Exploring and validating surrogate endpoints in colorectal cancer

Peer-reviewed author version

Exploring and Validating Surrogate Endpoints in Colorectal Cancer

Tomasz Burzykowski, PhD [1,2], Marc Buyse, ScD [1,3], Greg Yothers PhD [4], Junichi Sakamoto, PhD [5], Dan Sargent, PhD [6]

[1] Center for Statistics, Hasselt University, Diepenbeek, Belgium
[2] MSource Medical Development, Warsaw, Poland
[3] IDDI (International Drug Development Institute), Louvain-la-Neuve, Belgium
[4] NSABP (National Surgical Adjuvant Breast and Bowel Project) Biostatistical Center, Pittsburgh, USA
[5] Nagoya University Graduate School of Medicine, Japan
[6] Mayo Clinic College of Medicine, Rochester, Minnesota, USA

Address for correspondence and reprints:
Tomasz Burzykowski, Center for Statistics, Hasselt University, Agoralaan D, B-3590 Diepenbeek, Belgium
Tel: + 32 11268263, Fax: +32 003211268299, tomasz.burzykowski@uhasselt.be

Key words: surrogate endpoint, colorectal cancer, surrogate threshold effect

Introduction

Overall survival (OS) has been used as the primary endpoint in clinical trials aimed at evaluating benefits of adjuvant chemotherapy for colorectal cancer. However patient death can only be observed 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 disease-free survival (DFS) could replace OS as the primary endpoint in randomised colorectal cancer trials.

In an attempt to investigate the issue, Sargent et al (2005) analyzed individual patient data for 20,898 patients enrolled in 18 colon cancer adjuvant trials included in the Adjuvant Colon Cancer Endpoints (ACCENT) Group database. They concluded that 3-year disease-free survival (DFS) can be considered a valid surrogate (replacement) endpoint for 5-year overall survival (OS).

The analysis naturally leads to several questions. For instance, does the conclusion holds for trials involving other classes of treatment than those considered by Sargent et al? Is the 3-year cutpoint an optimal one?

In the present paper we aim to address these questions using individual patient data from three centrally randomized adjuvant colorectal cancer trials performed by the Japanese Foundation for Multidisciplinary Treatment for Cancer (JFMTC) to compare oral fluorinated pyrimidines with an untreated control group. The trials included 5,233 patients. Results of a meta-analysis of these trials were presented by Sakamoto et al (2004). In particular, we investigate whether the results observed in the ACCENT trials could have been used to predict treatment effects in the JFMTC trials. Moreover, we assess the dependence of the precision of prediction on the censoring cutpoint for DFS using the novel measure of surrogacy, the surrogate threshold effect (STE), developed by Burzykowski and Buyse (2006). STE is the minimum treatment effect on DFS required to predict a non-zero treatment effect on OS in a future (large) trial. The smaller STE, the higher the precision of prediction, and the more useful the proposed surrogate.

Material and methods

Material

Individual patient data were available for 20,898 patients enrolled in 18 randomized trials comparing experimental treatments with control treatments. All trials included at least one arm with a fluoro-uracil (FU) regimen. Nine trials included a no-treatment control arm. Several of the trials were multi-armed. In total there were 25 experimental vs. control treatment comparisons. In the analysis, each comparison was used as a separate trial. A more detailed description of the dataset can be found in Sargent et al (2005).

The three trials (identified as 7-1, 7-2 and 15) carried out by the JFMTC involved a total of 5,233 patients. All three trials had separate randomizations for patients with colon cancer (referred to as 7-1-C, 7-2-C and 15-C, respectively) and those with rectal cancer (referred to as 7-1-R, 7-2-R and 15-R, respectively). In colon cancer, two trials (7-2-C and 15-C) tested Carmofur and one (7-1-C) tested oral 5FU. In rectal cancer, two trials tested UFT (7-1-R and 15-R) and one trial tested Carmofur (7-2-R). All trials used an untreated control group. A more detailed description can be found in the paper by Sakamoto et al (2004).

For the purpose of the analysis, the within-cancer-location comparisons were considered separate studies. Additionally, trial 15 had a third treatment arm consisting of the non-specific immunopotentiator OK-432, which was discontinued. After the discontinuation, randomization in this study was performed in a 2:1 ratio (2 treatments to 1 control). The JFMTC 15 trial was therefore considered as two separate trials, labelled 15-1 (three-arm study) or 15-2 (two-arm study). Thus, in total, eight comparisons between an adjuvant oral fluoropyrimidine experimental treatment and an untreated control group were used in the analysis.

Methods

Analyses were based on all randomised patients. DFS was calculated from the time of randomization to first disease recurrence as defined in each individual trial, or death from any cause. OS was calculated from the time of randomization to death from any cause.

The ACCENT data were re-analyzed using a correlation approach (Buyse et al 2000). The method consisted of estimating (a) the rank correlation coefficient between DFS and OS, using a bivariate copula distribution for these endpoints and (b) the correlation coefficient between the treatment effects on DFS on OS (quantified through log hazard ratios (log HR), estimated through a proportional hazards (Weibull) model, stratified for trial, with treatment as the only factor), using an ordinary linear regression (Burzykowski et al 2001). The Plackett copula providing the best fit to the data (as determined by AIC) was chosen. The linear regression model was estimated with and without adjusting for the estimation error present in the treatment effects (Burzykowski & Cortiñas Abrahantes 2005). The model was used to generate predicted treatment effects on OS in the Japanese trials, based on the DFS results in those same trials. These predictions were compared with the actual results obtained in the trials. Also, the precision of the predictions and of the estimates were compared. The linear regression model between log hazard ratios was also used to compute the “surrogate threshold effect” (STE), which is the minimum treatment effect on DFS required to predict a non-zero treatment effect on OS in a future trial (supposed of infinite size to avoid the issue of estimation error in the future trial) (Burzykowski and Buyse 2006). The magnitude of the STE reflects the minimum width of the prediction limits for the treatment effect on OS in a new trial, obtained from the effect on DFS. The smaller the STE, the narrower the prediction limits, and the more useful the surrogate.

The analyses were performed initially using all available information. In order to investigate the influence of censoring on the validity of DFS as a surrogate for OS, the analyses were repeated with DFS censored (for all patients) at 1, 2, or 3 years, and OS censored at 5 years.

Results

All available data analysis

Figure 1 presents the (estimation-error adjusted) linear regression line, estimated using all available information in the ACCENT dataset, and used to predict treatment effects on OS from the observed treatment effects on DFS in the JFMTC trials. The regression equation was log HR_{OS} = 0.03 + 1.20 × log HR_{DFS} (standard errors: intercept 0.02, slope 0.16), indicating that the risk reductions were approximately 20% (= 1 + 0.20) higher on OS than on DFS. It is worth noting that the standard error of the slope does not allow to exclude the possibility that the true value can be equal to or smaller than 1. The correlation coefficient (resulting form an ordinary regression model) between the log hazard ratios was equal to 0.95 (CI 0.91 – 0.99).

Figure 1 here

The STE (based on the estimation-error corrected prediction limits) was (on the log-hazard scale) -0.09, what corresponded to a DFS hazard ratio of 0.91 (or 1.09). Thus, in order to predict a non-zero treatment effect on OS in a future trial, a hazard ratio of at most 0.91 or at least 1.09 would need to be ascertained.

Censored data analyses

Figures 2-4 show the (estimation-error adjusted) linear regression lines, estimated from the observed treatment effects on DFS and OS, with OS censored at 5 years and DFS censored at 1 (Figure 2), 2 (Figure 3), and 3 (Figure 4) years, respectively.

Figures 2-4 here

The correlation coefficients (from ordinary regression models) between the log hazard ratios were equal to 0.86 (CI 0.76 – 0.96) for censoring of DFS at 1 year, 0.92 (CI 0.90 – 0.94) for censoring at 2 years, and 0.92 (CI 0.90 – 0.95) for

censoring at 3 years. The (estimation error adjusted) linear regression lines used to predict treatment effects on OS (censored at 5 years) from the observed treatment effects on DFS (censored at 1, 2, or 3 years) in the JFMTC trials were log HR_{OS5} = 0.50 × log HR_{DFS1} (standard errors: intercept 0.04, slope 0.26), log HR_{OS5} = 0.02 + 0.77 × log HR_{DFS2} (standard errors: intercept 0.02, slope 0.18), and log HR_{OS5} = 0.03 + 0.90 × log HR_{DFS3} (standard errors: intercept 0.02, slope 0.20), respectively. It is worth noting that the last regression equation is almost identical to that reported by Sargent et al (2005) for a weighted linear regression model applied to the estimated treatment effects (but expressed as hazard ratios) obtained from marginal proportional hazard models. It suggests an attenuation of treatment effect on OS relative to the effect on DFS. However, contrary to the analysis conducted by Sargent et al, the standard error does not exclude the possibility that the true value of the slope is equal to 1, i.e., no attenuation.

The surrogate threshold effects (based on the measurement-error corrected prediction limits) were (on the log-hazard scale) -0.49, -0.26, and -0.21 at 1, 2, and 3 years, respectively. These values correspond to DFS hazard ratios of 0.61 (or 1.63), 0.77 (1.30), and 0.81 (1.23), respectively.

Figure 5 presents point predictions for the treatment effects on OS (censored at 5 years) obtained from the observed effects on DFS (censored at 1, 2, or 3 years); Table 1 provides detailed numerical results. In general, the later the censoring, the closer the point prediction to the actual estimate of the treatment effect. The poorest point predictions are obtained for trials 7-1-C, 15-1-R, and 15-2-R. Nevertheless, even for these trials the observed effect on OS at 5 years falls within the prediction limits implied by the estimated standard error of the prediction (see Table 1).

Figure 5 and Table 1 here

From Table 1 it can be observed that the standard error of predictions is similar or smaller than the standard error of the estimates obtained from the data. This indicates that it is possible to obtain an estimate of the treatment effect on

5-year OS earlier, by predicting it from the treatment effect on DFS, without losing precision of the estimation.

It is worth noting that the standard errors for the predicted log HRs for OS at 5 years remain similar (or even slightly increase) for DFS censored at 1, 2, and 3 years, while both the precision of the estimation of the effect on DFS (see the fourth column of Table 1) and the strength of the association between the treatment effects on DFS and OS seem to increase with time. (The latter can be inferred from the increasing correlation between the treatment effects and from the decreasing STE.) This counterintuitive result is due to the fact that the variance of the point prediction for the effect of treatment on OS contains three components: one comes from the residual variability in the regression of OS treatment effects based on DFS effects; one comes from the estimation of the regression line; and one comes from the estimation of the treatment effect on DFS (see equation (14) in Burzykowski and Buyse, 2006). The STE is based on the prediction limits constructed using the first two components. As the association between the treatment effects increases as the censoring point for DFS becomes later in time, - the sum of the two components gets smaller, and so does STE. For instance, for the predictions based on the estimation-error adjusted linear regression model, for censoring DFS at 1 year, the sum of the first two components is equal to about 0.14; for 2 years, it is about 0.009; and for 3 years, about 0.008. On the other hand, the component of variance in the estimated OS effect due to the estimation of the effect on DFS is equal to the expected value of

(slope of the regression line)^2 x (variance of the estimated effect on DFS).

The slope increases with the increasing cutpoint for censoring: it is 0.5 for censoring at 1 year, 0.77 at 2 years, and 0.90 at 3 years. This means that, for the individual prediction, we add to the prediction variance larger parts of the variance of the estimated treatment effect on DFS. Therefore, although the sum of the variance components associated with the regression line estimation gets smaller, the decrease is counterbalanced by the addition of the increasing part of the variance of the estimated treatment effect on DFS.

Thus, the main reason for a slightly increased individual prediction variance is the increased slope of the regression line. This indicates that, at earlier times within a trial, larger effects on DFS are observed, and that these early estimates must be attenuated more so than at later times to concur with the value of the treatement effect on OS at 5 years. This phenomena has been reported by Sargent et al (2007), who demonstrated a non-constant effect of treatment on DFS, with a highly significant benefit in only the first two years, compared to a constant benefit of treatment on OS over an 8 year follow-up period.

Discussion

There are several differences between the analysis of the ACCENT data presented in this paper and the one conducted by Sargent et al (2005).

First and foremost, in their primary analysis Sargent et al considered marginal hazard ratios estimates, obtained from a proportional hazard model. Thus, in their analysis the association between DFS and OS was ignored. They used simple linear regression weighted by the trial size to model the association between the marginal hazard ratios. Second, they censored data using the theoretical calendar time, at which the median follow up in a particular trial would have reached, e.g., 3 years. In our analyses, treatment effects were estimated using a bivariate copula with marginal Weibull models. This allowed us to take into account the association between the endpoints. We used measurement-error modelling techniques to adjust for the estimation error in the observed treatment estimates. Finally, as we did not have access to the actual dates of observations in the JFMTC trials, we uniformly censored DFS for all patients at 1, 2, or 3 years (OS at 5 years).

Despite the differences, we observed many similarities between the results obtained in our analysis for DFS censored at 3 years and OS censored at 5 years and those reported by Sargent et al (2005).

Regarding the first question posed in the Introduction - does the conclusion holds for trials involving other classes of treatments than those considered by Sargent et al? - our analysis shows that, for all JFMTC trials, the prediction limits for

trial-specific treatment effect on OS, constructed based on the models built using the ACCENT database, included the point estimate obtained directly from the data on OS available in the trial. In none of the cases did the prediction limits suggest a non-zero treatment effect on OS, which is in accordance with the results obtained directly from the OS data available in the JFMTC trials. Finally, in four cases (7-1-R, 7-2-R, 15-1-C, and 15-2-C) the point predictions were also very close to the direct estimates (when considering DFS censored at 3 years). Note that, while the prediction model was based solely on colon cancer (ACCENT) data, it also seemed to extrapolate well for predicting the association between 3 year DFS and 5 year OS also in rectal cancer. These results suggest that DFS censored at 3 years can be considered a surrogate for OS censored at 5 years for colorectal cancer. It should however be emphasized that the prediction intervals are wide, which implies that if a trial of a new experimental treatment uses DFS as a surrogate for OS, there will remain substantial uncertainty regarding the true survival benefit provided by that treatment.

With respect to the second question posed in the Introduction - is the 3-year cutpoint an optimal one? - our analysis shows that, with later censoring cutpoints for DFS, the association between treatment effects on DFS and 5-year OS becomes stronger, as indicated by the increasing correlation between the treatment effects and the decreasing STE. For the JFMTC trials, in general, the point-wise prediction for the treatment effect for 5-year OS improved with a later censoring cutpoint. At the same time, the precision of the prediction was similar to the precision of the estimate of the treatment effect obtained directly from the OS data. From this point of view, using DFS at 2 or 3 years would be the best option for the prediction of OS at 5 years.

In this paper, we have used the surrogate threshold effect to assess the validity of DFS, censored at a particular time, as a surrogate for OS censored at 5 years. An interesting feature of a surrogate threshold effect, apart from providing information relevant to the practical use of a surrogate endpoint, is its natural interpretation from a clinical point of view. It can be expressed in terms of treatment effect on the surrogate necessary to be observed to predict a significant treatment effect on the true endpoint. The use of STE might facilitate

communication between the statisticians and clinicians regarding results of a validation of a surrogate endpoint.

Acknowledgments

The authors are grateful to the ACCENT and the JFMTC collaborators for permission to use their data. Dr Burzykowski acknowledges financial support from the IAP Research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

1. Burzykowski T, Buyse M (2006). Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics* 5:173-186.
2. Burzykowski T, Cortiñas Abrahantes J (2005) Validation in case of two failure-time endpoints. In: *Evaluation of Surrogate Endpoints* (Burzykowski T, Molenberghs G, Buyse M, eds.) New York: Springer.
3. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D (2001). Validation of surrogate endpoints in multiple randomised clinical trials with failure-time endpoints. *Journal of the Royal Statistical Society C (Applied Statistics)* 50:405-422.
4. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H (2000). The validation of surrogate endpoints in meta-analyses of randomised experiments. *Biostatistics* 1:49-68.
5. Sakamoto J, Ohashi Y, Hamada C, Buyse M, Burzykowski T, Piedbois P for the Meta-Analysis Group of the Japanese Society for Cancer of the Colon and Rectum and the Meta-Analysis Group in Cancer (2004) Efficacy of oral adjuvant therapy after resection of colorectal cancer: 5-year results from three randomized trials. *Journal of Clinical Oncology* 22:484-492.
6. Sargent D, Wieand S, Haller DG, et al (2005) 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. *Journal of Clinical Oncology* 23:8664-8670.
7. Sargent DJ for the ACCENT Group (2007) Time-dependent patterns of failure and treatment benefit from adjuvant therapy for resectable colon cancer: Lessons from the 20,800 patient ACCENT dataset. 2007 Gastrointestinal Cancers Symposium, Abstract #274.

Tables

Table 1: Observed and predicted treatment effects (log HR = log hazard ratios) on disease-free survival (DFS) censored at 1, 2, or 3 years, and on overall survival (OS, censored at 5 years). SE = standard error; LS = ordinary (least squares) linear regression; ME = measurement-error regression.

Trial	Cutpoint (years)	Observed				Predicted (LS)		Predicted (ME)	
		$\log HR_{DFS}$ (x)	SE(x)	$\log HR_{OS5}$ (y)	SE(y)	y	SE(y)	Y	SE(y)
7-1-C	1	0.211	0.216	-0.050	0.149	0.118	0.178	0.096	0.146
7-1-R	1	-0.159	0.176	-0.140	0.123	-0.081	0.161	-0.082	0.137
7-2-C	1	-0.182	0.226	-0.226	0.133	-0.093	0.178	-0.093	0.148
7-2-R	1	-0.173	0.210	-0.051	0.134	-0.088	0.172	-0.088	0.145
15-1-C	1	0.097	0.321	-0.243	0.198	0.057	0.218	0.041	0.174
15-1-R	1	0.101	0.349	-0.091	0.214	0.059	0.230	0.043	0.182
15-2-C	1	-0.563	0.280	0.019	0.180	-0.298	0.203	-0.275	0.162
15-2-R	1	-0.485	0.285	-0.130	0.217	-0.256	0.203	-0.238	0.164
7-1-C	2	0.032	0.163	-0.050	0.149	0.046	0.169	0.043	0.154
7-1-R	2	-0.168	0.134	-0.140	0.123	-0.112	0.151	-0.111	0.138
7-2-C	2	-0.247	0.159	-0.226	0.133	-0.173	0.166	-0.171	0.152
7-2-R	2	-0.048	0.157	-0.051	0.134	-0.018	0.164	-0.019	0.150
15-1-C	2	-0.367	0.221	-0.243	0.198	-0.268	0.206	-0.263	0.189
15-1-R	2	-0.515	0.219	-0.091	0.214	-0.384	0.207	-0.377	0.189
15-2-C	2	-0.137	0.198	0.019	0.180	-0.087	0.189	-0.087	0.175
15-2-R	2	-0.235	0.227	-0.130	0.217	-0.164	0.209	-0.162	0.194
7-1-C	3	0.093	0.150	-0.050	0.149	0.111	0.168	0.111	0.157
7-1-R	3	-0.187	0.122	-0.140	0.123	-0.141	0.147	-0.141	0.137
7-2-C	3	-0.203	0.136	-0.226	0.133	-0.155	0.156	-0.155	0.147
7-2-R	3	-0.056	0.136	-0.051	0.134	-0.023	0.156	-0.023	0.147
15-1-C	3	-0.329	0.199	-0.243	0.198	-0.268	0.205	-0.267	0.194
15-1-R	3	-0.376	0.204	-0.091	0.214	-0.311	0.209	-0.310	0.198
15-2-C	3	-0.074	0.181	0.019	0.180	-0.039	0.189	-0.039	0.180
15-2-R	3	-0.338	0.195	-0.130	0.217	-0.276	0.201	-0.275	0.191

Figures

Figure 1: Correlation between treatment effects (log HR = log hazard ratios) on disease-free survival (DFS) and on overall survival (OS). Symbol size is proportional to the number of patients; the line of prediction and prediction limits come only from the ACCENT data.

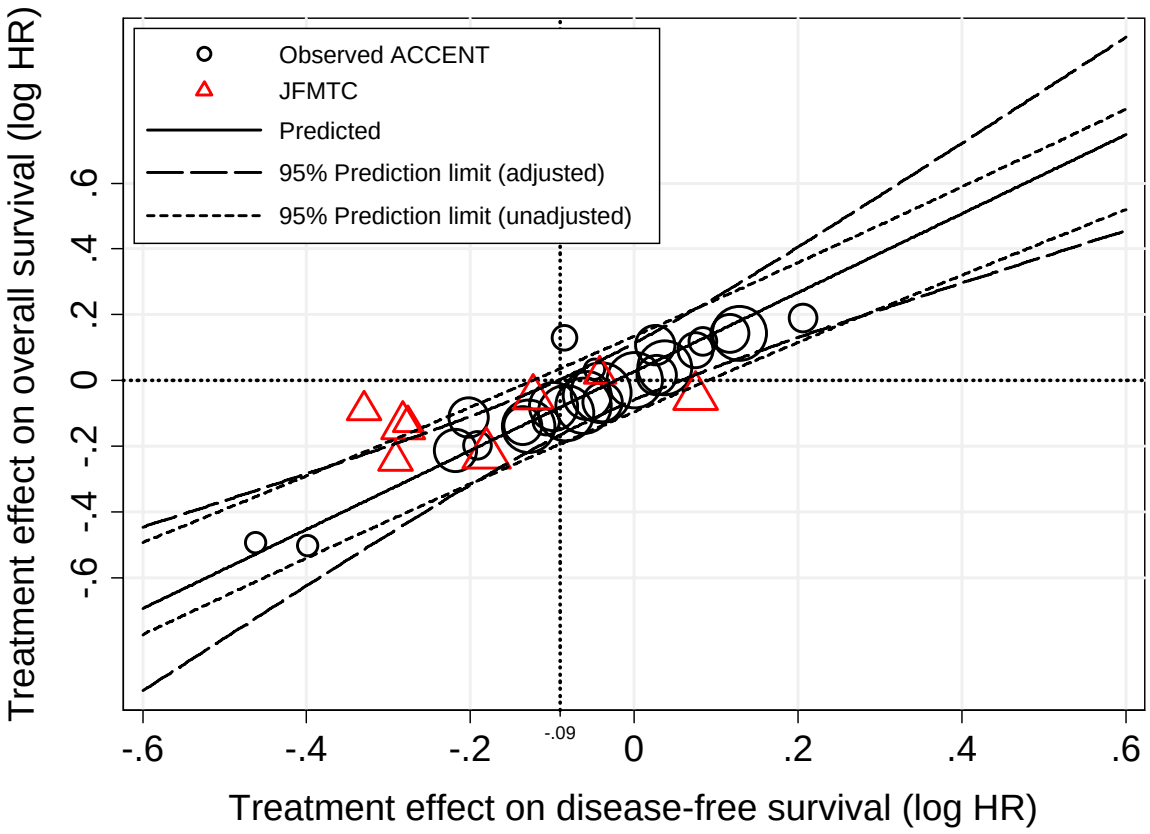

Figure 2: Correlation between treatment effects (log HR = log hazard ratios) on disease-free survival (DFS, censored at 1 year) and on overall survival (OS, censored at 5 years). Symbol size is proportional to the number of patients; the line of prediction and prediction limits come only from the ACCENT data.

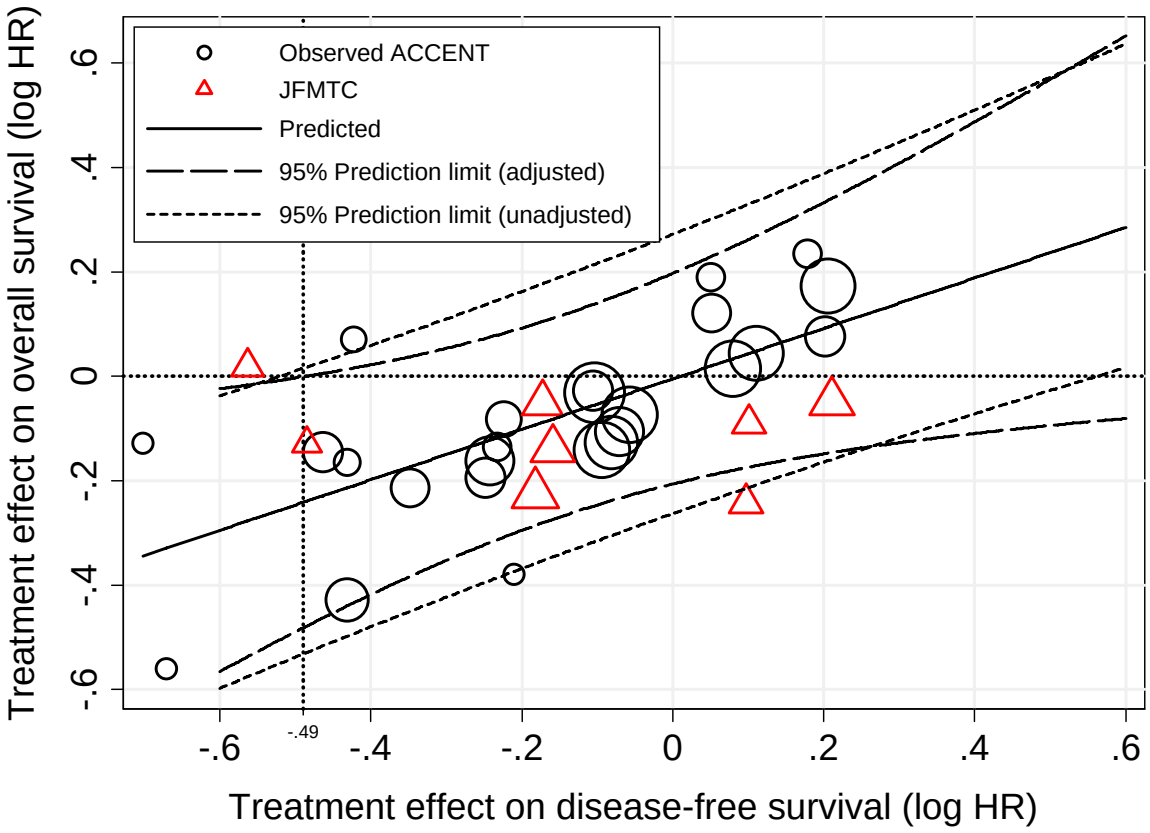

Figure 3: Correlation between treatment effects (log HR = log hazard ratios) on disease-free survival (DFS, censored at 2 years) and on overall survival (OS, censored at 5 years). Symbol size is proportional to the number of patients; the line of prediction and prediction limits come only from the ACCENT data.

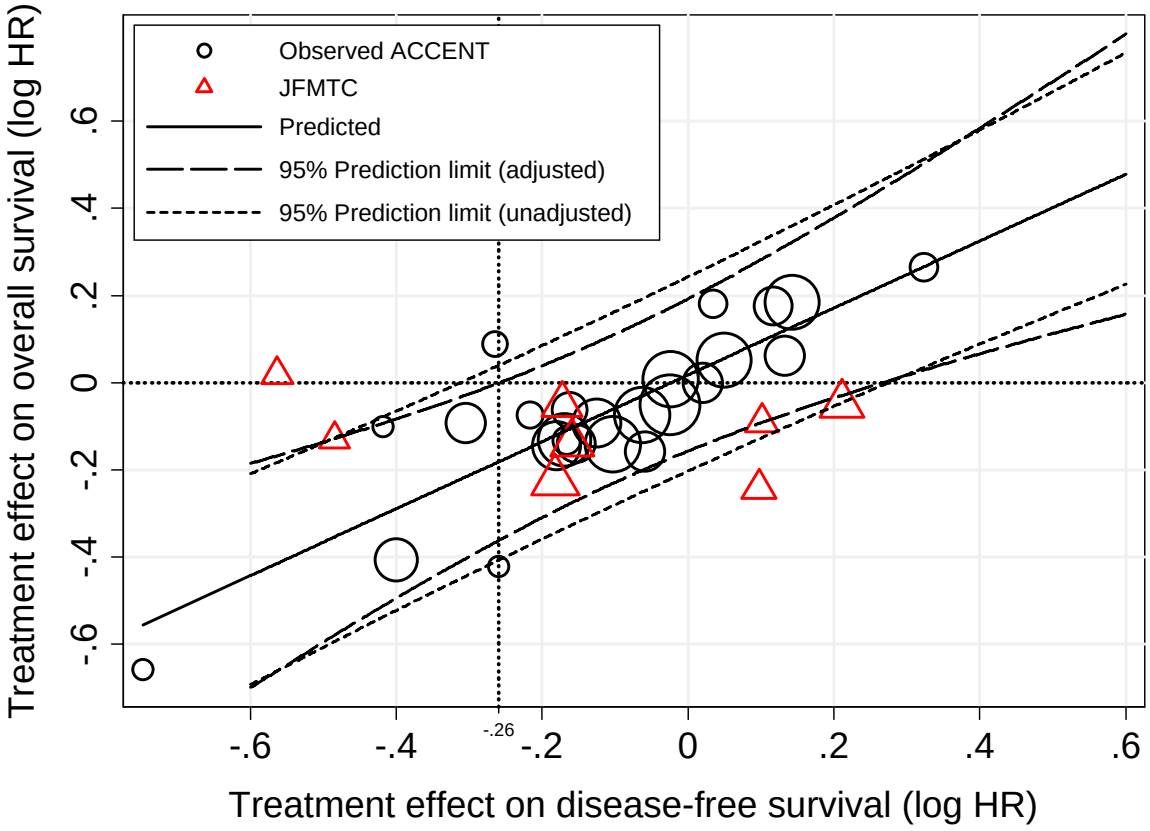

Figure 4: Correlation between treatment effects (log HR = log hazard ratios) on disease-free survival (DFS, censored at 3 years) and on overall survival (OS, censored at 5 years). Symbol size is proportional to the number of patients; the line of prediction and prediction limits come only from the ACCENT data.

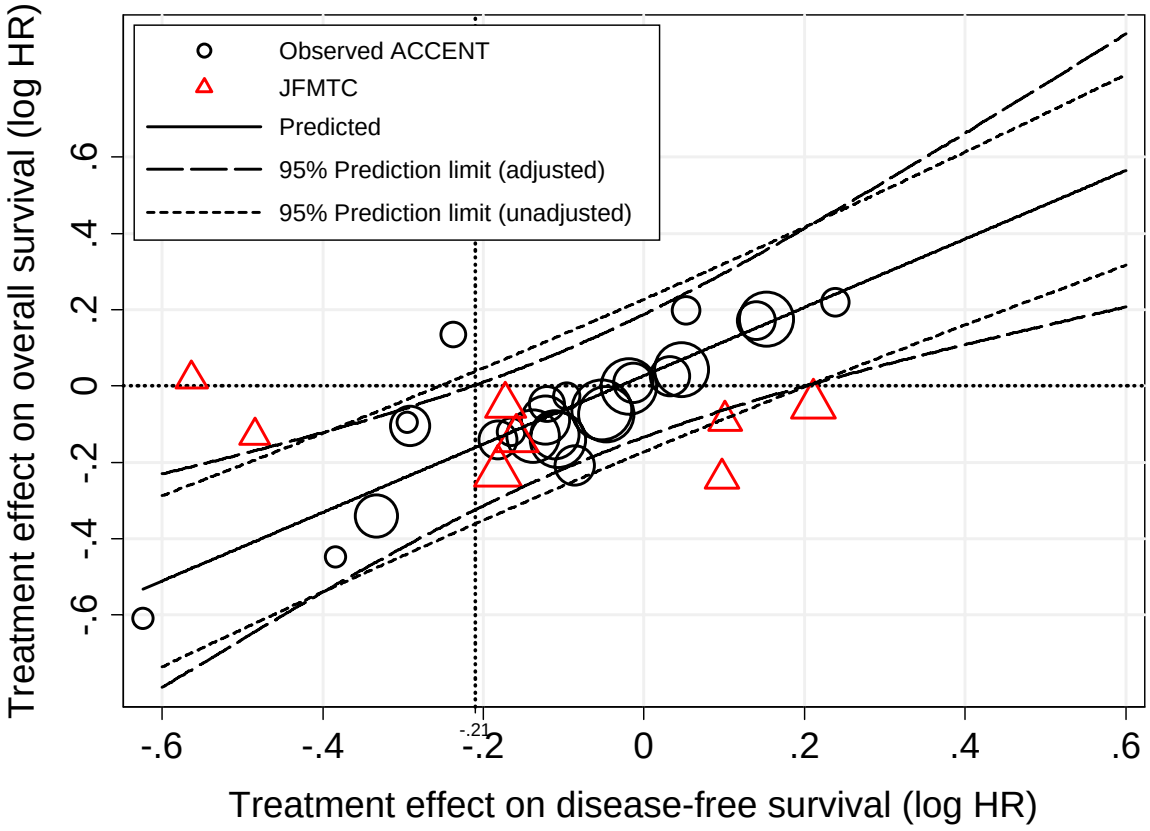

Figure 5: Predicted treatment effects (log HR = log hazard ratios) on overall survival (OS, censored at 5 years) for different censoring times of DFS. The red straight line is the y=x diagonal.