Individual- and trial-level surrogacy in colorectal cancer

Peer-reviewed author version

BUYSE, Marc; BURZYKOWSKI, Tomasz; Michiels, Stefan & Carroll, Kevin (2008) Individual- and trial-level surrogacy in colorectal cancer. In: STATISTICAL METHODS IN MEDICAL RESEARCH, 17(5). p. 467-475.

DOI: 10.1177/0962280207081864
Handle: http://hdl.handle.net/1942/8746

Statistical Methods in Medical Research 2008; 1–9

Individual- and trial-level surrogacy in colorectal cancer

Marc Buyse IDDI (International Drug Development Institute), Louvain-la-Neuve, Belgium and Center for Statistics, Hasselt University, Diepenbeek, Belgium, **Tomasz Burzykowski** Center for Statistics, Hasselt University, Diepenbeek, Belgium, **Stefan Michiels** Biostatistics and Epidemiology Unit, Institut Gustave Roussy, Villejuif, France and **Kevin Carroll** Oncology Therapy Area, Astra-Zeneca Research and Development, Macclesfield, UK

Two conditions must be fulfilled for an intermediate endpoint to be an acceptable surrogate for a true clinical endpoint: (1) there must be a strong association between the surrogate and the true endpoint, and (2) there must be a strong association between the effects of treatment on the surrogate and the true endpoint. We test whether these conditions are fulfilled for disease-free survival (DFS) and progression-free survival (PFS) on data from 20 clinical trials comparing experimental treatments with standard treatments for early and advanced colorectal cancer. The effects of treatment on DFS (or PFS in advanced disease) and OS were quantified through log hazard ratios (log HR), estimated through a Weibull model stratified for trial. The rank correlation coefficients between DFS and OS, and trial-specific treatment effects, were estimated using a bivariate copula distribution for these endpoints. A linear regression model between the estimated log hazard ratios was used to compute the "surrogate threshold effect", which is the minimum treatment effect on DFS required to predict a non-zero treatment effect on OS in a future trial. In early disease, the rank correlation coefficient between DFS and OS was equal to 0.96 (CI 0.95–0.97). The correlation coefficient between the log hazard ratios was equal to 0.94 (CI 0.87–1.01). The risk reductions were approximately 3% smaller on OS than on DFS, and the surrogate threshold effect corresponded to a DFS hazard ratio of 0.93. In advanced disease, the rank correlation coefficient between PFS and OS was equal to 0.82 (CI 0.82–0.83). The correlation coefficient between the log hazard ratios was equal to 0.99 (CI 0.94–1.04). The risk reductions were approximately 19% smaller on OS than on PFS, and the surrogate threshold effect corresponded to a PFS hazard ratio of 0.86. One trial with a large treatment effect on PFS and OS had a strong influence on the results in advanced disease. DFS (and PFS in advanced disease) are acceptable surrogates for OS in colorectal cancer.

1 Introduction

The validation of surrogate endpoints has been a topic of intense research and heated controversy over the last few years. Prentice laid the foundation of many subsequent efforts when he proposed a definition of, and validation criteria for, surrogate endpoints.[1] This paper covers the situation in which a surrogate endpoint (S) is proposed for a true endpoint (T), and data are available on both the surrogate and the true endpoints in a series of randomized clinical trials comparing an experimental treatment

Address for correspondence: Marc Buyse, IDDI, 30 Avenue Provinciale, 1340 Ottignies Louvain-la-Neuve, Belgium. E-mail: marc.buyse@iddi.com

10.1177/0962280207081864
Los Angeles, London, New Delhi and Singapore

(Trt) with a control treatment. Essentially, the Prentice criteria require that the following conditions be fulfilled (Figure 1): 1) the treatment has a statistically significant effect on the surrogate endpoint, 2) the surrogate endpoint has a significant impact on the true endpoint and 3) the effect of treatment on the true endpoint is 'fully captured' by the surrogate endpoint.[1,2]

Although the Prentice criteria were very useful to focus attention on the need for statistical criteria to be met before surrogate endpoints are used in practice, they were criticized for various reasons.[3] Several other avenues were subsequently explored to identify alternative approaches towards a practicable validation. The approach proposed in this paper is based on the strength of the association between the surrogate and the true endpoint (called the 'individual-level surrogacy'), and between the effects of treatment on the surrogate and the true endpoint (called the 'trial-level surrogacy').[4] Essentially, this approach requires that two simple conditions be fulfilled (Figure 2): 1) there is a strong association between the surrogate endpoint and the true endpoint and 2) there is a strong association between the effects of treatment on the surrogate endpoint and on the true endpoint.[4]

We illustrate this approach using data from clinical trials comparing experimental treatments with standard treatments for early[5] and advanced colorectal cancer.[6]

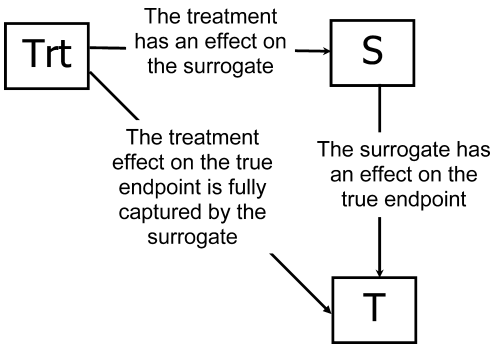

Figure 1 The 'full capture' approach to the validation of surrogate endpoints.

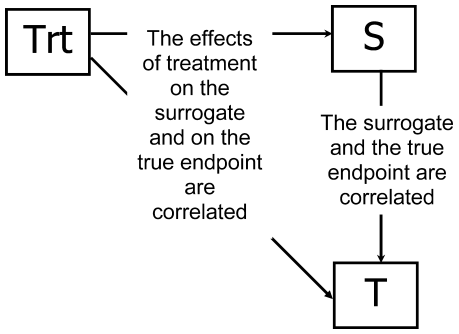

Figure 2 A correlation approach to the validation of surrogate endpoints.

Both of these situations are of interest because of the high incidence of colorectal cancer worldwide and the number of new agents that are being developed for this disease. In patients who are diagnosed early, surgical resection is possible, usually followed by adjuvant chemotherapy when lymph nodes are found to be involved at surgery. In patients diagnosed with advanced (metastatic) disease, several lines of chemotherapy are available. There has been considerable progress in chemotherapy for both early and advanced disease over the last decade, but many new drugs and biological agents still await clinical testing. Until recently, the overall survival (OS) was widely accepted as the primary endpoint to judge the efficacy of new treatments. However, it takes a long time to evaluate survival, especially in early disease, and given the availability of treatments that delay disease progression and death following progression, there has been considerable interest in replacing survival by earlier endpoints, such as disease-free survival (DFS) in early disease and progression-free survival (PFS) in advanced disease.[7]

2 Material and methods

2.1 Material

In early disease, data were available for 10 255 patients enrolled in 10 randomized trials comparing experimental treatments with control treatments.[5,8] In advanced disease, data were available for 3089 patients enrolled in 10 randomized trials comparing experimental treatments with control treatments.[6,8]

2.2 Survival analyses

Analyses were based on all randomised patients. In early disease, DFS was calculated from the time of randomization to first disease recurrence as defined in each individual trial, or death from any cause. In advanced disease, PFS was calculated from the time of randomization to first disease progression 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 distributions of time to events were estimated using the Kaplan–Meier method. The effects of treatment on DFS, PFS and OS were quantified through log hazard ratios (log HR), estimated through a proportional hazards (Weibull) model, stratified for trial, with treatment as the only factor.[9]

For small treatment effects, log HR $\approx 1 - \text{HR}$; hence log HR is an approximate estimate of the risk reduction due to experimental therapy. All confidence intervals (CI) had a 95% coverage.

2.3 Surrogacy criteria

A correlation approach was used to assess DFS or PFS as potential surrogates for OS.[4] The analysis proceeded in two stages: 1) the rank-correlation coefficients between DFS or PFS and OS, and trial specific treatment effects, were estimated using a bivariate copula distribution for these endpoints and 2) the correlation coefficient between the treatment effects on DFS or PFS and on OS was computed using an ordinary linear regression fitted on the treatment effects estimated in the first stage.[10] The use of a

copula model allows one to take into account the association between DFS and OS when estimating the treatment effects. At the same time, it naturally provides a measure of the strength of the association that takes into account both trial and treatment effects. The copula providing the best fit to the data as determined by Akaike's Information Criterion was chosen, respectively, a Plackett copula in early disease and a Hougaard copula in advanced disease.[10] In advanced disease, one trial exhibited extreme treatment benefits in terms of both PFS and OS, and therefore a robust linear regression model (using trimmed least squares) was also fitted by minimizing the sum of the smallest 5 ($=N/2$) least squared residuals, without adjustment for estimation errors.[11]

In the approach adopted in this paper, DFS or PFS could be claimed to be an acceptable surrogate endpoint for OS if 1) there was a strong correlation between the endpoints and 2) there was a strong correlation between the treatment effects on the endpoints. A linear regression model between log HRs was used at the second stage of the two-stage approach to compute the 'surrogate threshold effect', which is the minimum treatment effect on DFS or PFS required to predict a nonzero treatment effect on OS in a future trial. The surrogate threshold effect was given by the intersection of the 95% prediction limits obtained from the model and the x-axis (corresponding to no treatment effect on OS).[12] Two versions of the regression model and prediction limits were constructed: without and with adjustment for the estimation error present in the treatment effects estimated using a bivariate copula model at the first stage of the two-stage approach.[12] Note that the limits were computed assuming a perfect knowledge of the treatment effect on DFS or PFS, thus ignoring variability of the estimation in a future trial (but accounting for the estimation error of the linear regression model). Thus, they can be interpreted as providing the minimum width of a prediction interval for the treatment effect on OS. In order to account for the variability of the estimated treatment effect on DFS or PFS, one would require the appropriate limit (upper or lower) of the CI for this estimate to fall below or above the surrogate threshold effect.

3 Results

3.1 Early disease

Figure 3 shows the DFS and OS curves by treatment group: experimental (solid lines) versus control (dotted lines). The rank-correlation coefficient between DFS and OS was equal to 0.96 (CI 0.95–0.97).

The correlation coefficient between the log HRs was equal to 0.94 (CI 0.87–1.01). Figure 4 shows the linear regression line used to predict treatment effects on OS from the observed treatment effects on DFS. The regression equation was $\log \mathrm{HR}_{\mathrm{OS}} = 0.02 + 0.97 \times \log \mathrm{HR}_{\mathrm{DFS}}$, indicating that the risk reductions were $\sim$3% ($=1 - 0.97$) smaller on OS than on DFS.

The surrogate threshold effect (based on the measurement-error corrected prediction limits) corresponded to a DFS HR of 0.93 (for a beneficial treatment) or 1.07 (for a harmful treatment). Thus, in order to predict a non-zero treatment effect on OS in a future trial, a HR of at most 0.93 or at least 1.07 would need to be ascertained. In order to account for the variability of the estimation in a future trial, one would therefore

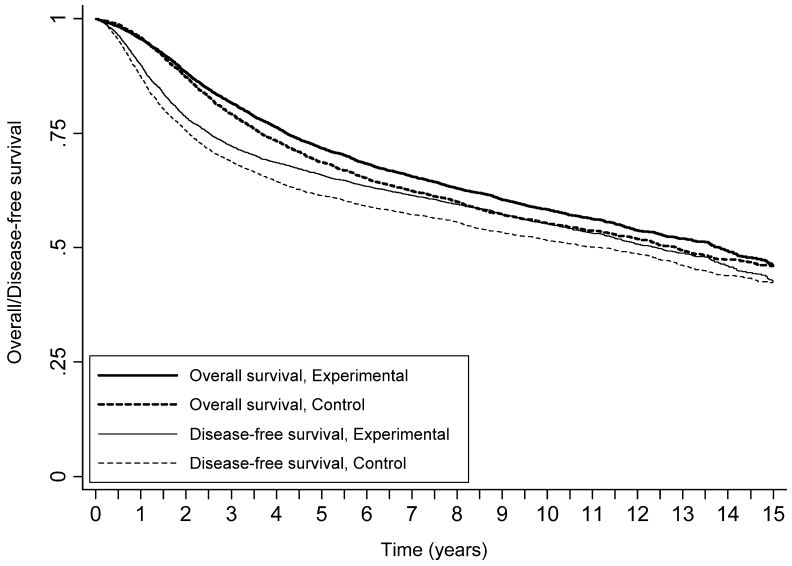

Figure 3 Kaplan–Meier estimates of OS and DFS in resectable colorectal cancer.

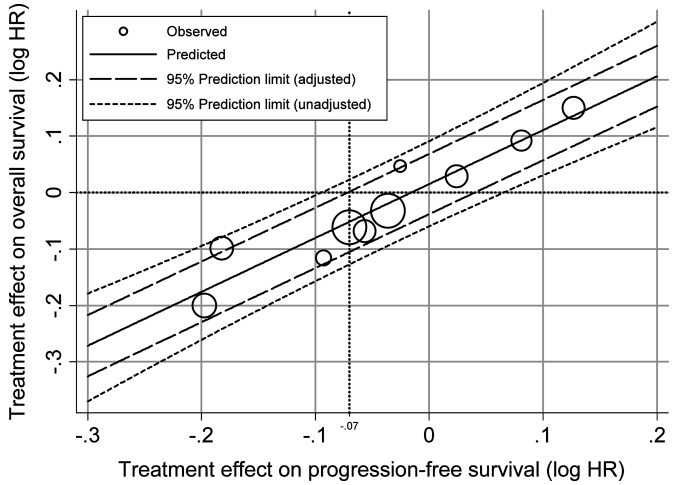

Figure 4 Correlation between treatment effects (log HR = log hazard ratio) on DFS and on OS in resectable colorectal cancer. Symbol size is proportional to the number of patients.

require the upper limit of the CI of the estimated HR to fall under 0.93, or the lower limit above 1.07.

3.2 Advanced disease

Figure 5 shows the PFS and OS curves by treatment group: experimental (solid lines) versus control (dotted lines). The rank-correlation coefficient between PFS and OS was equal to 0.82 (CI 0.82–0.83).

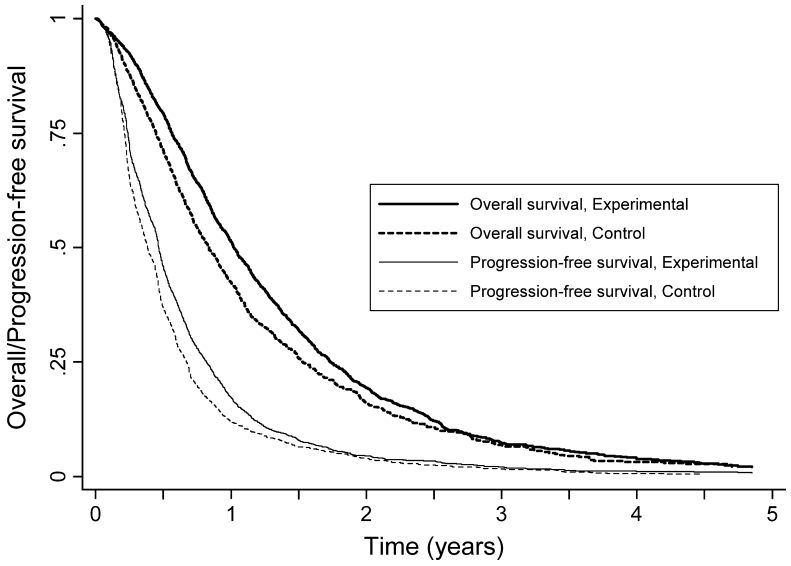

Figure 5 Kaplan–Meier estimates of OS and PFS in advanced colorectal cancer.

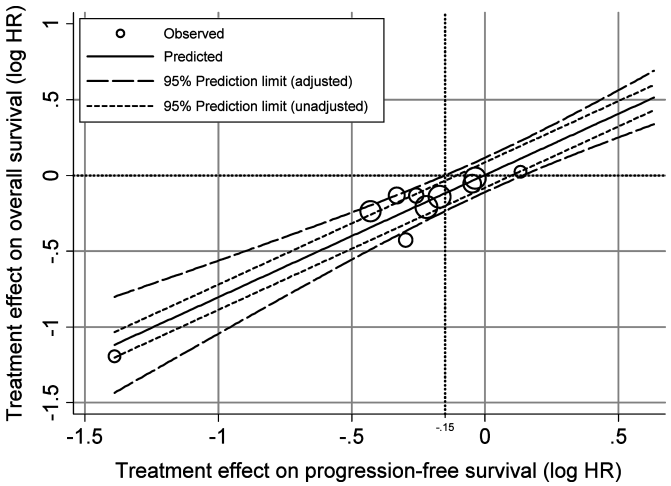

Figure 6 Correlation between treatment effects (log HR = log hazard ratio) on PFS and on OS in advanced colorectal cancer. Symbol size is proportional to the number of patients.

The correlation coefficient between the log HRs was equal to 0.99 (CI 0.94–1.04). Figure 6 shows the linear regression line used to predict treatment effects on OS from the observed treatment effects on PFS. The regression equation was $\log \mathrm{HR}_{\mathrm{OS}} = 0.03 + 0.81 \times \log \mathrm{HR}_{\mathrm{PFS}}$, indicating that the risk reductions were ~19% (=1 − 0.81) smaller on OS than on PFS.

The surrogate threshold effect (based on the measurement-error corrected prediction limits) corresponded to a PFS HR of 0.86 (or 1.16). Thus, in order to predict a non-zero treatment effect on OS in a future trial, a HR of at most 0.86 (or at least 1.14) would need to be ascertained.

Two trials appeared atypical in this analysis. One of the trials had randomized 310 patients and had a very long follow-up but surprisingly few events. Exclusion of this trial did not lead to any noticeable difference in model parameters. Another trial had randomized 150 patients and exhibited extreme treatment benefits in terms of both PFS and OS (Figure 6). Exclusion of this influential trial resulted in a much weaker association between the treatment effects, with a correlation coefficient R equal to 0.74 (CI 0.44–1.04) and a surrogate threshold effect of 0.77 (or 1.30). The two atypical trials were carefully scrutinized, but no obvious defect or methodological problem could be found to explain their discrepant results, and therefore their *post-hoc* exclusion does not seem justified. Applying a robust regression model to all 10 trials, the correlation coefficient was 0.88 and the regression equation was $\log \mathrm{HR}_{\mathrm{OS}} = 0.04 + 0.45 \times \log \mathrm{HR}_{\mathrm{PFS}}$, suggesting a much larger attenuation of the predicted treatment effects on OS.

4 Discussion

There is currently no consensus regarding the theoretical conditions required for a surrogate endpoint to be useful in practice, let alone 'valid'. In addition, some question whether a true statistical validation will ever be possible, casting considerable doubts on any attempts to identify surrogates.[13] In this paper, we show that quantitative approaches can be used to shed light on the potential value of candidate surrogate endpoints. In doing so, we also identify a few essential requirements for the validation process.

In early colorectal cancer, our analyses show that DFS is an excellent surrogate for OS, since these endpoints are tightly correlated, as are the effects of 5FU-based experimental treatments upon these endpoints. These analyses complement those previously published on the same dataset, which showed three-year DFS to be an excellent surrogate for five-year OS.[5] In advanced colorectal cancer, our analyses also show that PFS is a good surrogate for OS. These results are at variance with those we published previously, based on two trials that investigated the efficacy of interferon-α in advanced colorectal cancer.[10] In these trials, there was a very poor correlation between the treatment effects in the participating sites considered as the units of analysis. However, interferon-α had no overall effect on either PFS or OS, which suggests that trial-level surrogacy can only be established under departures from the null hypothesis, as proposed by Prentice long ago.[1] In another set of trials, tumour shrinkage ('response') was found not to be an acceptable surrogate for OS, in spite of highly significant treatment effects on response. Indeed, although patients who achieved a response had a significantly prolonged OS (indicating adequate individual-level surrogacy), treatment effects on response were very poorly correlated with treatment effects on OS (suggesting inadequate trial-level surrogacy).[14]

The analyses presented in this paper underscore the importance of the two levels of surrogacy. Individual-level surrogacy reflects the natural history of the disease and is useful for patient management. Trial-level (or group-level) surrogacy reflects the treatment mechanism and is useful to predict therapeutic benefits. The question of which of these levels is more important is not resolved, but it is likely that an acceptable surrogate needs to meet some minimum requirements at both levels. Further analyses in various medical fields will provide a better understanding of the role and limitations of the statistics that we suggest here to quantify surrogacy.

One caveat about the present analyses in advanced disease is that the results were very dependent on the presence of one trial with extreme treatment effects. When this trial was excluded, the treatment effects were still correlated, but less impressively so ($R = 0.74$). When robust regression was used instead of ordinary linear regression, the slope of the regression line decreased and suggested a larger attenuation of treatment effects on OS predicted from the treatment effects on PFS. These findings lead us to hypothesize that for a validation to be effective, a range of treatment effects is desirable on both the surrogate and the true endpoints. As a consequence, it may be desirable to carry out the validation using trials that test several treatments with different efficacy, in order to show that the relationship between treatment effects on the surrogate and true endpoints are not treatment-dependent. This sets the bar for an effective validation very high, since one ideally requires a set of randomized experiments testing different treatments for the disease under consideration.

In the present paper, we also show that the trial-level correlation approach naturally leads to quantifying the predictive value of a surrogate endpoint. In early disease, the surrogate threshold effect was 0.93, and therefore an adjuvant treatment that showed at least a 7% reduction in risk of tumour recurrence would be expected to yield a significant treatment effect on OS in a future trial. In advanced disease, the surrogate threshold effect was 0.86 and therefore a treatment that showed at least a 14% reduction in risk of tumour progression would be expected to yield a significant treatment effect on OS in a future trial. In advanced disease, these results are sensitive to exclusion of a single trial, with the surrogate threshold effect going down to a HR of 0.77, suggesting the need for larger (but still achievable) treatment effects on PFS. HRs in the range of 0.8 for PFS are realistic and have, in fact, been achieved by several treatments recently approved for the treatment of advanced colorectal cancer.[6]

The prediction approach presented in this paper can be used for the purposes of designing a trial using a surrogate rather than a true endpoint. One could use the surrogate threshold effect to determine the sample size required in a future trial to predict a benefit on the true endpoint, allowing for the estimation error in the surrogate. An important limitation of such an approach is that the surrogate threshold effect assumes that the relationship observed so far will hold true in the future. This may or may not be true for agents yet to be tested in future trials. In oncology, biologicals have substantially different modes of action than the cytotoxic chemotherapies used in the trials analysed here. Hence, it might be useful to validate our current results using data from trials of new agents, for instance the monoclonal antibodies targeting the vascular endothelial growth factor (bevacizumab) and epidermal growth factor receptor (cetuximab), both of which have recently been approved, respectively, for the first-line and second-line treatment of metastatic colorectal cancer.[6]

Acknowledgments

The authors are grateful to the MAGIC and ACCENT collaborators for the permission to use their data[8] and to Dr Stuart Baker for very constructive comments on draft versions of this paper. Dr Burzykowski acknowledges financial support from the IAP research network nr P6/03 of the Belgian government (Belgian Science Policy).

References

1 Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine* 1989; **8**: 431–40.
2 Buyse M, Molenberghs G. Criteria for the validation of surrogate end-points in randomized experiments. *Biometrics* 1998; **54**: 1014–29.
3 Burzykowski T, Molenberghs G, Buyse M ed. *Evaluation of Surrogate Endpoints*. Springer Verlag, 2005.
4 Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomised experiments. *Biostatistics* 2000; **1**: 49–68.
5 Sargent D, Wieand S, Haller DG, Gray R, Benedetti J, Buyse M, Labianca R, Seitz JF, O'Callaghan CJ, Francini G, Grothey A, O'Connell M, Catalano PJ, Blanke CD, Kerr D, Green E, Wolmark N, Andre T, Goldberg RM, de Gramont A. 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* 2005; **23**: 8664–70.
6 Buyse M, Burzykowski T, Carroll K, Michiels S, Sargent D, Miller LL, Elfring GL, Pignon JP, Piedbois P. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *Journal of Clinical Oncology* 2007; 25: 5218–24.
7 DiLeo 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. *Annals of Oncology* 2004; **14**: 545–9.
8 Burzykowski T. Introduction. (To appear in this volume of *Statistical Methods in Medical Research*).
9 Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**: 187–220.
10 Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate endpoints in multiple randomised clinical trials with failure-time endpoints. *Journal of the Royal Statistical Society, Series C* (*Applied Statistics*) 2001; **50**: 405–22.
11 Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*, John Wiley & Sons, 1987.
12 Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics* 2006; **5**: 173–86.
13 Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine* 1996; **125**: 605–13.
14 Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P, for the Meta-Analysis Group In Cancer. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. *Lancet* 2000; **356**: 373–8.