Randomized Phase II Designs when Phase III is Based on Overall Survival: Comparing Rank-Based with Progression-Free Survival Designs

Tian Chen[1], Yun Zhang[2], Weichao Bao[3], Fei Ma[3], Yunro Chung[4], William L. Mietlowski[3]
[1] University of Toledo, Toledo, OH
[2] University of Rochester, Rochester, NY
[3] Novartis Pharmaceutical Corporation, East Hanover, NJ
[4] University of North Carolina at Chapel Hill, Chapel Hill, NC

Abstract

Gan (2013) reported that approximately 27% of 120 randomized Phase III trials with a primary end- point of OS had statistically significant outcomes. Ratain (2005) suggested that the low Phase III success rate in Oncology may stem from a low positive predictive value (PPV) in Phase II trials. We proposed two rank-based (RB) randomized Phase II designs that differentially weigh the risk of death by the type and time of disease progression and the percentage change in tumor burden. Both RB designs utilized the Wei-Lachin test (Lachin 1992); one (MI-WL) utilized multiple imputation prior to applying the Wei-Lachin test (Mogg and Mehrotra 2007), the other did not (WL). We then compared these designs with one based on progression-free survival (PFS) by simulating 2500 randomized Phase II studies from Phase III trials with known OS outcome with respect to sensitivity, specificity, and PPV.

The MI-WL test had the greatest PPV among the three methods considered. The increase relative to PFS reflected a gain in specificity, whereas the increase relative to the WL method was due to increased sensitivity. The decreased specificity of the WL method may be due to a biased missingness which is addressed by the MI-WL method.

Key Words: randomized Phase II, overall survival, progression-free survival, positive predictive value, Wei- Lachin test, multiple imputation

1. Introduction

Overall survival (OS) is regarded as the ”gold standard” for demonstrating clinical benefit for Phase III oncology (Ellis et al 2014). However the rate of statistically significant treatment effects in OS-based Phase III trials is only approximately 27% (Gan, 2013). Ratain (2005) proposed that a low positive predictive value (PPV) in Phase II may be a contributing factor to the low Phase III success rate. With a success rate of 27%, the specificity of a Phase II trial may be a major determinant of the PPV.

Because OS as the primary efficacy end point generally requires more events and longer periods of follow-up, due to time limitations, most randomized Phase II studies do not rely on an OS endpoint but on an earlier intermediate endpoint that is considered to be correlated with OS. Progression-free survival (PFS), which measures the time from randomization to documented tumor progression or death, has been advocated as the standard Phase II endpoint.

Furthermore, there is emerging evidence that the components of the tumor measurement process (occurrence of new lesions, non-target progressive disease (PD), change in target lesion tumor burden, baseline target lesion tumor burden) may have independent prognostic value for OS [5-8]. Table 1 at the end of the manuscript shows the findings and summary information for these 4 independent publications. We note that PFS only measures the the time of documented disease progression, ignoring the type of disease progression, the extent of change of the target lesion tumor burden and post-progression overall survival information.

Motivated by the low success rate in OS-based Phase III trials, we proposed a rank based approach which integrates the information from the components of the tumor assessment process and overall survival, including post-progression overall survival. It may be more informative than PFS and possibly enable an earlier Phase II assessment (before PFS data is mature).

We describe construction of the rank-based tests and the impact of missing data on the analysis in Section 2. We describe the datasets used to evaluate the diagnostic properties of the rank-based tests compared to PFS and the design of the simulations in Section 3 and present the results of the simulations in Section 3. Section 4 discusses our findings.

2. Rank-Based Designs and Tests

2.1 Proposed Rank-Based Design

We propose to combine overall survival status with tumor status at each scheduled tumor assessment. Based on the tumor assessment schedule, we construct disjoint visit windows with a half schedule allowance for a delayed assessment. For example, if day 1 represents the day of randomization and tumor assessments are every 8 weeks, the first visit window covers day 1 to day 85, the second visit window covers day 86 to day 141, etc. We combine the overall survival status and tumor assessment information at each visit window and calculate a score based on this information. The earliest death has the lowest score and a patient with the maximal tumor shrinkage at the tumor assessment for the current visit window has the highest score. Figure 1 presents a schematic diagram of the ordinal score calculation. Rank-based methods (Wei-Lachin test [9,10]) are then applied to the resultant repeated measures ordinal data.

2.2 Impact of Missing Values

There are two types of missingness. Patients may be missing partial data but a score can be computed for the interval using the midrank principle. The midrank principle applied to missing data is an average of the best possible and worst possible rank. For example, if some baseline target lesions are not measured at follow-up, the best possible rank is that the non-measured lesions are absent and the worst possible rank scenario assumes that the non-measured lesions have the largest possible growth.

Another example involves censoring of overall survival during the interval but the tumor assessment data for the interval is known. This case uses weighted midrank imputation wherein the best case is the patient is alive at the end of the interval and the worst case is the patient is dead at the censoring date and the weight is the conditional probability of surviving to the end of the interval (BCW) and (1-BCW), respectively, assuming overall survival has a Weibull distribution.

The second type of missingness is an entirely missing score during the interval due to staggered entry and administrative censoring of overall survival or tumor assessment. Such missing data arises in two ways in the proposed rank-based procedure. One occurs when overall survival is censored before the start of a visit window and the other occurs when the tumor assessment information for an ongoing patient is censored, i.e. the patient is scheduled to have a tumor assessment during the interval, but the data cutoff occurs prior to the scheduled tumor assessment.

We note that the missing value process may be biased and differential against treatments increasing overall survival, leading to potential decreased sensitivity for the Phase II rank-based design. Effective therapies may have fewer deaths and disease progressions with more patients remaining on treatment, i.e. more potential for missing values in healthier patients. We performed two kinds of analyses to investigate the potential biased and differential missingness. We assumed that the patient's percentile from a combined ranking at the previous visit (or last non-missing visit) measures the patient's relative health. We compared the mean previous percentiles for missing and non-missing patients by visit and treatment and modeled the probability of missingness with previous percentile as a covariate using logistic regression.

To address the potentially biased and differential missingness, we use the multiple imputation (MI) technique with 50 completed datasets [11,12] before applying the Wei-Lachin rank-based test (MI - WL) [13]. We then compare it to the Wei-Lachin test without the use of multiple imputation (WL). We model the logit of the percentile of the current visit as a function of the logits of the percentiles from the previous visits.

3. Simulation

In this section, we evaluate the performance of our proposed rank-based design using the randomized Phase II trials as a diagnostic test for a Phase III trial based on OS and compare it to PFS. All results are based on 2500 iterations.

3.1 Data Preparation

The sample size needed for our proposed rank-based design is a total of 150 patients (75 per group) to achieve 80% power for a one-sided test at the 10% level of significance at a Mann-Whitney probability alternative of 0.6 or equivalently, Mann-Whitney odds of 1.5 [14]. On the other hand, PFS requires 110 events to achieve 80% power for a one-sided test at the 10% level of significance at an alternative hazard ratio of 0.67 [15].

By finding databases for Phase III trials designed to detect differences in OS which have a known OS outcome (statistically significant treatment effect or not), we can simulate thousands of randomized Phase II trials of size 150 from these Phase III trials and compare the specificity, sensitivity and positive predictive value of the rank-based randomized Phase II trials vs. those based on PFS.

We had access to the raw data for a Novartis Phase III trial with a primary endpoint of OS that had virtually no OS differences between the study medication and comparator. This was the CONFIRM-2 trial which compared the VEGFR inhibitor PTK787 + FOLFOX4 with FOLFOX4+ placebo in second-line metastatic colorectal cancer (HR=1.00, n=855)[16]. This negative Phase III study was used for assessment of specificity.

There are two more design elements required to evaluate specificity. They are the data analysis cutoff date and the accrual distribution. Since Phase II studies should not unduly delay the start of Phase III, we propose an early time-driven data analysis cutoff date for the rank-based analyses. We require all patients to have at least one post-treatment tumor assessment, allowing for delays in assessment of up to 50% of the time between the first and second post-treatment. For CONFIRM-2, tumor assessments were every 8 weeks, so the data analysis cutoff date is date of randomization of the 150th patient + 12 weeks.

We investigated four accrual distributions.The first was a piecewise uniform distribution of 10 patients per month for months 1-5 and 25 patients per month for months 6-9. This approximates the accrual pattern observed in an actual Phase II trial in 2nd line colorectal cancer. The other three are all uniform accrual distributions with enrollments of 25, 15 and 10 patients per month for 6, 10 and 15 months, respectively.

To determine the sensitivity and positive predictive value, of the randomized Phase II study as a diagnostic test, we need to find the individual patient data in the same indication for Phase III studies with a significant treatment effect on overall survival. The external trial of bevacizumab + FOLFOX4 vs. FOLFOX4 alone in second line metastatic colorectal cancer did achieve a statistically significant treatment effect in overall survival (hazard ratio=0.75, p=0.0011) [17]. By using the midrank transformation [18] and oversampling patients with an overall survival beyond the median (58% of longer survivors allocated to the experimental arm, 42% allocated to the control arm) , we created a Phase III trial (CONFIRM-2 P1) from the CONFIRM-2 data set that had OS hazard ratios similar to study referenced [17]. A summary of this study is listed in Table 2. Figure 2 describes the oversampling process in a schematic fashion.

3.2 Results

3.21 Diagnostic operating characteristics

As Figure 3 indicates, both rank-based procedures appear to be relatively stringent

with specificities ranging from 0.91-0.93 for the MI-WL test and 0.88-0.90 for the WL test compared to 0.73-0.77 for PFS.

With regard to sensitivity under the CONFIRM-2 P1 positive trial, the Wei-Lachin test appears to have a marked decrease (0.16-0.20) compared with PFS (Figure 4). This may reflect a differential and biased missingness in OS-positive trials. If multiple imputations are applied to the repeated measurements data to address missingness issues and the Wei-Lachin test subsequently applied to the completed data, the two-stage procedure (MI-WL) is more similar in sensitivity to PFS (decrease of 0- 0.12) (Figure 4).

Assuming a prevalence of 27% for a statistically significant treatment effect in a Phase III trial based on OS, Figure 5 shows that the MI-WL procedure may have a marked increase in PPV relative to PFS (0.21-0.27). The MI-WL procedure also has an increased PPV compared to the WL test (0.07-0.15).

3.22 Time to data cutoff for analysis

Table 3 summarizes the Phase II study duration and number of PFS events based on the data analysis cutoff dates of last patient randomized + 84 days and 110 PFS events for the 2500 simulated randomized Phase II trials. There is an approximate two to four month earlier mean study completion date for the rank-based methods representing decreases of 9 to 31 per cent.

3.23 Analysis of missingness

To investigate biased missingness, we compared the mean previous percentiles for missing and non-missing patients by visit and treatment assuming piecewise enrollment (Figure 6). We found that mean previous percentile is greater in missing patients than non-missing patients for all visits and both treatment groups. Mean previous percentile is greater in treated patients than in control patients for both missingness groups. Notice that the degree of missingness can be substantial, especially at the later visits. We also modeled the probability of missingness with previous percentile as a covariate using logistic regression assuming piecewise enrollment (see Table 4). Previous percentile was a significant predictor of missingness in more than 97.5% of simulations and adding the treatment in the logistic regression model had virtually no additional effect. All these results illustrated the presence of biased missingness and justified the use of history of previous percentiles to predict current percentile in the multiple imputation model.

Figures 7 and 8 present boxplot displays of the distributions of the Mann-Whitney probabilities over the 2500 simulations by enrollment rate in descending order of missingness (approximately 70% at the last visit to approximately 40% at the last visit) for the MI-WL and WL methods for CONFIRM-2 and CONFIRM-2 P1, respectively. The WL method has decreasing variability as missingness decreases while MI - WL has considerable robustness over enrollment. This may suggest a good model fit with the MI model.

4. Discussion

We have proposed rank-based endpoints for randomized Phase II Oncology trials where the Phase III design is based on overall survival. These may be more informative than the traditional intermediate endpoints of progression-free survival and changes in target lesion tumor burden. They incorporate the time and type of disease progression (including death without documented progression, occurrence of new lesions, clear worsening of non-target disease, change in target lesion tumor burden), and post-progression overall survival information. At each tumor assessment, deaths and patients with early progression are given lower ranks and patients with ongoing tumor response are given higher ranks. Since the prevalence of statistically significant treatment effects in Phase III trials based on overall survival is 27%, specificity will be the major determinant of positive predictive value. The rank-based methods appear to have good specificity (90%) in the CONFIRM-2 trial. If the components of the tumor measurement process are independent prognostic factors for overall survival in an indication, then a lack of treatment effect on overall survival implies a lack of treatment effect on the components of tumor measurement (logical contrapositive). Thus, there might be greater reproducibility of specificity and PPV for the rank-based methods than in the CONFIRM-2 case study.

On the other hand, we lacked the raw data for an actual OS-positive trial, so we simulated a potential OS-positive trial from CONFIRM-2 (i.e. CONFIRM-2 P1). Further investigations of the MI-WL method are warranted especially in Phase III trials with a positive overall survival outcome.

We chose to use a minimum follow-up time for each patient.. to determine study completion rather than an event-based method used by PFS. This led to a two to four months or 9%-31% earlier study completion date which might make randomized Phase II trials more attractive.

We have assumed that randomized Phase II trials may be considered as random subsamples from an underlying Phase III trial. However, patient selection in Phase II trials may differ from that in Phase III[19]. This may suggest that estimates of PPV (from the rank-based methods as well as from PFS) based on random subsampling from the Phase III trial may overestimate the PPV in practice.

5. References

1) Ellis LM, Bernstein DS, Voest EE, et al., “American Society of Clinical Oncology perspective: Raising the bar for clinical trials by defining clinically meaningful outcomes.” *J Clin Oncol*, 2014, 32:1277-1280.
2) Gan HK, You B, Pond GR, et al. “Assumptions of expected benefits in randomized Phase III trials evaluating systemic treatments for cancer.” *J Natl Cancer Inst*, 2012; 104: 590-598.
3) Gan HK. Personal communication to William Mietlowski 2013
4) Ratain MJ. “Phase II Oncology trials: let’s be positive.” *Clin Cancer Res*, 2005; 11:5661-5662.
5) Litiere S, De Vries EGE, Seymour L, et al. “The components of progression as explanatory variables for overall survival in the Response Evaluation Criteria in Solid Tumours 1.1 database”, *Eur J Cancer* 2014; 50: 1847-1853.
6) Mietlowski WL, Bao W, Wood PA, et al. “Clinical importance of including new and non-target lesion assessment of disease progression (PD) to predict overall survival (OS): implications for randomized Phase II study design”. *J Clin Onco*l 2012; 30 (suppl; abstr 2543).

7) Stein A, Bellmunt J, Escudier B et al. "Survival prediction in Everolimus-treated patients with metastatic renal cell carcinoma incorporating tumor burden response in the RECORD-1 trial*". Eur Urol* 2013; 64:994-1002.
8) Suzuki C, Blomqvist L, Sundin A, et al. "The initial change in tumor size *predicts response and survival in patients with metastatic colorectal cancer treated with combination chemotherapy". Ann Oncol 2012; 23: 948-954.*
9) Lachin JM. "Some large-sample distribution-free estimators and tests for multivariate partially incomplete data from two populations". *Statist Med* 1992; 11:1151-1170.
10) Wei LJ., Lachin JM. "Two-sample asymptotically distribution-free tests for incomplete multivariate observations". *J Amer Statist Assoc* 1984; 79: 653-661.
11) Horton NJ, Lipsitz SR. "Multiple imputation in practice". *The Amer Statist* 2001; 55: 244-254.
12) Zhao Y, Herring AH, Zhou H, et al. "A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring". *J Biopharm Statist* 2014; 24: 229-253.
13) Mogg R, Mehrotra D. "Analysis of antiretroviral immunotherapy trials with potentially non-normal and incomplete longitudinal data". *Statist Med* 2007; 26:484497.
14) Noether GE. "Sample size determination for some common nonparametric tests". *J Amer Statist Assoc* 1987; 82: 645-647.
15) Rubinstein L, Crowley J, Ivy P, et al. "Randomized Phase II designs". *Clin Cancer Res* 2009; 15: 1883-1890.
16) Van Cutsem E, Bajetta E, Valle J, et al. "Randomized, placebo-controlled, Phase III Study of Oxaliplatin, Fluorouracil, and Leucovorin with or without PTK787/ZK 222584 in patients with previously treated metastatic colorectal adenocarcinoma". *J Clin Oncol* 2011; 29: 2004- 2010.
17) Giantonio BJ, Catalano PJ, Meropol NJ, et al. "Bevacizumab in combi-nation With Oxaliplatin,Fluorouracil, and Leucovorin (FOLFOX4) for previously treated metastatic colorectal cancer: Results from the Eastern Cooperative Oncology Group Study E3200". *J Clin Oncol* 2007; 25:1539-1544.
18) Hudgens M, Satten G. "Midrank unification of rank tests for exact, tied, and censored data". *Nonparametric Statistics*, 2002; 14: 569-581.
19) Sonpavde G, Galsky MD, Hutson, TE et al."Patient selection for Phase II trials". *Am J Clin Oncol* 2009;32: 216-219.

Table 1. Support for importance of components of tumor assessment to predict overall survival in Oncology trials

Publication	Number of studies	Tumor types (number of patients analyzed)	New lesion occurrence Hazard ratio: median, min-max (p-value: median, min-max)	Non-target worsening Hazard ratio: median, min-max (p-value: median, min-max)
Litiere et al (2014 Eur J Cancer)	12 (random sample of 60% stratified by study)	7 MBC (n=1069) 3 NSCLC (n=1776) 2 MCRC (n=682)	1.93[a] (1.58-2.22) (<0.001, <0.001-<0.001)	1.49[a] (1.47-1.65) (<0.001, <0.001-0.005)
Mietlowski et al (ASCO 2012)	5	2 MCRC (n=1847) 2 NSCLC (n=1804) 1 OVCA (n=524)	3.02[b] (2.20-3.91) (<0.001, <0.001-<0.001)	1.67[b] (1.19-1.95) (<0.001, <0.001-0.332)
Stein et al (2013 Eur Urology)	1	1 RCC (n=246)	1.56 (0.053)	1.86 (0.005)
Suzuki et al (2012 Ann Oncology)	1	1 MCRC (n=506)	3.77[c] (<0.001)	3.77[c] (<0.001)

Multivariate Cox proportional model with target lesion data in the model (various functional forms); MBC=metastatic breast cancer; MCRC=metastatic colorectal cancer; NSCLC=non-small cell lung cancer; OVCA=advanced ovarian cancer; RCC=renal cell cancer
a=median across 3 tumor types; b=median across 5 studies; c=new and/or non-target lesion PD combined.

Table 2. Results of OS and PFS analyses of Phase III Oncology trials conducted or simulated

Study	Indication	Treatments	OS HR (p value) Event rate	PFS HR (p value) Event rate
CONFIRM-2 *	2nd line MCRC †	FOLFOX4 ± PTK787	HR=1.00 (p=0.96) 732/855=86%	HR=0.83 (p=0.01) 723/855=85%
CONFIRM-2 * (P1)	2nd line MCRC †	E58 vs. C42 (dummy treatments) ♯	HR=0.75 (p<0.0001) 732/855=86%	HR=0.74 (p<0.0001) 723/855=85%

* Tumor measurements every 8 weeks; Phase II data cutoff LPFV + 84 days;
† MCRC=metastatic colorectal cancer
♯ Simulated trial from CONFIRM-2 with 58% of patients with overall survival > median assigned to "experimental" treatment (E58) and 42% of patients with overall survival > median assigned to "control" treatment (C42) to obtain an overall survival hazard ratio of approximately 0.75 (as reported by Giantonio et al 2007)

Table 3. Study durations (first patient randomized to cutoff date) and number of PFS events for rank-based methods vs. PFS, CONFIRM-2 trial (2500 simulated randomized Phase II trials)

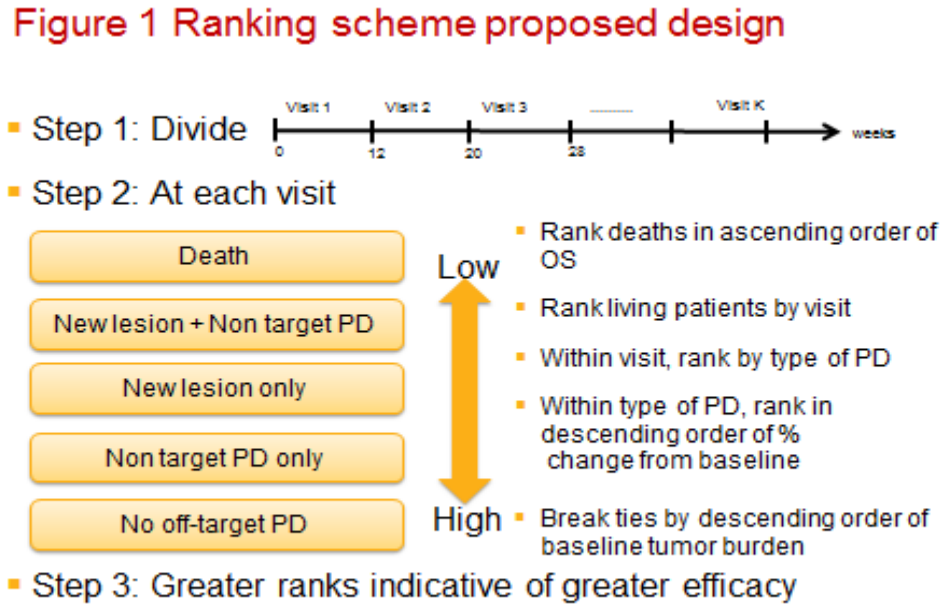
Enrollment rate	Maximum number of visits (months)	Data cutoff date	Median study duration in months (s.d. of median)	Mean number of PFS events (s.d.)
Uniform 25	5 (9)	LPFV* + 84 days **	8.7 (0.6)	77 (5)
		Mature PFS (110 events)	12.6 (1.7)	110 (0.3)
Piecewise	6 (12)	LPFV* + 84 days **	11.6 (0.1)	81 (5)
		Mature PFS (110 events)	15.5 (1.2)	110 (0.3)
Uniform 15	7 (13)	LPFV* + 84 days **	12.8 (0.1)	92 (5)
		Mature PFS (110 events)	15.6 (1.4)	110 (0.2)
Uniform 10	9 (18)	LPFV* + 84 days **	17.8 (0.1)	102 (5)
		Mature PFS (110 events)	19.5 (1.5)	110 (0.3)

* LPFV=last patient first visit=date of randomization of the 150th patient
** Tumor assessments every eight weeks + 50% allowance for delays
Rank based data cutoffs, in order of increasing LPFV date, are approximately 4 to 2 months earlier than PFS event based cutoffs;
Approximately 31%, 25%, 18% and 9% earlier respectively.

Table 4 Odds ratios from logistic model to predict probability of missingness from previous percentile assuming piecewise enrollment (2500 simulations)

Visit	Statistic	Covariate(s) in logistic model	
		Previous %ile	Previous %ile + treatment
4	Median (min,max)	1.331* (1.085, 1.682)	1.328* (1.083, 1.681)
	% p<0.05	97.6	97.3
5	Median (min,max)	1.385* (1.139, 1.778)	1.385* (1.141, 1.812)
	% p<0.05	98.9	98.7
6	Median (min,max)	1.636* (1.285, 2.422)	1.628* (1.283, 2.414)
	% p<0.05	100.0	100.0

* Odds ratio for a 13 point increase from 0.50 to 0.63 in the previous percentile
This is approximately 0.5 s.d., frequently used to establish a clinically meaningful effect size



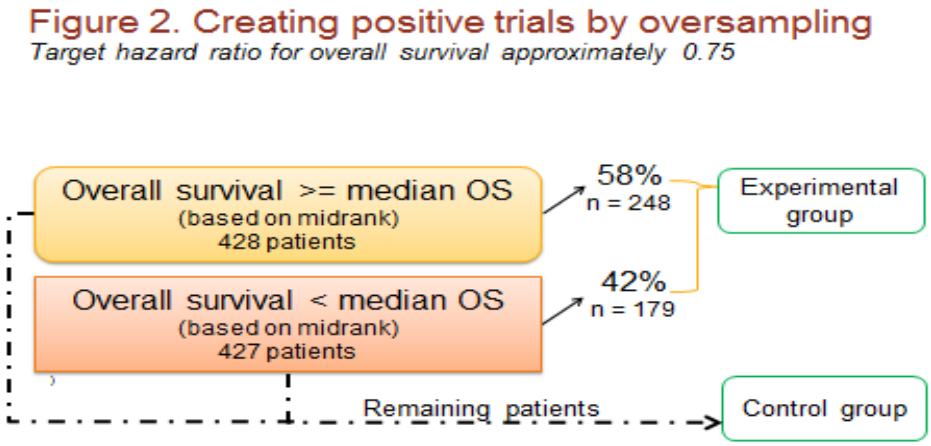


Figure legends for Figures 3-5
PFS= Progression-free survival, WL=Wei-Lachin test, MI-WL=Two stage test (multiple imputation followed by Wei-Lachin test on completed data sets)
Uniform 25=25 patients per month for six months
Piecewise= 10 patients per month for months 1-5, 25 patients per month for months 6-9
Uniform 15= 15 patients per month for 10 months
Uniform 10= 10 patients per month for 15 months

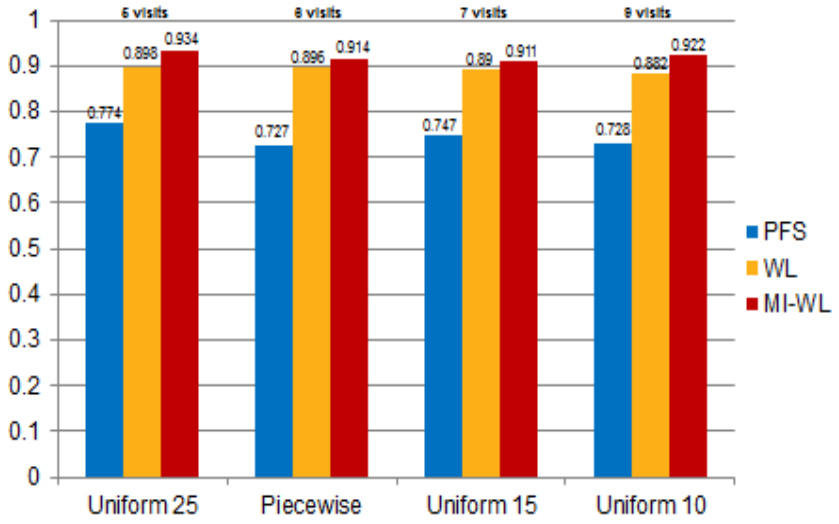


Figure 3 Specificity for Phase II studies generated from CONFIRM-2 by analysis method and enrollment distribution

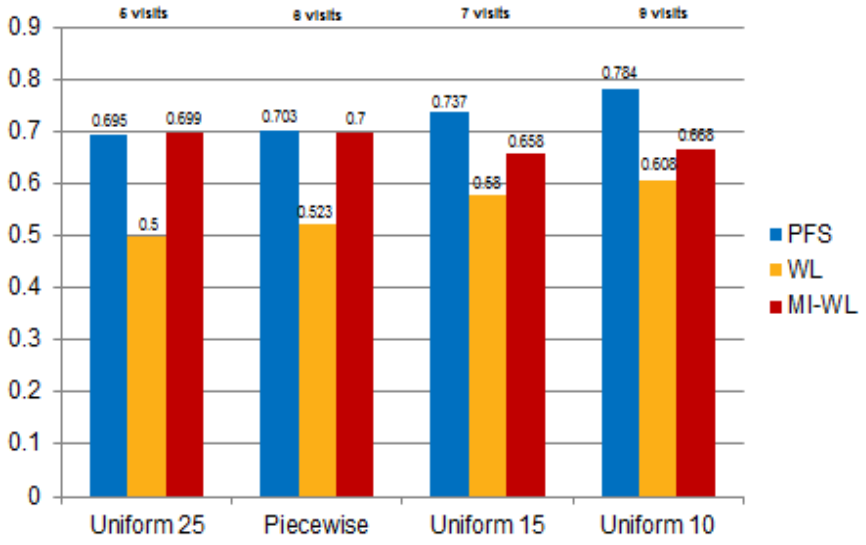


Figure 4 Specificity for Phase II studies generated from CONFIRM-2 P1 by analysis method and enrollment distribution

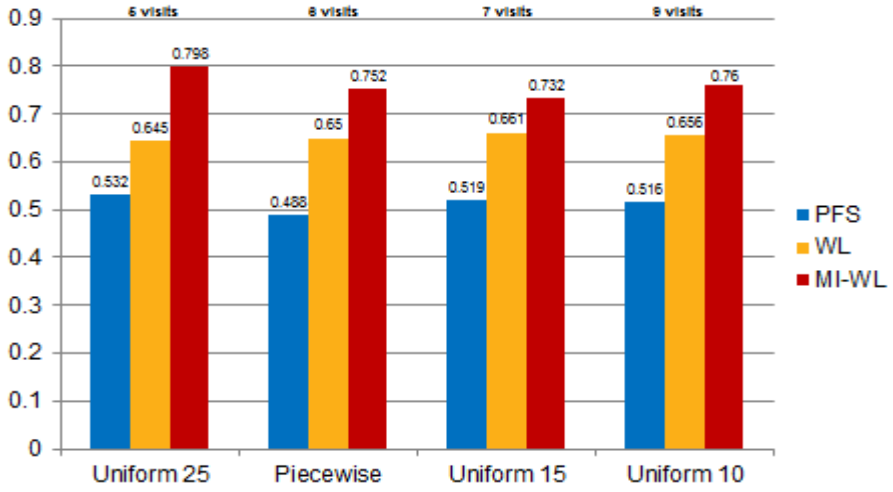
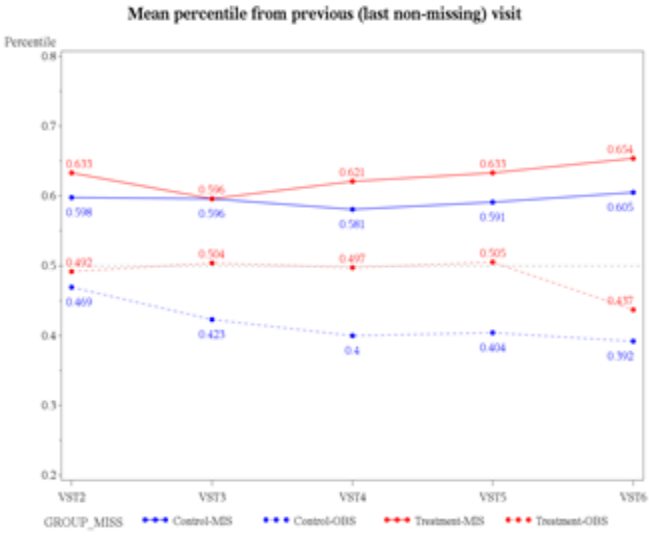


Figure 5 Positive predictive value (PPV) for Phase II studies generated from CONFIRM-2 and CONFIRM-2 P1 by analysis method and enrollment distribution Assuming a prevalence (Phase III success rate) of 27%

Figure 6 Mean previous percentile by missingness

and by visit and treatment group with piecewise enrollment

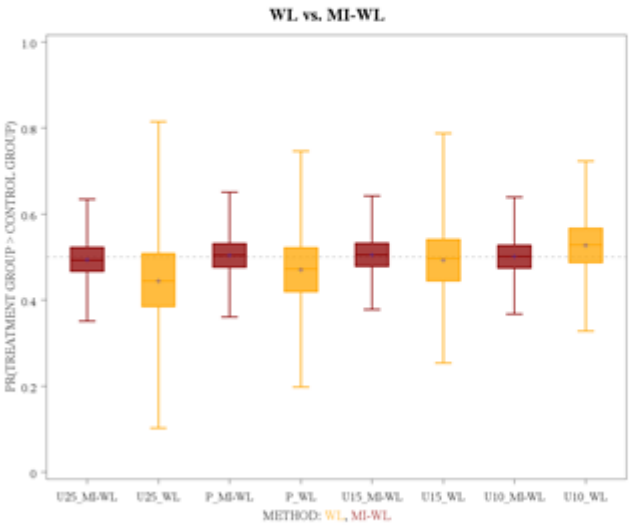


Visit	Group	N *
2	Control	64
	Treatment	64
3	Control	45
	Treatment	46
4	Control	32
	Treatment	29
5	Control	26
	Treatment	23
6	Control	22
	Treatment	16

* N = average sample size across simulations rounded to nearest integer

Figure 7 Mann-Whitney Probabilities – CONFIRM-2

Boxplot for WL vs. MI-WL by enrollment rate (decreasing missingness)



- WL shows decreasing variability as missingness decreases
- Also true for PFS (increasing number of events as missingness decreases)
- MI→ WL is robust over enrollment rate and degree of missingness, indicates good model fit with the MI model

Figure 8 Mann-Whitney probabilities CONFIRM-2 P1

Boxplot of WL vs. MI-WL by enrollment rate (decreasing missingness)