bone metastases. After a median follow-up of 3.25 years, 1,018 patients (52.9%) had died, 815 (71.3%) as a result of PCa (Table 4). The median OS was 2.2 years (95% CI, 2.1 to 2.5) for the Casodex-treated patients and 2.3 years (95% CI, 2.1 to 2.6) in the pooled control groups (Fig 1). The average number of PSA assessments per patient was 6.9 (range, 1 to 23)

PSA Response (≥ 50% Decline From Baseline) and PSA Normalization

PSA response could be assessed for 1,853 patients. A total of 974 (89.4%) and 687 (90.0%) assessable patients on the Casodex and control groups, respectively, achieved a PSA response (Table 4). Only thirteen trial units representing 1,606 patients were used in the analysis: two trial units were removed because no deaths were observed in the castration group, and six were removed because all patients responded in one or both treatment arms. At the individual level, PSA response was a strong predictor of prolonged survival with a survival odds ratio θ_{patient} of 1.94 (SE, 0.33), representing a two-fold increase in the odds of surviving beyond any specified time t for the PSA responders compared to the nonresponders. At the trial level, the effects of hormonal intervention on PSA response and on OS were poorly correlated with $R^2_{trial} = 0.08$ (SE, 0.14; 95% CI, 0.0 to 0.49). Figure 2A presents the estimated treatment effects on the response (log odds ratio) and OS (log HR).

One should be careful in interpreting these results, because eight trial units with extreme results were excluded from the analysis.

In 399 (36.6%) and 380 (49.8%) of the assessable patients, the PSA declined to a value ≤ 4 ng/mL. Seventeen trial units representing 1,778 patients could be used for this analysis: four were excluded for same reasons as above. At the individual level, the survival odds ratio θ_{patient} for patients with PSA normalization compared to those without was 4.90 (SE, 0.52), indicating a 4.9-fold greater odds of surviving any specified time t for the patients whose PSA normalized. At the trial level, the treatment effects on PSA and on OS were moderately correlated with $R^2_{trial} = 0.41$ (SE, 0.18; 95% CI, 0.05 to 0.72; Table 5). Figure 2B presents the estimated treatment effects on PSA normalization and OS.

Table 2. Trial Units Available for the Analysis (N = 2,161)

Trial	Country	N
US	Canada	114
US	United States	647
301	Denmark	158
301	Norway	75
301	Sweden	63
302	Austria	46
302	The Netherlands	29
302	United Kingdom	159
306	Denmark	83
306	Finland	69
306	Norway	83
306	Sweden	86
307	Australia	35
307	Austria	14
307	Belgium	95
307	Germany	47
307	The Netherlands	35
307	Italy	11
307	Republic of South Africa	48
307	Spain	22
307	United Kingdom	242

PSA Progression

Nineteen trial units (2,070 patients) and 18 trial units (2,043 patients) could be used for the analysis of TTPP-1 and TTPP-2, respectively (two trial units were excluded from both analyses because of absence of deaths in the castration arm and one from the TTPP-2 analysis because of the absence of PSA progressions in both treatment arms).

The TTPP-1 is presented in Figure 3A: 54.6% of the patients progressed according to this definition (Table 4) within a median time of 11.1 months after being randomly assigned. TTPP-1 was somewhat shorter for the pooled Casodex group than for the control group. TTPP-1 was moderately associated with OS at the individual patient level: the concordance coefficient $\tau_{\text{patient}} = 0.52$ (SE, 0.004) indicates that for each individual patient there is an approximately 50% chance to observe a long (short) OS given a long (short) TTPP. At the trial-unit level, the association between the effects of Casodex on TTPP-1 and on OS was low, with $R^2_{trial} = 0.21$ (SE, 0.17; 95% CI, 0.0 to 0.56; Table 5). This analysis is depicted in Figure 4A, where the treatment effect on survival is regressed against the treatment effect on TTPP-1: the size of the circles represents the trial-unit size. The low trial-level association may be partly because of the outlying data from one trial unit. Excluding this unit from the analysis leaves the individual-level association unchanged ($\tau_{\text{patient}} = 0.52$; SE, 0.004) but increases R^2_{trial} to 0.58 (SE, 0.15; 95% CI, 0.20 to 0.81).

Only 31.8% of the patients met the more stringent criterion TTPP-2 (Table 4) at a median time of 24.9 months (Fig 3B). At the patient level, the association of TTPP-2 and OS was somewhat stronger than for TTPP-1, with a

Information downloaded from jco.ascopubs.org and provided by at BOEHRINGER on June 30, 2013 from 148.188.1.60
Copyright © 2005 American Society of Clinical Oncology. All rights reserved.

Table 3. Patient Characteristics

	Age					Baseline PSA			
	Mean	SD	Median	First and Third Quartiles	Performance Status 0/1/2/3/4, %	Mean	SD	Median	First and Third Quartiles
301/302									
Total (N = 530)			Data not available		Data not available	839.1	1,551.3	267.9	98.6, 784.7
Casodex 50 mg (N = 262)			Data not available		Data not available	811.2	1,477.8	273.2	98.3, 840.0
Castration (N = 268)			Data not available		Data not available	866.3	1,622.2	266.7	99.4, 713.3
306/307 (UICC M1 pts.)									
Total (N = 870)	71.6	8.2	72	66, 78	53.8/32.8/13.3/0/0.1	747.3	1,657.2	179.1	65.7, 634.7
Casodex 100/150 mg (N = 617)	71.2	8.2	72	66, 77	54.0/31.9/14.1/0/0	772.6	1,772.5	189.8	64.5, 658.4
Castration (N = 253)	72.7	8.1	73	67, 78	53.4/34.8/11.5/0/0.4	685.6	1,336.0	156.0	67.0, 587.3
US (D2 pts.)									
Total (761)	70.2	8.7	70	65, 76	51.4/37.2/11.4/0/0	694.2	1,444.2	174.3	45.6, 580.6
Casodex + castration (N = 377)	69.8	8.2	70	65, 75	53.8/36.1/10.1/0/0	650.4	1,382.8	170.0	53.8, 588.1
Flutamide + castration (N = 384)	70.5	9.2	71	65, 77	49.0/38.3/12.8/0/0	737.3	1,502.6	178.3	38.7, 576.5

Abbreviations: PSA, prostate-specific antigen; SD, standard deviation.

concordance coefficient $\tau_{\text{patient}} = 0.61$ (SE, 0.02). The association between the treatment effects on TTPP-2 and OS was somewhat higher than for TTPP-1, with $R^2_{trial} = 0.66$ (SE, 0.13; 95% CI, 0.30 to 0.85; Fig 4B and Table 5).

Longitudinal Measurements of PSA

All previously considered PSA-based end points are summary measures derived from the longitudinal PSA measurements and use only a limited amount of the available information. It thus seemed logical to investigate if the longitudinal series of PSA measurements would not be a better surrogate end point for OS. Figure 5A presents the mean profiles of log-transformed PSA measurements for groups of patients with similar observation time: all profiles eventually end with a PSA increase (progression), and patients with an early progression tend to have a higher initial PSA that does not decrease as much early on.

Figure 5B displays the mean PSA profiles per treatment group: starting from week 52 the curves show a relatively stable linear decrease rather than the increasing curvature observed in Figure 4A. This distortion results from attrition: progressive patients, in whom PSA increases, tend to leave the study, and thus the curve in Figure 5B reflects only those with stable PSA.

In view of Figure 5A, the treatment effect on the log-transformed PSA levels was expressed as a function of time and its square root in a joint model of PSA measurements and survival times. In that model, the individual patient-level association between the PSA process and the hazard of dying is a function of time and cannot be easily summarized into a single measure.[35] The results indicated that the correlation between the individual PSA and mortality hazard processes was > 0.90 at any time > 7 months, which suggests a strong association between the PSA profile and the hazard of dying for individual

Table 4. Survival and Prostate-Specific Antigen Outcome

	Casodex (n = 1,256)		Control (n = 905)		Total (N = 2,161)	
	No.	%	No.	%	No.	%
Alive	571		447		1,018	
Dead	685		458		1,143	52.9
Because of prostate cancer	496		319		815	71.3
Because of another cause	189		139		328	28.7
PSA response						
Evaluable	1,090		763		1,853	
Decline to ≤ 4 ng/mL	399	36.6	380	49.8	779	42.0
Decline by ≥ 50% of baseline	575	52.8	307	40.2	882	47.6
No response	116	10.6	76	10.0	192	10.4
Not evaluable	142		166		308	
PSA progression (TTPP-1)	415		729		1,144	
PSA progression (TTPP-2)	432		233		665	
Not evaluable for PSA progression	35		32		67	

Abbreviations: PSA, prostate-specific antigen; TTPP, time to PSA progression.

Information downloaded from jco.ascopubs.org and provided by at BOEHRINGER on June 30, 2013 from 148.188.1.60
Copyright © 2005 American Society of Clinical Oncology. All rights reserved.

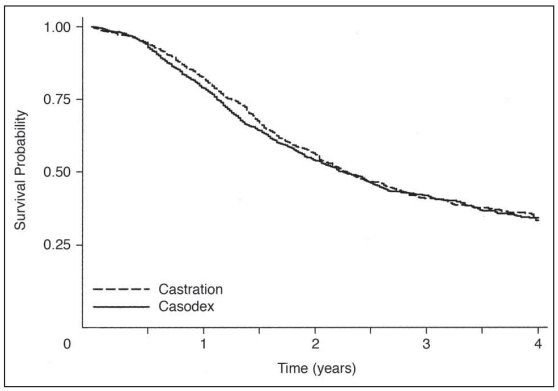


Fig 1. Overall survival by randomized treatment.

patients. At the trial-unit level, the association between the effect of Casodex on the longitudinal PSA and OS was slightly higher than that for TTPP-2 ($R^2_{trial} = 0.68$; SE, 0.12; Table 5).

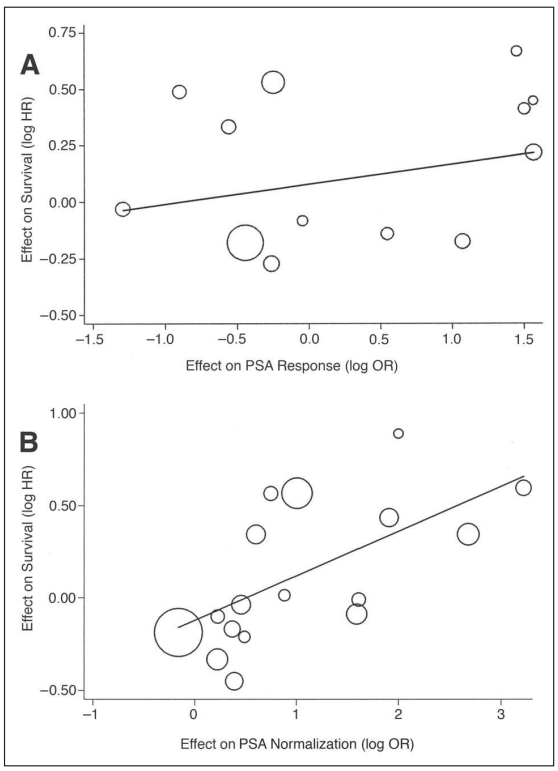


Fig 2. The treatment effects on survival and prostate-specific antigen (PSA) response. The circles represent the observations in the trial units, and their size is proportionate to the trial-unit sample size. The line represents the prediction from an estimated (weighted) regression line. (A) ≥ 50% decline from baseline level: $R^2_{trial} = 0.08$. (B) PSA normalization (PSA ≤ 4 ng/mL): $R^2_{trial} = 0.41$. HR, hazard ratio; OR, odds ratio.

DISCUSSION

Using data from the Casodex Development Program, we investigated whether the biomarker PSA could be used to define a valid surrogate for OS in patients with metastatic PCa. The analyses confirm the known value of PSA as a biomarker of prognosis and disease activity (individual-level association). When comparing groups of patients treated with Casodex-based or control treatment, however, the association between the treatment effect on any PSA-based end point and the treatment effect on OS was low in general ($R^2_{trial} < 0.69$ with wide confidence intervals).

The choice of the threshold for R^2_{trial} required for a valid surrogate is still a matter of debate.[6] Nevertheless, one can argue that the precision of the prediction of the treatment effect on OS from the effect on the PSA-based end points, indicated by the R^2_{trial} values observed in the present study, is insufficient to claim any of the assessed PSA-based end points as a statistically valid surrogate end point for OS in phase III clinical trials of hormonal treatment in metastatic PCa.

To illustrate the problem, let us consider a new trial with TTPP as the primary end point (defined as TTPP-2), where data analysis occurs after 400 events and yields an HR of 0.75 for PSA progression (with 400 events, SE [log{HR}] would be of the order of 0.10, resulting in $P < .01$). Without adjusting for the estimation error in the parameters of the prediction model, one could predict with approximately 95% confidence that the corresponding survival HR would lie within the interval 0.48 to 1.12. Adjustment for the estimation error would widen the confidence interval even further; thus, non-null treatment effects on survival would potentially be identifiable only in large new trials showing a large effect on the PSA end point (eg, HR approximately 0.50 with SE = 0.10).

Buyse et al[19] assessed similar PSA-based end points as candidate surrogates for OS in patients with androgen-independent PCa treated with liarozole versus antiandrogen monotherapy. In their study, the association between treatment effects at the trial level were generally low, with $R^2_{trial} < 0.45$ for all tested PSA end points. They concluded that PSA end points could not be regarded as valid surrogates for OS. The reasons for the lack of association in their study may be different than ours; the disease was more advanced, and treatment mode of action differed. In early disease, for which the time savings of using PSA could be greater than in advanced disease, Newling at al[20] also found only moderate correlation between the effect of Casodex on PSA progression and objective clinical progression.

Unfortunately, in cancer and other diseases, biomarkers that are strong predictors of the clinical end point for the individual patient often proved to be poor surrogate end points.[37-43] Several authors have discussed biologic and medical reasons why biomarkers often fail to validate as

Information downloaded from jco.ascopubs.org and provided by at BOEHRINGER on June 30, 2013 from 148.188.1.60
Copyright © 2005 American Society of Clinical Oncology. All rights reserved.

Table 5. Summary of the Results

PSA End Point	Patient-Level Association Between PSA and Survival	SE	Trial-Level Association Between PSA and Survival R^2_{trial}	SE	95% CI
PSA response (decline by ≥ 50% from baseline)	$\theta_{patient} = 1.94$	0.33	0.08	0.14	0 to 0.49
PSA normalization (≤ 4 ng/mL)	$\theta_{patient} = 4.90$	0.52	0.41	0.18	0.05 to 0.72
TTPP-1	$\tau_{patient} = 0.52$	0.004	0.21	0.17	0 to 0.56
TTPP-2	$\tau_{patient} = 0.61$	0.02	0.66	0.13	0.30 to 0.85
Longitudinal PSA measurements	$R^2_{patient} > 0.9$ at all times > 7 mo	—	0.68	0.12	Undetermined

Abbreviations: PSA, prostate-specific antigen; TTPP, time to PSA progression; $\theta_{patient}$, survival odds ratio; $\tau_{patient}$, concordance coefficient between time to PSA progression and duration of survival.

surrogate end points.[2,37,39,44] The principal explanation is that only a part of the treatment effect on the true clinical end point will be reflected in the biomarker, which may lead to over- or underestimation of the treatment effect on the true end point from the observed effect on the biomarker. Baker and Kramer[45] mention that perfect predictors of the true end point at the patient level do not necessarily make good surrogate end points, because the prediction function could differ between randomized treatments and thus would induce incorrect inference on the true end point.

The inability thus far to demonstrate surrogacy for PSA can be explained by several biologic mechanisms. PSA is also produced by normal prostatic tissue, and the amount present may vary between patients. Poorly differentiated tumors

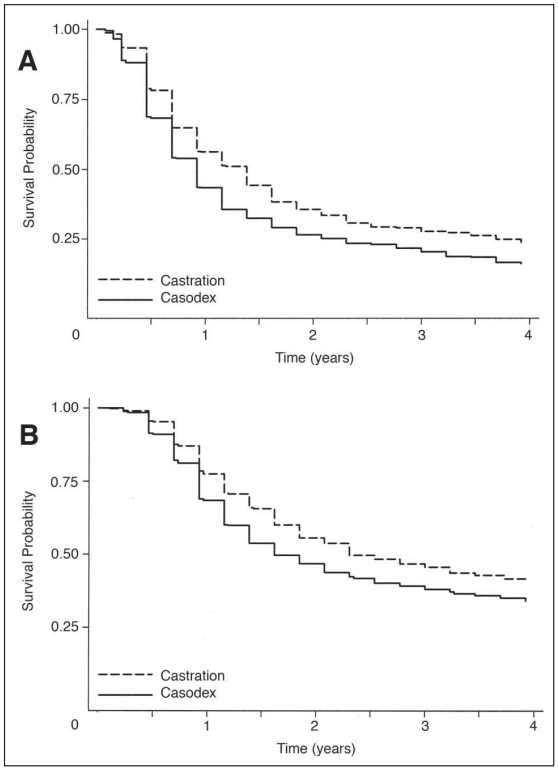


Fig 3. Time to prostate-specific antigen (PSA) progression (TTPP) by randomized treatment. (A) TTPP-1: time to the first 20% increase of PSA over previously observed nadir to a value above the upper limit of the normal PSA range (4 ng/mL). (B) TTPP-2: time to the first 50% increase of PSA over previously observed moving average nadir to a value > 2.5 times the upper limit of the normal PSA range (10 ng/mL), sustained for at least 4 weeks.

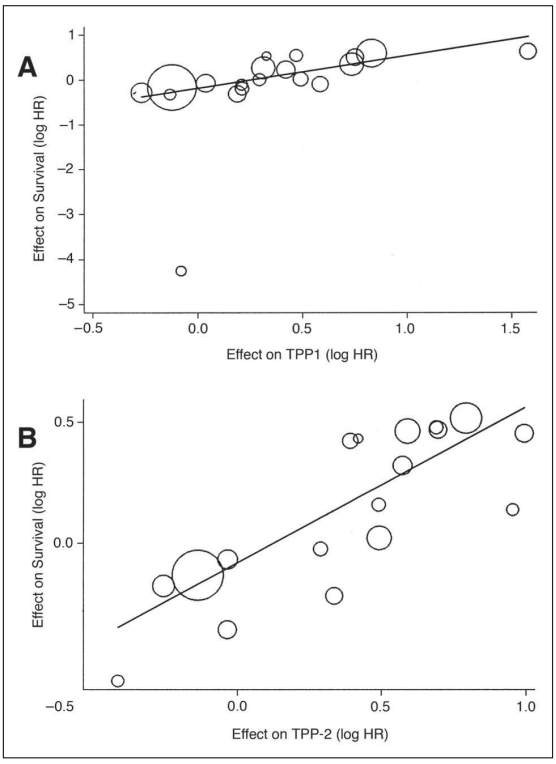


Fig 4. The treatment effects on time to prostate-specific antigen (PSA) progression (TTPP) and on overall survival. The circles represent the observations in the trial units, and their size is proportionate to the trial-unit sample size. The line represents the prediction from an estimated (weighted) regression line. (A) TTPP-1: $R^2_{trial} = 0.21$. (B) TTPP-2: $R^2_{trial} = 0.66$. HR, hazard ratio.

Information downloaded from jco.ascopubs.org and provided by at BOEHRINGER on June 30, 2013 from 148.188.1.60
Copyright © 2005 American Society of Clinical Oncology. All rights reserved.

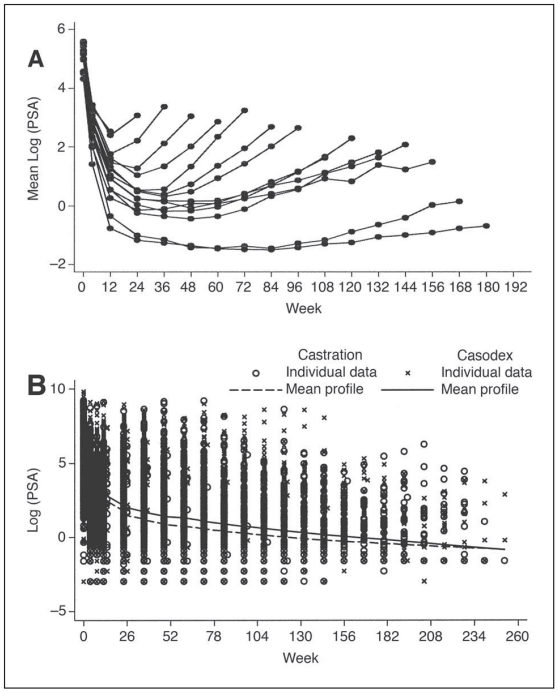


Fig 5. Log(prostate-specific antigen [PSA]) measurements over time. (A) Average log(PSA) profiles by drop-out time. (B) Average log(PSA) profile by treatment arm.

may produce proportionally less PSA for the level of tumor burden compared with better differentiated tumors. In addition, PSA is studied in the serum, although the source is prostatic tissue. Conceptually, serum levels can be related to unknown factors promoting or inhibiting leakage from PCa cells into the blood, to cellular levels of PSA and their interindividual variation, and obviously to the total tumor mass present in a given patient.[46,47] PSA in itself is an endocrine-dependent enzyme, and its expression is regulated by a promoter that contains androgen-responsive elements.[48,49] The treatment effects seen on PSA in trials of endocrine treatment of PCa thus may result, at least in part, from a direct, nontumor-mass–related effect. Such considerations led Scher et al[18] to conclude that PSA may not be an appropriate end point for clinical trials of first-line hormonal treatment.

Part of the imprecision in the prediction achieved in our study may be because of the limited number of observations available in each trial unit. The database we used, however, is the largest available. There is also some heterogeneity in trial design between the two monotherapy trials and the combined androgen-blockade trial. However, reanalysis excluding the latter did not change the results.

One also could argue about the use of overall rather than disease-specific survival as the true end point in our study. Analyses using disease-specific survival as the true end point, however, led to essentially similar conclusions (R^2_{trial} for TTPP-2 was then 0.49; SE, 0.17; 95% CI, 0.11 to 0.76).

Finally, it was not possible to assess dynamic measures of PSA such as PSA doubling time or PSA velocity in these analyses; as suggested by Kelloff et al[50] and D'Amico et al,[51] such measures of PSA may carry more information than the ones we could assess.

Our study indicates that PSA surrogacy could not be statistically validated in trials of hormonal treatments against metastatic PCa. However, if large effects on time to PSA end point (HR, < 0.50) could be demonstrated with high precision in a new trial, the results of the present study would still provide evidence of a likely non-null effect (upper bound of the 95% prediction interval for HR, < 1) on OS. This suggests that, in such an instance, TTPP could potentially serve as a basis for accelerated drug approval, together with other trial data documenting safety and other measures of patient benefit, until firm evidence on the basis of the true end point becomes available. Nevertheless, additional research for more powerful surrogate end points in PCa is still needed. Such research should probably focus on dynamic PSA measurements or new, hopefully more specific markers or combinations of markers.

Acknowledgment

We thank AstraZeneca Pharmaceuticals for permission to use the data from the Casodex Development Program from patients with advanced prostate cancer.

Information downloaded from jco.ascopubs.org and provided by at BOEHRINGER on June 30, 2013 from 148.188.1.60
Copyright © 2005 American Society of Clinical Oncology. All rights reserved.

Authors' Disclosures of Potential Conflicts of Interest

Although all authors completed the disclosure declaration, the following authors or their immediate family members indicated a financial interest. No conflict exists for drugs or devices used in a study if they are not being evaluated as part of the investigation. For a detailed description of the disclosure categories, or for more information about ASCO's conflict of interest policy, please refer to the Author Disclosure Declaration and the Disclosures of Potential Conflicts of Interest section in Information for Contributors.

Authors	Employment	Leadership	Consultant	Stock	Honoraria	Research Funds	Testimony	Other
Kevin J. Carroll	AstraZeneca			AstraZeneca (A)				
Don Newling			AstraZeneca (C)					
Tom Morris	AstraZeneca			AstraZeneca (A)				
Dollar Amount Codes (A) < $10,000 (B) $10,000-99,999 (C) ≥ $100,000 (N/R) Not Required								

REFERENCES

1. Lesko LJ, Atkinson AJ: Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: Criteria, validation, strategies. Annu Rev Pharmacol Toxicol 41:347-366, 2001

2. Biomarkers Definitions Working Group: Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clin Pharmacol Ther 69:89-95, 2001

3. Prentice RL: Surrogate endpoints in clinical trials: Definitions and operational criteria. Stat Med 8:431-440, 1989

4. Freedman LS, Graubard BI, Schatzkin A: Statistical validation of intermediate endpoints for chronic diseases. Stat Med 11:167-178, 1992

5. Buyse M, Molenberghs G: Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics 54:1014-1029, 1998 [Erratum: Biometrics 56:324, 2000]

6. Buyse M, Molenberghs G, Burzykowski T, et al: Statistical validation of surrogate endpoints: Problems and proposals. Drug Inf J 34:447-454, 2000

7. Molenberghs G, Buyse M, Geys H, et al: Statistical challenges in the evaluation of surrogate endpoints in randomized trials. Control Clin Trials 23:607-625, 2002

8. Daniels MJ, Hugues MD: Meta-analysis for the evaluation of potential surrogate markers. Stat Med 16:1515-1527, 1997

9. Buyse M, Molenberghs G, Burzykowski T, et al: The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics 1:49-68, 2000

10. Gail MH, Pfeiffer R, Van Houwelingen HC, et al: On meta-analytic assessment of surrogate outcomes. Biostatistics 1:231-246, 2000

11. Polascik TJ, Oesterling JE, Partin AW: Prostate specific antigen: A decade of discovery—what we have learned and were we are going. J Urol 162:293-306, 1999

12. Ornstein DK, Pruthi RS: Prostate-specific antigen. Expert Opin Pharmacother 1:1399-1411, 2000

13. Schroeder FH. Evidence in advanced disease: The challenge of new therapies. Eur Urol 1:11-16, 2002 (suppl)

14. Small EJ, Roach M III: Prostate-specific antigen in prostate cancer: A case study in the development of a tumor marker to monitor recurrence and assess response. Semin Oncol 29:264-273, 2002

15. Sridhara R, Eisenberger MA, Sinibaldi VJ, et al: Evaluation of prostate-specific antigen as a surrogate marker for response of hormone-refractory prostate cancer to suramin therapy. J Clin Oncol 13:2944-2953, 1995

16. Smith DC, Dunn RL, Stawderman MS, et al: Change in serum prostate-specific antigen as a marker of response to cytotoxic therapy for hormone-refractory prostate cancer. J Clin Oncol 16:1835-1843, 1998

17. Kelly WK, Scher HI, Mazumdar M, et al: Prostate-specific antigen as a measure of disease outcome in metastatic hormone-refractory prostate cancer. J Clin Oncol 11:607-615, 1993

18. Scher HI, Kelly WK, Zhang ZF, et al: Post-therapy serum prostate-specific antigen level and survival in patients with androgen-independent prostate cancer. J Natl Cancer Inst 91:244-251, 1999

19. Buyse M, Vangeneugden T, Bijnens L, et al: Validation of biomarkers and surrogates for clinical endpoints, in JC Bloom (ed): Biomarkers in Clinical Drug Development. New York, NY, Springer-Verlag, 2003, pp 149-168

20. Newling D, Carroll K, Morris T: Is prostate-specific antigen progression a surrogate for objective clinical progression in early prostate cancer? J Clin Oncol 22:4652, 2004 (suppl 15; abstr 4652)

21. Matzkin H, Soloway MS, Schellhammer PF, et al: Prognostic factors in stage D2 prostate cancer treated with a pure nonsteroidal antiandrogen. Cancer 72:1286-1290, 1993

22. Miller JI, Ahmann FR, Drach GW, et al: Serum PSA levels predict duration of remission and survival post hormone therapy of metastatic prostate cancer. J Urol 145:384A, 1991

23. Newling DW, Denis L, Vermeylen K. Orchiectomy versus goserelin and flutamide in the treatment of newly diagnosed metastatic prostate cancer. Analysis of the criteria of evaluation used in the European Organization for Research and Treatment of Cancer—Genitourinary Group Study 30853. Cancer 72:3793-3798, 1993 (suppl)

24. Cooper EH, Armitage TG, Robinson MRG, et al: Prostatic specific antigen and the prediction of prognosis in metastatic prostatic cancer. Cancer 66:1025-1028, 1990 (suppl 5)

25. Collette L, de Reijke TM, Schroeder FH, et al: Prostate specific antigen: a prognostic marker of survival in good prognosis metastatic prostate cancer? (EORTC 30892). Eur Urol 44:182-189, 2003

26. Eisenberger MA, Blumenstein BA, Crawford ED, et al: Bilateral orchiectomy with or without flutamide for metastatic prostate cancer. N Engl J Med 339:1036-1042, 1998

27. Newling D, Denis L, Vermeylen K: Orchiectomy versus goserelin and flutamide and the treatment of newly diagnosed metastatic prostate cancer. Cancer 72:3793-3798, 1993

28. Pound CR, Partin AW, Eisenberger MA, et al: Natural history of progression after PSA elevation following radical prostatectomy. JAMA 281:1591-1597, 1999

29. Iversen P, Tveter K, Varenhorst E: Randomised study of Casodex 50 mg monotherapy vs orchidectomy in the treatment of metastatic prostate cancer. The Scandinavian Casodex Cooperative Group. Scand J Urol Nephrol 30:93-98, 1996

30. Kaisary AV, Tyrrell CJ, Beacock C, et al: A randomised comparison of monotherapy with Casodex 50 mg daily and castration in the treatment of metastatic prostate carcinoma. Casodex Study Group. Eur Urol 28:215-222, 1995

31. Tyrrell CJ, Kaisary AV, Iversen P, et al: A randomised comparison of 'Casodex' (bicalutamide) 150 mg monotherapy versus castration in the treatment of metastatic and locally advanced prostate cancer. Eur Urol 33:447-456, 1998

32. Schellhammer P, Sharifi R, Block N, et al: A controlled trial of bicalutamide versus flutamide, each in combination with luteinizing hormone-releasing hormone analogue therapy, in patients with advanced prostate cancer. Casodex Combination Study Group. Urology 45:745-752, 1995

33. Schellhammer PF, Sharifi R, Block NL, et al: Clinical benefits of bicalutamide compared with flutamide in combined androgen blockade for patients with advanced prostatic carcinoma: Final report of a double-blind, randomized, multicenter trial. Casodex Combination Study Group. Urology 50:330-336, 1997

34. Burzykowski T, Molenberghs G, Buyse M: The validation of surrogate endpoints using data from randomized clinical trials: A case study in advanced colorectal cancer. J R Stat Soc [Ser A] 167:103-124, 2004

35. Burzykowski T, Molenberghs G, Buyse M, et al: Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. Appl Stat 50:405-422, 2001

36. Renard D, Geys H, Molenberghs G, et al: Validation of a longitudinally measured surrogate marker for time-to-event endpoint. J Appl Stat 30:235-247, 2003

Information downloaded from jco.ascopubs.org and provided by at BOEHRINGER on June 30, 2013 from 148.188.1.60
Copyright © 2005 American Society of Clinical Oncology. All rights reserved.

37. Fleming TR: Evaluating therapeutic interventions: Some issues and experiences. Stat Sci 7:428-456, 1992

38. Temple RJ. A regulatory authority's opinion about surrogate endpoints, in Nimmo W, Ticker G (eds), Clinical Measurement in Drug Evaluation. Chichester, United Kingdom, John Wiley, 1995

39. Fleming TR, DeMets DL: Surrogate end points in clinical trials: Are we being misled? Ann Intern Med 125:605-613, 1996

40. Buyse M, Thirion P, Carlson RW, et al: Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: A meta-analysis. Meta-Analysis Group in Cancer. Lancet 356:373-378, 2000

41. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators: Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. N Engl J Med 321:406-412, 1989

42. The Cardiac Arrhytmia Suppression Trial (CAST) II Investigators: Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. N Engl J Med 327:227-233, 1992

43. Burzykowski T, Molenberghs G, Buyse M, et al: The validation of surrogate endpoints using data from randomized clinical trials: A case study in advanced colorectal cancer. J R Stat Soc [Ser A] 167:103-124, 2004

44. Frank R, Hargreaves R: Clinical biomarkers in drug discovery and development. Nat Rev Drug Discov 2:566-580, 2003

45. Baker SG, Kramer BS: A perfect correlate does not a surrogate make. BMC Med Res Methodol 3:16, 2003

46. Stege RH, Tribukait B, Calrlstron KA, et al: Tissue PSA from fine-needle biopsies of prostatic carcinoma as related to serum PSA, clinical stage, cytological grade and DNA ploidy. Prostate 38:183-188, 1999

47. Schmid HP, McNeal JE, Stamey TA: Observations on the doubling time of prostate cancer. The use of serial prostate-cancer antigen in patients with untreated disease as a measure of increasing cancer volume. Cancer 71:2031-2040, 1993

48. Cleutjens KB, van Eekelen CC, van der Kroput HA, et al: Two androgen response regions cooperate in steroid hormone regulated activity of the prostate-specific antigen promoter. J Biol Chem 271:6379-6388, 1996

49. Riegman PH, Vlietstra RJ, van der Korput JA, et al: The promoter of the prostate-specific antigen gene contains a functional androgen responsive element. Mol Endocrinol 5:1921-1930, 1991

50. Kelloff GJ, Coffey DS, Chabner BA, et al: Prostate-specific antigen doubling time as a surrogate marker for evaluation of oncologic drugs to treat prostate cancer. Clin Cancer Res 10: 3927-3923, 2004

51. D'Amico AV, Moul JW, Carroll PR, et al: Surrogate end point for prostate cancer-specific mortality after radical prostatectomy or radiation therapy. J Natl Cancer Inst 95:1376-1383, 2003

Information downloaded from jco.ascopubs.org and provided by at BOEHRINGER on June 30, 2013 from 148.188.1.60
Copyright © 2005 American Society of Clinical Oncology. All rights reserved.

4. Buyse M, Burzykowski T, **Carroll K** et al. Progression-Free Survival Is a Surrogate for Survival in Advanced Colorectal Cancer. Journal of Clinical Oncology, 2007; 25:5218-5224.

VOLUME 25 · NUMBER 33 · NOVEMBER 20 2007

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Progression-Free Survival Is a Surrogate for Survival in Advanced Colorectal Cancer

Marc Buyse, Tomasz Burzykowski, Kevin Carroll, Stefan Michiels, Daniel J. Sargent, Langdon L. Miller, Gary L. Elfring, Jean-Pierre Pignon, and Pascal Piedbois

From the International Drug Development Institute, Louvain-la-Neuve; Center for Statistics, Hasselt University, Diepenbeek, Belgium; Oncology Therapy Area, AstraZeneca Research and Development, Macclesfield, United Kingdom; Biostatistics and Epidemiology Unit, Institut Gustave Roussy, Villejuif; Oncology Therapy Area, AstraZeneca, Rueil Malmaison, France; Division of Biostatistics, Mayo Clinic, Rochester, MN; and PTC Therapeutics, South Plainfield, NJ.

Submitted March 24, 2007; accepted July 30, 2007.

Supported by the IAP Research Network P6/03 of the Belgian Government (Belgian Science Policy; T.B.).

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Address reprint requests to Marc Buyse, ScD, IDDI, 30 avenue Provinciale, 1340 Louvain-la-Neuve, Belgium; e-mail: marc.buyse@iddi.com.

© 2007 by American Society of Clinical Oncology

0732-183X/07/2533-5218/$20.00

DOI: 10.1200/JCO.2007.11.8836

ABSTRACT

Purpose
The traditional end point for assessing efficacy of first-line chemotherapies for advanced cancer is overall survival (OS), but this end point requires prolonged follow-up and is potentially confounded by the effects of second-line therapies. We investigated whether progression-free survival (PFS) could be considered a valid surrogate for OS in advanced colorectal cancer.

Patients and Methods
Individual patient data were available from 10 historical trials comparing fluouracil (FU) + leucovorin with either FU alone (1,744 patients) or with raltitrexed (1,345 patients) and from three validation trials comparing FU + leucovorin with or without irinotecan or oxaliplatin (1,263 patients). Correlation coefficients were estimated in historical trials between the end points of PFS and OS, and between the treatment effects on these end points. Treatment effects on OS were predicted in validation trials, and compared with the observed effects.

Results
In historical trials, 1,760 patients (57%) had progressed or died at 6 months, and 1,622 (52%) had died at 12 months. The rank correlation coefficient between PFS and OS was equal to 0.82 (95% CI, 0.82 to 0.83). The correlation coefficient between treatment effects on PFS and on OS ranged from 0.99 (95% CI, 0.94 to 1.04) when all trials were considered to 0.74 (95% CI, 0.44 to 1.04) after exclusion of one highly influential trial. In the validation trials, the observed OS hazard ratios were within the 95% prediction intervals. A hazard ratio of 0.77 or lower in terms of PFS would predict a benefit in terms of OS.

Conclusion
PFS is an acceptable surrogate for OS in advanced colorectal cancer.

J Clin Oncol 25:5218-5224. © 2007 by American Society of Clinical Oncology

INTRODUCTION

Approximately 50% of patients diagnosed with colorectal carcinoma have metastatic or nonresectable disease at time of diagnosis or will develop metastases or a locoregional recurrence after their initial diagnosis. Substantial progress has been made during the last 20 years in the use of fluorouracil (FU) to treat advanced colorectal cancer, with a doubling of tumor response through modulation of FU by leucovorin or methotrexate, or through continuous intravenous infusion of FU instead of a bolus injection.[1] These therapeutic approaches were shown to yield a modest but statistically significant impact on overall survival (OS).[2] More recently, the chemotherapeutic agents irinotecan and oxaliplatin have become available after randomized trials showed they increased response rates, progression-free survival (PFS), and OS.[3-6]

Most patients with metastatic colorectal cancer still die as a result of their disease. The ultimate goal of chemotherapy is to cure the disease, or failing that, to improve patient symptoms, quality of life, and OS. It seems justified, therefore, to use OS to assess the efficacy of chemotherapies for advanced colorectal cancer. However, patient death can be observed only after prolonged follow-up, and with the increasing number of active compounds available in this disease, any effect of first-line therapies on OS may be confounded or diminished by the effects of subsequent therapies. It is therefore of interest to investigate whether PFS could replace OS as the primary end point in randomized trials for the treatment of patients with advanced colorectal cancer.

In this article, we quantify the relationship between PFS and OS in a set of historical trials, and we

Information downloaded from jco.ascopubs.org and provided by AstraZeneca - Discovery Information on November 19, 2007 from 193.132.159.169.
Copyright © 2007 by the American Society of Clinical Oncology. All rights reserved.

investigate whether the results observed in these trials could have been used to predict the effects of irinotecan and oxaliplatin in a set of more recently conducted validation trials that played a pivotal role in the development of these newer drugs.

PATIENTS AND METHODS

Trials

Individual patient data were available on 10 historical trials[2,7-9] and three validation trials[3-5] that all had a FU + leucovorin treatment group (Table 1). Historical trials consisted of all trials comparing FU + leucovorin with FU alone (seven trials, 1,744 patients), and all trials comparing FU + leucovorin with raltitrexed (three trials, 1,345 patients). They accrued patients between 1981 and 1990 (median follow-up, 30.4 months). Validation trials (1,263 patients in total) consisted of all trials comparing FU + leucovorin with the same plus irinotecan (two trials, 843 patients) or with the same plus oxaliplatin (one trial, 420 patients). They accrued patients between 1995 and 1998 (median follow-up, 22.0 months). A meta-analysis of trials comparing FU + leucovorin with FU was previously reported.[2] The other trials were those carried out for the registration of the new drugs raltitrexed,[7-9] irinotecan,[3,4] and oxaliplatin.[5]

Data

The following data were requested for individual patients in all trials: patient identifier, center identifier, randomization date, treatment assigned by randomization, tumor measurability (ie, measurable or nonmeasurable tumors), age, sex, performance status, primary tumor site (colon or rectum), site of metastases, overall response status with the first assigned treatment, date of response, date of progression with the first allocated treatment, date of death or last visit, survival status, and cause of death if applicable.

Survival Analyses

PFS and OS analyses were based on all randomly assigned patients using the intention-to-treat approach. PFS was defined as the time from random assignment to progressive disease (as assessed in each individual trial) or death from any cause. OS was defined as the time from random assignment to death from any cause. The distributions of PFS and OS were estimated using the Kaplan-Meier method. Estimation procedures and hypothesis tests were stratified for trial. The effect of treatment on PFS and on OS was quantified through hazard ratios (HRs), respectively HR_{PFS} and HR_{OS}, estimated through a proportional hazards model with treatment as the only factor.[10]

Surrogacy Criteria

A correlation approach was used to assess the validity of PFS as a surrogate for OS, which is appropriate when multiple randomized trials are available in which both end points are measured.[11] The method comprised estimation of ρ, the rank correlation coefficient between PFS and OS, and of R, the correlation coefficient between the treatment effects on PFS and on OS, expressed respectively as $\log HR_{PFS}$ and $\log HR_{OS}$. For small treatment effects, $\log HR \approx 1 - HR$; hence, log HR is an approximate estimate of the risk reduction. PFS would be claimed an acceptable surrogate end point for OS if (a) ρ were close to 1, indicating a strong correlation between PFS and OS, and (b) R were close to 1, indicating a strong correlation between treatment effects on PFS and on OS.[12]

Correlation Coefficients

The rank correlation coefficient ρ between PFS and OS was estimated through a Hougaard bivariate copula distribution of these end points over the entire time range,[13] or using the Kaplan-Meier estimates of PFS at 6 months and OS at 12 months. The correlation coefficient R between treatment effects on PFS and OS was estimated though a linear regression model, using all events over the entire time range, or up until 6 months for PFS and 12 months for OS.

Validation Strategy

Correlations between treatment effects were estimated in the historical trials, and used to predict the treatment effects on OS in the validation trials, based on the treatment effects on PFS actually observed in the validation trials.

Table 1. Clinical Trials Included in the Analyses

Trial	Treatments	No. of Patients
Historical		
FU		
Crema[2,14]	FU 370 bolus + LV 200 bolus, days 1-5, every 28 days	100
	FU 370 bolus, days 1-5, every 28 days	50
NCCTG[2]	FU 370-425 bolus + LV 20 bolus, or FU 500 bolus + LV 200 bolus, days 1-5, every 28-35 days	142
	FU 500 bolus, days 1-5, every 35 days	70
Siena[2]	FU 400 + LV 200, days 1-5, every 21 days	94
	FU 400, days 1-5, every 28 days	91
EORTC[2]	FU 2,600 24-hour infusion + LV 500 2-hour infusion, once weekly, for 6 weeks followed by a 2-week rest	165
	FU 2,600 24-hour infusion, once weekly, for 6 weeks followed by a 2-week rest	166
SWOG[2]	FU 425 bolus + LV 20 bolus, days 1-5, every 28-35 days, or FU 600 bolus + LV 500 3-hour infusion, once weekly, for 6 weeks, followed by a 2-week rest	178
	FU 500 bolus, days 1-5, every 35 days	93
SAKK[2,13]	FU bolus 400 + LV 20 bolus, days 1-5, every 28 days	152
	FU 400 bolus, days 1-5, every 28 days	158
HECOG[2]	FU 500 1-hour infusion + LV 200 2-hour infusion, once weekly, for 6 weeks followed by a 2-week rest	70
	FU 600 2-hour infusion, once weekly, for 6 weeks followed by a 2-week rest	68
Raltitrexed		
TCCSG-EU1[7]	FU 400 bolus + LV 200 bolus, days 1-5, every 28 days	248
	Raltitrexed 3 bolus, every 21 days	247
TCCSG-US[8]	FU 425 bolus + LV 20 bolus, days 1-5, every 28-35 days	210
	Raltitrexed 3 bolus, every 21 days	217
TCCSG-EU2[9]	FU 425 bolus + LV 20 bolus, days 1-5, every 28 days	216
	Raltitrexed 3 bolus, every 21 days	223
Validation		
Irinotecan		
Irinotecan-US[3]	Irinotecan 125 + FU 500 bolus + LV 20 bolus, once weekly for 4 weeks, followed by a 2-week rest	231
	FU 425 bolus + LV 20 bolus, days 1-5, every 28 days	226
Irinotecan-EU[4]	Irinotecan 80 + FU 2300 24-hour infusion + LV 500, weekly OR irinotecan 180 + FU 400 bolus and 600 22-hour infusion + LV 200 days 1-2, every other week	198
	FU 2600 24-hour infusion + LV 500, weekly, OR FU 400 bolus and 600 22-hour infusion + LV 200 days 1-2, every other week	187
Oxaliplatin		
Oxaliplatin-EU[5]	Oxaliplatin 85 + LV 200 2-hour infusion + FU 400 22-hour infusion, day 1, LV 200 2-hour infusion + FU 400 22-hour infusion, day 2, every other week	210
	LV 200 2-hour infusion+ FU 400 22-hour infusion, days 1-2, every other week	210

NOTE. All dosages are in mg/m^2. Control treatment is FU + LV.
Abbreviations: FU, fluouracil; LV, leucovorin; EU, European Union; NCCTG, North Central Cancer Treatment Group; EORTC, European Organisation for Research and Treatment of Cancer; SWOG, Southwest Oncology Group; SAKK, Swiss Group for Clinical Cancer Research; HECOG, Hellenic Cooperative Oncology Group; TCCSG, Tomudex Colorectal Cancer Study Group.

Information downloaded from jco.ascopubs.org and provided by AstraZeneca - Discovery Information on November 19, 2007 from 193.132.159.169.
Copyright © 2007 by the American Society of Clinical Oncology. All rights reserved.

The observed treatment effect on OS (and its 95% CI) was compared with the predicted effect (and its 95% prediction interval). The linear regression model between treatment effects was used to compute the surrogate threshold effect, the minimum treatment effect on PFS required to predict a nonzero treatment effect on OS in a future trial.[12]

RESULTS

Correlation Between End Points

In historical trials, similar numbers of events were observed for PFS at 6 months (1,760 events) as for OS at 12 months (1,622 events). The degree of association between Kaplan-Meier estimates of 6-month PFS and 12-month OS was weak, with a rank correlation coefficient ρ equal to only 0.32 (95% CI, −0.14 to 0.67; Fig 1). There was no evidence that the correlation between PFS and OS differed between treatments (Fig 1). In contrast, PFS and OS over the entire time range were reasonably well correlated, with a rank correlation coefficient ρ equal to 0.82 (95% CI, 0.82 to 0.83).

Treatment Effects

Figure 2 shows the PFS and OS curves by treatment group: FU + leucovorin (solid lines) versus FU or raltitrexed (dotted lines) versus irinotecan or oxaliplatin (dashed lines).

Figure 3 shows good overall agreement between the HRs for PFS and for OS in both the historical trials and the validation trials; trials tended to show large treatment benefits for both end points or small benefits for both end points. The FU + leucovorin group tended to fare better than FU alone or raltitrexed (HRs < 1) but worse than FU + leucovorin with either irinotecan or oxaliplatin.

Correlation Between Treatment Effects

The correlation coefficient between log HR_{PFS} and log HR_{OS} over the entire time range, R, was equal to 0.99 (95% CI, 0.94 to 1.04). Figure 4 shows the linear regression line used to predict treatment effects on OS from the observed treatment effects on PFS. The regression equation was log $HR_{OS} = 0.003 + 0.81 \times \log HR_{PFS}$, indicating that the risk reductions were approximately 19% ($= 1 - 0.81$) lower on OS than on PFS. The correlation coefficient between log HR_{PFS} up until 6 months and log HR_{OS} up until 12 months was equal to 0.94 (95% CI, 0.87 to 1.01).

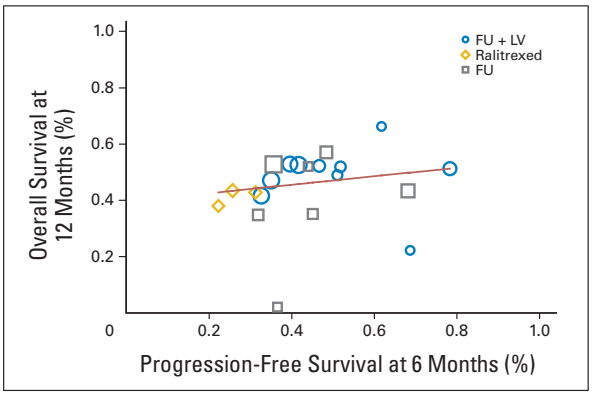


Fig 1. Correlation between 6-month progression-free survival and 12-month overall survival in different treatment groups of historical trials. Symbol size is proportional to the number of patients. FU, fluouracil; LV, leucovorin.

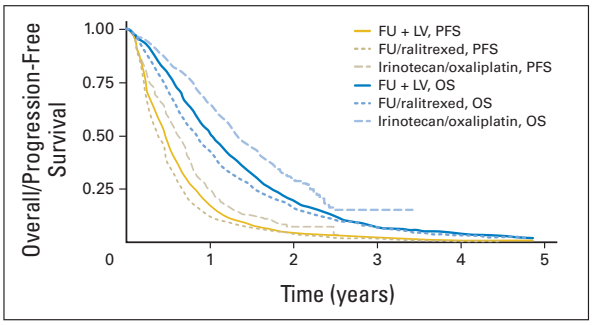


Fig 2. Progression-free survival (PFS) and overall survival curves (OS). FU, fluouracil; LV, leucovorin.

Sensitivity Analyses

The Swiss Group for Clinical Cancer Research (SAKK) trial (310 patients)[14] had a very long follow-up but surprisingly few events, whereas the Crema trial (150 patients)[15] exhibited extreme treatment benefits in terms of both PFS and OS. Exclusion of the SAKK trial, and of other trials than the Crema trial, had little impact on the results (data not shown). Exclusion of the Crema trial resulted in a much weaker association between the treatment effects, with a correlation coefficient between log HR_{PFS} and log HR_{OS} equal to 0.74 (95% CI, 0.44 to 1.04). Both of these trials were further scrutinized, but no obvious defect or methodological problem could be found to explain their atypical behavior.

Surrogate Threshold Effect

The surrogate threshold effect, as shown on Figure 4, corresponds to PFS HRs of 0.86 (or 1.16 if new treatment was worse). Thus, in order to predict a nonzero treatment effect on OS in a future trial, a hazard ratio of at most 0.86 (or at least 1.16) would need to be ascertained. After exclusion of the Crema trial, the surrogate threshold effect corresponded to hazard ratios of 0.77 (or 1.30).

Predicted Effects on OS

In validation trials, the observed treatment effect on OS was compared with the predicted treatment effect on OS, on the basis of the observed treatment effect on PFS. Table 2 compares predicted with observed treatment effects, and also shows the proportion of patients receiving second-line therapy after experiencing failure of their randomized first-line treatment: any second-line therapy, a second-line regimen containing the same new drug as their first-line therapy (irinotecan or oxaliplatin), or a second-line therapy with crossover to the other new drug (oxaliplatin or irinotecan).

DISCUSSION

Fast progress has been made in the treatment of advanced colorectal cancer during the last decade, and a number of promising new drugs are now entering clinical development for this condition. In just a few years since the approval of irinotecan and oxaliplatin, monoclonal antibodies targeting the vascular endothelial growth factor (bevacizumab) and epidermal growth factor receptor (cetuximab) have been approved as additional therapies, respectively, for the first- and second-line treatment of metastatic colorectal cancer.[16-17] It is clearly

Information downloaded from jco.ascopubs.org and provided by AstraZeneca - Discovery Information on November 19, 2007 from 193.132.159.169.
Copyright © 2007 by the American Society of Clinical Oncology. All rights reserved.

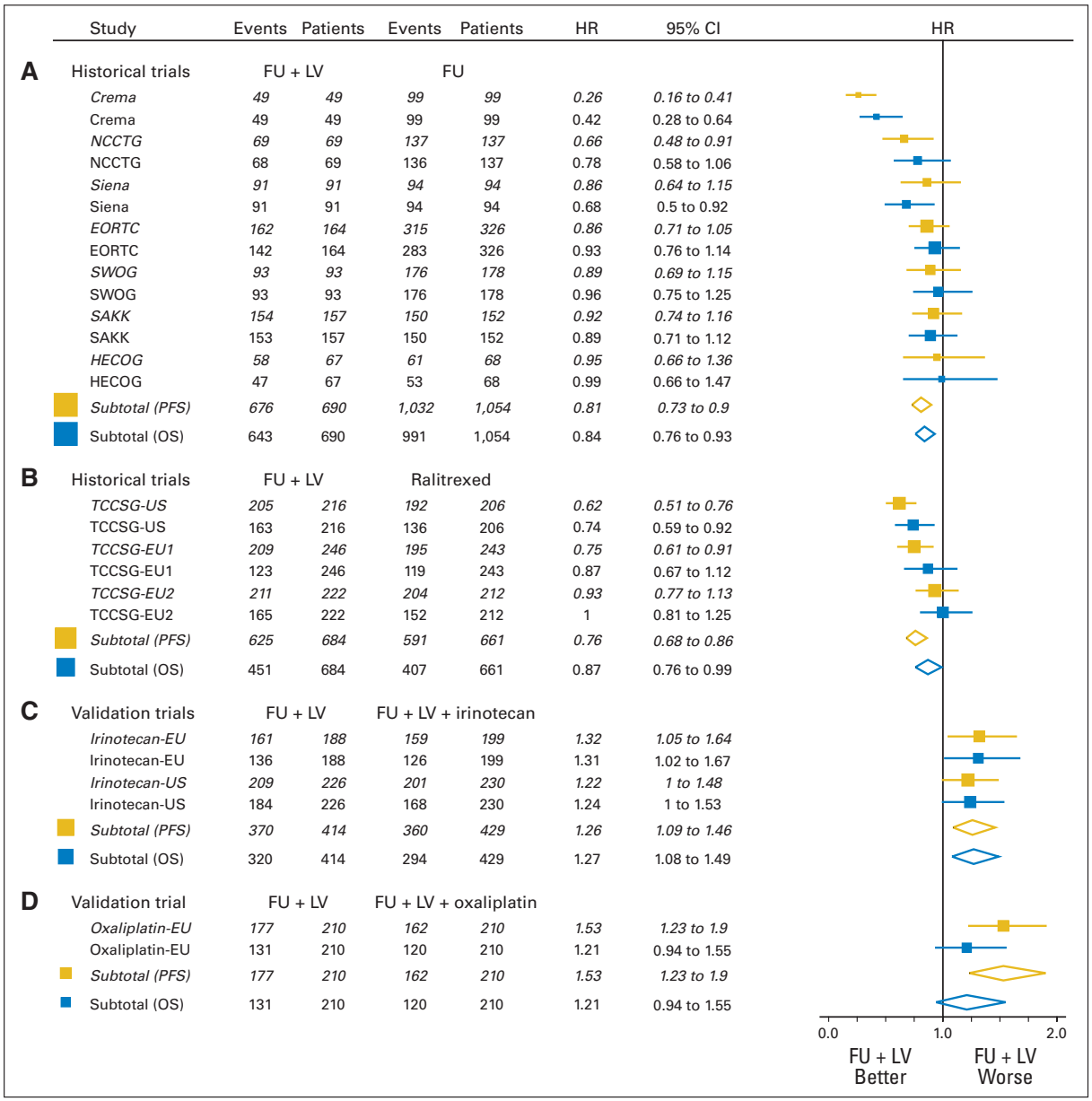


Fig 3. Forest plots of hazard ratios (ratio of hazard in FU + LV group to hazard in experimental group) for progression-free survival (PFS; yellow lines) and overall survival (OS; blue lines). Symbol size is proportional to the number of patients. NCCTG, North Central Cancer Treatment Group; EORTC, European Organisation for Research and Treatment of Cancer; SWOG, Southwest Oncology Group; SAKK, Swiss Group for Clinical Cancer Research; HECOG, Hellenic Cooperative Oncology Group; TCCSG, Tomudex Colorectal Cancer Study Group; EU, European Union.

in the best interest of future patients that new drugs be made available as soon as their efficacy is established beyond doubt on the most clinically meaningful end point, or on some earlier end point that can be considered a surrogate for it. Our analyses indicate that in advanced colorectal cancer, an analysis of PFS at 6 months would include approximately the same number of events as an analysis of OS at 12 months and would, therefore, have approximately the same statistical power to detect any given risk reduction.

Reviews of the literature suggest a tight correlation between PFS and OS in advanced colorectal cancer,[18] but this observation alone does not make PFS a good surrogate for survival.[19] Although there is no consensus regarding the theoretical conditions required for a surrogate end point to be valid, recent work suggests that surrogacy can be assessed through the correlation between the end points and the treatment effects on these end points in a series of trials.[11-12] In resectable colorectal cancer, this approach was used successfully to show that

Information downloaded from jco.ascopubs.org and provided by AstraZeneca - Discovery Information on November 19, 2007 from 193.132.159.169.
Copyright © 2007 by the American Society of Clinical Oncology. All rights reserved.

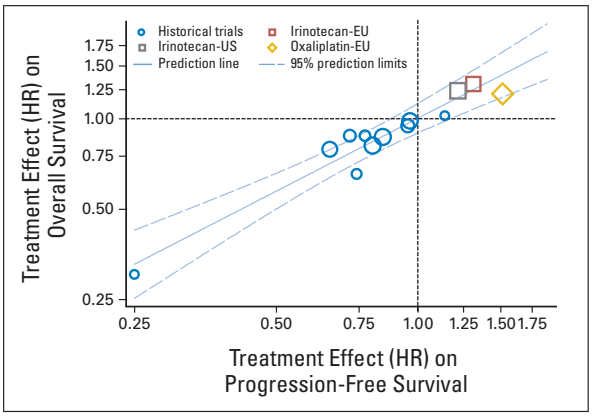


Fig 4. Correlation between treatment effects on progression-free and on overall survival in historical trials (circles), in irinotecan trials (squares), and in oxaliplatin trial (diamond). A logarithmic scale is used for both axes. Symbol size is proportional to the number of patients. HR, hazard ratio; EU, European Union.

3-year DFS was an excellent surrogate for 5-year OS.[20] In metastatic colorectal cancer, this approach was also used to investigate the relationship between tumor response and OS. Although patients who achieved a response had a significantly prolonged OS, treatment effects on response were poorly correlated with treatment effects on OS, making tumor response an unacceptable surrogate in this disease.[1,21] The analyses presented here show that, in historical trials comparing FU + leucovorin with single-agent FU or with raltitrexed, PFS was an acceptable surrogate for OS. Indeed, the two end points were well correlated ($\rho = 0.82$), and so were the effects of treatment on the two end points ($R = 0.99$), although the latter finding was heavily influenced by one trial that had a much larger treatment effect than all others. When this trial was excluded, the treatment effects were still correlated, but less impressively so ($R = 0.74$). This observation is in line with previously reported analyses of another data set, wherein the lack of treatment effect on either end point made it impossible to validate PFS as a surrogate for OS.[13] Hence for the validation to be effective, a range of treatment effects is desirable on both the surrogate and the true end points.

When our analyses were censored at 6 months for PFS and at 12 months for OS, the effects of treatment on the two end points remained highly correlated ($R = 0.94$), but the correlation between the Kaplan-Meier estimates of 6-month PFS and 12-month OS was much lower ($\rho = 0.32$; Fig 1). This finding indicates that using summary statistics for time-related end points (such as medians or Kaplan-Meier estimates at a given time point) is insensitive and potentially misleading. This approach should not be used to validate potential surrogate end points when complete individual patient data are available.[22] In practice, however, the estimate of a surrogate end point at a single time point will often be used to predict the true end point at a later time point.

In the present article, we extended the validation methodology to investigate the predictive value of PFS as a surrogate end point for OS. We calculated the surrogate threshold effect and showed that if a treatment achieved an HR for PFS of 0.86 or less, it would be expected to ultimately achieve a benefit in terms of OS (Fig 4). After exclusion of the Crema trial, the surrogate threshold effect corresponded to HRs of 0.77, suggesting the need for much larger but still achievable treatment effects on PFS. HRs in the range of 0.7 to 0.8 for PFS are realistic and have, in fact, been achieved by several treatments recently approved for the treatment of advanced colorectal cancer.[3-6,16-17] Similar conclusions were reached independently in a meta-analysis of a large number of published trials.[23] In practice, the threshold effect depends on the size of the future trial, because the estimation error of the PFS HR must be accounted for in the construction of the prediction limits for the OS HR.

The trials used in our analyses were conducted over a long period of time. The correlation between PFS and OS could largely be explained by a trend for both PFS and OS to improve over time, the former as a result of more effective first-line treatments, and the latter as a result of more effective second-line treatments. However, our claim of surrogacy is also based on a high correlation between the effects of treatment on PFS and OS, and there is no plausible way in which the correlation between treatment effects could simply be a result of time trends.

To validate results from historical trials in independent, more recent trials, we also investigated whether treatment effects on OS were reliably predicted by the treatment effects on PFS in three validation trials testing the new drugs irinotecan and oxaliplatin. In these trials, the prediction intervals of the predicted effect were narrower than the CI of the observed effect (Table 2), which underscores the potential gain arising from the use of surrogate end points for which more

Table 2. Observed Versus Predicted OS HRs

Trial	Observed OS HR	95% CI	Predicted OS HR*	95% Predicted Interval	% FU + LV + New Drug: Total Receiving Second-Line Treatment	FU + LV + New Drug: Receiving New Drug	FU + LV + New Drug: Crossed Over to New Drug	FU + LV + New Drug: Other Second-Line Treatment	% FU + LV: Total Receiving Second-Line Treatment	FU + LV: Receiving New Drug	FU + LV: Crossed Over to New Drug	FU + LV: Other Second-Line Treatment
Irinotecan-EU[4]	1.31	1.02 to 1.67	1.25	1.00 to 1.55	39	0	16	23	58	31	13	14
Irinotecan-US[3]	1.24	1.00 to 1.53	1.17	0.96 to 1.43	52	0	5†	47	70	56	5†	9
Oxaliplatin-EU[5]	1.21	0.94 to 1.55	1.40	1.12 to 1.75	58	0	30	28	61	28	20‡	23

NOTE. HRs are HR of FU + LV *v* FU + LV + new drug.
Abbreviations: OS, overall survival; HR, hazard ratio; FU, fluouracil; LV, leucovorin; EU, European Union.
*Prediction based on observed progression-free survival HR.
†Reported as < 5%.
‡10% of patients received both oxaliplatin and irinotecan in second-line treatment and are counted only once in the overall 61%.

Information downloaded from jco.ascopubs.org and provided by AstraZeneca - Discovery Information on November 19, 2007 from 193.132.159.169.
Copyright © 2007 by the American Society of Clinical Oncology. All rights reserved.

events are available than for the true end point. Such a gain would be even more pronounced for trials with less mature OS data, although those trials would also have less mature PFS data, and therefore more uncertainty in the estimated treatment effect on PFS. The predicted effects agreed extremely well with the observed effects in trials testing irinotecan, but less well in the trial testing oxaliplatin, in which the predicted effect overestimated the observed effect (Table 2). However, all of the observed effects fell within the prediction limits (Fig 4), and these differences could be a chance finding. They could also be a result of the effect of second-line treatments. In a recently reported US Intergroup trial testing oxaliplatin added to FU + leucovorin versus irinotecan added to FU + leucovorin (not included in our analyses), a much higher proportion of patients were crossed over to the other drug on disease progression in the first-line oxaliplatin arm (60%) than in the first-line irinotecan arm (24%). The observed OS benefit of first-line oxaliplatin compared with first-line irinotecan was larger (HR = 0.57) than would have been predicted from the benefit on PFS (predicted HR = 0.79), possibly because of the larger number of crossovers.[6] Second-line use of new agents is likely to produce lesser antitumor effect than first-line use,[24] but recent observations strongly suggest that use of effective second-line therapies may extend the time between first-line disease progression and death.[25,26] Thus, the ultimate OS benefits of improvements in first-line PFS may be reduced as greater numbers of effective second-line therapies are introduced. Our analyses indicate that PFS would have been an acceptable surrogate for OS in developing the drugs considered here. Similar analyses should be repeated with data from randomized trials testing newer drugs in patients receiving effective second- or even third-line therapies.

PFS offers a direct measure of new drug activity that is not obscured by subsequent therapies. Unlike response rate, PFS also has the intrinsic advantage of assessing the time of tumor control. As more active drugs enter the clinic, PFS will become an even more desirable end point than OS as the primary efficacy end point for trials in colorectal cancer.[27,28] Its use can reduce sample size, shorten accrual time, and speed time until first analysis, besides serving as an appropriate indicator of clinical benefit. Of course, increasing reliance on assessment of PFS raises the challenge of ensuring that ascertainment of tumor progression in clinical trials is reliable and unbiased.[29]

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Although all authors completed the disclosure declaration, the following author(s) indicated a financial or other interest that is relevant to the subject matter under consideration in this article. Certain relationships marked with a "U" are those for which no compensation was received; those relationships marked with a "C" were compensated. For a detailed description of the disclosure categories, or for more information about ASCO's conflict of interest policy, please refer to the Author Disclosure Declaration and the Disclosures of Potential Conflicts of Interest section in Information for Contributors.

Employment or Leadership Position: Kevin Carroll, AstraZeneca (C); Pascal Piedbois, AstraZeneca (C) **Consultant or Advisory Role:** None **Stock Ownership:** None **Honoraria:** None **Research Funding:** None **Expert Testimony:** None **Other Remuneration:** None

AUTHOR CONTRIBUTIONS

Conception and design: Marc Buyse, Tomasz Burzykowski, Pascal Piedbois
Financial support: Marc Buyse, Tomasz Burzykowski
Administrative support: Marc Buyse, Tomasz Burzykowski
Collection and assembly of data: Marc Buyse, Tomasz Burzykowski, Kevin Carroll, Stefan Michiels, Jean-Pierre Pignon, Pascal Piedbois
Data analysis and interpretation: Marc Buyse, Tomasz Burzykowski, Kevin Carroll, Stefan Michiels, Daniel J. Sargent, Langdon L. Miller, Gary L. Elfring, Jean-Pierre Pignon, Pascal Piedbois
Manuscript writing: Marc Buyse, Tomasz Burzykowski, Kevin Carroll, Stefan Michiels, Daniel J. Sargent, Langdon L. Miller, Gary L. Elfring, Jean-Pierre Pignon, Pascal Piedbois
Final approval of manuscript: Marc Buyse, Tomasz Burzykowski, Kevin Carroll, Stefan Michiels, Daniel J. Sargent, Langdon L. Miller, Gary L. Elfring, Jean-Pierre Pignon, Pascal Piedbois

REFERENCES

1. Buyse M, Thirion P, Carlson RW, et al: Tumour response to first line chemotherapy improves the survival of patients with advanced colorectal cancer. Lancet 356:373-378, 2000

2. Thirion P, Michiels S, Pignon JP, et al: Modulation of fluorouracil by leucovorin in patients with advanced colorectal cancer: An updated meta-analysis. J Clin Oncol 22:3766-3775, 2004

3. Saltz LB, Cox JV, Blanke C, et al: Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer. N Engl J Med 343:905-914, 2000

4. Douillard JY, Cunningham D, Roth AD, et al: Irinotecan combined with fluorouracil compared with fluorouracil alone as first-line treatment for metastatic colorectal cancer: A multicentre randomised trial. Lancet 355:1041-1047, 2000

5. de Gramont A, Figer A, Seymour M, et al: Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. J Clin Oncol 18:2938-2947, 2000

6. Goldberg RM, Sargent DJ, Morton RF, et al: A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. J Clin Oncol 22:23-30, 2004

7. Cunningham D, Zalcberg JR, Rath U, et al: Final results of a randomised trial comparing 'Tomudex' (raltitrexed) with 5-fluorouracil plus leucovorin in advanced colorectal cancer. Ann Oncol 7:961-965, 1996

8. Pazdur R, Vincent M: Raltitrexed (Tomudex) versus 5-fluorouracil and leucovorin (5-FU + LV) in patients with advanced colorectal cancer (ACC): Results of a randomized, multicenter, North American trial. Proc Am Soc Clin Oncol 16:228a, 1997 (abstr 801)

9. Cocconi G, Cunningham D, Van Cutsem E, et al: Open, randomized, multicenter trial of raltitrexed versus fluorouracil plus high-dose leucovorin in patients with advanced colorectal cancer. J Clin Oncol 16:2943-2952, 1998

10. Cox DR: Regression models and life tables (with discussion). JR Stat Soc B 34:187-220, 1972

11. Buyse M, Molenberghs G, Burzykowski T, et al: The validation of surrogate endpoints in meta-analyses of randomised experiments. Biostatistics 1:49-67, 2000

12. Burzykowski T, Molenberghs G, Buyse M Evaluation of Surrogate Endpoints. Heidelberg, Germany, Springer Verlag, 2005

13. Burzykowski T, Molenberghs G, Buyse M, et al: Validation of surrogate endpoints in multiple randomised clinical trials with failure-time endpoints. J Royal Stat Soc C (Applied Statist) 50:405-422, 2001

14. Borner MM, Castiglione M, Bacchi M, et al: The impact of adding low-dose leucovorin to monthly 5-fluorouracil in advanced colorectal carcinoma: Results of a phase III trial. Ann Oncol 9:535-541, 1998

15. Bobbio-Pallavicini E, Porta C, Moroni M, et al: Folinic acid does improve 5-fluorouracil activity in vivo: Results of a phase III study comparing 5-fluorouracil to 5-fluorouracil and folinic acid in advanced colon cancer patients. J Chemother 5:52-55, 1993

16. Hurwitz H, Fehrenbacher L, Novotny W, et al: Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. N Engl J Med 350:2335-2342, 2004

17. Cunningham D, Humblet Y, Siena S, et al: A randomised comparison of cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. N Engl J Med 351:337-345, 2004

18. Louvet C, de Gramont A, Tournigand C, et al: Correlation between progression free survival and response rate in patients with metastatic colorectal carcinoma. Cancer 91:2033-2038, 2001

Information downloaded from jco.ascopubs.org and provided by AstraZeneca - Discovery Information on November 19, 2007 from 193.132.159.169.
Copyright © 2007 by the American Society of Clinical Oncology. All rights reserved.

19. Fleming TR, DeMets DL: Surrogate endpoints in clinical trials: Are we being misled? Ann Intern Med 125:605-613, 1996

20. Sargent D, Wieand S, Haller DG, et al: Disease-free survival (DFS) vs overall survival (OS) as a primary endpoint for adjuvant colon cancer studies: Individual patient data from 20,898 patients on 18 randomized trials. J Clin Oncol 23:8664-8670, 2005

21. Burzykowski T, Molenberghs G, Buyse M, et al: The validation of surrogate endpoints using data from randomised clinical trials: A case-study in advanced colorectal cancer. J Royal Statist Soc A 167:103-124, 2004

22. Michiels S, Piedbois P, Burdett S, et al: Meta-analysis when only the median survival times are known: A comparison with individual patient data results. Intl J Technol Assess Health Care 21:119-125, 2005

23. Johnson KR, Ringland C, Stokes BJ, et al: Response rate or time to progression as predictors of survival in trials of metastatic colorectal cancer or non-small-cell lung cancer: A meta-analysis. Lancet Oncol 7:741-746, 2006

24. Tournigand C, Andre T, Achille E, et al: FOLFIRI followed by FOLFOX6 or the reverse sequence in advanced colorectal cancer: A randomized GERCOR study. J Clin Oncol 22:229-237, 2004

25. Grothey A, Sargent D, Goldberg RM, et al: Survival of patients with advanced colorectal cancer improves with the availability of fluorouracil-leucovorin, irinotecan, and oxaliplatin in the course of treatment. J Clin Oncol 22:1209-1214, 2004

26. Grothey A, Sargent D: Overall survival of patients with advanced colorectal cancer correlates with availability of fluorouracil, irinotecan, and oxaliplatin regardless of whether doublet or single-agent therapy is used first line. J Clin Oncol 23:9441-9442, 2005

27. Di Leo A, Bleiberg H, Buyse M: Is overall survival a realistic primary endpoint in advanced colorectal cancer? A critical assessment based on four clinical trials comparing fluorouracil plus leucovorin with the same treatment combined either with oxaliplatin or with irinotecan. Ann Oncol 14:545-549, 2004

28. Di Leo A, Bleiberg H, Buyse M: Overall survival is not a realistic endpoint for clinical trials in advanced solid tumors: A critical assessment based on recently reported phase III trials in colorectal and breast cancer. J Clin Oncol 21:2045-2047, 2003

29. US Food & Drug Administration: Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. www.fda.gov/cder/guidance/6592dft.htm

Acknowledgment

The authors are grateful to the Meta-Analysis Group in Cancer and to the principal investigators of the validation trials for permission to reanalyze individual patient data. The datasets analyzed were provided by the Biostatistics and Epidemiology Unit of Institut Gustave-Roussy, AstraZeneca, Sanofi-aventis, and Pfizer.

The complete Acknowledgment is included in the full-text version of this article, available online at www.jco.org. It is not included in the PDF version (via Adobe® Reader®).

Information downloaded from jco.ascopubs.org and provided by AstraZeneca - Discovery Information on November 19, 2007 from 193.132.159.169.
Copyright © 2007 by the American Society of Clinical Oncology. All rights reserved.

5. **Carroll KJ**. Biomarkers in drug development: friend or foe? A personal reflection gained working within oncology. Pharmaceutical Statistics, 2007; 6(4): 253–260.

PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2007; **6**: 253–260
Published online 9 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.269

WILEY InterScience®
DISCOVER SOMETHING GREAT

Biomarkers in drug development: friend or foe? A personal reflection gained working within oncology

VIEWPOINT

Kevin J. Carroll*,†
AstraZeneca Pharmaceuticals, Global Clinical Function, Alderley Park, Macclesfield, UK

Hopes and expectations for the use and utility of new, emerging biomarkers in drug development have probably never been higher, especially in oncology. Biomarkers are exalted as vital patient selection tools in an effort to target those most likely to benefit from a new drug, and so to reduce development costs, lessen risk and expedite developments times. It is further hoped that biomarkers can be used as surrogate endpoints for clinical outcomes, to demonstrate effectiveness and, ultimately, to support drug approval. However, I perceive that all is not straightforward, and, particularly in terms of the promise of accelerated drug development, biomarker strategies may not in all cases deliver the advances and advantages hoped for. Copyright © 2007 John Wiley & Sons, Ltd.

Keywords: ***biomarkers; surgery; prognostic; predictive; oncology***

1. INTRODUCTION

With the advent of ever more sophisticated proteomic, genomic and genetic technologies, the era of personalized medicine is dawning, or at least that appears to be the view coalescing across industry, academia and regulatory health authorities alike. In addition, efforts to gain a more in-depth biologic understanding of disease, particularly in oncology, is simultaneously leading to the identification of a whole host of biomarkers that reflect underlying biologic processes and the aetiology of disease. But what exactly is a 'biomarker'? For the purposes of this short article, the terms 'biomarker,' 'surrogate endpoint' and 'clinical endpoint' will be defined as per Gruttola *et al.* [1]:

Biomarker: A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.

Clinical endpoint (*or outcome*): A characteristic or variable that reflects how a patient feels or functions, or how long a patient survives.

Surrogate endpoint: A biomarker intended to substitute for a clinical endpoint.

*Correspondence to: Kevin J. Carroll, AstraZeneca Pharmaceuticals, Global Clinical Function, Alderley Park, Macclesfield, UK.
†E-mail: kevin.carroll2@astrazeneca.com

Copyright © 2007 John Wiley & Sons, Ltd.

Pharmaceutical STATISTICS

Despite clear definitions in the literature, I, like many of my fellow statistical colleagues, have not found it uncommon in practice for biomarkers and surrogate endpoints to be confused, with the terms often (and incorrectly) used interchangeably. Whereas biomarkers can perhaps help us define patient populations that may stand to benefit to a greater extent or provide some reassurance that the drug is biologically active, interfering with the underlying disease process as was hoped for, biomarkers are infrequently true surrogates for a relative improvement in clinical outcome. The distinction between a biomarker and a true surrogate endpoints is particularly important in drug development and more will be said about this later.

As evidenced by FDA's Critical Path Initiative and EMEA's Pipeline Program, regulatory authorities are looking at the drug development process, hoping to see and encourage the prospective identification of patients who will gain most benefit from new medicines while reducing ever-lengthening development times.

With an increased focus on biomarkers come high expectations. Herceptin (trastuzumab) and Gleevec (imatinib mesylate) are two often-sited biomarker-led development successes which others are encouraged to emulate. For such developments, one frequently encounters the general belief pivotal trials will be smaller in size (since effect sizes will be larger), less costly and more secure in terms carrying a lesser risk of failure. The benefit:risk ratio will be clearer and reimbursement arguments strengthened by virtue of a narrower indication and a larger treatment effect. The public is looking for the 'right drug, right patient, right time' and biomarkers seem to offer this promise.

But all is not rosy in the garden.

Biomarkers are too often and too quickly labelled as surrogate endpoints for clinical outcome without proper qualification, which can falsely raise expectations. Also, care needs to be taken to avoid the risk of missed opportunities in phase II trials designed to examine only biomarker 'positive' patients on the assumption of no possible chance of therapeutic efficacy in 'negative' patients. Further, care needs to be taken with early, apparently promising data showing separation in clinical outcomes between biomarker 'positive' and 'negative' patients treated with a new drug, since the observed difference might not in fact be due to the new drug. Over enthusiasm with such data will increase chance of failing in Phase III. And, finally, the widespread belief that biomarker-led developments will, in all cases, be cheaper and less risky is not necessarily true.

The remainder of this article is therefore structured as follows: in Section 2, the role of biomarkers as a part of a patient selection strategy is discussed including the issue of prognostic versus predictive biomarkers and issues in phase II and phase III trial design. Section 3 then briefly examines the use of biomarkers as endpoints in the drug approval process, examining what might be needed to elevate a biomarker to the status of surrogate endpoint and reflecting on the sponsors versus the regulators risk. Section 4 closes the article with a brief summary of the main points discussed.

2. BIOMARKERS TO SELECT PATIENTS MORE LIKELY TO BENEFIT FROM DRUG

A chief and increasing use of biomarkers in drug development is to select patients thought more likely to benefit from a new drug (or to deselect patients thought more likely to experience unwanted side effects of drug). This is nothing new *per se* since all approved therapies are selective to a lesser or greater extent by virtue of their indication. Rather, biomarkers offer the opportunity for a more sophisticated and refined selection of patients on the basis of disease or target biology. It is important to note that the aim is to select patients *more likely to benefit* from drug; the aim is not, as is commonly perceived, to identify 'responders' (and therefore exclude 'non-responders') to drug. Such absolute dichotomization rarely occurs in nature because, in many instances, underlying biology is a continuum. Prostate-specific antigen (PSA) level in prostate cancer or epidermal growth factor receptor (EGFR) protein

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

expression level in NSCLC or even PANSS score in depression or NIHSS score in stroke are markers and scales that measure and reflect a continuum of disease status and, consequently, there is no magic cut point that separates patients into 'responders' and 'non-responders'.

Along similar lines, and as gratefully highlighted by a reviewer, Senn points out that while selection strategies assume a patient is either responsive or unresponsive to drug, it is just as possible that all patients are responsive but to a varying and variable degree [2]. As a simple example, he considers a hypothetical 1000 patient trial in which 700 patients achieve a clinical response and the remaining 300 do not. The commonplace interpretation of such data is that the treatment works for 70% of patients 100% of time and for 30% patients 0% of time, whereas another equally valid and plausible interpretation is the drug works for 100% of patients 70% of time. He goes on to correctly point out that we can never really know which patients are more likely to benefit from a given intervention unless we have within-patient repeat assessments of response to drug (and control) from the likes of repeat period crossover trials. However, such trial designs are impossible in oncology, so researchers seemingly have little option but to think in terms of drugs working for some fraction patients 100% of the time.

Still, it is easy to see that over simplification of biomarkers in this fashion could result an increased risk of both type I and type II errors depending on the stage of development. What we are actually looking to do statistically with a biomarker is to ascertain if the apparent variability in patient response to treatment – or more precisely *variability in the treatment effect*, *drug relative to control*, – is likely to be dictated on the average by the biomarker.

2.1. Prognostic versus predictive biomarkers

Before discussing trial design issues it is important to emphasize the crucial distinction between prognostic and predictive biomarkers.

Prognostic markers tell you something about clinical outcome independent of therapeutic intervention. An example of a biomarker that was thought to be predictive but was found to be more prognostic might be epidermal growth factor receptor (EGFR) mutations in non-small cell lung cancer (NSCLC). Initially data emerged that, in Western patients receiving EGFR inhibitors, clinical outcomes were better in patients with a mutation than in patients without a mutation [3,4]. Subsequently, however, further data emerged in Western patients treated with chemotherapy which showed that clinical outcomes were better in those with a mutation relative to those without a mutation [5]. So it would seem that having the mutation is probably a good thing *per se*, irrespective of therapeutic intervention, meaning that mutations are probably prognostic for outcome in Western patients.

While prognostic biomarkers may have utility in patient enrichment strategies, predictive biomarkers are arguably more important to successful drug development. A predictive biomarker tells you that the *effect of a new drug relative to control is related to the biomarker*. Examples might be high EGFR gene copy number in NSCLC, which appears to be predictive for the effect of EGFR inhibitors relative to control, and high her-2 gene copy number in advanced breast cancer, which is predictive for the effect of trastuzumab.

If a predictive biomarker can be found early in the development process for a given drug, then this may well provide a sensible and secure direction for further development. The challenge is how do we do this? What early work is needed and how should we design phase II trials to help us understand whether a biomarker is likely to be predictive as opposed to prognostic and so provide data and information to help guide further development?

2.2. Some design issues in phase II and phase III

Where a new drug is not first in class, it is likely that the path for the use of a biomarker as a selection tool will already have been established. However, for a new drug with a novel mechanism of action, the situation is likely to be different. While pre-clinical and translational science work

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

may shed light on the biology of the disease and therapeutic target and thus, in doing so, may further suggest a biomarker that might identify patients more likely to benefit from drug, this body of work is only ever hypothesis generating – it does not prove that a given biomarker-based selection strategy will be successful. What is needed is a well-designed phase II trial(s) to examine the biomarker hypothesis and, in doing so, to guide the shape and direction of large-scale phase III trials.

Phase II designs can be complex but, broadly speaking, options range as follows:

- trials where all patients are treated with the new drug and outcomes in biomarker 'positive' patients are compared to outcomes in biomarker 'negative' patients – these are poor, relatively uninformative non-randomized designs;
- trials including only biomarker 'positive' patients randomized to drug and control – these are better designs but assume the new drug will benefit only biomarker 'positive' patients;
- trials including both biomarker 'positive' and biomarker 'negative' patients randomized to drug and control – these are the preferred designs which attempt to maximize information about the predictive value of the biomarker.

Breaking with traditional statistical thought, it may be better to design phase II trials not to look for $p<0.05$ in the comparison of drug to control, but to provide data that allows the chance of improved efficacy to be gauged both overall and in relation to the biomarker. Knowing, for example, from phase II that the chance of an improved outcome across all patients is, say, 80%, rising to 90% in those who were biomarker positive and slipping to 75% in those who are biomarker negative, a rational and informed decision could be made about how to proceed in phase III. The design of phase II would then be driven not by hypothesis testing and concerns about showing significant differences, but more by the quantity of information it was desired to generate and the fraction of patients expected to be biomarker 'positive.'

With respect to phase III programme design, there is a widespread belief that a biomarker selection strategy will result in smaller, more efficient and lower risk developments. However, this is not necessarily true in all cases. Two crucial assumptions frequently made to support this notion are (i) that the selected, biomarker 'positive' patients will experience a treatment effect while the unselected, biomarker 'negative' patients will be associated with no treatment effect and (ii) that the diagnostic that evaluates the biomarker is perfect, with 100% sensitivity and specificity. Again, as highlighted by a reviewer, several authors have previously considered similar concepts in the context of assessing the value of placebo run-in periods, where the aim is to maximize the fraction of compliant patients randomized while recognizing the classification of patients as 'compliant' and 'non-compliant' is not error free [6–8]. More recently, Maitournam and Simon have looked specifically at biomarker or genomic patient selection strategies, reaching similar conclusions [9].

To appreciate the issues a little better, consider the example situation described in Table I.

Assuming a median follow-up of 18 months, it can be shown that a trial in all patients will require approximately 1000 patients to provide 90% power for a one-sided 2.5% α level. However, if

Table I. True median survival for new and control anti-cancer drugs.

	Median survival on control (months)	Median survival on new (months)	Treatment effect HR* new:control
Biomarker positive (25%)	6	12	0.50
Biomarker negative (75%)	6	6	1.00
All patients	6	7.5	0.80

*HR = hazard ratio.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

Table II. The impact of an imperfect test.

Sensitivity (%), specificity (%)	PPV* (%)	Median survival on control (months)	Median survival on new (months)	HR new:control	Number required to enter	Number required to screen
100, 100	100	6	12	0.50	117	468
95, 75	56	6	9.4	0.64	260	613
75, 95	83	6	11	0.55	149	663
75, 75	50	6	9	0.68	317	845

*PPV = positive predictive value.

Table III. The impact of a small, non-zero effect in biomarker 'negative' patients.

	Median survival on control (months)	Median survival on new (months)	Treatment effect HR* new:control	Number of patients required (screened)
Biomarker positive (25%)	6	12	0.50	117 (468)
Biomarker negative (75%)	6	7.5	0.80*	—
All patients	6	8.7	0.69	384 (384)

*Effect size in 'negative' patients = 1/3 effect size in 'positive' patients.

the trial selects only biomarker 'positive' patients, then only 117 patients would be required to achieve the same power, but 468 patients would have to be screened to allow for only 1 in 4 being biomarker 'positive.'

Table II shows how the efficiency gain in selecting only biomarker 'positive' patients is highly dependent on the performance of the diagnostic. With 75% sensitivity and specificity, the treatment effect is diluted by the erroneous inclusion of biomarker 'negative' patients, which consequently pushes up the sample size and the number needed to screen so that, at 845 patients, we begin to approach the trial size required for the unselected approach.

In a similar fashion, Table III shows the impact of assuming a small treatment effect in biomarker 'negative' patients.

With a modest, non-zero treatment effect in biomarker 'negative' patients, the number of patients required in the unselected trial reduces from 1000 to 384 so that the unselected trial requires fewer patients than are required to be screened for the selected design.

Thus, while it is clear that a biomarker-directed development strategy can deliver efficiencies and greater security, much depends on how confident we are that the biomarker is predictive as opposed to prognostic, the size of the treatment effect in biomarker 'negative' patients and the performance of the diagnostic. Violation of any one of these assumptions can quickly erode the value of a selected approach. This tends to underscore the importance of well-designed phase II trials to provide data to investigate these crucial assumptions.

3. MOVING ON FROM PATIENT SELECTION – BIOMARKERS AS ENDPOINTS TO EVALUATE THE RELATIVE EFFECTIVENESS OF NEW DRUGS

Moving on from patient selection, another key use for biomarkers in drug development is to employ the biomarker as endpoint. In considering the issues associated with this use of a biomarker, it is helpful to clarify (i) what stage of development are we in, what question is being asked and (ii) to what extent to drug-induced effects on biomarker reflect a drug-induced effect on clinical outcome.

The first of these questions is related to whether the decision we are making and, hence, the risk

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

being taken is largely the sponsors (internal) risk or the regulators (external) risk. For example, using a biomarker to screen drug candidates early in development for likely efficacy or to choose between doses in phase II is largely the sponsor's risk. If an error is made, the use of the biomarker results in taking forward an ineffective agent or the wrong dose, the burden falls on the sponsor to rethink the development strategy. Little risk is borne by the regulator.

However, the use of a biomarkers as substitute for clinical outcome for purpose of directly supporting approval is more troublesome. Here we are using the biomarker as a surrogate endpoint and a large part of the risk falls on the regulator. The burden to demonstrate true surrogacy of an endpoint is known to be considerable. Statisticians know that simple correlation between the (assumed) surrogate and clinical outcome is a necessary, but insufficient condition to show surrogacy. Prentice [10] and Freedman [11] and, more recently, Buyse and Molenbergs [12,13] have provided a framework for assessing surrogacy, a framework that requires large, randomized controlled trials which capture both the (assumed) surrogate and clinical outcome – i.e., that require the very trials drug developers hope to avoid by use of the surrogate. PSA in prostate cancer illustrates how very difficult it can be to establish a biomarker as a surrogate endpoint. Despite more than 15 years of routine use in the management of patients with prostate cancer coupled with multiple large randomized, controlled trials looking at both PSA and clinical outcome and, more recently, formal published analyses [14] to examine surrogacy via the Buyse and Molenbergs method, PSA is still not accepted as an endpoint for drug approval. However, there is some hope – recent FDA workshops on cancer endpoints, including endpoints in prostate cancer, that may yet result in at least a composite endpoint that includes some component of PSA change as a measure of disease progression [15].

In a similar vein, longstanding clinical endpoints like progression-free survival in colorectal cancer are only just being accepted as endpoints for drug approval. This again is after years of use in the clinical management of patients and on the basis of multiple, well-controlled randomized trials that have recently been analysed to look formally at surrogacy for overall survival outcome [16].

So, set against this background, what realistic hope is there for novel biomarkers in new disease areas for their use as substitutes for clinical outcome in the evaluation of efficacy and safety and, ultimately, for drug approval?

The road for new biomarkers would therefore seem very difficult unless the level of evidence required to elevate a biomarker to surrogate endpoint status is lowered in some fashion. Many biomarkers in routine use today did not undergo rigorous evaluation for surrogacy using any kind of statistical criteria. For example, blood pressure, lipid lowering and response rate have all been used as primary endpoints to support drug approval, yet the evidence base for the surrogacy of these endpoints versus clinical outcome was not in hand. Rather, a judgement was applied at the time that a drug effect on these endpoints, whilst not being the ultimate clinical goal, was likely to reflect a (long-term) benefit to the patient.

For some of the newer biomarkers emerging today and given current concerns over the long-term safety of drugs, there may be some reluctance to make the leaps of faith made in the past. This suggests that the best we can realistically hope for, at least initially, is biomarker endpoints to *support* approval meaning that trials which examine biomarker endpoints will likely still have to be designed and powered to examine accepted clinical endpoints. However, it might be possible to argue for conditional approval on the basis of the biomarker endpoint, with a commitment to either conduct further trials to confirm clinical benefit or to continue trial follow-up beyond the point of having obtained biomarker endpoint data to collect long-term clinical outcome data and thus confirm clinical benefit.

4. SUMMARY

In my experience, founded mainly in oncology, biomarkers are increasingly seen as the route to

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

faster, cheaper and more secure drug development. They are seen by academia and industry alike as important tools in the drug development process and in helping to formulate drug development strategies. With the right biomarker in hand and well-designed early phase clinical trials, biomarker patient selection strategies can confer development advantages but there are crucial assumptions that must be highlighted and tested. Experienced statistical input to aid in the design phase II trials to evaluate these assumptions and guide phase III development is crucial. The use of biomarkers as endpoints *per se* is also important going forward. Where biomarkers are use for candidate drug screening for intrinsic activity or proof of mechanism or for any internal decision-making purpose, there seems little impediment to their use and the burden falls squarely on the sponsor to be sure the biomarker endpoint helps to make the right, not the wrong decisions. However, using biomarkers as surrogate endpoints for clinical outcome to support drug approval is more troublesome. Establishing a new biomarker as a true surrogate endpoint using published statistical criteria is extremely demanding, if not impossible. This suggests acceptance of a lower burden of evidence is required and, consequently, that greater risks must be taken, in order to use new biomarkers as substitutes for clinical outcome. At the present time and in the present climate, the prospect of this seems rather remote. In due course, however, with experience and completed trials in hand, we might be able to move to a position where confidence with novel biomarker endpoints is such that they can form the basis of drug approval though, based on past performance, the time frame may not be a quick as some would wish.

ACKNOWLEDGEMENTS

The author would like to thank the reviewer of this article for their helpful and constructive comments and for the provision of a number of valuable references, all of which helped to improve the overall content and clarity of the material presented.

REFERENCES

1. De Gruttola VG, Clax P, DeMets DL *et al*. Considerations in the evaluation of surrogate endpoints in clinical trials: summary of a National Institutes of Health Workshop. *Controlled Clinical Trials* 2001; **22**:485–502.
2. Senn SJ. Individual response to treatment: is it a valid assumption? *British Medical Journal* 2004; **329**:966–968.
3. Paez JG, Janne PA, Lee JC *et al*. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 2004; **304**:1497–1500.
4. Lynch TJ, Bell DW, Sordella R *et al*. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *New England Journal of Medicine* 2004; **350**:2129–2139.
5. Eberhard DA, Johnson BE, Amler LC *et al*. Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non-small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. *Journal of Clinical Oncology* 2005; **23**:5900–5909.
6. Blackwelder WC, Hastings BK, Lee ML, Deloria MA. Value of a run-in period in a drug trial during pregnancy. *Controlled Clinical Trials* 1990; **11**:187–198.
7. Brittain E, Wittes J. The run-in period in clinical trials. The effect of misclassification on efficiency. *Controlled Clinical Trials* 1990; **11**:327–338.
8. Schechtman KB, Gordon ME. A comprehensive algorithm for determining whether a run-in strategy will be a cost-effective design modification in a randomized clinical trial. *Statistics in Medicine* 1993; **12**:111–128.
9. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine* 2005; **24**:329–339
10. Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine* 1989; **8**:431–440.
11. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* 1992; **11**:167–178.
12. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; **54**:1014–1029.
13. Molenberghs G, Geys H, Buyse M. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine* 2001; **20**:3023–3038.
14. Collette L, Burzykowski T, Carroll KJ, Newling D, Morris T, Schröder FH. Is prostate-specific antigen

Copyright © 2007 John Wiley & Sons, Ltd.

a valid surrogate end point for survival in hormonally treated patients with metastatic prostate cancer? Joint Research of the European Organisation for Research and Treatment of Cancer, the Limburgs Universitair Centrum, and AstraZeneca Pharmaceuticals. *Journal of Clinical Oncology* 2005; **23**:6139–6148.

15. FDA Public Workshop on Clinical Trial Endpoints in Prostate Cancer, June 21–22, 2004. Available at: http://www.fda.gov/cder/drug/cancer_endpoints/default.htm#prostate (last accessed 19 January 2007).
16. Buyse M, Burzykowski T, Carroll K, Michiels S, Pignon JP, Piedbois P, Meta-Analysis Group in Cancer (MAGIC). Progression-free survival (PFS) as a surrogate for overall survival (OS) in patients with advanced colorectal cancer: an analysis of 3159 patients randomized in 11 trials. *Journal of Clinical Oncology*, 2005. ASCO Annual Meeting Proceedings. Vol 23, No. 16S, Part I of II (June 1 Supplement), 2005, 3513pp. Available at: http://www.asco.org/portal/site/ASCO/menuitem.34d60f5624ba07fd506fe310ee37a01d/?vgnextoid=76f8201eb61a7010VgnVCM100000ed730ad1RCRD&vmview=abst_detail_view&confID=34&abstractID=31926 (last accessed 22 February 2007).

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

6. Stone A, Wheeler C, **Carroll K** and Barge A. Optimizing randomized phase II trials assessing tumor progression. Contemporary Clinical Trials, 2007; 28(2):146-52.

Available online at www.sciencedirect.com
ScienceDirect
Contemporary Clinical Trials 28 (2007) 146–152
Contemporary Clinical Trials
www.elsevier.com/locate/conclintrial

Optimizing randomized phase II trials assessing tumor progression

Andrew Stone [a,*], Catherine Wheeler [b], Kevin Carroll [a], Alan Barge [a]

[a] AstraZeneca, Alderley Park, Macclesfield, Cheshire, SK10 4TG, UK
[b] AstraZeneca, Boston, MA, USA

Received 16 September 2005; accepted 8 May 2006

Abstract

The traditional development paradigm for phase II trials in oncology has been challenged in recent years by the introduction of cytostatic therapies. These agents slow the growth of tumors rather than cause high rates of shrinkage, this argues for the use of endpoints that measure growth inhibition such as progression free survival. We have previously argued the need for randomized trials in this setting. Here we discuss methodological solutions to enhance the development decision at the end of phase II in the context of progression endpoints employed in randomized trials. There are well recognized issues associated with progression endpoints relating to bias in the timing and interpretation of assessments. In this paper we present design and analysis solutions that will minimize bias by using methods that are either partially or completely time independent. We also discuss other design features to maximize the information yielded in a phase II setting. We advocate the creation of progression endpoints that utilize all available progression data rather than early fixed timepoint analyses and show that little is to be gained by assessing progression status any more frequently than would be required in routine clinical practice. Such design and analysis measures will optimize the development decision made at the end of phase II clinical evaluation.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Phase II trials; Cancer; Bias; Randomization; Sensitivity

1. Introduction

Recent advances in molecular biology are making available an increasing number of cancer therapies, these agents often demonstrate tumor growth inhibition in preclinical models, rather than cytoreduction (tumor shrinkage) the classic action of cytotoxic therapies. This suggests that a delay in tumor progression may be an appropriate clinical endpoint for assessment of efficacy in early development of these agents [1–4]. Indeed, recent data appear to confirm this hypothesis [5–7]. For example, five times as many patients would be required to replicate the observed difference in response rate compared with the difference in progression-free survival in the trial reported by Hurwitz et al. [5]. Time to progression, however, has been regarded with suspicion by investigators and regulatory authorities.

Progression-free survival differs qualitatively from overall survival in that the exact time a patient progresses (unlike death) is never actually observed: the event is recorded as having occurred in an interval between two visits. The time

* Corresponding author. Tel.: +44 1625 515969; fax: +44 1625 518537.
E-mail address: andrew.stone@astrazeneca.com (A. Stone).

1551-7144/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2006.05.003

of progression is usually assigned to the study visit at which the progression is detected. These properties mean that differential rates of assessment between treatment arms can lead to bias [8], which is of particular concern if the trial is not blinded, or if one of the treatment groups has a prevalent and distinctive side effect. To enable a treatment effect to be estimated free from bias, additional design and analysis measures need to be considered.

There is a perception that unless a time to event analysis technique is used, any analysis is hopelessly underpowered. However, a number of authors have noted that a comparison of the count of events can be nearly as powerful, which creates the option of using time independent endpoints [9–11]. As the power of statistical tests depends on the number of events observed and not the number of patients studied [12,13], any endpoint created should incorporate all available progression data. Additionally, it is often assumed that an increase in the frequency of assessment is necessary to increase the chance of revealing a difference in the rate of progression. Another concern is lead-time bias, which is particularly problematic in single-arm trials in which there is no correction for potential variability in the time course of the patient's disease at enrolment but which is unlikely to be a problem in randomized trials. Finally, the relationship of disease progression to measures of clinical benefit, such as survival or quality of life, is poorly defined in many disease settings [14].

We believe that phase II assessment of cytostatic agents should be based on randomized and controlled trials, and have previously suggested that the most natural endpoint to assess the activity of an agent is the rate of progression of disease. Here we focus on issues in trial design with time to progression endpoints in phase II and suggest methods to control for bias that will optimize decision-making at the end of phase II development.

There are a variety of terms, variably defined, which are used to capture the notion of progression, such as progression-free survival, time to progression, and time to treatment failure. Here we will use the term 'progression-free survival' and we will assume patients who die in the absence of progression will be included in the analysis as having an event at the time of death.

2. Trial design and analysis

2.1. Study endpoints — fixed early timepoint analyses?

This section will compare the sensitivity of two endpoints that measure the rate of disease progression:

- Progression-free survival (PFS), where the earlier of the 'time to death' and 'observed time to progression' is analyzed.
- Progressive disease (PD) rate, where we analyze the proportion of patients who have progressed before a fixed timepoint.

Eisenhauer [15] has proposed using the PD rate, specifically at 8 weeks, as a rapid measure of assessing the effectiveness of a new therapy. This is based on an observed association between the rate of progression at 8 weeks and survival in a series of trials conducted at the National Cancer Institute of Canada (NCIC). In these trials, the survival of patients whose best response at 8 weeks was partial response or stable disease was similar to each other, but in marked contrast to those whose best response was PD. This led the authors to consider whether the rate of PD could be used in conjunction with the response rate to determine whether a new agent was active. Can a PD rate at 8 weeks therefore be considered the endpoint of choice in a randomized setting, particularly as patients would only need to be followed for 8 weeks? This would be a very attractive approach to early drug development.

When evaluating such an approach it should be recognized that the power of any statistical analysis depends on the number of events observed and not the number of patients recruited [12,13]. Therefore a PD rate endpoint is statistically inefficient as it does not incorporate data from patients who progress after 8 weeks. With staggered recruitment, there will always be many observed progression events excluded from the analysis if a PD rate endpoint is used.

In general and assuming proportional hazards, the correspondence between the PD rate and hazard ratio (HR) is described using the following relationship:

$$\mathrm{HR} = \log(1-\mathrm{PD}_0(t))/\log(1-\mathrm{PD}_1(t))$$

$\mathrm{PD}_0(t)$, proportion of patients progressing at time t for the new agent; $\mathrm{PD}_1(t)$, proportion of patients progressing at time t for the control; HR = hazard ratio (new:control) [13].

Using this relationship, it is easy to quantify the number of patients required to detect a common treatment effect dependent on endpoint using standard sample size calculations for time to event and proportions [12,16]. Examples are displayed in Table 1, where each row is an identical test of treatment effect measured by a common HR of 0.67. It is assumed that a convenient timepoint is chosen where the expected control PD rate is 10, 20, 30 or 50%. Therefore, the results are independent of the rate of progression and hence the timepoint chosen for PD rate comparison. Data in Table 1 show that PFS is a far more sensitive endpoint to treatment effects than an early PD rate. With the assumptions described, a trial sized on a 50% difference in medians would require 100 patients when using PFS, and 11 months would elapse before the required number of events had occurred. In contrast, if treatment arms were compared using the PD rate at the timepoint where it was anticipated 20% of the control group would have progressed, over 400 patients would need to be recruited to detect the same treatment effect. Furthermore, given the same rate of recruitment, the trial would take 11 months longer.

Any perceived benefit in speed using a PD rate endpoint (or early fixed timepoint analysis), is negated by the need to recruit larger numbers of patients to maintain sensitivity. The greatest source of the extra sensitivity with a PFS endpoint is the inclusion of all available progression events and not the use of the time of the event [9]. We therefore recommend the use of a PFS endpoint that utilizes all available data. Examples of such endpoints are presented in the next section.

2.2. Differential follow-up

As patient progression is assessed at study visits only, the exact timing is never documented and an estimate of the size of the true therapeutic benefit can be subject to bias. Here we present alternative analyses, event counts and grouped survival methods, which can minimize or remove time assessment bias.

Williams et al. [8] have demonstrated that if there is an asymmetric visit schedules between treatment arms, bias in the underlying assessment of time to progression can occur if data are analyzed using the log–rank test or Cox proportional hazard model. Even if by design, visit schedules are identical between treatment arms, there is still a concern regarding bias if there is a tendency for investigators to look earlier, or more frequently, in one treatment group. This may be of particular concern if the trial is not blinded, or if one of the treatment groups has a prevalent and distinctive side effect. For example, in the absence of an effect on progression, if patients in one treatment arm present more frequently on the basis of symptoms there is the opportunity to assess the progression status in that treatment arm sooner and hence increase the chance of falsely revealing a treatment effect.

Instead of using a log–rank test, an alternative approach would be to ignore time altogether and compare a count of the number of progression events (including deaths in the absence of progression). Unlike in a fixed timepoint analysis, events would be counted regardless of how long after randomization they occurred. The only stipulation required for the timing of assessments would be that all non-progressing should be assessed within a time-window at the end of follow-up. Results can be approximated to HRs by using a binary analysis with a complementary log–log link [17]. As long as the number of non-progressing patients completing the end of follow-up assessment was comparable between arms, this method would minimize any potential for time assessment bias.

A number of authors have noted that a comparison of event counts results in a relatively small reduction in power compared with a traditional time to event analysis [9–11]. In cases where only a small proportion of patients have

Table 1
Number of patients and duration of trial to detect a hazard ratio of 0.67 with 80% power and a one-sided significance level of 20%

Endpoint[a]	No. of patients[b]	Duration[c] (months)
PFS	100[d]	11
PD rate 10% vs. 6.8%	852	43
PD rate 20% vs. 13.8%	414	22
PD rate 30% vs. 21.2%	278	16
PD rate 50% vs. 37.0%	164	12

[a] For PD rate — comparison that results in a hazard ratio of 0.67.
[b] Total number of patients for a two-arm trial.
[c] Assumes patients are recruited uniformly at 20/month; events follow an exponential distribution with medians of 4 and 6 months, respectively.
[d] Sixty-nine events are required, a greater number of patients would reduce the duration for this endpoint.

progressed, such as in an early disease setting, event count and time to event analyses are virtually indistinguishable in terms of power. In situations normally encountered in an advanced setting, as long as the data are analyzed when the proportion of patients with an event is not too large, for example 75% or less, sample size is increased by less than 25% when using an event count [10].

Another alternative would be to use a grouped survival method, which categorizes the events according to the period in which the events are known to occur [17,18]. For example, if patients are scheduled to be assessed for disease progression every 8 weeks, time periods of 0–8 weeks, 8–16 weeks, 16–24 weeks, etc. are created. Progression events are then assigned to the period in which they occurred and for non-progressing patients, data are censored in the period in which the latest scan was performed. The proportion of patients progressing within each period is then compared between each treatment arm and the results combined across all of the periods and expressed as a HR. For example, if a patient presented with symptoms at 12 weeks and was found to have progressed, their event would be assigned to the same period (8–16 weeks) as an event observed at 16 weeks from a patient who had adhered to the protocol. This is in contrast to using a classical time to event assessment, which would differentiate between these patients (the times to event being 12 and 16 weeks, respectively) when in fact we cannot determine who progressed first.

Differential schedule frequencies, and hence visit frequencies, would not present a problem to either analysis method. For example, if a 2-weekly schedule is compared with a 3-weekly schedule, to prevent any bias, progression assessments could be made every 6 weeks.

Given concerns regarding time assessment bias, alternative design and analysis strategies should be considered when evaluating, in randomized phase II trials, whether new agents alter the rate of progression of disease. The loss in power or increase in sample size may be considered small in comparison to the extra reassurance given to the end of phase II development decision.

2.3. Assessment frequency

The belief that it is necessary to scan patients as frequently as practically possible to optimize the comparison of the rate of progression is widely held. We have addressed this question by simulation and considering both a Cox proportional hazards model and the grouped survival method proposed earlier (Table 2).

For the Cox proportional hazards model, there is only a marginal loss of power (<3%) when assessments are made at a frequency that is half the control median. Less frequent assessments are, however, associated with a greater loss of power. Similar results have been described by others, the extent of the effect on power being affected by the proportion of patients censored [19]. In contrast, for the grouped survival method there is no appreciable deterioration of power as the frequency of assessments and hence the number of periods reduces. However, given the results in the previous section, if data were analyzed after a high proportion (>75%) of patients had progressed then we would also expect the power of this method to reduce with frequency.

Also shown in Table 2 is the time taken for the required number of events to be observed. Not surprisingly, when patients are being assessed less frequently a greater time elapses before the progression information becomes available. Only when there is very infrequent assessment, approaching the median event times, would these adversely effect the duration of the trial.

A number of conclusions can be drawn concerning an optimum assessment frequency. First, there are other considerations apart from statistical efficiency. When patients are being treated outside of a clinical trial, they would not remain on treatment indefinitely without monitoring whether their disease had progressed. Therefore, there will be a minimum frequency at which it would be acceptable to assess disease status. However, these results do show that assessing patients in the context of a clinical trial more frequently, adds only burden with little gain in sensitivity.

2.4. Informed censoring

Censored data arises when the event of interest has not been observed (i.e. progression or death) at the time of analysis. Censored data should be considered as missing data and therefore we should always consider carefully the potential for resultant bias and take steps to minimize this potential; just because censoring is permitted by analysis techniques it does not mean reliable analyses necessarily follow. For example, observation may stop because the patient is considered to be progressing symptomatically and their tumor dimensions are getting close, but not quite achieving protocol-defined criteria. Survival analysis methods assume that such a patient is no closer to protocol-defined

Table 2
Simulated power for a trial designed to have 80% power to detect a hazard ratio of 0.67 at a one-sided significance level of 20%, for various scanning frequencies

Visit frequency	Cox proportional hazards			Grouped analysis		
	Power[a]	Hazard ratio[b]	Follow-up[c] (weeks)	Power[a]	Hazard ratio[b]	Follow-up[c] (weeks)
a) Comparison of treatments with medians of 4 and 6 months						
Constant	79.8%	0.67	50	NA	NA	NA
2 weeks	79.6%	0.67	51	79.9%	0.66	51
1 month	79.0%	0.67	52	79.8%	0.66	52
2 months	78.1%	0.67	54	80.1%	0.66	54
4 months	74.5%	0.69	59	79.4%	0.66	59
b) Comparison of treatments with medians of 8 and 12 months						
Constant	79.6%	0.67	86	NA	NA	NA
2 weeks	79.6%	0.67	87	80.0%	0.66	86
1 month	79.2%	0.67	88	80.0%	0.66	87
2 months	78.8%	0.67	91	80.0%	0.66	90
4 months	77.5%	0.68	95	79.7%	0.66	94
6 months	76.0%	0.69	100	79.6%	0.66	98
8 months	73.4%	0.70	104	79.5%	0.66	103

Each row is the result of 5000 simulations, each with 50 observations per treatment arm and waiting for 69 events to occur, assuming an exponential distribution with the stated medians.
NA, not applicable.
[a] Proportion of simulations with a one-sided p-value <0.2 in favor of the more effective therapy.
[b] Calculated as the geometric mean of the hazard ratios estimated from each dataset.
[c] Calculated as the time from start of recruitment to the 69th event observed, patients recruited over 26 weeks.

progression than patients who have not progressed. This assumption would clearly not hold and is termed informative censoring. In this example, if censoring occurs more frequently in one arm than another then bias will result as the number of events will be differentially undercounted in one treatment arm.

If these issues are problematic what alternatives do we have? We could create a treatment failure endpoint where withdrawal of therapy in the absence of progression is also counted as an event. This would remove the need to censor but would be at the expense of creating a more subjective endpoint that is further removed from eventual approval endpoints used in phase III trials. A second alternative would be to impute a date of progression on the basis of the serial tumor measurements already observed. This analysis would, however, require a whole new set of assumptions to be made instead. The best solution is to continue to follow up as many patients as possible until they achieve protocol-defined criteria for progression [20]. This would mean that regardless of whether patients have withdrawn from randomized therapy or taken other therapy prior to progression, they should continue to be assessed until they meet the protocolled definition of progression. In cases where patients take another cancer therapy prior to progression, there is the possibility that this therapy could influence the outcome, particularly if it is administered more frequently in one treatment arm and is efficacious. However, only by collecting data up to the point of progression can we hope to tease out the potential biases at play. This approach is consistent with an intention-to-treat philosophy that attempts to address the effect of a new therapy when it is introduced into clinical practice. The need to obtain complete follow-up should be stressed to investigators at trial initiation, as only with the actual dates of progression can we hope to assess what impact intervening events have had on the outcome of the trial.

3. Discussion

We previously have argued, with others, that the phase II assessment of cytostatic agents should be based on randomized and controlled trials [2–4]. Here we propose that the most natural endpoint to assess the activity of these agents is the rate of disease progression. We have compared the statistical efficiency of this endpoint to an early fixed timepoint analysis. Despite the apparent attractiveness of an early timepoint analysis, we found that a classic PFS endpoint will always be more sensitive and often substantially so. The intuitive explanation is that as the number of progression events drives the power of any analysis, an early timepoint analysis will always result in relevant information being excluded.

Randomization alone is not sufficient to obtain an estimate of the treatment effect free from bias. Other authors have highlighted deficiencies and potential pitfalls in the use of PFS as an endpoint due to the imprecise nature of documentation and the potential for differential follow-up [8,14]. If there is potential to scan one treatment arm more frequently than another due to the differential presence of signs and symptoms, we have recommended alternative analyses that either group or count events. The event count analysis should be distinguished from a fixed timepoint analysis as all events are included regardless of how long an individual patient has been in the trial. Which of these alternatives should be preferred? Compared to an event count analysis, the grouped survival method has the advantage of fully maintaining power and describing the treatment effect over the whole period of follow-up and not just at the end of follow-up. Its drawback is the need for investigators still to follow a protocolled schedule. If this schedule is not followed, its application will be complicated by decisions regarding the handling of missing assessments and bias, although reduced, may still persist. The advantage of an event count lies in its simplicity, as the only constraint is that non-progressing patients are assessed at the end of follow-up. Careful planning would be required to prevent the possibility that either the treatment arms were compared when nearly all patients had progressed or when so few patients had progressed that the trial was greatly underpowered.

We recognize neither the grouped survival method or event count represent the standard approach to the analysis of time-to-event data. However, grouped methods have been available for some time [17] and the event count has been the subject of recent debate. The use of PFS and the consequences for design and analysis have been the subject of regulatory debate at recent open discussions initiated by the U.S. Food and Drug Administration (FDA) [21] and in a Draft Guidance for Industry on Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics [22]. The event count has been presented as an alternative by Carroll [23] who demonstrated the high concordance in results obtained using an event count analysis and a log–rank test in a large randomized controlled trial. Further, such an approach, in this case referred to as a single timepoint assessment, is discussed as a possible future method in the recent draft FDA guidance. This approach is also being prospectively applied in the design of recently initiated phase II trials sponsored by AstraZeneca. A more detailed assessment of the event count analysis will be presented in a separate paper. We believe, either of the analyses presented has the potential to minimize, and possibly remove, any bias and should give extra reassurance to the end of phase II development decision.

We have also presented results showing the minimal gain in statistical sensitivity by scanning patients more frequently, and therefore recommend that protocols do not place an undue burden in terms of assessment frequency, especially as this increases the likelihood of poor protocol compliance and the resultant problems in analysis interpretation. Finally, we re-affirmed as with any phase III survival trial, one should continue to assess all patients until they have the event of interest regardless of intervening events.

There are also more radical alternatives to the analysis of data that we have not discussed, which might be valuable in a phase II setting. In our analysis we reduce tumor volume data, recorded on a continuous scale and on multiple occasions, into a binary assessment of whether or not the patient progressed. This can be regarded as statistically wasteful as there exist more powerful statistical techniques that could utilize the continuous nature of the data recorded and assess the relative rates of the increase in tumor burden. This should be the subject of further research as it has the potential to enable the same questions to be answered with fewer patients.

Another advantage of a randomized phase II trial is that an initial assessment can be made of the impact of the new therapy on survival. This would require prolonged follow-up of the trial after the progression data are available but may allow the trial to act as a supportive trial to any subsequent pivotal trial and may prove helpful in the end of phase II development decision if the progression data are equivocal.

4. Recommendations

The following recommendations are proposed:

- A PFS endpoint that incorporates all available progression data is preferred to an early fixed timepoint analysis.
- In order to minimize any potential bias in the estimate of the treatment effect, trials should be double-blind where possible and the primary analysis technique should either be a grouped survival method or a comparison of the count of progression events.
- All patients should be assessed until progression, regardless of intervening events.
- Protocols should not impose an undue assessment frequency in the false belief that this will improve sensitivity.

References

[1] Korn EL, Arbuck SG, Pluda JM, et al. Clinical trial designs for cytostatic agents: are new approaches needed? J Clin Oncol 2001;19:265–72.
[2] Stadler WM, Ratain MJ. Development of target-based antineoplastic agents. Invest New Drugs 2000;18:7–16.
[3] Simon RM, Steinberg SM, Hamilton M, et al. Clinical trial designs for the early clinical development of therapeutic cancer vaccines. J Clin Oncol 2001;19:1848–54.
[4] Eskens FA, Verweij J. Clinical studies in the development of new anticancer agents exhibiting growth inhibition in models: facing the challenge of a proper study design. Crit Rev Oncol Hematol 2000;34:83–8.
[5] Hurwitz H, Fehrenbacher L, Novotny W, et al. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. N Engl J Med 2004;350:2335–42.
[6] Yang JC, Haworth L, Sherry RM, et al. A randomized trial of bevacizumab, an anti-vascular endothelial growth factor antibody, for metastatic renal cancer. N Engl J Med 2003;349:427–34.
[7] Escudier B, Szczylik C, Eisen T, et al. Randomized Phase III trial of the Raf kinase and VEGFR inhibitor sorafenib (BAY 43-9006) in patients with advanced renal cell carcinoma (RCC). Proc Am Soc Clin Oncol 2005 (abst 4510).
[8] Williams G, He K, Chen G, Chi G, Pazdur R. Operational bias in assessing time to progression (TTP). Proc Am Assoc Cancer Res 2002 (abst 975).
[9] Cuzick J. The efficiency of the proportions test and the logrank test for censored survival data. Biometrics 1982;38:1033–9.
[10] Gail MH. Applicability of sample size calculations based on a comparison of proportions for use with the logrank test. Control Clin Trials 1985;6:112–9.
[11] Gart JJ, Tarone RE. On the efficiency of age-adjusted tests in animal carcinogenicity experiments. Biometrics 1987;43:235–44.
[12] Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics 1983;39:499–503.
[13] Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev 2002;24:39–53.
[14] Hirschfeld S, Pazdur R. Oncology drug development: United States Food and Drug Administration perspective. Crit Rev Oncol Hematol 2002;42:137–43.
[15] Eisenhauer EA. Phase I and II trials of novel anti-cancer agents: endpoints, efficacy and existentialism. The Michel Clavel lecture, held at the 10th NCI-EORTC Conference on new drugs in cancer therapy, Amsterdam, 16–19 June 1998. Ann Oncol, vol. 9. 1998. p. 1047–52.
[16] Fleiss JL, Tytun A, Ury SHK. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 1980;36:343–6.
[17] Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. Biometrics 1978;34:57–67.
[18] Whitehead J. The analysis of relapse clinical trials, with application to a comparison of two ulcer treatments. Stat Med 1989;8:1439–54.
[19] Wallenstein S, Wittes J. The power of the Mantel–Haenszel test for grouped failure time data. Biometrics 1993;49:1077–87.
[20] Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 1976;34:585–612.
[21] Colorectal Cancer Endpoints Workshop Summary. http://www.fda.gov/cder/drug/cancer_endpoints/colonEndpointsSummary.htm.
[22] Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics DRAFT GUIDANCE, 2005. http://www.fda.gov/cder/guidance/6592dft.htm.
[23] Carroll K. Progression as an endpoint in CRC http://www.fda.gov/ohrms/dockets/ac/04/slides/4037OPH2_01_Carrol_files/frame.htm.

7. **Carroll KJ**. Analysis of progression-free survival in oncology trials: some common statistical issues. Pharmaceutical Statistics, 2007; Vol 6(2): 99-113.

PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2007; **6**: 99–113
Published online 22 January 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.251

Analysis of progression-free survival in oncology trials: Some common statistical issues

MAIN PAPER

Kevin J. Carroll*,†
AstraZeneca Pharmaceuticals, Global Clinical Information Science, Alderley Park, Macclesfield, UK

With the advent of ever more effective second and third line cancer treatments and the growing use of 'crossover' trial designs in oncology, in which patients switch to the alternate randomized treatment upon disease progression, progression-free survival (PFS) is an increasingly important endpoint in oncologic drug development. However, several concerns exist regarding the use of PFS as a basis to compare treatments. Unlike survival, the exact time of progression is unknown, so progression times might be over-estimated and, consequently, bias may be introduced when comparing treatments. Further, it is not uncommon for randomized therapy to be stopped prior to progression being documented due to toxicity or the initiation of additional anti-cancer therapy; in such cases patients are frequently not followed further for progression and, consequently, are right-censored in the analysis. This article reviews these issues and concludes that concerns relating to the exact timing of progression are generally overstated, with analysis techniques and simple alternative endpoints available to either remove bias entirely or at least provide reassurance via supportive analyses that bias is not present. Further, it is concluded that the regularly recommended manoeuvre to censor PFS time at dropout due to toxicity or upon the initiation of additional anti-cancer therapy is likely to favour the more toxic, less efficacious treatment and so should be avoided whenever possible.
Copyright © 2007 John Wiley & Sons, Ltd.

Keywords: ***oncology; event count analysis; progression-free survival informative censoring; interval censoring***

*Correspondence to: Kevin J. Carroll, AstraZeneca Pharmaceuticals, Global Clinical Information Science, Alderley Park, Macclesfield, UK.
†E-mail: kevin.carroll2@astrazeneca.com

INTRODUCTION

Progression-free survival (PFS) is an important endpoint in oncologic drug development. With the

Copyright © 2007 John Wiley & Sons, Ltd.

advent of a new generation of biologically targeted cytostatic anti-cancer agents, drug developers and researchers can no longer rely on uncontrolled phase II trials and response rate to screen new medicines for clinical utility [1–3]. For drugs designed to stabilize disease, the most sensible phase II approach has been argued to be randomized trials with PFS as the primary endpoint [4,5]. Further, with the advent of ever more effective second and third line cancer treatments and the growing use of 'crossover' designs, in which patients switch to the alternate randomized treatment upon disease progression, detecting an improvement in survival in confirmatory phase III trials has been recognized as an increasingly difficult goal [6–8]. The recently issued EMEA anti-cancer guideline acknowledges these issues and states that either survival or PFS can be used as a primary endpoint in pivotal trials seeking approval for a new drug; when PFS is used and justified as the primary endpoint, survival should be a stated secondary endpoint with follow-up sufficient 'to ensure that there are no relevant negative effects on this [survival] endpoint' [9]. In light of the issues, there has recently been a number of open discussions initiated by the US Food and Drug Administration (FDA) to examine the utility of progression and other measures as endpoints for oncologic drug approval [6–8,10]. In particular, at the Oncologic Drugs Advisory Committee (ODAC) discussion in December 2003 on approval endpoints in non-small cell lung cancer, the vote was 18 'yes', 0 'no' and 1 abstention for the use of PFS as an endpoint to support accelerated approval in the advanced disease setting [6]. Similarly, at the ODAC discussion in May 2004 on approval endpoints in colorectal cancer, the vote was 8 'yes', 5 'no' for the use of PFS as an endpoint to support full approval in the advanced setting [8]. More recently, PFS has been used as the sole basis to provide full drug approval for sorafenib in the treatment of advanced renal cell carcinoma and panitumumab in the treatment of advanced colorectal cancer [11,12].

However, despite such support for the use of PFS as a primary endpoint to support drug approval, key concerns remain regarding the use of PFS to compare treatments for relative effectiveness:

(i) Unlike survival, the exact timing of progression is unknown. Discrete clinic visit schedules for disease assessment means that progressions that occur in between visits are commonly assigned to the visit at which progression was detected, leading to over-estimation of the time to progression [13]. Consequently bias may be introduced in the comparison of treatments as was suggested in the FDA's review of oblimersen sodium [14]. This concern has led to a consensus emerging that clinic visits need to be frequent and identically scheduled between treatments to ensure an accurate determination of the time of progression and a fair comparison of treatments.

(ii) It is not uncommon for randomized therapy to stop (say, due to toxicity) or for additional anti-cancer therapies to be initiated prior to progression being documented. Handling of such patients in the analysis is problematic; recent FDA draft guidelines have suggested that progression time should be censored at the time of the intermediate event [15]. However, this view does not appear to be entirely shared by EU regulatory authorites based on the recently issued appendix to the CHMP's anti-cancer guideline [9,16].

This paper discusses these issues, their practical implications and importance when comparing treatments, and explores if there are ways in which they might be addressed or ameliorated. The remainder of the paper is therefore structured as follows: Section 2 describes a typical oncology trial design with PFS as the primary endpoint. Section 3 examines the practice of assigning the time of progression to the visit at which it was detected and Section 4 examines censoring PFS time on drop-out due to toxicity or the initiation of additional anti-cancer therapy. Section 5 then closes the paper with recommendations for trial design and analysis and a brief discussion of some other, key issues.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

TYPICAL ONCOLOGY TRIAL DESIGN

Suppose two treatments, experimental (E) and control (C), are to be compared in terms of PFS time in a clinical trial powered to detect an underlying hazard ratio, E:C, of size θ, with a 1-sided Type I error rate of α and power $1-\beta$ so that a total of d events are required [17]. Assuming event rates of λ_E and λ_C, uniform accrual over a period of R months and a minimum follow-up period of F months (giving a maximum follow-up of $R+F$ months, which is hereafter referred to as the 'trial follow-up period'), a total of $2N$ patients are to be randomized on a 1:1 basis [18–20]. Assume also that disease status is assessed at regular, scheduled clinic visits, every V months, say. For simplicity, further assume that V is chosen such that $\frac{R}{V}$ and $\frac{F}{V}$ are both integer so that $\frac{F}{V}$ is the minimum and $\frac{R+F}{V}$ the maximum number of scheduled assessments per patient and a clinic visit always takes place at the end of the trial follow-up period.

In advanced disease, PFS time is defined as the interval from randomization to the first of either disease progression or death from any cause. The equivalent measure in adjuvant settings is disease-free survival (DFS), being the interval from randomization to the first of either disease recurrence or death from any cause. The discussion that follows is framed in terms of PFS but can equally be applied to DFS. Since disease is normally assessed at regular, scheduled clinic visits, the exact time of progression is typically unknown. The time of progression is therefore usually assigned to the date of the clinic visit at which it is detected. Patients who are lost to follow-up prior to progression or who reach the end of the trial follow-up period without progression are right censored in the analysis. Patients may also stop randomized therapy during the trial follow-up period prior to reaching a confirmed progression event due toxicity or the addition of further anti-cancer therapy. Such patients are commonly not followed further for progression status, being censored at the time of the event associated with the cessation of randomized therapy.

THE IMPACT OF NOT KNOWING THE EXACT TIMING OF PROGRESSION

When the time of progression is assigned to the visit at which progression was first detected, the extent to which bias is introduced can be gauged directly for exponentially distributed lifetimes using maximum likelihood methods (see Appendix A). If T_i denotes the observed PFS time, event or censored, for the ith patient then

$$\overline{T} = \frac{\sum_{i=1}^{N} T_i}{d} = \frac{1}{\text{observed event rate}}$$

is approximately

$$N\left(\frac{V}{1-\mathrm{e}^{-\lambda V}}, \frac{V^2\mathrm{e}^{-\lambda V}}{d(1-\mathrm{e}^{-\lambda V})^2}\right)$$

or, more conveniently,

$$\ln\left(\overline{T}\right) \sim N\left(\ln\left(\frac{V}{1-\mathrm{e}^{-\lambda V}}\right), \frac{\mathrm{e}^{-\lambda V}}{d}\right)$$

If comparing E to C, then the observed hazard ratio

$$\hat{\theta} = \frac{\text{observed event rate}_E}{\text{observed event rate}_C} = \frac{\overline{T}_C}{\overline{T}_E} \tag{1}$$

is a biased estimate since

$$E[\hat{\theta}] = \frac{V_C\left(1-\mathrm{e}^{-\lambda_E V_E}\right)}{V_E\left(1-\mathrm{e}^{-\lambda_C V_C}\right)} \neq \frac{\lambda_E}{\lambda_C} \tag{2}$$

Note that bias is introduced even if visits are scheduled symmetrically between treatments ($V_E = V_C$) and, as one would expect, the degree of bias depends upon the ratio of the interval between visits and the expected PFS time (that is on $\lambda_E V_E$ and $\lambda_C V_C$). The bias in the hazard ratio erodes the power of the standard log rank test to

$$\phi^{-1}\left[\left(z_\alpha + z_\beta\right)\omega - z_\alpha\right] \tag{3}$$

where $\omega = \mathrm{abs}\left(\frac{\ln(\hat{\theta})}{\ln(\theta)}\right)$ and $\phi^{-1}(.)$ is the inverse of the cumulative normal distribution; to restore power, the target number of events would have to increase to

$$d' = \frac{d}{\omega^2} \tag{4}$$

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

(see Appendix B). Table I illustrates the degree of bias that can be introduced by assigning progression time to the clinic visit at which it was detected and the consequences on power.

Table I shows that as the interval between visits lengthens and the number of clinic visits declines, then the hazard ratio is increasingly biased toward the null. Consequently, power falls and, in the examples given, the number of events required to maintain 90% power increases by between 7% and 16% even when visits are as frequent as every month.

Assuming a common visit schedule between treatments, it is of interest to note that to ensure retention of at least $100(1-\gamma)\%$ power $\gamma > \beta$, visits must be scheduled approximately every V' months where

$$V' = (\text{median PFS on } C) \times \frac{2}{\ln(2)} \frac{1}{\theta} \left(\frac{\theta^k - \theta}{1 - \theta^k} \right) \quad (5)$$

and $k = \text{abs}\left(\frac{z_\alpha + z_\gamma}{z_\alpha + z_\beta}\right)$. This result follows since $\frac{V}{1-e^{-\lambda V}} \approx \frac{1}{\lambda} + \frac{V}{2}$ for small λV so that $\hat{\theta} \approx \frac{\theta(1+\tau)}{1+\tau\theta}$, where $\tau = \frac{V\lambda_C}{2}$ (see Appendix B). Table II provides V' for varying θ and median PFS values.

Table II suggests that, for hazard ratios between 0.80 and 0.667, the interval between visits can afford to be no more than about $\frac{1}{2}$ the median PFS time on control to ensure power does not fall below 80%. With a larger hazard ratio of 0.50, the interval between visits can be longer, up to approximately $\frac{2}{3}$ the median PFS time on control.

The Type I error is not inflated providing the scheduling of visits is the same on E and C. However, Table III illustrates the degree to which the Type I error can be increased when clinic visits are asymmetric between treatments.

While it is unlikely that clinic visits would intentionally be scheduled asymmetrically, Table III serves to illustrate the importance in practice of closely matching visit schedules when performing routine log rank analyses of PFS time.

Possible design and analysis strategies when the exact timing of progression is unknown and assigned to the clinic visit at which it was detected

When assigning the time of progression to the visit at which it is detected, bias is introduced and

Table I. Bias and loss of power associated with assigning time of progression to the scheduled clinic visit at which it was detected.

Hazard ratio, θ	Median PFS on E (months)	Median PFS on C (months)	Interval between clinic visits, V, (months)	Expected HR[a], $\hat{\theta}$	Log rank power[b] (%)	Relative increase in d to compensate for loss in power[c]
0.667	6	4	0.5	0.677	87.8	1.07
			1	0.686	85.4	1.16
			2	0.705	80.0	1.34
			4	0.740	67.2	1.81
0.75	8	6	0.5	0.755	88.5	1.05
			1	0.761	86.9	1.11
			2	0.771	83.3	1.23
			4	0.792	75.0	1.51
0.80	12	9.6	0.5	0.803	89.1	1.03
			1	0.806	88.1	1.07
			2	0.811	85.9	1.14
			4	0.822	81.1	1.30

[a] HR = hazard ratio via equation (2).
[b] Log rank power via equation (3); assuming trial originally powered at 90% ($\beta = 0.1$), 2.5% 1-sided α level to detect a HR size θ.
[c] Via equation (4).

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

Table II. Maximum inter-visit interval length to maintain at least 80% power[a] for varying θ and median PFS values.

Hazard ratio, θ	$\frac{2}{\ln(2)}\frac{1}{\theta}\left(\frac{\theta^k-\theta}{1-\theta^k}\right)$	Median PFS on C (months)	Visits at least every V' months
0.8	0.5058	4	2.0
		6	3.0
		9	4.6
		12	6.1
0.75	0.5219	4	2.1
		6	3.1
		9	4.7
		12	6.3
0.667	0.5520	4	2.2
		6	3.3
		9	5.0
		12	6.6
0.50	0.6315	4	2.5
		6	3.8
		9	5.7
		12	7.6

[a]Assuming trial originally powered at 90% ($\beta=0.1$), 2.5% 1-sided α level to detect a HR size θ.

power is decreased. Some options to address this problem are as follows:

(i) Do nothing. If clinic visits are scheduled symmetrically between treatments, ensure they occur at least every V' months as per equation (5) and accept some loss in power. This is the approach most commonly taken in the analysis of PFS times. Some other approaches that might be adopted are given below.

(ii) Increase the target number of events to $d'=\frac{d}{\omega^2}$. Note that since PFS is a mixture of assumed progression times and known times to death, this increase will be somewhat conservative.

(iii) Since $\hat{\theta}$ is known to be biased, use rather

$$\breve{\theta}=\frac{\ln\left(1-\frac{V}{\overline{T}_E}\right)}{\ln\left(1-\frac{V}{\overline{T}_C}\right)} \tag{6}$$

as the asymptotically unbiased maximum likelihood estimate of the hazard ratio with estimated variance

$$\hat{\mathrm{V}}\mathrm{ar}[\ln(\breve{\theta})]=\frac{V^2}{d_E\overline{T}_E^2\left\{\ln\left(1-\frac{V}{\overline{T}_E}\right)\right\}^2\left(1-\frac{V}{\overline{T}_E}\right)}+\frac{V^2}{d_C\overline{T}_C^2\left\{\ln\left(1-\frac{V}{\overline{T}_C}\right)\right\}^2\left(1-\frac{V}{\overline{T}_C}\right)} \tag{7}$$

(see Appendix A). The lack of bias in this estimate is illustrated by simulation in Table IV.

This approach represents an interval-censored analysis as described by Stone *et al.* [21] and Whitehead [22]. Note that, with SE $\breve{\theta}$ being close to that expected from a log rank analysis of actual PFS times $\left(\sqrt{\frac{4}{200}}=0.1414\right)$, this approach requires little if any increase in trial size. It is also important to note that both $\breve{\theta}$ and SE$\breve{\theta}$ vary little as V increases. This suggests, contrary to common belief, there is little to be gained by the imposition of very frequent clinic visits – when data are analysed on an interval-censored basis, frequent visit scheduling is unnecessary and would serve only to impose an unnecessary burden on patients and investigators alike. Note that while simulations in Table IV are based on exponentially distributed PFS times, distribution-free interval-censored analyses are possible via PROC LIFETEST in SAS [23] and Prentice and Gloeckler provide a method for interval-censored analyses via Cox regression [24]; both approaches require a common visit schedule between treatments.

(iv) If, despite protocol intent, clinic visits are not executed exactly as planned leading to variable spacing between visits and asymmetry in schedules between treatments, PROC LIFEREG can be used to estimate event rates on E and C assuming exponentially distributed PFS times (and alternative distributions), and thus provides an unbiased comparison of treatments. In practice, PFS times will not always follow an exponential distribution making interval-censored analyses in these circumstances difficult. However, Sun *et al.* give a generalized formulation of the log rank test applicable to interval-censored data that provides a score statistic to test equality of survival

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Table III. Inflation in Type I error resulting from asymmetric visit scheduling in a trial with 508 events (sized to detect an assumed hazard ratio of 0.75, 90% power, 2.5% 1 sided α).

Median PFS on E and C (HR = 1)	Interval between visits on C (months)	Interval between visits on E (months)	Expected HR[a], $\hat{\theta}$	Type I error[b] (1-sided)
4	0.5	1	0.959	0.069
	1	2	0.920	0.152
6	1	1.5	0.972	0.050
	1	2	0.945	0.092
	2	3	0.946	0.090
9	1	1.5	0.981	0.040
	2	3	0.963	0.062
	3	4	0.964	0.061
12	1	2	0.973	0.050
	2	3	0.972	0.050
	3	4	0.972	0.050
	4	6	0.946	0.090

[a] HR = hazard ratio via equation (2).

[b] Type I error $= \phi^{-1}\left[-1.96 - \frac{\ln\left(\hat{\theta}\right)}{\sqrt{\frac{4}{508}}}\right]$.

Table IV. Hazard ratio estimates resulting from 1000 simulations of a trial with 200 patients (100 per arm) in which all patients achieve an event.

Hazard Ratio, θ	Median PFS on E (months)	Median PFS on C (months)	Interval between clinic visits, V, (months)	Expected value of $\hat{\theta}$[a]	$\hat{\theta}$[b]	$\breve{\theta}$[c]	SE $\ln\left(\breve{\theta}\right)$[d]
0.667	6	4	0.5	0.677	0.679	0.672	0.1438
			1	0.686	0.681	0.669	0.1418
			2	0.705	0.690	0.664	0.1388
			4	0.740	0.718	0.668	0.1428
0.75	8	6	0.5	0.755	0.751	0.746	0.1483
			1	0.761	0.759	0.750	0.1438
			2	0.771	0.764	0.748	0.1376
			4	0.792	0.782	0.751	0.1433
0.80	12	9.6	0.5	0.803	0.804	0.802	0.1425
			1	0.806	0.808	0.803	0.1440
			2	0.811	0.811	0.802	0.1377
			4	0.822	0.817	0.799	0.1474

[a] Expected value of $\hat{\theta}$ via equation (2) to illustrate the closeness of the simulation to the theoretical result.

[b] $\hat{\theta}$ = geometric mean of 1000 hazard ratios based on analysis of PFS time where timing of progression is assigned to the visit at which is was detected.

[c] $\breve{\theta}$ = geometric mean of 1000 simulated hazard ratios based on equation (6).

[d] SE $\breve{\theta}$ = standard deviation 1000 simulated hazard ratios based on equation (6).

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

distributions [25]. While there is no means given for estimating an overall treatment effect such as the hazard ratio, at least a p-value can be obtained to assess the strength of evidence against the null.

(v) As an alternative to the analysis of PFS time, treatments could be compared on the basis of the overall number of PFS events occurring at any time during the trial follow-up period thus circumventing issues associated with the over-estimation of PFS time, scheduling of visits and any asymmetry between treatments.

Under proportionality, an analysis with a complementary log-log link function [21,24,26,27] would provide an unbiased estimate of the hazard ratio as

$$\tilde{\theta} = \frac{\ln(1-p_E)}{\ln(1-p_C)} \tag{8}$$

where p_E and p_C are the proportions of patients with a PFS event on treatment E and C. The estimated variance of $\ln \tilde{\theta}$ by Taylor series expansion is

$$\hat{\mathrm{V}}\mathrm{ar}[\ln(\tilde{\theta})] = \frac{p_E}{N(1-p_E)\{\ln(1-p_E)\}^2} + \frac{p_C}{N(1-p_C)\{\ln(1-p_C)\}^2} \tag{9}$$

It is interesting to note that $\tilde{\theta}$ and $\hat{\mathrm{V}}\mathrm{ar}[\ln(\tilde{\theta})]$ coincide with $\tilde{\theta}$ and $\hat{\mathrm{V}}\mathrm{ar}[\ln(\tilde{\theta})]$ when $V = R + F$, that is, when there is just one assessment of progression coinciding with the end of the trial period.

Further, by noting that $S_C(t)^{\theta} = S_E(t)$ where θ is the true hazard ratio and, with no censoring, that $S_E(R+F) = p_E$ and $S_C(R+F) = p_C$ so that $\tilde{\theta} = \theta$ and the number of events expected on E and C are $N[S_E(R+F)]$ and $N[S_C(R+F)] = Np_E$ and Np_C, it is possible to compare the power of the log rank test on exact PFS times with the power of an analysis with a complementary log–log link based only on the total number of events occurring over the trial follow-up period. Under these circumstances, the relative efficiency of the two tests is given by

$$\frac{\frac{1}{Np_E} + \frac{1}{Np_C}}{\mathrm{Var}[\ln(\tilde{\theta})]} = \left[\frac{1}{p_E} + \frac{1}{p_C}\right] \times \left[\frac{p_E}{(1-p_E)\{\ln(1-p_E)\}^2} + \frac{p_C}{(1-p_C)\{\ln(1-p_C)\}^2}\right]^{-1}$$

Assuming 90% power in the log rank test, Figure 1 plots the relative efficiency for values of $S_C(t)$ from 0.05 to 0.95.

In line with work earlier work, Figure 1 indicates that, under proportionality, a comparison on the overall number of PFS events over the follow-up period is associated with little loss of power relative to the log rank test on exact PFS times providing fewer than around 50% of patients have reached an event [28,29]. If fewer than 75% of patients have reach an event, the loss in power is, at most, 5%. It is not until 90% or more have reached an event that the power of the relative risk test dips below 80% to around 77%. For exponentially distributed times to event, since the probability of an event over the trial follow-up period $\approx 1 - \mathrm{e}^{-\lambda(0.5R+F)}$, fewer than 75% events will in general be assured if the median follow-up at the time of the analysis is not more than two times the median PFS time [20].

This suggests that in those trial settings where progression of disease is the primary focus but significant concerns persist regarding the assumed time of progression, a supportive analysis based on the number of patients with a PFS event over the trial follow-up period can provide reassurance. This analysis is unbiased under proportionality and suffers relatively little loss of power under common trial circumstances. It also offers the opportunity to simplify clinical trial design. It might be possible, for example, to envisage a trial where progression is assessed as per clinical practice with a requirement for objective verification of any suspected disease progression. At a minimum and in addition to the baseline assessment of disease, a single mandatory assessment at the end of the trial follow-up period would be

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

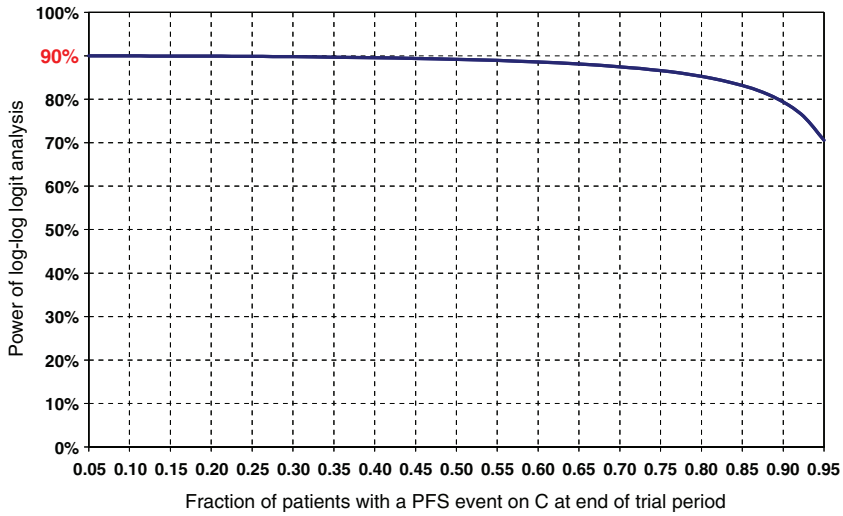


Figure 1. Power of a comparison based on the number of patients with a PFS event over the trial period relative to the log rank test on PFS time with 90% power.

required in patients who had not previously progressed in order to catch any missed progressions. PFS events would then be counted over the trial follow-up period and treatments compared via an analysis with a complementary log-log link function.

CENSORING ON DROPOUT DUE TO AE OR ADDITIONAL ANTI-CANCER THERAPY

FDA's recent Draft Guidance for Industry on Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics recommends that patients who stop taking randomized therapy prior to documented progression are censored at the time randomized treatment is stopped [15]. The rationale for this recommendation is not provided explicitly, but seems to be related to a concern that PFS times may be over-estimated otherwise. Patients who die in the absence of documented progression remain an event, irrespective of whether the death occurred whilst the patient was still receiving or some time after stopping randomized therapy.

This approach is, unfortunately, highly problematic since it ignores the issue of informative censoring. Patients who stop taking randomized therapy prior to documented progression frequently do so due to either toxicity of the drug or due to a deterioration in the status of their disease. In such cases, the treating physician often judges that immediate intervention, commonly in terms of the introduction of a new cancer treatment, is in the best interests of the patient without necessarily waiting for confirmatory, radiographic evidence of progressive disease.

Time to progression therefore cannot be censored in the analysis as the censoring mechanism is self evidently informative. In such circumstances, if the prevalence of censoring differs between arms, naive censoring could lead to extremely biased results and, ultimately, incorrect licensing decisions [30]. Figure 2 provides a simple illustration of the problem.

Suppose E is compared to C in a trial of 100 patients, 50 per arm. Suppose on C that 25 patients progress whilst taking drug at a mean

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

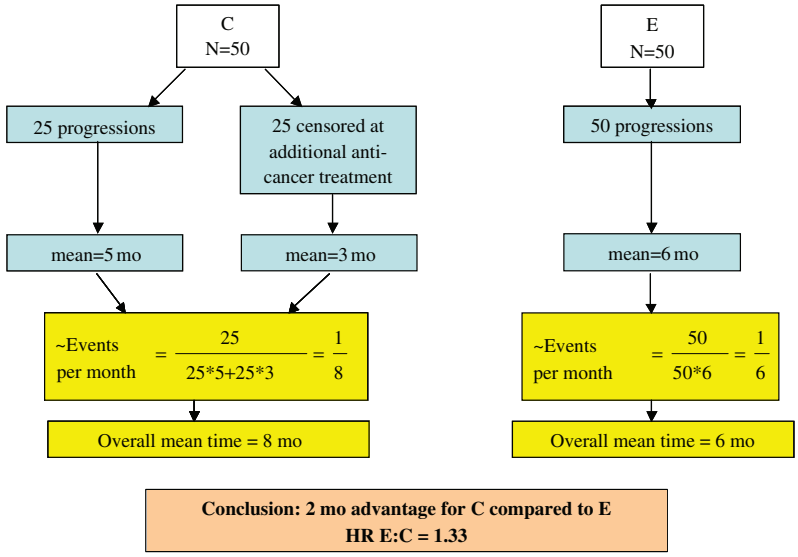


Figure 2. Censoring on the addition of further anti-cancer therapy.

time of 5 months and the other 25 patients receive additional anti-cancer treatment prior to documented progression at a mean time of 3 months. Suppose on E that all 50 patients progress at a mean time of 6 months; no patients received additional anti-cancer treatment. It is obvious that E is the better treatment, with a longer time to progression and no need for additional anti-cancer treatment. However, suppose now that the data are subject to formal statistical analysis with the 25 patients on C who received additional anti-cancer treatment censored for progression. The progression event rate on drug C is therefore $\frac{1}{8}$ progressions per patient per month compared to $\frac{1}{6}$ progressions per patient per month on drug E, giving mean PFS times of 8 and 6 months for C and E, respectively, and a hazard ratio of 1.33, leading to a conclusion that, in fact, drug C is better than E. Clearly, there is a problem with a recommended statistical analysis when it leads to a conclusion that the less efficacious and more toxic treatment is better. If, however, those on C who received additional anti-cancer therapy are treated rather as failures as in Figure 3, a more sensible conclusion is reached that E is in fact better than C.

Hence, recommendations to censor patients who stop taking randomized treatment prior to documented progression, perhaps due to the use of additional anti-cancer therapy owing to a deterioration in their condition or due to toxicity, are inherently flawed and should be avoided. This practice, if adopted, not only results in informative censoring but also contravenes the basic principle of an intent-to-treat analysis which is the accepted standard for the comparison of treatments for survival. If a similar approach was applied to the analysis of survival, then only those deaths occurring on randomized treatment would be considered when comparing treatments with all other deaths censored. The interpretation of such an analysis is, at best, unclear and its relevance to the assessment of treatment policies questionable. Overall, it would seem better and more consistent to apply a common standard to important efficacy variables such as PFS and survival to allow both to be interpreted within the same framework. For

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

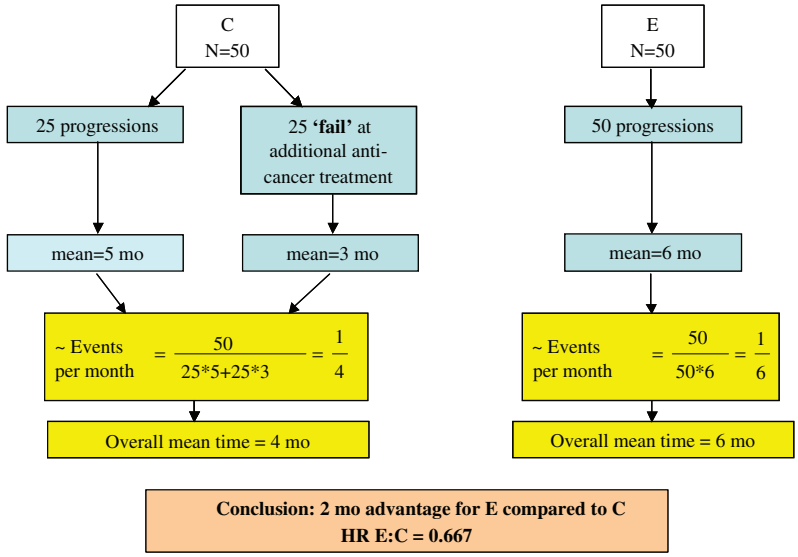


Figure 3. Addition of further anti-cancer therapy considered as a 'failure'.

progression (like survival) this would mean the routine follow-up patients for documented evidence of progression irrespective of when and why they stop taking randomized treatment so that treatment policies could be compared on the basis of data that reflect actual clinical practice. This is essentially the same approach as forwarded in the recently published appendix to the CHMP anti-cancer guideline [16].

SUMMARY, RECOMMENDATIONS AND DISCUSSION

This paper has focused on some key statistical issues associated with the analysis of PFS in oncology trials. The routine practice of assigning the time of progression to the clinic visit at which it was first detected results in a downwardly biased estimate of the hazard ratio and, thus, reduces power. Further, if clinic visit schedules are not closely matching between treatments, the Type I error can also be increased. Fortunately, these issues can be addressed as follows:

- Size the trial to detect a true HR of size θ via the log rank test. Assume E events are required to provide power of $1-\beta$ with a 1-sided Type I error rate of α Plan for, and maintain during conduct, a common clinic visit schedule.
- Employ an interval-censored analysis of PFS.
- Alternatively, PFS times can be assigned to clinic visit at which they were first detected and PFS time analysed via the usual log rank test; however, to maintain power at $1-\beta$ the target number of events should be increased to d' (as defined in equation (4)).
- If, as is common in practice, the visit schedule is not as closely adhered to as intended, resulting in variability in the interval between visits and, possibly, between treatments also, a supportive analysis based on the number of PFS events over the trial period will provide for an

Copyright © 2007 John Wiley & Sons, Ltd.

Pharmaceutical
STATISTICS

unbiased comparison between treatments. An analysis with a complementary log–log link will provide an estimate of the hazard ratio with reasonable power so long as no more than around 75% of patients have an event.

To illustrate the problems associated with assigning the time of progression to the visit at which it was first detected, it has been assumed event times are exponentially distributed since this allows the reader to most easily appreciate the extent to which bias can be introduced and how alternative estimators might be formulated to eliminate this bias. An area of further work might be to look at how rank based estimators of the hazard ratio (such as the Pike estimator or the exponent of the ratio of observed minus expected deaths to the variance from the log rank test) perform when PFS times are not known exactly [31,32]. It might also be of interest to examine other distributions for PFS times, such as the Weibull or log Normal and the performance of $\breve{\theta}$ when proportionality holds but the underlying distribution of PFS times is not exponential.

With respect to patients who stop randomized therapy during the trial follow-up period prior to reaching a confirmed progression event due to toxicity or the addition of further anti-cancer therapy, the common practice of censoring at the time of the intermediate event is highly problematic and is likely to favour the less efficacious, more toxic treatment. Adopting the ITT approach used in the analysis of survival, whereby all patients are followed for a documented evidence of progression irrespective of when and why they stop taking randomized treatment, would provide (i) a better basis for comparing treatment policies and (ii) data that more closely mimic actual clinical practice. If desired, a supportive analysis could still be conducted censoring dropouts in the absence of documented progression, though considerable care would be needed when interpreting the results.

The issues raised in this article are not the only concerns that impact the use of PFS as an endpoint to demonstrate drug effectiveness. Two key issues worth raising briefly are (a) whether an improvement in PFS is a clinical benefit in and of itself or is at least reasonably likely to predict clinical benefit in terms of symptomatic improvement and/or overall survival and (b) the need for independent review of radiological data relating to disease progression. With respect to the first of these issues, recent work in prostate and colorectal cancer has seen the question of surrogacy of PFS for survival carefully and formally examined using contemporary statistical methodology [33–35]. This work supports the use of PFS as a true surrogate endpoint in these disease settings and, in doing so, lends support to the view offered by Williams, that few in the oncology community doubt that delaying the growth of a cancer is of benefit to patients; rather, issues relate to whether progression can be reliably measured in trials and, if so, what a given improvement in progression means clinically [6, transcript p. 30].

The use of open trials in oncology raises the possibility of bias in the assignment of progression status by the treating investigator. As evidenced in both FDA and CHMP guidelines, this concern frequently results in a request from regulatory agencies for independent review of radiographic and imaging data in patients said to have progressed by the investigator [9,15,16]. While this may make sense in open, small scale trials with few investigational sites, the value of independent review in large-scale international trials with possibly hundreds of sites is questionable – when seeking a large effect on progression, a false claim would seem rather unlikely in the absence of a systematic intent to defraud across multiple countries and sites. A further difficulty introduced when incorporating an independent review, is how to handle patients where the investigator and independent review disagree on progression status and where the investigator believes the patient has progressed. In this situation, radiological assessment will cease and any censoring of this data will be informative as such patients will be closer on average to progressing than patients neither the investigator or independent reviewer believe have progressed.

Even when an independent review is deemed worthwhile, the common practice to review only

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

data in patients who have progressed is unsatisfactory at best and misleading at worst. This approach will always lead to a less precise estimate of the treatment effect since the number of progression events can only go down. A more satisfactory approach would be to also take a random sample of non-progressing patients to estimate the fraction of patients without progression reclassified as progressive by independent review. The overall number of progression events could then be estimated under independent review and treatment groups compared accordingly. This approach was used in the review of progression events in the bicalutamide early prostate cancer programme where it was concluded that there was no evidence of bias in the investigator assessment of progression [36].

The trend toward the use of PFS as a primary endpoint to assess the effectiveness of new anticancer treatments is, on the whole, beneficial to drug development and consistent with the aim of FDA's Critical Path and EMEA's Road Map initiatives which actively seek ways to accelerate the drug development process [37,38]. It is hoped that this article will help to address some of the perceived statistical issues related to trial design and analysis and, in doing so, will help to alleviate the concerns and barriers that might otherwise discourage or even prevent the use of PFS as a primary endpoint in oncologic drug development.

ACKNOWLEDGEMENTS

The author would like to thank Andrew Stone, Biostatistics Group, AstraZeneca Pharmaceuticals and three anonymous reviewers for their helpful and constructive comments that served to improve the content and clarity of this paper.

REFERENCES

1. Stadler WM, Ratain MJ. Development of target-based antineoplastic agents. *Investigational New Drugs* 2000; **28**:7–16.
2. Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical trial designs for cytostatic agents: are new approaches needed? *Journal of Clinical Oncology* 2001; **19**:265–272.
3. Simon RM, Steinberg SM, Hamilton M, Hildesheim A, Khleif S, Kwak LW, Mackall CL, Schlom J, Topalian SL, Berzofsky JA. Clinical Trial Designs for the Early Clinical Development of Therapeutic Cancer Vaccines. *Journal of Clinical Oncology* 2001; **19**:1848–1854.
4. Eskens FA, Verweij J. Clinical studies in the development of new anticancer agents exhibiting growth inhibition in models: facing the challenge of proper study design. *Oncology Haematology* 2000; **34**:83–88.
5. Stone A, Wheeler C, Barge A. Improving the design of phase II trials of cytostatic anticancer agents. *Contemporary Clinical Trials* 2006. In press, Corrected Proof, Available online 14 July 2006 at www.scincedirect.com (last accessed 1 October 2006).
6. FDA Oncologic Drugs Advisory Committee: Endpoints in clinical cancer trials and endpoints in lung cancer clinical trials. 16th December 2003. http://www.fda.gov/ohrms/dockets/ac/03/transcripts/4009T1.DOC (last accessed 1 October 2006).
7. FDA Public Workshop on Clinical Trial Endpoints in Prostate Cancer, 21–22 June 2004. http://www.fda.gov/cder/drug/cancer_endpoints/default.htm#prostate (last accessed 1 October 2006).
8. FDA Oncologic Drugs Advisory Committee: Colorectal cancer endpoint discussion. 4th May 2004, transcript page 218. http://www.fda.gov/ohrms/dockets/ac/04/transcripts/4037T2.DOC (last accessed 1 October 2006).
9. CHMP Guideline on the evaluation of anticancer medicinal products in man. December 2005. http://www.emea.eu.int/pdfs/human/ewp/020595en.pdf (last accessed 1 October 2006).
10. FDA/CDER. FDA Project on Cancer Drug Approval Endpoints, 2003: http://www.fda.gov/cder/drug/cancer_endpoints/default.htm (last accessed 1 October 2006).
11. FDA/CDER New and Generic Drug Approvals: NEXAVAR (sorafenib) product labelling, December 2005. Label: http://www.fda.gov/cder/foi/label/2005/021923lbl.pdf (last accessed 1 October 2006). Approval letter: http://www.fda.gov/cder/foi/appletter/2005/021923ltr.pdf (last accessed 1 October 2006).
12. FDA/CDER New and Generic Drug Approvals: VECTIBIX (panitumumab) product labelling, September 2006. Label: http://www.fda.gov/cder/foi/label/2006/125147s0000lbl.pdf (last accessed 1 October 2006).
13. Williams G, He K, Chen G, Chi G, Pazdur R. Operational Bias in assessing time to progression (TTP). *Proceedings of the American Society of Clinical Oncology* 2002; abstr 975.
14. FDA Oncologic Drugs Advisory Committee: Genasense™ (oblimersen sodium) Injection for

Copyright © 2007 John Wiley & Sons, Ltd.

Pharmaceutical
STATISTICS

Advanced Melanoma in Combination With Dacarbazine (DTIC). 4th May 2004. Statistical Review. http://www.fda.gov/ohrms/dockets/ac/04/briefing/4037B1_04_C-FDA-Statistical%20Review-RSR13.doc and http://www.fda.gov/ohrms/dockets/ac/04/slides/4037S1_04_FDA-Sridhara-Ridenhour.ppt (last accessed 1 October 2006).
15. FDA/CDER Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics DRAFT GUIDANCE, 2005. http://www.fda.gov/cder/guidance/6592dft.htm (last accessed 1 October 2006).
16. CHMP Draft appendix 1 to the guideline on the evaluation of anticancer medicinal products in man. London, 27 July 2006. http://www.emea.eu.int/pdfs/human/ewp/26757506en.pdf#search = %22emea%20anti-cancer%20guideline%20appendix%22 (last accessed 1 October 2006).
17. Schoenfeld DA. The asymptotic properties of nonparamteic tests for comparing survival distributions. *Biometrika* 1981; **68**:316–318.
18. Rubinstein LV, *et al*. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases* 1981; **34**:469–479.
19. Yateman NA, Skene AM. Sample sizes for proportional hazards survival studies with arbitrary patient entry and loss to follow-up distributions. *Statistics in Medicine* 1992; **11**:1103–1113.
20. Bernstein D, Lagakos SW. Sample size and power determination for stratified clinical trials. *Journal of Statistical Computer Simulations* 1978; **8**:65–73.
21. Stone A, Wheeler C, Carroll K, Barge A. Optimizing randomized phase II trials assessing tumor progression. *Contemporary Clinical Trials* 2006. In press, Corrected Proof, Available online 19 May 2006 at www.scincedirect.com (last accessed 1 October 2006).
22. Whitehead J. The analysis of relapse clinical trials, with application to a comparison of two ulcer treatments. *Statistics in Medicine* 1989; **8**:1439–1454.
23. SAS/STAT® *User's Guide*. *Version 6* (4th edn), Vol. 2. SAS Institute Inc.: Cary, NC, 1989.
24. Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 1978; **34**:57–67.
25. Sun J, Zhao Q, Zhao X. Generalised log-rank tests for interval-censored failure time data. *Scandinavian Journal of Statistics* 2005; **32**:49–57.
26. Cox DR, Oakes D. *Analysis of Survival Data*. Chapman & Hall: London, 1984.
27. Collett D. *Modeling Survival Data in Medical Research*. Chapman & Hall: London, 1994.
28. Cuzick J. The efficiency of the proportions test and the log rank test for censored survival data. *Biometrics* 1982; **38**:1033–1039.
29. Gail MH. Applicability of sample size calculations based on a comparison of proportions for use with the log rank test. *Controlled Clinical Trials* 1985; **6**:112–119.
30. DiRienzo AG. Nonparametric comparison of two survival-time distributions in the presence of dependent censoring. *Biometrics* 2003; **59**:497–504.
31. Berry G, Kitchin RM, Mock PA. A comparison of two simple hazard ratio estimators based on the logrank test. *Statistics in Medicine* 1991; **10**: 749–755.
32. Sellke T, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika* 1983; **70**:315–326.
33. Collette L, Burzykowski T, Carroll KJ, *et al*. Is prostate-specific antigen a valid surrogate end point for survival in hormonally treated patients with metastatic prostate cancer? Joint Research of the European Organisation for Research and Treatment of Cancer, the Limburgs Universitair Centrum, and AstraZeneca Pharmaceuticals. *Journal of Clinical Oncology* 2005; **23**:6139–6148.
34. Buyse M, Burzykowski T, Carroll K, *et al*. Progression-free survival (PFS) as a surrogate for overall survival (OS) in patients with advanced colorectal cancer: an analysis of 3159 patients randomized in 11 trials. *Proceedings of the American Society of Clinical Oncology* 2005; abstr 3513.
35. Molenberghs G, Buyse M, Geys H, *et al*. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials* 2002; **23**:607–625.
36. FDA/CDER Public Workshop on Clinical Trial Endpoints in Prostate Cancer, 21–22 June 2004. Re-evaluation of Radiographic Outcomes: The Casodex Early Prostate Cancer Trial Program Experience. http://www.fda.gov/cder/drug/cancer_endpoints/ProstatePresent/Carroll.ppt (last accessed 1 October 2006).
37. FDA Challenge and opportunity on the Critical Path to New Medical Products. March 2004. http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html (last accessed 1 October 2006).
38. The European Medicines Agency Road Map to 2010: Preparing the Ground for the Future. March 2005. http://www.emea.eu.int/pdfs/general/direct/directory/3416303enF.pdf (last accessed 1 October 2006).

APPENDIX A

Suppose N patients are followed for some event of interest with the time to the event denoted as t.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Suppose t is an exponentially distributed random variable with parameter λ. Suppose the process is monitored at r equally spaced intervals of length V for a total follow-up time of rV. Events therefore occur in intervals $(0, V]$, $(V, 2V]$,...,$((r-1)V, rV]$ with the event for the ith patient occurring in $((k_i-1)V, k_iV]$. Events times are assigned to the start of the interval in which they are detected. There are d patients with an event. Furthermore, c patients are randomly censored prior to time rV with censoring times of k_gV for the gth patient. The $N-d-c$ remaining patients who are without an event at the end of the follow-up period are right censored at time rV. The likelihood function is therefore given by

$$L = \prod_{i=1}^{d} \left(e^{-\lambda(k_i-1)V} - e^{-\lambda k_i V}\right) \prod_{g=1}^{c} e^{-\lambda k_g V} \prod_{j=1}^{N-d-c} e^{-\lambda rV} \tag{A1}$$

$$\ell = \ln(L) = d\ \ln\left(e^{\lambda V} - 1\right) - \lambda\left[\sum_{i=1}^{d} k_i V + \sum_{g=1}^{c} k_g V + \sum_{j=1}^{N-d-c} rV\right] \tag{A2}$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{dVe^{\lambda V}}{e^{\lambda V} - 1} - \left[\sum_{i=1}^{d} k_i V + \sum_{g=1}^{c} k_g V + \sum_{j=1}^{N-d-c} rV\right] \tag{A3}$$

$$\frac{\partial \ell}{\partial \lambda} = -d\left[\frac{\sum_{i=1}^{N} T_i}{d} - \frac{V}{1-e^{-\lambda V}}\right] \tag{A4}$$

where T_i denotes the observed time, event or censored, of the ith patient. Thus

$$\overline{T} = \frac{\sum_{i=1}^{N} T_i}{d} = \frac{1}{\text{observed event rate}}$$

is the MLE for

$$\frac{V}{1-e^{-\lambda V}} \text{ and } \text{Var}\left[\overline{T}\right] = \frac{V^2 e^{-\lambda V}}{d(1-e^{-\lambda V})^2}$$

If comparing two treatments, experimental (E) and control (C), the estimated hazard ratio

$$\hat{\theta} = \frac{\text{observed event rate}_E}{\text{observed event rate}_C} = \frac{\overline{T}_C}{\overline{T}_E}$$

is biased since

$$E[\hat{\theta}] = \frac{1-e^{-\lambda_E V}}{1-e^{-\lambda_C V}} \neq \frac{\lambda_E}{\lambda_C} \tag{A5}$$

Further,

$$\text{Var}[\ln(\hat{\theta})] = \frac{e^{-\lambda_E V}}{d_E} + \frac{e^{-\lambda_C V}}{d_C} \tag{A6}$$

and, thus,

$$\hat{\text{Var}}[\ln(\hat{\theta})] = \frac{1-\frac{V}{\overline{T}_E}}{d_E} + \frac{1-\frac{V}{\overline{T}_C}}{d_C} \tag{A7}$$

An approximately unbiased estimate of the HR is given by

$$\breve{\theta} = \frac{\ln\left(1-\frac{V}{\overline{T}_E}\right)}{\ln\left(1-\frac{V}{\overline{T}_C}\right)} \tag{A8}$$

with variance

$$\text{Var}[\ln(\breve{\theta})] = \frac{e^{\lambda_E V}\left(1-e^{-\lambda_E V}\right)^2}{d_E \lambda_E^2 V^2} + \frac{e^{\lambda_C V}\left(1-e^{-\lambda_C V}\right)^2}{d_C \lambda_C^2 V^2} \tag{A9}$$

and, thus,

$$\hat{\text{Var}}[\ln(\breve{\theta})] = \frac{V^2}{d_E \overline{T}_E^2 \left\{\ln\left(1-\frac{V}{\overline{T}_E}\right)\right\}^2 \left(1-\frac{V}{\overline{T}_E}\right)} + \frac{V^2}{d_C \overline{T}_C^2 \left\{\ln\left(1-\frac{V}{\overline{T}_C}\right)\right\}^2 \left(1-\frac{V}{\overline{T}_C}\right)} \tag{A10}$$

APPENDIX B

Suppose two treatments are to be compared in a clinical trial on a time to event endpoint using the log rank test. To test the hypotheses H_0: hazard ratio = 1 vs H_1: hazard ratio = θ (< 1) with a 1-sided

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

Type I error rate of α and power $1-\beta$ a total of

$$d = \frac{4\left(z_\alpha + z_\beta\right)^2}{\ln(\theta)^2} \tag{B1}$$

events are required [18]. It therefore follows immediately that

(i) if the power to detect $\hat{\theta}$, $\hat{\theta} > \theta$, with d events is $1-\gamma$, then

$$z_\gamma = \left(z_\alpha + z_\beta\right)\omega - z_\alpha \tag{B2}$$

so that

$$1 - \gamma = \phi^{-1}\left[\left(z_\alpha + z_\beta\right)\omega - z_\alpha\right] \tag{B3}$$

where $\omega = \mathrm{abs}\left(\frac{\ln(\hat{\theta})}{\ln(\theta)}\right)$ and $\phi^{-1}(.)$ is the inverse of the cumulative normal distribution.

(ii) there is a simple relationship between θ and $\hat{\theta}$ such that

$$\hat{\theta} = \theta^k \tag{B4}$$

where $k = \mathrm{abs}\left(\frac{z_\alpha + z_\gamma}{z_\alpha + z_\beta}\right)$

(iii) to maintain power of $1-\beta$ to detect $\hat{\theta}$ a total of

$$d' = \frac{d}{\omega^2} \tag{B5}$$

events are required.

Suppose now that $\hat{\theta} = \frac{1-\mathrm{e}^{-\lambda_E V}}{1-\mathrm{e}^{-\lambda_C V}}$. If λV is small, then

$$\begin{aligned}\frac{V}{1-\mathrm{e}^{-\lambda V}} &\approx \frac{1}{\lambda}\left[\frac{1}{1-\frac{\lambda V}{2}+O(\lambda^2 V^2)}\right] \\ &= \frac{1}{\lambda}\left(1+\frac{\lambda V}{2}+O(\lambda^2 V^2)\right) \\ &\approx \frac{1}{\lambda}+\frac{V}{2}\end{aligned} \tag{B6}$$

so that

$$\hat{\theta} = \frac{1-\mathrm{e}^{-\lambda_E V}}{1-\mathrm{e}^{-\lambda_C V}} \approx \frac{\frac{1}{\lambda_E}+\frac{V}{2}}{\frac{1}{\lambda_C}+\frac{\lambda_C}{2}} = \frac{\theta(1+\tau)}{1+\tau\theta} \tag{B7}$$

where $\tau = \frac{V\lambda_C}{2}$. Substitution of (B7) in to (B4) reveals that if a common visit schedule is used when assessing PFS such that PFS times are assigned to the visit at which progression was first detected, to ensure retention of at least $1-\gamma$ power, $\gamma > \beta$, visits must be scheduled approximately every V' months where

$$V' = (\text{median PFS on } C) \times \frac{2}{\ln(2)}\,\frac{1}{\theta}\left(\frac{\theta^k - \theta}{1-\theta^k}\right) \tag{B8}$$

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

8. **Carroll KJ**. Decision Making from Phase II to Phase III and the Probability of Success: Reassured by 'Assurance'? Journal of Biopharmaceutical Statistics, 2013; Vol 23(5):1188-1200.

Journal of Biopharmaceutical Statistics, 23: 1188–1200, 2013
Copyright © Taylor & Francis Group, LLC
ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2013.813527

Taylor & Francis
Taylor & Francis Group

DECISION MAKING FROM PHASE II TO PHASE III AND THE PROBABILITY OF SUCCESS: REASSURED BY "ASSURANCE"?

Kevin J. Carroll
Independent Statistical Consultant, Cheshire, United Kingdom

With Phase III failure rates of 50%, better ways of predicting late-stage success are needed. One concept that has been used is "assurance." Rather than conventional power calculations hypothesizing a known effect of a drug, assurance provides an expected power calculation based on some prior distribution for the treatment effect. It therefore has appeal in Phase III planning and decision making, especially when the prior is based on Phase II data. However, assurance has counterintuitive properties that can serve to confuse and concern the nonstatistician. Appreciation of these properties is helpful to ensure an informed use of assurance in strategic drug development.

Key Words: Assurance; Decision making; Phase II to phase III.

1. INTRODUCTION

An ongoing and serious challenge facing the pharmaceutical industry is the high failure rate in the latter stages of drug development, resulting in low research and development (R&D) productivity. Failure rates of 80% in Phase II and 50% in Phase III have been reported (Arrowsmith, 2011a,b). Two-thirds of Phase III failures are reported as due to not demonstrating a positive treatment effect reflecting poorly on the quality of Phase II design and decision making (Arrowsmith, 2011b). While Prinz et al. (2011) point to unreliable models used in drug screening programs, broader issues in late stage development are highlighted by Arrowsmith (2011b) where he cites:

> a result of the pressure on companies to replenish pipelines with drugs that have high potential for approval and reimbursement, particularly in a period during which patent expiries for major products are threatening future revenues. Owing to this urgency, it seems that companies have progressed drugs into Phase III trials even though they only displayed marginal statistically significant efficacy in Phase II proof-of-concept studies; consequently, these drugs carry a greater than average risk of failure.

Arrowsmith (2011b) further notes:

> The way to improve Phase III success rates is to avoid wishful thinking and to rely on high-quality scientific evidence by fully testing mechanisms against each

Received November 29, 2012; Accepted March 19, 2013
Address correspondence to Kevin J. Carroll, 79 Albany Road, Bramhall, Cheshire SK7 1NE, UK; E-mail: kevincarroll2@sky.com

target indication, using well-defined endpoints in the right patient population in Phase II trials. Initially, this may lead to higher failure rates in Phase II trials, but some companies have already shown that good science can deliver a steady flow of robust positive proof-of-concept data.

While the lack of statistical significance in Phase II is, arguably, not a major contributor to Phase III failure, Phase II design and decision making are self-evidently critical. Given that the cost of developing a new drug is estimated to be $800 million to $2 billion, high Phase II and Phase III failure rates are clearly unsustainable and pressure is on to improve the situation (Chuang-Stein et al., 2011). Over recent years this has led to the demand for more adaptive Phase II and Phase III trial designs and alternative statistical methodologies such as Bayes that purport to offer greater flexibility, lower cost, and better success rates. Despite the promise, to date very few new medicines have been approved based upon a truly adaptive Phase III trial design. Another area of focus is to find better ways of decision making when transitioning from Phase II to Phase III and, in particular, better ways using Phase II data to predict the chance of success in Phase III.

One statistical concept that has been used in this regard is "assurance" (Chuang-Stein, 2006; O'Hagan et al., 2005). Rather than a traditional power calculation that hypothesizes a fixed treatment effect, assurance provides an unconditional expected power calculation based on some prior distribution for the treatment effect. As such, it has a natural appeal in late-stage decision making and Phase III design, particularly when the prior is not arbitrarily chosen but rather based on objective data generated in preceding Phase II trials. Chuang-Stein et al. (2011) consider the use of assurance in late-stage development decision making, and Kirby et al. (2012) consider its use in designing Phase III studies given prior, discounted Phase II data. Despite its increasing prominence in the literature, assurance has some strange and counterintuitive properties that can serve to confuse and concern the nonstatistician decision maker. It is helpful for these properties to be understood by statisticians and nonstatisticians alike in order to ensure an informed use of the concept in strategic drug development. Aside from assurance, there are some basic steps the drug developer can take to enhance and improve late-stage decision making. In the end it is probably these basic steps, coupled with experience and good judgment, that are more likely to result in an improved Phase III success rate than the application of concepts like assurance per se, since without the former, the latter is entirely academic.

The remainder of this article is structured as follows. Section 2 reviews power and assurance, making several observations regarding the nature of assurance when the prior is defined by preceding Phase II data. Section 3 discusses assurance for two independent Phase III trials. Section 4 provides an example, and Section 5 summarizes the key features of assurance. Section 6 offers some basic recommendations for good decision making and Section 7 closes with a brief discussion.

2. POWER AND ASSURANCE

Consider the planning of a Phase III clinical trial. Typically, the hypothesis to be tested is $H_0 : \theta_{TRUE} = 0$ vs. $H_1 : \theta_{TRUE} \neq 0$, where θ_{TRUE} denotes the true

treatment effect. For the purposes of sizing, a positive effect $\theta_{\text{TRUE}} = \theta$ (>0) is usually assumed under the alternative. Let x be a sufficient statistic for θ with distribution $f(x \mid \theta) \sim N(\theta, \sigma^2)$, where $N(.,.)$ represents the normal distribution. Trial size is then governed by Type I and Type II errors, α and β, and the need to deliver the required information content, $1/\sigma^2 = (z_\alpha + z_\beta)^2/\theta^2$, where $z_u = \Phi^{-1}(1 - u)$ and $\Phi^{-1}(\cdot)$ represents the inverse standard Normal distribution function. The null hypothesis is rejected in favor of the alternative when $x > c = z_\alpha\sigma$. Suppose further that a preceding Phase II trial has been conducted using the same patient population and endpoint; if there are two or more similar Phase II trials, assume results are statistically consistent and the data combined. Suppose these Phase II data provide an estimate of the treatment effect, m, with variance s^2, and the data are then used to define a prior for θ: $f(\theta) \sim N(m, s^2)$.

It is important to note at this stage, and as highlighted by an anonymous referee, that choices for $f(\theta)$ other than as defined by preceding Phase II data could be considered. However, the intent of this article is to examine assurance in the context of a prior for θ specifically determined by preceding Phase II data since this most accurately reflects the decision-making process in practice. It is the Phase II data themselves that guide and heavily influence Phase III planning, and in some pharmaceutical companies these data are used to directly predict the chances of success in Phase III. With well-conducted and controlled Phase II data in hand, there seems little rationale in introducing an assumed, informative prior for Phase III. To do so would serve to render the decision-making process somewhat academic as the real (Phase II) data may be diluted (or strengthened) or even overlooked by the choice of some possibly arbitrary prior. Arguably, it is the Phase II data in and of themselves that best inform the decision makers regarding the true treatment effect and, hence, form the most appropriate prior for Phase III planning and decision making.

With $f(\theta) \sim N(m, s^2)$ as the prior for θ, it then follows that the joint (posterior) distribution of x and θ is given by $f(x, \theta) = f(x \mid \theta)f(\theta)$. As shown by O'Hagan et al. (2005),

$$f(x) = \int_{-\infty}^{+\infty} f(x, \theta)d\theta = \int_{-\infty}^{+\infty} f(x \mid \theta)f(\theta)d\theta = N(m, s^2 + \sigma^2) \tag{1}$$

"Assurance" is then defined as the probability of a "successful" Phase III trial, that is, the probability of achieving $p < 0.025$ (O'Hagan et al., 2005). This is given by

$$pr(x > c) = pr\left(z > \frac{c - m}{\sqrt{s^2 + \sigma^2}}\right) = 1 - \Phi\left(\frac{c - m}{\sqrt{s^2 + \sigma^2}}\right) \tag{2}$$

where $\Phi(\cdot)$ represents the standard Normal cumulative distribution function. Therefore assurance is the expected value of conventional power over the prior for θ. It should be noted at this point that other definitions of "success" could be used, such as $p < 0.025$, and the point estimate of the treatment effect attaining some minimum, prespecified value (Chuang-Stein et al., 2011; Kirby et al., 2012). However, for the purposes of exploring the properties of "assurance" in terms of expected power, "success" is defined as achieving $p < 0.025$. Some observations can now be made regarding assurance.

(i) When Phase II is Small For a small Phase II, that is, as $s^2 \to \infty$, then, regardless of the size of Phase III,

$$pr\left(z > \frac{c-m}{\sqrt{s^2+\sigma^2}}\right) \to pr(z > 0) = 0.50 \tag{3}$$

Equation (3) tells us that even the very largest Phase III trial has only a 50% probability of success (PoS) when the Phase II data are very few. This may seem obvious to the statistician, considering that the conventional power function is integrated over an essentially flat prior for θ such that all values on the real line are equally plausible. However, this situation can be rather confusing and counterintuitive to the nonstatistician who would consider the chance of achieving $p < 0.025$ to be virtually guaranteed in the largest of Phase III trials where there would be essentially 100% power for any $\theta_{\text{TRUE}} > 0$, even if the observed difference was very small.

(ii) When Phase II is Large For a large Phase II, that is, as $s^2 \to 0$, then

$$pr\left(z > \frac{c-m}{\sqrt{s^2+\sigma^2}}\right) \to pr\left(z > \frac{c-m}{\sqrt{s^2+\sigma^2}}\right) = pr\left(z > z_\alpha - \frac{m}{\theta}(z_\alpha + z_\beta)\right) \tag{4}$$

Hence, equation (4) is the power of the Phase III study under H_1: $\theta_{\text{TRUE}} = m$. This clearly makes sense since the larger the Phase II is, the closer the estimated effect is to truth. If the Phase III is sized with $\theta = m$, the chance of success by assurance is $1 - \beta$, that is, regular power. However, it is unlikely in practice that s^2 would be smaller than σ^2. At best one might expect $s^2 \approx \sigma^2$ so that assurance would be given by $pr(z > z_\beta/\sqrt{2}) = 82\%$ if $\beta = 0.1$.

(iii) When Phase III is Large As $\sigma^2 \to 0$, then

$$pr\left(z > \frac{c-m}{\sqrt{s^2+\sigma^2}}\right) \to pr\left(z > \frac{-m}{s}\right) \tag{5}$$

Hence, equation (5) $= 1 - \{\text{one-sided Phase II } p\text{-value}\}$. Consequently, if the proposed Phase III is very large, the chance of success by assurance is maximally 1 minus the one-sided p-value from Phase II. For example, a Phase II with $p = 0.20$ one-sided means that assurance in Phase III cannot exceed 80% even with infinitely many patients entered. More generally, as $\sigma^2 \to 0$ then $c \to 0$ and the conventional power curve approaches a step function. Then any observed difference > 0 rejects the null with 100% probability, and any difference ≤ 0 rejects with probability zero. Therefore, for *any* prior $f(\theta)$, assurance $= \int_{-\infty}^{+\infty} (power \mid \theta) f(\theta) d\theta \to \int_0^{+\infty} f(\theta) d\theta$, which is the fraction of the prior density that lies to the right of zero.

(iv) If Phase III is *k* Times Larger than Phase II More commonly, $\sigma^2 > 0$ and $s^2 > 0$ with $\sigma^2 \leq s^2$. In this circumstance, if the size of the Phase III is k times that of Phase II then

$$pr\left(z > \frac{c-m}{\sqrt{s^2+\sigma^2}}\right) \to pr\left(z > \frac{-m}{\sqrt{\sigma^2(k+1)}}\right) - pr\left(z > \frac{z_\alpha - \frac{m}{\theta}(z_\alpha + z_\beta)}{\sqrt{k+1}}\right) \tag{6}$$

Table 1 Summary of key features of assurance

		Phase III	
		Small	Large
Phase II	Small	$s^2 \to \infty, s^2 \gg \sigma^2$ Assurance = Phase II one-sided p-value (= 50%)	$\sigma^2 \to 0, s^2 \gg \sigma^2$ Assurance = Phase II one-sided p-value
	Large	$s^2 \to 0, s^2 \ll \sigma^2$ Assurance = Power of Phase III trial at $\theta_{\text{TRUE}} = \text{m}$	$s^2 > 0, \sigma^2 > 0$ with $s^2 = k\sigma^2$ Assurance = $\Phi^{-1}\left(\frac{z_\beta}{\sqrt{k+1}}\right)$ if Phase III sized in region of $\theta_{\text{TRUE}} = \text{m}$

Therefore, if Phase III is sized with θ hypothesized in the region of m, then then PoS by assurance is approximately $pr(z > \frac{z\beta}{\sqrt{k+1}})$. Hence, for Phase III trials 3 to 5 times larger than Phase II, which inthe authors' experience is not uncommon, assurance will routinely be in the region of 70% since equation (6) = 0.74 for $k = 3$, 0.72 for $k = 4$, and 0.70 for $k = 5$. Table 1 summarizes key observations thus far regarding assurance.

3. POWER AND ASSURANCE FOR TWO PHASE III TRIALS

In most new drug applications, the regulatory requirement is for substantial evidence from adequate and well-controlled trials, leading many developments to have at least two Phase III trials. When asked what is the chance that two independent Phase III trials are successful the answer is clearly $(1 - \beta)^2$. However, this is no longer the case for assurance as the incorporation of prior information induces a correlation between the outcomes of the two Phase III's.

To see this, consider

$$E[x, \theta] = \int_{+\infty}^{-\infty}\int_{+\infty}^{-\infty} x\theta f(x, \theta)d\theta dx = \int_{+\infty}^{-\infty} \theta f(\theta)\int_{+\infty}^{-\infty} x f(x \mid \theta)dx d\theta$$
$$= \int_{+\infty}^{-\infty} \theta^2 f(\theta)d\theta = s^2 + m^2$$

Hence,

$$Cov(x, \theta) = E[x, \theta] - E[x]E[\theta] = s^2 + m^2 - m \cdot m = s^2$$

so that

$$\rho(x, \theta) = \sqrt{\frac{s^2}{s^2 + \sigma^2}} \tag{7}$$

In relation to a second Phase III trial, let y represent a sufficient statistic for θ with probability distribution $f(y \mid \theta) \sim N(\theta, \tau^2)$. Then, as shown in Appendix A,

$$\rho(x, y) = \rho(x, \theta)\rho(y, \theta) = \sqrt{\frac{s^2}{s^2 + \sigma^2}}\sqrt{\frac{s^2}{s^2 + \tau^2}} \tag{8}$$

When $\sigma^2 = \tau^2$,

$$\rho(x, y) = \frac{s^2}{s^2 + \sigma^2} \tag{9}$$

Furthermore, as shown in Appendix B and consistent with equation (9), when integrated over the prior for θ outcomes x and y from the two Phase III trials are bivariate normal:

$$\begin{aligned} f(x, y) &= \int_{+\infty}^{-\infty} f(x, y, \theta) d\theta = \int_{+\infty}^{-\infty} f(x \mid \theta) f(y \mid \theta) f(\theta) d\theta \\ &= N\left[\begin{pmatrix} m \\ m \end{pmatrix} \begin{pmatrix} s^2 + \sigma^2 & s^2 \\ s^2 & s^2 + \tau^2 \end{pmatrix} \right] \end{aligned} \tag{10}$$

Therefore equation (10) allows the PoS by assurance associated with two Phase III trials to be calculated.

Further observations are as follows.

(v) $\rho(x, y) \to 1$ As $\sigma^2/s^2 \to 0$ then $\rho(x, y) \to 1$. This occurs when either Phase III is very large relative to Phase II or, equally, when Phase II is very small. In such cases, assurance for the two trials individually and combined are again maximally $1 - \{$one-sided Phase II p-value$\}$.

This means if Phase II delivers a one-sided p-value of 0.2, $2 \times$ Phase III's of any size will each have at most an 80% chance of success by assurance, and, further, the overall chance of success for both by assurance is also 80%.

(vi) $\rho(x, y) \to 0$ If the Phase II is large relative to Phase III such that $s^2 \to 0$, then $\rho(x, y) \to 0$. If this is the case, two Phase III's sized with $m = \theta$ will each have $1 - \beta$ chance of success, and the overall chance of success for both will be $(1 - \beta)^2$. However, as noted earlier, in practice $s^2 \geq \sigma^2$ so that typically $\frac{1}{2} \leq \rho(x, y) \leq 1$.

4. AN EXAMPLE

Consider a randomized Phase II oncology trial with 70 events (corresponding to a design with $\alpha = \beta = 0.2$, hypothesized hazard ratio [HR] of 0.667) and an observed HR of 0.75, $p = 0.11$. Suppose a Phase III trial is planned with a hypothesized HR of 0.75, requiring 508 PFS events for $\alpha = 0.025$ one-sided and power 90%. The assurance via equation (1) is 67%. Hypothesizing an HR of 0.8 in the Phase III and increasing the number of events by more than 80% to 844 increases assurance only by 6%, to 73%. As noted earlier, the assurance cannot exceed $1 - p = 89\%$ even with a vast number of events in Phase III. The Phase II data and power in Phase III are displayed graphically in Fig. 1. Examination of this figure quickly reveals why a large jump in the number of events in Phase III has little impact on assurance. As noted by King (2009), it's generally not the steep Phase III power curve that most drives assurance but rather the spread of the prior based on Phase II where a sizable portion of the density for the HR lies to the right of unity. To meaningfully impact assurance requires either a larger Phase II or a better result, shifting the density for the HR to the left. Also, the maximum assurance attainable

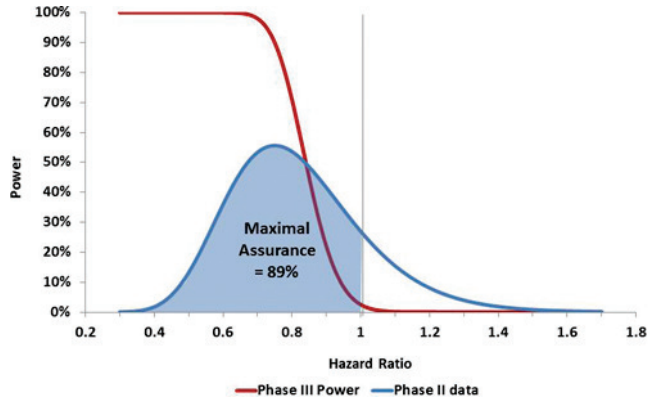


Figure 1 Phase II data, HR = 0.75 on 70 events, and Power in Phase III with 508 events and $\alpha = 0.025$ 1-sided. (Color figure available online.)

is represented by the shaded area shown, which is equal to 1 minus the Phase II one-sided p-value. As Phase III size increases, the power curve becomes ever steeper until, in the limit, it becomes a step function equaling 1 for all hazard ratio values less then unity, and 0 otherwise. Hence, when averaged over the Phase II data prior, assurance is represented by the area of prior density to the left of unity. Finally, recalling via equation (3) that Phase III assurance results from averaging the power curve over the prior distribution for θ, this is more readily seen in Table 2, which provides a rough approximate derivation of assurance. The approximate figure of 67.01% is a very close match to the exact figure of 67%.

Table 2 Approximate value of assurance for a single Phase III and two identical Phase IIIs

	Column 1	Column 2		
Range for θ_{TRUE}	Phase III[a] power $\|\theta_{TRUE}$ = midpoint of range	Pr(θ_{TRUE} in range) based on PII[b] data	Assurance for single PIII 1×2	Assurance for two PIIIs $1^2 \times 2$
0.15–0.25	100.00%	0.00%	0.00%	0.00%
0.25–0.35	100.00%	0.07%	0.07%	0.07%
0.35–0.45	100.00%	1.56%	1.56%	1.56%
0.45–0.55	100.00%	8.09%	8.09%	8.09%
0.55–0.65	99.99%	17.75%	17.75%	17.74%
0.65–0.75	98.03%	22.53%	22.08%	21.65%
0.75–0.85	71.05%	19.97%	14.19%	10.08%
0.85–0.95	21.99%	13.89%	3.05%	0.67%
0.95–1.05	2.50%	8.17%	0.20%	0.01%
1.05–1.15	0.12%	4.28%	0.01%	0.00%
1.15–1.25	0.00%	2.06%	0.00%	0.00%
		Total	67.01%	59.88%

[a]Phase III: 508 events to provide 90% power to test the hypothesis $\theta_{TRUE} = 0.75$ at the 0.025 one-sided α-level.

[b]Phase II: Observed HR = 0.75 on 70 events.

Now suppose a second, duplicate Phase III is planned. Then, in averaging over the Phase II data, outcomes for the two Phase IIIs are highly correlated with $\rho = 0.88$. The probability of success in both Phase IIIs via equations (9) and (10) is 60%. Again, Table 2 provides a rough approximation of assurance for two identical Phase III trials. The approximate figure is 59.88%, as compared to the exact figure of 60%.

5. ASSURANCE SUMMARY

Taken at face value, assurance is an appealing concept in relation to Phase III decision making as it helps to inform regarding the likelihood of success given the data observed in Phase II. With Phase III failure rates continuing to be on the order of 50%, pharmaceutical and biotech companies and their statisticians or statistical advisors are increasingly turning to concepts like assurance in an attempt to improve decision making.

However, when looked at carefully, assurance has several strange and counterintuitive properties that may give the nonstatistician good reason to pause and think. These include:

- PoS = 50% when Phase II is very small regardless of the size of Phase III.
  - While not surprising to the statistician, to the nonstatistician, this says a small and relatively uninformative Phase II will deliver approximately 50% assurance regardless of whether Phase III has 100 or 100,000+ patients.
- PoS = 1 − {one-sided Phase II p-value} when Phase III is large relative to Phase II
  - So if $p = 0.2$ one-sided in Phase II, assurance cannot exceed 80% even if Phase III was huge in the extreme with, say, 100,000+ patients and a conventional power $> 99.99\%$
- The outcomes of two independent Phase III trials are in fact correlated when integrated over the prior Phase II data, with correlation typically $\geq 1/2$.
  - When either the Phase III's are large relative to Phase II, or when Phase II is small relative to Phase III, the correlation between Phase III outcomes is 1 and the assurance for the two trials individually and combined is, again, maximally 1 − {one-sided Phase II p-value}.

In the author's experience, these observations when presented to medical and other nonstatistical colleagues are invariably met with a mix of confusion and disbelief. They understandably lead to concern and wariness in using assurance to help guide Phase III decision making. The confusion boils down to a failure, often on behalf of the statistician, to recognize the fundamental switch that takes place with assurance: the move away from considering the treatment effect as a fixed parameter to assuming it is a random variable; and then a further failure to realize that assurance only really makes sense if the Phase III is analyzed in the same way it was designed—by formally incorporating the Phase II data as a prior and performing a Bayesian analysis.

With perhaps the exception of noninferiority trials using the synthesis method, Phase III trials are predominantly analyzed using frequentist methodology by sponsors and regulators alike. As such, one may fairly ask, if not assurance to guide better Phase III decision making, then what?

6. SUGGESTED PRINCIPLES FOR GOOD PHASE III DECISION MAKING: 10 BASIC STEPS

Before seeking to employ concepts like assurance, there is much that can and should be done in terms of basic good science and common sense. While some of what follows may seem obvious, a cursory glance through the medical literature and published Phase II studies will quickly suggest that not all decision makers, both academic and big pharma alike, appreciate the basics.

Design

Step 1. Ensure Phase II is well controlled, randomized, and double-blinded where possible. If blinding is not possible, consider having the primary endpoint(s) assessed or adjudicated by an independent third party to minimize the potential for bias. Where possible, include an active arm reflective of the comparator anticipated in Phase III. Finally, include more than one dose of drug where possible and ensure the highest dose provides the maximum possible kinetic exposure to minimize type II error.

Step 2. Consider performing two Phase II trials. To support feasibility, use α and β levels consistent with Phase II decision making and not a confirmatory Phase III trial. Positive outcomes from two Phase II's of moderate size are generally more compelling and reliable than a single larger Phase II, especially in settings with softer endpoints like CNS. Any concerns regarding cost and feasibility may be allayed if the following is considered. If a single Phase II is designed with $\theta^2 = \sigma^2(z_\alpha + z_\beta)$, as compared to two Phase IIs each with $\theta^2 = 2\sigma^2(z_\alpha + z_\gamma)$, then $\gamma = 1 - \Phi\{(z_\alpha + z_\beta)/\sqrt{2} - z_\alpha\}$. Therefore, if $\alpha = 0.1$ and $1 - \beta = 0.90(0.80)$, then $1 - \gamma^2 = 0.911(0.829)$, meaning that two Phase II's with 10% significance and $N/2$ patients together carry greater power than a single Phase II with N patients. Similarly, if $\alpha = 0.05$ and $1 - \beta = 0.90(0.80)$, then $1 - \gamma^2 = 0.887(0.793)$ so, again, two Phase II's with $N/2$ patients provide similar power to a single Phase II with N patients.

Step 3. Choose the most sensitive patient population relevant to the purported mechanism of action for efficacy. Failure to do so obviously increases the probability of a type II error. Ensure the patient population also reflects that anticipated in Phase III.

Step 4. Predefine the Phase III go/no-go decision rule for "success" and, importantly, stick to it.

Step 5. In addition to the use of assurance or expected power in sizing Phase III, also present sample size based upon a regular conditional power calculation. This will help Phase III decision makers understand the impact on size, cost, and risk of seemingly minor deviations in the value of the hypothesized treatment effect.

Analysis and Interpretation

Step 6. Analyze efficacy variables on an intention-to-treat (ITT) basis to ensure results are reflective what is expected in Phase III. Do not ignore drop-outs; include them as failures to respond in the analysis.

Step 7. Try to avoid multiple interim analyses in Phase II, particularly when the study is open and/or where results maybe seen by the sponsor and/or investigator at midstream. The potential for serious (if unintentional) bias is obvious, as is the unfortunately tendency for researchers to see what they hope to see. If interims are desired, ensure they are properly planned in advance with clear decision rules and, preferably, governed by a fully independent IDMC. The advantage of the latter is that objective judgements regarding study conduct are made possible, thereby helping to preserve the integrity of the study.

Step 8. Stick to your predefined go/no-go decision rule. If the primary fails the go/no-go, do not look for "signals" elsewhere in data; do not look to salvage by means of extensive (post hoc) analyses in subgroups; do not retrospectively substitute the primary endpoint with some other secondary endpoint; do not seek to substitute primary with a subset of some secondary endpoint; and do not look to retrospectively alter or fudge the predefined go/no-go criteria.

Step 9. Ensure that senior leaders and decision makers are talented, well-experienced drug developers with a proven track record of getting a drug through development to approval with the likes of the Food and Drug Administration (FDA)/European Union (EU)/Japan. Decision makers must have relevant experience and a sound appreciation of good experimental design, data analysis, and interpretation. Excellent, enthusiastic scientists, project leaders, and product champions are not the best decision makers—evangelic belief often supplants a rational, sober assessment of the data and impairs objective judgement.

Step 10. Include an experienced, technically expert statistician in the heart of the decision making process. It is not helpful if the statistician is no longer practicing, having "left the bench early," and consequently possesses little or no contemporary statistical knowledge or technical skill. Sadly, without a highly capable statistician, strange decisions can be, and are, very easily made. While Step 8 may seem obvious to the statistician, these kinds of issues do occur. Making retrospective excuses for failure is sadly not uncommon, especially when under pressure to replenish an ailing pipeline (Arrowsmith, 2011b). The notable learning point here is, few drugs have failed well-designed, well-conducted, properly analyzed and appropriately interpreted Phase II trials only to proceed to Phase III and be positive. But many Phase IIIs have failed on the back of questionable Phase II data, design, and analysis. Examples of convincing Phase III failures on the back of seemingly impressive Phase II data include TC-5214 (*s*-mecamylamine) in depression (Dunbar, 2009; AstraZeneca Plc., 2012), dimebon (latrepirdine) in Alzheimer's disease (Doody et al., 2008; Medivation, Inc, 2010), AGI-1067 (succinobucol) in acute coronary syndromes (AstraZeneca Plc., 2004; Tardif et al., 2008), and iniparib in breast cancer (O'Shaughnessy et al., 2011a,b).

7. CONCLUSIONS

The concept of assurance is appealing in the context of Phase II to Phase III decision making. It makes sense to formally incorporate prior Phase II data into Phase III design. Averaging Phase III power over the distribution of treatment effect observed in Phase II seems like a reasonable approach. However, assurance has several idiosyncrasies and counterintuitive properties that make it difficult for the nonstatistician, and even some statisticians, to trust and understand. Examples of this are Phase III assurance being capped at the 1 minus one-sided p-value observed in Phase II even when Phase III includes an infinite number of patients, and two independent Phase III trial outcomes being correlated when integrated over the Phase II data. A more fruitful approach to solving the high failure rate in Phase III may be a return to the basics of drug development, and good Phase II trial design, conduct, analysis, and interpretation. Statistics and statistical methodology can go a long way to enhancing sound decision making, but not all the way. In the end it boils down to a matter of relying on the experience and sound scientific judgment of the decision makers to make rational, data-driven decisions in the best interests of their organizations, patients and physicians.

One way to assist this from a statistical perspective is to have highly experienced, technically capable statisticians sitting at the decision-making table. In this way it might just be that some poor Phase II to Phase III, decisions are averted resulting in higher success rates, albeit among fewer Phase III programs.

APPENDIX A

$$corr(w_2, w_3) = corr(w_1, w_2)corr(w_1, w_3)$$

Let v_1, v_2 and v_3 be i.i.d N(0,1) random variables. Define

$$\begin{aligned} w_1 &= v_1 \\ w_2 &= \gamma v_1 + v_2\sqrt{1-\gamma^2} \\ w_3 &= \rho v_1 + v_3\sqrt{1-\rho^2} \end{aligned}$$

Then

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \gamma & \rho \\ \gamma & 1 & \gamma\rho \\ \rho & \gamma\rho & 1 \end{bmatrix} \right)$$

APPENDIX B

$$f(x, y) = \int_{-\infty}^{+\infty} f(x, y, \theta)d\theta = \int_{-\infty}^{+\infty} f(x \mid \theta)f(y \mid \theta)f(\theta)d\theta$$

$$f(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{\sqrt{2\pi s^2}}e^{-\frac{x^2}{2\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}e^{-\frac{m^2}{2s^2}} \times \int_{-\infty}^{+\infty} e^{-\frac{\theta^2-2x\theta}{2\sigma^2}}e^{-\frac{\theta^2-2y\theta}{2\sigma^2}}e^{-\frac{\theta^2-2m\theta}{2s^2}}d\theta$$

$$f(x, y) = A\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left\{\frac{\theta^2(2s^2+\sigma^2)-2\theta(xs^2+ys^2+m\sigma^2)}{\sigma^2 s^2}\right\}}d\theta$$

where

$$A = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{\sqrt{2\pi s^2}}e^{-\frac{x^2}{2\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}e^{-\frac{m^2}{2s^2}}$$

$$f(x,y) = A\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\frac{(2s^2+\sigma^2)}{\sigma^2 s^2}\left\{\left(\theta-\frac{xs^2+ys^2+m\sigma^2}{2s^2+\sigma^2}\right)^2-\left(\frac{xs^2+ys^2+m\sigma^2}{2s^2+\sigma^2}\right)^2\right\}}d\theta$$

$$f(x,y) = Ae^{\frac{1}{2}\frac{(2s^2+\sigma^2)}{\sigma^2 s^2}\left(\frac{xs^2+ys^2+m\sigma^2}{2s^2+\sigma^2}\right)^2}\sqrt{2\pi\frac{s^2+\sigma^2}{2s^2+\sigma^2}}$$

$$f(x,y) = \frac{1}{2\pi(s^2+\sigma^2)\sqrt{1-\rho^2}}e^{\frac{1}{2}\frac{(2s^2+\sigma^2)}{\sigma^2 s^2}\left(\frac{xs^2+ys^2+m\sigma^2}{2s^2+\sigma^2}\right)^2}e^{-\frac{x^2}{2\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}e^{-\frac{m^2}{2s^2}}$$

where

$$\rho = \frac{s^2}{s^2+\sigma^2}$$

$$f(x,y) = \frac{1}{2\pi(s^2+\sigma^2)\sqrt{1-\rho^2}}e^{-\frac{1}{2\sigma^2 s^2}\left\{\frac{(xs^2+ys^2+m\sigma^2)^2}{2s^2+\sigma^2}-x^2s^2-y^2s^2-m^2\sigma^2\right\}}$$

$$f(x,y) = Be^{-\frac{1}{2}\left[\frac{(s^2+\sigma^2)^2}{\sigma^2(2s^2+\sigma^2)}\right]\left\{\frac{(x-m)^2}{s^2-\sigma^2}+\frac{(y-m)^2}{s^2+\sigma^2}-\frac{2s^2(x-m)(y-m)}{(s^2+\sigma^2)^2}\right\}}$$

where

$$B = \frac{1}{2\pi(s^2+\sigma^2)\sqrt{1-\rho^2}}$$

$$f(x,y) = \frac{1}{2\pi(s^2+\sigma^2)\sqrt{1-\rho^2}}e^{-\frac{1}{2}\left[\frac{1}{1-\rho^2}\right]\left\{\frac{(x-m)^2}{s^2-\sigma^2}+\frac{(y-m)^2}{s^2+\sigma^2}-2\rho\frac{2s^2(x-m)(y-m)}{s^2+\sigma^2}\right\}}$$

Hence

$$f(x,y) \sim N\left[\begin{pmatrix} m \\ m \end{pmatrix}\begin{pmatrix} s^2+\sigma^2 & s^2 \\ s^2 & s^2+\sigma^2 \end{pmatrix}\right]$$

REFERENCES

Arrowsmith, J. (2011a). Phase II failures: 2008–2010. *Nature Reviews Drug Discovery* 10: 328–329.

Arrowsmith, J. (2011b). Phase III and submission failures: 2007–2010. *Nature Reviews Drug Discovery* 10:87–88.

AstraZeneca Plc. (2012). Remaining TC-5214 Phase III Efficacy Studies Do Not Meet Endpoint, Regulatory Filing Will Not Be Pursued. March 20. Available at: http://www.astrazeneca.com/Media/Press-releases/Article/20032012tc5214-failed-Phase-iii-endpoint

AtheroGenics, Inc. (2004). AtheroGenics Reports Positive Final Results from CART-2 Clinical Trial of AGI-1067. AGI-1067 Achieves Statistically Significant Plaque Regression Versus Baseline. November 22. Available at: http://www.sec.gov/Archives/edgar/data/1107601/000110760104000056/exhibit99 prfinalresults.htm

Chuang-Stein, C. (2006). Sample size and the probability of a successful trial. *Pharmaceutical Statististics* 5:30–309.

Chuang-Stein, C., Kirby, S., French, F., Kowalski, K., Marshall, S., Smith, M. K., Bycott, P., Beltangady, M. (2011). A quantitative approach for making go/no go decisions in drug development. *Drug Information Journal* 45:187–202.

Doody, R. S., Gavrilova, S. I., Sano, M., Thomas, R. G., Aisen, P. S., Bachurin, S. O., Seely, L., Hung, D. (2008). Effect of dimebon on cognition, activities of daily living, behaviour, and global function in patients with mild-to-moderate Alzheimer's disease: A randomised, double-blind, placebo-controlled study. *Lancet* 372:207–215.

Dunbar, G. (2009). Positive effects of the nicotinic channel blocker TC-5214 as augmentation treatment in patients with major depressive disorder who are inadequate responders to a first-line SSRI. nAChR2009 satellite meeting of the 39th annual meeting of the Society for Neuroscience; 2009. Slides available at http://www.faqs.org/sec-filings/091016/TARGACEPT-INC_8-K/dex991.htm

King, M. (2009). Evaluating probability of success in oncology clinical trials. BASS XVI, November. Available at: http://www.bassconference.org/PDFs/Bass%202009%20Martin%20King.pdf

Kirby, S., Burke, J., Chuang-Stein, C., Sin, C. (2012). Discounting phase 2 results when planning phase 3 clinical trials. *Pharmaceutical Statistics* 11:373–385.

Medivation, Inc. (2010). Pfizer And Medivation Announce Results From Two Phase 3 Studies in Dimebon (latrepirdine∗) Alzheimer's Disease Clinical Development Program. March 3. Available at: http://investors.medivation.com/releasedetail.cfm?ReleaseID=448818

O'Hagan, A., Stevens, J. W., Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics* 4:18–201.

O'Shaughnessy, J., Osborne, C., Pippen, J. E., Yoffe, M., Patt, D., Rocha, C., Koo, I. C., Sherman, B. M., Bradley, C. (2011a). Iniparib plus chemotherapy in metastatic triple-negative breast cancer. *New England Journal of Medicine* 364:205–214.

O'Shaughnessy, J., Schwartzberg, L. S., Danso, M. A., Rugo, H. S., Miller, K., Yardley, D. A., Carlson, R. W., Finn, R. S., Charpentier, E., Freese, M., Gupta, S., Blackwood-Chirchir, S., Winer, E. P. (2011b). A randomized phase III study of iniparib (BSI-201) in combination with gemcitabine/carboplatin (G/C) in metastatic triple-negative breast cancer (TNBC). *Journal of Clinical Oncology* 29(suppl.):abstract 1007.

Prinz, F., Schlange, T., Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Review Drug Discovery* 10:712–713.

Tardif, J.-C., McMurray, J. J., Klug, E., Small, R., Schumi, J., Choi, J., Cooper, J., Scott, R., Lewis, E. F., L'Allier, P. L., Pfeffer, M. A. (2008). Effects of succinobucol (AGI-1067) after an acute coronary syndrome: A randomised, double-blind, placebo-controlled trial. *Lancet* 371: 1761–1768.

9. **Carroll K** and Milsted R. Barriers to clinical development in oncology: The impact of new thinking around non-inferiority. 2004, Journal of Clinical Oncology, ASCO Annual Meeting Proceedings (Post-Meeting Edition). Vol 22, No 14S (July 15 Supplement), 2004: 6082.

Journal of Clinical Oncology, 2004 ASCO Annual Meeting Proceedings (Post-Meeting Edition).
Vol 22, No 14S (July 15 Supplement), 2004: 6082
© 2004 American Society of Clinical Oncology

Abstract

Barriers to clinical development in oncology: The impact of new thinking around non-inferiority

K. Carroll and R. Milsted

AstraZeneca, Macclesfield, United Kingdom

6082

Background: While it is hoped that novel, biologically targeted anticancer agents will offer significant advantages over standard therapies in terms of improved tolerability, they may not always demonstrate increased efficacy. Therefore, active-control, non-inferiority trials to compare the new agent with a standard agent are likely to be necessary, the conventional aim being to show no clinically relevant loss of efficacy. **Methods:** Rothmann et al (Stat Med 2003; 22: 239–64) recently described non-inferiority sample size and data analysis methods, which are increasingly being used by regulators in the USA and Europe. We illustrate the impact of Rothmann's approach on trial size and, thus, development of novel anticancer agents. **Results:** For example, if a standard treatment was previously shown to double survival (eg hazard ratio=0.5, p=0.02), and the goal for a new, better-tolerated therapy is to retain at least half of this effect, a conventional sample size calculation shows that a total of 350 deaths is required to provide 90% power at the 1-sided, 2.5% significance level. Applying Rothmann's method increases the number of deaths required, and so sample size, by almost 9-fold to 3082 deaths, which would result in impractical trials in many settings. **Conclusions:** As shown, use of Rothmann's method has enormous consequences. We suggest the first purpose of non-inferiority trials should be to prove that a new agent would have been better than placebo, had placebo been included. The second purpose should be to estimate (indirectly) the size of the effect of the new agent relative to placebo. Both aims are achievable with Rothmann's approach, with some small modifications. This philosophy does not require prespecification of a percentage effect retention (although the result can be displayed as the likely fraction of standard effect retained) and concentrates on estimating the degree of benefit over placebo. Fisher et al (Am Heart J 2001; 141: 26–32) describe a related approach in patients at risk of ischaemic events. This approach focuses on absolute rather than relative efficacy, and, in future, may be a more appropriate model to apply to trials of oncology agents.

Publication 9: Carroll, Milsted 2004

BARRIERS TO CLINICAL DEVELOPMENT IN ONCOLOGY: THE IMPACT OF NEW THINKING AROUND NON-INFERIORITY

K CARROLL, B MILSTED

ASTRAZENECA, MACCLESFIELD, UK

6082

ABSTRACT

Background: While it is hoped that novel, biologically targeted anticancer agents will offer significant advantages over standard therapies in terms of improved tolerability, they may not always demonstrate increased efficacy. Therefore, active-control, non-inferiority trials to compare the new agent with a standard agent are likely to be necessary, the conventional aim being to show no clinically relevant loss of efficacy.

Methods: Rothmann et al (Stat Med 2003; 22: 239-64) recently described non-inferiority sample size and data analysis methods, which are increasingly being used by regulators in drug evaluation. We illustrate the impact of Rothmann's approach on trial size and, thus, development of novel anticancer agents.

Results: For example, if a standard treatment was previously shown to double survival (eg hazard ratio=0.5, p=0.02), and the goal for a new, better-tolerated therapy is to retain at least half of this effect, a conventional sample size calculation shows that a total of 350 deaths is required to provide 90% power at the 1-sided, 2.5% significance level. Applying Rothmann's method increases the number of deaths required, and so sample size, by almost 9-fold to 3082 deaths, which would result in impractical trials in many settings.

Conclusions: As shown, use of Rothmann's method has considerable consequences. We suggest the first purpose of non-inferiority trials should be to prove that a new agent would have been better than placebo, had placebo been included. The second purpose should be to estimate (indirectly) the size of the effect of the new agent relative to placebo. Both aims are achievable with Rothmann's approach, with some small modifications. This philosophy does not require prespecification of a percentage effect retention (although the result can be displayed as the likely fraction of standard effect retained) and concentrates on estimating the degree of benefit over placebo. Fisher et al (Am Heart J 2001; 141: 26-32) describe a related approach in patients at risk of ischaemic events. This approach, which is in line with recent European regulatory guidelines on non-inferiority trials, focuses on absolute rather then relative efficacy, and, in future, may be a more appropriate model to apply to trials of oncology agents.

INTRODUCTION

- While novel, well-tolerated, biologically targeted (cytostatic) anticancer agents may offer significant tolerability advantages in comparison with standard cytotoxic agents, they may not demonstrate increased efficacy.
- When a direct comparison with placebo is impossible or unethical, indirect comparisons with historical data using active-control, 'non-inferiority' trials are necessary in the evaluation of novel agents.
- One impact of using non-inferiority trials is that the number of events needed to show non-inferiority is often more than in the original placebo-controlled trial. Recent publications on non-inferiority trial methodology have an important impact on trial size and feasibility.[1-4] This poster will address these issues.

CONVENTIONAL METHOD

- Using data from earlier trials that established the efficacy of a standard treatment (standard effect) relative to placebo, non-inferiority trials have traditionally aimed to demonstrate that a new treatment suffers no clinically relevant loss of efficacy.
- In drug evaluation and approval, 'no clinically relevant loss' is often translated into the goal of proving that ≥50% of the standard effect is retained by the new drug. For example, if a standard treatment had previously doubled survival (hazard ratio [HR] = 0.5, p=0.02) and the goal for the new, better tolerated therapy is to retain at least half of this effect (HR=0.5), using conventional methodology 350 deaths would be required to provide 90% power at the 1-sided, 2.5% significance level.
- However, the effectiveness of the standard treatment (when compared with previous trials) is not known with complete certainty and, as the conventional method ignores any uncertainty around the standard effect, there is an increased chance of approving an inferior drug.

RECENT METHODOLOGY

- A recently published methodology by Rothmann *et al* [1] explicitly allows for uncertainty around the standard effect, reflected by the standard error and p value of the standard effect.
- Rather than using the point estimate for the standard effect, this methodology uses a lesser standard effect that is somewhere between the point estimate and its lower 97.5% confidence interval, thereby managing the regulatory risk by reducing the chance of approving an inferior drug to exactly 2.5%.
- The impact of this methodology on trial size can be seen using the hypothetical scenario mentioned above (HR=0.5, p=0.02) [Table 1], where the number of deaths increases to 3082. Indeed, in some situations this methodology may actually demand more events than there are patients with the disease.

Table 1. The number of deaths required using recent methodology[1] in a hypothetical scenario[a]

Historical p value for standard vs placebo	Deaths required to demonstrate non-inferiority using recent methodology[1]
0.049	3,000,000
0.02	3082
0.01	1563
0.001	735
0.0001	572
0.00001	505
0.000001	459
<0.0000001[b]	350

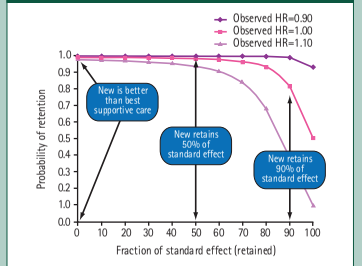
[a]HR for standard, 0.50; trial with 90% power at the 1-sided, 2.5% significance level
[b]Equivalent to the conventional approach, ie the standard effect is known with (almost) complete certainty

- Therefore, while the design of non-inferiority trials needs to take into account the uncertainty of the standard effect, they also need to be clinically feasible in so far as patient numbers are concerned.

ALTERNATIVE APPROACH

- Alternative approaches that lie between conventional methodology and the Rothmann methodology described above may provide a way forward.
- An alternative approach that focuses on the need to establish absolute efficacy:
  - calculates the number of deaths required to indirectly establish that the new treatment is better than placebo with 90% power at the 2.5% significance level
  - uses Rothmann[1] to take into account the uncertainty of standard effect and in doing so makes no assumptions over those described
  - avoids the need for/use of an arbitrary prespecified non-inferiority limit.
- By incorporating these modifications, the alternative approach avoids clinically infeasible trials, eg using the scenario mentioned above (HR=0.5, p=0.02) [Table 1], 770 patients would be required as opposed to the 3082 required by Rothmann.[1]
- In addition, using this approach enables the benefit of the new treatment to be gauged on a continuum rather than in a simplistic dichotomy of 'success' and 'failure.' Results are shown in a natural hierarchy using an 'effect retention likelihood' plot, which quantifies the probability that (i) the new drug is effective (ie would have beaten placebo) and (ii) that the new drug retains a given fraction (≥0% to 100%) of the standard effect (Figure 1).

Figure 1. The effect retention likelihood for a trial conducted with 770 events[a]



[a]Based on a hypothetical standard effect of HR 2.0, p=0.02, with 90% power at the 1-sided 2.5% significance level

- Using an effect retention likelihood plot means that data are much more informative with regard to understanding the efficacy of the new drug in relation to both a putative placebo and the standard treatment. This is highlighted in Table 2 using data from Figure 1.

Table 2. Effect retention likelihood for a trial conducted with 770 events[a]

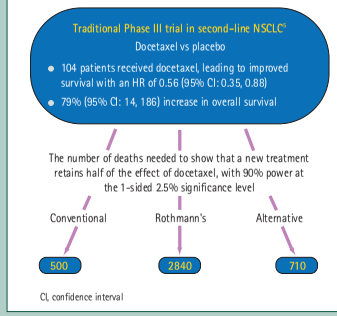
Fraction of standard effect retained, %	Probability, % (HR=1.00)
0[b]	99
50[c]	98
90[d]	81

[a]Historical control effect HR=2.0, p=0.02; active-control trial design with 90% power at the 1-sided 2.5% significance level; data derived from Figure 1
[b]New is better than best supportive care
[c]New retains 50% of standard effect
[d]New retains 90% of standard effect

DATA COMPARISON OF APPROACHES

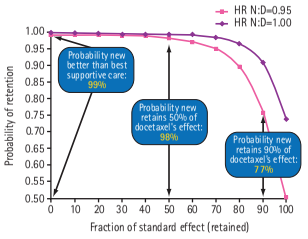
- The effect of the different approaches described above on patient numbers is shown in Figure 2, using data from a Phase III clinical trial of docetaxel in 104 patients with locally advanced or metastatic non-small-cell lung cancer who had failed prior platinum-based chemotherapy:[5]
  - recent methodology[1] increases the size of the non-inferiority trial 5.7-fold when compared with the conventional method
  - the alternative approach increases the size of the non-inferiority trial 2.2-fold when compared with the conventional method.

Figure 2. Differences in patient number using non-inferiority trials



- Therefore, the impact of the alternative approach on patient numbers is much lower when compared with recent methodology.[1]
- The effect retention likelihood plot that would be generated for the evaluation of a new drug compared with the standard drug (in this case, docetaxel) using 710 events is shown in Figure 3.

Figure 3. The effect retention likelihood for a trial conducted with 710 events[a]



[a]Based on the standard effect of docetaxel described in the Food and Drug Administration labels[6]
N:D, new:docetaxel

CONCLUSIONS

- Non-inferiority trials are needed for the evaluation of new drugs in situations where direct comparisons with placebo are impossible or unethical.
- The conventional method has a low impact on trial size but does not take into account the uncertainty around the standard effect.
- Rothmann[1] accounts for uncertainty around the standard effect but also has a considerable impact on oncological drug development.
- An alternative approach intermediate between the conventional and recent methodologies[1] accounts for uncertainty around the standard effect but also focuses on demonstrating absolute rather than relative benefit.
- In doing so, patient numbers are reduced, thereby lessening the impact on drug development.
- Importantly, this alternative way of thinking is consistent with recent European regulatory guidelines on non-inferiority trials.[6]
- The debate (scientific and statistical) on how best to draw inferences from active-control, non-inferiority trials is ongoing.

References

1. Rothmann M, et al. Stat Med 2003; 22: 239-264.
2. Hung HMJ, et al. Stat Med 2003; 22: 213-225.
3. Wang S-J, Hung HMJ. Control Clin Trials 2003; 24: 147-155.
4. Wang S-J, Hung HMJ. Stat Med 2003; 22: 227-238.
5. Docetaxel ('Taxotere') label. FDA 2001. Available at: http://www.fda.gov/cder/approval/index.htm
6. Committee for Proprietary Medicinal Products. EMEA 2004. Available at: http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf

10. **Carroll K**, Milsted B and Lewis JA. Design and analysis of non-inferiority mortality trials in oncology. Letter to the Editor. Statistics in Medicine 2004, Vol 23(17): 2771-2774.

STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:2771–2778 (DOI: 10.1002/sim.1897)

LETTER TO THE EDITOR

Design and analysis of non-inferiority mortality trials in oncology

by M. Rothman, N. Li, G. Chen and G.Y.H. Chi, *Statistics in Medicine* 2003; **22**:239–264

We would like to draw your attention to the implications for oncologic drug development of the article by Rothmann *et al*. published in the January 2003 special edition of SIM on non-inferiority trials [1]. The methods described in this article are increasingly used by regulators in the U.S.A. and Europe to evaluate the design and analysis of trials of new agents. The consequences for trial size are enormous. Shlaes and Moellering have expressed closely related concerns for anti-infective drug development [2].

There has been something of a paradigm shift in the approach to cancer treatment over recent years. Academia and industry alike are now fully engaged in the discovery, research and development of novel, well tolerated, biologically targeted (cytostatic) anticancer agents. It is hoped that these new treatments will offer significant advantages to patients in terms of improved tolerability, but they may not always demonstrate increased efficacy. This naturally leads to the use of active-control, non-inferiority trials to compare the new agent with a standard agent, the conventional aim being to show no clinically relevant loss of efficacy.

Such trials are often designed to demonstrate that the new treatment retains some fraction of the established effect of the standard, say at least 1/2. Note that this fraction is essentially arbitrary and no regulatory guidance currently mandates this as the minimum amount either to demonstrate clinical non-inferiority or to secure regulatory approval. If the standard treatment was previously shown to double survival in a particular disease setting (hazard ratio $= 0.50$, $p = 0.02$, say), and the goal for a new, better tolerated therapy is to retain at least 1/2 of this effect, a routine sample size calculation shows that a total of 350 events is required to provide 90 per cent power at the one-sided, 2.5 per cent significance level.

There are several important issues associated with the design and analysis of non-inferiority trials, including 'constancy'—the extent to which the standard treatment performs as it did in previous trials—and 'assay sensitivity'—the ability of a non-inferiority trial to detect a real difference between the treatments compared. Much has been published in this area. The regulatory guidelines ICH E9 and E10 describe the issues in detail and provide some general guidance with respect to trial design and conduct [3, 4].

An issue not addressed in these guidelines arises from the fact that the standard effect is an estimate from earlier work and so is not known with certainty. Sample size calculations often ignore this uncertainty. Hung *et al*. have shown that this approach increases the probability of erroneously accepting the efficacy of a truly inferior drug [5].

The approach offered by Rothmann tackles this issue. Assuming constancy of the effect of the standard and accepting assay sensitivity, Rothmann proposes a formal statistical com-

Copyright © 2004 John Wiley & Sons, Ltd.

parison between the historical data characterising the standard effect and the data arising in the non-inferiority trial, thereby explicitly incorporating the uncertainty (i.e. SE) around the standard effect estimate. This is in fact akin to the putative or virtual placebo comparison approach described by Wang *et al*. [6]. Operationally, Rothmann's approach is equivalent to conventional methodology using *not* the point estimate for the standard effect, but, rather, a lesser effect somewhere between the point estimate and its lower 97.5 per cent confidence limit. This reduced effect is chosen so that the chance of falsely approving an inferior drug is exactly 2.5 per cent, thereby managing the regulatory risk.

The key problem for researchers, physicians and patients alike is that, with Rothmann's approach, there is a dramatic increase in the size of the trial required, often rendering the trial completely infeasible. Applying this methodology to the example above increases the number of events required from 350 to 3082, a near nine-fold increase. This size of the increase derives from a combination of the (arbitrary) effect retention fraction (50 per cent in this example) and the strength of prior characterization of the standard effect, which is reflected in the (historical) p-value. As illustrated in the table below, the application of this methodology may actually require more events than there are patients with the disease Table I:

This serves to illustrate that even with a highly significant standard effect estimate, $p \sim 0.001$ say, Rothmann's approach can double the size of a non-inferiority trial. Importantly, if the standard treatment has only just reached statistical significance, this approach implies that no new drug can ever be approved via the non-inferiority route in that group of patients on the basis of clinical benefits other than efficacy; superiority in efficacy to the standard treatment would have to be shown.

Thus, the use of Rothmann's approach, coupled with the arbitrary 50 per cent effect retention requirement, would result in impracticable trials in many settings, notably those where the standard was approved on the basis of a relatively small evidence base. Assuming that direct comparisons with placebo are unethical, we are forced to contemplate non-inferiority trials too large ever to be mounted, effectively removing non-inferiority as a viable tool in the evaluation and ultimate approval of new cancer medicines. This outcome is identical to the one faced by those developing anti-infective agents which has contributed toward a decline

Table I. The number of deaths required to prove a new treatment retains 1/2 of the effect of standard treatment (HR = 0.50) using Rothmann's methodology (true HR new:standard is unity, 90 per cent power, α 2.5 per cent one-sided).

(Historical) p-value for standard vs placebo	Upper 95 per cent CI for HR of new-to-standard must be less than:	Approx No. deaths required to prove 50 per cent retention
0.049	1.004	3 000 000
0.02	1.12	3082
0.01	1.18	1563
0.001	1.27	735
0.0001	1.31	572
0.00001	1.33	505
0.000001	1.35	459
$\ll$0.000001*	1.41	350

*Equivalent to the conventional approach i.e. the standard effect is known with (virtually) complete certainty.

Copyright © 2004 John Wiley & Sons, Ltd.

in the number of companies investing in antibacterial research and development and, consequently, in the number of new antibiotics to treat serious infections [2]. Some fundamental re-thinking in this area is called for to avoid the obvious adverse impact on the future development of new cancer medicines.

One way forward is to argue that there should be no difference between the standards of evidence required in a superiority setting and in a non-inferiority (or active-control) setting. Hence the first purpose of the non-inferiority trials in question should be to prove that a new agent would have been better than placebo if placebo had been included. The second purpose should be to estimate (indirectly) the size of the effect of the new agent relative to placebo. Both of these aims are achievable with Rothmanns' approach and assumptions, with some small modifications. In well-conducted trials with hard endpoints, little non-compliance and complete follow up, there should be no need to require 50 per cent retention of effect to demonstrate superiority to placebo. When estimating the size of the effect, attention could focus on the point estimate in the usual manner, and not on the lower confidence limit alone. An approach along these lines has been nicely illustrated by Fisher *et al*. [7] and does not require pre-specification of a percentage effect retention though, having obtained the data, the result can easily be displayed in relation to the likely fraction of the standard effect retained. Concentrating on estimating the degree of benefit over placebo, albeit through indirect measures, seems more in line with the efficacy standards required by U.S. and European law, both of which call for substantial evidence of efficacy to be established, with no requirement on relative efficacy with respect to existing agents. This approach is in fact consistent with the recently issued draft CPMP guidance on non-inferiority trials [8].

The scientific and statistical debate on how best to draw inferences from active-control, non-inferiority trials should not be considered complete. Rothmann's approach serves to highlight that considerable statistical, methodological and philosophical issues remain. Failure to consider these issues constructively will, at the very least, lead to ever increasing drug development times and, thus, delay the availability of new therapeutic options to patients with life-threatening diseases. At worst, the barriers posed will discourage drug development where it otherwise might have been feasible and so prevent potentially useful new medicines becoming available to patients. We sincerely hope that the scientific community together with regulatory bodies worldwide will give this important area further careful thought.

REFERENCES

1. Rothmann M, Li N, Chen G, Chi GYH, Temple R, Tsou H-H. Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine* 2003; **22**:239–264.
2. Shlaes DM, Moellering RC. The United States Food and Drug Administration and the end of antibiotics. *Clinical Infectious Diseases* 2002; **34**:420–422.
3. ICH Topic E9. *Note for Guidance on Statistical Principles for Clinical Trials*. ICH Technical Coordination, EMEA: London, 1998.
4. ICH Topic E10. *Note for Guidance on Choice of Control Group for Clinical Trials*. ICH Technical Coordination, EMEA: London, 2000.
5. Hung J, Wang S-J, Tsong Y, Lawrence J, O'Neil RT. Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**:213–225.
6. Wang S-J, Hung J, Tsong Y. Utility and pitfalls of some statistical methods in active controlled clinical trials. *Controlled Clinical Trials* 2002; **23**:15–28.

Copyright © 2004 John Wiley & Sons, Ltd.

7. Fisher LD, Gent M, Büller HR. Active-control trials: how would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin, and placebo. *American Heart Journal* 2001; **141**:26–32.
8. Committee for Propriety Medicinal Products: '*Points to consider on the choice of non-inferiority margin*'. EMEA. London, 2004, CPMP/EWP/2158/99draft

From: *Kevin Carroll*[1], *Bob Milsted*[2] *and John A. Lewis*[3]
[1]*Statistical Expert for Global Oncology*
[2]*VP, Regulatory Director for Global Oncology, AstraZeneca Pharmaceuticals Alderley Park, Macclesfield, U.K.*
[3]*Visiting Professor University of Leicester, U.K. and Consultant to AstraZeneca*

Author's Reply[†,‡]

Sir—I would like to thank you for the opportunity to reply in writing to comments of Mr Carroll, Dr Milsted and Professor Lewis [1]. This reply will respond to the premise of the Carroll *et al.* commentary that it is difficult to design a non-inferiority trial based on 50 per cent retention of the survival effect of a standard therapy using the methodology in Rothmann *et al.* [2] when the estimate of the standard therapy vs placebo survival hazard ratio is 0.5, discuss powering a non-inferiority trial for survival, discuss the adequacy of comparing survival between a test therapy and placebo when a standard therapy has been approved for survival, and elaborate on the content of some of the papers referenced in Reference [1].

With respect to the article by Shlaes and Moellering [3], I invite the readers to read the comments by Powers *et al.* [4], the comments by Gilbert *et al.* [5], the reply by Shlaes [6], and the transcript of the February 19–20, 2002 Division of Anti-Infective Drug Products Advisory Committee meeting [7], which Dr Shlaes was a participant. Much of the commentary of both References [5, 6] centers on the discussions of that advisory committee meeting and how these discussions clarified concerns given previously in Shlaes and Moellering. Dr Shlaes states in their reply 'The FDA should be congratulated for organizing a very informative and extremely useful meeting. At this meeting, the FDA succeeded in defining a number of issues and in achieving some early consensus for several of these issues.' Dr Shlaes then lists some of the highlights of the discussions.

According to Carroll *et al.* it is too difficult to design a non-inferiority trial when the estimated standard therapy vs placebo survival hazard ratio is 0.5. Examples that they provide attribute very little precision to such an estimate and less precision than required for a regulatory approval based on survival. An example Carroll *et al.* bring up twice in text would have this standard therapy vs placebo survival hazard ratio estimate of 0.5 based on 45 events (p-value = 0.02; 45 events based on a one-to-one randomization). Carroll *et al.* discuss and apparently prefer that the estimate of 0.5 based on 45 events from one trial should be treated with complete certainty as the true theoretical value of the standard therapy vs placebo

[†]The views expressed in this article do not necessarily represent those of the U.S. Food and Drug Administration.
[‡]This article is a U.S. Government work and is in the public domain in the U.S.A.

11. **Carroll KJ**. Active-controlled, non-inferiority trials in oncology: arbitrary limits, infeasible sample sizes and uninformative data analysis. Is there another way? Pharmaceutical Statistics, 2006; Vol 5(4): 283-293.

PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2006; **5**: 283–293
Published online 18 May 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.218

WILEY InterScience®
DISCOVER SOMETHING GREAT

Active-controlled, non-inferiority trials in oncology: arbitrary limits, infeasible sample sizes and uninformative data analysis. Is there another way?

MAIN PAPER

Kevin J. Carroll*,†
AstraZeneca Pharmaceuticals, Global Clinical Information Science, Alderley Park, Macclesfield, UK

In oncology, it may not always be possible to evaluate the efficacy of new medicines in placebo-controlled trials. Furthermore, while some newer, biologically targeted anti-cancer treatments may be expected to deliver therapeutic benefit in terms of better tolerability or improved symptom control, they may not always be expected to provide increased efficacy relative to existing therapies. This naturally leads to the use of active-control, non-inferiority trials to evaluate such treatments. In recent evaluations of anti-cancer treatments, the non-inferiority margin has often been defined in terms of demonstrating that at least 50% of the active control effect has been retained by the new drug using methods such as those described by Rothmann et al., Statistics in Medicine 2003; 22:239–264 and Wang and Hung Controlled Clinical Trials 2003; 24:147–155. However, this approach can lead to prohibitively large clinical trials and results in a tendency to dichotomize trial outcome as either 'success' or 'failure' and thus oversimplifies interpretation. With relatively modest modification, these methods can be used to define a stepwise approach to design and analysis. In the first design step, the trial is sized to show indirectly that the new drug would have beaten placebo; in the second analysis step, the probability that the new drug is superior to placebo is assessed and, if sufficiently high in the third and final step, the relative efficacy of the new drug to control is assessed on a continuum of effect retention via an 'effect retention likelihood plot'. This stepwise approach is likely to provide a more complete assessment of relative efficacy so that the value of new treatments can be better judged. Copyright © 2006 John Wiley & Sons, Ltd.

Keywords: *active-control trials; non-inferiority; oncology; percent preservation; effect retention likelihood*

*Corresponding to: Kevin J. Carroll, AstraZeneca Pharmaceuticals, Global Clinical Information Science, Alderley Park, Macclesfield, UK.
†E-mail: kevin.carroll2@astrazeneca.com

Copyright © 2006 John Wiley & Sons, Ltd.

Pharmaceutical STATISTICS

1. INTRODUCTION

The design and analysis of active-controlled, non-inferiority trials continues to be a topic of discussion and methodological development in the literature [1,2] with recent issues of both *Statistics in Medicine* and *Biopharmaceutical Journal* being largely devoted to papers relating to non-inferiority [3,4]. Articles by Hung *et al.* and Wang and Hung and, most notably, by Rothmann *et al* deal with non-inferiority by requiring that at least a predefined percentage of the established control effect is retained by a new drug [1,2,5–8]. The effect preservation or retention method described by these authors explicitly incorporates the uncertainty around the estimate of the historical control effect and, in doing so, maintains the one-sided type I error rate, α, at 2.5%. This is achieved under the assumption of 'constancy', that is assuming the control treatment performs as it did in previous trials, an assumption which is implicit in all active-controlled trials. This approach is less conservative than the use of the lower 95% confidence limit of the point estimate as the historical control effect, but, nevertheless, can still lead clinical trials of an infeasible size [5,9]. The consequences for oncologic drug development can be considerable. Shlaes and Moellering [10] have expressed closely related concerns for anti-infective drug development.

The purpose of this article is to briefly review the percent effect retention method and, in doing so, to suggest an alternative philosophy to sizing active-controlled, non-inferiority trials and also offer a simple yet informative way of analysing and presenting the resulting data, so that better informed decisions can be made regarding the efficacy of a new drug relative to the active control.

2. SIZING AN ACTIVE-CONTROLLED TRIAL

Suppose an active-controlled trial is proposed to compare a new drug (T) with an active control drug (C) in terms of overall survival. It is assumed that effectiveness of C has been previously established in a placebo (P) controlled clinical trial(s). The estimate of the effect of C is captured in the hazard ratio, HR[P:C], with an associated standard error, SE $\ln\{\mathrm{HR}[P{:}C]\}$. The aim, as typically stated in such trials, is to assess whether T is non-inferior to C and, in line with Jones *et al.* and recent examples in oncology, the hypothesis to be tested is that T retains at least $\frac{1}{2}$ of the active control effect [11–13]. Thus, the null and alternative hypotheses are

$$H_0 : \ln \mathrm{HR}(T:C) \geqslant \tfrac{1}{2} \ln \mathrm{HR}(P:C) \quad \textit{vs.}$$
$$H_1 : \ln \mathrm{HR}(T:C) < \tfrac{1}{2} \ln \mathrm{HR}(P:C)$$

More generally, the fraction of the control effect to be retained, δ, say, can be defined as per Rothmann *et al.* [1] as

$$\delta = 1 - \frac{\ln \mathrm{HR}(T:C)}{\ln \mathrm{HR}(P:C)}$$

so that the null and alternative hypotheses become

$$H_0 : \ln \mathrm{HR}(T:C) \geqslant (1-\delta) \ln \mathrm{HR}(P:C) \quad \textit{vs.}$$
$$H_1 : \ln \mathrm{HR}(T:C) < (1-\delta) \ln \mathrm{HR}(P:C) \qquad (1)$$

Suppose C was previously shown to significantly increase survival in a particular disease setting with $\mathrm{HR}(P:C) = 1.5$, $p = 0.005$, say. It is desired to demonstrate that the new treatment T retains $\frac{1}{2}$ of the control effect. Assuming T and C are truly equal in effectiveness, the number of deaths required in the trial of T compared to C to achieve this aim can be derived in a number of ways:

(i) If the uncertainty in the estimate of the control effect is ignored, a routine sample size calculation shows that a total of 1023 deaths is required to provide 90% power at the one-sided, 2.5% significance level [14]. Non-inferiority will be concluded if the upper 95% confidence limit (CL) for HR (T:C) is 1.23 or less, where $1.23 = 1.50^{1/2}$.

(ii) Clearly, ignoring the uncertainty in the estimate of the control effect is problematic and inflates the type I error [5]. An alternative approach that has been used in oncologic drug evaluation has been to use the lower 95% CL for the HR(P:C) as the estimate of the control effect and demonstrate preservation of $\frac{1}{2}$ of this effect [10–12]. Here the lower 95% CL for $\mathrm{HR}(P:C) = 1.13$. If used, then just over 11 200

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

deaths are required to provide 90% power at the one-sided, 2.5% significance level and non-inferiority will be concluded if the upper 95% CL for HR (T:C) is 1.06 or less, where $1.06 = 1.13^{1/2}$. However, this approach has been shown to be very conservative and should be avoided if at all possible [5, 6].

(iii) Rothmann *et al* offers a simple linear combination of the data arising in the active-control trial and the historical data on the control to test the null [1]. The test statistic, z^*, is given as

$$z^* = \frac{\ln \hat{\mathrm{HR}}(T{:}C) - (1-\delta)\ln \hat{\mathrm{HR}}(P{:}C)}{\sqrt{\hat{\mathrm{SE}}^2(\ln \hat{\mathrm{HR}}(T{:}C)) + (1-\delta)^2\hat{\mathrm{SE}}^2(\ln \hat{\mathrm{HR}}(P{:}C))}} \quad (2)$$

Under the assumption of constancy, $|z^*| > 1.96$ leads to a rejection of the null and so a conclusion that T retains more than 100δ% of the control effect and the type I error is maintained at 2.5% one-sided. The number of deaths required to achieve a given power can be calculated via [Reference 1, equation (12), p. 254].

Using this approach, a total of 3220 events are required to provide 90% power at the one-sided, 2.5% significance level. Non-inferiority will be concluded if the upper 95% CL for HR (T:C) is 1.12 or less, being intermediate between (i) and (ii). The 1.12 limit is derived using equation (9), p. 251 of Reference [1].

Thus the above approach appears attractive. Assuming constancy of the effect of the control and accepting assay sensitivity, a formal statistical comparison is proposed between the historical data characterizing the control effect and the data arising in the trial comparing new to control, thereby explicitly incorporating the uncertainty (i.e. SE) around the control effect estimate. Operationally, this approach is equivalent to conventional methodology using *not* the point estimate for the control effect, but, rather, a lesser effect somewhere between the point estimate and its lower (two-sided) 95% confidence limit. This reduced effect is chosen so that the chance of falsely approving an inferior drug is exactly 2.5%, thereby managing the regulatory risk.

However, while sample sizes are smaller than would be the case with the use of the lower 95% CL for the control effect estimate, this approach can still result in very large trials. This derives from a combination of the (arbitrary) effect retention fraction (50% in this example) and the strength of prior characterisation of the control effect, which is reflected in the (historical) p-value (0.005 in this example). As illustrated in Table I, the application of this methodology may actually require more events than there are patients with the disease [9].

In reply to these concerns, it has been suggested that, rather than assuming T and C are equal in effectiveness, active control trials could be powered at alternatives where T is a little better than C and it has been further recommended that survival trials intended to support regulatory approval should not have fewer than 200 deaths [15]. While both suggestions would result in a lower sample size requirement for the active-controlled trial, they are somewhat problematic. In the absence of

Table I. The number of deaths required to prove a new treatment retains $\frac{1}{2}$ of the effect of control treatment (HR = 0.50) using Rothmann *et al* methodology (true HR new:control is unity, 90% power, α 2.5% one-sided).

(Historical) p-value for control vs placebo	Upper 95% CI for HR(T:C) must be less than	Approx No. of deaths required to prove 50% effect retention
0.049	1.004	3 000 000
0.02	1.12	3082
0.01	1.18	1563
0.001	1.27	735
0.0001	1.31	572
0.00001	1.33	505
0.000001	1.35	459
$<<0.000001$[a]	1.41	350

[a] Equivalent to the control effect being known with (virtually) complete certainty.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

data to the contrary, to assume T is a little better than C would be to reduce power and so increase the risk of an equivocal result and, therefore, risk exposing patients to trial procedures without a realistic prospect of obtaining a clear answer. With respect to a recommendation that survival trials should not have fewer than 200 deaths, there does not appear to be any regulatory guidance, peer reviewed publication or statistical text that argues for a threshold on the number of events to support a time-to-event analysis. It seems more reasonable to argue that considerations of plausible effect size, type I and II error are, and should remain, the determinants of the number of deaths needed to secure a meaningful analysis in a survival trial.

3. AN ALTERNATIVE APPROACH

While it is hard to argue that, in both planning and analysing an active-control, non-inferiority trial, the uncertainty around the estimate of the control effect can be ignored, it has been argued by Senn that pre-specification of a non-inferiority limit (here captured in the form of a given percent of the control effect to be retained) is of little value in analysing and interpreting such trials [16]. Rather, the judgement as to what is and is not an unacceptable loss of effectiveness of the control treatment should lie with the 'consumer', that is with physicians and their patients or the regulatory authority acting on the patients behalf. This is akin to the usual approach to interpreting data arising in superiority trials, where the hypothesis sensibly advanced in the planning stage to size the trial is seldom, if ever, taken into account when judging the clinical value of an observed, significant difference. Further, Senn provides a simple example where two sponsors compare their new drugs (say T1 and T2) against the active control (C), the first sponsor specifying a more liberal non-inferiority limit (say a HR(T1:C) = 1.30) than the second (say HR(T2:C) = 1.15). If the upper 95% CL is 1.25 for HR(T1:C) and 1.20 for HR(T2:C) then the rather odd conclusion is that T1 is non-inferior while T2 is not, despite being able to rule out a lesser for disadvantage for T2.

In the context of approaches where an indirect comparison is made between historical data and data arising in the trial of T compared to C, the application of a margin imposes a burden on the active-control trial greater than that which would be applied if a direct comparison to placebo was feasible. It could be argued that there should be no difference between the standards of evidence required in a superiority setting and in a non-inferiority (or active-control) setting. As stated in the recent CHMP guidance on non-inferiority trials, the first purpose of an active-control trial should be to provide evidence that a new agent would have been better than placebo if placebo had been included [17]. The second purpose of an active-control trial should be to estimate (indirectly) the size of the effect of the new agent relative to control. Both of these aims are achievable with the effect preservation approach and assumptions, with some small modifications.

This thinking leads to an alternative approach which does not require pre-specification of a percentage effect retention (and hence, a non-inferiority limit) though, having obtained the data, the result can be easily displayed in relation to the likely fraction of the standard effect retained. Examples that mirror this kind of approach have been given by Fisher and by Simon where the effect of drug relative to placebo, T:P, has been estimated by combining the effect of T:C, $\hat{\beta}_{T:C}$ say, estimated from the active-control trial with historical data estimating the effect of P:C, $\hat{\beta}_{P:C}$ say, [18,19]. Fisher considers only the contrast $\hat{\beta}_{T:C} - \hat{\beta}_{P:C}$ as an estimate of the effect of drug relative to placebo with variance $\hat{V}(\hat{\beta}_{T:C}) + \hat{V}(\hat{\beta}_{P:C})$ whereas Simon also considers the contrast $\hat{\beta}_{T:C} - (1-\delta)\hat{\beta}_{P:C}$ with variance $\hat{V}(\hat{\beta}_{T:C}) + (1-\delta)^2\hat{V}(\hat{\beta}_{P:C})$ for different values of effect retention, $0 \leqslant \delta \leqslant 1$.

Expanding on the previous example, to prove efficacy over placebo indirectly with 90% power and 2.5% one-sided significance, 800 events would be required in a head-to-head comparison of the new treatment with the standard. To understand the origin of this calculation, examination of (2) reveals that if $\delta = 0$ the null hypothesis being tested is HR$(T : P) = 1$. Thus, effect preservation

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

methodology with a zero effect retention is equivalent to testing indirectly whether drug is superior to placebo. The number of events required to achieve a desired power can therefore easily be obtained either via equation (12), p. 254 of Reference [1] or, equivalently, via equation (8), p. 487 of Reference [19]. Working again through equation (9), p. 251 of Reference [1], it is straightforward to show that indirect efficacy of T over P would be concluded if the upper 95% CL for the HR (T:C) was 1.26 or less.

4. THE EFFECT RETENTION LIKELIHOOD

Once complete, the data from the active-control trial can be displayed in the form of an 'effect retention likelihood' plot, through which the probability of the new drug retaining a given fraction of the control effect can be gauged.

Operationally, the effect retention likelihood is easily obtained from (2). Given the historical data on the control effect and the data arising in the trial of T compared to C, a range of values can be inserted for δ, the fraction of the control effect to be retained, from $\delta = 0$ which is equivalent to demonstration that T is superior to (putative) placebo, through to $\delta = 1$ which is equivalent to demonstration that T is superior to C. For each value of δ inserted, $\phi^{-1}(z)$ gives the likelihood that T has retained at least this fraction of the efficacy of C, where $\phi^{-1}(.)$ is the inverse cumulative density function for the standard normal distribution.

Continuing the example above, examples of possible effect retention likelihood plots are displayed in Figure 1.

Graphical display of the data in this fashion is more informative than the usual analysis associated with active-control, non-inferiority trials. For example, it can be discerned from Figure 1 that, if the observed hazard ratio was unity, then there is a 99.4% chance that the T would have beaten placebo if a placebo controlled trial would have been possible, a 97.8% chance that 50% of the effect of C has been retained and a 90% chance that 75% of the effect of C has been retained. In line with CHMP guidance, it may be more informative to interpret effects not as percentages, but in terms of a unique HRs and differences in median survival [17]. To see how this might be

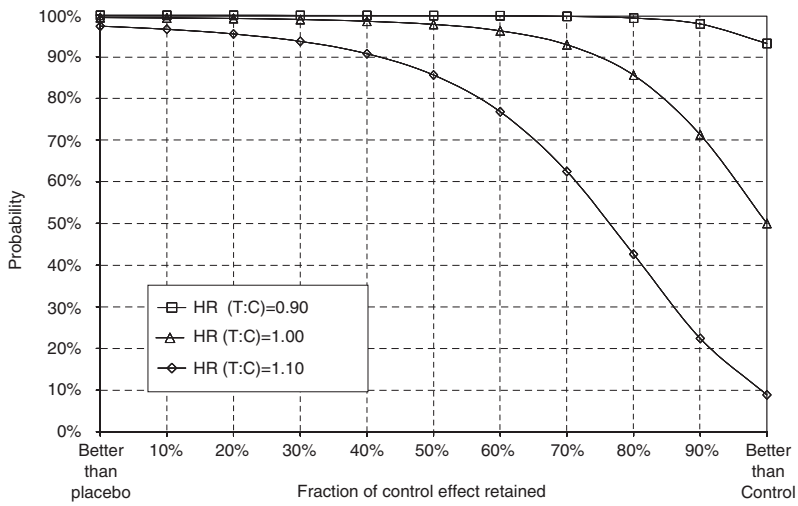


Figure 1. Effect retention likelihood plot: historical control effect, $HR = 1.5$, $p = 0.005$; active-control trial design with 800 events, 90% power, 2.5%, one-sided α level.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

achieved, the null hypothesis in (1) can be restated as H_0: $\mathrm{HR}(T{:}P) \geqslant \mathrm{HR}(P{:}C)^{\delta}$. Hence, if we insert the historical point estimate for HR(P:C), zero, 50% and 75% effect retentions correspond to the likelihood that the hazard ratio HR (T:P) is <1, <0.82 and <0.74 respectively. Further, if, for example, median survival on placebo was 6 months and assuming exponentially distributed survival times, then these hazard ratios would translate to the likelihood that the difference between T and P in median survival was >0, >1.4 and >2.1 months respectively.

5. EXAMPLES

5.1. 2nd line treatment of advanced non-small cell lung cancer (NSCLC)

In terms of a more concrete example, consider a new drug, T, for the treatment of 2nd line advanced NSCLC. The current approved standard of care in the USA is docetaxel 75 mg which, on the basis of a 104 patient trial, was shown to significantly improve survival over best supportive care (BSC), with HR (docetaxel:BSC) = 0.56, 95% CI (0.35, 0.88), $p = 0.01$ so that the SE ln HR (docetaxel:BSC) is approximately 0.23 [20,21]. Ignoring uncertainty in the estimate of the docetaxel effect, a clinical trial comparing T to docetaxel which aims to show retention of $\frac{1}{2}$ of the docetaxel effect with 90% power and 2.5%, one-sided significance would require around 500 deaths, already substantially larger than the original trial supporting approval. Application of Reference [1], equation (12), p. 254, to show retention of the same amount of the docetaxel effect would result in a trial requiring 2840 deaths. The alternative method proposed above would require 625 deaths to prove efficacy over BSC indirectly with the same power and significance level. Assuming a range of outcomes for the trial, examples of the possible associated effect retention likelihood plots are displayed in Figure 2.

For example, we can discern from Figure 2 that, if the observed hazard ratio was unity, then there is a 99.1% chance T would have beaten BSC if a BSC controlled trial would have been possible, a

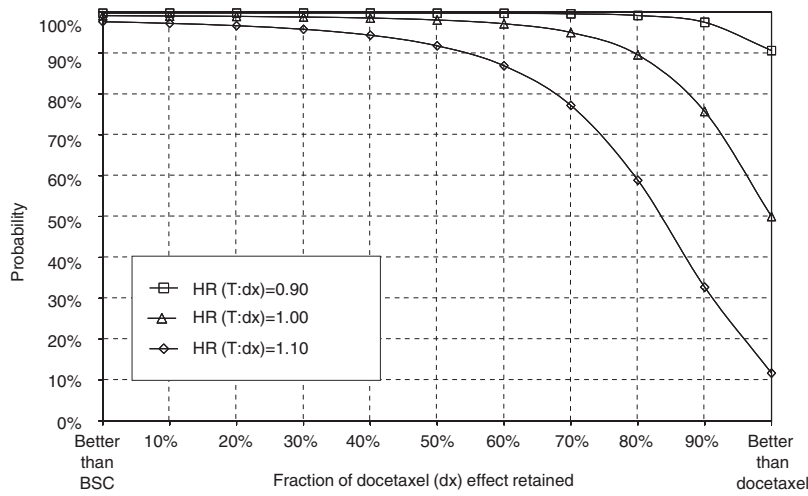


Figure 2. Effect retention likelihood plot: historical control effect $HR = 0.56$ $p = 0.01$; Active control trial design with 625 events, 90% power, 2.5%, one-sided α level.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

98.1% chance that 50% of the docetaxel effect has been retained and a 92.9% chance that 75% of the docetaxel effect has been retained. In terms of hazard ratios these numbers correspond to the likelihood the hazard ratio HR (T:P) is <1, <0.75 and <0.65, respectively.

5.2. Pemetrexed in the treatment of 2nd line advanced NSCLC

Building on the previous example, it is informative to consider pemetrexed which was compared to docetaxel in a randomized phase III trial in 571 patients with advanced NSCLC [22]. The aim of the protocol was to show non-inferiority of pemetrexed to docetaxel in terms of survival. As prospectively stated in the protocol, non-inferiority was to be concluded if the upper 95% CL was less than 1.11. A rationale for the 1.11 limit is not provided in the cited paper but corresponds to observing at least 78% retention of the docetaxel effect. The reported analysis was conducted with 409 deaths (target number of deaths was 385). The HR (pemetrexed:docetaxel) was 0.99, 95% CI (0.82–1.20) and median survival for pemetrexed was 8.3 vs. 7.9 months for docetaxel.

These data were reviewed in an open Oncologic Drugs Advisory Committee (ODAC) in July 2004 [23]. The discussion and debate on the data and non-inferiority issues in general was extensive. On the basis of the data, FDA concluded that pemetrexed was not proven to be non-inferior to docetaxel; the upper 95% CL for the HR (pemetrexed:docetaxel) was 1.20 and so exceeded the pre-defined non-inferiority limit of 1.11. However, ODAC panel members voted in favour of approval of the drug on the basis that it appeared to have similar survival to docetaxel and a different, more favourable side effect profile. This apparent divergence of views might have been avoided to some degree if the data had been presented as in Figure 3.

Figure 3 indicates that there is a 99.1% chance pemetrexed would have beaten BSC if a BSC controlled trial would have been possible, a 97.6% chance that 50% of the docetaxel effect has been retained and a 91.2% chance that 75% of the docetaxel effect has been retained. As before, these numbers correspond to the likelihood the hazard ratio HR (pemetrexed:BSC) is <1, <0.75 or <0.65 or better, respectively. With respect to the 1.11 limit in the protocol, corresponding to

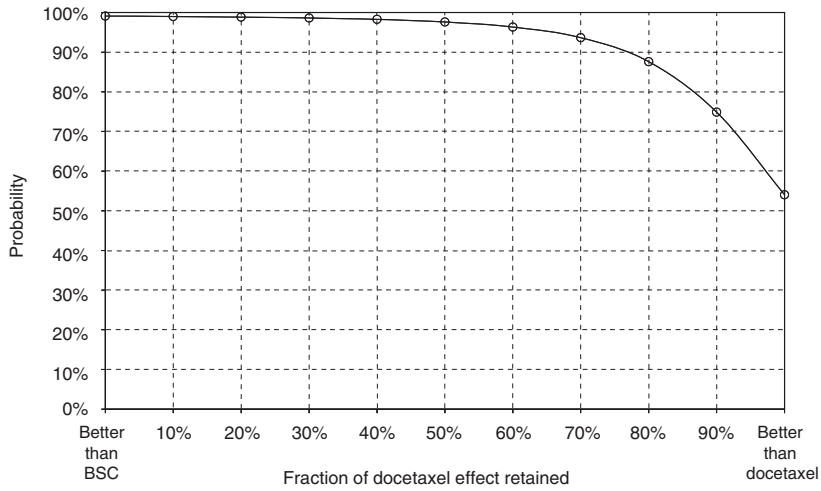


Figure 3. Effect retention likelihood plot for pemetrexed vs. docetaxel.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

observing 78% effect retention, the data show there is an 89% chance that this degree of the docetaxel effect has been retained by pemetrexed. Hence, by examining the data in terms of the effect retention likelihood, it is clear that the trial data provide confidence that pemetrexed is a efficacious drug and, further, that there is a high chance that pemetrexed retains a large fraction of the docetaxel effect. It is interesting to note that the interpretation of data that results from this approach is entirely consistent with the European Public Assessment Report on pemetrexed which states that 'although non-inferiority was not formally established, the data submitted are robust enough to conclude that a clinically significant inferiority of pemetrexed to docetaxel in terms of efficacy in this population is unlikely' which is a rather more complete assessment of the data than a statement that non-inferiority has not been met [24].

6. SITUATIONS WHERE THERE ARE NO TRIAL DATA DEMONSTRATING THE EFFECTIVENESS OF THE CONTROL RELATIVE TO PLACEBO

In all of the above scenarios it is assumed that there are historical data on the control treatment to provide an estimate of the effectiveness of the control over placebo (or some other agent). However, there may be some situations, most frequently in oncology, where treatments are prescribed in practice but no trial data exist that demonstrate efficacy relative to placebo. An example of such a situation might be the drug methotrexate which is commonly used to treat recurrent head and neck cancer. How one should proceed in such situations is problematic though the newly issued CHMP guidance does suggest an interesting way forward in which a trial is powered to compare T to C for modest superiority, but with a more liberal significance level than the conventional 2.5% one-sided [17]. For example, if it was hypothesized that the HR (T:C) was 0.80, then 528 deaths would be required for 90% power with one-sided α level of 10%. This means that an observed HR of 0.89 or better would be required to show T was superior to C with 90% probability. The situation is illustrated graphically in Figure 4 via a Normal probability density function with mean $\ln(0.89)$ and variance $\frac{4}{528}$.

In practice this approach will always require the HR (T:C) to be in favour of the new drug, but, in an attempt to balance the risk of failing to offer likely therapeutic advances to patients with few treatment options in truly life threatening situations, it does not demand $p<0.025$. This is clearly an important philosophical change in thinking that will hopefully prompt healthy and productive discussion in both the oncologic and regulatory communities alike and encourage development

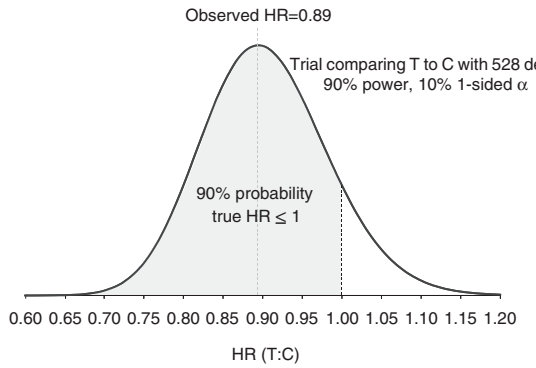


Figure 4. Probability density function for T vs C with 528 deaths, observed HR(T:C) 0.89, one-sided α level of 10%.

Copyright © 2006 John Wiley & Sons, Ltd.

Pharmaceutical
STATISTICS

of medicines in oncologic indications where treatments are used but have not been proven to be effective in placebo controlled trials.

7. DISCUSSION

This paper describes a stepwise approach to design and analysis of active-controlled, non-inferiority trials. In the first design step, the trial is sized to show indirectly that the new drug would have beaten placebo; in the second analysis step, the probability that the new drug is superior to placebo is assessed and, if sufficiently high, in a final third step, the relative efficacy of the new drug to control is assessed on a continuum of effect retention via an 'effect retention likelihood plot'. This approach should be considered prospectively since it is likely to provide a more meaningful assessment of relative efficacy so that the value of new treatments can be better judged.

However, issues do remain. Arguably, the most difficult, often intractable, issues are 'constancy' and 'assay sensitivity' – the latter being the ability of a non-inferiority trial to detect a real difference between the treatments compared. Constancy is inherently an unverifiable assumption and inflation of the type I error when this assumption is not met has been well documented [5, 7, 25]. In fact 'constancy' and 'assay sensitivity' are closely related concepts, since the presence of 'assay sensitivity' guarantees that the control treatment will have a non-zero effect and the assumption of constancy guarantees the size of that effect. The latter is a quantitative version of the former. Since ICH E9 and E10 describe these issues in detail and provide some general guidance with respect to trial design and conduct, there is little more to add, though the issue of constancy is worthy of further, brief comment [26,27].

When a direct comparison to placebo is impossible or unethical, researchers have no other choice, but to make indirect comparisons with historical data. Whether the simple conventional, effect preservation or the conservative use of the lower 95% CL for the estimate of the control effect is used, the problems of indirect comparisons apply equally. Given that some form of formal or informal indirect assessment is inescapable for active-control trials, whether the aim is to show superiority or to demonstrate efficacy indirectly versus historical data, it would seem appropriate to argue that, in drug development, concentrating on estimating the degree of benefit over placebo, albeit through indirect measures, is more in line with the statutory efficacy standard, which calls for substantial evidence of efficacy to be established, with no requirements on relative efficacy with respect to existing agents. Further, it is often overlooked that the issue of constancy applies equally to superiority trials, where, while $p<0.05$ will always indicate a difference between treatments, this does not always imply efficacy. In such trials, if the comparator has not behaved as expected and is considered to have underperformed relative to historical data, then, even though $p<0.05$, doubts may persist that efficacy has been shown.

Nevertheless, showing that the new drug is superior in efficacy to placebo indirectly might still be considered insufficient, especially by regulatory authorities when making licensing decisions. The motivation for requiring demonstration that the new drug has retained some positive fraction of the historical control effect is often a way of compensating for uncertainty about the constancy assumption. Thus, if the constancy assumption is in doubt, then showing that you have retained, say, at least 50% of the historical control effect provides increased confidence that the new drug is truly better than placebo, even if constancy is violated to some degree with the control treatment underperforming. Unfortunately, attempting to address concerns about constancy in this way leads back to the arbitrary pre-specification of a percent effect to be retained and, hence, to large and often infeasible trial sizes. A better approach would be to size the active-control trial to demonstrate the efficacy of the new drug relative to placebo indirectly, and then use the effect retention likelihood plot to judge how much better than placebo the new drug is likely to be.

Another issue relevant to constancy in oncology trials is the impact of improvements in the

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

standard of care and the introduction of new, effective therapies for late-stage disease, both of which are likely to result in better apparent performance of the control relative to historical data. There is little that can be done directly to address this issue but it is worth noting that since the performance of BSC has typically remained unchanged in many oncologic settings, this failure of constancy for the control does not necessarily increase regulatory risk. Hence, if non-inferiority is concluded despite a somewhat better performance of the control than expected, confidence may actually be increased rather than decreased that the new drug is truly efficacious.

As highlighted by a reviewer, recent CHMP guidelines state that it is preferable to consider relative efficacy some absolute measure, like the hazard ratio, since, for example, 50% effect retention in an adjuvant setting is inherently different to 50% effect retention in late-stage disease [17]. As described above, it is possible to plot the effect retention likelihood on the hazard ratio or median difference scale and so address, at least in part, the issue raised in the guidance. This guidance also suggests that, over and above indirect demonstration that the new drug is efficacious relative to placebo, consideration should also be given to the traditional 'clinically unimportant difference' since this is likely to be valuable when interpreting trial data when a label claim of non-inferiority is sought. The approach offered in this paper does not imply that the usual approaches to determining this quantity via literature review, examination of national treatment guidelines and discussion with treating physicians are not worthwhile but, rather, only that this does not necessarily have to form the basis for sizing a trial and, further, should not be binding in terms of the achievement or not of non-inferiority since this can be more comprehensively assessed via the effect retention likelihood plot.

The scientific and statistical debate on how best to draw inferences from active-control, non-inferiority trials is ongoing. The recent literature serves to highlight that considerable statistical, methodological and philosophical issues remain. Failure to consider these issues constructively within the broader oncologic community could result in ever increasing drug development times and, thus, delay the availability of new therapeutic options to patients with life threatening diseases. The recently issued CHMP guidance on non-inferiority trials is highly constructive in this regard, and offers valuable new thinking especially in situations where a drug is commonly prescribed but no trial data exist to confirm its efficacy relative to placebo. It is hoped that the scientific community together with regulatory bodies worldwide will continue to give active-controlled, non-inferiority trial design and analysis further careful thought and that the approach offered in this paper will be of some value in this regard.

ACKNOWLEDGEMENTS

The author would like to thank Prof. John A. Lewis, Dr. Bob Milsted and three anonymous reviewers for their helpful and constructive comments on the ideas offered in this paper.

REFERENCES

1. Rothmann M, Li N, Chen G, Chi GYH, Temple R, Tsou H-H. Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine* 2003; **22**:239–264.
2. Wang S-J, Hung J. Assessing treatment efficacy in non inferiority trials. *Controlled Clinical Trials* 2003; **24**:147–155.
3. *Biometrical Journal* 2005; **47**(1):5–107.
4. Special issue: non-inferiority trials: advances in concepts and methodology. *Statistics in Medicine* 2003; **22**(2):165–336.
5. Hung J, Wang S-J, Tsong Y, Lawrence J, O'Neil R. Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**:213–225.
6. Snapinn SM. Alternatives for discounting in the analysis of non inferiority trials. *Journal of Biopharmaceutical Statistics* 2004; **14**:263–273.
7. Hung J, Wang S-J, O'Neil R. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal* 2005; **47(**1):28–36.
8. Wang S-J, Hung J. TACT method for non-inferiority testing in active controlled trial. *Statistics in Medicine* 2003; **22**:227–238.

Copyright © 2006 John Wiley & Sons, Ltd.

Pharmaceutical
STATISTICS

9. Carroll K, Milsted B, Lewis JAL. Letter to the editor, 'Design and analysis of non-inferiority mortality trials in oncology'. *Statistics in Medicine* 2004; **23**:2771–2774.
10. Shlaes DM, Moellering RC. The United States Food and Drug Administration and the end of antibiotics. *Clinical Infectious Diseases* 2002; **34**: 420–422.
11. FDA/CDER New and Generic Drug Approvals: Taxotare (docetaxol for injection concentrate) statistical review, February 2002. Available at: http://www.fda.gov/cder/foi/nda/2002/20-449S018_Taxotere_statr.pdf (last accessed 10 March 2006).
12. FDA/CDER New and Generic Drug Approvals: Xeloda (capecitabine tablets) product labelling. Available at: http://www.fda.gov/cder/foi/label/2003/20896slr012_xeloda_lbl.pdf (last accessed 10 March 2006).
13. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *British Medical Journal* 1996; **313**:36–39.
14. Schoenfeld DA. The asymptotic properties of nonparamteic tests for comparing survival distributions. *Biometrika* 1981; **68**:316–318.
15. Rothmann M. Author's reply, 'Design and analysis of non-inferiority mortality trials in oncology'. *Statistics in Medicine* 2004; **23**:2774–2778.
16. Senn S. 'Equivalence is different' – some comments on therapeutic equivalence. *Biometrical Journal* 2005; **47(**1):104–107.
17. Committee for Medicinal Products for Human use (CHMP). *Guideline on the Choice of the Non-inferiority Margin*. EMEA: London, 2005. Available at: http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf (last accessed 10 March 2006).
18. Fisher LD, Gent M, Buller HR. Active-control trials: how would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin, and placebo. *American Heart Journal* 2001; **141**:26–32.
19. Simon R. Bayesian design and analysis of active control clinical trials. *Biometrics*, 1999; **55**:484–487.
20. FDA/CDER New and Generic Drug Approvals: 1998–2004. Taxotare (docetaxol for Injection concentrate) product labeling, February 2003. Available at: http://www.fda.gov/cder/foi/label/2005/020449s033lbl.pdf (last accessed 10 March 2006).
21. Shepherd F, Dancey J, Ramlau R, *et al*. Prospective randomized trial of docetaxel versus best supportive care in patients with non–small-cell lung cancer previously treated with platinum-based chemotherapy. *Journal of Clinical Oncology* 2000; **18**: 2095–2103.
22. Hanna N, Shepherd F, Fossella F, *et al*. Randomized phase III trial of pemetrexed versus docetaxel in patients with non small cell lung cancer previously treated with chemotherapy. *Journal of Clinical Oncology* 2004; **22**:1589–1597.
23. US Food and Drug Administration Division of Oncologic Drug Products Advisory Committee meeting transcript. 27 July 2004. Available at: http://www.fda.gov/ohrms/dockets/ac/04/transcripts/2004-4060T1.htm (last accessed 10 March 2006).
24. *European Public Assessment Report, Alimta: Scientific Discussion*. EMEA: London, 2005. Available at: http://www.emea.eu.int/humandocs/Humans/EPAR/alimta/alimta.htm (last accessed 10 March 2006).
25. Wang S-J, Hung J. Utility and pitfalls of some statistical methods in active controlled clinical trials. *Controlled Clinical Trials* 2002; **23**:15–28.
26. ICH Topic E9. *Note for Guidance on Statistical Principles for Clinical Trials*. ICH Technical Coordination. EMEA: London, 1998. Available at: http://www.emea.eu.int/pdfs/human/ich/036396en.pdf (last accessed 10 March 2006).
27. ICH Topic E10. *Note for Guidance on Choice of Control Group for Clinical Trials*. ICH Technical Coordination. EMEA: London, 2000. Available at: http://www.emea.eu.int/pdfs/human/ich/036496en.pdf (last accessed 10 March 2006).

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

12. **Carroll KJ**. Statistical issues and controversies in active-controlled, 'non-inferiority' trials. Statistics in Biopharmaceutical Research 2013; 5:3, 229-238

Statistical Issues and Controversies in Active-Controlled, "Noninferiority" Trials

Kevin J. Carroll

Active-controlled, "noninferiority" (NI) trials continue to raise many issues and controversies. With placebo-controlled trials becoming increasingly difficult in areas like oncology, infection, arthritis and respiratory illness, the use of active-controlled, "NI" trials to evaluate new treatments is likely to continue to be an important feature of drug development. Such trials continue to pose fundamental issues, many of which remain without broad scientific or regulatory consensus. These issues range from the fundamental purpose of an active-controlled NI trial to determination of the effectiveness of control and sample size, to issues of assay sensitivity and trial quality, to the statistical methodologies to be used. In this article, these matters are reviewed and discussed and observations are offered regarding the relative merits of the most common methodologies currently in use for NI assessment. Opinions are also included occasionally, some perhaps controversial, with the intention of generating discussion and debate.

Key Words: Fixed margin; Percent preservation; Putative placebo; Standard of evidence; Synthesis.

1. Introduction

Active-controlled, "noninferiority" (AC, NI) trials have been a feature of drug development programs for many years, yet they continue to raise issues and controversies among researchers and regulators alike. With placebo-controlled trials becoming increasingly difficult in areas like oncology and infection, and also in diseases like rheumatoid arthritis and respiratory illness, AC, NI trials are likely to remain an important feature of drug development. Such trials continue to pose fundamental issues, many of which remain without broad scientific or regulatory consensus. And this is despite such trials being subject to multiple regulatory guidances from EMA and FDA (ICH E9 1998; ICH E10 2000; EMA 2005; FDA 2010). Issues range from the fundamental purpose of an active-controlled NI trial to determination of the effectiveness of control via historical data, to sample size determination and matters relating to assay sensitivity and trial quality. A further and critical area of debate is what statistical methodology should be used to determine, indirectly, the effectiveness of a new experimental drug relative to historical control data.

These matters are crucial to the reliability and informative value of AC, NI trials. In this review article, these matters are discussed and observations are offered regarding the relative merits of the most common methodologies currently in use for "NI" assessment, namely, the "fixed" margin, preservation of effect, and synthesis methods. Opinions are also included occasionally, some perhaps controversial, with the intention of generating discussion and debate.

© American Statistical Association
Statistics in Biopharmaceutical Research
August 2013, Vol. 5, No. 3
DOI: 10.1080/19466315.2013.786651

Statistics in Biopharmaceutical Research: August 2013, Vol. 5, No. 3

2. Is There Really Any Such Thing as a "Noninferiority" Trial?

The statutory requirement for regulatory approval of a new drug is (i) that the drug is effective and (ii) that there is judged to be a positive benefit:risk. Note that (i) requires demonstration that the drug is more effective than placebo. This is consistent with the CHMP guidance on NI trials, which states that the first purpose of an active-controlled trial should be to provide evidence that a new drug would have been better than placebo, if placebo could have been included (EMA 2005). Therefore, while comparative effectiveness might be a later requirement to secure pricing and reimbursement, new drugs do not have to be shown to have superior effectiveness to existing, approved agents to secure licensure (FDA Oncologic Drugs Advisory Committee 2004). If that was the requirement then, logically, and in the absence of some other tangible benefit, only one drug therapy could ever be approved and available at any given time since the approval of a new drug would logically necessitate the currently approved standard to be withdrawn.

There are essentially two ways to demonstrate drug effectiveness: (i) directly, via a placebo-controlled trial or (ii) indirectly via an AC trial when a placebo-controlled trial is either unethical or impractical.

In the latter case, effectiveness is established by either showing that the drug is superior to control and, therefore, the drug is better than placebo, or by showing indirectly, by reference to historical data, that the drug is better than placebo. When considered in these terms, and in line with Brown (2008), it might be argued that there is no such thing as an NI trial in which we can conclude "*drug is noninferior to control*," rather there are only AC trials with differing objectives. In terms of establishing effectiveness, the first objective of an active-controlled NI trial is therefore not NI, but rather is to establish indirectly that the new drug would have beaten placebo if a placebo-controlled trial could have been included. As discussed later, this is directly equivalent to the FDA's M1 margin, defined as "the entire effect of the active control" estimated from historical data (FDA 2010). Examination of the relative effectiveness of a drug to control by, for example, showing a given fraction of the control effect (often times 50%; Simon 1999; Hasselblad and Kong 2001; Wang and Hung 2002; Hung et al. 2003; Rothmann et al. 2003; Wang and Hung 2003a; Wang and Hung 2003b; Snapinn 2004; Hung et al. 2005) has been retained is an additional, arbitrarily higher hurdle, and is broadly equivalent to FDA's M2 margin defined as "the largest clinically acceptable" loss of effectiveness of drug relative to control (FDA 2010).

The author would argue that this higher hurdle represents a differentially higher effectiveness standard based merely on study design (Carroll, Milsted, and Lewis 2004; Peterson et al. 2010). It would seem preferable to maintain a single standard and address critical issues such as constancy using appropriate methodology such as discounting (though this itself is equivalent to the use of a lower *p*-value for "NI" and, hence, a higher standard). Nevertheless, examination of relative effectiveness is valuable in the judgment of benefit/risk, and may further be of value to reimbursement authorities (FDA 2010). However, in terms of establishing the effectiveness of a new drug, it could be considered as a subsidiary objective (Carroll 2006).

3. Some Disadvantages of Indirect Comparisons

The fundamental problem with AC, NI trials as a vehicle to establish drug effectiveness is that many experienced scientists, statisticians, and regulators have an understandable mistrust of indirect comparisons. The absence of randomization and blinding affords little protection against bias and increases the probability of erroneous licensing decisions. The situation can be compounded when the historical evidence of control effectiveness is relatively little (as can be the case in end stage cancer treatment), or uncontrolled (again as is not uncommon in oncology), or antiquated (Fleming and Powers 2008; Dane 2011). And when there are more substantial historical data, often in terms of several clinical trials of similar design, the evidence has to be aggregated using meta analytic methods, which raises a further set of issues including possible selection and publication biases in addition to debate around the appropriateness of fixed versus random effects meta analyses. Crucially, one must assume that the historical data not only are relevant to the population studied in the current active-controlled trial, but that the true effectiveness of control is identical in the historical setting and the current trial. This "constancy" assumption is inherently unverifiable and potentially poses a serious problem to *any* AC trial regardless of objective (Carroll 2006). And, finally, there is a closely related concept of "assay sensitivity," which guarantees control will have a nonzero effect, whereas the assumption of constancy guarantees the size of that effect. "Assay sensitivity" therefore relies on the quality of trial design, conduct, and completeness of patient follow-up (Carroll 2006). A poorly conducted trial with, for example, many protocol violators and deviators, and/or where there are many dropouts who are not followed for key trial endpoints, obviously provides little, if any, reliable evidence upon which to make credible inferences (Fleming 2008). And while the argument is often that such trials tend to bias toward the conclusion of NI, this may be not so

since confidence intervals widen as variability increases. In summary, the underlying issues relating to trial quality apply to any AC trial, since any AC trial of manifestly poor conduct will be of little, if any, inferential value.

4. Establishing the Control Effect via Historical Evidence

The upfront provision of data in support of the relevant effectiveness of control is an absolute necessity when planning an active-controlled NI trial, and is a task that requires considered, experienced statistical input (Fleming and Powers 2008). In the author's experience, when discussing trial design, regulators often raise concerns if there are only limited data supporting the effectiveness of an approved control drug (FDA Oncologic Drugs Advisory Committee 2004). This is not infrequently the case in advanced oncological disease, an example being docetaxol in advanced lung cancer (FDA/CDER New and Generic Drug Approvals, Taxotare 2011). While this may seem a reasonable regulatory concern, one has to ask if the prior trial data were judged sufficient to support approval of control, then why are the same data not sufficient to support an analysis to demonstrate, indirectly, a new drug is efficacious? If the weight of evidence for control is considered too weak to do so, then one might legitimately ask on what evidentiary basis was control approved in the first place?

4.1 Comparison of Current Approaches to NI Assessment

To explore the issues in AC, NI trials further, the following notation will be used. Let θ_{cp} denote the true effect of control relative to placebo estimated as $\hat{\theta}_{cp}$ with variance V_{cp}. Similarly, let θ_{ec} denote the true effect of experimental to control estimated as $\hat{\theta}_{ec}$ with variance V_{ec}. Under constancy, the true effect of experimental to placebo θ_{ep} is therefore estimated indirectly as $\hat{\theta}_{cp} + \hat{\theta}_{ec}$ with variance $V_{cp} + V_{ec}$. The fundamental hypothesis to be tested is therefore:

$$H_0 : \theta_{ep} = 0 \quad \text{versus} \quad H_1 : \theta_{ep} > 0. \tag{1}$$

In the context of powering for an "NI" assessment, the alternative in (1) is replaced by $H_1 : \theta_{ep} > \Delta$. Further, the hypothesis in terms of θ_{ec} may be stated as

$$H_0 : \theta_{ec} < 0 \quad \text{versus} \quad H_1 : \theta_{ec} = 0, \tag{2}$$

where, again, for the purposes of powering, the null is replaced by $H_0 : \theta_{ec} < -\Delta$. If $\Delta = \theta_{cp}$, then hypotheses (1) and (2) correspond directly to the *synthesis approach.* In this approach, the historical control effect estimate and its SE are combined directly with the control versus drug effect estimate from the current AC trial under the assumption of constancy (EMA 2005; FDA 2010; Schumi and Wittes 2011).

To determine effectiveness of the experimental treatment with $(1-\beta)$ power and one-sided Type I error rate α using the synthesis approach, and assuming $\hat{\theta}_{cp}$ and $\hat{\theta}_{ec}$ are Normally distributed, then

$$V_{ec(\text{synthesis})} = \left[\frac{\theta_{ec} + \theta_{cp}}{z_\alpha + z_\beta}\right]^2 - V_{cp}, \tag{3}$$

where z_ω represents a standard Normal deviate and α is one-sided. Equation (3) follows directly from the basic power equation $(\delta_1 - \delta_0)^2 = V(z_\alpha + z_\beta)^2$ relating to the hypotheses $H_0 : \delta = \delta_0$ versus $H_1 : \delta = \delta_1 (> \delta_0)$, where H_0 is rejected when $T > z_\alpha\sqrt{V}$, where T a sufficient test statistic such that $T \sim N(\delta, V)$. In practice in Equation (3), θ_{cp} and V_{cp} would be replaced by their estimates $\hat{\theta}_{cp}$ and $\hat{V}_{cp}$.

Due to concerns regarding constancy, Equation (1) is often modified to demonstrate that experimental drug retains some fraction f of the historical control effect so that the alternative is replaced by $H_1 : \theta_{ep} \geq f\theta_{cp}$ with $0 \leq f \leq 1$. In this case, the synthesis approach morphs into the *preservation of effect method* with $\theta_{ec} - (1 - f)\theta_{cp}$ in Equation (2) (Simon 1999; Hasselblad and Kong 2001; Wang and Hung 2002; Rothmann et al. 2003; Hung et al. 2003; Wang and Hung 2003a; Wang and Hung 2003b; Snapinn 2004; EMA 2005; Snapinn and Jiang 2008; FDA 2010). Then,

$$V_{ec(\text{preservation})} = \left[\frac{\theta_{ec} + (1-f)\theta_{cp}}{z_\alpha + z_\beta}\right]^2 - (1-f)^2 V_{cp}. \tag{4}$$

And, again, θ_{cp} and V_{cp} would be replaced by their estimates in Equation (4).

For the "*fixed*" *margin* approach, Δ is set to $\eta(\hat{\theta}_{cp} - z_\alpha\sqrt{\hat{V}_{cp}})$, where $0 \leq \eta \leq 1$ is arbitrarily chosen to insert conservatism in terms of the effectiveness of control (EMA 2005; FDA 2010; Schumi and Wittes 2011). Most commonly $\eta = 0.5$. For this approach, it follows that

$$V_{ec(\text{fixed})} = \left[\frac{\eta(\hat{\theta}_{cp} - z_\alpha\sqrt{\hat{V}_{cp}})}{z_\alpha + z_\beta}\right]^2. \tag{5}$$

Examining (3)–(5), we can see that, under the usual assumption $\theta_{ec} = 0$,

- The synthesis method is always more efficient than the preservation of effect method for testing indirectly the hypothesis a new drug is efficacious. This follows since $V_{ec(\text{preservation})} \leq V_{ec(\text{synthesis})}$; that is,

the synthesis method requires a lower N (or number of events) to test hypotheses (1) and (2).

- To determine effectiveness of experimental with $(1-\beta)$ power, both synthesis and preservation of effect methods require $\hat{\theta}_{cp} - (z_\alpha + z_\beta)\sqrt{V_{cp}} > 0$; that is, that the historical data provide an estimate of the control effect with $p < 2\Phi^{-1}\{-(z_\alpha + z_\beta)\}$, meaning a historical control effect with $p < 0.0012$ is required to achieve 90% power.
- The maximum power achievable to test indirectly the hypothesis that a new drug is efficacious via either the synthesis or preservation of effect methods is $\Phi^{-1}\{|z_{CP}| - z_\alpha\}$, where z_{CP} is the z-value for the historical control effect estimate.
- The same issues and power cap do not apply to the "fixed" margin approach where any level of power can theoretically be achieved providing $\hat{\theta}_{cp} - z_\alpha\sqrt{V_{cp}} > 0$; that is, the estimated control effect is significant with $p < 0.05$.
- For the common choice $f = \eta = 0.5$, the "fixed" approach is less efficient than the preservation of effect method when the control effect has significance $p < 2\Phi^{-1}\{-0.5z_\alpha\{(1 + z_\beta/z_\alpha)^2 + 1\}\}$, or $p < 0.00025$ when 90% power is desired.
- Otherwise, the "fixed" approach is more efficient.

A key observation is therefore that while the synthesis method is always more efficient than the preservation of effect approach, both require a historical control estimate with a p-value ≤ 0.0012 to provide at least 90% power to test, indirectly, whether experimental drug is effective. This may suggest that there is little need to further discount data on historical control (Snapinn 2004; Snapinn and Jiang 2008). And, with $0.00025 < p \leq 0.0012$ for historical control, the "fixed" margin approach is more efficient than preservation of effect approach, and this reverses with $p \leq 0.00025$.

To illustrate these observations, consider, without loss of generality, a control drug with hazard ratio (HR) versus placebo of 0.667 95% CI (0.524, 0.849), $p = 0.0010$ for overall survival, representing a 50% increase in the event rate for placebo relative to control based on 264 events. The "fixed" NI limit would typically be $= 0.849^{0.5} = 0.921$. Via Equation (5) and assuming constancy, an active-controlled trial would require 6270 events ($\approx$24$\times$ more than the 264 events characterizing the historical effect of control) to deliver 90% power to test, indirectly, the effectiveness of drug. The preservation of effect method (4) with $f = 0.5$ would require 35,073 events ($\approx$130$\times$ more than for historical control) while the synthesis method (3) would require 8768 events ($\approx$33$\times$ more than for historical control). The sensitivity of the power calculation to the strength of historical evidence characterizing the control effect estimate is further illustrated in Table 1.

In most oncological and infection settings, AC trials using either a "fixed" margin or a preservation of effect method would be infeasible, and the synthesis approach would be a severe challenge (Carroll, Milsted, and Lewis 2004; Dane 2011). In cardiovascular (CV) outcome settings, again the "fixed" margin or preservation of effect approaches would prove very challenging, while the synthesis method may be more feasible.

It should be noted that some authors have advocated hypothesizing some small benefit for drug relative to control under the alternative (Fleming 2008). If so, then Equation (2) becomes $H_0 : \theta_{ec} = -\Delta$ versus $H_1 : \theta_{ec} = +\xi\Delta$, where ξ is some small, positive number such that $\xi \ll \Delta$. Then the required sample size is reduced by a factor $(1 + \xi)^{-2}$. While this approach may at

Table 1. Comparison of sample size requirements in an AC trial to test indirectly at the one-sided 2.5% α level the hypothesis that a new drug is efficacious

Historical data								Number of events required in AC trial			
Historical HR for C:P[a]	No. of Events	V_{CP}[b]	95% CI	Two-sided p-value	Assumed true HR for E:C[c]	F or η	"Fixed" NI limit	Synthesis	"Fixed"	Preservation	Max power for efficacy[e]
0.667	100	0.0400	(0.451, 0.987)	0.04289	1	0.5	0.994	–[d]	1,000,128	–	53%
0.667	200	0.0200	(0.506, 0.880)	0.00419	1	0.5	0.938	–	10,297	–	82%
0.667	264	0.0152	(0.524, 0.849)	0.00100	1	0.5	0.921	8768	6273	35,073	91%
0.667	327	0.0122	(0.537, 0.828)	0.00025	1	0.5	0.910	1185	4747	4740	96%
0.667	500	0.0080	(0.560, 0.795)	0.00001	1	0.5	0.892	526	3188	2103	99%
0.667	1000	0.0040	(0.589, 0.755)	<0.00001	1	0.5	0.869	345	2129	1378	100%

NOTE: [a]Historical estimate of the control effect, C = control, P = placebo.
[b]V_{CP} = variance of the historical control estimate.
[c]Hypothesized effect for drug versus control effect, E = drug.
[d]Variability of historical control effect estimate too great to achieve 90% power to test indirectly the hypothesis that a new drug is efficacious.
[e]Maximum power achievable to test indirectly the hypothesis that a new drug is efficacious versus placebo.

first seem appealing, unless there is a very good scientific rationale for assuming drug is, in truth, marginally better than control, the author would caution the use of this approach since artificially manipulating the sample size downward in this way results in a reduced power of $\Phi^{-1}\{(z_\alpha + z_\beta)(1+\xi)^{-1} - z_\alpha\}$ to test the arguably more realistic hypothesis (2).

Overall, it can be seen that both the "fixed" margin and preservation of effect standard approaches transfer a considerable burden onto sponsors and academic organizations trying to bring forward new medicines with equal effectiveness but different or improved tolerability or other perceived benefits (Snapinn and Jiang 2008; 2011; Peterson et al. 2010). The synthesis method offers a more reasonable alternative and, under constancy, directly assesses the fundamental hypothesis of interest, (1) or (2), with Type I error α (Simon 1999; Hung et al. 2003; Rothmann et al. 2003; Peterson et al. 2010).

4.2 When There are No or Very Old Historical Data

While the situation is difficult when some historical data do exist, it is near impossible when there are no prior data quantifying the effectiveness of control, or where the historical data are from a completely different era, as is the case in for some infectious diseases where the only placebo-controlled data are around six decades old (Fleming and Powers 2008; Dane 2011). In such situations the historical evidence is subjectively discounted, down-weighting the observed effectiveness of control to such an extent that the resultant NI margin becomes something of a guess, forcing trials that are ever higher in size (Snapinn 2004; Fleming and Powers 2008; Dane 2011). Another approach might be to consider the EMEA guidance that encourages the use of active-controlled trials with a superiority objective with perhaps a somewhat more liberal α level for "significance" in those situations where using the usual 0.025 one-sided level results in impractical trial sizes. In this way, licensing decisions can be made on the basis of some randomized, controlled evidence of effectiveness rather than no evidence at all (EMA 2005; Carroll 2006).

5. A Comment on Pre-Protocol Versus ITT Analyses

It is well known and captured in regulatory guidelines that the primary analysis set in an NI trial is preferred to be per-protocol (PP) with an intent-to-treat (ITT) analysis being supportive (ICH E9 1998; ICH E10 2000; EMA 2005; FDA 2010). In the author's view, this may be challenged for several reasons: PP (or "modified ITT" which is not, incidentally, ITT) results in a comparison that is data driven and unsupported by randomization. In trials with morbidity and mortality endpoints, a PP analysis makes little sense as it could result in the exclusion of patient deaths and/or other morbid events. It is crucial that phase III confirmatory clinical trials are generalizable, and reflect what might happen in clinical practice. Analyses that attempt to "clean up" the patient population and exclude those who deviated from the protocol in some way, or even who never received randomized treatment, do not necessarily reflect reality and, hence, are of questionable value and meaning. And while it is often required that the PP and ITT analyses give qualitatively similar results, as highlighted by a referee, this seems odd given that these analyses have differing objectives, namely, to estimate efficacy (PP) and effectiveness (ITT) (Sheiner and Rubin 1995).

Rather, what is needed are valid, unbiased analyses based on the randomization. In relation to AC, NI trials, ICH E9 (1998) stated that

> "*. . . it is especially important to minimise the incidence of violations of the entry criteria, noncompliance, withdrawals, losses to follow-up, missing data and other deviations from the protocol, and also to minimise their impact on the subsequent analyses*."

So, for those who rightly worry about violators, deviators, and dropouts, the solution is not to cut them out of the analysis in a PP approach, but rather to execute AC trials to rigorous and exacting standards, to minimize protocol nonadherence and ensure full ITT follow-up of all randomized patients so the trial evidence generated is of the highest completeness and quality (Fleming 2008). In this way, regulators and the scientific community can rely upon the data and what they show. This goes to the heart of "assay sensitivity," as mentioned previously, ensuring that the AC trial is of the highest possible scientific standard, regardless of whether the objective is superiority or "NI."

6. A Comment on FDA's Draft NI Guidance

FDA's recent draft NI guidance lays out the challenges with AC, NI trials and, in so doing, raises many issues for debate (FDA 2010). Unfortunately the guidance is rather long and perhaps a little confusing at times and, consequently, misses the opportunity to provide very clear, concise, and consistent guidance to sponsors. However, FDA has begun to tackle the longstanding issues associated with AC, NI trials, and that is to be applauded. The most contentious part of the current draft guidance is, arguably, the introduction of not one, but two NI margins: M1 and M2, being as defined previously.

M1 is constructed such that it establishes the effectiveness of drug indirectly versus putative placebo, which, in principle, makes sense. M2, however, is a different matter. It is a "fixed" margin based upon clinical judgment, shaped by prior historical data. As raised previously, the problem with this margin is that it represents an arbitrarily higher standard of effectiveness based on trial design (Carroll 2006; Peterson et al. 2010). Furthermore, the notion that it is a "fixed" margin representing some known and acceptable loss of effectiveness, and therefore is treated as a constant in both powering and analysis seems rather strange. Since the margin is obviously based on the available evidence quantifying the effectiveness of control, and since that evidence has uncertainty associated with it, the margin itself is hard to call "fixed." ICH guidance calls for a justification of the margin, which must take us back to the historical data and their uncertainty. To ignore the underlying uncertainty in the data giving rise to the margin is arguably improper statistically. Use of the synthesis method is more satisfying in this regard as the uncertainty in the control effect estimate is directly accounted for. Further, there should only be one standard for drug effectiveness, being effectiveness versus placebo at the 0.05 level, which equates to M1 (Peterson et al. 2010). If a higher standard is required, say akin to the single trial level of evidence, then this boils down to M1 with $p < 0.01$, or lower, say, and hence there appears to be little need, per se, for M2 in establishing effectiveness of drug. If, for the likes of reimbursement, it is desired to *describe* the relative effectiveness of drug to control, then this can be achieved using methodology such as the effect retention likelihood (Carroll 2006); this is discussed further in Section 8.

7. Logical Problems With Preservation of Effect and "Fixed" Margin Approaches

Putting aside for now arguments relating to the need for a "fixed" margin M2, if one decides that this is the route to take then it should be noted that there are rather serious problems with both this approach and the related preservation of effect method, problems that could lead to rather odd licensure decisions.

Consider Figure 1 based on Snapinn and Jiang (2008) and Peterson et al. (2010).

Here a fictitious drug developer, EfficsFarma, say, has been required by the regulator to use the preservation of effect method in order to show that their new drug, Bettapill, retains a fraction f of the effectiveness of the control drug, Mediocredex. Mediocredex was previously shown to be better than placebo by TitanicFarma in a small Phase III study. The effectiveness of Mediocredex relative to the (indirect) effectiveness of Bettapill is illustrated in Figure 1 by their respective 95% confidence intervals, and the desired fraction of the Mediocredex effect to be retained is also shown, labeled $f \times \hat{\theta}_{\text{mp}}$. The results indicate that both Bettapill and Mediocredex are superior to placebo, and they suggest that Bettapill may be even better than Mediocredex, but since Mediocredex was approved first, the preservation of effect criterion logically requires that Bettapill cannot be approved since its 95% CI does not exclude the desired fraction of the Mediocredex historical effect.

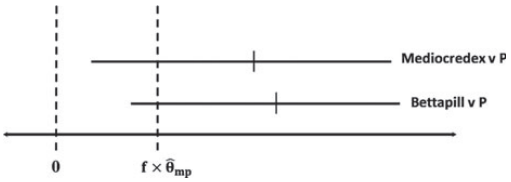


Figure 1. 95% confidence intervals for Mediocredex versus placebo and Bettapill versus placebo.

A similar kind of illogical consequence for the "fixed" margin approach was highlighted by Senn (2005). Suppose now that TitanicFarma and EfficsFarma compare their new drugs Mediocredex and Bettapill against some established active control, Proventrt say. Assume, without loss of generality, the endpoint is a time to event with treatment effects expressed as HRs. Suppose TitanicFarma specifies a fixed limit of 1.30 and EfficsFarma a limit of 1.15. Further, suppose the upper 95% CL for the estimated Mediocredex:Proventrt effect is 1.25, and for Bettapill:Proventrt is 1.20. The unfortunate conclusion is that Mediocredex is noninferior, while Bettapill is not, despite being able to rule out a lesser disadvantage for Bettapill relative to Mediocredex.

8. Example Design, Analysis, and Presentation of an AC NI Trial

Recognizing the issues highlighted above, the obvious question is: If not the standard "fixed" and preservation of effect approaches, then what? The most straightforward answer is, unsurprisingly, to use the synthesis method to both size and analyze the AC, NI trial since, under constancy, this approach provides a test of exactly size 2α to test the most relevant hypothesis, (1) (Peterson et al. 2010).

In terms of design, Equation (3) above provides the number of patients or events required to demonstrate effectiveness of drug relative to putative placebo at the one-sided α level with $(1-\beta)$ power. Also critically required at the design stage (as an appendix to the protocol), and not several months or years after the trial has commenced,

is a clear and transparent documentation of the historical body of evidence supporting the effectiveness of control and the clinical relevance of this evidence to the setting of the proposed AC, NI study. This approach is consistent with ICH E9 (1998), which states that

> *Active comparators should be chosen with care. An example of a suitable active comparator would be a widely used therapy whose effectiveness in the relevant indication has been clearly established and quantified in well designed and well documented superiority trial(s) and which can be reliably expected to exhibit similar effectiveness in the contemplated active control trial. To this end, the new trial should have the same important design features (primary variables, the dose of the active comparator, eligibility criteria, etc.) as the previously conducted superiority trials in which the active comparator clearly demonstrated clinically relevant effectiveness, taking into account advances in medical or statistical practice relevant to the new trial.*

Having conducted the AC trial to exacting standards of quality and patient follow-up, standards that are laid out in advance as part of the study protocol in line with Fleming (2008), the statistic $z = \frac{\hat{\theta}_{ec} - \hat{\theta}_{cp}}{\sqrt{V_{ep} + V_{cp}}}$ provides an indirect test of effectiveness via the synthesis method (Peterson et al. 2010). Clearly, constancy is a major concern and inclusion of analyses that discount the historical control data can be helpful and informative. In a manner similar to that proposed by Rothmann et al. (2003), one simple approach might be to discount $\hat{\theta}_{cp}$ by some amount based on judgment taking into consideration the amount, age, and relevance of the historical data. In this case, it may be of interest to calculate the maximum degree of discounting that would still provide $\hat{\theta}_{ep}$ with one-sided $p < 0.025$. Finally, it is helpful if the presentation of results is simple and transparent, as is hoped by example in Table 2 and Figure 2.

Beyond Table 2 and Figure 2, there is often interest in the relative effectiveness of drug to control. An approach to display the full range of relative effectiveness on a continuum, from drug better than placebo to drug better than control, has previously been described in terms of an effect retention likelihood plot (Carroll 2006). Figure 3

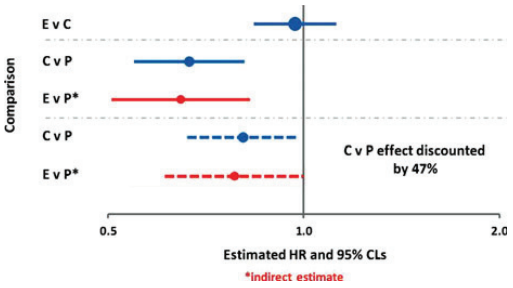


Figure 2. Active-controlled, NI analysis.

shows the likelihood effect retention for the preceding example, both for discounted and nondiscounted historical data.

Graphical display of the data in this fashion is more informative than the usual analysis associated with active-control, "NI" trials. We can discern from Figure 3 that there is a 99.98% chance that the drug is effective versus putative placebo; a 99.5% chance that 50% of the effect of control has been retained; and a 65.8% chance that 100% of the effect of control has been retained; that is, that the drug is superior to control. Similarly, if the historical control effect is discounted by as much as 43%, there is a 97.5% chance that the drug is effective versus putative placebo and a 93.7% chance that 50% of the effect of control has been retained. Discounting further would not allow rejection of $H_0 : \theta_{ep} = 0$. Hence, by examining the data in terms of the effect retention likelihood, it seems clear that, in this example, the AC trial data provide confidence that drug is efficacious and, further, there is a high chance that drug retains a large fraction of the control effect.

9. Summary and Recommendations

This article attempts to draw out and discuss issues and controversies associated with AC, "NI" trials. In line with Brown (2008), the issue of whether there is really such a thing as an AC, NI trial in which it can be stated that "*noninferiority was established,*" is not merely an

Table 2. Example presentation of active-controlled, "NI" analyses

	HR	Discount factor	Discounted HR	No. of events	SE logHR	Lower 95% CI	Upper 95% CI	*p*-Value
C v P	0.667	1	0.667	400	0.1000	0.548	0.811	0.00005
E v C	0.970	1	0.970	713	0.0749	0.838	1.123	0.68426
E v P*	0.647		0.647		0.1249	0.506	0.827	0.00049
C v P	0.667	0.53	0.807	400	0.1000	0.663	0.982	0.03185
E v P*	0.647		0.783		0.1249	0.613	1.000	0.04980

NOTE: *Indirect estimate.

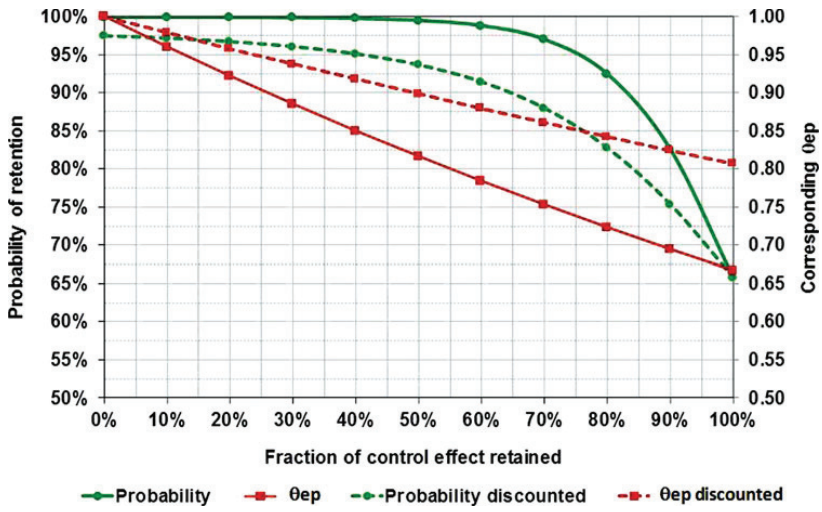


Figure 3. Effect retention likelihood plot for C v P HR = 0.67 (0.55, 0.81) and E v C HR = 0.97 (0.84, 1.12).

idiosyncrasy of nomenclature; it goes directly to the heart of what is trying to be achieved by conducting an AC trial. In terms of the statutory requirement for licensure, the primary goal is to establish effectiveness of the experimental drug, and this is achieved by indirect comparison via an AC trial in settings where it is unethical or infeasible to incorporate placebo. The dislike of indirect comparisons to historical data is entirely understandable—experienced statisticians would typically prefer the security of a randomized comparison to ensure both comparability and an absence of bias.

Without randomization, regulators are left uneasy, and consequently have tended to impose punitively conservative methods such as a "fixed" margin based on the 50% of the lower 95% CI, or a requirement for at least 50% retention of the control effect. Both of these approaches are subject to serious illogicalities, which arguably render them unsuitable for licensing decisions (Peterson et al. 2010). Even if the shortcomings are ignored and these methods are used, two undesirable outcomes flow as a consequence. First, a disproportionately large burden is transferred to the sponsor or research organization who is required to conduct large, often infeasible AC trials to offset the quantity of historical evidence, which ultimately could lead to fewer therapeutic advances in the future. Second, these methods represent an arbitrarily higher standard for effectiveness based purely upon trial design.

These and many other issues are covered in FDA's draft guidance. While much could be discussed in relation to this guidance, the single most important issue is the proposal for two NI margins, M1 and M2. While M1 makes sense in terms of establishing effectiveness of experimental drug, M2 imposes an arbitrarily higher hurdle based upon clinical judgment. The latter again represents a higher standard of effectiveness, this time determined subjectively by clinical judgment.

It would seem the only solution to the plethora of issues assailing AC, NI trials is pragmatism. In terms of design and analysis, it makes most sense to employ the synthesis method, thereby foregoing the need for a "fixed" margin or percent preservation target. Under constancy, this approach provides a test of size 2α to evaluate the most relevant hypotheses relating to drug effectiveness. While concerns regarding assay sensitivity and constancy are well founded, the latter cannot be tested statistically and is inherently unverifiable (Carroll 2006). Nevertheless, some discounting of the historical control estimated effect can be explored to ascertain how much discounting could be imposed and yet still demonstrate indirectly that experimental is effective. Further, critical to instilling confidence in the reliability of an AC, NI assessment is careful, a priori examination of the relevance of historical control data to the setting of the proposed AC trial (ICH E9 1998; Fleming 2008). And, with respect to assay sensitivity, the solution is not to resort to patient exclusions and nonrandomized PP analyses, but rather to establish stringent standards for trial conduct and execution that serve to enhance close adherence to the protocol and deliver full patient follow-up (ICH E9 1998; Fleming 2008).

Statistical leadership can help substantially in the matters, by ensuring the right thinking takes place upfront,

and that this thinking is reflected in the protocol. This means ensuring that

(i) all the relevant data on the historical control effect are laid out objectively and transparently, and are combined using statistically appropriate analytic techniques,

(ii) the relevance of the historical data to the setting of the proposed AC trial is carefully evaluated and documented, and

(iii) that metrics are laid down in advance to deliver the highest quality AC data by enhancing protocol compliance and minimizing violations, deviations and patient loss to follow-up.

In addition, statisticians can make a further, telling contribution by ensuring the primary hypotheses to be tested, (1) and (2) as provided in Section 4.1, are clearly stated in the protocol at the outset and are understood by nonstatistical colleagues. Statisticians can also contribute by ensuring the sample size and powering to test, indirectly, for drug effectiveness and the associated method of analysis are also captured and understood by all. Finally, the statistician needs to guarantee a transparent presentation of the analysis, where direct and indirect estimates of effect are clearly displayed side-by-side, and perhaps where methodologies are employed that illustrate the relative effectiveness of drug to control on a continuum from drug better than placebo to drug superior to control.

The scientific discussion around AC, NI trials is very active due to the many issues and controversies that remain unresolved, some of which are highlighted in this article. The intention is not to offer a set of answers to these issues, but rather to try cast some of these issues in a newer light, in the hope that this might spark further debate and discussion between researchers, sponsors, and regulatory agencies. In so doing, it is hoped that pragmatic approaches and solutions might be found that enable AC, "NI" trials of feasible dimension and excellent design to be conducted as part of drug development programs seeking to bring forward therapeutic alternatives to existing treatments.

Acknowledgments

The author thanks two anonymous referees and the Editor for the helpful and constructive comments offered in their review of this article.

[Received September 2012. Revised December 2012.]

References

Brown, D. (2008), "Non-inferiority Trials: A Regulator's Perspective," Available at *http://www.efspi.org/PDF/activities/international/non_inf_2008/1 3_regulator_DB.pdf*. [230,235]

Carroll, K., Milsted, B., and Lewis, J. A. L. (2004), "Letter to the Editor. Design and Analysis of Non-Inferiority Mortality Trials in Oncology," *Statistics in Medicine*, 23, 2771–2774. [230,232]

Carroll, K. J. (2006), "Active-Controlled, Non-Inferiority Trials in Oncology: Arbitrary Limits, Infeasible Sample Sizes and Uninformative Data Analysis. Is There Another Way?" *Pharmaceutical Statistics*, 5, 283–293. [230,233,235,236]

Dane, A. (2011), "Active Controlled Studies in Antibiotic Drug Development," *Pharmaceutical Statistics*, 10, 454–460. [230,232,233]

EMA Committee for Medicinal Products for Human Use (CHMP). (2005), *Guideline on the Choice of the Non-inferiority Margin*, London: EMEA. Available at *http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000366.jsp&mid=WC0b01ac0580032ec4*. [229,230,231,233]

FDA (2010), "FDA Guidance for Industry Non-Inferiority Clinical Trials," Available at *http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf*. [229,230,231,233]

FDA/CDER New and Generic Drug Approvals. (2011), "Taxotare (docetaxol injection concentrate) Product Labeling," Available at *http://www.accessdata.fda.gov/drugsatfda_docs/label/2011/020449s065lbl.pdf*. [231]

Fleming, T. R. (2008), "Current Issues in Non-inferiority Trials," *Statistics in Medicine,* 27, 317–332. [230,232,233,235,236]

Fleming, T. R., and Powers, J. H. (2008), "Issues in Non Inferiority Trials: The Evidence in Community-Acquired Pneumonia," *Clinical Infectious Diseases,* 47, S108–S120. [230,231,233]

FDA Oncologic Drugs Advisory Committee. (2004), "Food and Drug Administration Division of Oncologic Drug Products Advisory Committee Meeting Transcript," Available at *http://www.fda.gov/ohrms/dockets/ac/04/transcripts/2004-4060T1.htm*. [230,231]

Hasselblad, V., and Kong, D. F. (2001), "Statistical Methods for Comparison to Placebo in Active-Control Trials," *Drug Information Journal,* 35, 435–449. [230,231]

Hung, J., Wang, S.-J., and O'Neil, R. (2005), "A Regulatory Perspective on Choice of Margin and Statistical Inference Issue in Non-inferiority Trials," *Biometrical Journal,* 47, 28–36. [230]

Hung, J., Wang, S.-J., Tsong, Y., Lawrence, J., and O'Neil, R. (2003), "Some Fundamental Issues With Non-Inferiority Testing in Active Controlled Trials," *Statistics in Medicine,* 22, 213–225. [230,231,233]

ICH E9 (1998), *Note for Guidance on Statistical Principles for Clinical Trials. ICH Technical Coordination*, London: EMEA. Available at *www.ema.europa.eu/pdfs/human/ich/036396en.pdf*. [229,233,235,236]

ICH E10. (2000), *Note for Guidance on Choice of Control Group for Clinical Trials. ICH Technical Coordination*, London: EMEA. Available at *www.ema.europa.eu/pdfs/human/ich/036496en.pdf*. [229,233]

Peterson, P., Carroll, K., Chuang-Stein, C., Ho, Y.-Y., Jiang, Q., Li, G., Sanchez, M., Sax, R., Wang, Y.-C., and Snapinn, S. (2010), "PISC Expert Team White Paper: Toward a Consistent Standard of Evidence When Evaluating the Effectiveness of an Experimental Treatment From a Randomized, Active-Controlled Trial," *Statistics in Biopharmaceutical Research,* 2, 522–531. [230,233,234,235,236]

Rothmann, M., Li, N., Chen, G., Chi, G. Y. H., Temple, R., and Tsou, H.-H. (2003), "Design and Analysis of Non-inferiority Mortality Trials in Oncology," *Statistics in Medicine,* 22, 239–264. [230,231,233,235]

Schumi, J., and Wittes, J. T. (2011), "Through the Looking Glass: Understanding Non-inferiority," *Trials,* 12, 106. Available at *http://www.trialsjournal.com/content/pdf/1745-6215-12-106.pdf*. [231]

Senn, S. (2005), "'Equivalence is Different' – Some Comments on Therapeutic Equivalence," *Biometrical Journal,* 47, 104–107. [234]

Sheiner, L. B., and Rubin, D. B. (1995), "Intention-to-Treat and the Analysis of Clinical Trials," *Clinical Pharmacology and Therapy,* 57, 6–15. [233]

Simon, R. (1999), "Bayesian Design and Analysis of Active Control Clinical Trials," *Biometrics*, 55, 484–487. [230,231,233]

Snapinn, S., and Jiang, Q. (2008), "Preservation of Effect and the Regulatory Approval of New Treatments on the Basis of Non-inferiority Trials," *Statistics in Medicine,* 27, 382–391. [231,232,233,234]

——— (2011), "Indirect Comparisons in the Comparative Effectiveness and Non-Inferiority Settings," *Pharmaceutical Statistics,* 10, 420–426. [233]

Snapinn, S. M. (2004), "Alternatives for Discounting in the Analysis of Non inferiority Trials," *Journal of Biopharmaceutical Statistics,* 14, 263–273. [230,231,232,233]

Wang, S.-J., and Hung, J. (2002), "Utility and Pitfalls of Some Statistical Methods in Active Controlled Clinical Trials," *Controlled Clinical Trials,* 23, 15–28. [230,231]

——— (2003a), "Assessing Treatment Effecacy in Non Inferiority Trials," *Controlled Clinical Trials,* 24, 147–155. [230,231]

——— (2003b), "TACT Method for Non-inferiority Testing in Active Controlled Trial," *Statistics in Medicine,* 22, 227–238. [230,231]

About the Author

Kevin J. Carroll, Independent Statistical Consultant, 79 Albany Road, Bramhall, Cheshire, SK7 1NE, UK (E-mail: *kevincarroll2@sky.com*).

13. **Carroll KJ**. On the use and utility of the Weibull model in the analysis of survival data. Controlled Clinical Trials 2003; 24: 682–701

ELSEVIER

Controlled Clinical Trials 24 (2003) 682–701

Controlled Clinical Trials

On the use and utility of the Weibull model in the analysis of survival data

Kevin J. Carroll, M.Sc.*

AstraZeneca Pharmaceuticals, Biostatistics Group, Global Clinical Science, Alderley Park, Macclesfield, UK

Manuscript received September 9, 2002; manuscript accepted March 17, 2003

Abstract

In the analysis of survival data arising in clinical trials, Cox's proportional hazards regression model (or equivalently in the case of two treatment groups, the log-rank test) is firmly established as the accepted, statistical norm. The wide popularity of this model stems largely from extensive experience in its application and the fact that it is distribution free—no assumption has to be made about the underlying distribution of survival times to make inferences about relative death rates. However, if the distribution of survival times can be well approximated, parametric failure-time analyses can be useful, allowing a wider set of inferences to be made. The Weibull distribution is unique in that it is the only one that is simultaneously both proportional and accelerated so that both relative event rates and relative extension in survival time can be estimated, the latter being of clear clinical relevance. The aim of this paper is to examine the use and utility of the Weibull model in the analysis of survival data from clinical trials and, in doing so, illustrate the practical benefits of a Weibull-based analysis. © 2003 Elsevier Inc. All rights reserved.

Keywords: Survival data; Proportional hazards regression; Weibull model; Hazard ratio; Event time ratio

Introduction

Cox's proportional hazards regression model (or equivalently in the case of two treatment groups, the log-rank test) has become the statistician's mainstay in the analysis of survival data [1–6]. Its predominance stems from 3 decades of application and experience, together with the fact that it is distribution free; no assumption has to be made about the underlying distribution of survival times to make inferences about relative death rates. While this is a key

* Corresponding author: Kevin Carroll, AstraZeneca Pharmaceuticals, Biostatistics Group, Global Clinical Science, Alderley Park, Macclesfield, SK10 4TG, UK. Tel.: +44-1625-515234; fax: +44-1625-518537.

E-mail address: kevin.carroll2@astrazeneca.com

0197-2456/03/$—see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0197-2456(03)00072-2

strength of the model, it does introduce some limitations. Specifically, a direct quantification of the improvement in survival time is not possible, except in the special case of truly exponentially distributed lifetimes where the reciprocal of the hazard ratio estimates the ratio of median times to event [7]. However, lifetimes are seldom truly exponential in their distribution, so statisticians have tended to rely on Kaplan-Meier estimates of the underlying survivor function to read off estimated percentiles. The reliability and precision of these estimates depends upon the number of deaths and patients remaining at risk at any given time point on the curve. Median survival time is often used as a measure of improvement in time, though this measure is often unavailable at the earlier analyses of longer-term trials with relatively low event rates. Even when median survival can be estimated from the Kaplan-Meier curve, tests for differences in medians between treatments are generally approximate and do not directly link with tests for parity of hazard rates [8].

The Weibull model provides an alternative, fully parametric approach to the Cox model. Both these models are, in fact, closely related; both assume proportional hazards and both provide asymptotically unbiased, equally efficient estimates of the hazard ratio between two treatments. The Weibull model, in addition to being proportional, is simultaneously an accelerated failure-time model (AFT), and is the only parametric distribution to possess both properties [4,9]. AFT models simply examine survival times via a log-linear model so that treatment effects are expressed in terms of the relative increase or decrease in survival time. The Weibull, being both accelerated and proportional, therefore allows the simultaneous description of treatment effects both in terms of hazard ratios and also in terms of the relative increase or decrease in survival time; we might conveniently refer to this latter quantification of treatment effect as an "event time ratio," if only to illustrate the close parallel with the better known hazard (or event) rate ratio. Cox has suggested that these kinds of analyses are most favorable when a direct interpretation of the treatment effect is desired [10].

It is important to recognize that the Weibull and other AFT models are not new, having previously been described in the literature [11]. A good, accessible overview can be found in Colette [4]. Prentice and Kalbfleisch [9] and Wei [12] have discussed the potential use of AFT models in survival analyses and, more recently, Chen and Wang [13,14] have discussed AFT models alongside a new class of models, the "accelerated hazards model," which models how the underlying hazard changes over time.

Despite such coverage in the literature, the Weibull model is rarely used in the routine analysis and reporting of clinical trial data. Given that the Weibull allows simultaneous estimation of both the usual hazard ratio and an event time ratio, in addition to allowing a more thorough examination of proportionality and providing a means for predicting how data might mature over time, further consideration of its use and usefulness seems worthwhile.

The remainder of this paper is therefore structured as follows: the next section provides an overview of the Weibull model, including its form, estimation of hazard and event time ratios, examination of proportionality, and prediction of data maturation. After this a comparison of Cox and Weibull models in the analysis of real clinical trial data is made, followed by a brief discussion on the need for an exact distributional match when using the Weibull model. A brief summary of key results is then followed by the final section discussing the practical value and application of the Weibull and related models in the analysis of survival data in arising clinical trials.

The Weibull model

Before describing the Weibull model, it is helpful to consider a general distribution for lifetimes for which proportionality holds.

Let $T = t$ denote the time to some event of interest; this could be time to death or progression-free survival in an oncology setting. If $f(t)$ denotes the probability density function of T, $S(t)$ the survivor function, and $h(t)$ the hazard function, then, as is well known,

$$f(t)=h(t)e^{-\int_0^t h(u)du}$$

Under proportionality, $h_A(t) = \theta h_B(t)$, so that $S_A(t) = [S_B(t)]^\theta$, where θ is the hazard ratio and A and B denote two independent treatment groups.

The maximum likelihood estimate of the hazard ratio is the easily derived:

$$\hat{\theta}_{para} = e^{\hat{\gamma}_{para}} = \frac{\sum^{N_B}\int_0^{t_i} h(u)du}{\sum^{N_A}\int_0^{t_i} h(u)du}\,\frac{d_A}{d_B} \tag{1}$$

and

$$\hat{V}(\hat{\gamma}_{para}) = \frac{1}{d_A} + \frac{1}{d_B} \tag{2}$$

where d_A and d_B denote the total number of deaths observed in treatment groups A and B, respectively. Full details of this result are given in the appendix.

Armitage and Berry give an estimate of the hazard ratio associated with the Cox (log-rank) model [6],

$$\hat{\theta}_{\mathrm{Cox}} = e^{\hat{\gamma}\,\mathrm{Cox}} = \frac{d_A}{E_A}\frac{E_B}{d_B} = \frac{d_A}{\sum r_{iA}\dfrac{d_i}{r_i}}\,\frac{\sum r_{iB}\dfrac{d_i}{r_i}}{d_B} \tag{3}$$

where, at time t_i, there are a total of d_i events out of r_i subjects at risk with d_{iA} events out of r_{iA} at risk in group A and d_{iB} events out of r_{iB} at risk in group B so that $E(d_{iA}) = (r_{iA})\, d_i/r_i$; and d_A and d_B denote the total number of deaths in groups A and B, respectively. If γ is small then,

$$\hat{V}(\hat{\gamma}_{\mathrm{Cox}}) = \frac{1}{d_A} + \frac{1}{d_B} \tag{4}$$

Under the assumption of proportionality, Eqs. (2) and (4) show that the standard error for the log hazard ratio is asymptotically the same for *all* underlying distributions, $f(t)$, and is the same as that for the Cox model estimate. Thus, the use of parametric analyses does not lead to any asymptotic loss of efficiency compared to the log-rank or Cox analysis under the assumption of proportionality [2,9]. Furthermore, upon close examination of Eqs. (1) and (3), we can see that, under the assumption of proportionality, both quantities are estimating the average risk of death on treatment A relative to that on B. Hence, any parametric analysis where proportionality is assumed to hold, such as the Weibull (or simpler exponential), will give rise to an estimated hazard ratio very similar to that from a conventional Cox analysis.

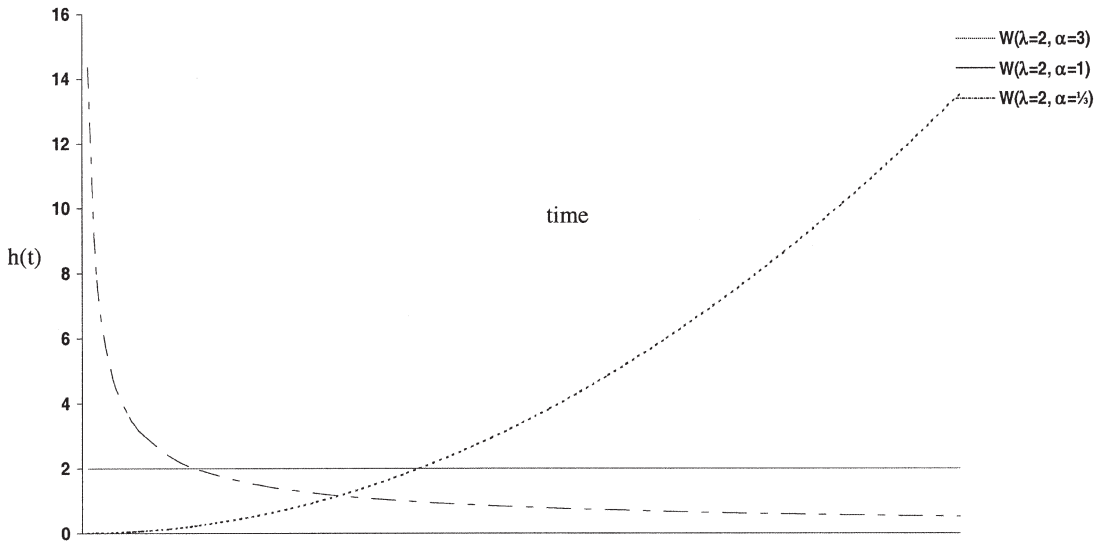


Fig. 1. The Weibull hazard function.

In general, therefore, we should not be concerned in employing parametric models, such as the Weibull, when proportionality holds.

Now, returning to the specifics of the Weibull model. If T represents the time-to-event variable, then the probability density function of the Weibull distribution is given by

$$f(t) = \alpha\lambda t^{\alpha-1} e^{-\lambda t^{\alpha}} \tag{5}$$

where $\lambda > 0$ is the event rate parameter and $\alpha > 0$ is the scale, or shape parameter. Thus, $S(t) = e^{-\lambda t^{\alpha}}$ and $h(t) = \alpha\lambda t^{\alpha-1}$. Note that the variable $Y = T^{\alpha}$ is a simple exponential with parameter λ. An illustration of $h(t)$ is given in Fig. 1. An alternative parameterization of the Weibull is given by setting

$$\alpha = \frac{1}{\sigma}, \quad \text{and} \quad \lambda_i = e^{-(\mu+\underline{\beta}'\underline{x}_i)/\sigma} \tag{6}$$

where the influence of the covariates, $\underline{x}_i$, for the ith individual is modeled through the event rate parameter, λ_i. (The software package SAS uses this parameterization when fitting the Weibull in the procedure PROC LIFEREG [15].) We shall now consider the important features of this distribution.

The hazard ratio

Based on the parameterization in Eq. (5), the hazard ratio for two treatments is given by

$$\theta(t) = \frac{\alpha_A\lambda_A}{\alpha_B\lambda_B} t^{\alpha_A-\alpha_B} \tag{7}$$

Hence, if $\alpha_A \neq \alpha_B$, hazards are not proportional. When proportionality does hold, $\theta = \lambda_A/\lambda_B$.

Based on the parameterization in Eq. (6), the log hazard ratio for an individual with covariates $\underline{x}_i$ relative to an individual with covariates $\underline{x}_j$ is

$$\frac{-\underline{\beta}'(\underline{x}_i - \underline{x}_j)}{\sigma} \tag{8}$$

In the case of just two treatments, the log hazard ratio is therefore given by $-\beta/\sigma$. In SAS the variance of the estimated log hazard ratio, $-\hat{\beta}/\hat{\sigma}$, is not given directly but can be easily derived from the variance-covariance matrix via Taylor's expansion [6],

$$\hat{V}\left(-\frac{\hat{\beta}}{\hat{\sigma}}\right) \cong \left(\frac{\hat{\beta}}{\hat{\sigma}}\right)^2 \left\{\frac{V(\hat{\beta})}{\hat{\beta}^2} - \frac{2Cov(\hat{\beta},\hat{\sigma})}{\hat{\beta}\hat{\sigma}} + \frac{V(\hat{\sigma})}{\hat{\sigma}^2}\right\}$$

As was noted above, under proportionality a Weibull analysis will give rise to an estimated hazard ratio and standard error very similar to that obtained from a conventional Cox analysis. This close matching of outcomes is easily verified by simulation. Table 1 shows the results of a simple simulation study where deviates from a range of Weibull distributions were randomly generated and analyzed in SAS by Cox's proportional hazards regression and also by assuming a Weibull distribution. For each Weibull shown, 1000 datasets were simulated, for sample sizes of 250, 100, and 25 in each of two treatment groups. A random amount of censoring (10%) was incorporated. A hazard ratio of 0.8 was used throughout.

Table 1. Simulated Weibull data: analysis by Cox and by Weibull

			Cox analysis			Weibull analysis		
n[a]	Shape	Event rate on treatment A	HR[b]	5th and 95th percentiles	SE[c] log HR	HR	5th and 95th percentiles	SE log HR
250	$\alpha = 1/3$	$\lambda_A = 0.5$	0.801	0.679, 0.934	0.0991	0.802	0.681, 0.935	0.0983
		$\lambda_A = 2$	0.801	0.685, 0.938	0.0966	0.800	0.686, 0.935	0.0955
	$\alpha = 3$	$\lambda_A = 0.5$	0.803	0.692, 0.937	0.0949	0.804	0.692, 0.937	0.0946
		$\lambda_A = 2$	0.796	0.685, 0.924	0.0914	0.796	0.685, 0.920	0.0912
100	$\alpha = 1/3$	$\lambda_A = 0.5$	0.805	0.624, 1.034	0.1529	0.804	0.624, 1.023	0.1516
		$\lambda_A = 2$	0.801	0.629, 1.027	0.1508	0.801	0.620, 1.024	0.1493
	$\alpha = 3$	$\lambda_A = 0.5$	0.794	0.612, 1.034	0.1550	0.794	0.692, 1.034	0.1539
		$\lambda_A = 2$	0.801	0.629, 1.047	0.1540	0.801	0.636, 1.040	0.1520
25	$\alpha = 1/3$	$\lambda_A = 0.5$	0.786	0.451, 1.345	0.3261	0.782	0.460, 1.334	0.3209
		$\lambda_A = 2$	0.799	0.483, 1.353	0.3127	0.800	0.494, 1.347	0.3079
	$\alpha = 3$	$\lambda_A = 0.5$	0.795	0.473, 1.363	0.3231	0.789	0.461, 1.349	0.3205
		$\lambda_A = 2$	0.810	0.485, 1.366	0.3202	0.805	0.474, 1.363	0.3151

[a] Number per group.
[b] Hazard ratio.
[c] Standard error.

Percentiles and the event time ratio

The percentiles of a Weibull are easily derived,

$$e^{-\lambda t^{\alpha}} = p \Rightarrow t_p = \left[\frac{\log\left(\frac{1}{p}\right)}{\lambda}\right]^{\frac{1}{\alpha}}$$

where t_p denotes the time taken to reach the p^{th} percentile. The relative difference in the time to achieving the p^{th} percentile between treatments A and B is

$$\frac{t_{Ap}}{t_{Bp}} = \frac{\lambda_B^{\frac{1}{\alpha_B}}}{\lambda_A^{\frac{1}{\alpha_A}}}$$

which, under proportional hazards, simplifies to the acceleration factor or event time ratio,

$$\kappa = \left[\frac{\lambda_B}{\lambda_A}\right]^{\frac{1}{\alpha}} = \left[\frac{1}{\theta}\right]^{\frac{1}{\alpha}} \tag{9}$$

Again, based on the parameterization in Eq. (6), the log event time ratio for an individual with covariates $\underline{x_i}$ relative to an individual with covariates $\underline{x_j}$ is $\underline{\beta}'(\underline{x_i}-\underline{x_j})$. In the case of just two treatments, the log event time ratio is simply given by β.

Note that Eqs. (8) and (9) demonstrate that, under proportionality, parameters describing changes in the log event time ratio are simply a scalar multiple of those describing changes in the log hazard ratio. As event time and event rate ratios are therefore linked by the shape parameter, it follows that if the hazard ratio can be estimated in a Weibull analysis, then so can the event time ratio.

Assessing proportionality in a Weibull analysis

In the analysis of survival data, graphical methods are routinely employed to assess the extent to which proportionality holds [4]. These methods may also be supplemented by a simple test for proportionality [16]. If data follow a Weibull distribution, then a direct, model-based test of proportionality can easily be achieved by comparison of shape parameters. If a Weibull is fitted separately for each treatment group, the two shape parameters, σ_1 and σ_2, say, together with their variances, can be independently estimated and compared.

To test the hypothesis $H_0 : \hat{\sigma}_1/\hat{\sigma}_2 \neq 1$, then

$$\frac{\left[\log\left(\frac{\hat{\sigma}_1}{\hat{\sigma}_2}\right)\right]^2}{\left[\frac{\hat{V}(\hat{\sigma}_1)}{\hat{\sigma}_1{}^2} + \frac{\hat{V}(\hat{\sigma}_2)}{\hat{\sigma}_2{}^2}\right]}$$

can be compared to a $\chi_1{}^2$ distribution. If shape parameters are found to differ significantly, then the null hypothesis of proportionality is rejected. In practice it may be more sensible to

examine the confidence interval for the possible extent of nonproportionality rather than relying on a significance test. This is because relatively mild departures from proportionality, such as late divergence of survivor functions, have little impact on inferences, especially if interpretation is confined solely to the time period of observation. Thus, in the analysis of clinical trial data, even when there is some modest departure from proportionality, it may still be reasonable to conclude that the event rate and event time ratio estimates show, on average, treatment differences over the period of study follow-up.

Assessing treatment differences when proportionality does not hold

While some interpretation of treatment effect estimates may be possible in the presence of modest nonproportionality, some statisticians will rightly feel unease in drawing conclusions. This being the case, the Weibull allows the hazard ratio to be plotted as a function of time, via Eq. (7). From this description of the hazard ratio, it is possible to compare treatments in terms of the average or integrated hazard over some time interval $(0-T)$. The integrated hazard is given by $\lambda T^{\alpha-1}$ so that the ratio of average hazards is given by $(\lambda_A/\lambda_B)T^{\alpha_A-\alpha_B}$. If a Weibull model is again fitted to each treatment group separately, the variance-covariance matrices can again be used to derive the standard error of the log of this quantity:

$$\hat{S}E\ log\left[\frac{\lambda_A}{\lambda_B}\,T^{\,\alpha_A-\alpha_B}\right] \cong \sqrt{\sum_{r=A,B} V[\log(\hat{\lambda}_r)] + T^2V[(\hat{\alpha}_r] + 2T\mathrm{Cov}[\log(\hat{\lambda}_r), \hat{\alpha}_r]}$$

The ratio of average hazards may then be plotted, with confidence limits, against time in order to explore how the averaged hazard ratio evolves with follow-up.

Predicting data maturation

The Weibull has been used in the field of engineering to predict the proportion of future failures after having observed a failure process to a given point in time [17]. In the context of clinical trials, predicting how deaths are likely to accumulate over time is often important, especially in the many trials designed with prespecified, event-driven interim analyses. In such trials, it is of great interest to accurately predict the time course of emerging deaths so that the appropriate resources can be put into place and to forewarn that perhaps additional follow-up beyond that envisaged at the outset, or at the previous analysis, is required to achieve the desired level of data maturity.

This can either be achieved in aggregate, for each treatment group, via simple extrapolation of the estimated survivor function, $e^{-\hat{\lambda}_r t^{\hat{\alpha}}}$, $r = A, B$, or by a more complex, individual patient based analysis as follows:

1. Assume an analysis has been performed with a mean follow-up time F, at which time d patients have died and $c = n-d$ are censored.
2. Consider the individual i with covariates $\underline{x}_i$, censored at time F. The probability that this individual survives to time $F+S$ is

$$p(T > F+S/T > F) = \frac{e^{-\lambda_i(F+S)^{\alpha}}}{e^{-\lambda_i F^{\alpha}}}$$

so that

$$F + S = \left[-\frac{ln(1-U)}{\lambda_i} + F^{\alpha}\right]^{\frac{1}{\alpha}} \tag{10}$$

where $U \sim U(0,1)$.

3. Survival times for the c censored individuals can be predicted if c deviates are randomly sampled from a $U(0,1)$ distribution, and substituted into Eq. (10). If, for the i^{th} patient, predicted survival exceeds $F + S$, then the patient remains censored; otherwise the patient is predicted to have died in the interval $(F, F + S]$.
4. Repeating 3, say, 1000 times, and averaging over repeats, provides an estimate of the number of additional deaths expected in the interval $(F, F + S]$.

This approach allows individual patient covariates to be used in predicting survival time, and so overall data maturity at a time S following the earlier analysis at time F. If a given level of maturity is required at the next analysis, the amount of additional follow-up needed can be estimated by trial and error.

An example

Analysis by both Cox's regression and the Weibull model is illustrated in the following example [18]. Patients with early prostate cancer were randomized to one of two treatments, active (bicalutamide 150 mg) or placebo. The primary endpoint was progression-free survival. The analysis took place at a minimum of 2 years and a median of 3 years follow-up. All patients were followed to disease progression or death irrespective of withdrawal of randomized therapy or addition of other, systemic therapies. Patients who remained progression free, or who were lost to follow-up at some earlier point in time, were censored. In addition to randomized treatment, four important, prospectively identified prognostic factors were included as covariates: these were primary background therapy (surgery, radiotherapy, or observation); log prostate-specific antigen level at diagnosis; stage of disease (either localized or locally advanced); and the degree of differentiation of disease (well, moderate, or poorly differentiated). The effect of primary therapy was captured in terms of contrasts between surgery versus radiotherapy and surgery versus observation. Similarly, the effect of degree of differentiation was captured in terms of contrasts between well versus moderately differentiated and well versus poorly differentiated.

A total of 1798 and 1805 patients were randomized to active and placebo treatments, respectively. At the time of the analysis, 181 and 293 events had accrued on active and placebo, respectively.

The results of the analysis are presented in Tables 2 and 3. Fig. 2 shows the Kaplan-Meier curves for the treatment groups, together with the fitted survivor function estimates from the Weibull analysis.

In Table 2, it is immediately obvious that both analyses provide very similar results, this being expected as discussed above. As the Weibull models log time, the parameter estimates, $\hat{\beta}$, represent event time ratios on the log scale. For example, patients with poorly differentiated disease were associated with a reduction in event time of approximately 33% (since $e^{-0.3973} = 0.67$) relative to those with well differentiated disease. As indicated in Eq. (8), division of $\hat{\beta}$ by $-\hat{\sigma}$ converts Weibull parameters from log event time ratios to log hazard ratios. Informal comparison of $-\hat{\beta}/\hat{\sigma}$ and $\hat{\gamma}$ indicates a close match between the Cox and Weibull models. This is as expected given that the Weibull distribution provides a close fit to the data.

Table 3 provides estimates of the treatment effect, both in terms of a hazard ratio and an event time ratio. From these results it can be seen that treatment with bicalutamide 150 mg significantly reduces the risk of progression compared to placebo by approximately 43% and, in doing so, significantly increases the progression-free survival interval by approximately 50%.

In terms of predicting how events might accrue over time, application of Eq. (10) indicates expected maturities of 21%, 28%, and 35% with additional follow-up of 1, 2, and 3 years, respectively.

The affect of departures from the Weibull distribution

Concerns may arise when using Weibull-based analyses in that the data collected may not conform exactly to a Weibull distribution. Simple graphical checks can be used to assess the extent to which data have a Weibull distribution and residual diagnostics can be also examined to assess goodness of fit [2,4].

Nevertheless, concerns may still be present that without a close distributional match, inferences based on a Weibull analysis may be misleading. However, for modest departures

Table 2. Results of Weibull and Cox analyses

	Weibull: modeling event time ratio				Cox: modeling log hazard ratio		
	$\hat{\beta}$	SE[a]$\hat{\beta}$	t	$-\hat{\beta}/\hat{\sigma}$	$\hat{\gamma}$	SE$\hat{\gamma}$	t
Intercept, μ	8.977	0.158					
Shape, σ[b]	0.7275	0.0307					
Randomized treatment[b]	0.4022	0.0706	5.70	−0.5529	−0.5544	0.0947	−5.85
Log PSA[c] at diagnosis	−0.2005	0.0352	−5.70	0.2756	0.2772	0.0471	5.89
Disease stage	0.3802	0.0746	5.10	−0.5226	−0.5265	0.1002	−5.25
Radiotherapy	−0.3184	0.0987	−3.23	0.4377	0.4382	0.1347	3.25
Observation	−0.6184	0.0837	−7.39	0.8500	0.8548	0.1096	7.80
Moderately differentiated	−0.1456	0.0891	−1.63	0.2001	0.1977	0.1222	1.62
Poorly differentiated	−0.3973	0.0937	−4.24	0.5461	0.5500	0.1275	4.31

[a] Standard error.
[b] Covariance between scale and treatment parameters was estimated to be 0.00047305.
[c] Prostate-specific antigen.

Table 3. Estimated hazard (HR) and event time ratios (ETR) for active relative to placebo

Cox			Weibull					
HR	SE[a]	95% CI[b]	HR	SE	95% CI	ETR	SE	95% CI
0.574	0.0947	0.477, 0.692	0.575	0.0947	0.477, 0.693	1.495	0.0706	1.302, 1.717

[a] Standard error.
[b] Confidence interval.

from a true Weibull, such concerns may largely be unwarranted, especially for hazard ratio estimation under proportionality.

To investigate hazard ratio estimates achieved via Weibull and Cox analyses irrespective of the true distribution for survival times, the following simulation approach was used.

Clinical trial data were simulated from lognormal, gamma, and piecewise exponential distributions. In each case, two treatments, *A* and *B*, say, were assumed and a random amount of censoring (10%) was incorporated. Parameters for each distribution were chosen so that the mean on treatment *A* was 6 months, say, and also so that variance of the lognormal and gamma distributions coincided. For the piecewise exponential, both treatments were assumed to have a common event rate for the first 3 months, diverging thereafter. Treatment differences, in terms of ratio of means, of 1.25 and 1.50 were used. To further reflect the clinical trial situation, uniform patient accrual over a 6-month period was simulated and a data cutoff

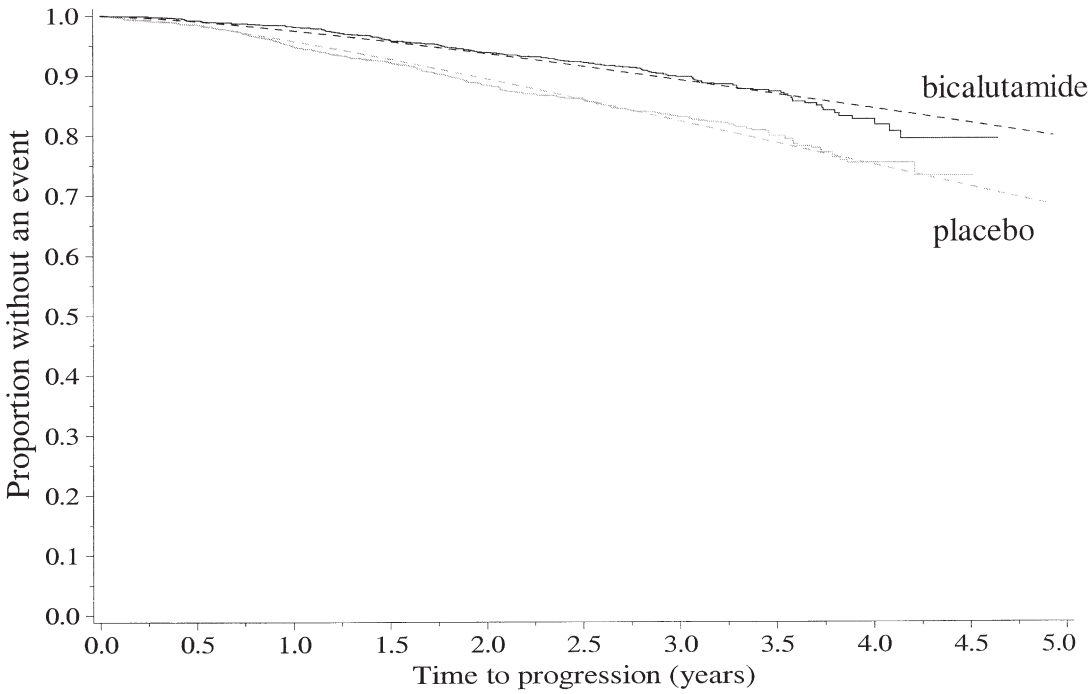


Fig. 2. Kaplan-Meier plot for progression-free survival with fitted Weibull survivor function estimates.

was employed whereby all event times were truncated at a given point in follow-up; the data cut-off time used was 12 months.

The lognormal and gamma survivor functions are illustrated in Figs. 3(a) and 3(b). Survivor function for the piecewise exponential distribution is illustrated in Fig. 3(c).

Data were then analyzed assuming a Weibull distribution via PROC LIFEREG in SAS. Both the hazard ratio and event time ratio were estimated. The data were also analyzed via PROC PHREG again in SAS to estimate the conventional Cox's hazard ratio. One thousand trial datasets, each of size $n = 400$ (200 in each treatment group), were simulated and the mean and variance of resulting log hazard and log event time ratios were calculated. A summary of results is provided in Tables 4–6.

These data suggest that even when data are known not to follow a Weibull distribution, analysis assuming a Weibull distribution provides results that are similar to those obtained by conventional Cox analysis. For the piecewise exponential, the Weibull analysis tends to give slightly larger log hazard ratio estimates (in absolute terms) compared to the Cox analysis, although the standard errors also tends to be slightly larger. Student t values were thus little different, perhaps being slightly higher for the Weibull analysis. With respect to the estimated event time ratio, this tends to be a little less than the known ratio of median times to event in the cases where the true median exceeds 3 months. If the 12-month data cutoff truncation is removed, then the resulting event time ratios are higher, as would be expected, and more in line with the known ratio of median times to event. For both lognormal and gamma data, results from Weibull and Cox analyses are virtually indistinguishable. The estimated event time ratio again tends to be less than the known ratio of median (or mean) times to event. Removal of the 12-month data cutoff truncation results in estimated event time ratios of 1.25 and 1.5, in exact concordance with the known differences in times to event.

This study suggests that hazard ratio estimates obtained via a Weibull analysis will tend to be similar to that obtained from a conventional Cox analysis, even when the Weibull does not provide an exact distributional match to the data. The importance of this is that for those data where it is considered reasonable to apply Cox regression to estimate the underlying hazard ratio, it should also be reasonable to apply a Weibull analysis to estimate the hazard ratio and, using the estimated scale parameter, to transform the hazard ratio to provide an estimated event time ratio. If the Weibull is considered only to provide a moderate fit to the data, then both the hazard and event time ratios can still be interpreted as the averaged risk of death and averaged increase in time on treatment A relative to treatment B. However, the estimation of percentiles, as described above, and the scale parameter are more dependent upon an adequate fit to the data. If a reasonable fit to the Weibull cannot be achieved, then it is recommended that percentiles be estimated directly from the Kaplan-Meier curves.

Summary

This paper has shown the Weibull model can provide a useful, parametric alternative to conventional Cox's regression modeling in the analysis of survival data. In addition to the hazard ratio, Weibull analysis provides a means of directly estimating the relative

improvement in survival time, the event time ratio. This quantification of treatment effect is of some clinical relevance and is likely to be better understood by some nonstatisticians than the conventional hazard ratio. Further, it has been shown that when data follow a Weibull distribution, Weibull analysis is asymptotically as efficient as Cox regression; both approaches give rise to similar hazard ratio estimates with the same standard error. Even when data are known not to follow a Weibull distribution, analysis assuming a Weibull distribution can

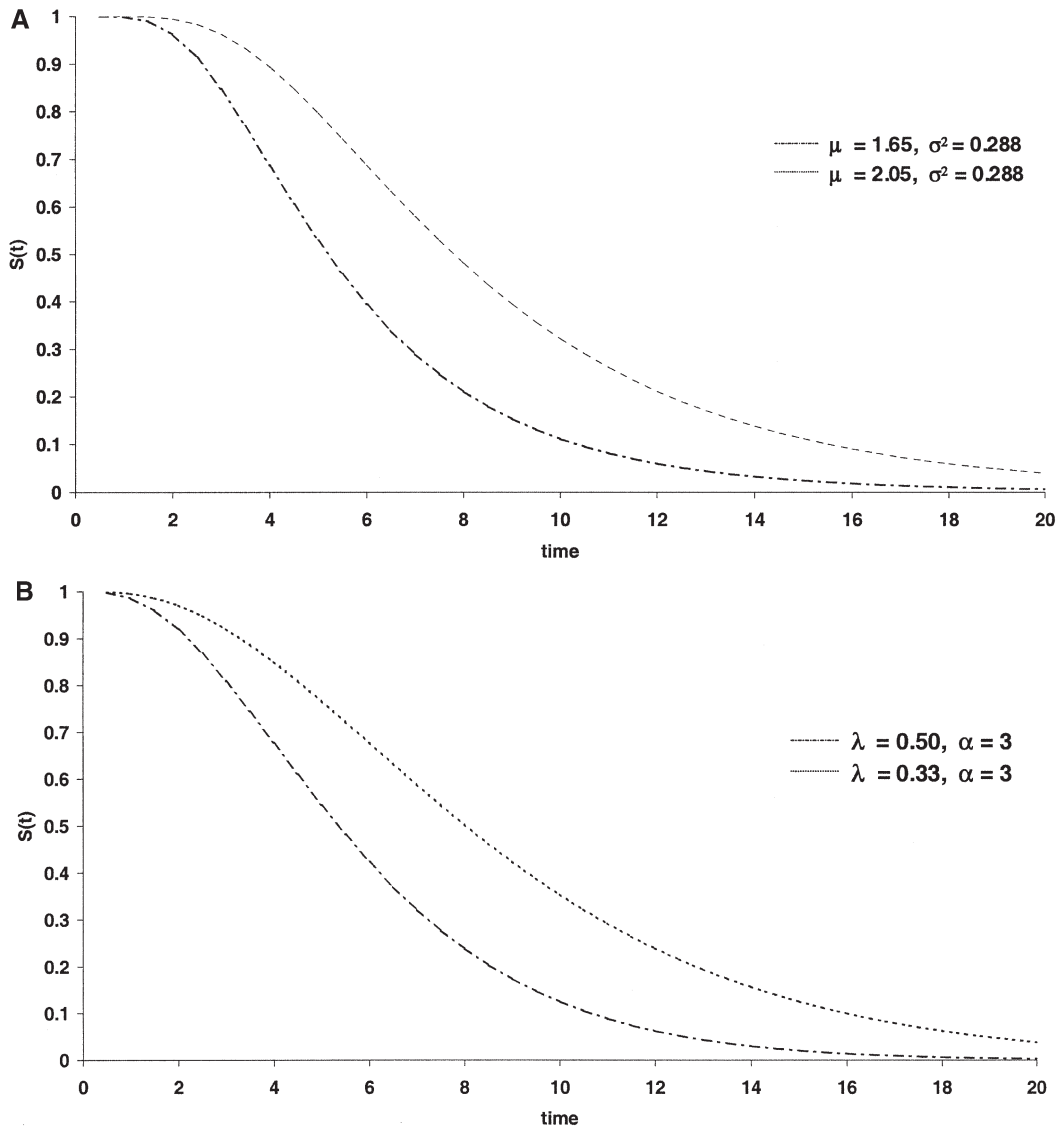


Fig. 3. (A) Lognormal survivor function. (B) Gamma survivor function.

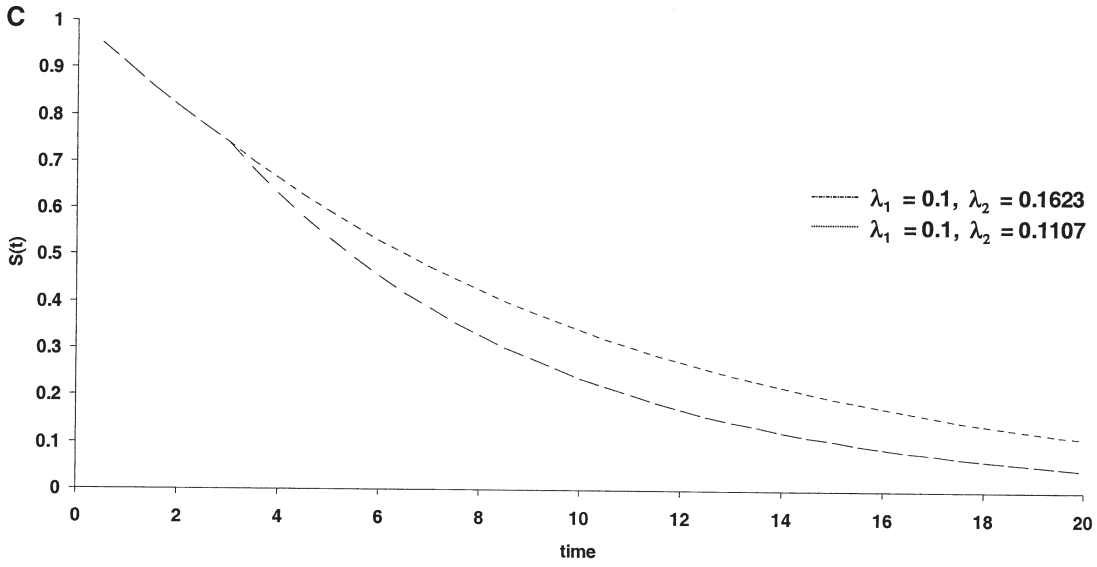


Fig. 3. (C) Piecewise exponential survivor function.

give very similar results, in terms of the hazard ratio, to those obtained by conventional Cox regression.

Key results in relation to the Weibull presented in this paper are thus:

1. Weibull analysis allows simultaneous characterization of the treatment effect in terms of the hazard ratio and the event time ratio, being a direct measure of the relative improvement in survival time.

Table 4. Simulation of piecewise exponential: analysis by Cox and by Weibull

			Cox analysis			Weibull analysis					
λ_1[a]	μ_A/μ_B[b]	$\tilde{\mu}_A/\tilde{\mu}_B$[c]	*HR*[d]	*SE*[e] *ln HR*	*t*	HR	SE ln HR	*t*	ETR[f]	SE In ETR	*t*
0.01	1.25	1.13	0.834	0.1199	−1.51	0.826	0.1185	−1.62	1.099	0.0585	1.62
0.01	1.50	1.26	0.716	0.1270	−2.64	0.702	0.1251	−2.83	1.191	0.0612	2.86
0.10	1.25	1.10	0.872	0.1142	−1.19	0.874	0.1073	−1.25	1.115	0.0868	1.25
0.10	1.50	1.21	0.783	0.1195	−2.05	0.784	0.1123	−2.16	1.221	0.0920	2.17
1	1.25	1.00	0.995	0.1096	−0.05	0.982	0.1341	−0.33	1.033	0.1568	0.14
1	1.50	1.00	0.987	0.1127	−0.12	0.967	0.1407	−0.25	1.043	0.1658	0.25

[a] Common event rate over first 3 months.
[b] Ratio of mean times to event; μ_A = *6 months throughout.*
[c] Ratio of median times to event.
[d] Hazard ratio.
[e] Standard error.
[f] Event time ratio.

Table 5. Simulation of longnormal: analysis by Cox and by Weibull

		Cox analysis			Weibull analysis					
μ_A/μ_B[a]	σ^2	HR[b]	SE[c] ln HR	t	HR	SE ln HR	t	ETR[d]	SE ln ETR	t
1.25	0.288	0.676	0.1156	−3.39	0.662	0.1236	−3.33	1.209	0.0559	3.40
1.5	0.288	0.489	0.1158	−6.23	0.472	0.1203	−6.24	1.397	0.0515	6.49
1.25	1.386	0.833	0.1111	−1.64	0.827	0.1165	−1.64	1.199	0.1104	1.64
1.5	1.386	0.720	0.1128	−2.91	0.711	0.1154	−2.95	1.376	0.1075	2.97

[a] Ratio of mean times to event = ratio of median time-to-event for lognormal data. $\mu_A = 6$ months throughout.
[b] Hazard ratio.
[c] Standard error.
[d] Event time ratio.

2. Weibull and Cox analyses coincide when data follow a Weibull distribution; both approaches are asymptotically equally efficient.
3. The Weibull provides an adequate fit in many situations. Even when data do not follow an exact Weibull distribution, a Weibull-based analysis can give results that are very similar to those obtained from a Cox analysis. However, the estimation of percentiles and the event time ratio is more dependent upon an adequate fit to the data. If a reasonable fit to the Weibull cannot be achieved, then it is recommended that percentiles be estimated directly from the Kaplan-Meier curves.
4. Weibull analysis allows direct assessment and quantification of proportionality, or lack thereof.
5. If the data display nonproportional hazards, then a Weibull analysis provides a description of the hazard ratio (and event time ratio) over time and, depending on the circumstances, an analysis of hazards averaged over time.
6. Weibull analysis offers the opportunity to predict how data might mature over time, something that is of great interest within oncology trials, especially where a series of interim analyses are planned.

Table 6. Simulation of gamma: analysis by Cox and by Weibull

		Cox analysis			Weibull analysis					
μ_A/μ_B[a]	α	HR[b]	SE[c] ln HR	t	HR	SE ln HR	t	ETR[d]	SE ln ETR	t
1.25	3	0.679	0.1075	−3.60	0.675	0.1084	−3.62	1.212	0.0540	3.63
1.5	3	0.494	0.1170	−6.03	0.490	0.1141	−6.24	1.415	0.0539	6.45
1.25	1/3	0.897	0.1125	−0.96	0.901	0.1038	−1.00	1.258	0.2302	1.00
1.5	1/3	0.821	0.1189	−1.65	0.828	0.1103	−1.72	1.522	0.2441	1.72

[a] Ratio of mean times to event = ratio of median time-to-event for gamma data. $\mu_A = 6$ months throughout.
[b] Hazard ratio.
[c] Standard error.
[d] Event time ratio.

Discussion

The key implication of this paper is that in those very frequent instances where two or more treatments are to be compared for survival (or some other time-to-event endpoint) with adjustment for one or more baseline prognostic factors, the Weibull is at least as informative as a corresponding Cox analysis, and probably more so. Use of the Weibull provides researchers and data analysts with an estimate of treatment effect as per routine Cox analysis but, furthermore, provides a clinically useful, alternative representation of the treatment difference in terms of the event time ratio—consistency, in terms of statistical significance, is assured as both measures of treatment effect have essentially the same *p*-value. On this basis, and as the primary objective of clinical trials with survival as the primary endpoint is the simple comparison of survival distributions, it seems reasonable to argue that a Weibull-based analysis would likely serve data analysts and clinical researchers better than a corresponding Cox-based analysis in most circumstances.

The assumption of proportionality is often an issue that rightly concerns statisticians when analyzing via Cox, being explored heuristically via graphical methods. The use of the Weibull offers the ability to explicitly examine the degree and nature of proportionality and, further, allows a simple, direct test for its presence. These model utilities are likely to be valuable tools in the routine analysis of clinical trial data.

The quantification of the treatment effect along the time axis is, in the author's experience, one of the most common requests from clinicians and other nonstatisticians in the analysis of survival and other time-to-event data and, thus, is one of the most common disappointments with Cox-based analyses. Simultaneous estimation of effects in terms of both rate and time is therefore a key strength of Weibull-based analyses. Unless data follow an exponential distribution, the common use of the reciprocal of the hazard ratio as an estimate of the relative difference in the median times to event is incorrect as easily evidenced via the above example discussed previously where the reciprocal of the hazard ratio would suggest a 1.7-fold increase in progression-free survival time, whereas a more appropriate Weibull analysis gives an event time ratio of 1.5 [7]. The routine use of Kaplan-Meier curves as a descriptive aid to Cox or log-rank analyses is both standard and sensible, but all too often is taken by nonstatisticians to be the literal interpretation of the analysis, such that the *p*-value is "attached" to the curves rather than to the hazard ratio, a practice which can be misleading. Weibull analysis allows the survivor function to be estimated, which, when plotted, more accurately reflects the estimated treatment effect. This in turn allows prediction versus data maturation, something that is of considerable practical value in the ongoing management of clinical trials with time-to-event endpoints, and yet another feature that does not readily flow from conventional Cox-based regression.

In applying a Weibull analysis, concerns may arise regarding degree of model fit. The simulations carried out in this paper would suggest that the Weibull provides an adequate fit in many situations such that even when data do not follow an exact Weibull distribution, a Weibull-based analysis gives results similar to those obtained from a corresponding Cox analysis. Upon reflection, it is not surprising nor unexpected that a time-to-event model with the flexibility of both a shape (λ) and a scale parameter (α) would provide a good fit in many situations just as it is not surprising that the normal distribution, with both location (μ)

and scale (σ) parameters, provides an adequate fit to interval data in a wide variety of applications.

While the majority of time-to-event analyses in clinical trials are univariate in nature, multivariate data often arises and a reasonable question is whether a Weibull-based approach can offer advantages here, too. Extensions of Cox-based regression for repeated event data have been developed, such as the commonly used Andersen and Gill model, which assumes the risk of a repeat event is unaffected by earlier events and follows the proportional hazards assumption [19]. A comprehensive overview of Cox-based models for multiple failure-time data has been offered by Wei and Glidden [20]. These authors note that while Cox-type regression has been widely used for multivariate failure-time data, it may not fit data well, and AFT models offer a useful alternative. They also note that AFT models can accommodate repeat events without natural ordering, being in contrast to the Andersen and Gill approach where natural ordering is assumed. Indeed, as the Weibull and other AFT models are simply log-linear models with error distributions reflective of time-to-event data, existing and well-developed theory in relation to generalized linear models and multivariate data analysis can be applied [12].

When wishing to explore the relationship between multiple events, random effects or "frailty" analyses can be considered. While extensions to Cox-based analyses in the form of time-dependant covariates are possible, Weibull and AFT models are preferred by some authors [20]. Keiding et al have suggested it would be advantageous to upgrade AFT approaches alongside conventional approaches for random effects survival analyses, emphasizing intuitive interpretation of the Weibull model [21].

In addition to extension to multivariate failure-time data, the Weibull and other parametric models have been found to be useful in other areas also, such as data monitoring and analysis of failure-time data when cure is possible [22]. Sposto has examined parametric cure models, concluding that they are at least as good as Cox-based approaches and are to be preferred when proportionality fails to hold, allowing simultaneous assessment of covariate effects on both the proportion cured and the failure rate among those not cured [23].

Despite the many appealing features of Weibull-based analyses, the author does advocate wholesale replacement of Cox's proportional hazards regression model for routine, univariate failure-time analyses. Albeit at the cost of assuming proportionality, Cox's regression offers the advantage of being distribution-free and can readily accommodate time-dependent covariate analysis. Rather, a Weibull analysis offers the statistician the opportunity to supplement and enrich routine Cox regression analyses, especially when a direct quantification of improvement in survival time is desired or a more thorough evaluation of proportionality is warranted. Indeed, when such matters are of primary interest, one may reasonably argue that a Weibull-based approach is to be preferred, being at least as informative as Cox regression with no loss of power or sensitivity under proportional hazards. Therefore, the use of the Weibull, or other parametric models, in the analysis of survival data in clinical trials, at very least, should not be overlooked and even be promoted to sit with equal status alongside routine Cox-based analyses.

It is interesting to note that model-based analyses are the norm and have been for many decades, in the analysis of normally distributed data, where analysis of variance to multivariate analysis of covariance to complex nonlinear mixed effects modeling approaches are routinely

employed, with nonparametric alternatives mainly taking a supporting role. Similar comments can be made in relation to binary and ordered categorical data where model-based data analysis prevails. For survival and other time-to-event data, however, the approach presently taken is the inverse in many ways; nonparametric analyses are considered standard while potentially more informative model-based approaches are seldom seen. This may, in part, be due to past difficulties in computationally applying these models.

However, widely available software packages, such as SAS and S-Plus, have simple procedures devoted to data analysis via the Weibull and other parametric models such as the gamma and lognormal [24]. Application of the Weibull-based analyses described in this paper is therefore very straightforward and not an area where specific, homegrown software has to be written to affect an analysis. Hence, it is fair to say that statisticians have simple and readily accessible software on their desks and are thus well poised and better equipped than ever before to reap the benefits that Weibull and other parametric-based approaches have to offer in the day-to-day, practical analysis of survival data arising in clinical trials.

Acknowledgments

The author would like to thank Mr. Stuart Ellis, AstraZeneca Pharmaceuticals, Alderley Park, UK, and two anonymous referees for their thoughtful comments that greatly helped to improve the clarity and focus of this manuscript.

Appendix: The hazard ratio for a parametric failure-time distribution under proportional hazards

Let T denote the time-to-event variable with probability distribution function $f(t)$, and survivor function $S(t)$. Under proportional hazards

$$h_x(t) = h(t)e^{\alpha+x\gamma}$$

where $x = 1$ for treatment group A, $x = 0$ for treatment group B.

Assuming N_A patients in group A with d_A events and $N_A - d_A$ censored. Employing similar notation for group B, the likelihood for the data observed is

$$L = \prod_i^{d_A} h(t_i)e^{d_A(\alpha+\gamma)}\, e^{-\sum^{N_A}\int_0^{t_i} h(u)e^{\alpha+\gamma}du} \prod_i^{d_B} h(t_i)e^{d_B\alpha}\, e^{-\sum^{N_B}\int_0^{t_i} h(u)e^{\alpha}du}$$

Thus,

$$\ell = \log(L) \propto d_A(\alpha+\gamma) - \sum^{N_A}\int_0^{t_i} h(u)e^{\alpha+\gamma}du + d_B\alpha - \sum^{N_B}\int_0^{t_i} h(u)e^{\alpha}du$$

$$\frac{\partial \ell}{\partial \gamma} = d_A - e^{\alpha+\gamma}\sum^{N_A}\int_0^{t_i} h(u)du \qquad (11)$$

$$\frac{\partial \ell}{\partial \alpha} = d_A - e^{\alpha+\gamma} \sum^{N_A} \int_0^{t_i} h(u)du + d_B - e^{\alpha} \sum^{N_B} \int_0^{t_i} h(u)du \tag{12}$$

$$\frac{\partial^2 \ell}{\partial \gamma \partial \alpha} = -e^{\alpha+\gamma} \sum^{N_A} \int_0^{t_i} h(u)du \tag{13}$$

$$\frac{\partial^2 \ell}{\partial \gamma^2} = -e^{\alpha+\gamma} \sum^{N_A} \int_0^{t_i} h(u)du \tag{14}$$

$$\frac{\partial^2 \ell}{\partial \alpha^2} = -e^{\alpha+\gamma} \sum^{N_A} \int_0^{t_i} h(u)du - e^{\alpha} \sum^{N_B} \int_0^{t_i} h(u)du \tag{15}$$

Hence, Eqs. (11) and (12) give

$$e^{\hat{\alpha}} = \frac{d_B}{\sum^{N_B} \int_0^{t_i} h(u)du}$$

$$e^{\hat{\alpha}+\hat{\gamma}} = \frac{d_A}{\sum^{N_A} \int_0^{t_i} h(u)du}$$

so that the hazard ratio, for any probability density function, $f(t)$, under the assumption of proportionality, is given by

$$e^{\hat{\gamma}} = \frac{\sum^{N_B} \int_0^{t_i} h(u)du}{\sum^{N_A} \int_0^{t_i} h(u)du} \frac{d_A}{d_B}$$

and Eqs. (13), (14), and (15) give

$$\underline{V} = E\left[-\underline{I}^{-1}\right]$$

$$-\underline{I}^{-1} = \begin{bmatrix} X+Y & Y \\ Y & Y \end{bmatrix}^{-1} = \frac{1}{XY} \begin{bmatrix} Y & -Y \\ -Y & X+Y \end{bmatrix}$$

where

$$X = e^{\alpha}\sum^{N_B}\int_0^{t_i} h(u)du \quad \text{and} \quad Y = e^{\alpha+\gamma}\sum^{N_A}\int_0^{t_i} h(u)du$$

Therefore, the variance of the log hazard ratio, under the assumption of proportionality, is given by

$$\hat{V}(\hat{\gamma}) = E\left[\frac{1}{X} + \frac{1}{Y}\right] = \left[\frac{1}{e^{\hat{\alpha}}\sum^{N_B}\int_0^{t_i} h(u)du}\right] + \left[\frac{1}{e^{\hat{\alpha}+\hat{\gamma}}\sum^{N_A}\int_0^{t_i} h(u)du}\right] = \frac{1}{d_A} + \frac{1}{d_B}$$

If γ is small, then

$\hat{V}(\hat{\gamma}) \cong 4/d$ where d denotes the total number of events across both groups.

References

[1] Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B 1972;34:187–220.
[2] Kalbfleisch JD, Prentice RL. The statistical analysis of failure-time data. New York: Wiley, 1980.
[3] Cox DR, Oakes D. Analysis of survival data. London: Chapman and Hall, 1984.
[4] Collett D. Modeling survival data in medical research. London: Chapman and Hall, 1994.
[5] Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). Journal of the Royal Statistical Society, Series A 1972;135:185–207.
[6] Armitage P, Berry G. Statistical methods in medical research. Oxford: Blackwell Scientific Publications, 1987.
[7] Sylvester R, Collette L. When do statistics lie? UroOncology 2001;1:185–194.
[8] Kim J. Confidence intervals for the difference of median survival times using the stratified Cox proportional hazards model. Biometrical Journal 2001;43:781–790.
[9] Prentice RL, Kalbfleisch JD. Hazard rate models with covariates. Biometrics 1979;35:25–39.
[10] Reid N. A conversation with Sir David Cox. Statistical Science 1994;9:439–455.
[11] Byar DP. Analysis of survival data: Cox and Weibull models with covariates. Statistics in Medical Research 1982;12:365–401.
[12] Wei LJ. The accelerated failure-time model: a useful alternative to the Cox regression model in survival analysis. Stat Med 1992;11:1871–1879.
[13] Chen YQ, Wang M-C. Analysis of accelerated hazards model. J Am Stat Assoc 2000;95:608–618.
[14] Chen YQ, Wang M-C. Estimating a treatment effect with the accelerated hazards model. Control Clin Trials 2000;21:369–380.
[15] SAS/STAT User's Guide. Version 6, 4th ed, Volume 2. Cary, NC: SAS Institute Inc., 1989.
[16] Gill RD, Schumacher M. A simple test of the proportional hazards assumption. Biometrika 1987;74:289–300.
[17] Nelson W. Weibull prediction of a future number of failures. Quality and Reliability Engineering International 2000;16:23–26.
[18] Wirth M, Tyrrell C, Wallace M, et al. Bicalutamide (Casodex) 150 mg as immediate therapy in patients with localized or locally advanced prostate cancer significantly reduces risk of disease progression (with Editorial comment). Urology 2001;58:146–151.
[19] Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Ann Stat 1982;10:1100–1120.

[20] Wei LJ, Glidden DV. An overview of statistical methods for multiple failure-time data in clinical trials. Stat Med 1997;16:833–839.
[21] Keiding N, Andersen PK, Klein JP. The role of frailty models and accelerated failure-time models in describing heterogeneity due to omitted covariates. Stat Med 1997;16:215–224.
[22] Lecoutre B, Mabika B, Derzko G. Assessment and monitoring in clinical trials when survival curves have distinct shapes. Stat Med 2002;21:663–674.
[23] Sposto R. Cure model analysis in cancer: an application to data from the Children's Cancer Group. Stat Med 2002;21:293–312.
[24] S-PLUS 6 Guide to Statistics, Vol. 2. Seattle, Washington: Insightful Corporation, 2001.

14. **Carroll KJ**. Back to basics: explaining sample size in outcome trials, are statisticians doing a thorough job? Pharmaceutical Statistics, 2009; 8: 333–345.

PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2009; **8**: 333–345
Published online 29 January 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.362

Back to basics: explaining sample size in outcome trials, are statisticians doing a thorough job?

MAIN PAPER

Kevin J. Carroll*,†
AstraZeneca Pharmaceuticals, CMOs Office, Alderley Park, Macclesfield, UK

Time to event outcome trials in clinical research are typically large, expensive and high-profile affairs. Such trials are commonplace in oncology and cardiovascular therapeutic areas but are also seen in other areas such as respiratory in indications like chronic obstructive pulmonary disease. Their progress is closely monitored and results are often eagerly awaited. Once available, the top line result is often big news, at least within the therapeutic area in which it was conducted, and the data are subsequently fully scrutinized in a series of high-profile publications. In such circumstances, the statistician has a vital role to play in the design, conduct, analysis and reporting of the trial. In particular, in drug development it is incumbent on the statistician to ensure at the outset that the sizing of the trial is fully appreciated by their medical, and other non-statistical, drug development team colleagues and that the risk of delivering a statistically significant but clinically unpersuasive result is minimized. The statistician also has a key role in advising the team when, early in the life of an outcomes trial, a lower than anticipated event rate appears to be emerging. This paper highlights some of the important features relating to outcome trial sample sizing and makes a number of simple recommendations aimed at ensuring a better, common understanding of the interplay between sample size and power and the final result required to provide a statistically positive and clinically persuasive outcome. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: ***outcome trials; sample size; power; hypothesized effect; critical value***

1. INTRODUCTION

An 'outcomes' trial has been defined as a large-scale, long duration clinical trial with hard clinical endpoints as outcomes, typically morbidity and mortality [1]. As such, they are invariably

*Correspondence to: Kevin J. Carroll, AstraZeneca Pharmaceuticals, CMOs Office, Alderley Park, Macclesfield, SK10 4TG, UK.
†E-mail: kevin.carroll2@astrazeneca.com

Copyright © 2009 John Wiley & Sons, Ltd.

Pharmaceutical STATISTICS

expensive and high-profile affairs. Such trials are commonplace in oncology and cardiovascular drug development but are also seen in other therapeutic areas such as respiratory in indications like chronic obstructive pulmonary disease. They tended to be closely monitored and results are often eagerly awaited. Once available, the top line results are often big news, at least within the therapeutic area in which the trial was conducted, and the data are subsequently fully scrutinized in a series of high-profile publications. In such circumstances, the statistician has a vital role to play in the design, conduct, analysis and reporting of the trial. In particular, in drug development it is incumbent on the statistician to ensure at the outset that the sizing of the trial is fully appreciated by their medical, and other non-statistical, drug development team colleagues and that the risk of delivering a statistically significant but clinically unpersuasive result is minimized. The statistician also has a key role in advising the team when, early in the life of an outcomes trial, a lower than anticipated event rate is emerging.

This paper highlights some of the key features relating to outcome trial sample sizing and makes a number of simple recommendations aimed at ensuring a better, common understanding of the interplay between sample size and power and the final result required to provide a statistically positive and clinically persuasive outcome. The remainder of the paper is therefore structured as follows: Section 2 examines the statistical issues relating to sample sizing in the context of a hypothetical dialogue between a medic and a statistician. Section 3 discusses how the statistician might better explain sample size calculations and their consequences to their non-statistical drug development colleagues and, in so doing, offers some simple recommendations. Section 4 then examines the role of the statistician in advising the drug development team when, some time after the trial is initiated, a lower than anticipated event rate appears to be emerging. Section 5 then closes paper with a brief summary of the main points raised.

2. A TYPICAL CONVERSATION REGARDING SAMPLE SIZE FOR AN OUTCOMES TRIAL

2.1. A typical, if somewhat simplified, conversation

Throughout this paper, issues in outcomes trial sizing are highlighted and discussed by means of a simple, hypothetical example. All that follows is therefore framed generically and the principles outlined apply equally to outcome trials in across therapeutic areas including oncology and cardiovascular where such trials are commonplace.

Suppose an outcomes trial is being considered for a new drug. A conversation ensues between the medic and statistician.

Medic: 'We need to show that 'Efektiv' is better than the current standard of care treatment in terms of clinical outcome. We need at least a 15% reduction in risk of the outcome to convince the medical community, regulators and formularies alike that 'Efectiv' is a genuine candidate to replace the current standard in managing patients. How big a trial are we looking at?'

After some thought, the statistician responds: 'We'll need 1,591 events to provide 90% power at the 2-sided 5% significance level. Further, assuming 10% of patients have an outcome event after 1 year, and with plans for a 1 year accrual period and a 1 year minimum follow-up period, 11 800 patients will need to be randomised'.

Hence, we are done. A total of 1591 events are needed and, assuming 10% of patients have an outcome event within 1 year, this translates to 11 800 pts. In this simple example, no allowance is made for dropouts as outcome studies are usually governed by the intent-to-treat principle so that outcomes are ascertained for all patients irrespective of dropout or early cessation of randomized therapy. That said, what follows would apply just as well if some allowance for dropouts were made. In addition, the 10% of patients with an event at 1 year is assumed to be a reasonable estimate, being based on relevant prior experience or literature.

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

2.2. Taking a closer look at the required number of events

The preceding calculation flows from the standard result for the log-rank test that the total number of events, d, is given by

$$d = \frac{4(z_\alpha + z_\beta)^2}{\theta^2} \quad (1)$$

where θ is the log hazard ratio (HR) hypothesized under the alternative, α is the 1-sided significance level, $1-\beta$ is power and $z_\gamma = \phi^{-1}(1-\gamma)$ is the $100(1-\gamma)$th percentile of the standard Normal distribution, this following from the result the estimated log HR is approximately $N(\theta, 4/d)$ [2–5]. Hence,

$$\mathrm{Var}(\hat{\theta}) = \frac{4}{d} \quad (2)$$

For planning purposes, if the expected fraction of patients with an event at some fixed time T is q_i, say, then

$$\theta = \log\left\{\frac{\log(1-q_E)}{\log(1-q_C)}\right\} \quad (3)$$

where E and C denote the experimental and control treatments [6]. Having conducted the trial, it is worth noting that the HR can be estimated directly from the statistics of the log-rank test as

$$\hat{\theta} = \frac{d}{4}\{d_E - E[d_E]\} \quad (4)$$

where d_E and $E[d_E]$ are the observed and expected number of events on the experimental treatment, E, respectively [7]. From (1) and (2) it immediately follows that the critical value of the test, the threshold value for the HR, $e^{\theta_{crit}}$, say, that flips the outcome between $p \le 0.05$ and $p > 0.05$, is given by

$$e^{\theta_{crit}} = e^{-2z_\alpha\sqrt{d^{-1}}} \quad (5)$$

For the example above, this critical value is 0.906, i.e. a 9.4% risk reduction. It is worth considering this a while, as this is where problems and misunderstandings can start to emerge.

In his conversation with the statistician, the medic is clear that he needs a 15% reduction in risk to convince physicians to change their usual practice. In response, the statistician has designed a trial to test the hypothesis H_0: $\theta = 0$ versus H_0: $\theta = \log(0.85)$. Now, it is not uncommon for the medic to assume that this means that if the trial delivers an HR of 0.85 or better, then $p \le 0.05$, otherwise the p-value will be 'NS', not significant. However, the trial will in fact yield $p \le 0.05$ for an HR of 0.906 or better, i.e. a 9.4% risk reduction or better; if a 15% risk reduction is observed, then it follows from (1) that

$$p = 2(1 - \phi(z_a + z_\beta)) = 0.0012 \quad (6)$$

Thus, there is a danger that the trial, as currently sized, will yield a statistically significant but clinically irrelevant result since the medic is blissfully unaware that differences smaller than the minimum difference required to be clinically persuasive will yield $p \le 0.05$.

This basic misunderstanding between what is minimally desired in terms of a convincing outcome and powering based on a hypothesis test that places this minimal requirement under the alternative is the main reason why so many times medics and other researchers turn to statisticians and say 'our trial was over-powered' or 'why do we need 1000 patients when a competitor achieved $p < 0.05$ in their trial with 500 patients?' It is the job of project statisticians working in drug development to explain as simply as possible why these apparent contradictions arise and to ensure complete transparency in terms of what hypothesis the trial is testing and what this translates to in terms of the threshold difference to yield a positive, $p \le 0.05$ outcome for the trial.

Hence, what might be done to improve this situation, to do a more thorough job at pointing out and communicating the issues and implications of our sample size calculations? This question is covered in Section 3 but, before then, it is informative to take a brief look at how, given the target number of events d, the overall sample size, N, is determined.

2.3. A brief comment on N

Given the number of target number of outcome events, d, the total number of patients to be randomized, N, is given by $N = d/\pi_{bar}$ where π_{bar} is the average probability of an event across E and

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

C over the trial period. Typically, for the purposes of calculating N, the time to an outcome event is assumed to be exponentially distributed, though other, non-parametric methods are available [8–10]. In addition, the accrual pattern is usually assumed to be smooth, with patient entry times following a Uniform distribution over the accrual period (0, A) [11–13]. Under these assumptions it has been shown that

$$\pi_i = 1 - \frac{e^{-\lambda_i F}[1 - e^{-\lambda_i A}]}{\lambda_i A} \quad (7)$$

where A is the length of the accrual period, F is the minimum follow-up period, λ_i is the outcome event rate with $i = E$, C and $\theta = \lambda_E/\lambda_C$ [14]. If λ_i is small, then

$$\pi_i \approx 1 - e^{-\lambda_i(\frac{A}{2}+F)} = 1 - e^{-\lambda_i(\text{mean follow-up time})} \quad (8)$$

Further, with exponentially distributed times to event, if the fraction of patients with an event at some fixed time T is q_i, say, then

$$\lambda_i = -\frac{\log(1 - q_i)}{T} \quad (9)$$

and

$$\pi_{\text{bar}} = 2(\pi_E^{-1} + \pi_C^{-1})^{-1} \quad (10)$$

i.e. π_{bar} is the harmonic mean of π_E and π_C [4]. In the example above, λ_E and λ_C are 0.0896 and 0.1054, respectively, and, thus, π_E and π_C are 0.1254 and 0.1458. Hence, $\pi_{\text{bar}} = 0.1348$, giving $N = 1591/0.1348 = 11\,803$. As is clear, the estimate for N is built on the two key assumptions of uniformly distributed entry times and exponentially distributed times to event. One or both of these can be relaxed to a degree to arrive at a more general form for π_i. For example, if times to event were assumed to follow a Weibull distribution [6] with shape parameter κ, say, so that $f(t) = \kappa\lambda t^{\kappa-1}e^{-\lambda t^{\kappa}}$, $t>0$, $\lambda>0$, $\kappa>0$, and entry times were assumed to follow some distribution $f(a)$, with $0 \le a \le A$, then

$$\pi_i = \int_{a=0}^{A}\int_{t=a}^{A+F} f(t|a)f(a)\,dt\,da$$
$$= 1 - E_a[e^{-\lambda_i(A+F-a)^{\kappa}}] \cong 1 - e^{-\lambda_i(A+F-E[a])^{\kappa}} \quad (11)$$

with the approximation applying when λ_i is small [15]. In this formulation, the form of $f(a)$ does not matter since all that is needed is $E[a]$ to provide an estimate of π_i Nevertheless, if so desired, a simple choice for $f(a)$ might be

$$f(a) = \frac{\eta a^{\eta-1}}{A^{\eta}}, \quad 0 \le a \le A \quad (12)$$

where η is the measure of non-uniformity in entry times, $0 < \eta < \infty$, and $E[a] = A\eta/(\eta + 1)$. Hence, $\eta = 1$ corresponds to uniformly distributed entry times and values $\eta > 1$ indicate slow initial recruitment that accelerates towards the end of the accrual period. If event times are exponential, then, for this choice of $f(a)$, evaluating equation (11) gives

$$\pi_i = 1 - e^{-\lambda_i F} \times \left\{(-1)^{\eta-1}\frac{\eta!}{(\lambda_i A)^{\eta}}\left[\sum_{s=0}^{\eta-1}\frac{(-\lambda_i A)^s}{s!} - e^{-\lambda_i A}\right]\right\}, \quad \eta = 1, 2, 3, \ldots \quad (13)$$

More generally, if times to event are exponential with λ_i small, so that the approximation (8) can be used, and accrual was expected to be less than perfect, as it almost always the case in practice [16,17], with mean entry times anticipated of, for example, $\frac{2}{3}A$ (corresponding to $\eta = 2$ in for $f(a)$ in equation (12)) rather than $\frac{1}{2}A$ for Uniform accrual, then the follow-up period F would have to be extended by around $\frac{1}{6}A$ to make up for the loss in patient exposure. In the example above, this would mean extending follow-up from 1 year to approximately 14 months. In this fashion, the impact of failing to meet assumptions of Uniform patient entry and/or exponential times to event can begin to be explored and explained.

3. SUGGESTIONS TO ENHANCE THE WAY WE DESCRIBE SAMPLE SIZE

3.1. So, what might we try do a little better?

In terms of communicating the issues and implications of sample size calculations, what we might do

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

a little differently? Perhaps a good start might be to routinely

(i) Point out that if a specific log HR advantage needs to be realized of at least θ , with a result less than θ not being clinically persuasive even if it reached statistical significance, then the need is to hypothesize not θ but $\theta' = (1 + z_\beta z_\alpha^{-1})\theta$ under the alternative.

From equations (1) and (5), if an observed advantage at least θ is required with $p \leq 0.05$ then it follows that

$$\theta' = (1 + z_\beta z_\alpha^{-1})\theta \tag{14}$$

needs to be hypothesized under the alternative. Thus, if a 15% risk reduction is required as in the example above, the need is to hypothesize not 0.85 but $0.85^{1.65} = 0.764$. The practical consequence of this is that fewer events are required; in fact

$$E' = E(1 + z_\beta z_\alpha^{-1})^{-2} \tag{15}$$

events are needed representing a saving of $100\{1 - (1 + z_\beta z_\alpha^{-1})^{-2}\}\%$ in trial size relative to hypothesizing θ. With respect to the example above, $E' = 582$, a saving of 63% in trial size. The implications of hypothesizing θ' as opposed to θ would need to be carefully articulated to the development team; the jump in expectation could be considered biologically implausible, even to such an extent as to render the entire trial non-viable. If so, reconsideration of the minimally desired outcome to be convincing will very likely be necessary.

(ii) Always provide θ_{crit} as in (5).

This is particularly informative when judging the merits of a range of trial size and power options. In the example above $\theta_{\text{crit}} = 0.906$.

(iii) Translate θ_{crit} into more meaningful terms by stating what this means in terms of the anticipated split of events between E and C.

From (3), and given N, θ and d, it follows that

$$d_C = \frac{N}{2}\left\{\left(1 - \frac{2d_C}{N}\right)^{e^\theta} - 1\right\} + d \tag{16}$$

with $d_E = d - d_C$. If the expected event rate is low, then (16) simplifies to provide the approximation

$$d_C \cong \frac{d}{e^\theta + 1} \tag{17}$$

Using the example above, substituting θ_{crit} for θ in (16) gives 832 events on C versus 759 events on E, a difference of 73 events. Hence, to achieve $p \leq 0.05$ an excess of at least 73 events needs to be observed on C relative to E. If the approximation in (17) is used, then the spilt is slightly over stated at 835 events on C versus 756.

3.2. Saying a little more about the calculation

Returning to the hypothetical dialogue between the medic and the statistician, given the recommendations above, it would be hoped that after '…11 800 patients will need to be randomised' the statistician would consider adding a little more along the following lines: '…though you should realise that in this trial I'm hypothesising a risk reduction of 15% which means that a lesser observed difference at the end of the trial would give $p \leq 0.05$; a risk reduction of at least 9.4%, corresponding to a difference in events of at least 832 (14.1%) versus 759 (12.9%), i.e. a difference of at least 73 (1.2%) events, would be significant. If the end result was actually a 15% risk reduction [corresponding to 855 (14.5%) versus 736 (12.5%) events, a difference of 119 (2%) events] the p-value would be well below 0.05; it would be around 0.0012. I've prepared a table with the details and a few other scenarios to look at…' Table I illustrates the kind of information that would most likely be of value to the medic and broader team.

This simple dialogue and illustration raises the question as to whether the non-statisticians in drug development, in particular medical, regulatory and commercial leaders and overall product development team leaders, conceptually appreciate what hypothesis testing and power really are, and how these concepts relate to sample size and the p-value. The situation is not helped by the regular and unfortunate use of loose language in the statistical community when referring to trials as being 'sized [or designed] to detect a difference of θ

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical STATISTICS

Table I. Trial size scenarios for differing assumptions.

Total number of events required, d, and θ_{crit}					The overall probability of an event, π_{bar}, and the total number of patients, N					The number (%) of events on experimental, E, and control, C, to give θ_{crit}		
1-sided α level (%)	Power (%)	HR hypothesized under H_1, θ	Number of events required, d	HR (RR) result, θ_{crit}, required to achieve $p \leq 2\alpha$	Expected fraction of control patients with an event at 1 year, q_t (%)	Accrual, A (yrs)	Minimum follow-up, F (yrs)	Prob. of an event over the trial period, π_{bar}	Total number of patients required, N	Number (%) of events on C	Number (%) of events on E	Difference (%) in events
2.5	90	0.750	508	0.840 (16.0%)	10	1	1	0.1264	4019	275 (13.7%)	233 (11.6%)	42 (2.1%)
2.5	90	0.800	844	0.874 (12.6%)	10	1	1	0.1307	6458	448 (13.9%)	396 (12.3%)	52 (1.6%)
2.5	90	0.850	1591	0.906 (9.4%)	10	1	1	0.1348	11803	832 (14.1%)	759 (12.9%)	73 (1.2%)
2.5	90	0.764	580	0.850 (15%)	10	1	1	0.1276	4545	312 (13.7%)	268 (11.8%)	44 (1.9%)

with 90% power at the 2-sided 5% significance level' rather than more correctly stating that the trial is 'sized [or designed] to test the hypothesis H_0: the true difference = 0 versus H_1: the true difference = θ with Type I and II errors of 5% and 10% respectively'.

In trying to help colleagues come to grips with these matters, it can be helpful to frame the problem in simple terms, perhaps as follows: It might be supposed that there is a very large jar with over 100 000 beads, say. Beads are black or white and the fraction of black beads is unknown. A statistician hypothesizes that the fraction is 10% (the null) and his medic colleague hypothesizes it is 30% (the alternative). A sample of 100 beads is (blindly) taken from the jar and the number of black beads noted. The beads are returned and the procedure repeated a further nine times, say. Now, if the 10 trials result in a consistently low number of black beads, say, mostly in the range 0–15, with a mean fraction of, for example, 12%, then it seems likely that the statistician is correct and the true fraction of black beads is 10%. However, if the trials result in a consistently higher number of black beads, say, mostly in the range 25–40, with a mean fraction of, for example, 28%, then it seems likely that the medic is correct and the true fraction of black beads is 30%. It's important to note that (i) due to the play of chance, trials with fewer than 30 beads and a mean fraction of less than 30% are not necessarily inconsistent with the medics view – one does not have to observe a fraction of exactly 30% (or better) to conclude the medic's hypothesis is most likely; (ii) there must therefore be critical pivot point for the observed fraction of black beads somewhere between 10% and 30% that favours the medic's hypothesis over the statistician's hypothesis. Intuitively this might be 20%; an observed fraction lower than this supporting the statistician's hypothesis, and higher the medic's. This pivot point is corresponds conceptually to θ_{crit} (as in equations (5)) and is why a trial does not have to yield a result equal to the hypothesized effect to give a positive ($p \leq 0.05$) outcome. See Figure 1 for a simple graphic that

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

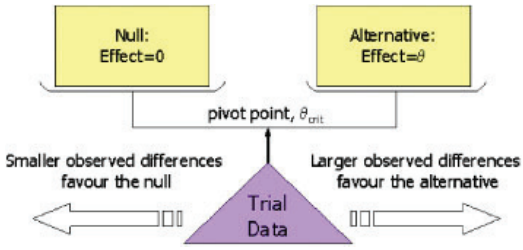


Figure 1. The weight of evidence: hypotheses, trial data and the critical value.

has proven useful in describing the basic idea to non-statisticians.

With respect to power, it may further be explained that if the true fraction of black beads in the jar is 30%, power is the number of times the observed fraction in repeated trials crosses the pivot point in favour of the medic (i.e. the alternative). The larger the sample size, the less likely the play of chance will throw up trials with a low number of black beads, so that higher sample sizes mean higher power.

Finally, there remains the question as to how large a sample should be taken? One answer might be not so small that you wish it would have been larger, and not so large than you know it could have been smaller. In practice it can be explained that the sample size is calculated by the statistician to control the probability of false-positive (concluding the fraction of black beads in the jar is 30% when the true fraction is actually 10%) and false-negative (concluding the fraction of black beads in the jar is 10% when the true fraction is actually 30%) findings at some pre-determined levels, typically 5% and 10%, respectively, corresponding to 5% (two sided) significance and 90% power. If the probability of false-positive and false-negative findings are set to be equal, then it follows from (14) that $\theta_{\text{crit}} = \theta/2$; if, as is more usual, false-positive and false-negative findings are fixed at 5% and 10%, $\theta_{\text{crit}} \approx 0.6 \times \theta$; whereas if they are fixed at 5% and 20%, $\theta_{\text{crit}} \approx 0.7 \times \theta$.

Taking a little time upfront to refresh memories on the key concepts of hypothesis testing, power and sample size is likely to prove invaluable further down the line. In particular, a common appreciation of these ideas is likely to be very useful when tackling the thorny and not infrequent problem of what to do when, several months into the trial, events appear to be accumulating at a lower rate than expected and concerns are growing that the target number of events, d, will not be achieved over the planned duration of the trial, $A+F$. How the statistician might help in working through this problem is the topic of the next section.

4. THE BEST LAID PLANS OF MICE AND MEN...

The trial discussed between the medic and statistician is underway. A total of 1591 events are targeted and 11 800 patients are to be recruited over 1 year at a rate of just fewer than 1000 patients per month, with an additional year of follow-up after the last patient is entered. However, a problem has arisen. Nine months into the trial, the (blinded) event rate has been assessed over the first 6 months of accrual for which up-to-date data are available (all patient visits over the first 6 months have been monitored and the presence/absence of events verified) and it is lower than expected. Of the 5900 patients randomized over the first 6 months, there are 104 events (1.77%). Based on initial assumptions, the statistician reveals the expected number of events at this time is 141 (2.38%). This represents a shortfall in the 1-year rate from 10% to 7.5%. Extrapolating forward, this means a total of 1200 events are now expected by the end of the trial. The question for the drug development team is what should be done? The options appear to be

1. Keep N at 11 800 but extend follow-up to compensate to achieve 1591 events – question for the statistician is thus how much should follow-up be extended?
2. Keep the overall trial duration at $A+F$ and but increase N to compensate to achieve 1591 events – question for the statistician is how much should N be increased?

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

3. Some combination of 1 and 2.
4. Do nothing. Check again in 6 months and hope the event rate has picked up.
5. Accept the lower event rate, do not change N or extend follow-up, and settle for the lower number of events, now expected to be 1200 – question for the statistician is what are the implications in terms of achieving a positive outcome for the trial?

Firstly, given r events out of N patients over some time period T, the event rate easily derived *via* (7) or (8) when the event rate is small. In practice, the Kaplan–Meier curve would also be drawn to show how events have emerged over time and a best-fit model could be fitted to predict how future events were likely to accrue [6]. However, for the purposes of what follows, exponentially distributed times to event are assumed as they were in the original sample size calculation.

With respect to Option 1, if the event rate is small and reduced by $100\omega\%$, then, from (8), to compensate minimum follow-up F must be increased by approximately $(\omega-1)^{-1}$ times the overall duration of the trial $A+F$, i.e.

$$F \to F + \left(\frac{A}{2} + F\right)(\omega - 1)^{-1} \tag{18}$$

Alternatively, with respect to Option 2, if the overall duration of the trial remains fixed, then N must be increased by approximately $100(\omega-1)^{-1}\%$ to $N(\omega-1)^{-1}$, i.e.

$$N \to N(1 - \omega)^{-1} \tag{19}$$

An intermediate question for the team is now which of Options 1 and 2 is worse? According to (18), minimum follow-up should increase by approximately $0.33 \times 18 = 6$ months, taking the overall duration from 24 to 30 months. On the other hand, (19) says the number of patients entered should increase by approximately 33%, i.e. by 3900 patients, to around 15 700 pts. In practice, an extension of follow-up is likely to be favoured over an increase in N if only due to issues relating to the feasibility of boosting recruitment in existing sites within the remaining accrual period, and logistical issues relating to the speed with which new sites could be added and made operational. A hybrid approach might be to increase N to 13 500 and follow for an additional 3 months which, based on the observed event rate, would deliver 1591 events by the end of the trial.

Option 4 is certainly a realistic option, given that relatively little data are available. However, in face of what appears to be a potentially serious shortfall in the event rate, it is unlikely little else would be done in practice. It is more likely that some change to N and/or minimum follow-up would be proposed with the event rate continuously monitored such that if it was to pick up, plans to increase trial size and/or follow-up of could be revisited.

Option 5 is likely to be viewed as least favourable by the typical project team. The resultant lost of power is often difficult to accept and external factors, like trial Steering Committees (usually consisting of the Principal Investigators and senior sponsor personnel) and other thought leaders in the scientific community, tend to want to meet the predefined event total. However, anxiety around settling for fewer events is often founded upon a poor appreciation of what is really being lost in terms of outcomes to yield significance or, conversely, what is really to be gained by pushing for the original target number.

Table II shows that while power is reduced to approximately 80%, there is actually little practical difference between 1591 and 1200 events in terms of the hypothesized effect to provide 90% power and θ_{crit}, the threshold value to achieve $p \leq 0.05$. With 1200 events, an observed HR of 0.893 (10.7% risk reduction) will provide a positive result, which is little different to the HR of 0.906 (9.4% risk reduction) for 1591 events. In terms of the anticipated spilt of events, these HRs translate to 832 (14.1%) versus 759 (12.9%) events, an absolute difference of 73 (1.2%) events, for 1591 total events as compared with 632 (10.7%) versus 568 (9.6%) events, an absolute difference of 64 (1.1%), events for 1200 total events. Further, with 1200 events, if the HR observed was 0.906, then the split of events would be 628 (10.6%) versus 572 (9.7%), a difference of

Copyright © 2009 John Wiley & Sons, Ltd.

Pharmaceutical
STATISTICS

Table II. Comparing the practical impact of achieving fewer events than originally planned for.

	Number of events at end of trial	HR (RR) result required to achieve $p \leq 0.05$	Number (%) of events on control	Number (%) of events on experimental	Difference (%) in events	Power for hypothesized HR = 0.85	Hypothesized HR for power = 90%
As originally planned	1591	0.906 (9.4%)	832 (14.1%)	759 (12.9%)	73 (1.2%)	90%	0.85
Accepting fewer events	1200	0.893 (10.7%)	632 (10.7%)	568 (9.6%)	64 (1.1%)	80.4%	0.83

56 (0.9%). Hence, while the additional 391 events from 1200 to 1591 buy an extra 10% power, they do not provide as great a gain in practical terms as one might at first think; these extra events provide a slack of just eight events in terms of the outcome to provide a positive, $p \leq 0.05$ result. Seen in these terms, the drug development team may now view Option 5 as a more realistic option in the face of a lower than expected event rate.

Figures 2 and 3 illustrate the degree of gain in accumulating events in an outcomes trial. From (1), once around 1000–1200 events are exceeded, Figure 2 shows the relationship between the 'detectable' HR and the number of events is shallow so that relatively large jumps in the number of additional events provide little extra in terms of the HR that can be detected with a given power, which is in line with the example above. Figure 3 perhaps illustrates the situation better, plotting the first derivative of (1) with respect to e^{θ} to give the rate of change in the HR as a function of total events. Again, it is clear that once around 1000–1200 events have been exceeded, the rate of change is close to zero, meaning again that substantial jumps in the number of additional events beyond around 1000 events are required to make meaningful inroads into the 'detectable' HR. The situation is of course much different for smaller numbers of events where relatively small increments have a larger impact.

5. SUMMARY AND RECOMMENDATIONS

Outcome trials are a very important component of drug development for new products across a number of therapeutic disease areas. They are typically very large-scale trials of long duration with hard clinical endpoints as outcomes, typically morbidity and mortality. As such, they are invariably expensive and high-profile trials with a critical impact on the clinical, regulatory and commercial fate of the drug. The drug develop-

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

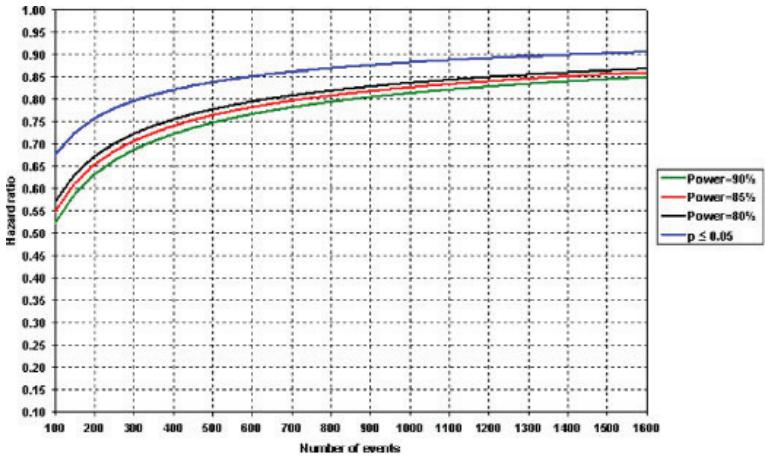


Figure 2. The 'detectable' hazard ratio versus total number of events.

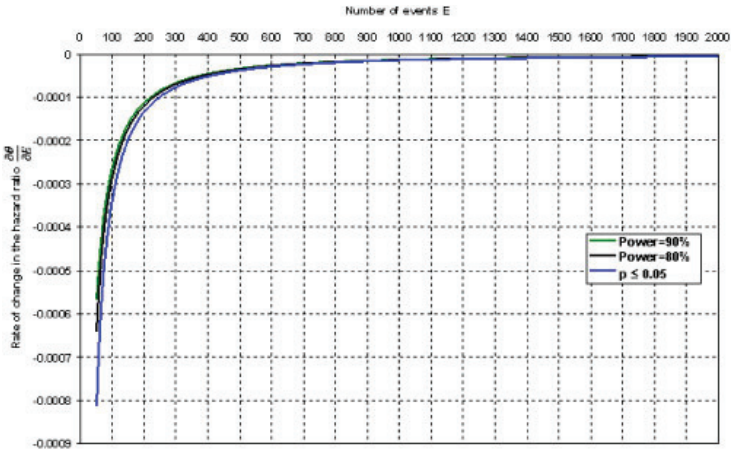


Figure 3. Rate of change in the 'detectable' hazard ratio versus total number of events.

ment statistician therefore has a vital role to play in the design, conduct, analysis and reporting of such trials. In particular, the statistician must do all they can at the outset to ensure that the sizing of the trial is fully appreciated by their medical, and other non-statistical, drug development team colleagues and that the risk of delivering a statistically significant but clinically unpersuasive result is minimized. Equally, the statistician has a key leadership role in guiding the team when a lower than anticipated event rate appears to be emergent.

In order to make the most effective contribution possible to the drug development team and process, it is therefore recommended that statisticians

1. Avoid the use of loose and imprecise language when describing basic sample size and power;

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

in particular, statisticians should avoid statements such as 'the trial is sized [or designed] to detect a difference of θ with 90% power at the 2-sided 5% significance level' and rather use more correct language such as ' the trial is sized [or designed] to test the null hypothesis H_0: the true difference = 0 versus the alternative H_1: the true difference = θ with Type I and II errors of 5% and 10%, respectively [or with a 5% 2-sided significance level and 90% power]'.

2. Take a little time upfront to refresh their colleagues appreciation of the true nature of hypothesis testing, sample size and power and, thus, help avoid confusion when the trial subsequently delivers a significant result for an observed outcome less than that which was hypothesized.
3. Always and routinely provide the critical value of the test, θ_{crit}, equation (5), when describing sample size and power.
4. Point out that if a specific advantage of, say, at least θ, needs to be realized to be persuasive in clinical, regulatory and commercial terms, with a result less than θ not being meaningful even if it reached statistical significance, then the need is to hypothesize not θ but rather $\theta' = (1 + z_\beta z_\alpha^{-1})\theta$, equation (14); and that this results in a reduction in the number of events required from E if θ is hypothesized to $E' = E(1 + z_\beta z_\alpha^{-1})^{-2}$ when θ' is hypothesized, equation (15); this being a reduction of $100\{1 - (1 + z_\beta z_\alpha^{-1})^{-2}\}$% in the required number of events. The implications of hypothesizing θ' as opposed to θ should to be carefully explained to the development team as the jump in expectation could be considered biologically implausible to such an extent as to render the entire trial non-viable. If so, reconsideration of the minimally desired outcome to be convincing will be necessary.
5. State that, if θ is hypothesized and is realized, the resultant p-value will be considerably less than $p = 0.05$; it will in fact be $p = 0.0012$, equation (6).
6. Translate θ_{crit} into more meaningful terms by stating what this means in terms of the anticipated split of events between E and C, equations (16) and (17).
7. Point out that, when the event rate is relatively low, and early blinded trial data suggest the event rate may be reduced by 100ω% relative to initial expectations, then to compensate either the minimum follow-up F must be increased by approximately $(\omega - 1)^{-1}$ times the overall duration of the trial, $A+F$, equation (18), or the target number of events must be increased by approximately $100(\omega - 1)^{-1}$%, equation (19).
8. State that, once around 1000–1200 events are achieved, the practical gain in accumulating further events in is marginal, Figure 3, such that substantial jumps in the number of additional events beyond 1000–1200 events are required to make a meaningful difference to the 'detectable' HR.

By following these simple recommendations it is hoped that statisticians involved in drug development will make an even more thorough contribution to the drug development team.

However, issues do, and will, remain. It could be argued that, despite best intentions, the drug development team cannot know at the outset the minimum degree of efficacy that must be observed to make for a persuasive result in the future. However, given the link between the threshold value for significance and the effect hypothesized under the alternative (as in (14)), if the minimum degree of efficacy to be observed to make for a persuasive result cannot be set in advance, then, logically, it is not possible to postulate a meaningful alternative hypothesis, since these quantities are two sides of the same coin. A similar sort of problem was gratefully highlighted by an anonymous reviewer who pointed out that while a clinician or commercial colleague might state at the outset what they feel is the minimally acceptable for a new drug, that is, what the new drug needs to 'deliver' at a minimum to be both clinically worthwhile and commercially viable, this 'target drug profile' often changes as the development programme unfolds. It is not uncommon for the target drug profile to demand high efficacy from a new drug at the outset, only for

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical STATISTICS

expectations to be lowered sometime later in response to accumulating data or changes in the competitive landscape. Then a lesser, yet still clinically meaningful, threshold for effectiveness may well be viewed as acceptable so that, in turn, a smaller point estimate for efficacy might become acceptable. If, in the end, the observed treatment effect was somewhat less than that stated as important at the outset but was nevertheless consistent with revised expectations around efficacy, the sample size will not be large enough to produce a significant *p*-value, and the drug development team may well be very unhappy. It is indeed possible for target drug profile expectations to change during a development programme, but, at the start, it is impossible to know if, or when, expectations will change, and, if they do, whether they will be weaker or stronger. All the statistician can reasonably do at the outset is to ensure that the appropriate dialogue takes place within the team to design a trial to deliver what is felt, at that time, to be a clinically and commercially persuasive result. If expectations change during the course of the trial, then the obvious and correct course is for the Team to reconsider the trial design, making adjustments to size and/or duration of follow-up as necessary.

Before closing, another common, event-related generic issue in outcome trials briefly worth a mention is the practice of analysing adverse event (AE) data as 'time to event data', where patients who withdraw early from the trial or die without experiencing the AE of interest are censored despite the obvious issues of informative censoring and competing risks. This is particularly problematic in outcome trials where drug is found to significantly improve overall survival relative to control. For example, if the 'time to event' analysis for a given AE provides a HR of 1.33, drug:control, and yet the HR for overall survival is 0.75, drug:control, how do we interpret the data? Given that patients are living longer on drug, they are obviously more likely to experience a higher proportion of other events than patients treated with control and, thus, the HR of 1.33 for the AE does not necessarily mean that the drug is associated in a 33% increase in the risk of the AE. It is therefore generally better, and arguably more meaningful and easier to interpret, to plan to look rather at 'event-free survival' for an AE, i.e. the time to the first of the AE or death. The resulting analysis is free from the complexity of competing risk and informative censoring due to death and can be meaningfully interpreted unconditionally, as the length of time patient is alive and free from the AE of interest, which is quantity likely to be of interest to both patients and physicians alike.

Overall, and notwithstanding the issues that remain, it is hoped that the statistician, in following the simple recommendations offered in this paper, will help their clinical, commercial, regulatory and other non-statistical drug development team leaders to better appreciate the nuances of outcome trial size determination and powering. It is therefore hoped the whole team will have a more complete appreciation of what the outcome trial will and will not deliver and, thus, will be better equipped to deal with those tricky and not infrequent situations when the initial event rate assumption seems perhaps to have been a little optimistic, requiring careful consideration of the options going forward.

ACKNOWLEDGEMENTS

The author would like to thank the anonymous reviewer for their helpful, thoughtful comments and advice.

REFERENCES

1. Stern MP. On the need for outcome trials in preventive pharmacology. *Diabetes Care* 1999; **22**(5):844–845.
2. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* 1966; **50**(3):163–170.
3. Peto R, Peto J. Asymptotically Efficient Rank Invariant Test procedures. *Journal of the Royal Statistical Society Series A* 1972; **135(**2):185–207.
4. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 1981; **68**:316–319.
5. Rubinstein RV, Gail MH, Santner JT. Planning the duration of a comparative clinical trial with loss to

Copyright © 2009 John Wiley & Sons, Ltd.
DOI: 10.1002/pst

Pharmaceutical
STATISTICS

follow-up and a period of continued observation. *Journal of Chronic Diseases* 1991; **34**:469–479.
6. Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. *Controlled Clinical Trials* 2003; **24**:682–701.
7. Sellke T, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika* 1983; **70**:315–326.
8. Schoenfeld DA. Sample size formula for the proportional-hazards regression model. *Biometrics* 1983; **39**:449–1983.
9. Morgan TM. Nonparametric estimation of duration of accrual and total study length for clinical trials. *Biometrics* 1987; **43**:903–912.
10. Cantor AB. Power calculation for the Log Rank Test using historical data. *Controlled Clinical Trials* 1996; **17**:111–116.
11. Rubinstein LV *et al*. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases* 1981; **34**:469–479.
12. Gross AJ *et al*. Sample size determination in clinical trials with an emphasis on exponentially distributed responses. *Biometrics* 1987; **43**:875–883.
13. Elashoff JD. *nQuery advisor version 4.0. Users guide*. Statistical Solutions Ltd.: Boston, 2000.
14. Bernstein D, Lagakos SW. Sample size and power determination for stratified clinical trials. *The Journal of Statistical Computation and Simulation* 1978; **8**: 65–73.
15. Yateman NA, Skene AM. Sample sizes for proportional hazards survival studies with arbitrary patient entry and loss to follow-up distributions. *Statistics in Medicine* 1992; **11**:1103–1113.
16. Williford WO *et al*. The 'Constant Intake Rate' assumption in interim recruitment goal methodology for multicenter clinical trials. *Journal of Chronic Diseases* 1987; **40**:297–307.
17. Hunninghake DB, Darby CA, Probstfield JL. Recruitment experience in clinical trials: literature summary and annotated bibliography. *Controlled Clinical Trials* 1987; **8**:6S–30S.

Copyright © 2009 John Wiley & Sons, Ltd.
DOI: 10.1002/pst

15. Ellis S, **Carroll KJ** and Pemberton K. Analysis of Duration of Response in Oncology Trials. Contemporary Clinical Trials, 2008; 29: 456–465

ScienceDirect
Contemporary Clinical Trials 29 (2008) 456–465
Contemporary Clinical Trials
www.elsevier.com/locate/conclintrial

Analysis of duration of response in oncology trials

Stuart Ellis [a,*], Kevin J. Carroll [b,*], Kristine Pemberton [c]

[a] *Independent Statistical Consultant, 31 Rowan Drive Cheadle Hulme, Cheshire, SK8 7DU, UK*
[b] *AstraZeneca Pharmaceuticals, CMOs Office, Alderley Park, Macclesfield, SK10 4TG, UK*
[c] *AstraZeneca Pharmaceuticals, Biostatistics Group, Macclesfield, UK*

Received 17 January 2007; accepted 29 October 2007

Abstract

The fraction of patients who respond to treatment and the duration of response in the subset of responding patients are commonly evaluated in oncology trials of cytotoxic compounds. While formal, comparative analysis of the fraction of patients responding to treatment is straightforward in a randomised trial, analyses that attempt to compare treatments in terms of the duration of response in responding patients are likely to be biased since the groups being compared are defined by the post-treatment outcome of response rather than by randomisation. Subsets of responding patients may not be comparable with respect to baseline prognostic factors and, consequently, formal comparative analysis is discouraged by the European Medicines Evaluation Agency. In an attempt to combine both the fraction of patients responding to treatment and the duration of response in responding patients, Temkin considered the probability of being in response function (PBRF) as a description of the treatment difference. Begg and Larson subsequently developed a parametric version of the PBRF under the exponential assumption. This paper briefly considers the PBRF as a means of estimating the expected duration of response across all randomised patients, thereby allowing a formal and unbiased comparison of treatments for duration of response. Building on earlier work, a more general and flexible approach to estimating the expected duration of response is offered to generalise beyond the exponential distribution.
© 2007 Elsevier Inc. All rights reserved.

Keywords: Duration of response (DoR); Probability of being in response function (PBRF); Expected DoR (EDoR)

1. Introduction

It is common in oncology trials of cytotoxic drugs that some patients will respond to treatment, as defined by some percentage reduction in tumour mass, say, as per RECIST criteria [1]. Responding patients will subsequently experience progressive disease (or die in the absence of progression) or reach the end of the trial without progression. Hence, the duration of response in responding patients will be known for some and censored for others. The fraction of patients who respond to treatment and the duration of response in responding patients are widely evaluated in oncology trials and both measures are considered clinically important determinants of therapeutic value by practising oncologists and regulatory authorities alike [2–5]. While comparative analysis of the fraction of patients responding to treatment is straightforward in a randomised trial, analyses that attempt to compare treatments in terms of the duration of response in responding patients are likely to be biased since the groups being compared are defined by the post-treatment outcome of

* Corresponding authors.
E-mail addresses: stuart.ellis@dsl.pipex.com (S. Ellis), kevin.carroll2@astrazeneca.com (K.J. Carroll).

1551-7144/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2007.10.008

response rather than by the randomisation schedule [6]. Subsets of responding patients may not be comparable with respect to baseline prognostic factors and, consequently, formal comparative analysis is discouraged by the European Medicines Evaluation Agency [5,7]. Another problem is that response rates and duration of response may show trends in opposite directions, for example a higher response rate in the control therapy but a longer duration of response in the experimental therapy. Such separate results may be difficult to interpret in terms of which treatment is preferred.

In an attempt to combine the fraction of patients responding to treatment and the duration of response in responding patients, Temkin considered the probability of being in response function (PBRF) [8]. Subsequently, Begg and Larson examined the PBRF using a simplified parametric model in which each of the different sojourn-time distributions is assumed to be exponential [9].

This paper attempts to build on this earlier work and is structured as follows: Section 2 considers the PBRF as a means of estimating the expected duration of response across all randomised patients, thereby allowing a formal and unbiased comparison of treatments for duration of response. In Section 3, an alternative, more general approach to estimating the expected duration of response is offered. Section 4 then closes with a summary of the ideas discussed and offers recommendations for the presentation and analysis of response data in Oncology trials.

2. The probability of being in response function (PBRF) and the expected duration of response

2.1. The expected duration of response

Temkin defined the PBRF as the fraction of patients who would be in response as a function of time after study entry if censoring were eliminated [8]. This provides a useful, descriptive way of visualising the difference between treatments in terms of the likelihood of being in response at any point during follow-up. The area under the PBRF, if available to infinity, measures the mean or expected duration of response (EDoR) across all patients and so could be used to formally compare treatments. Resampling methods could then be used to provide a confidence interval for the difference in expected response durations together with a p-value [10]. In practice, however, the PBRF is unlikely to be well defined at later follow-up times making estimation of the EDoR by area under the PBRF potentially unreliable.

Begg and Larson build on Temkin's approach by considering a stochastic process in which a patient must start in an initial state 0 and eventually progress to an absorbing state, 2 (progression or death in the absence of progression), possibly passing through a transient state, 1 (response) [9]. Response and duration of response are assumed to be independent and different sojourn-time distributions are assumed to be exponential so that transition between states is governed by constant hazards:

$$\lambda_1\colon 0\rightarrow 2, \quad \lambda_2\colon 0\rightarrow 1, \quad \lambda_3\colon 1\rightarrow 2,$$

Begg and Larson show the time to first event has an exponential distribution with hazard $\lambda_1+\lambda_2$ and the probability that this first event is a response is $\lambda_2/(\lambda_1+\lambda_2)$, being independent of the time the event occurred. Further, if $P(t)$ is the probability that a patient is in response at time t, it is shown that

$$P(t) = \lambda_2 \exp(-\lambda_3 t)\{1 - \exp(-\beta t)\}/\beta \quad \text{if } \beta = 0$$

$$t\lambda_2 \exp(-\lambda_3 t) \quad \text{if } \beta = 0$$

where $\beta=\lambda_1+\lambda_2-\lambda_3$.

Importantly, the area under $P(t)=\lambda_2/\{(\lambda_1+\lambda_2)\lambda_3\}$ is the EDoR based on all patients, not just the subset of responding patients. Hence, the EDoR, if estimated together with its standard error for both treatments in a randomised trial, could form the basis of a formal, unbiased comparison of treatments across all patients for their relative effect on response duration.

2.2. Estimation

To provide an estimate of the EDoR, define U to be the sum of all observed times to the first event, including the observed times of patients censored while in the initial state and let W be the sum of all response durations including the times of patients censored while in response. Then, let

n_1	number of patients who progress without response
n_2	number of patients who respond and are then censored
n_3	number of patients who respond and then progress
n_4	number of patients who are censored without response or progression

The maximum likelihood estimates (MLEs) of λ_1, λ_2 and λ_3 are provided by Begg and Larson; they are, respectively, n_1/U, $(n_2+n_3)/U$ and n_3/W. The estimated EDoR, $\hat{\text{E}}$DoR, say, is therefore given by $\frac{n_2+n_3}{n_1+n_2+n_3}\times\frac{W}{n_3}$.

Note that, in the formulation offered by Begg and Larson patients who remain in state 0 who neither progress

nor achieve a response (and thus are censored without response or progression at the end of the trial follow-up period), are considered uninformative and do not contribute when estimating the EDoR. Bearing this in mind, it can be seen that ÊDoR is a product of the estimated fraction of patients with a response, $p = \frac{n_2+n_3}{n_1+n_2+n_3}$, and the mean duration of response in responding patients, $\frac{W}{n_3}$.

Since the distribution of event times is assumed to be exponential, it is preferable to consider the EDoR on the log scale. Since $\ln(\text{EDoR}) = \ln(\lambda_2) - \ln(\lambda_1+\lambda_2) - \ln(\lambda_3)$, then the estimated variance of $\ln(\hat{\text{E}}\text{DoR})$ is given by $\hat{\text{V}}\text{ar}\left[\ln\left(\hat{\text{E}}\text{DoR}\right)\right] \cong \sum_{i=1}^{3} \text{Var}\left(\hat{\lambda}_i\right) \times \frac{\partial \ln(\text{EDoR})}{\partial \lambda_i}\Big|_{\lambda_i = \hat{\lambda}_i}$. Substituting the MLEs for λ_1, λ_2 and λ_3,

$$\hat{\text{V}}\text{ar}\left[\ln\left(\hat{\text{E}}\text{DoR}\right)\right] \cong \frac{1}{n_3} + \frac{1}{n_2+n_3} - \frac{1}{n_1+n_2+n_3} \qquad (1)$$

Thus, in a randomised trial comparing a new drug, E, with a control, C, the hypothesis that the EDoR is equal for E and C can be assessed by testing

$$H_0 : R = \frac{\text{EDoR}_E}{\text{EDoR}_C} = 1 \text{ vs. } H_1 : R = \frac{\text{EDoR}_E}{\text{EDoR}_C} \neq 1$$

using

$$z = \frac{\ln\left(\hat{R}\right)}{\sqrt{\hat{\text{V}}\text{ar}\left[\ln\left(\hat{R}\right)\right]}} = \frac{\ln\left(\hat{\text{E}}\text{DoR}_E\right) - \ln\left(\hat{\text{E}}\text{DoR}_C\right)}{\sqrt{\hat{\text{V}}\text{ar}\left[\ln\left(\hat{\text{E}}\text{DoR}_E\right)\right] + \hat{\text{V}}\text{ar}\left[\ln\left(\hat{\text{E}}\text{DoR}_C\right)\right]}} \qquad (2)$$

as the test statistic and comparing to a standard Normal (0,1) distribution.

2.3. An example

Suppose 200 patients are randomised to either treatment E or C on a 1:1 basis. Suppose further that, on treatment E, 70 patients progress without a response and the remaining 30 patients experience a response and, of these 30 responders, 10 are censored in response at the end of the trial follow-up period. Thus $n_1=70$, $n_2=10$, $n_3=20$ and $n_4=0$. The mean duration of response in the 30 responding patients is 6 months. Suppose further that the corresponding figures on C are 80 patients progressing without a response, 20 patients experience a response and, of these, 5 are censored in response. Thus $n_1=80$, $n_2=5$, $n_3=15$ and $n_4=0$. The mean duration of response in the 20 responding patients is 3 months. Hence, the estimated EDoR is $\frac{30}{100} \times 6 = 1.8$ for treatment E and $\frac{20}{100} \times 3 = 0.6$ for treatment C. The approximate standard errors (SEs) for the log of these estimates are $\sqrt{\frac{1}{20} + \frac{1}{30} - \frac{1}{100}} = \sqrt{0.0733}$ for $\ln(\hat{\text{E}}\text{DoR}_\text{E})$ and $\sqrt{\frac{1}{15} + \frac{1}{20} - \frac{1}{100}} = \sqrt{0.1067}$ for $\ln(\hat{\text{E}}\text{DoR}_\text{C})$. Therefore, $z = \frac{\ln(1.8) - \ln(0.6)}{\sqrt{0.0733+0.1067}} = 2.589$, so that the 2-sided p-value is 0.0096.

This example is extended in Table 1 to illustrate the interplay between the fraction of patients responding, the duration of response in responding patients and the EDoR.

3. An alternative approach to estimating the EDoR

3.1. Duration of response as mixture distribution

In the above formulation of the EDoR, event times are assumed to be exponentially distributed. However, it is possible to consider a more flexible approach that generalises to other distributions as follows. Let x represent duration of response and let p be the probability of response. Let $g(x)$ denote the probability density function of

Table 1
Examples of comparing treatments on the basis of the expected duration of response

Example	Treatment E (N=100) [a]					Treatment C (N=100) [a]					Comparison	
	Number responding	Number responding censored in response	Mean DoR [b] in responders (months)	EDoR [c]	SE [d] ln EDoR	Number responding	Number responding censored in response	Mean DoR in responders (months)	EDoR	SE ln EDoR	Ratio of EDoR and 95% CI [e]	p-value
1	30	10	6	1.80	0.271	30	7	3	0.90	0.258	2.00 (0.96, 4.17)	0.0641
2	30	15	6	1.80	0.300	20	5	6	1.20	0.327	1.50 (0.63, 3.58)	0.3606
3	30	15	6	1.80	0.300	20	5	3	0.60	0.327	3.00 (1.26, 7.16)	0.0132
4	30	15	6	1.80	0.300	40	5	3	1.20	0.209	1.50 (0.73, 3.07)	0.2672
5	30	15	3	0.90	0.300	20	5	6	1.20	0.327	0.75 (0.31, 1.70)	0.5165

[a] All non-responding patients are progressors; there are no patients without a response or progression, hence $n_4=0$ throughout.
[b] DoR = Duration of response in responding patients; exponential distribution assumed.
[c] EDoR = Expected duration of response.
[d] SE = standard error.
[e] CI = Confidence interval.

x, $f_p(x)$ denote probability density function of x in responding patients and $f_{1-p}(x)$ denote probability density function of x in non-responding patients. Then $g(x)$ is a simple mixture distribution such that $g(x)=pf_p(x)+(1-p)f_{1-p}(x)$. The EDoR is therefore $E_g(x)=pE_{f_p}(x)+(1-p)E_{f_{1-p}}(x)$ where $E_s(x)$ denotes expectation over $s(x)$. If the duration of response in non-responding patients is defined as zero, then $x=0$ with probability 1 and $f_{1-p}(x)=0$ for $x>0$ and hence $E_{f_{1-p}}(x)=0$. Consequently, the EDoR reduces to $pE_{f_p}(x)$ which is, again, the product of the estimated fraction of patients with a response and the mean duration of response in responding patients. If the EDoR is considered on the log scale, then $\ln(\text{EDoR})=\ln(p)+\ln(E_{f_p}(x))$ and, further, if $\hat{p}$ and $\hat{E}_{f_p}(x)$ denote the MLEs of p and $E_{f_p}(x)$, then

$$\hat{\text{Var}}\left[\ln\left(\hat{\text{EDoR}}\right)\right] \cong \frac{1}{\hat{p}^2}\hat{\text{Var}}\,(\hat{p}) + \frac{1}{\left[\hat{E}_{f_p}(x)\right]^2}\hat{\text{Var}}\left[\hat{E}_{f_p}(x)\right] \tag{3}$$

Consider as before a randomised trial comparing treatments E and C in N_E and N_C patients per group. Let p_E and p_C denote the true response rates and M_E and M_C the true mean DoR in responding patients. Then $R=\frac{\text{EDoR}_E}{\text{EDoR}_C}=\frac{p_E M_E}{p_C M_C}$ so that $\ln(R)=\ln\left(\frac{p_E}{p_C}\right)+\ln\left(\frac{M_E}{M_C}\right)$. Substituting estimates for p_E, p_C, M_E and M_C,

$$\ln\left(\hat{R}\right) = \ln\left(\frac{\hat{p}_E}{\hat{p}_C}\right) + \ln\left(\frac{\hat{M}_E}{\hat{M}_C}\right) \tag{4}$$

and, thus, from Eq. (3),

$$\hat{\text{Var}}\left[\ln\hat{R}\right] = \frac{1-\hat{p}_E}{N_E\hat{p}_E} + \frac{1-\hat{p}_C}{N_C\hat{p}_C} + \frac{1}{\hat{M}_E^2}\hat{\text{Var}}\left[\hat{M}_E\right] + \frac{1}{\hat{M}_C^2}\hat{\text{Var}}\left[\hat{M}_C\right] \tag{5}$$

Note the last two terms in Eq. (5) represent the variance of the estimated ratio of mean response durations in responding patients and applies when each treatment group is examined separately. The hypothesis that the EDoR is equal for E and C can be tested by:

$$H_0 : R=\frac{\text{EDoR}_E}{\text{EDoR}_C}=1 \text{ vs. } H_1 : R=\frac{\text{EDoR}_E}{\text{EDoR}_C}\neq 1$$

using

$$z=\frac{\ln\left(\hat{R}\right)}{\sqrt{\hat{\text{Var}}\left[\ln\left(\hat{R}\right)\right]}} \tag{6}$$

as the test statistic which is equivalent to Eq. (2) above.

3.2. Comparison to Begg and Larson

It is of interest to compare the estimate of the EDoR achieved via a mixture distribution to that achieved by Begg and Larson via a stochastic model. If $f_p(x)$ is exponential with hazard rate λ, then $E_{f_p}(x)=1/\lambda$ so that the EDoR $=p/\lambda$. If it is assumed that subjects who do not respond or progress are uninformative, then $\hat{p}=(n_2+n_3)/(n_1+n_2+n_3)$ and $\hat{E}_{f_p}(x)=n_3/W$, being the same as the estimate for λ_3 above. Hence, $\hat{\text{EDoR}}=(n_2+n_3)/\{(n_1+n_2+n_3)\}\times n_3/W$, which is the same as the estimate achieved by Begg and Larson. Further, since $\text{Var}(p)=\frac{1-p}{p(n_1+n_2+n_3)}$, then

$$\frac{1}{\hat{p}^2}\hat{\text{Var}}\,(\hat{p}) = \frac{n_1}{(n_2+n_3)(n_1+n_2+n_3)} = \frac{1}{(n_2+n_3)} - \frac{1}{(n_1+n_2+n_3)}$$

and $\hat{\text{Var}}\left[\ln\left(\hat{E}_{f_p}(x)\right]=\frac{1}{n_3}$, so that

$$\hat{\text{Var}}\left[\ln\left(\hat{\text{EDoR}}\right)\right] = \frac{1}{(n_2+n_3)} - \frac{1}{(n_1+n_2+n_3)} + \frac{1}{n_3}$$

which is, again, as per Begg and Larson.

3.3. Application in practice

Since the above formulation gives the EDoR to be the product of the fraction of patients with a response and the mean duration of response in responding patients, it is straightforward to compare treatments in practise as follows:

(i) Estimate p_E as $\frac{r_E}{N_E}$ and p_C as $\frac{r_C}{N_C}$, where r_E and r_C are the number of patients responding to E and C respectively.

Note the exclusion of patients who have neither responded nor progressed from the denominator when estimating the fraction of patients responding to treatment is problematic. For example, suppose of 50 patients are treated, 10 respond, 20 progress without a response and 20 neither progress nor respond. Typically in an oncology trial, the fraction of responding patients is calculated as the number responders divided by the number of patients treated. Thus, in the current example, this fraction would be 20% (10/50). If, however, those who neither progress nor respond during follow-up are excluded from the denominator, the apparent fraction of responding patients is increased, in this example to 33% (10/30). If all 40 of the non-responding patients did not progress during follow-up, the fraction of responding patients would be quoted as 100%. Excluding patients who neither progress nor respond during follow-up when estimating the fraction of patients responding to

treatment allows the number of patients in the denominator to be determined by a post-treatment event (or, more precisely, by the absence of an event). This approach results in an inflated estimate for the fraction of patients responding to treatment (and thus an inflated estimate for the EDoR) and should be avoided.

(ii) Estimate M_E and M_C and their standard errors.

This will depend on the probability distribution for the duration of response in responding patients, $f_p(x)$. Given a sample of responding patients, some of whom may be censored in response, estimates of M_E and M_C and their standard errors can be calculated separately for each treatment group for a range of commonly used time to event probability distributions such as the simple exponential, the Weibull, the gamma, the Normal and the log Normal. This is easily done using software such as SAS® PROC LIFEREG, where the mean duration of response together with its variance can readily be estimated for any member of the generalised gamma family of distributions [11]. Derivations for the Weibull and log Normal are provided in the Appendix.

(iii) Combine the estimates in (i) and (ii) to provide estimates of R and Var[ln(R)] and the difference between E and C is then assessed using Eq. (6).

3.4. A theoretical example

To illustrate the above approach, a 1000 patient dataset was simulated with patients randomised on a 1:1 basis to one of two treatment groups, E and C. Underlying time to response, duration of response and time to progression were simulated from exponential distributions with medians for the control group of 4, 3 and 5 months respectively. The maximum follow-up time was 10 months. Of those patients randomised to E, 253 responded and of those randomised C, 153 patients responded. Fig. 1 displays the PBRF which shows that initially patients treated with E are more likely to be in response than those randomised to C [8] although towards the end of the curves this is reversed. Response durations were analysed assuming exponential, Weibull and log Normal densities. Fig. 2a–c show the Kaplan–Meier curves for the duration of response in responding patients with curves assuming an exponential, Weibull and log Normal distributions for duration of response superimposed. Given the fraction of patients responding to each treatment, the estimated mean durations of response in responding patients and their associated standard errors, Table 2 shows how the data can be displayed in a simple, transparent fashion so that the calculation of the EDoR can be easily seen and understood. In this example, assuming exponentially distributed response durations, the EDoR is 3.1 months and 2.3 months for E and C respectively and comparing gives a ratio of 1.37, 95% CI (0.98 to 1.90), $p=0.065$. An alternative analysis assuming a Weibull distribution gives a broadly similar result. As evident in Fig. 2c, the log Normal distribution provides a poor fit to the data and consequently overestimates the EDoR.

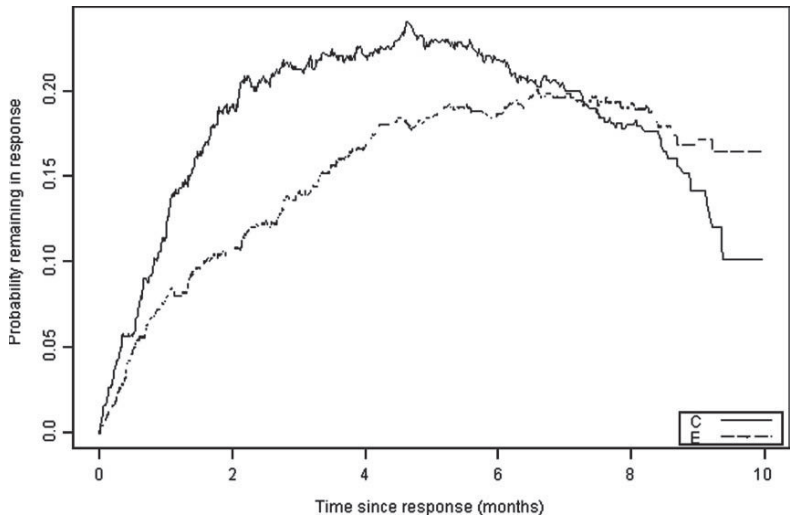


Fig. 1. Probability of being in response as a function of follow-up time. Example data.

3.5. A real example

Response and duration of response were measured in a randomised, double-blind trial of gefitinib + doublet chemotherapy vs. placebo + doublet chemo therapy in the treatment of advanced lung cancer [12]. A total of 345 patients were randomised to 250 mg gefitinib, 347 patients to 500 mg gefitinib and 345 patients to placebo. For the purposes of illustration, we shall focus on the 500 mg and placebo arms. The data are shown in Table 3, together with the EDoR in all patients assuming exponential, Weibull and log Normal densities. The PRBF for 500 mg vs. placebo is shown in Fig. 3 and Kaplan–Meier curves for response duration in responding patients are

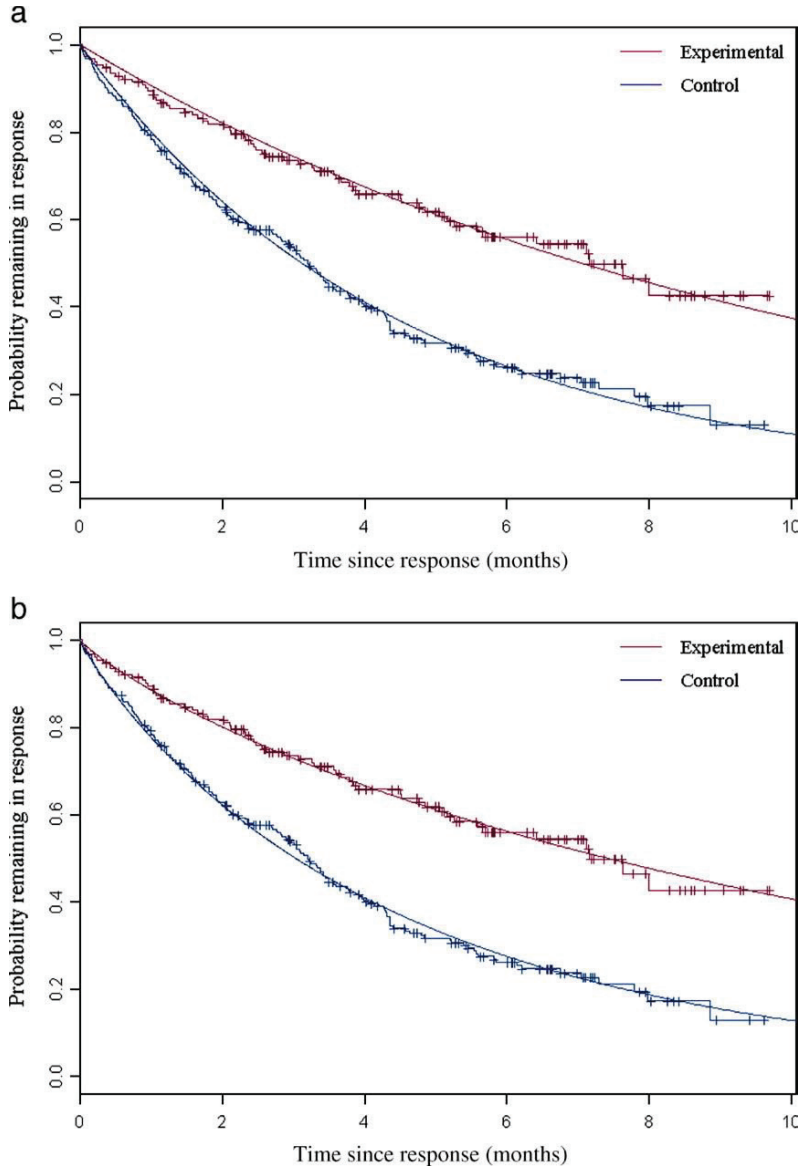


Fig. 2. (a) Duration of response for responding patients with an exponential distribution: Simulated example data. (b) Duration of response for responding patients with a Weibull distribution: Simulated example data. (c) Duration of response for responding patients with a log Normal distribution: Simulated example data.

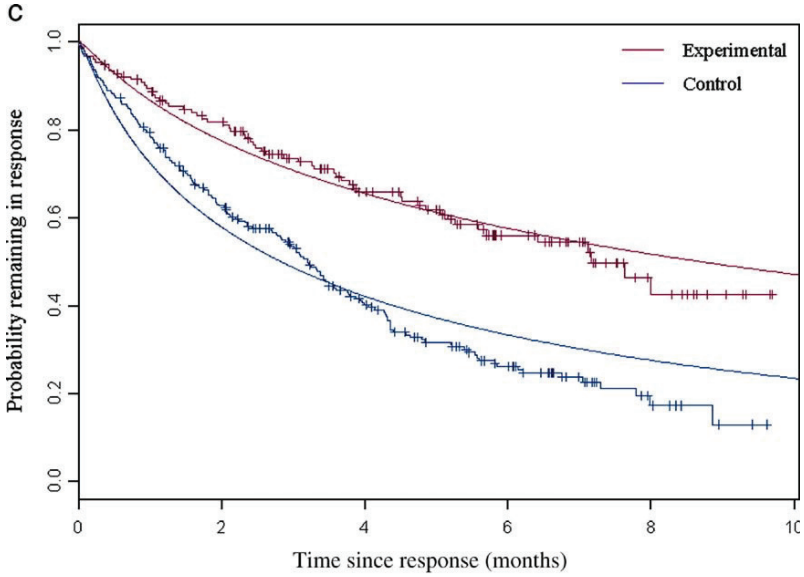


Fig. 2 (*continued*).

shown in Fig. 4a–c. It is clear from these data that the exponential distribution is a poor fit and therefore does not provide a reasonable estimate of mean DoR in responding patients, and thus, the EDoR. The Weibull and log Normal distributions are clearly better, with the log Normal fitting the latter part of the survival curves a little better than the Weibull. For the Weibull, EDoR is estimated to be 40 days on placebo and 53 days on gefitinib. The difference between treatments approaches significance with an EDoR ratio of 1.32, 95% CI (0.98 to 1.78), $p=0.07$. For the log Normal, EDoR is estimated to be 42 days on placebo and 62 days on gefitinib. The difference between treatments now reaches significance with an EDoR ratio of 1.49, 95% CI (1.03 to 2.16), $p=0.04$.

4. Summary and recommendations

The major concern with comparing the duration of response between treatments in oncology trials is that the comparison is frequently based on a subset of patients determined after randomisation. Subsets of

Table 2
Comparison of treatments for expected duration of response using simulated data

	Exponential		Weibull		Log normal	
	Treatment *C* (*N*=500)	Treatment *E* (*N*=500)	Treatment *C* (*N*=500)	Treatment *E* (*N*=500)	Treatment *C* (*N*=500)	Treatment *E* (*N*=500)
Response rate, % [1]	50.6%	30.6%	50.6%	30.6%	50.6%	30.6%
Mean DoR[a] [2]	4.5	10.2	4.7	12.1	12.9	57.6
SE[b] DoR	0.08	0.13	0.09	0.22	0.23	0.53
EDoR[c] [1]x[2]	2.27	3.11	2.40	3.71	6.57	17.62
Ratio of EDoR and 95% CI[d]	1.366 (0.981 to 1.903) *P*=0.065		1.547 (0.944 to 2.534) *P*=0.08		2.681 (0.859 to 8.365) *P*=0.09	

[a]DoR = Duration of response in responding patients, months.
[b]SE = standard error.
[c]EDoR = Expected duration of response, months.
[d]CI = Confidence interval.

Table 3
Gefitinib vs. placebo, INTACT 2. Comparison of treatments for Expected Duration of Response using exponential, Weibull and log Normal densities

	Exponential		Weibull		Log Normal	
	Gefitinib $N=347$	Placebo $N=345$	Gefitinib $N=347$	Placebo $N=345$	Gefitinib $N=347$	Placebo $N=345$
Response rate, % [1]	30.6%	29.9%	30.6%	29.9%	30.6%	29.9%
Mean DoR[a] [2]	221.6	148.8	173.7	134.7	202.6	139.5
SE[b] DoR	0.137	0.115	0.083	0.057	0.131	0.074
EDoR[c] [1]x[2]	67.7	44.4	53.1	40.2	61.9	41.7
Ratio of EDoR and 95% CI[d]	1.524 (1.003 to 2.313) $P=0.048$		1.320 (0.977 to 1.783) $P=0.07$		1.486 (1.025 to 2.155) $P=0.04$	

[a]DoR = Duration of response in responding patients, days.
[b]SE = standard error.
[c]EDoR = Expected duration of response, days.
[d]CI = Confidence interval.

responding patients may not be comparable with respect to important prognostic factors and, consequently, a fair and unbiased comparison of treatments cannot be guaranteed. Furthermore, response rate and duration of response are particularly difficult to interpret if, for example, one treatment gives a high response rate with a short duration of response but the other treatment gives a low response rate but those patients that do respond achieve a durable, long lasting response. The question as to which is the better drug is then unclear.

The approach discussed in this paper has been to build on earlier work that attempted to tackle this problem by combining information on the response rate with information on the duration of response in responding patients. By considering the problem in terms of a simple mixture distribution, the expected duration of response can be calculated based on all patients and not only those who responded. A formal comparison of treatments can then be affected by assuming some underlying distribution for the duration of response in responding patients. The suitability of the assumed probability distribution can be assessed using published diagnostics. To avoid post-hoc choice of the probability distribution and the associated criticisms, the choice should be made based upon grouped, overall data prior to unblinding or data analysis.

Therefore, in situations where a formal statistical comparison of treatments is desired in terms of impact on duration of response, it is recommended that

(i) the PBRF is used to display the data
(ii) response rates are provided with N_E and N_C used in the denominator

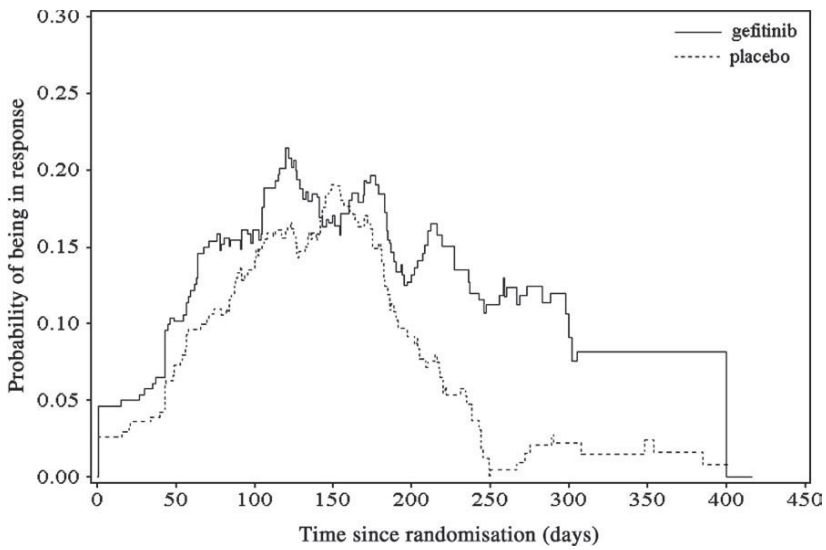


Fig. 3. Probability of being in response as a function of follow-up time: gefitinib 500 mg vs. placebo, INTACT 2.

(iii) descriptive data are provided for the duration of response in responding patients, including the associated Kaplan–Meier curves (without any formal comparison or p-value attached)
(iv) the expected duration of response is derived as the sole basis to formally compare treatments statistically and
(v) the data are laid out as per the example provided in Table 2.

In following these recommendations the aim is to ensure treatments are formally compared not in the subset of patients who responded but on the basis of all randomised patients and, further, to ensure the data are

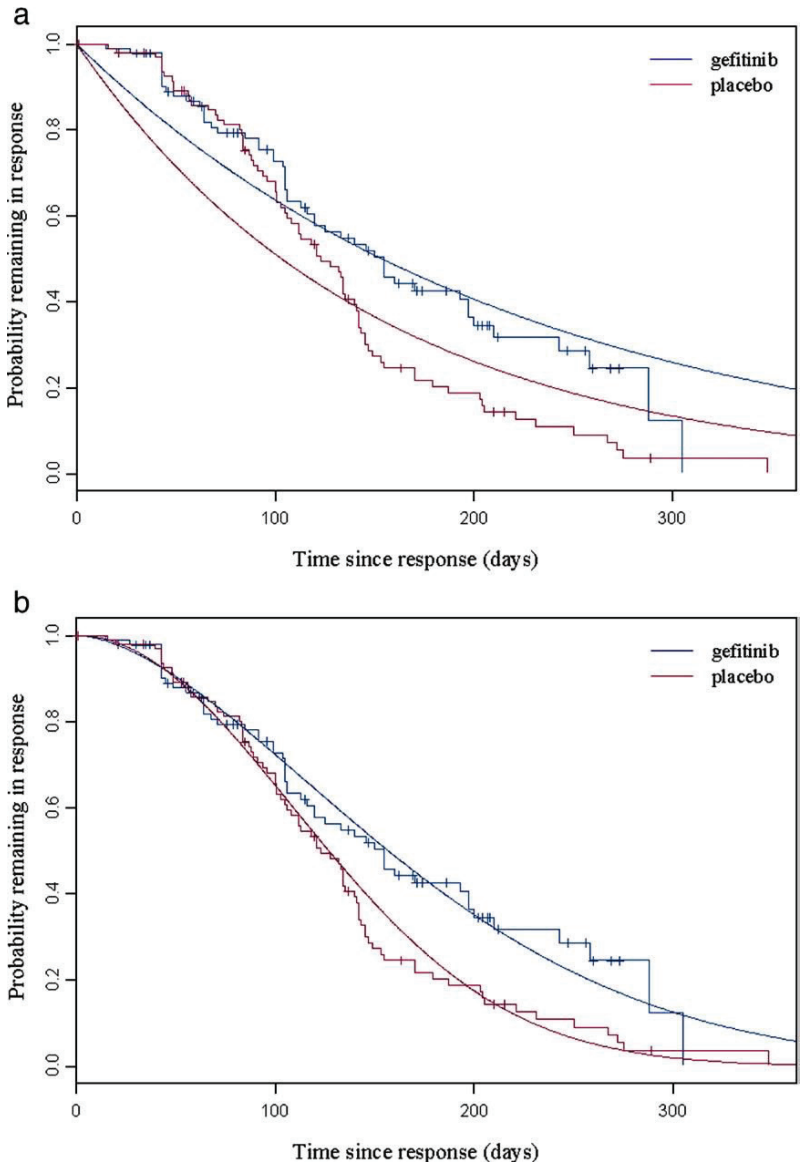


Fig. 4. (a) Duration of response for responding patients with an exponential distribution: gefitinib 500 mg vs. placebo, INTACT 2. (b) Duration of response for responding patients with a Weibull distribution: gefitinib 500 mg vs. placebo, INTACT 2. (c) Duration of response for responding patients with a log Normal distribution: gefitinib 500 mg vs. placebo, INTACT 2.

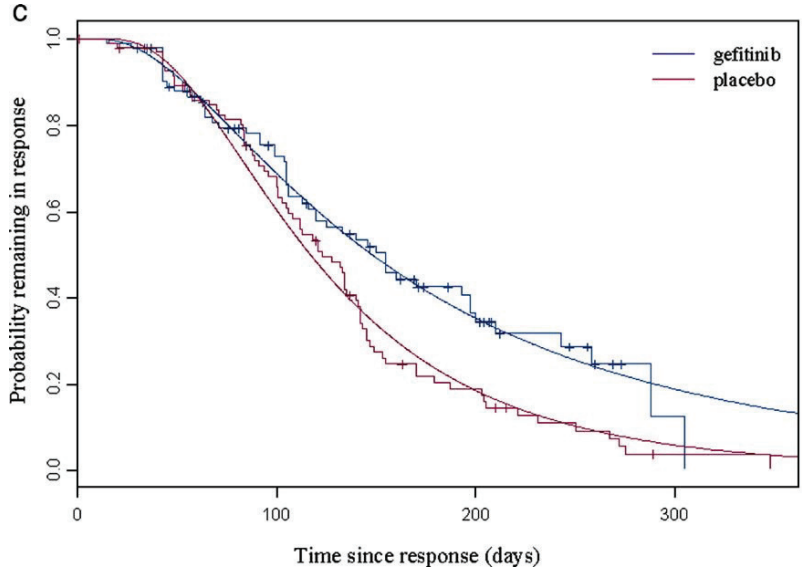


Fig. 4 (*continued*).

presented and displayed in a transparent and intuitive fashion so that statisticians and non-statisticians alike can better appreciate the relative difference between treatments.

Appendix. Derivation of mean duration of response and estimated variance for selected distributions when analysing using SAS® PROC LIFEREG

For the patients who have responded the SAS® procedure LIFEREG provides a straightforward means of analysing the duration of response assuming a range of common distributions. The procedure provides parameter estimates, typically the intercept, μ, and scale, σ, together with the estimated covariance matrix. Depending on the distribution selected for analysis, these parameter estimates can be combined to provide estimates of the mean duration of response together and its variance as follows:

(i) For the Weibull distribution, $f(t)=\lambda\alpha t^{\alpha-1}e^{-\lambda t\alpha}$, LIFEREG parameterises such that $\alpha=\frac{1}{\sigma}$ and $\lambda=e^{-\frac{\mu}{\sigma}}$ so that mean duration of response is estimated as $e^{\hat{\mu}}\Gamma(1+\hat{\sigma})$ with $\hat{\mathrm{V}}\mathrm{ar}\left[\ln\left\{e^{\hat{\mu}}\Gamma(1+\hat{\sigma})\right\}\right]=\mathrm{Var}(\hat{\mu})+\{\mathrm{digamma}(1+\hat{\sigma})\}^2\mathrm{Var}(\hat{\sigma})+2\times\mathrm{digamma}(1+\hat{\sigma})\mathrm{Cov}_{(\hat{\mu},\hat{\sigma})}$.

(ii) For the log Normal distribution, $f(t)=\frac{1}{\sqrt{2\pi}\sigma t}e^{-\frac{1}{2\sigma^2}\left\{\frac{\ln(t)-\mu}{\sigma}\right\}^2}$, the mean duration of response is estimated as $e^{\hat{\mu}+\frac{1}{2}\hat{\sigma}^2}$ with $\hat{\mathrm{V}}\mathrm{ar}\left[\ln\left\{e^{\hat{\mu}+\frac{1}{2}\hat{\sigma}^2}\right\}\right]=\mathrm{Var}(\hat{\mu})+\hat{\sigma}^2\mathrm{Var}(\hat{\sigma})+2\,\hat{\sigma}\mathrm{Cov}(\hat{\mu},\hat{\sigma})$.

References

[1] Therasse P, Arbuck S, Eisenhauer E, et al. New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst 2000;92:205–16.

[2] FDA. Oncologic drugs advisory committee meeting proceedings; March 12–13, 2003. http://www.fda.gov/ohrms/dockets/ac/cder03.html#OncologicDrugs.

[3] Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics; 2007. http://www.fda.gov/cder/guidance/7478fnl.pdf.

[4] Johnson J, Williams G, Pazdur R. End points and United States Food and Drug Administration, approval of oncology drugs. J Clin Oncol 2003;21:1404–11.

[5] CHMP. Guideline on the evaluation of anticancer medicinal products in man; December 2005. London http://www.emea.eu.int/pdfs/human/ewp/020595en.pdf.

[6] Schroder T, Schumaker M. Methodoligical. Problems in evaluating duration of response to therapy in cancer clinical trials. Onkologie 1997;20:393–9.

[7] CHMP. Points to consider on adjustment for baseline covariates; May 2003. London http://www.emea.eu.int/pdfs/human/ewp/286399en.pdf.

[8] Temkin NR. An analysis for transient states with application to tumour shrinkage. Biometrics 1978;34:571–80.

[9] Begg BB, Larson M. A study of the use of the Probability-of-being-in-response function as a summary of Tumour response data. Biometrics 1982;38:59–66.

[10] Westfall P, Young S. Resampling based multiple testing: examples and methods for *p*-value adjustment. New York: Wiley; 1993.

[11] SAS/STAT User's Guide, 4th ed., vol. 2. Cary, NC: SAS Institute Inc.; 1989. Version 6.

[12] Herbst RS, Giaccone G, Schiller JS, et al. Gefitinib in combination with paclitaxel and carboplatin in advanced non-small-cell lung cancer: a phase III trial- INTACT 2. J Clin Oncol 2004;22:785–94.