This appendix has been provided by the authors to give readers additional information about their work:

Circulating Tumor Cell Biomarker Panel As an Individual-Level Surrogate for Survival in Metastatic Castration-Resistant Prostate Cancer

Howard I. Scher, Glenn Heller, Arturo Molina, Gerhardt Attard, Daniel C. Danila, Xiaoyu Jia, Weimin Peng, Shahneen K. Sandhu, David Olmos, Ruth Riisnaes, Robert McCormack, Tomasz Burzykowski, Thian Kheoh, Martin Fleisher, Marc Buyse, and Johann S. de Bono

Howard I. Scher and Daniel C. Danila, Genitourinary Oncology Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center (MSKCC), New York, NY; Glenn Heller, Department of Epidemiology and Biostatistics, MSKCC, New York, NY; Martin Fleisher, Department of Laboratory Medicine, MSKCC, New York, NY; Xiaoyu Jia, Department of Biostatistics, Columbia University, New York, NY; Arturo Molina and Weimin Peng, Janssen Research & Development (formerly Cougar Biotechnology, Inc.), San Diego, CA; Robert McCormack, Janssen Research & Development (formerly Veridex), Raritan, NJ; Gerhardt Attard, Shahneen K. Sandhu, David Olmos, and Johann S. de Bono, The Institute of Cancer Research and The Royal Marsden Hospital, Sutton, Surrey, UK; Ruth Riisnaes, The Institute of Cancer Research, Sutton, Surrey, UK; Tomasz Burzykowski and Marc Buyse, International Drug Development Institute, Louvain-la-Neuve, Belgium, Thian Kheoh, Janssen Research & Development, San Diego, CA.

Supplementary Appendix Content

Methods S1. Individual-Level Surrogacy: Criterion P4

The Prentice (P) criteria[1] that were applied to assess the surrogate outcome measure at the individual level are shown in Table 2 of the main paper. Below we show further detail on testing criterion P4.

> Criterion P4: The full effect of treatment on the clinical end point must be captured by the biomarker. To show that after accounting for the surrogate, treatment has no residual effect on survival, a test for conditional independence was undertaken between a survival model that includes the patient surrogate (w) and treatment assignment (z), and a survival model based solely on the surrogate (w).

To assess the fourth Prentice criterion requires a test for equivalence between the survival function that includes the patient surrogate classification (w) and treatment assignment (z), and the survival function based solely on the surrogate. Let $S(t|w_i,z_i)$ represent the stratified Cox model survival estimate for subject *i*, with surrogate classification w_i and treatment assignment z_i. The protocol-specified baseline stratification factors were: ECOG status, pain index, prior chemotherapy, and type of disease progression. The statistic used to test for conditional independence is:

$$D(t) = n^{-1}\sum_i |S(t|w_i,z_i) - S(t|w_i)|$$

The statistic D(t) provides a measure of the distance between the full model S(t|w,z) and the reduced model S(t|w), which uses the surrogate class but not the treatment assignment to predict survival. This formalizes the concept that treatment provides no added value to the survival function after accounting for the surrogate. The test for conditional independence is carried out using a test for equivalence. Let $\Delta(t)$ represent the population value for D(t). The null and alternative hypotheses for this test of equivalence at the t-month mark:

$$H_0: \Delta(t) \geq .05$$

$$H_a: \Delta(t) < .05$$

Note that the concept of no added value has been enlarged to indicate that the survival functions differ by less than .05.

A test of equivalence was performed monthly in the time interval 6 to 24 months from the start of treatment. To account for the multiplicity of these tests, Holm's step-down adjusted *P*-value and the Bonferroni 95% upper confidence bound for Δ were computed at each time point (Table S1).

Table S1. Criterion P4 — Equivalence Test or Test of Conditional Independence (N=711)

Month	D(t)	$\sigma_{D(t)}$	Holm's adjusted *P* value	Bonferroni 95% upper confidence bound for Δ(t)
6	.0027	.0023	< .0001	.0090
7	.0040	.0033	< .0001	.0133
8	.0055	.0046	< .0001	.0182
9	.0068	.0056	< .0001	.0226
10	.0075	.0062	< .0001	.0249
11	.0083	.0068	< .0001	.0273
12	.0091	.0075	< .0001	.0300
13	.0091	.0076	< .0001	.0304
14	.0094	.0078	< .0001	.0312
15	.0097	.0081	< .0001	.0323
16	.0099	.0083	< .0001	.0331
17	.0099	.0083	< .0001	.0332
18	.0098	.0083	< .0001	.0329
19	.0098	.0082	< .0001	.0328
20	.0097	.0082	< .0001	.0325
21	.0096	.0081	< .0001	.0322

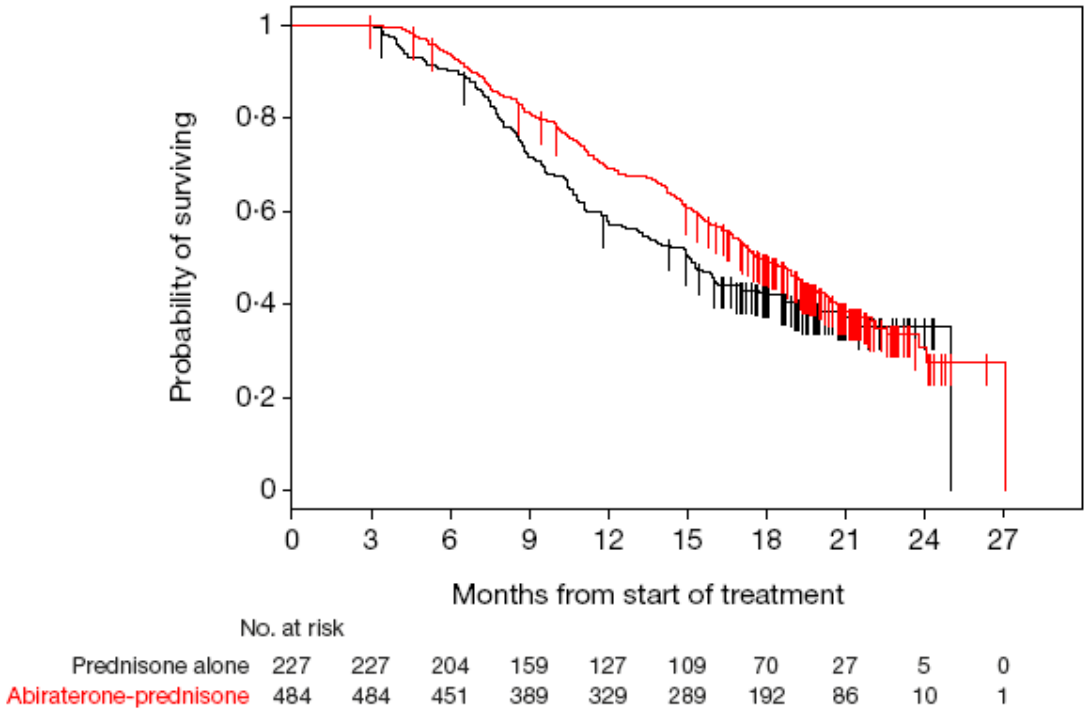
22	.0088	.0075	< .0001	.0298
23	.0086	.0073	< .0001	.0289
24	.0085	.0062	< .0001	.0258

Table S2. Stratified Cox Models —Model With Treatment Alone and The Model With Treatment + Biomarker

	Coeff	SE(Coeff)	*P* value
Treatment	–.225	.106	.035
Treatment	.093	.111	.402
Biomarker			< .001
High risk	1		
Intermediate risk	–.639	.145	
Low risk	–1.762	.128	

Abbreviations: Coeff, coefficient; SE, standard error.

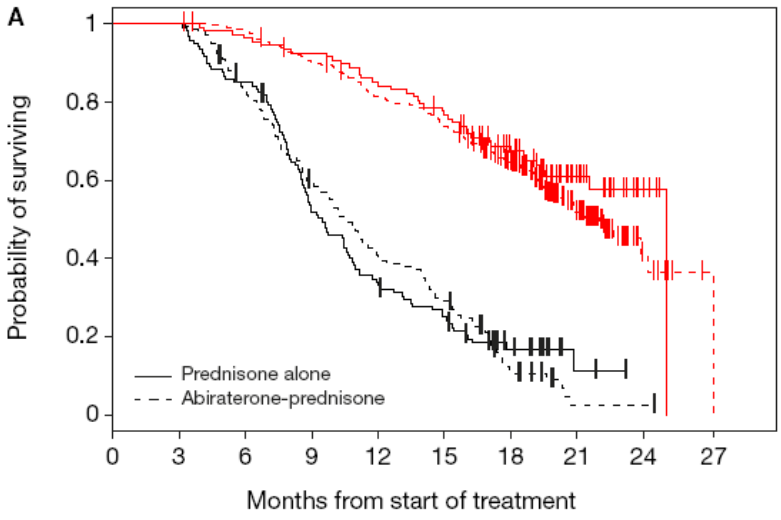
Figure S1. Kaplan-Meier estimates of overall survival based on treatment group for patients with 12-week biomarker data (N = 711)



Results S1. Analysis of CTC-Only and LDH-Only Surrogate Biomarkers

Our initial analysis examined CTC alone as a potential surrogate for survival. However, the survival estimates used in the assessment of criterion P4 could not be derived from the Cox proportional hazards model because the proportional hazards assumption was not satisfied for a model that included treatment assignment and CTC. As a result, we used the within-group Kaplan-Meier estimates of survival for the test of equivalence in criterion P4, but they did not satisfy criterion P4. The analyses used to examine CTC alone for surrogacy are provided in Figure S3A and Table S3A. Similar analyses were used to examine LDH alone for surrogacy; these results are provided in Figure S3B and Table S3B.

Figure S2A. Kaplan-Meier estimates of overall survival for CTC-only biomarker

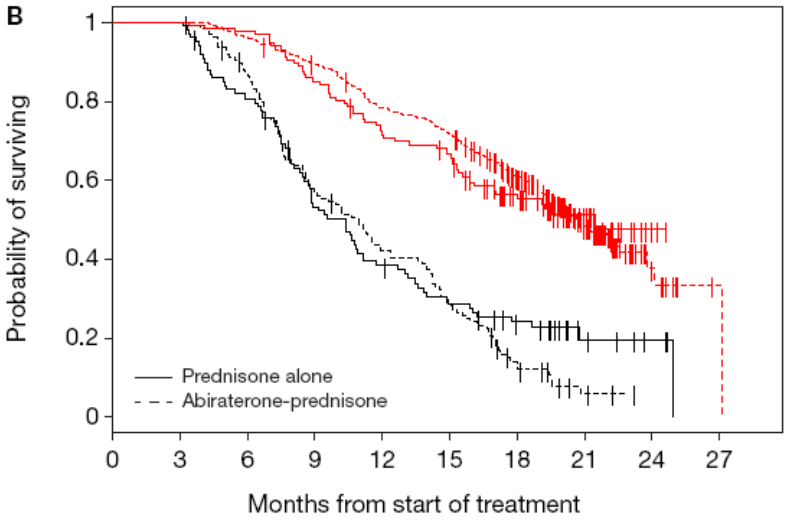


No. at risk

CTC ≥ 5, Prednisone alone	120	120	102	61	38	28	11	2	0	0
CTC ≥ 5, Abiraterone-prednisone	144	144	117	84	55	41	13	1	1	0
CTC < 5, Prednisone alone	109	109	104	99	90	81	59	25	5	0
CTC < 5, Abiraterone-prednisone	341	341	335	306	274	248	179	85	9	1

Abbreviation: CTC, circulating tumor cell.

Figure S2B. Kaplan-Meier estimates of overall survival for LDH-only biomarker



No. at risk

LDH > 250, Prednisone alone	99	99	79	50	36	26	18	7	3	0
LDH > 250, Abiraterone-prednisone	126	126	107	70	51	34	14	3	0	0
LDH ≤ 250, Prednisone alone	126	126	123	107	89	82	51	20	2	0
LDH ≤ 250, Abiraterone-prednisone	350	350	336	312	271	248	171	79	9	1

Abbreviation: LDH, lactate dehydrogenase.

Table S3A. Test of Equivalence in Criterion P4 for CTC-Only Biomarker Using Within-Group Kaplan-Meier Estimates of Overall Survival

Month	D(t)	SE[D(t)]	Adjusted *P* value	Bonferroni 95% upper confidence bound for Δ(t)
6	.0099	.0067	< .0001	.0282
7	.0089	.0081	< .0001	.0310
8	.0025	.0079	< .0001	.0241
9	.0186	.0101	.0153	.0462
10	.0161	.0106	.0135	.0451
11	.0237	.0115	.0662	.0551
12	.0160	.0115	.0212	.0474
13	.0192	.0115	.0362	.0506
14	.0146	.0110	.0135	.0447
15	.0134	.0104	.0049	.0418
16	.0108	.0102	.0014	.0387
17	.0076	.0098	.0002	.0344
18	.0165	.0114	.0212	.0477
19	.0215	.0116	.0564	.0532
20	.0256	.0129	.1183	.0609
21	.0361	.0151	.3568	.0774

Table S3B. Test of Equivalence in Criterion P4 for LDH-Only Biomarker

Using Within-Group Kaplan-Meier Estimates of Overall Survival

Month	D(t)	SE[D(t)]	Adjusted *P* value	Bonferroni 95% upper confidence bound for Δ(t)
6	.0145	.0078	.0001	.0376
7	.0015	.0074	< .0001	.0234
8	.0047	.0090	< .0001	.0313
9	.0185	.0117	.0847	.0531
10	.0219	.0125	.2735	.0588
11	.0268	.0135	.4265	.0667
12	.0240	.0134	.4166	.0636
13	.0228	.0135	.3989	.0627
14	.0253	.0134	.4546	.0649
15	.0139	.0130	.0711	.0523
16	.0282	.0137	.4446	.0687
17	.0283	.0144	.2664	.0709
18	.0302	.0140	.1566	.0716
19	.0220	.0127	.2783	.0595
20	.0295	.0129	.3409	.0676
21	.0283	.0124	.4878	.0649

Results S2. Individual-Level Surrogacy: Sensitivity Analysis (N = 899)

To explore the sensitivity of the individual-level surrogacy analysis to the missing 12-week biomarker data (188 of 899 patients), we reran the analysis using a last observation carried forward approach to fill in the missing CTC and LDH 12-week data values. We used the latest post-baseline biomarker data recorded ≤ 12 weeks from the start of treatment to record the surrogate value for each patient. This resulted in the sensitivity analysis including 899 patients who had been followed for survival for at least 12 weeks. Importantly, the proportional hazards assumption no longer held for this expanded cohort and as a result, nonmodel-based measures were used to assess the Prentice criteria (see Results S3 for more detail on testing the proportional hazards assumption).

Criterion P1: Treatment must have a significant effect on the clinical end point, ie, survival. The log-rank test produced a *P*-value equal to .01, indicating that there is a survival difference between the two treatment groups.

Criterion P2: Treatment must have a significant effect on the proposed biomarker. The chi-square test to determine whether the surrogate distribution after treatment differs by treatment group produced a *P*-value <.001. This indicates that there was a significant effect on the surrogate marker, the panel consisting of CTC + LDH.

Surrogate	**Abiraterone Acetate Plus Prednisone (N=602)**	**Prednisone Alone (N=297)**
High-risk	104 (17%)	107 (36%)
Intermediate-risk	91 (15%)	54 (18%)
Low-risk	407 (68%)	136 (46%)

Criterion P3: The biomarker must have a significant impact on the clinical end point. The log-rank test for the surrogate effect on survival produced a *P*-value < .001.

Criterion P4: The full effect of treatment on the clinical end point must be captured by the biomarker; it must be shown that after accounting for the surrogate, treatment has no residual effect on survival. The statistic used to test for conditional independence was

$$D(t) = n^{-1}\sum_i | S(t|w_i,z_i) - S(t|w_i) |$$

where S(t|.) represents the within-group Kaplan-Meier estimates of survival. The within-group Kaplan-Meier estimates are computed up to 21 months, which represents the largest follow-up time, where all the group estimates are well-defined and greater than zero. The Kaplan-Meier estimates are not as smooth and have greater variability at the later time points relative to the estimates based on the Cox model; this is due to the smaller number at risk. However, the evaluation of time points prior to month 21 demonstrates that criterion P4 remains satisfied for this expanded cohort of 899 patients in the last observation carried forward analysis (see Table S4).

Table S4. Criterion P4 — Equivalence Test or Test of Conditional Independence (N = 899): Last Observation Carried Forward Sensitivity Analysis

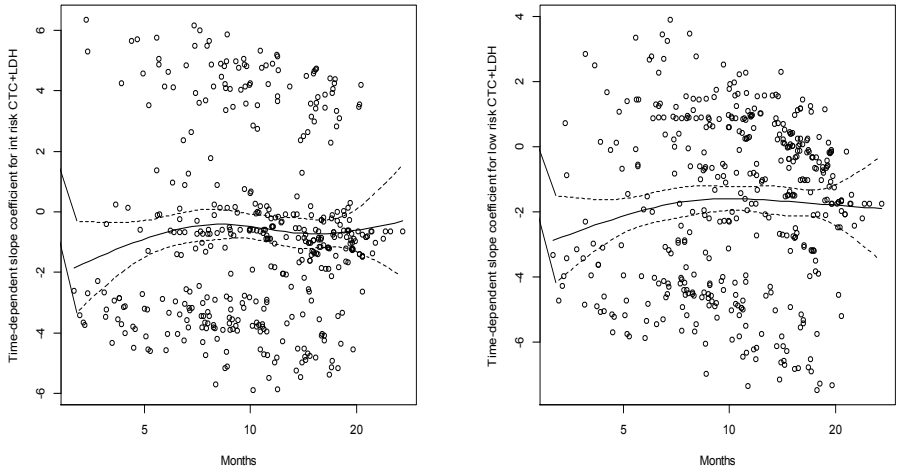
Month	D(t)	$\sigma_{D(t)}$	Holm's Adjusted *P* value	Bonferroni 95% upper confidence bound for Δ(t)
6	.0235	.0075	.0025	.0440
7	.0216	.0094	.0105	.0473
8	.0130	.0089	.0002	.0373
9	.0071	.0079	< .0001	.0287
10	.0077	.0080	< .0001	.0296
11	.0100	.0085	< .0001	.0332
12	.0116	.0088	.0001	.0357
13	.0073	.0088	< .0001	.0314
14	.0123	.0087	.0001	.0361
15	.0044	.0080	< .0001	.0263
16	.0078	.0081	< .0001	.0299
17	.0054	.0080	< .0001	.0273
18	.0189	.0088	.0025	.0430
19	.0220	.0098	.0106	.0488
20	.0204	.0101	.0106	.0480
21	.0306	.0125	.1222	.0648

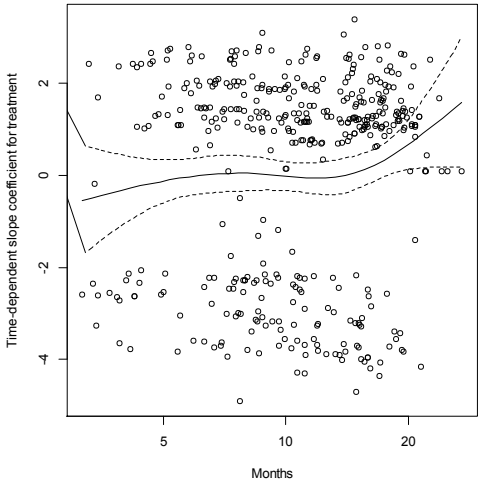
Results S3. Test of Proportional Hazards Assumption

A test is performed to determine whether the addition of a time-dependent covariate into the Cox model is significant. A small *P*-value is an indication that the coefficient associated with the additional covariate is nonzero, implying that the proportionality assumption is incorrect. In addition to the test, a graphical assessment of the proportionality is provided using smoothed scaled Schoenfeld residuals.[2] Within each figure, the solid line represents the time-dependent slope estimate and the dashed lines a 95% confidence interval. A nonzero slope indicates a violation of the proportional hazards assumption.

For the CTC + LDH three-level surrogate marker, the *P*-values indicate that the proportionality assumption cannot be rejected (Figure S4A). However, for the CTC-alone and LDH-alone surrogate markers, the global test produced *P*-values less than .05, indicating that the proportionality assumption is incorrect (Figure S4B and Figure S4C). Similarly, for a surrogate marker based on a 50% decline in PSA from baseline, the global p-value is less than .001, indicating that the proportionality assumption is incorrect (Figure S4D).

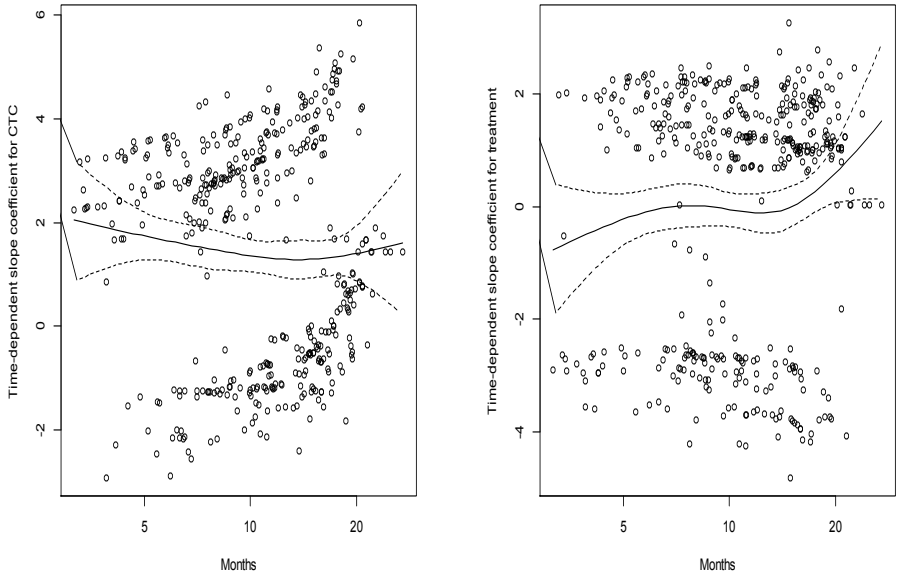
Figure S3A. Proportional hazards model based on the CTC + LDH three-level surrogate marker





	Rho	Chi square	*P* value
Intermediate risk (factor)	.026	.28	.596
Low risk (factor)	.050	1.00	.317
Treatment	.091	3.39	.066
Global	NA	5.56	.135

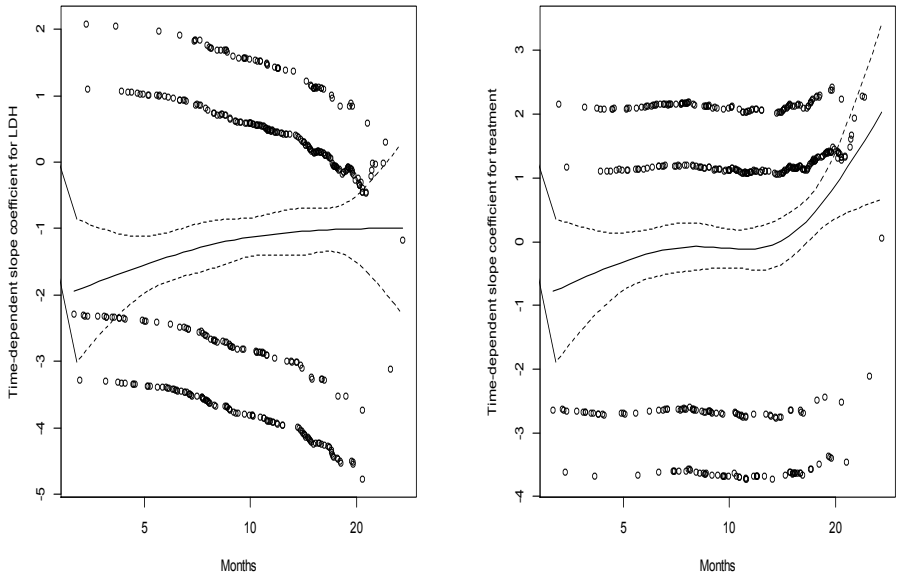
Figure S3B. Proportional hazards model based on CTC alone



	Rho	Chi square	*P* value
CTC	–.063	1.57	.211
Treatment	.095	3.70	.055
Global	NA	6.38	.041

Abbreviation: NA, not applicable.

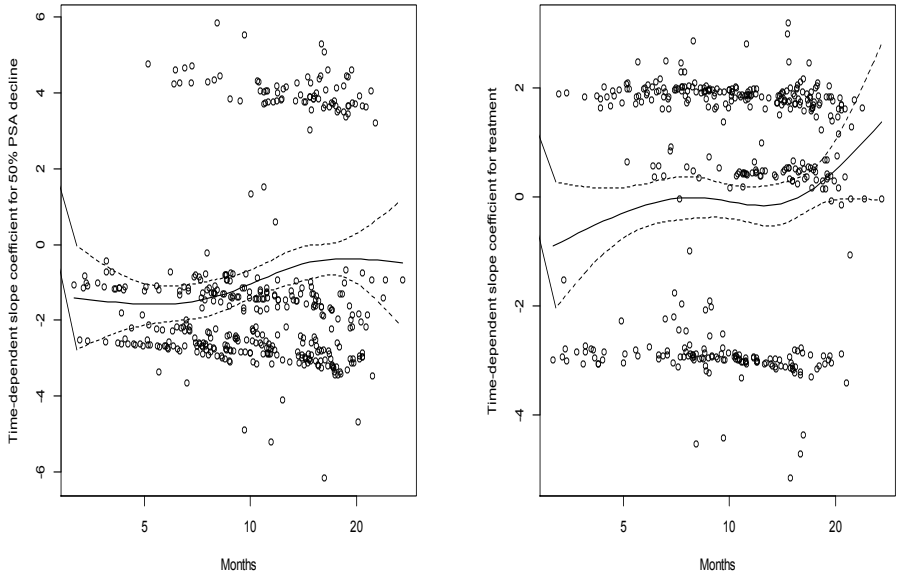
Figure S3C. Proportional hazards model based on LDH alone



	Rho	Chi square	*P* value
LDH	.092	3.43	.064
Treatment	.147	9.24	.002
Global	NA	15.76	< .001

Abbreviation: NA, not applicable.

Figure S3D. Proportional hazards model based on a 50% decline in PSA from baseline



	Rho	Chi squared	*P* value
PSA_{50}	.167	1.99	<. 001
Treatment	.092	3.27	.071
Global	NA	18.38	<. 001

Abbreviation: NA, not applicable.

REFERENCES

1. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 1989;8:431-40.

2. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994;81:515-26.