Supplementary material

Effect of the weighting on the amount of imbalance between the treatment groups

In order to quantify the amount of imbalance between the two treatment groups, balance diagnostics were calculated. A commonly used diagnostic is the standardized difference [Austin, 2009]. For continuous variables, the standardized difference is defined as

$$d = \frac{(\bar{x}_A - \bar{x}_B)}{\sqrt{\frac{s_A^2 + s_B^2}{2}}}, \tag{1}$$

which measures the difference in means in units of the pooled standard deviation. Here $\bar{x}_A$ and $\bar{x}_B$ are the sample means in the two treatment groups and s_A^2 and s_B^2 are the sample variances. This diagnostic is often used to see if after matching, baseline covariates are more equally distributed across the treatment groups.

The above formula was extended in order to apply the same idea to weighted data instead of matched data. The extension consists of incorporating the case weights into the formula, and taking into account that the sizes of the treatment groups need not be equal.

The weighted standard deviation was defined as

$$s_w = \sqrt{\frac{\sum_{i=1}^{n} w_i (x_i - \hat{\mu}^*)^2}{\mathrm{V}_1 - (\mathrm{V}_2/\mathrm{V}_1)}} \tag{2}$$

where $\hat{\mu}^* = \frac{\sum_i w_i x_i}{\sum_i w_i}$ is the weighted arithmetic mean, and

$$\mathrm{V}_1 = \sum_{i=1}^{n} w_i \tag{3}$$

$$\mathrm{V}_2 = \sum_{i=1}^{n} w_i^2 \tag{4}$$

are the sums of the weights and the squared weights, respectively.

The pooled version of the weighted standard deviations $s_{w,A}$ in treatment A and $s_{w,B}$ in treatment group B was then defined as

$$s_{w,\text{pooled}} = \sqrt{\frac{(\mathrm{V}_{1,A} - \mathrm{V}_{2,A}/\mathrm{V}_{1,A}) s_{w,A}^2 + (\mathrm{V}_{1,B} - \mathrm{V}_{2,B}/\mathrm{V}_{1,B}) s_{w,B}^2}{(\mathrm{V}_{1,A} - \mathrm{V}_{2,A}/\mathrm{V}_{1,A}) + (\mathrm{V}_{1,B} - \mathrm{V}_{2,B}/\mathrm{V}_{1,B})}} \tag{5}$$

and the weighted standardized difference as

$$d_w = \frac{\hat{\mu}_A^* - \hat{\mu}_B^*}{s_{w,\text{pooled}}}. \tag{6}$$

For categorical variables, the standardized difference was calculated by coding one category as a 1 and the other as a zero. For categorical variables with three or more category levels, the difference was calculated separately for each category compared to the other categories combined. For example, the outcome of the surgery had categories ‘no residual tumour’, ‘residual tumour $<$ 1 cm’, and ‘residual tumour $>$ 1 cm’. So for this variable three differences were calculated, each comparing one of the categories with the two other categories.

The results presented in table S1 are for observed observations only (no imputation done). A value of NaN in this table indicates that there were no patients in that group for whom a weight could be calculated because of missing values. Results of the balance diagnostics across multiple imputations are presented in table S2.

Table S1: Standardized difference between the groups before and after weighting (complete observations)

Variable or category	Standardized difference without weighting	Standardized difference after weighting
Age	-0.3229	-0.0133
BMI	0.0335	0.0397
Post-menopausal	-0.3082	-0.2487
World Health Organization (WHO) performance status	-0.2787	-0.0307
Tumour grade by Silverberg	-0.0745	-0.0382
Histology: sereus adenocarcinoom	-0.1908	0.0370
Histology: mucineus adenocarcinoom	0.1262	-0.1305
Histology: endometriod adenocarcinoom	0.3823	0.2531
Histology: clearcell adenocarcinoom	0.1097	-0.0741
Histology: undifferentiaded adenocarcinoom	-0.2238	-0.1214
Histology: mixed epithelial adenocarcinoom	0.0425	NaN
Histology: other epithelial tumor	0.1865	NaN
Serum CA-125 before treatment	-0.1379	-0.0233
Amount of ascites before treatment	-0.1228	0.0082
Clinical stage	-0.5165	-0.0388
Metastatic tumour size	-0.0153	0.0667
Residual tumour after surgery: no macroscopic residual tumour	-0.3355	-0.1904
Residual tumour after surgery: residual tumour < 1 cm	-0.3487	-0.2763
Residual tumour after surgery: residual tumour > 1 cm	0.8018	0.4846

Table S2: Standardized difference between the groups after weighting (averaged across multiple imputations)

Variable or category	Mean standardized difference after weighting
Age	-0.0130
BMI	0.0214
Post-menopausal	-0.0672
World Health Organization (WHO) performance status	-0.0637
Tumour grade by Silverberg	0.0078
Histology: sereus adenocarcinoom	-0.0663
Histology: mucineus adenocarcinoom	0.0567
Histology: endometriod adenocarcinoom	0.2887
Histology: clearcell adenocarcinoom	0.0091
Histology: undifferentiaded adenocarcinoom	-0.2738
Histology: mixed epithelial adenocarcinoom	0.0107
Histology: other epithelial tumor	0.1952
Serum CA-125 before treatment	0.0414
Amount of ascites before treatment	-0.0757
Clinical stage	-0.0857
Metastatic tumour size	-0.0157
Residual tumour after surgery: no macroscopic residual tumour	-0.3985
Residual tumour after surgery: residual tumour < 1 cm	-0.3642
Residual tumour after surgery: residual tumour > 1 cm	0.8472

Histograms of the Propensity Scores

The distribution of the propensity scores across the two treatment goups was inspected graphically using histograms [Garrido et al., 2014].

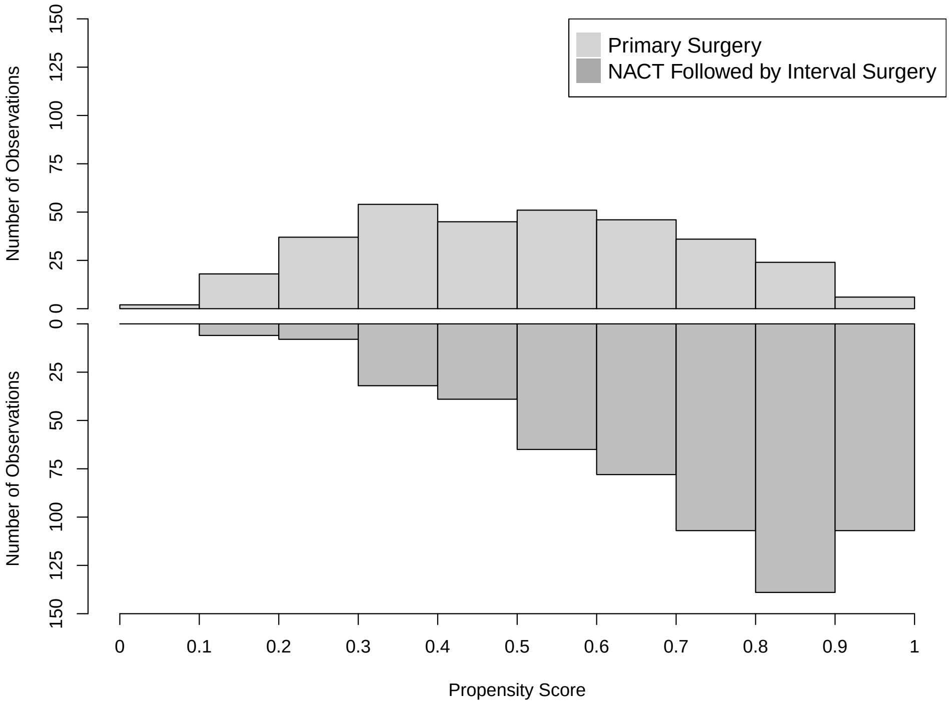

Figure S1: Histogram of propensity scores per treatment group

Overall survival by treatment per biomarker subgroup without imputation or weighting

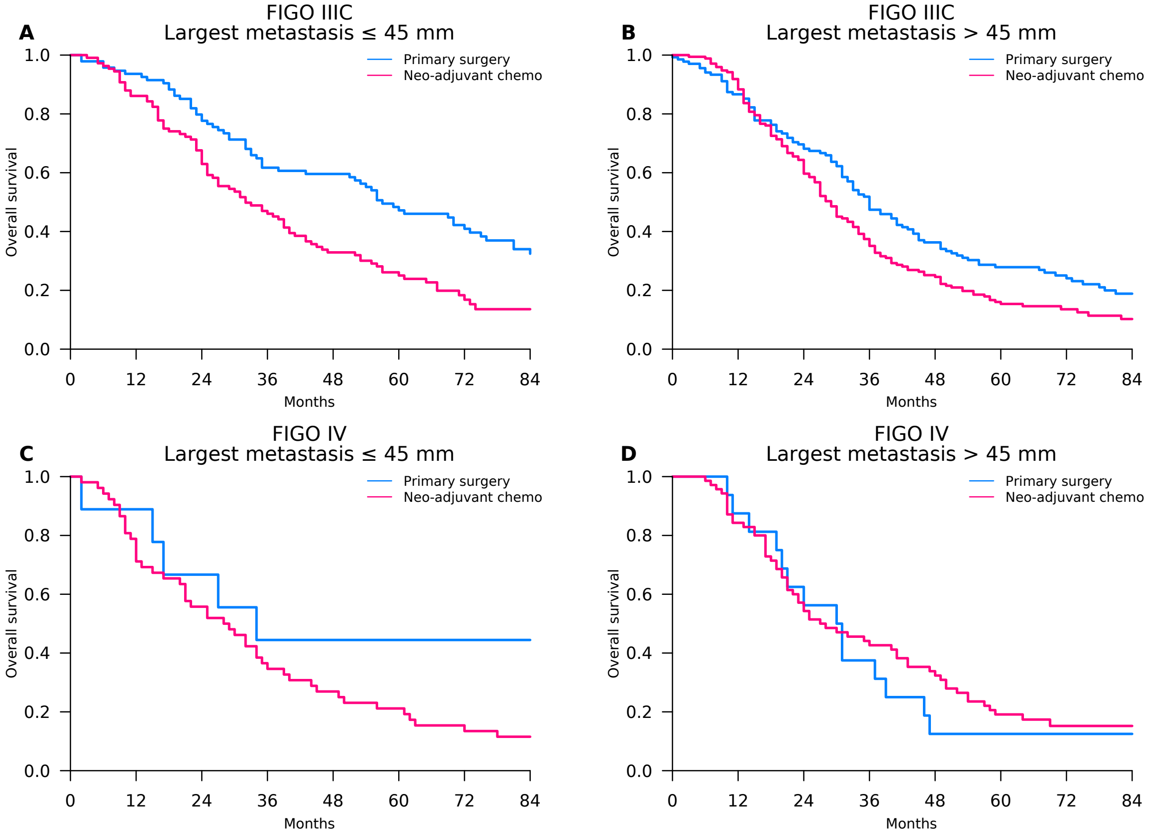

Figure S2: Overall survival by treatment group in the biomarker subgroups without imputation or weighting, using only patients with known metastatic tumour size

A. Subgroup of patients with FIGO stage IIIC and metastatic size $\leq$ 45mm, HR 1.87, 95% CI (1.35–2.59) p=0.0002;

B. Subgroup of patients with FIGO stage IIIC and metastatic size $>$ 45mm, HR 1.37, 95% CI (1.07–1.75) p=0.013;

C. Subgroup of patients with FIGO stage IV and metastatic size $\leq$ 45mm, HR 1.75, 95% CI (0.74–4.10) p=0.20;

D. Subgroup of patients with FIGO stage IV and metastatic size $>$ 45mm, HR 0.90, 95% CI (0.50–1.62) p=0.72

The heterogeneity of the treatment effect across the four subgroups was statistically significant (p-value for interaction 0.025).

References

[Austin, 2009] Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28:3083–3107.

[Garrido et al., 2014] Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., and Aldridge, M. D. (2014). Methods for constructing and assessing propensity scores. *Health services research*, 49:1701–1720.