Can logistic regression better predict the significance of the overall survival from surrogate endpoints for randomized controlled trials in oncology? Insights from a cross-tumor case study

Oguzhan Alagoz,[1] Swetha Srinivasan,[2] Inkyu Kim,[2] Murat Kurt[2]

[1]University of Wisconsin-Madison, Madison, WI, USA; [2]Bristol Myers Squibb, Princeton, NJ, USA

Introduction

- Overall survival (OS) outcomes data are often delayed due to the need for a long follow-up time and a low rate of events
- Surrogate measures, such as tumor response rate, time to progression, and progression-free survival (PFS), have been suggested as alternative endpoints to OS as they provide more convenient shorter-term outcomes
- Validation of a surrogate endpoint (SE) requires a meta-analysis of all randomized clinical trials (RCTs) on the target intervention and estimation of the correlation between the hazard ratio (HR) of the SEs (eg, PFS) and OS to demonstrate the strength of the surrogacy measure
  - Linear regression (LiR) is often used to relate the HR of OS to the HR of the surrogate measure
- No existing framework has used the HR of SEs to predict whether the 95% confidence interval (CI) of the OS HR includes 1, thereby indicating whether the OS HR of the intervention under consideration will be statistically significant
- Accurate prediction of whether the OS HR of an intervention will be statistically significant is crucial for approval by regulatory agencies as well as for clinical guideline development

Objectives

- To develop a logistic regression (LoR) model for a direct prediction of the significance of OS from SEs and demonstrate the statistical utility of our approach over indirect prediction via LiR
- To evaluate the impact of the number of RCTs included in a meta-analysis on the predictive performance of LoR and LiR

Methods

- A weighted LiR with weights based on the sample sizes of the individual RCTs was built (**Figure 1**) and trained, to be used in the absence of a specific model
- The trained LiR was used to estimate the 95% prediction interval for the OS HR and determine whether it includes 1
- A weighted LoR was built (**Figure 1**), which predicts the probability that the upper bound of the 95% CI of OS HR includes 1
- Per the conventional approach, a cutoff value of 0.5 was used to determine whether the predicted outcome for LoR included 1 or 0; that is, if the predicted LoR value is greater than 0.5, then the OS HR of the corresponding RCT is classified as nonsignificant
- The LoR was trained by defining a binary variable (OS_{sig}) that takes values of 0 and 1 using the 95% CI of the reported OS HR data
- Leave-one-out cross-validation (LOOCV) was used to compare the performance of the 2 approaches in a case study; LOOCV uses all but 1 RCT to train the LoR or LiR, and then tests the trained model on remaining RCTs and repeats this approach for all RCTs
- The R^2 and P value of the F-test were calculated to measure the performance of the LiR, and the area under the receiver operator curve (AUC) was calculated to measure the predictive performance of the LoR
- Sensitivity analysis on the cutoff value used for classifying outcomes in the LoR model was conducted
- The case study included 30 meta-analysis publications that reported on 35 instances of SEs in 13 different cancers that included a total of 556 RCTs (**Table 1**); these 35 instances formed the individual meta-analyses (test cases) for the case study
  - SEs were PFS or its analogs in 22 instances, DFS in 12 instances, and MFS in 1 instance
- The meta-analyses included in the case study were selected after an extensive literature review
  - A meta-analysis was included in the case study only if information on the 95% CI of the HRs of OS and SE, and the number of patients for each RCT, were reported
  - Meta-analyses covered a variety of tumors
  - **Figure 2** shows a comparison of the models for PFS from one of the meta-analyses in advanced CRC

Figure 1. Models used to estimate the significance of OS HR

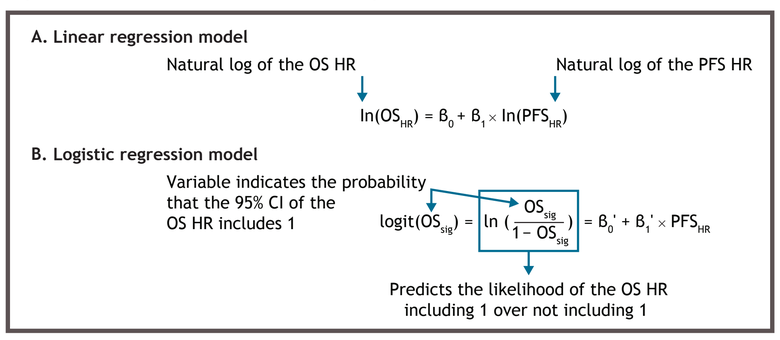
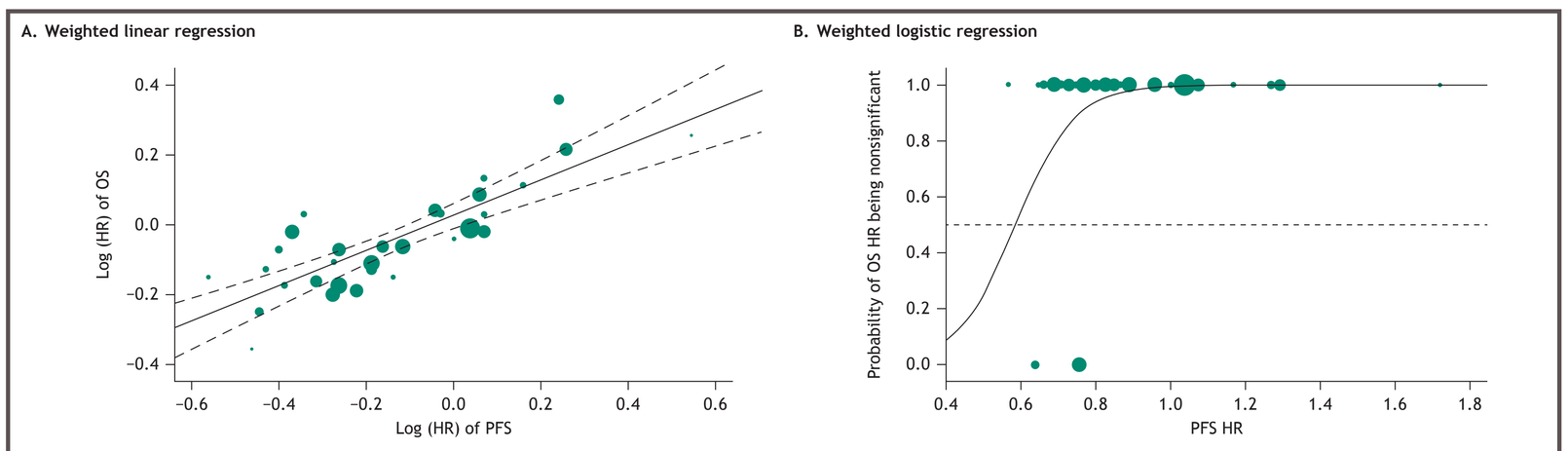


Table 1. Meta-analyses included in the case study

Test case	Study	Tumor type	Treatment setting	RCTs, n	Surrogacy measure
1	Moriwaki et al.[1]	Biliary tract	Advanced	9	PFS
2	Liang et al.[2]	Bladder	Advanced	26	PFS
3	Burzykowski et al.[3]	Breast	Advanced	11	PFS
4	Burzykowski et al.[3]	Breast	Advanced	11	TTP
5	Michiels et al.[4]	Breast	Advanced	9	PFS
6	Saad et al.[5]	Breast	Adjuvant	12	PFS
7	Gogate et al.[6]	Breast	Neoadjuvant	32	EFS/DFS
8	Burzykowski et al.[7]	CRC	Adjuvant	8	DFS
9	Buyse et al.[8]	CRC	Advanced	12	PFS
10	Montagnani et al.[9]	CRC	Advanced	10	PFS
11	Sidhu et al.[10]	CRC	Advanced	31	PFS
12	Kataoka et al.[11]	EC/GC	Neoadjuvant	10	PFS
13	Oba et al.[12]	EC/GC	Adjuvant	20	DFS
14	Paoletti et al.[13]	EC/GC	Advanced	12	PFS
15	Ronellenfitsch et al.[14]	EC/GC	Neoadjuvant	8	DFS
16	Ajani et al.[15]	EC/GC	Neoadjuvant/adjuvant	11	DFS
17	Ajani et al.[16]	EC/GC	Neoadjuvant/adjuvant	27	DFS/PFS
18	Fu et al.[17]	Glioblastoma	Adjuvant	11	PFS
19	Lee et al.[18]	HCC	Advanced	9	TTP
20	Llovet et al.[19]	HCC	Advanced	16	TTP
21	Llovet et al.[19]	HCC	Advanced	16	PFS
22	Flaherty et al.[20]	Melanoma	Advanced	14	PFS
23	Suciu et al.[21]	Melanoma	Adjuvant	13	RFS
24	Leung et al.[22]	Melanoma	Advanced	27	PFS
25	Chen et al.[23]	Nasopharyngeal	Advanced	16	FFS
26	Chen et al.[23]	Nasopharyngeal	Advanced	9	PFS
27	Paoletti et al.[24]	Ovarian	Advanced	17	PFS
28	Makris et al.[25]	Pancreatic	Advanced	22	PFS
29	Nie et al.[26]	Pancreatic	Adjuvant	23	DFS
30	Petrelli et al.[27]	Pancreatic	Adjuvant	11	DFS
31	Xie et al.[28]	Prostate	Neoadjuvant/adjuvant	16	EFS
32	Xie et al.[28]	Prostate	Adjuvant	31	DFS
33	Xie et al.[28]	Prostate	Adjuvant	21	MFS
34	Bria et al.[29]	RCC	Advanced	14	PFS
35	Harshman et al.[30]	RCC	Adjuvant	11	DFS

CRC, colorectal cancer; DFS, disease-free survival; EC, esophageal cancer; EFS, event-free survival; FFS, failure-free survival; GC, gastric cancer; HCC, hepatocellular carcinoma; MFS, metastasis-free survival; RCC, renal cell carcinoma; RFS, relapse-free survival; TTP, time to progression.

Figure 2. Comparison of the models for the example meta-analysis by Sidhu et al.[10]



Green dots represent the natural logarithm of the HRs of OS and PFS from individual RCTs, where the sizes of the dots correspond to the weights used for the LiR based on sample sizes. The solid line represents the LiR and the dotted lines represent the 95% predictive interval of the LiR equation.

Green dots represent the HRs of PFS and whether the reported 95% CI of the OS HR includes 1 from individual RCTs, where the sizes of the dots correspond to the weights used for the LiR based on sample sizes. The curve line represents the logistic regression function of PFS HR and the dashed line represents the cutoff point used to classify whether OS HRs are significant.

Results

- LiR and LoR correctly predicted the significance of 355 (64%) and 444 (80%) individual OS HR values, respectively, among a total of 556 RCTs (**Table 2**)
- LoR outperformed LiR in 26 meta-analyses whereas LiR outperformed LoR in 4 meta-analyses
- Among the 8 meta-analyses with < 10 individual RCTs, LiR and LoR correctly predicted the significance of 38 (53%) and 52 (72%) individual OS HR values, respectively, from a total of 72 RCTs
- Among the 15 meta-analyses with > 15 individual RCTs, LiR and LoR correctly predicted the significance of 219 (64%) and 275 (81%) individual OS HR values, respectively, from a total of 341 RCTs
- The mean R^2 for LiR was 0.59 and the P value of the F-test for LiR was < 0.05 in 30 of 35 test cases
- The mean AUC for LoR was 0.80 across all test cases
- Sensitivity analysis on the cutoff value used in LoR demonstrated that the performance of LoR was robust to the cutoff values < 0.7 (**Table 3**)
- Performance of LoR and LiR varied across individual cases (**Table 4**)

Table 2. Performance summary of the case study

Test case data	Meta-analyses included in the test set, n	Studies LoR performed better, n	Studies LiR performed better, n	Trials included in test cases, n	Correct predictions with LiR, n (%)	Correct predictions with LoR, n (%)	Average R^2 for LiR, %	Test cases where the F-test P value of the LiR is < 0.05, n	Average value of the AUC for LoR
Full test data	35	26	4	556	355 (64)	444 (80)	59	30	0.80
Meta-analyses with nRCT ≤ 10	8	6	2	72	38 (53)	52 (72)	51	5	0.79
Meta-analyses with 11 ≤ nRCT ≤ 15	12	7	1	143	98 (69)	117 (82)	55	11	0.78
Meta-analyses with nRCT ≥ 16	15	13	1	341	219 (64)	275 (81)	66	14	0.82
Meta-analyses with nRCT ≥ 20	10	10	0	260	156 (60)	205 (79)	71	10	0.77

Average values for AUC do not include test cases where AUC was not available.
nRCT, no. of randomized controlled trials.

Table 3. Sensitivity analysis on the cutoff point used for the LoR

Cutoff value	Meta-analyses included in the test set, n	Studies LoR performed better, n	Studies LiR performed better, n	Trials included in test cases, n	Correct predictions with LiR, n (%)	Correct predictions with LoR, n (%)
0.1	35	25	6	556	355 (64)	446 (80)
0.2	35	27	3	556	355 (64)	453 (81)
0.3	35	27	4	556	355 (64)	450 (81)
0.4	35	27	5	556	355 (64)	448 (81)
0.5	35	26	4	556	355 (64)	444 (80)
0.6	35	26	3	556	355 (64)	437 (79)
0.7	35	21	6	556	355 (64)	422 (76)
0.8	35	16	13	556	355 (64)	389 (70)
0.9	35	13	16	556	355 (64)	359 (65)

Table 4. Performance of LiR and LoR for each meta-analysis (test case) included in the case study

Test case	Study	Tumor	RCTs, n	RCTs with significant OS HR, n	Correct predictions with LiR, n	Correct predictions with LoR, n	R^2 for LiR, %	P value of the F-test of the LiR	AUC for LoR
1	Moriwaki et al.[1]	Biliary tract	9	2	5	7	55	0.02	0.43
2	Liang et al.[2]	Bladder	26	4	16	21	62	0	0.66
3	Burzykowski et al.[3]	Breast	11	2	9	11	39	0.04	1.00
4	Burzykowski et al.[3]	Breast	11	2	9	11	41	0.03	1.00
5	Michiels et al.[4]	Breast	9	1	4	7	16	0.29	1.00
6	Saad et al.[5]	Breast	12	5	7	7	65	0	0.63
7	Gogate et al.[6]	Breast	32	3	26	27	83	0	0.86
8	Burzykowski et al.[7]	CRC	8	0	2	8	37	0.11	0.86
9	Buyse et al.[8]	CRC	12	3	10	9	90	0	0.59
10	Montagnani et al.[9]	CRC	10	5	5	6	45	0.03	0.72
11	Sidhu et al.[10]	CRC	31	3	16	27	66	0	0.70
12	Kataoka et al.[11]	EC/GC	10	5	6	5	28	0.11	0.60
13	Oba et al.[12]	EC/GC	20	5	8	16	81	0	0.89
14	Paoletti et al.[13]	EC/GC	12	3	6	9	52	0.01	0.63
15	Ronellenfitsch et al.[14]	EC/GC	8	1	5	6	90	0	1.00
16	Ajani et al.[15]	EC/GC	11	1	9	10	93	0	0.90
17	Ajani et al.[16]	EC/GC	27	4	13	20	79	0	0.62
18	Fu et al.[17]	Glioblastoma	11	3	8	8	55	0.01	0.58
19	Lee et al.[18]	HCC	9	2	7	5	57	0.02	0.71
20	Llovet et al.[19]	HCC	16	5	12	13	69	0	0.89
21	Llovet et al.[19]	HCC	16	3	13	16	72	0	1.00
22	Flaherty et al.[20]	Melanoma	14	4	8	11	79	0	0.75
23	Suciu et al.[21]	Melanoma	13	1	5	11	43	0.02	0.83
24	Leung et al.[22]	Melanoma	27	11	17	18	62	0	0.74
25	Chen et al.[23]	Nasopharyngeal	16	4	10	14	79	0	0.98
26	Chen et al.[23]	Nasopharyngeal	9	3	4	8	81	0	1.00
27	Paoletti et al.[24]	Ovarian	17	0	17	17	23	0.05	1.00
28	Makris et al.[25]	Pancreatic	22	4	15	18	74	0	0.90
29	Nie et al.[26]	Pancreatic	23	8	15	17	79	0	0.73
30	Petrelli et al.[27]	Pancreatic	11	3	7	7	48	0.02	0.50
31	Xie et al.[28]	Prostate	16	4	11	10	35	0.02	0.62
32	Xie et al.[28]	Prostate	31	7	18	25	75	0	0.82
33	Xie et al.[28]	Prostate	21	6	12	16	53	0	0.81
34	Bria et al.[29]	RCC	14	2	9	12	21	0.10	1.00
35	Harshman et al.[30]	RCC	11	0	11	11	37	0.05	1.00

Conclusions

- LoR can serve as a robust and accurate alternative to LiR in predicting the significance of OS, particularly for meta-analyses that include a large number of RCTs with a balanced sample of OS HRs with potential outliers
- LoR performed substantially better than LiR in predicting the significance of the OS HR when all test cases were considered
- The performance of LoR was robust to the cutoff value used for classifying OS HRs as significant or nonsignificant
- LoR may not work well for meta-analyses that include a small number of RCTs
- More work is needed to estimate the performance of LoR and LiR in meta-analyses with balanced datasets, that is, when the numbers of significant and nonsignificant OS HRs are similar

References

1. Moriwaki T, et al. *Br J Cancer* 2016;114:881-888.
2. Liang F, et al. *Crit Rev Oncol Hematol* 2018;130:36-43.
3. Burzykowski T, et al. *J Clin Oncol* 2008;26:1987-1992.
4. Michiels S, et al. *Ann Oncol* 2016;27:1029-1034.
5. Saad ED, et al. *Lancet Oncol* 2019;20:361-370.
6. Gogate A, et al. *Cancer Res* 2022;82(suppl 4). Abstract P2-12-09.
7. Burzykowski T, et al. *Lifetime Data Anal* 2008;14:54-64.
8. Buyse M, et al. *J Clin Oncol* 2007;25:5218-5224.
9. Montagnani F, et al. *Anticancer Res* 2016;36:4259-4265.
10. Sidhu R, et al. *Clin Cancer Res* 2013;19:969-976.
11. Kataoka K, et al. *Eur J Surg Oncol* 2017;43:1956-1961.
12. Oba K, et al. *J Natl Cancer Inst* 2013;105:1600-1607.
13. Paoletti X, et al. *J Natl Cancer Inst* 2013;105:1667-1670.
14. Ronellenfitsch U, et al. *Eur J Cancer* 2019;123:101-111.
15. Ajani JA, et al. Data on file, Bristol Myers Squibb. 2022.
16. Ajani JA, et al. *Eur J Cancer* 2022;170:119-130.
17. Fu A, et al. *Value Health* 2020;23(suppl):S4.
18. Lee DW, et al. *Br J Cancer* 2016;115:1201-1205.
19. Llovet JM, et al. *J Hepatol* 2019;70:1262-1277.
20. Flaherty KT, et al. *Lancet Oncol* 2014;15:297-304.
21. Suciu S, et al. *J Natl Cancer Inst* 2018;110.
22. Leung L, et al. *Value Health* 2022;25(suppl):S22.
23. Chen YP, et al. *Radiother Oncol* 2015;116:157-166.
24. Paoletti X, et al. *JAMA Netw Open* 2020;3:e1918939.
25. Makris EA, et al. *Ann Surg Oncol* 2017;24:2371-2378.
26. Nie RC, et al. *BMC Cancer* 2020;20:421.
27. Petrelli F, et al. *HPB (Oxford)* 2017;19:944-950.
28. Xie W, et al. *J Clin Oncol* 2020;38:3032-3041.
29. Bria E, et al. *Crit Rev Oncol Hematol* 2015;93:50-59.
30. Harshman LC, et al. *Cancer* 2018;124:925-933.

Acknowledgments

- This study was supported by Bristol Myers Squibb
- All authors contributed to and approved the presentation; editorial support was provided by Russell Craddock, PhD, of Parexel, and was funded by Bristol Myers Squibb

Presented at ISPOR Europe 2022; November 6-9, 2022; Vienna, Austria, and virtual

Email: alagoz@engr.wisc.edu