Statistical modelling of time-varying covariates for survival data

Spreafico, M.

Chapter 7

Investigating the causal effects of joint-exposure on survival outcome in presence of time-varying confounders

The content of this chapter is based on the work by M. Spreafico, C. Spitoni, C. Lancia, F. Ieva and M. Fiocco "Causal effects of chemotherapy regimen intensity on survival outcome in osteosarcoma patients through Marginal Structural Cox Models" 2022.

Although multidisciplinary management of chemotherapy has improved clinical outcomes in patients with osteosarcoma, over the past 40 years there have been no further improvements in survival [15]. The strongest prognostic factor of both event-free and overall survival known so far in osteosarcoma is Histological Response (HRe) [31], i.e., improvement in the appearance of microscopic tissue specimens in a patient after pre-operative chemotherapy, whereas the impact of chemotherapy dose modification on patients' survival is still unclear [111].

As mentioned in the previous chapters, in cancer trials the relationship between chemotherapy regimen intensity and survival is problematic to analyse due to the presence of negative feedback between exposure to cytotoxic drugs and consequent toxic side effects. Chemotherapy is usually modelled by different allocated regimens, i.e., by Intention-To-Treat (ITT) analysis [70]. Since ITT ignores anything that happens after randomization, such as protocol deviations or changes in drug intake over time [110], the Received Dose Intensity (RDI) [86] indicator has been introduced to analyse how close the *actual treatment* delivered is to the planned treatment, marking a significant departure from ITT in the direction of a closer description of the actual clinical practice. Lancia *et al.* (2019) [111] showed that there is mismatch between target and achieved chemotherapy-RDI in osteosarcoma due to toxic side effects developed by patients through therapy. Toxicities affect subsequent exposure by delaying the next cycle or reducing chemotherapy doses [112], representing one of the principal reasons for treatment discontinuation [186]. Being at the same time risk factors for mortality and predictors of future exposure levels, toxicities hence represent *time-dependent confounders* for the effect of chemotherapy on patient's survival. In the presence of confounders, classical survival approaches, such

as the Cox model [46], have limitations in causally interpreting the hazard ratio of the treatment variable, even if the treatment assignment is randomized [2].

Time-dependent confounding of the exposure-outcome association represents a specific challenge for estimating the effect of a treatment on an outcome of interest. Standard analyses fail to give consistent estimators in the presence of time-varying confounders if those confounders are themselves affected by the treatment [48]. For this reason, different statistical methods to control for exposure-affected time-varying confounding have been proposed, including, among others, g-computation formula [169], g-estimation of structural nested models [170] or Marginal Structural Models (MSMs) estimated using Inverse Probability of Treatment Weighting (IPTW) [171]. In case of time-to-event outcomes, Clare *et al.* (2019) [39] found that the *Cox-type Marginal Structural Model* (Cox MSM, or marginal structural Cox model) approach is by far the most commonly used method in practice.

Cox MSMs were introduced by Hernán *et al.* (2000) [78] as a class of methods for estimating the causal effect of therapy modifications on survival in presence of time-dependent confounders through IPTW. Making use of marginal (population average) rather than conditional hazard models [105], Cox MSMs target *counterfactual* (or *potential*) time-to-event variables, i.e., variables indicating when an event would have been observed if the patient had been administered a specific exposure level. IPTW is a propensity score-based method that creates a pseudo-population by weighting each subject with the inverse probability of observing a certain treatment allocation given past exposure and confounders. In such a new pseudo-population, confounders no longer predict exposure and the causal effects of treatment modifications on survival can be just obtained by a crude analysis. IPTW construction requires a thoughtful process that includes the determination of an adequate set of confounding covariates which enter into the decision-making process of allocating a treatment modification and on which the four main assumptions of causal inference with MSMs (i.e., no unmeasured confounding, consistency, positivity, no model misspecification) [77] can be tolerated [41]. Compared to a standard propensity score matching, IPTW has the advantages of retaining all eligible patients in the analysis, which may be preferred if there are limitations in terms of sample size, as well as the ability to include more than two treatment comparisons simultaneously [10].

Motivated by a clinical question concerning the effect of changes in therapy intensity on survival for osteosarcoma patients, in this chapter treatment-administration data are used to assess the causal effect on Event-Free Survival (EFS) of chemotherapy-exposure seen in terms of both (i) improvement in the appearance of microscopic tissue specimens in a patient after pre-operative treatment, i.e., by HRe, and (ii) reductions in actual versus anticipated/planned dose intensity, i.e., by RDI reductions. Data from the control arms of two clinical trials of chemotherapy in osteosarcoma, namely, European Osteosarcoma Intergroup studies MRC BO03/EORTC 80861 [120] and MRC BO06/EORTC 80931 [119] are analysed. These data are complex because the drug administration is longitudinal while only the most severe side-effects are recorded. The analysis of such mixed longitudinal/non-longitudinal data requires both an original analytical strategy and an unconventional model formulation. Moreover, since adjustments in treatment allocation

are determined by the overall toxic burden of each patient, the different types and number of side effects must be adequately summarized and quantified [190]. Suitable IPTW-based techniques and Cox MSMs are hence designed to mimic a randomized trial where joint-exposure intensity is no longer confounded by toxicities or other confounders, and a crude analysis suffices to estimate the causal effect of exposure modifications. This requires (i) a proper (time-dependent) definition of the joint-exposure, (ii) a tailor-made identification of all possible (time-dependent) confounders, and (iii) a suitable characterisation of the causal structure of the chemotherapy data. In particular, two alternative definitions of joint-exposure, based on *time-fixed final RDI* or *time-dependent pre/post-operative RDI* [120] combined with HRe, are proposed along with their relative confounding factors and Direct Acyclic Graphs (DAGs) [67, 77] to characterized the causal exposure-confounders-outcome structure. To the best of our knowledge, this is the first application of IPTW-based techniques to survival data from randomized trials of chemotherapy in order to eliminate the *toxicity-treatment-adjustment* bias.

The aim of this chapter is hence presenting an all-round RDI-based analysis of complex chemotherapy data, with tutorial-like explanations of the difficulties encountered and the problem-solving strategies deployed. Data from BO03 and BO06 trials are presented in Section 7.1. The process of building proper causal models based on joint-exposure (difficult due to lack of longitudinal confounders) using two alternative strategies is shown in detail in Section 7.2. Sections 7.3 and 7.4 are devoted to discussing the Cox MSMs results, in contrasts with their standard Cox analogues fitted on the unweighted original population, and drawing final conclusions, respectively.

7.1. Data description

Data from control arms (i.e., conventional regimen *Reg-C*) of the Randomized Controlled Trials (RCTs) MRC BO03/EORTC 80861 and MRC BO06/EORTC 80931 (*International Standard Randomised Controlled Trial Number*: ISRCTR 11824145 and ISRCTR 86294690 respectively, `https://www.isrctn.com`) were analysed. Both RCTs were funded by the Medical Research Council (MRC) (`https://www.ukri.org/councils/mrc/`) and the European Organisation for Research and Treatment of Cancer (EORTC) (`https://www.eortc.org`). In both trials, control arms were characterized by the standard European Osteosarcoma Intergroup (EOI) treatment structured in 6 cycles of 3-weekly Cisplatin (CDDP) (100 mg/m^2) plus Doxorubicin (DOX) (75 mg/m^2), and compared to a different therapy regimen (i.e., variant of Rosen's T10 regimen [178] in BO03 and a 2-weekly dose-intensified version of CDDP+DOX [119] in BO06). Results of the primary analyses on BO03 and BO06 data can be found in Lewis *et al.* (2000; 2007) [120, 119].
In Section 7.1.1 the selected cohort of patients from BO03 and BO06 trials is illustrated. Longitudinal chemotherapy data and patient characteristics are presented in Section 7.1.2.

7.1.1. Control arms protocol and Cohort selection

As the control arms design in Figure 7.1 shows, in both RCTs chemotherapy was administered before and after surgical removal of the primary osteosarcoma. At the end of the pre-operative treatment, with a nominal duration of 3 cycles in BO03 and 2 in BO06, the tumour was surgically resected, and the levels of tumour necrosis and HRe evaluated. Variations to the planned surgery-schedule happened quite often due to administrative reason (delayed surgery) or disease progression (premature surgery), in a limited number of cases surgery was delayed due to haematological toxicity (low platelets count). Post-operative chemotherapy was intended to resume 2 weeks after surgery.

Originally, 444 patients were enrolled in the control arms of BO03 (199) and BO06 (245). In this sample, 106 (23.9%) patients were excluded due to missing HRe. Of the remaining 338 patients, 58 terminated the chemotherapy treatment prematurely or without surgery, while 4 completed the treatment but experienced an event throughout. The final cohort of 276 patients (114 from BO03 and 162 from BO06, respectively) included in the analyses (62.2% of the initial sample) is shown in the consort diagram in Figure 7.2.

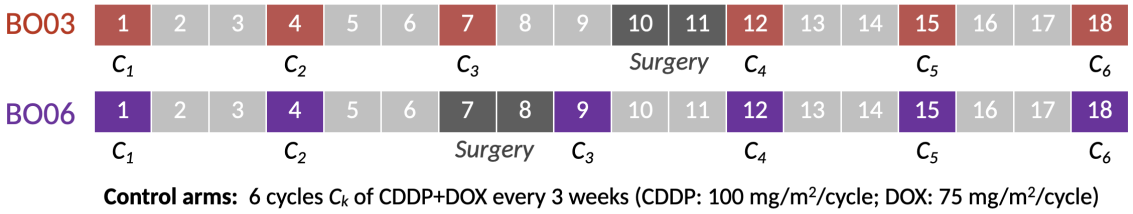

Figure 7.1. Control arms design for BO03 and B006 randomised clinical trials, characterized by the standard European Osteosarcoma Intergroup treatment structured in 6 cycles of 3-weekly Cisplatin (CDDP) (100 mg/m^2) plus Doxorubicin (DOX) (75 mg/m^2).

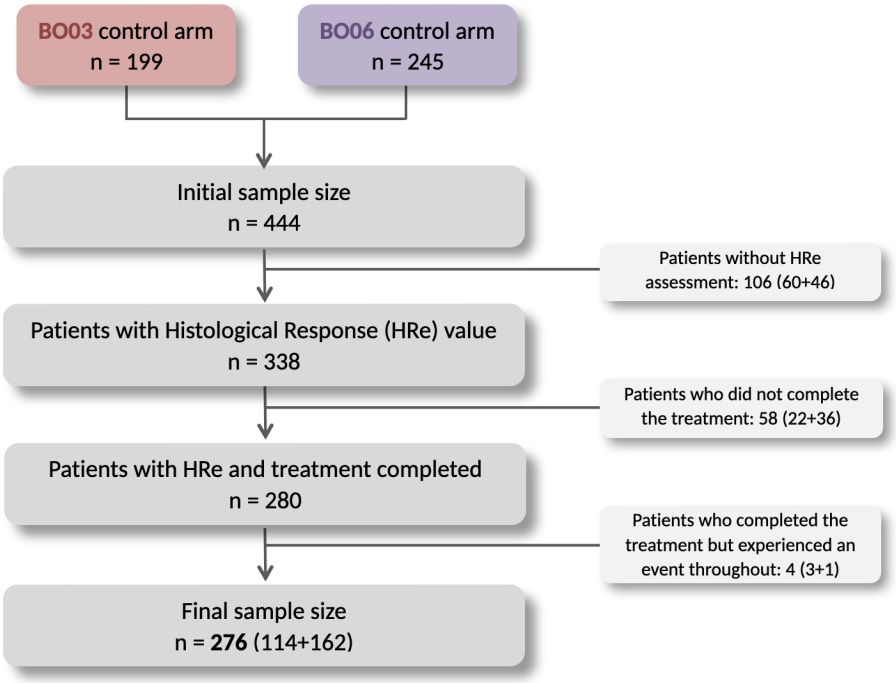

Figure 7.2. Flowchart of cohort selection.

7.1.2. Complexity of chemotherapy data

In cancer trials, therapy administration is usually complicated by the dynamical adjustment of the treatment on patients' clinical picture. Exposure to chemotherapy is likely to produce multi-systemic side effects, e.g. organ toxicity or myelosuppression. These side effects are a threat to patient's life and must be controlled by allocating either dose reductions/discontinuations or delays in the administration of the next course [112].

In BO03 and BO06 trials, case report forms were used to document across cycles all the information required by protocols for each patient. Patients baseline characteristics (age, gender, allocated chemotherapy regimen, site and location of the tumour) were registered at randomization. Therapy starting day was usually on the day of randomization or the day after, but could be postponed in case of administrative or clinical reasons. Treatment-related factors (administered dose of chemotherapy, cycles delays, haematological parameters, chemotherapy-induced toxicity and histological response to pre-operative chemotherapy) were collected prospectively during therapy.

A summary of baseline and trial characteristics over the entire dataset and by trial is shown in Table 7.1. Among 276 patients, 167 (60.5%) were males. Median age was 15.1 years (IQR [11.7; 18.2]). Therapy started *on time* in 71.0% of patients and surgery was performed *on time* since the start of the first cycle in 29.0% of patients.

In both studies, toxic side effects were recorded using the Common Terminology Criteria for Adverse Events Version 3 (CTCAE v3.0) [208], with grades ranging from 0 (none) to 4 (life-threatening) (see Table 7.2). Toxicity were collected longitudinally in BO06 trial, whereas in BO03 only the highest CTCAE grade (i.e., the most severe) was recorded for each toxicity in both the pre-operative and post-operative periods. According to protocols, the following side effects were linked to specific dose reduction or delay rules: *leucopenia* (i.e., a decrease in the number of white blood cells), *thrombocytopenia* (i.e., a decrease in the number of neutrophils), *oral mucositis*, *ototoxicity*, *cardiotoxicity* and *neurotoxicity*. If different *rule-specific* conditions co-existed and more than one dose reduction (or cumulative delays) applied, the lowest dose (or the highest delays) calculated was employed. According to expert knowledge, although not directly related to a specific adjustment rule, the patient's *generic* conditions of *nausea/vomiting* and *infections* was also taken into account during therapy. Treatment adjustments were hence determined as a combination of overall toxic burden related to both rule-specific and generic conditions, representing the confounding mechanisms due to toxicities.

To let pre- and post-operative toxicities be properly considered as confounding covariates and included in the analyses, individual side effects had to be appropriately summarized in order to quantify the overall toxic burden. For this purpose, the longitudinal Multiple Overall Toxicity (MOTox) score [190] introduced in Chapter 5 can be exploited. Since toxicity data over cycles were not recorded for BO03, MOTox computation was based on pre-operative and post-operative periods, considering the highest CTCAE grade recorded for each toxicity during pre/post-operative cycles.

Table 7.1. Patients baseline and trial characteristics.

	All	**BO03**	**BO06**
Patients	276	114 (41.3%)	162 (58.7%)
Age [years]			
*child**	76 (27.5%)	26 (22.8%)	50 (30.9%)
*adolescent**	117 (42.4%)	49 (43.0%)	68 (42.0%)
*adult**	83 (30.1%)	39 (34.2%)	44 (27.1%)
Median [IQR]	15.1 [11.7;18.2]	16.0 [12.8;19.0]	14.6 [11.3;17.7]
Min/Max	3.6/37.5	4.7/32.6	3.6/37.5
Gender			
Female	109 (39.5%)	43 (37.7%)	66 (40.7%)
Male	167 (60.5%)	71 (62.3%)	96 (59.3%)
Starting day**			
on time (day 0-1)	196 (71.0%)	63 (55.3%)	133 (82.1%)
low-delay (day 2-3)	43 (15.6%)	23 (20.2%)	20 (12.3%)
delay (day ≥ 4)	37 (13.4%)	28 (24.5%)	9 (5.6%)
Median [IQR]	1 [0;2]	1 [0;3]	0 [0,1]
Min/Max	0/15	0/15	0/7
Surgery time‡**			
on time	80 (29.0%)	29 (25.4%)	51 (31.5%)
delayed	196 (71.0%)	85 (74.6%)	111 (68.5%)
Median [IQR]	11 [4;22]	14 [5.25;22]	10 [4;21]
Min/Max	-39/132	-39/103	-3/132

* Age groups were defined according to Collins *et al.* (2013) [43]: *child* (male: 0–12 years; female: 0–11 years), *adolescent* (male: 13–17 years; female: 12–16 years) and *adult* (male: 18 or older; female: age 17 years or older).
** Starting day since randomization date. P-value of two-sided Mann-Whitney U test for starting day in BO03 vs BO06: 7.571e-08; p-value of chi-squared test among starting day category and trial: 1.096e-06.
‡ Surgery time (i.e., days since start of the first cycle) with respect to schedule is considered *on time* if performed from at most at the end of the scheduled week (BO03: week 10 – day 63 since start of first cycle; BO06: week 7 – day 42 since start of first cycle), or *delayed* if performed 7 or more days after scheduled date. P-value of two-sided Mann-Whitney U test for surgery time wrt schedule in BO03 vs BO06: 0.0899; p-value of chi-squared test among surgery time category and trial: 0.3397.

Multiple Overall Toxicity score. Let $\mathcal{T}$ and k denote the set of different toxicities and the time-period index, respectively. Let $tox_{ij,k}$ (with value from 0 to 4) be the most severe CTCAE grade of the j-th toxicity (with $j = 1, ..., |\mathcal{T}|$) measured during period k for the i-th patient. The Multiple Overall Toxicity (MOTox) score for the i-th patient during period k is defined as:

$$MOTox_{i,k} = \frac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} tox_{ij,k} + \max_{j \in \mathcal{T}} (tox_{ij,k}).$$

In particular, for each patient i four different MOTox scores could be computed considering as time-period index the pre-operative and post-operative periods, i.e., $k \in \{pre, post\}$, and two disjoint sets of toxicities related to *rule-specific* and *generic* conditions, i.e, $\mathcal{T}^{(rule)} = \{$*leucopenia*, *thrombocytopenia*, *oral mucositis*, *ototoxicity*, *cardiotoxicity*, *neurotoxicity*$\}$ and $\mathcal{T}^{(gen)} = \{$*nausea*, *infection*$\}$.

Table 7.2. Toxicity coding based on Common Terminology Criteria for Adverse Events (CTCAE) v3.0 by [208] for *rule-specific* (i.e., leucopenia, thrombocytopenia, oral mucositis, ototoxicity, cardiotoxicity and neurotoxicity) and *generic* (i.e., nausea/vomiting and infections) toxicities.

Toxicity	**Grade 0**	**Grade 1**	**Grade 2**	**Grade 3**	**Grade 4**
Rule-specific					
Leucopenia					
White Blood Cells	$\geq 4.0 \times 10^9/L$	$[3.0 - 4.0) \times 10^9/L$	$[2.0 - 3.0) \times 10^9/L$	$[1.0 - 2.0) \times 10^9/L$	$< 1.0 \times 10^9/L$
Thrombocytonpenia					
Neutrophils	$\geq 100 \times 10^9/L$	$[75 - 100) \times 10^9/L$	$[50 - 75) \times 10^9/L$	$[25 - 50) \times 10^9/L$	$< 25 \times 10^9/L$
Oral Mucositis	No change	Soreness or erythema	Ulcers: can eat solid	Ulcers: liquid diet only	Alimentation not possible
Cardiac toxicity	No change	Sinus tachycardia	Unifocal PVC arrhythmia	Multifocal PVC	Ventricular tachycardia
Ototoxicity	No change	Slight hearing loss	Moderate hearing loss	Major hearing loss	Complete hearing loss
Neurological toxicity	None	Paraesthesia	Severe paraesthesia	Intolerable paraesthesia	Paralysis
Generic					
Nausea/Vomiting	None	Nausea	Transient vomiting	Continuative vomiting	Intractable vomiting
Infection	None	Minor infection	Moderate infection	Major infection	Major infection with hypotension

PVC = Premature Ventricular Contraction

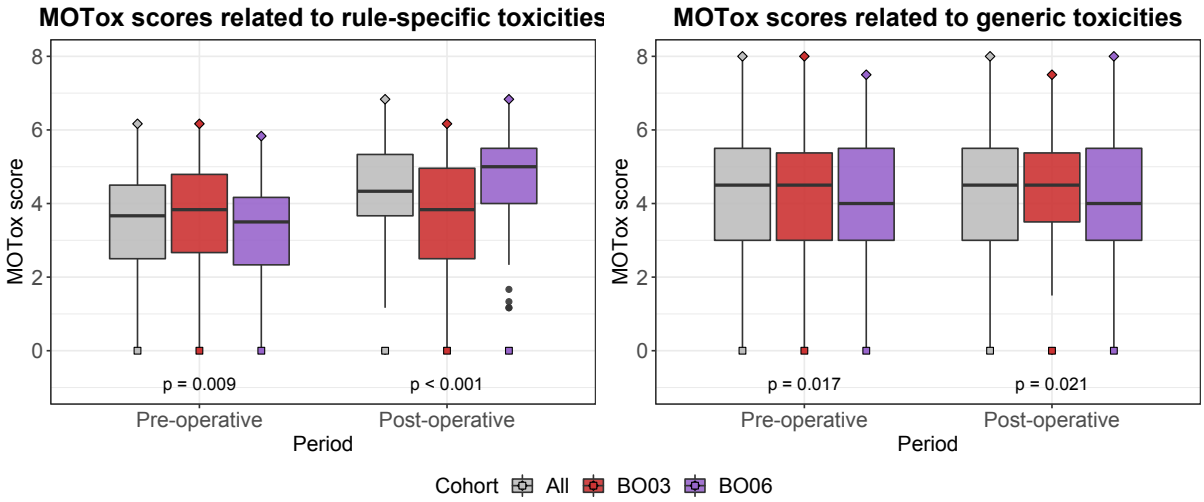

Figure 7.3. Left panel: boxplots of pre-operative and post-operative MOTox scores related to *rule-specific* toxicities, i.e., $MOTox_{i,k}^{(rule)}$ with $k \in \{pre, post\}$ and $\mathcal{T}^{(rule)} = \{leucopenia, thrombocytopenia, oral\ mucositis, ototoxicity, cardiotoxicity, neurotoxicity\}$. Right panel: boxplots of pre-operative and post-operative MOTox scores related to *generic* toxicities, i.e., $MOTox_{i,k}^{(gen)}$ with $k \in \{pre, post\}$ and $\mathcal{T}^{(gen)} = \{nausea, infection\}$.
Boxplots are grouped by cohorts (gray: *All*; red: *BO03*; purple: *BO06*). Squares and diamonds represent minimum and maximum values, respectively. P-values p refer to Mann-Whitney U tests for the distribution of MOTox scores in BO03 vs BO06 cohorts.

Figure 7.3 displays a summary of pre/post-operative MOTox characteristics for both *rule-specific* (left panel) and *generic* (right panel) conditions. Overall (gray boxes), *generic* MOTox scores were high: pre/post-operative median MOTox values were equal to 4.5 meaning that in median patients experienced at least one generic side effect of CTCAE-grade 3, i.e., severe or medically significant. This is not surprising because nausea is the most common chemotherapy-induced adverse event. *Rule-specific* MOTox resulted higher in the post-operative period than in the pre-surgery one. This indicates that toxicity levels have accumulated over time resulting in more severe overall toxic burden in the second phase of treatment.

7.1.3. Chemotherapy exposure characteristics

Data on chemotherapy administration (administered dose of chemotherapy, cycle starting dates, delays) were collected prospectively at each treatment cycle in both trials. After pre-operative treatment cycles, surgery was performed and data about HRe were measured. Chemotherapy exposure can hence be evaluated in terms of both (i) reductions in the actual dose intensity with respect to anticipated/planned one (i.e., by RDI reduction) and (ii) improvement in the appearance of microscopic tissue specimens in a patient after pre-operative treatment (i.e., by HRe).

As mentioned in Section 7.1.1, control arm patients in both BO03 and BO06 underwent the standard EOI treatment structured in 6 cycles of 3-weekly CDDP plus DOX. Reductions of CDDP and/or DOX dosage at each cycle may be assessed considering the *cycle-standardized dose*, defined as follows:

Cycle-standardized dose. The cycle-standardized dose of drug d for patient i at cycle j is

$$\delta_{ij}^{d} = \frac{\text{actual dose of drug } d \text{ assumed at cycle } j \text{ by patient } i \ [mg/m^2]}{\text{anticipated dose of drug } d \ [mg/m^2]} \tag{7.1}$$

where d is the type of drug (CDDP or DOX). As established by trial protocols (see Figure 7.1), anticipated doses of CDDP and DOX are 100 mg/m^2 and 75 mg/m^2, respectively.

Figures 7.4 shows the longitudinal nature of drug-dosage data and how treatment modifications were differently deployed in the two studies. Reductions were usually allocated in the last cycles. This is in line with the common understanding that toxicity levels are more severe towards the end of the treatment and tend to cumulate over time.

To evaluate both dose reductions/discontinuations, time-delays, and their impact in reducing the intensity of the whole therapy, the so-called Received Dose Intensity [86] approach can be adopted. RDI method is able to summarize information on treatment adjustments during the whole therapy, considering both *standardized dose* and *standardized time*.

Standardized dose. The standardized dose for patient i at the end of the treatment is

$$\Delta_i = \frac{1}{2}\left(\Delta_i^{CDDP} + \Delta_i^{DOX}\right) = \frac{1}{12}\left(\sum_{j=1}^{6}\delta_{ij}^{CDDP} + \sum_{j=1}^{6}\delta_{ij}^{DOX}\right). \tag{7.2}$$

$\Delta_i < 1$ indicates dose-reduced therapies, whereas $\Delta_i > 1$ corresponds to dose-augmented therapies.

Standardized time. The standardized time for patient i at the end of the treatment is

$$\Gamma_i = \frac{\text{actual treatment time}}{\text{anticipated treatment time}} \tag{7.3}$$

where

- *actual treatment time* is the difference in days between the starting date of cycle 1 and the 3rd day after the start of cycle 6,
- *anticipated treatment time* is $21 \times 5 + 14 + 3 = 122$ days, i.e., 5 cycles lasting 21 days each, 14 days of surgery and 3 days after the start of cycle 6.

$\Gamma_i > 1$ indicates delayed therapies, whereas $\Gamma_i < 1$ corresponds to accelerated treatments.

Received Dose Intensity. The Received Dose Intensity at the end of the treatment (i.e., *final* RDI) for patient i is defined as the ratio between standardized dose Δ_i and standardized time Γ_i, as follows

$$RDI_i = \frac{\Delta_i}{\Gamma_i}. \tag{7.4}$$

Figure 7.4. Scatter plots of cycle-standardised CDDP-dose (δ_{ij}^{CDDP}) vs. cycle-standardised DOX-dose (δ_{ij}^{DOX}) over cycles $j = 1, ..., 6$ (see Equation (7.1)) conditional on cycle and trial (*BO03*: top panels; *BO06*: bottom panels). Solid lines mark the planned dose, while dotted lines mark reductions of 20%. In order to improve the readability of the plot, very large values are trimmed off to 1.25.

In general, $\Delta_i \leq 1$ and $\Gamma_i \geq 1$ due to dose reductions and delays, respectively, and so $RDI_i \leq 1$.

Instead of considering the whole treatment from cycle 1 to 6, standardized dose, time and RDI could be computed for *pre-operative* and *post-operative* periods separately. For each patient, the pre-operative period is made up of cycles performed before surgery, while the post-operative period of cycles performed after surgery. Appendix D.1 reports how definitions in Equations (7.2), (7.3) and (7.4) can be adapted to consider *pre-operative* and *post-operative* periods separately, i.e., $\Delta_{i,k}$, $\Gamma_{i,k}$ and $RDI_{i,k}$ with $k \in \{pre, post\}$. Note that $RDI_i \neq RDI_{i,pre} + RDI_{i,post}$.

As mentioned in Section 7.1.1, the level of tumour necrosis for each patient was assessed after surgical resection (planned at the end of cycle 3/2 in BO03/BO06 – see Figure 7.1) and used to define HRe, as follows:

Histological Response. Histological Response (HRe) to pre-operative chemotherapy is defined as *poor* if tumour necrosis is less than 90% (i.e., $\geq 10\%$ of viable tumour) or *good* if tumour necrosis is greater than or equal to 90% (i.e., $< 10\%$ of viable tumour).

Figure 7.5 reports a summary of treatment exposure characteristics for the whole cohort and conditional on trials. The percentages of patients with a *good* HRe after surgical resection were 34.1% (94 patients) in the whole cohort, 32.5% in BO03 and 35.2% in BO06. Overall, median value of final RDI was 0.759 (IQR=[0.649; 0.857]), with minimum and maximum values of 0.376 and 1.121, respectively. Median percentages of pre-operative and post-operative RDI were 0.810 (IQR=[0.727; 0.901]) and 0.723 (IQR=[0.584;0.870]), respectively, confirming that reductions and delays are usually allocated in the post-operative cycles.

Figure 7.6 shows a scatter plot of RDI at the end of treatment (final RDI_i) against the final standardized dose of CDDP+DOX (Δ_i) conditional on trial (left panel: *BO03*; right panel: *BO06*) and HRe (blue: *poor*; green: *good*). The solid horizontal lines in pink vertically divide patients with *normal* RDI levels ($RDI_i \geq 0.85$) from *low* reduction ($0.70 \leq RDI_i < 0.85$) and *high* reduction ($RDI_i < 0.70$) patients. The solid diagonal line in black satisfies equation $RDI_i = \Delta_i$, dividing the group of patients with standardized time $\Gamma_i > 1$ (delayed therapy, below the line) from the group, almost void, of patients with $\Gamma_i < 1$ (anticipated therapy, above the line). The dotted diagonal line in black satisfies equation $RDI_i = \Delta_i/1.2$, dividing the group of patients with therapy delayed by more than 20% of anticipated time (below the dotted line) from the group of patients with therapy delayed by less than 20% of anticipated time (between solid and dotted black lines). This figure shows the lack of a clear association between HRe and RDI. Analogous figures for pre/post-operative RDI against their relative standardized doses can be found in Appendix D.1 . Both Figures 7.5 and 7.6 clearly display the difference of treatment delivery in BO03 and BO06 trials.

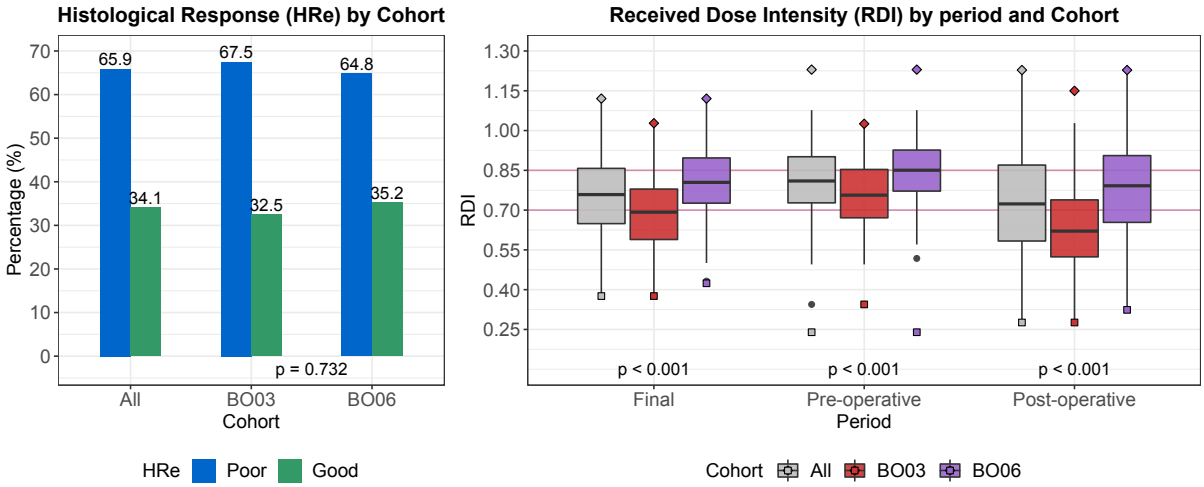

Figure 7.5. Patients treatment exposure characteristics. Left panel: barplots of Histological Response (HRe) by cohort (All, BO03, BO06) coloured according to HRe level (blue: *poor*; green: *good*). P-value p refers to the chi-squared test for the association between HRe and BO03/BO06 trial. Right panel: boxplots of final, pre-operative and post-operative Received Dose Intensity (i.e., RDI_i, $RDI_{i,pre}$, $RDI_{i,post}$) grouped by cohort (gray: *All*; red: *BO03*; purple: *BO06*). Squares and diamonds represent minimum and maximum values, respectively. P-values p refer to Mann-Whitney U tests for the distribution of RDI values in BO03 vs BO06 cohorts.

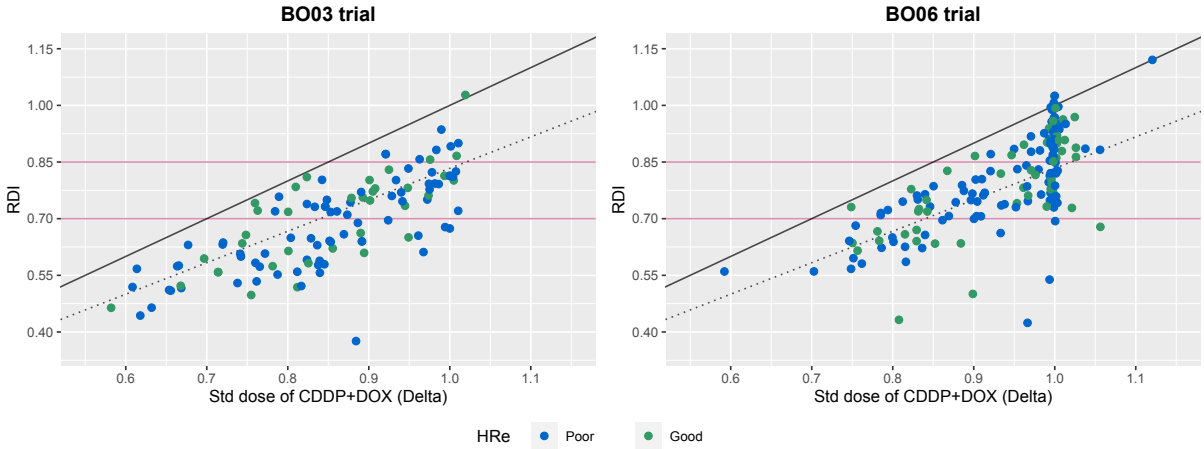

Figure 7.6. Scatter plots of RDI at the end of treatment (i.e., RDI_i in Equation (7.4)) against the final standardized dose of CDDP+DOX (Δ_i) conditional on trial (*BO03*: left panel; *BO06*: right panel) and HRe (blue points: *poor*; green points: *good*).

7.2. Causal inference structure and methods

Since negative feedback between therapy administration and toxicities acts as a (generally time-dependent) confounder for the effect of chemotherapy exposure on outcome, the idea of this study is to create a pseudo-population in which medical history no longer predicts exposure through IPTW. In that framework, Cox MSMs can be used to estimate the joint causal effect of HRe and dose intensity on Event-Free-Survival (EFS). In order to create such a pseudo-population, outcome, exposure, confounders and their mutual relationships have to be defined. EFS outcome is defined in Section 7.2.1. Causal inference assumptions for MSMs are introduced in Section 7.2.2. A suitable characterisation of the causal structure of the chemotherapy data is given in Section 7.2.3. Two alternative

definitions of joint-exposure with their relative models are finally introduced in Sections 7.2.4 and 7.2.5.

7.2.1. Event-Free Survival Outcome

The endpoint of this study is Event-Free Survival (EFS), defined as time from the end of therapy until the first event (local recurrence, evidence of new or progressive metastatic disease, second malignancy, death, or a combination of those events) or censoring at last contact. In particular:

EFS outcome. The time-to-event outcome for patient $i \in \{1, ..., N\}$ is denoted as (T_i, D_i), where $T_i = \min(T_i^*, C_i)$ is the observed EFS time, T_i^* is the true event time, C_i is the censoring time (i.e., the time from the end of the therapy until the last visit) and $D_i = I(T_i^* \leq C_i)$ is the event indicator, with $I(\cdot)$ being the indicator function that takes the value 1 when $T_i^* \leq C_i$, and 0 otherwise.

7.2.2. Causal inference assumptions for marginal structural models

Marginal structural Cox models allow the estimation of the causal associations between treatment exposure $\boldsymbol{A}$ and time-to-event response T in the presence of time-dependent covariates $\boldsymbol{L}$ that may be simultaneously confounders and intermediate variables [78, 79, 100]. Cox MSMs target *counterfactual* (or *potential*) time-to-event variables $T^{\boldsymbol{a}}$, i.e., the time at which an event would be observed had the subject, possibly contrary to fact, been administered a treatment exposure $\boldsymbol{A} = \boldsymbol{a}$. There exist four main assumptions for causal inference with (Cox) MSMs through IPTW [41, 77].

1. ***Exchangeability*** or ***No unmeasured confounding***

   Exchangeability (or conditional exchangeability) implies the well-known assumption of *no unmeasured confounding* [41]. It states that exposure allocation is independent of the potential outcomes conditional on pre-treatment covariates (i.e., $T^{\boldsymbol{a}} \perp\!\!\!\perp \boldsymbol{A}|\boldsymbol{L}$) or, in a longitudinal setting, that treatment is sequentially randomized given the past [41]. This assumption is often referred as "ignorable treatment assignment" or "sequential randomization" in statistics, "selection on observables" in the social sciences or "no omitted variable bias" in econometrics [77].

   The main limitation is that, in absence of randomization such as in observational studies, exchangeability is not be testable so there is no guarantee that it holds. Experts knowledge is then necessary for the identification of enough joint predictors of exposure and outcome such that, within the levels of these predictors, associations between exposure and outcome that are due to their common causes will disappear [41].

2. ***Consistency***

   Consistency means that the outcome observed for each individual is precisely the counterfactual outcome under their observed treatment history, that is $T^{\boldsymbol{a}} = T$ for every individual with $\boldsymbol{A} = \boldsymbol{a}$. This assumption would be violated in the presence of misclassification bias [217] and has two requirements [77]:

   i. since one must be able to explain how a certain level of exposure could hypothetically be assigned to a person exposed to a different level, the exposure must be *defined unambiguously* so that the counterfactual outcomes are well-defined;

   ii. there is a need to *link* the counterfactuals with observed data and thus to reasonably assume that the equality is valid for at least some individuals.

   Although consistency can not be empirically verified, it is assumed plausible in observational studies of medical treatments, because one can imagine how to hypothetically manipulate an individual's treatment status [40].

3. ***Positivity***

   Positivity states that there is a non-zero (i.e., positive) probability of receiving every level of exposure for every combination of values of exposure and covariate histories that occur among individuals in the population [41]. If this assumption is violated, then the weights in IPTW are undefined leading to biased estimates of the causal effect.

   If someone cannot be exposed to one or more levels of the confounders (e.g., it cannot be treated in the presence of recommendations from guidelines or established contraindications), then positivity is violated due to a *structural* zero probability of receiving the exposure. A solution is to restrict the inference to the subset with a positive probability of exposure, whenever possible [40]. Even in the absence of structural zeros, *random* zeros may occur by chance due to small sample sizes or highly stratified data by numerous confounders. The inclusion of weak or highly-stratified confounders can provide a better confounding adjustment but may cause severe non-positivity, increasing the bias and variance of the estimated effect. An indication of non-positivity may be the presence of estimated weights with the mean far from one or very extreme values [40].

4. ***No misspecification of both weighting and outcome models***

   The final assumption of MSMs is that both the weighting model for IPTW and the structural outcome model, which links the outcome to the exposure history, must be correctly specified. This assumption has similar roots in essentially all statistical models [217], as model misspecification leads to instability in the Cox MSM estimator [100, 101].

   Since the presence of estimated stabilized weights with the mean far from one or very extreme values are indicative of non-positivity or misspecification of the weight

model [40], correctness of the weighting model specifications can be checked by exploring the distribution of weights [41]. In addition, quantitative (e.g., weighted standardized difference to compare means or prevalences) and qualitative graphical methods can be used to assess whether measured covariates are balanced between treatment groups in the weighted sample [19].

If these assumptions hold, causal inference is possible from MSMs through IPTW. In particular, IPTW creates a pseudo-population by weighting each patient with the inverse probability of observing a certain treatment allocation given the past treatment and confounders history. In the context of chemotherapy treatment, a pseudo-population created in this way has the following two properties:

i. the past history of pseudo-patients no longer predicts exposure to chemotherapy in the next cycle;

ii. the association between exposure and outcome is the same in both the original and the pseudo-population, so that causal effect of treatment modifications can be just obtained by a crude analysis on the pseudo-population.

In the following sections, joint-exposure, confounders and Cox MSMs are introduced through a thoughtful process designed to make the four assumptions acceptable. Section 7.2.3 describes a suitable characterisation of the causal structure of the chemotherapy data through the introduction of appropriate Direct Acyclic Graphs (DAGs) that identify all possible (time-dependent) confounders and their relationships with exposure and outcome. In fact, once defined the appropriate DAGs according to clinical and statistical knowledge, it can be reasonably assumed that *exchangeability* is approximately true within confounding strata. Sections 7.2.4 and 7.2.5 introduce two alternative unambiguous definitions of exposure which meet *consistency* according to experts, along with their corresponding counterfactual EFS outcomes and relative proposed Cox MSMs to estimate the association between them. *Positivity* and *no misspecification* will be finally checked for data application results in Section 7.3.

7.2.3. Causal structure of chemotherapy data

Relationships between random variables (i.e., exposure, confounders and outcome) is usually represented using DAGs in causal inference [67, 77]. Both clinical/oncological expertise in osteosarcoma treatment and statistical competence in variables definitions and mathematical modelling are required to construct an appropriate DAG for the problem under analysis, where the main interest is to estimate the joint causal effect of HRe and dose intensity reduction on EFS.

In both trials, HRe level was measured after surgery and can be considered as a consequences of patient's pre-operative characteristics. Only the most severe CTCAE grades were recorded in BO03, while data from BO06 are fully longitudinal in both exposure and side-effects (see Section 7.1.2). This fact posed a modelling issue, because the therapy

adjustment cannot be modelled cycle-by-cycle. Two alternative options are then plausible for dose intensity:

1. *time-fixed final RDI*: the final value of RDI (i.e., the value at the end of treatment) can be seen as the result of the most severe toxicities experienced by the patient throughout the therapy;

2. *time-dependent pre/post-operative RDI*: therapy adjustment can be modelled by pre- and post-operative periods, considering the values of pre/post-operative RDI as results of the most severe overall toxicities experienced by the patient during pre/post-operative cycles.

The first option leads to a *time-fixed joint-exposure* of HRe and final RDI, whereas the second one to a *time-varying joint-exposure* given by HRe and time-varying pre/post-operative RDI.

Confounders were identified according to protocol guidelines and oncological experts knowledge. Conditioning chemotherapy administration over treatment as mentioned in Section 7.1.2, both *rule-specific* and *generic* multiple overall toxicities represent time-dependent confounders. Influencing the drug metabolism, and so being risk factors for increased toxicity, age and gender are baseline confounders because they were also clinically considered independent predictors of mortality. Although the trial does not represent a proper risk factor for failures (p-value of log-rank test for Kaplan-Meier estimators stratified by trial is about 1), it can be considered as a baseline confounder, being both an independent predictor for HRe (through number of preoperative cycles [119] and therapy starting days) and for dose intensity (see Figures 7.4, 7.5, 7.6), and influencing EFS through the way CTCAE grades were assessed and therapy modifications allocated (see Section 7.1.2). Furthermore, since there is usually a tendency not to delay surgery in the case of disease progression, the surgery timing may influence HRe.

According to the literature on MSMs, where the roman capital letter L is used to indicate a confounder, the following variables denote the characteristics of the i-th patient that influence both exposure and outcome.

Time-fixed confounders for the i-th patient are represented by vectors of baseline and surgery characteristics, i.e., $\boldsymbol{L}_i^{base}$ and L_i^{surg} with elements:

- $L_i^{base,1}$: trial number (BO03; BO06);
- $L_i^{base,2}$: gender (*female*; *male*);
- $L_i^{base,3}$: age group defined according to Collins *et al.* (2013) [43] (*child*: 0–12/0-11 years for males/females; *adolescent*: 13–17/12–16 years for males/females; *adult*: 18/17 or older for males/females);
- L_i^{surg}: surgery time category with respect to schedule (0: *delayed*; 1: *on time* – see Table 7.1).

Time-varying confounders for the i-th patient are represented by the vectors of Multiple Overall Toxicity burden during pre/post-operative periods $k \in \{pre, post\}$, i.e., $\boldsymbol{L}_{i,pre}^{tox}$ and $\boldsymbol{L}_{i,post}^{tox}$ with elements

- $L_{i,k}^{tox,1} = MOTox_{i,k}^{(rule)}$: MOTox score related to period k based on *rule-specific* conditions $\mathcal{T}^{(rule)}$ (see Section 7.1.2)
- $L_{i,k}^{tox,2} = MOTox_{i,k}^{(gen)}$: MOTox score related to period k based on *generic* conditions $\mathcal{T}^{(gen)}$ (see Section 7.1.2).

The choice of MOTox scores instead of individual CTCAE grades for the various toxicities is motivated both by the *positivity/confounders trade-off* and by the clinical protocols. By considering the individual grades for each toxicity, the number of possible confounders combinations would be too high leading to non-positivity. This choice also meets the clinical rationale, in the case of multiple toxicities, of adapting treatment according to the overall toxic burden of the patient (see Section 7.1.2).

According to experts knowledge, these characteristics have been believed to form a set of variables that satisfies the hypothesis of *no unmeasured confounding*. In particular, baseline and pre-operative MOTox confounders affect both HRe and RDI. As the delay in the surgery time already included in the calculation of the RDI (it concurs to standardized time), surgery confounder only affects HRe (p-value of chi-squared test for association is 0.023). Being HRe the response to pre-operative treatment, post-operative MOTox confounders only influence RDI.

Figure 7.7 shows two alternative DAGs resulting from the causal structure described above. DAG-1 (top panel) is characterized by EFS outcome, aforementioned confounders, and the time-fixed joint-exposure given by both HRe and final RDI. DAG-2 (bottom panel) identifies a relationship among EFS outcome, confounders and a time-varying joint-exposure given by HRe and pre/post-operative RDIs. Both DAGs rely upon the hypothesis that HRe and RDI(s) are conditionally independent on the patient's toxicity-history. In other words, given two patients with the same toxicity history but different values of HRe, the probability of observing a reduction in RDI, say of 15%, is the same in the two patients regardless of one being *poor* responder and the other *good* responder. This assumption can be defended on the following two facts:

i. HRe is typically not known until several weeks since chemotherapy is resumed after surgery, i.e., HRe could influence the decision to reduce therapy intensity only in the very last cycles;

ii. in a randomized trial clinicians can be expected to be rather committed to following the trial protocol.

Moreover, both modelling choices do not allow for a fine continuous analysis of RDI, as this would not guarantee the assumptions of *consistency* and *positivity*. Therefore, an unambiguous well-defined categorization according to a clinical rational of RDI exposure variables must be introduced.

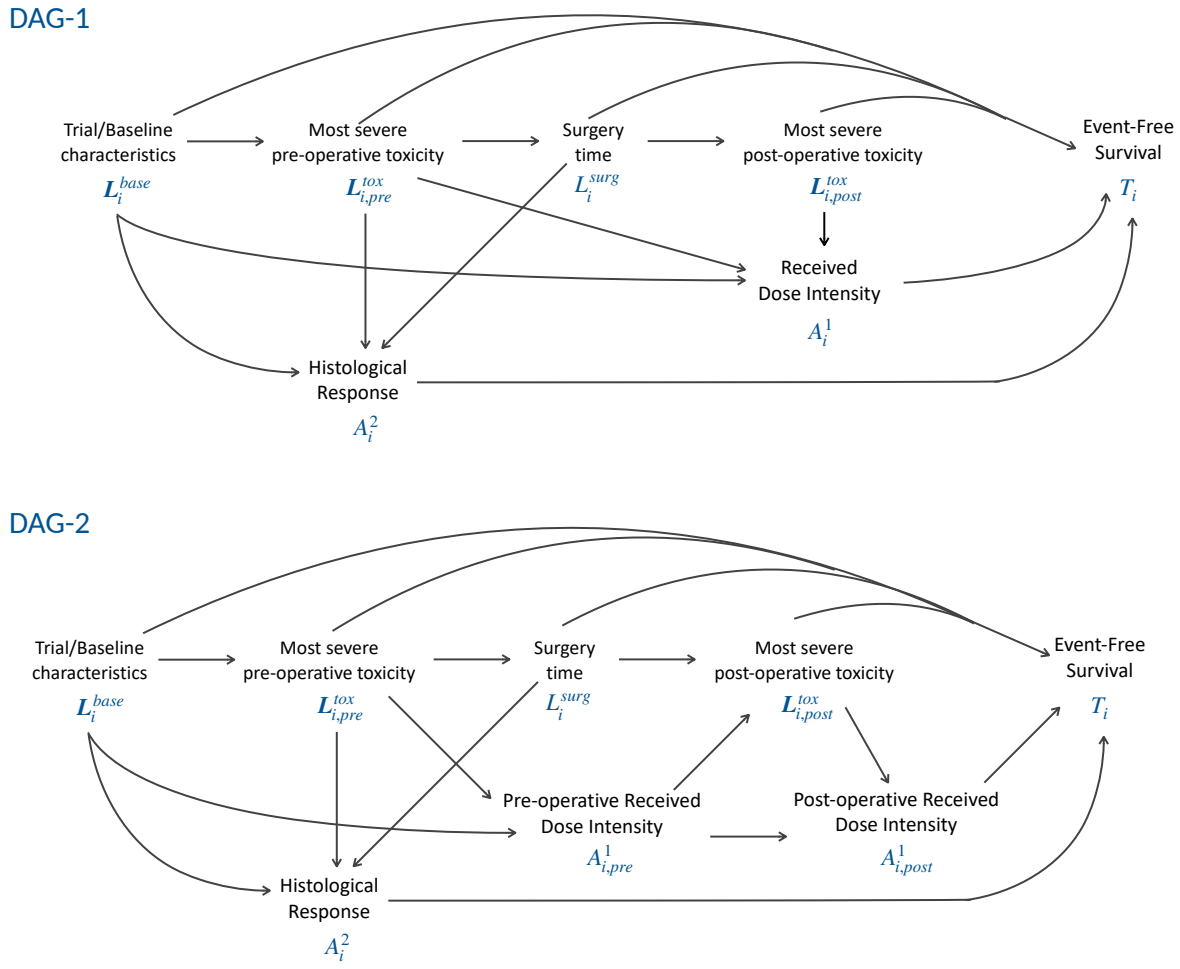

Figure 7.7. Directed Acyclic Graphs (DAGs) used to represent the causal relationships between event free survival outcome T_i, joint-exposure $\boldsymbol{A}_i$, time-fixed confounders $\boldsymbol{L}_i$ (baseline and surgery) and time-varying confounders $\boldsymbol{L}_{i,k}$ (pre/post-operative multiple overall toxicities). Top panel (DAG-1): joint exposure is characterized by HRe and time-fixed final RDI. Bottom panel (DAG-2): joint-exposure is characterized by HRe and time-varying pre/post-operative RDI.

7.2.4. Joint-exposure and marginal structural Cox model for DAG-1

DAG-1 (top panel in Figure 7.7) is characterized by the EFS outcome T_i, the time-fixed and time-varying confounders $\left(\boldsymbol{L}_i^{base}, L_i^{surg}\boldsymbol{L}_{i,pre}^{tox}, \boldsymbol{L}_{i,post}^{tox}\right)$, and a joint-exposure $\boldsymbol{A}_i$ given by HRe and final RDI, both time-fixed. According to expert knowledge, a *normal* RDI level (i.e., $RDI_i \geq 0.85$) can be analysed in contrast to *low-reduction* (from 15% to 30%) and *high-reduction* (more than 30%) categories. Joint-exposure $\boldsymbol{A}_i$ for DAG-1 can hence be defined as follows.

Joint-exposure. The *time-fixed joint-exposure* administered for subject i is denoted by the vector

$$\boldsymbol{A}_i = \left(A_i^1, A_i^2\right) \tag{7.5}$$

where

- A_i^1 is the three-level exposure related to final RDI

$$A_i^1 = \begin{cases} 0 & \text{if } RDI_i \geq 0.85 \\ 1 & \text{if } 0.70 \leq RDI_i < 0.85 \\ 2 & \text{if } RDI_i < 0.70 \end{cases}$$

- A_i^2 is the binary exposure related to HRe

$$A_i^2 = \begin{cases} 0 & \text{if tumour necrosis}_i < 90\% \\ 1 & \text{if tumour necrosis}_i \geq 90\% \end{cases}$$

  that is, $A_i^2 = 1$ is equivalent to a *"good"* HRe, while $A_i^2 = 0$ denotes a *"poor"* HRe.

Once joint-exposure is defined unambiguously, the counterfactual EFS outcome, i.e., the outcome that would be observed had the subject followed, possibly contrary-to-fact, a given treatment, is also well-defined:

Counterfactual outcome. Let $T_i^{\boldsymbol{a}} = T_i^{(a_1,a_2)}$ denote the counterfactual EFS time that would be observed in a subject i with joint-exposure treatment

$$A_i^1 = a_1, \quad a_1 \in \{0,1,2\}, \qquad \text{and} \qquad A_i^2 = a_2, \quad a_2 \in \{0,1\}.$$

In particular, there are exactly six joint-exposure (a_1, a_2) that can be realised according definition in Equation (7.5):

- (0,0): *poor* responder without significant reduction (i.e., *normal* RDI level);
- (1,0): *poor* responder with final *low-reduction* of 15-30%;
- (2,0): *poor* responder with final *high-reduction* of more than 30%;
- (0,1): *good* responder without significant reduction;
- (1,1): *good* responder with final *low-reduction* of 15-30%;
- (2,1): *good* responder with final *high-reduction* of 30%.

Within a counterfactual framework, i.e., in the pseudo-population, Cox MSMs enable the conceptual comparison of the hazard functions for different treatment level $\boldsymbol{a} = (a_1, a_2)$. No baseline/experimental covariates are included in the model because there is no clinical interest in assessing the causal effect of changes in chemotherapy exposure within specific population strata. The main interest consists in proposing a Cox MSM that represents the causal RDI analogue of the Intention-To-Treat (ITT) Cox model presented by Lewis *et al.* (2007) [119], which included HRe, intended treatment, and their interaction. A Cox MSM with interactions between a_1 and a_2, where the treatment binary variable is replaced by the actual final RDI level, is hence proposed as follows:

Cox MSM 1. The Marginal Structural Cox Model for EFS time under treatment level $\boldsymbol{a} = (a_1, a_2)$ is

$$h_{T_i^{\boldsymbol{a}}}(t) = h_0(t) \exp\left\{\beta_1 \mathbb{1}_{(a_1=1)} + \beta_2 \mathbb{1}_{(a_1=2)} + \beta_3 a_2 + \beta_4 \mathbb{1}_{(a_1=1)} a_2 + \beta_5 \mathbb{1}_{(a_1=2)} a_2\right\} \quad (7.6)$$

To estimate the causal parameters $\boldsymbol{\beta}$ of the Cox MSM in Equation (7.6), a weighted Cox model [30, 126] can be fitted to the pseudo-population obtained through IPTW, as follows

$$h_{T_i}^{SW_i}(t|\boldsymbol{A}_i) = h_0(t) \exp\left\{\theta_1 \mathbb{1}_{(A_i^1=1)} + \theta_2 \mathbb{1}_{(A_i^1=2)} + \theta_3 A_i^2 + \theta_4 \mathbb{1}_{(A_i^1=1)} A_i^2 + \theta_5 \mathbb{1}_{(A_i^1=2)} A_i^2\right\} \quad (7.7)$$

with subject-specific stabilized weights

$$SW_i = SW_i^{A^1} \cdot SW_i^{A^2} \quad (7.8)$$

where

$$SW_i^{A^1} = \frac{P\left(A_i^1\right)}{P\left(A_i^1 \middle| \boldsymbol{L}_i^1\right)} = \frac{P\left(A_i^1\right)}{P\left(A_i^1 \middle| \boldsymbol{L}_i^{base}, \boldsymbol{L}_{i,pre}^{tox}, \boldsymbol{L}_{i,post}^{tox}\right)};$$
$$SW_i^{A^2} = \frac{P\left(A_i^2\right)}{P\left(A_i^2 \middle| \boldsymbol{L}_i^2\right)} = \frac{P\left(A_i^2\right)}{P\left(A_i^2 \middle| \boldsymbol{L}_i^{base}, \boldsymbol{L}_{i,pre}^{tox}, L_i^{surg}\right)}.$$

In both $SW_i^{A^1}$ and $SW_i^{A^2}$ cases, numerators are the probability that a subject i received observed exposures A_i^1 and A_i^2 respectively, whereas denominators are the probability that the subject received observed exposures given relative time-fixed and time-dependent confounders. Regression models have to be chosen appropriately, according to the type of of exposure. In particular, multinomial logistic regression models are used for both numerator and denominator of $SW_i^{A^1}$, whereas binary logistic regression models are adopted for $SW_i^{A^2}$.

Under causal inference assumptions (see Section 7.2.2), association is causation in the pseudo-population and the estimates of the associational parameters $\boldsymbol{\theta}$ are consistent for the causal parameters $\boldsymbol{\beta}$. In applying this methodology to the chemotherapy data, different model specifications in terms of confounding covariate features must be compared to satisfy the final assumptions of *positivity* and *no misspecification of the weight-generating models* and guarantee an unbiased estimation of the results.

7.2.5. Joint-exposure and marginal structural Cox model for DAG-2

DAG-2 (bottom panel in Figure 7.7) is characterized by the EFS outcome T_i, the time-fixed and time-varying confounders $\left(\boldsymbol{L}_i^{base}, L_i^{surg} \boldsymbol{L}_{i,pre}^{tox}, \boldsymbol{L}_{i,post}^{tox}\right)$, and a joint-exposure $\bar{\boldsymbol{A}}_i$ given by HRe and time-varying pre/post-operative RDI. As in the previous section, a *normal* RDI level can be analysed in contrast to *low* and *high* reductions. Time-varying joint-exposure $\bar{\boldsymbol{A}}_i$ for DAG-2 is hence defined as follows.

Joint-exposure. The *time-varying joint-exposure* administered for subject i is denoted by the vector

$$\bar{\boldsymbol{A}}_i = \left(\bar{\boldsymbol{A}}_i^1, A_i^2\right) = \left(\left(A_{i,pre}^1, A_{i,post}^1\right), A_i^2\right) \tag{7.9}$$

where

- $\bar{\boldsymbol{A}}_i^1$ is the time-varying three-level exposure vector related to pre/post-operative RDI with elements

$$A_{i,k}^1 = \begin{cases} 0 & \text{if } RDI_{i,k} \geq 0.85 \\ 1 & \text{if } 0.70 \leq RDI_{i,k} < 0.85 \\ 2 & \text{if } RDI_{i,k} < 0.70 \end{cases}$$

  where $k \in \{pre, post\}$ indicating the *pre-operative* and *post-operative* periods, respectively;

- A_i^2 is the binary exposure related to HRe

$$A_i^2 = \begin{cases} 0 & \text{if tumour necrosis}_i < 90\% \\ 1 & \text{if tumour necrosis}_i \geq 90\% \end{cases}$$

  that is, $A_i^2 = 1$ is equivalent to a *"good"* HRe, while $A_i^2 = 0$ denotes a *"poor"* HRe.

Once joint-exposure is defined unambiguously, the counterfactual EFS outcome, i.e., the outcome that would be observed had the subject followed – possibly contrary-to-fact – a given treatment, is also well-defined:

Counterfactual outcome. Let $T_i^{\bar{\boldsymbol{a}}} = T_i^{((a_{11},a_{12}),a_2)}$ denote the counterfactual EFS time that would be observed in a subject i with time-varying joint-exposure

$$A_{i,pre}^1 = a_{11}, \quad A_{i,post}^1 = a_{12}, \quad a_{11}, a_{12} \in \{0,1,2\}, \qquad \text{and} \qquad A_i^2 = a_2, \quad a_2 \in \{0,1\}.$$

In particular, there are exactly 18 time-varying joint-exposure combinations $\bar{\boldsymbol{a}} = (\bar{\boldsymbol{a}}_1, a_2) = ((a_{11}, a_{12}), a_2)$ that can be realised according definition in Equation (7.9). To avoid too many combinations, we specify a model that combines information from many strategies to help estimate the causal effects. For example, we can hypothesize a cumulative treatment effects under sub-strategy $\bar{\boldsymbol{a}}_1$, named *cumulative RDI-exposure*

$$\text{cum}\left(\bar{\boldsymbol{a}}_1\right) = \sum_{k=1}^{2} a_{1k}$$

which could takes value

- 0: if no reduction, neither pre nor post surgery;
- 1: if only one *low* reduction pre or post surgery;
- 2: if *low* reductions both pre and post surgery or *high* reduction pre or post surgery;

- 3: if both pre-operative (or post-operative) *low* reduction and post-operative (or pre-operative) *high* reduction;
- 4: if *high* reductions both pre and post surgery.

Therefore $\text{cum}\left(\bar{\boldsymbol{a}}_1\right)$ represents the number of reductions by a value of 15-30%, where a single *high* reduction of at least 30% can be seen as twice a *low* reduction of 15-30%. In the following, the *time-varying joint-exposure* levels and values for patient i (with 18 different possible combinations) are indicated by $\bar{\boldsymbol{a}} = \left(\bar{\boldsymbol{a}}_1, a_2\right)$ and $\bar{\boldsymbol{A}}_i = \left(\bar{\boldsymbol{A}}_i^1, A_i^2\right)$ respectively, whereas the *cumulative joint-exposure* levels and values for patient i (with 10 different possible combinations) are indicated by $\tilde{\boldsymbol{a}} = \left(\text{cum}\left(\bar{\boldsymbol{a}}_1\right), a_2\right)$ and $\tilde{\boldsymbol{A}}_i = \left(\text{cum}\left(\bar{\boldsymbol{A}}_i^1\right), A_i^2\right)$, respectively.

Within a counterfactual framework, Cox MSMs enable the conceptual comparison of the hazard functions for different treatment exposure $\bar{\boldsymbol{a}} = \left(\bar{\boldsymbol{a}}_1, a_2\right)$. As in Section 7.2.4, no baseline/trial covariates are included in the proposed structural model. Since the interests is in analysing the causal RDI analogue of the ITT Cox model presented by Lewis *et al.* (2007) [119] according to pre/post-operative RDI definitions, a Cox MSM with interactions between $\text{cum}\left(\bar{\boldsymbol{a}}_1\right)$ and a_2 is hence proposed.

Cox MSM 2. The Marginal Structural Cox Model for EFS time under cumulative treatment level $\tilde{\boldsymbol{a}} = \left(\text{cum}\left(\bar{\boldsymbol{a}}_1\right), a_2\right)$ is

$$h_{T_i^{\tilde{\boldsymbol{a}}}}(t) = h_0(t) \exp\left\{\beta_1 \text{cum}\left(\bar{\boldsymbol{a}}_1\right) + \beta_2 a_2 + \beta_3 \text{cum}\left(\bar{\boldsymbol{a}}_1\right) a_2\right\} \tag{7.10}$$

To estimate the causal parameters $\boldsymbol{\beta}$ of the Cox MSM in Equation (7.10), a weighted Cox model [30, 126] can be fitted to the pseudo-population obtained through IPTW, as follows

$$h_{T_i}^{SW_i}\left(t|\bar{\boldsymbol{A}}_i\right) = h_0(t) \exp\left\{\theta_1 \text{cum}\left(\bar{\boldsymbol{A}}_i^1\right) + \theta_2 A_i^2 + \theta_3 \text{cum}\left(\bar{\boldsymbol{A}}_i^1\right) A_i^2\right\} \tag{7.11}$$

where $\text{cum}\left(\bar{\boldsymbol{A}}_i^1\right)$ is the cumulative RDI-exposure vector

$$\text{cum}\left(\bar{\boldsymbol{A}}_i^1\right) = \sum_{k \in \{pre, post\}} A_{i,k}^1$$

and SW_i are the subject-specific stabilized weights given by

$$SW_i = SW_i^{\bar{\boldsymbol{A}}^1} \cdot SW_i^{A^2} \tag{7.12}$$

with

$$SW_i^{\bar{\boldsymbol{A}}^1} = SW_i^{A_{pre}^1} \cdot SW_i^{A_{post}^1} = \frac{P\left(A_{i,pre}^1\right)}{P\left(A_{i,pre}^1 \middle| \boldsymbol{L}_i^{base}, \boldsymbol{L}_{i,pre}^{tox}\right)} \cdot \frac{P\left(A_{i,post}^1 \middle| A_{i,pre}^1\right)}{P\left(A_{i,post}^1 \middle| A_{i,pre}^1, \boldsymbol{L}_i^{base}, \boldsymbol{L}_{i,pre}^{tox}, \boldsymbol{L}_{i,post}^{tox}\right)};$$

$$SW_i^{A^2} = \frac{P\left(A_i^2\right)}{P\left(A_i^2 \middle| \boldsymbol{L}_i^{base}, \boldsymbol{L}_{i,pre}^{tox}, L_i^{surg}\right)}.$$

As in the previous section, multinomial logistic regression models can be used for both numerators and denominators of $SW_i^{\bar{\boldsymbol{A}}^1}$, whereas binary logistic regression models can be adopted for $SW_i^{A^2}$.

Under causal inference assumptions, association is causation in the pseudo-population and the estimates of the associational parameters $\boldsymbol{\theta}$ are consistent for the causal parameters $\boldsymbol{\beta}$. In applying this methodology to the chemotherapy data, different model specifications must be compared to satisfy *positivity* and *no misspecification of the weight-generating models* and guarantee an unbiased estimation of the results.

7.3. Results

IPTW-based causal methodologies introduced in Section 7.2 are now applied to BO03-BO06 chemotherapy data presented in Section 7.1. In Section 7.3.1 joint-exposures for DAG-1 and DAG-2 are explored in terms of percentages of patients in each exposure-level and association with EFS in the original population. Different IPTW model specifications to determine the subject-specific standardized weights for the pseudo-population are presented in Section 7.3.2. Results of causal Cox MSMs fitted on the pseudo-population are presented in Section 7.3.3, along with their relative unweighted Cox results to show the *toxicity-treatment-adjustment* bias present in the original data. Statistical analyses were performed in the R-software environment [161], in particular using ipw [210] and survival [201] packages. R code for the current study is provided here: `https://github.com/mspreafico/BO0x-CoxMSM`.

7.3.1. Joint-exposure descriptive and association with EFS

Once computed the *time-fixed* and *time-varying* joint-exposures for each subject, the percentage of patients in each level and the naive association with survival were observed. Overall, median EFS time computed using the reverse Kaplan-Meier method by Schemper and Smith (1996) [182] was 89.59 months (IQR = [50.33; 146.30]) and 155 patients (55.1%) experienced an event after the end of the therapy. Figure 7.8 shows both *time-fixed* $\boldsymbol{A}_i$ (top panels) and *cumulative* $\tilde{\boldsymbol{A}}_i$ (bottom panels) joint-exposure characteristics. In both cases, left panels report percentage of patients according to the various joint-exposure levels and right panels display Kaplan-Meier estimators for EFS curves stratified by joint-exposure levels with time expressed in months since end of therapy. As expected, *Good* Responders (GRs) (green curves) presented a better survival with respect to *Poor* ones (PRs – blue curves). In particular, in GRs an increased final/cumulative RDI level seemed associated with better survival, whereas a reversed association was observed in the group of PRs. However, in both cases the curve of GRs with the highest reduction overlapped PRs curves, suggesting the possibility of a non-negligible interaction between the joint-exposure components and validating the Cox MSMs proposed in Equations (7.6) and (7.10).

To further investigate these findings and analyse the causal effect of time-fixed/time-varying joint-exposure on EFS through Cox MSMs, subject-specific standardized weights must be computed from correctly specified IPTW models which take into account all the confounding factors identified in Section 7.2.3.

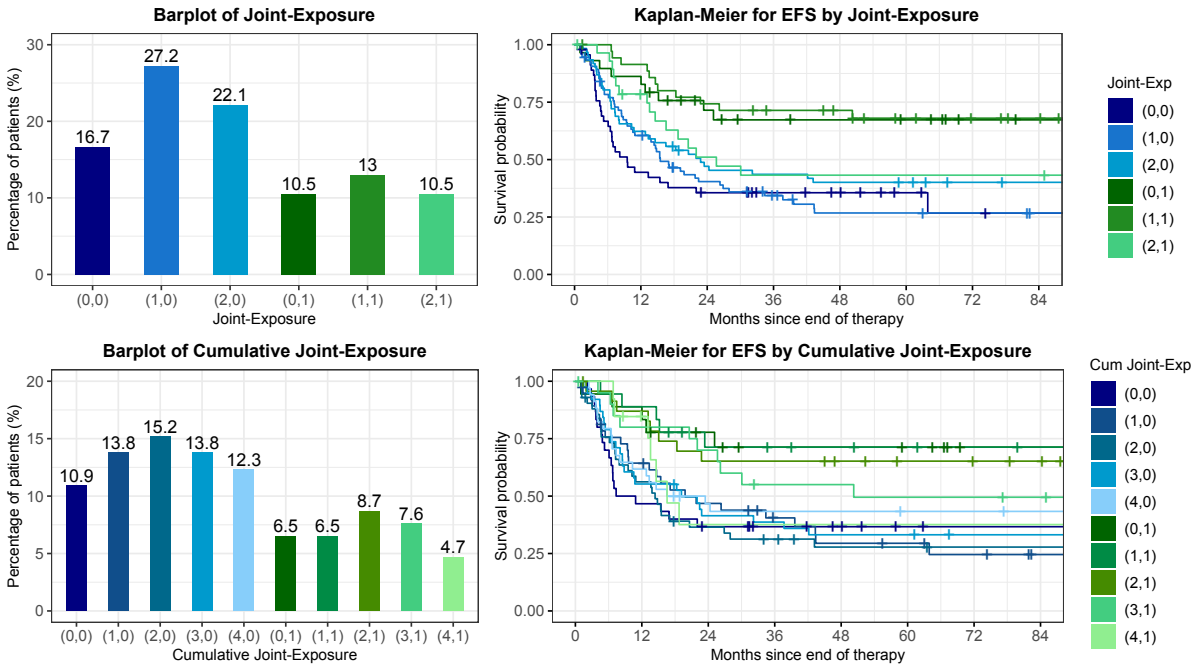

Figure 7.8. Joint-exposure characteristics. Top panels refer to *time-fixed joint-exposure* $\boldsymbol{A}_i = \left(A_i^1, A_i^2\right)$ introduced in Section 7.2.4 , where A_i^1 is the final RDI level (0: *normal* $RDI_i \geq 0.85$; 1: *low-reduction* $0.70 \leq RDI_i < 0.85$; 2: *high-reduction* $RDI_i < 0.70$) and A_i^2 is the HRe (0: *poor*: 1: *good*). Bottom panels refer to *cumulative joint-exposure* $\tilde{\boldsymbol{A}}_i = \left(\text{cum}\left(\bar{\boldsymbol{A}}_i^1\right), A_i^2\right)$, where $\text{cum}\left(\bar{\boldsymbol{A}}_i^1\right)$ is the cumulative pre/post-operative RDI level described in Section 7.2.5 and A_i^2 is the HRe (0: *poor*: 1: *good*). In both cases, left panels report percentage of patients by joint-exposure levels and right panels display Kaplan-Meier estimators for EFS curves stratified by joint-exposure levels.

7.3.2. IPTW diagnostics

Different specifications of the subject-specific standardized weights for final RDI level $SW_i^{A^1}$, HRe category $SW_i^{A^2}$ and pre/post-operative RDI levels $SW_i^{\bar{\boldsymbol{A}}^1} = SW_i^{A_{pre}^1} \cdot SW_i^{A_{post}^1}$ were investigated in order to check whether and which models best satisfied *positivity* and *no misspecification*. As mentioned in Sections 7.2.4 and 7.2.5, multinomial logistic regression models were used for both numerators and denominators of $SW_i^{A^1}$ and $SW_i^{\bar{\boldsymbol{A}}^1}$, whereas binary logistic regression models were adopted for $SW_i^{A^2}$. In all cases, the following four different model specifications in terms of confounding features for the denominators were compared:

1. each confounding covariate entered the IPTW model as a main effect only and the MOTox scores were linearly related to the log-odds;

2. specification 1 + two interaction terms linearly related to the log-odds, that is (i) interaction between pre-operative rule-specific and generic MOTox scores and (ii) interaction between post-operative rule-specific and generic MOTox scores

3. specification 1 + four interaction terms between the four MOTox scores and the trial assumed linearly related to the log-odds;

4. each categorical/binary confounding covariates entered the IPTW model as a main effect only and cubic smoothing B-splines with 3 internal knots were used to model

the relationship between each of the continuous MOTox scores and the log-odds of treatment.

Table 7.3 reports the summaries of the stabilized weights obtained with the different specifications for final RDI level $SW_i^{A^1}$, HRe category $SW_i^{A^2}$ and pre/post-operative RDI levels $SW_i^{\bar{\boldsymbol{A}}^1}$.

By examining the distributions of the standardized weights for final RDI, there was no evidence of non-positivity or of misspecification for IPTW methods 1 and 2 (mean values of about 0.99 without extreme values), whereas methods 3 and 4 presented lower mean values and higher standard deviations. The same was confirmed by the diagnostics balance plot in top-left panel of Figure 7.9, where the mean absolute standardized differences for final RDI confounders in the unweighted sample (black points) always exceeded those in the weighted samples, and the lowest values were observed for IPTW 1 and 2. IPTW model 1 was finally selected among the two as it had a mean value closer to 1 and lower standard deviation.

Similarly, according to the distributions of the standardized weights for HRe models, there was no evidence of non-positivity or misspecification in the four IPTW methods: they all presented a mean value of 1 with standard deviation from 0.22 to 0.25. In terms of covariates balance (top-right panel in Figure 7.9, all IPTW methods performed better than the unweighted sample (black) but IPTW 4 (blue) was worse than the others. In the absence of any particular contraindications, IPTW model 1 was finally selected as it was simpler (in terms of features) and had lower standard deviation weights.

Table 7.3. Inverse Probability of Treatment Weighting (IPTW) diagnostics based on summaries of stabilized weights related to final RDI level $SW_i^{A^1}$, HRe category $SW_i^{A^2}$ and pre/post-operative RDI levels $SW_i^{\bar{\boldsymbol{A}}^1}$ computed using the four different specifications listed in Section 7.3.2.

	Final RDI level: $SW_i^{A^1}$	
Specification	*Mean (s.d.)*	*Min/Max*
IPTW 1	0.988 (0.668)	0.330/5.252
IPTW 2	0.987 (0.682)	0.354/5.189
IPTW 3	0.979 (0.700)	0.324/5.469
IPTW 4	0.968 (0.797)	0.326/6.946
	HRe: $SW_i^{A^2}$	
Specification	*Mean (s.d.)*	*Min/Max*
IPTW 1	1.001 (0.200)	0.598/1.746
IPTW 2	1.001 (0.201)	0.603/1.780
IPTW 3	1.001 (0.242)	0.578/2.116
IPTW 4	0.999 (0.250)	0.531/2.373
	Pre/Post RDI levels: $SW_i^{\bar{\boldsymbol{A}}^1}$	
Specification	*Mean (s.d.)*	*Min/Max*
IPTW 1	0.988 (0.839)	0.285/7.109
IPTW 2	0.994 (0.910)	0.267/8.555
IPTW 3	0.998 (1.101)	0.266/11.438
IPTW 4	0.998 (1.245)	0.161/12.959

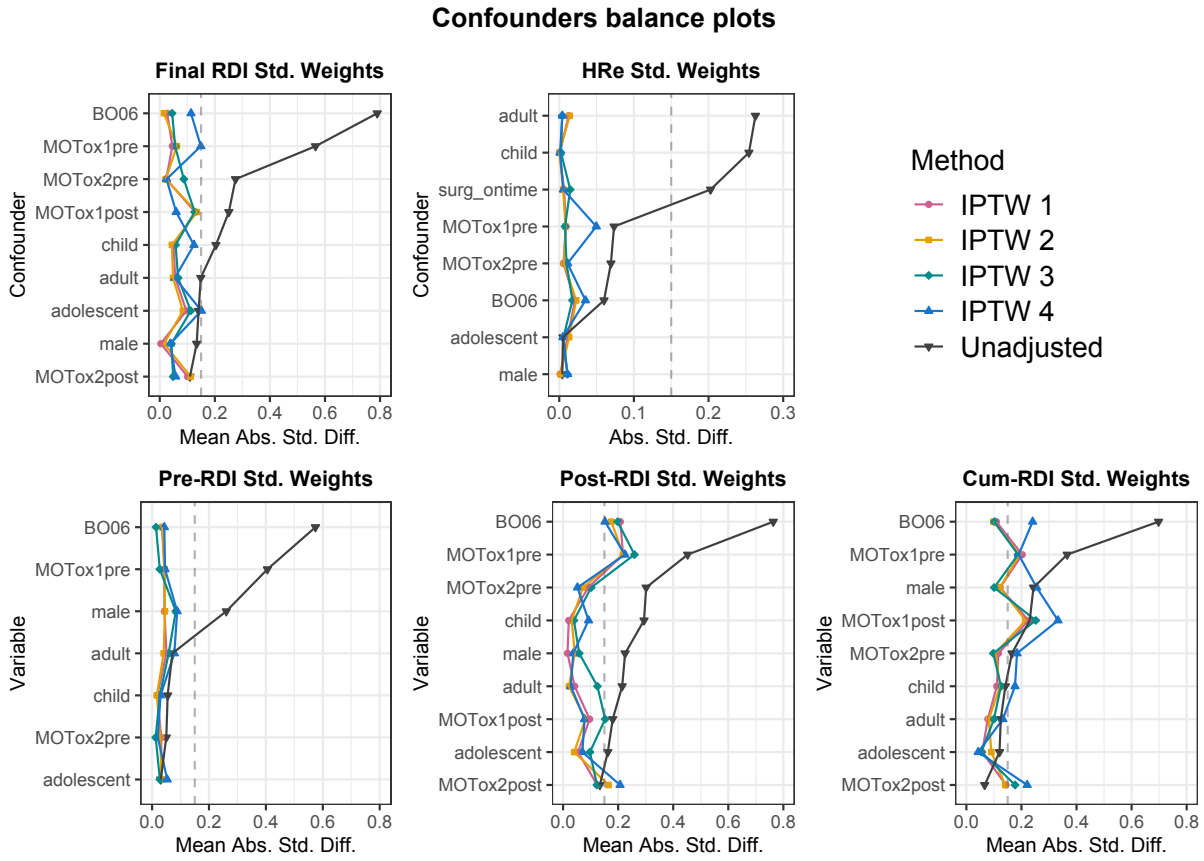

Figure 7.9. Diagnostic balance plot for Inverse Probability of Treatment Weighting (IPTW). Lines represent the (mean) absolute standardized differences for each exposure-related confounder according to the four different specification methods introduced in Section 7.3.2 and their unadjusted versions (pink: *IPTW 1*; orange: *IPTW 2*; green: *IPTW 3*; blue: *IPTW 4*; black: *Undajusted*).

IPTW methods for pre/post-operative RDI levels was selected as trade-off between the two product components. No evidence of assumptions violation was observed according to the distributions of the standardized weights $SW_i^{\bar{\boldsymbol{A}}^1}$. IPTW 4 method resulted in a worse balance of confounders in terms of mean absolute standardised differences for cumulative-RDI levels based on $SW_i^{\bar{\boldsymbol{A}}^1}$ (see bottom-right panel in Figure 7.9). The same was valid for post-RDI levels using $SW_i^{A^1_{post}}$ obtained through IPTW 3 (bottom-centre panel). Between IPTW methods 1 and 2, both with a mean value of about 0.99, IPTW 1 was selected as it was simpler (in terms of features) and had lower standard deviation weights.

The formulas of the denominators of $SW_i^{A^1}$, $SW_i^{A^1_{pre}}$ $SW_i^{A^1_{post}}$ and $SW_i^{A^2}$ related to the selected IPTW specifications are reported in Appendix D.2.

Left panel of Figure 7.10 shows the standardized weights SW_i in Equation (7.8) obtained as product of $SW_i^{A^1}$ and $SW_i^{A^2}$ to be used for create the pseudo-population in case of *time-fixed joint exposure*. The y-axis is in logarithmic scale. Mean value of SW_i was 0.983 ($s.d. = 0.694$) with minimum and maximum values of 0.272 and 4.849. Analogously, right panel of Figure 7.10 shows the standardized weights SW_i in Equation (7.12) obtained as product of $SW_i^{\bar{\boldsymbol{A}}^1}$ and $SW_i^{A^2}$ to be used for create the pseudo-population in case of *time-*

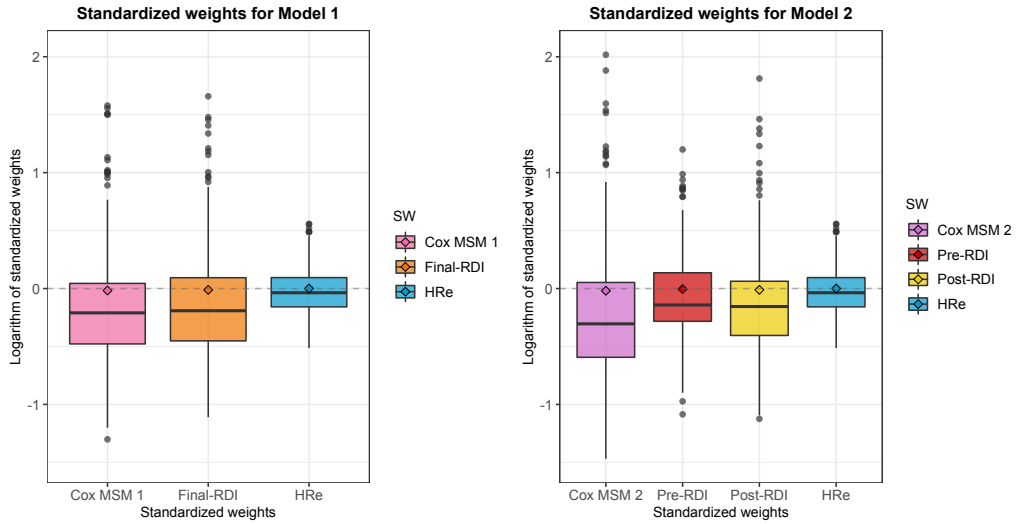

Figure 7.10. Diagnostic boxplots of subject-specific standardized weights computed via Equations (7.8) (left panel) and (7.12) (right panel). The scale on the y-axis is lgarithmic. Diamonds represent the mean values (in logarithmic scale).

varying joint exposure. Mean value of SW_i was 0.981 ($s.d. = 0.865$) with minimum and maximum values of 0.230 and 7.518. These weights satisfied all the required assumptions and were finally used to fit on the relative pseudo-populations the IPT weighted Cox models in Equations (7.7) and (7.11).

7.3.3. Causal inference through marginal structural Cox models

Once met causal inference assumptions, association was causation in both pseudo populations. The causal parameters $\boldsymbol{\beta}$ in Cox MSMs (7.6) and (7.10) were hence estimated through their consistent associational parameters $\boldsymbol{\theta}$ in IPT weighted Cox models (7.7) and (7.11) fitted on the relative pseudo-populations. Obtained estimates were finally compared to the results obtained by fitting the corresponding standard (i.e., unweighted) Cox models on the original population.

Estimated parameters for both Cox MSMs and their unweighted versions are reported in Table 7.4. In Cox MSM 1 and 2 robust standard errors for computing the confidence interval of each coefficient were obtained via the option robust=TRUE in R function coxph [201]. Figure 7.11 graphically displays the Hazard Ratios related to the different joint-exposure levels for Cox MSMs in 7.6 and 7.10 fitted on the pseudo population (left panels) and the results for corresponding unweighted models (right panels).

Top panels refers to the causal structure of DAG-1 presented in Section 7.2.4. Reference level was PRs with *normal* RDI level at the end of treatment, i.e., $(a_1, a_2) = (0, 0)$. Considering the unweighted Cox model 1 (top-right), which represents the RDI-analogue of the ITT Cox model presented by Lewis *et al.* (2007) [119] without considering IPT weights, in PRs the RDI reductions appeared associated with an improvement in EFS, even if not statistically significant. With respect to GRs receiving a *normal* RDI, GRs receiving a *low-reduction* experienced an event 12% slower ($HR = 0.273/0.309 = 0.88$) whereas those

Table 7.4. Estimated parameters $\hat{\boldsymbol{\beta}}$ along with their 95% Confidence Intervals (CIs) for Cox MSMs 1 and 2 in Equation 7.6 and 7.10, respectively, and for their corresponding unweighted versions.

	Cox MSM 1		**Unweighted Cox model 1**	
Treatment	$\hat{\boldsymbol{\beta}}$	*95% CIs*	$\hat{\boldsymbol{\beta}}$	*95% CIs*
$a_1 = 1$	-0.498	$[-0.986; -0.010]$	-0.116	$[-0.568; 0.335]$
$a_1 = 2$	-0.833	$[-1.409; -0.257]$	-0.359	$[-0.844; 0.127]$
$a_2 = 1$	-1.914	$[-2.880; -0.948]$	-1.175	$[-1.921; -0.429]$
$a_1 = 1 \times a_2 = 1$	0.762	$[-0.399; 1.923]$	-0.006	$[-0.997; 0.984]$
$a_1 = 2 \times a_2 = 1$	1.850	$[0.643; 3.057]$	0.979	$[0.020; 1.938]$
	Cox MSM 2		**Unweighted Cox model 2**	
Treatment	$\hat{\boldsymbol{\beta}}$	*95% CIs*	$\hat{\boldsymbol{\beta}}$	*95% CIs*
$\text{cum}\,(\bar{\boldsymbol{a}}_1)$	-0.181	$[-0.370; 0.009]$	-0.062	$[-0.197; 0.072]$
$a_2 = 1$	-1.823	$[-2.714; -0.932]$	-1.461	$[-2.215; -0.707]$
$\text{cum}\,(\bar{\boldsymbol{a}}_1) \times a_2 = 1$	0.397	$[0.052; 0.743]$	0.305	$[0.010; 0.601]$

receiving a *high-reduction* experienced an event 86% faster ($HR = 0.574/0.309 = 1.86$). However, these results were affected by the toxicity-treatment-adjustment bias and could not be interpreted in a causal way. In fact, the final value of RDI was the realisation of the treatment trajectory as result of both the severity of the overall toxicity experienced by each patient and the side-effects handling operated by physicians. To overcome these issues, Cox MSM 1 (top-left) represented a clear improvement with respect to its unweighted version. At the same final RDI level, a *good* response caused an 85.2% decrease in the risk of an event ($\exp(\hat{\beta}_3) = 0.148$) with respect to a *poor* one. Reductions in the final RDI caused better EFS for PRs, whereas a reverse causal association was founded in GRs. In particular, the higher the final reduction, the better the survival for PRs (estimated HRs were 0.608 and 0.435 for *low* and *high* reduction PRs, respectively). On the contrary, the higher the final reduction, the worsen the survival for GRs: GRs with *low* or *high* reduction experienced an event 1.30 ($HR = 0.192/0.148 = 1.30$) or 2.76 ($HR = 0.408/0.148 = 2.76$) times faster than GRs with *normal*-RDI.

Bottom panels refers to the causal structure of DAG-2 presented in Section 7.2.5, where reference level was PRs without reduction, neither pre nor post surgery, i.e., $(\text{cum}\,(\bar{\boldsymbol{a}}_1), a_2)$ is $(0, 0)$. Results were in line with previous results: (i) GRs presented better survival with respect to PRs; (ii) an increasing number of pre/post-operative reductions in RDI showed opposite trends for PRs and GRs, improving and worsening EFS, respectively. This was even more evident in the Cox MSM 2 (bottom-left) than in its unweighted version (bottom-right) affected by the *toxicity-treatment-adjustment* bias: point estimates with respect to reference level dramatically improved even if statistical significance did not change, again showing the bias due to *toxicity-treatment-adjustment*. Considering parameter estimates for Cox MSM 2 (see Table 7.4), at the same RDI level, a *good* response caused an 83.8% decrease in the risk of an event ($\exp(\hat{\beta}_3) = 0.162$) with respect to a *poor* one. Moreover, 1-unit increase in the number of reductions of 15-30% (i.e., 1-unit increase in $\text{cum}\,(\bar{\boldsymbol{a}}_1)$) caused a decrease of 16.5% in the risk of an event for PRs ($\exp(\hat{\beta}_1) = 0.835$) and an increase of 24.1% for GRs ($\exp(\hat{\beta}_1 + \hat{\beta}_3) = 1.241$).

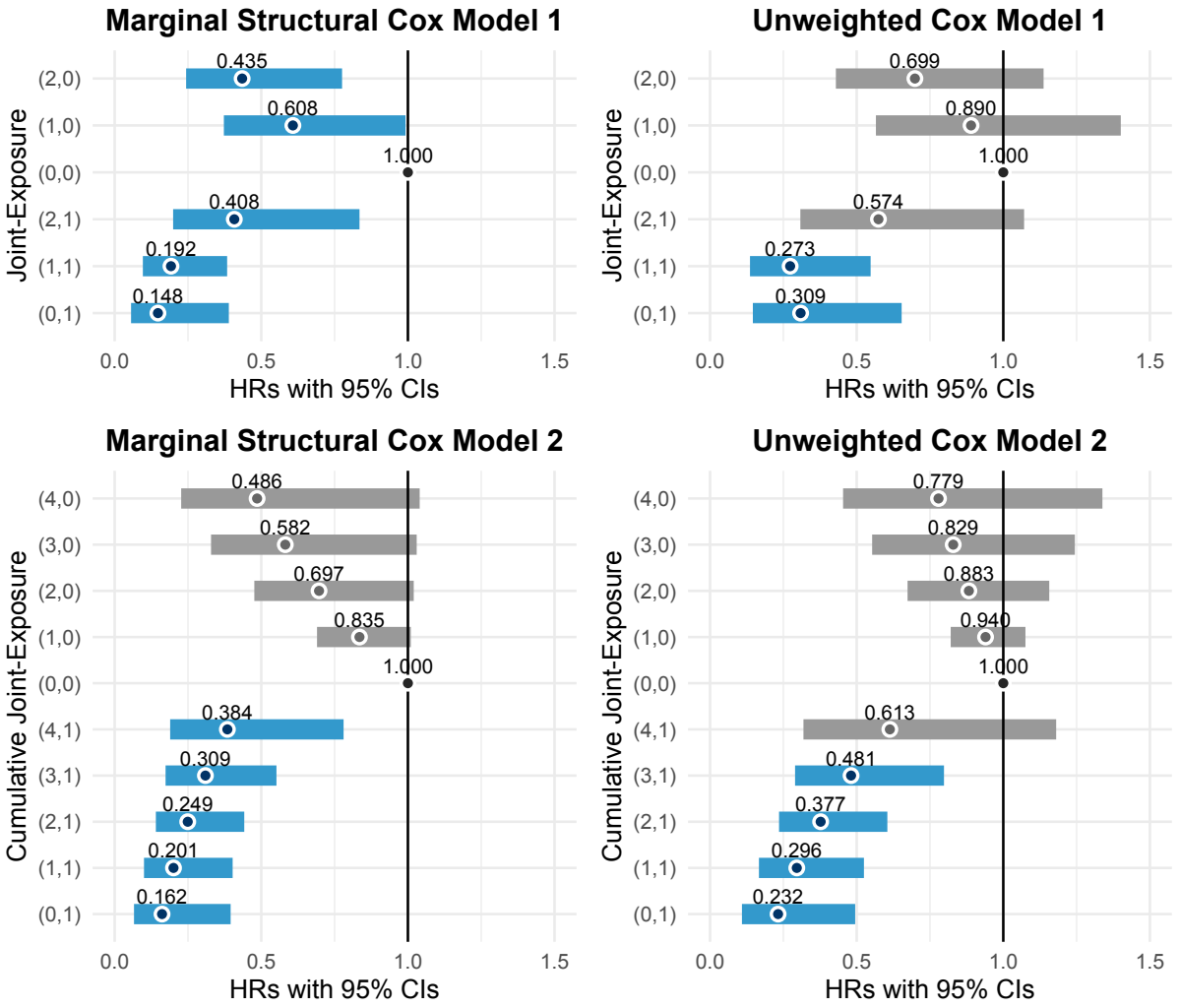

Figure 7.11. Graphical displays of Hazard Ratios (HRs) along with their 95% Confidence Intervals (CIs) for Marginal Structural Cox Models (left panels) and corresponding unweighted Cox models (right panels). Top panels refer to Cox MSM 1 in Equation (7.6), where reference level is *poor* responder with *normal* RDI level at the end of treatment (i.e., $(a_1, a_2) = (0, 0)$). Bottom panels refer to Cox MSM 2 in Equation (7.10), where reference level is *poor* responder without reduction, neither pre nor post surgery, i.e., $(\text{cum}\,(\bar{\boldsymbol{a}}_1)\,, a_2) = (0, 0)$.

One possible clinical explanation for these reverse behaviours could lie in the fact that chemotherapy also damages non-cancerous cells and processes of the immune system that can detect and kill cancer cells. In PRs, for whom chemotherapy is less effective, this negative effect is not offset by treatment efficacy, and an increase in RDI may be detrimental to survival due to the impact on the immune system.

7.4. Final remarks

In cancer trials, longitudinal chemotherapy data are problematic to analyse due to the presence of negative feedback between exposure to cytotoxic drugs and consequent toxic side effects. Therapy administration is usually complicated by the dynamical adjustment of the treatment based on patients' clinical picture, especially on chemotherapy-induced

multi-systemic toxicities. For this reason, chemotherapy is usually modelled by Intention-To-Treat (ITT) analysis [70], although the introduction of the Received Dose Intensity (RDI) [86] marked a significant departure from ITT in the direction of a closer description of the actual clinical practice. The main issue in analysing actual treatment lies in the fact that toxicities act as time-dependent confounders for the effect of chemotherapy intensity exposure on survival, determining the *toxicity-treatment-adjustment* bias if not properly considered. Suitable methodologies are hence needed to control for exposure-affected (time-varying) toxicity confounding in longitudinal chemotherapy data. In addition, since the assignment of dose reductions/interruptions or delays in administration during treatment is determined not by individual toxicities but by the overall toxic burden of each patient, the different types and number of side effects must be adequately summarized to be included in the analysis.

Motivated by a sharp yet delicate clinical question on the effect of treatment modifications on Event-Free Survival (EFS) in osteosarcoma patients, this chapter proposed Marginal Structural Models (MSMs) in combination with Inverse-Probability-of-Treatment Weighted (IPTW) estimators to assess the causal effects of chemotherapy intensity exposure seen in terms of both Histological Response (HRe) and RDI reductions compared to protocol. Control arms data from BO03 and BO06 trials for osteosarcoma were analyzed. Since only the most severe side-effects were recorded in BO03, the analysis of such mixed longitudinal/non-longitudinal data required both an original analytical strategy and an unconventional model formulation. First, pre and post-operative toxicity data were summarized using a Multiple Overall Toxicity (MOTox) approach [190] based on most severe CTACE grades of both *rule-specific* and *generic* side effects. This allowed (i) to reduce the number of possible confounders combinations, avoiding non-positivity and highly-correlated data, and (ii) to meet the clinical rationale of tailoring treatment according to the patient's overall toxic burden in the case of multiple toxic side effects. Then, two different joint-exposure characterizations – which met *consistency* according to experts – were defined unambiguously based on *time-fixed final RDI* or *time-dependent pre/post-operative RDI*) combined with HRe. This led to the introduction of two alternative Direct Acyclic Graphs (DAGs) to identify all possible (time-dependent) confounders and their relationships with both joint-exposure and EFS outcome, validating the assumption of *no unmeasured confounding*. Suitable IPTW-based techniques and Cox MSMs, representing the causal RDI analogues of the ITT Cox model presented by Lewis *et al.* (2007) [119], were finally designed to mimic a randomized trial where the joint-exposure intensity was no longer confounded by toxicities. Once *positivity* and *no misspecification* were satisfied, in the pseudo-population thus created, a crude analysis sufficed to estimate the causal effect of joint-exposure modifications on EFS.

Regardless of RDI-level, in both Cox MSMs all estimated HRs were lower for *Good* Responders (GRs) than for *Poor* ones (PRs), showing that GRs presented a better EFS than PRs in all cases. This was not surprising because HRe is the strongest prognostic survival factor known to date in osteosarcoma [31]. Increasing RDI-reductions created two opposite trends for PRs and GRs: the higher the reduction in final or pre/post-operative RDI, the better (worsen) was the EFS in PRs (GRs). One possible clinical explanation for

these inverse behaviours could lie in the effect of chemotherapy on non-cancerous cells. By targeting a broad spectrum of cells, chemotherapy also damages the processes and mechanisms of the immune system that can detect and kill cancer cells. While in GRs this negative effect may be largely offset by the efficacy of the tumour therapy, in PRs – for whom chemotherapy is less effective – an increased RDI may be harmful to survival due to the impact on the immune system.

This study highlighted both the confounding nature of toxicity data and the detrimental effect of not considering them in the analysis, showing the potential pitfalls of a naive RDI-based analysis of chemotherapy data. When ITT models were translated into RDI-based ones by simply neglecting the role of toxicities as in the unweighted Cox models, results were clearly affected by the *toxicity-treatment-adjustment* bias. The use of Cox MSMs allowed to model the contribution of patient's toxicity history to EFS through the realisation of the (cumulative) joint-exposures. In other words, the use of IPTW-based Cox MSMs broke the feedback between side effects and therapy adjustments, resulting in unbiased estimates of the effect of treatment modifications on EFS and describing better the effect of low-intensity regimens.

The presented IPTW-based MSMs have clear limitations. The property of MSMs to give unbiased estimates relies on the four main assumptions presented in Section 7.2.2, which are often unverifiable and mostly based on experts knowledge. This is really the potential weakness of both the analysis presented above and the methodology based on IPTW and MSMs in general. In addition, the lack of longitudinal confounders in BO03 forced the causal structures represented by the DAGs in Figure 7.7. These DAGs relied on the assumption that the most severe CTCAE grades of each toxicity in pre- and post-operative treatment predicted well the final and pre/post-operative RDI values, thus flattening the toxicity history. This assumption might still be challenged, since severe toxicities might look simultaneous producing some significant interactions. However, there is no guarantee that severe CTCAE grades occurred simultaneously, so these interactions were not considered. The development of appropriate causal structures and methodologies for studying chemotherapy data using a cycle-by-cycle longitudinal perspective would be of great interest for future analyses, as it would overcome this issue, but the need for adequate toxicity data collection still remains.

In summary, this chapter showed the difficulty of analysing chemotherapy data on a RDI-based approach, mostly originated from data quality. The main contribution of this work is the presentation of an all-round analysis of complex chemotherapy data, with tutorial-like explanations of the difficulties encountered and the problem-solving strategies deployed. Focusing on a way of analysing chemotherapy data that is RDI-based rather than ITT-based, it illustrated the key role played by toxicities in this transition and showed the detrimental effect of neglecting them. To the best of our knowledge, this is the first application of IPTW-based techniques to survival data from a randomized trial of chemotherapy in order to eliminate the *toxicity-treatment-adjustment* bias.

D. Appendix to Chapter 7

D.1. Pre/Post-operative Received Dose Intensity definitions

For each patient i, let np_i denote the number of cycles performed before the surgery. *Pre-operative* period is made up of cycles performed before surgery, i.e., $j \in \{1, ..., np_i\}$. *Post-operative* period is made up of cycles performed after surgery, i.e., $j \in \{np_i+1, ..., 6\}$. To consider the two periods separately, definitions in Equations (7.2), (7.3) and (7.4) in Section 7.1.3 can be adapted as follows.

Pre/Post-operative standardized dose. The pre-operative and post-operative standardized doses $\Delta_{i,pre}$ and $\Delta_{i,post}$ for patient i are defined as

$$\Delta_{i,pre} = \frac{1}{2}\left(\Delta_{i,pre}^{CDDP} + \Delta_{i,pre}^{DOX}\right) = \frac{1}{2 \cdot np_i}\left(\sum_{j=1}^{np_i} \delta_{ij}^{CDDP} + \sum_{j=1}^{np_i} \delta_{ij}^{DOX}\right),$$

$$\Delta_{i,post} = \frac{1}{2}\left(\Delta_{i,post}^{CDDP} + \Delta_{i,post}^{DOX}\right) = \frac{1}{2 \cdot (5 - np_i)}\left(\sum_{j=np_i+1}^{6} \delta_{ij}^{CDDP} + \sum_{j=np_i+1}^{6} \delta_{ij}^{DOX}\right).$$

Pre/Post-operative standardized time. The pre-operative standardized time for the i-th patient is

$$\Gamma_{i,pre} = \frac{\text{actual pre-operative time}}{\text{anticipated pre-operative time}}$$

where

- *actual pre-operative time* is the difference in days between the starting date of cycle 1 and the date of the surgery,
- *anticipated pre-operative time* is $21 \times np_i$ days, i.e., np_i cycles lasting 21 days each.

Similarly, the post-operative standardized time for patient i is

$$\Gamma_{i,post} = \frac{\text{actual post-operative time}}{\text{anticipated post-operative time}}$$

where

- *actual post-operative time* is the difference in days between the surgery date and the 3rd day after the start of cycle 6,
- *anticipated post-operative time* is $14 + (5 - np_i) \times 21 + 3$ days, i.e., 14 days of surgery, $5 - np_i$ cycles lasting 21 days each and 3 days after the start of cycle 6.

Pre/Post-operative Received Dose Intensity. The pre-operative and post-operative Received Dose Intensities (RDIs) for patient i are defined as

$$RDI_{i,pre} = \frac{\Delta_{i,pre}}{\Gamma_{i,pre}}, \tag{7.13}$$

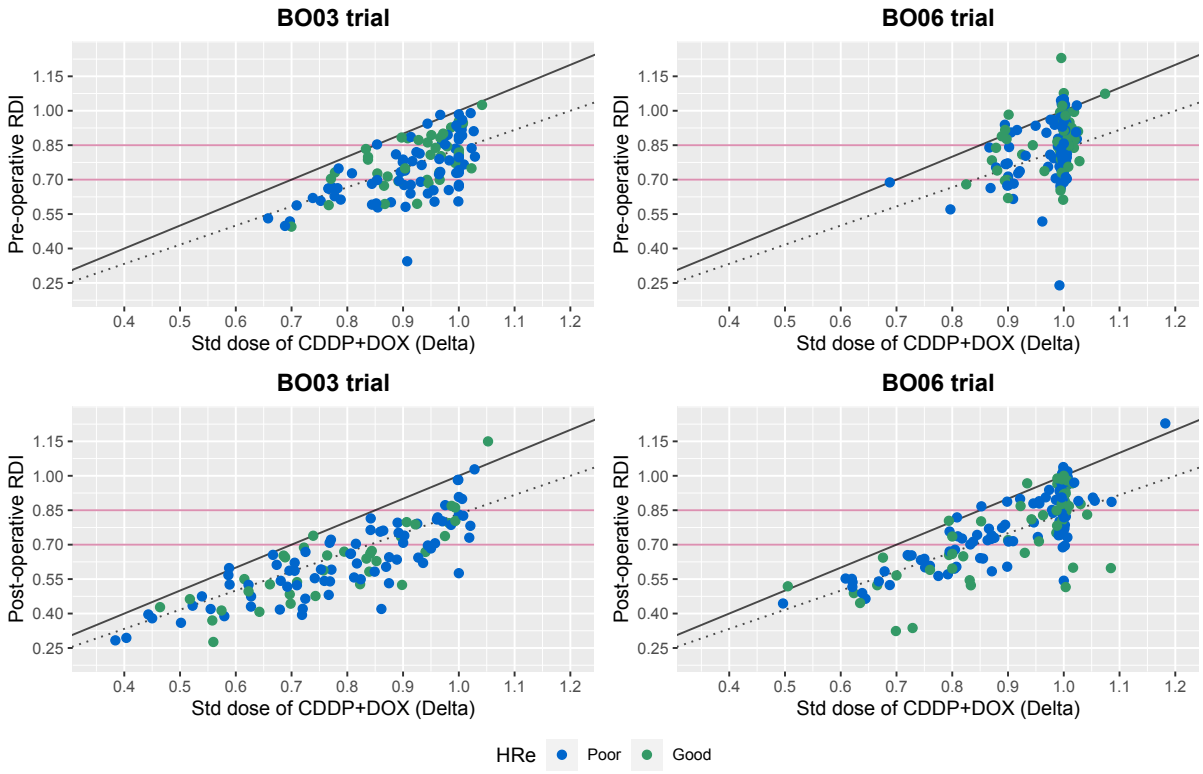

Figure 7.12. Top panels: Scatter plots of pre-operative RDI (i.e., $RDI_{i,pre}$ in Equation (7.13)) against pre-operative standardized dose of CDDP+DOX ($\Delta_{i,pre}$) conditional on trial (*BO03*: left panel; *BO06*: right panel) and HRe (blue: *poor*; green: *good*).
Bottom panels: Scatter plots of post-operative RDI (i.e., $RDI_{i,post}$ in Equation (7.14)) against post-operative standardized dose of CDDP+DOX ($\Delta_{i,post}$) conditional on trial (*BO03*: left panel; *BO06*: right panel) and HRe (blue: *poor*; green: *good*).

$$RDI_{i,post} = \frac{\Delta_{i,post}}{\Gamma_{i,post}}. \tag{7.14}$$

The RDI computed on the whole treatment as in Equation (7.4) is not the sum of pre-operative and post-operative RDIs, i.e., $RDI_i \neq RDI_{i,pre} + RDI_{i,post}$.

A summary of $RDI_{i,pre}$ and $RDI_{i,post}$ exposure characteristics for the whole cohort and conditional on trials is reported in Figure 7.6. Figure 7.12 shows the scatter plots of pre- (top panels) and post- (bottom panels) operative RDI ($RDI_{i,k}$) against their relative standardized doses of CDDP+DOX ($\Delta_{i,k}$) conditional on trial (left panel: *BO03*; right panel: *BO06*) and HRe (blue: *poor*; green: *good*). The solid horizontal lines in pink vertically divide patients with *normal* RDI levels ($RDI_{i,k} \geq 0.85$) from *low* reduction ($0.70 \leq RDI_{i,k} < 0.85$) and *high* reduction ($RDI_{i,k} < 0.70$) patients. The solid diagonal line in black satisfies equation $RDI_{i,k} = \Delta_{i,k}$, dividing the group of patients with standardized time $\Gamma_{i,k} > 1$ (delayed therapy, below the line) from the group, almost void, of patients with $\Gamma_{i,k} < 1$ (anticipated therapy, above the line). The dotted diagonal line in black satisfies equation $RDI_{i,k} = \Delta_{i,k}/1.2$, dividing the group of patients with therapy delayed by more than 20% of anticipated time (below the dotted line) from the group of patients with therapy delayed by less than 20% of anticipated time (between solid and dotted black lines). This figure clearly shows the difference of treatment delivery in BO03

vs. BO06, also considering pre/post-operative periods separately. It also shows the lack of a clear association between HRe and pre/post-operative RDI.

D.2 Denominator specifications for selected IPTW models

In Section 7.3.2, different model specifications (in terms of confounding covariates $\boldsymbol{L}$ for the denominators) to determine the subject-specific standardized weights for final RDI level $SW_i^{A^1}$, HRe category $SW_i^{A^2}$ and pre/post-operative RDI levels $SW_i^{\bar{\boldsymbol{A}}^1} = SW_i^{A^1_{pre}} \cdot SW_i^{A^1_{post}}$ were investigated. The following denominator formulas were selected:

- multinomial logistic regression model for denominator of final RDI level $SW_i^{A^1}$:

$$\begin{aligned}\log \frac{\Pr\left(A_i^1 = a \middle| \boldsymbol{L}_i^1\right)}{\Pr\left(A_i^1 = 0 \middle| \boldsymbol{L}_i^1\right)} = {} & \gamma_{a0} + \gamma_{a1} \cdot \mathbb{1}_{(\texttt{trial}_i = BO06)} + \gamma_{a2} \cdot \mathbb{1}_{(\texttt{age}_i = adolescent)} + \\ & \gamma_{a3} \cdot \mathbb{1}_{(\texttt{age}_i = adult)} + \gamma_{a4} \cdot \mathbb{1}_{(\texttt{gender}_i = male)} + \\ & \gamma_{a5} \cdot \texttt{MOTox}_{i,pre}^{(gen)} + \gamma_{a6} \cdot \texttt{MOTox}_{i,pre}^{(rule)} + \\ & \gamma_{a7} \cdot \texttt{MOTox}_{i,post}^{(gen)} + \gamma_{a8} \cdot \texttt{MOTox}_{i,post}^{(rule)}\end{aligned}$$

  where confounding covariates are

$$\begin{aligned}\boldsymbol{L}_i^1 &= \left(\boldsymbol{L}_i^{base}, \boldsymbol{L}_{i,pre}^{tox}, \boldsymbol{L}_{i,post}^{tox}\right) = \\ &= \left(\texttt{trial}_i, \texttt{age}_i, \texttt{gender}_i, \texttt{MOTox}_{i,pre}^{(gen)}, \texttt{MOTox}_{i,pre}^{(rule)}, \texttt{MOTox}_{i,post}^{(gen)}, \texttt{MOTox}_{i,post}^{(rule)}\right);\end{aligned}$$

- binary logistic regression model for denominator of HRe category $SW_i^{A^2}$

$$\begin{aligned}\log \frac{\Pr\left(A_i^2 = 1 \middle| \boldsymbol{L}_i^2\right)}{1 - \Pr\left(A_i^2 = 1 \middle| \boldsymbol{L}_i^2\right)} = {} & \alpha_0 + \alpha_1 \cdot \mathbb{1}_{(\texttt{trial}_i = BO06)} + \alpha_2 \cdot \mathbb{1}_{(\texttt{age}_i = adolescent)} + \\ & \alpha_3 \cdot \mathbb{1}_{(\texttt{age}_i = adult)} + \alpha_4 \cdot \mathbb{1}_{(\texttt{gender}_i = male)} + \\ & \alpha_5 \cdot \texttt{MOTox}_{i,pre}^{(gen)} + \alpha_6 \cdot \texttt{MOTox}_{i,pre}^{(rule)} + \\ & \alpha_7 \cdot \mathbb{1}_{(\texttt{surgery}_i = on\,time)}\end{aligned}$$

  where confounding covariates are

$$\begin{aligned}\boldsymbol{L}_i^2 &= \left(\boldsymbol{L}_i^{base}, \boldsymbol{L}_{i,pre}^{tox}, L_i^{surg}\right) = \\ &= \left(\texttt{trial}_i, \texttt{age}_i, \texttt{gender}_i, \texttt{MOTox}_{i,pre}^{(gen)}, \texttt{MOTox}_{i,pre}^{(rule)}, \texttt{surgery}_i\right);\end{aligned}$$

- multinomial logistic regression model for denominator of pre-operative RDI level $SW_i^{A^1_{pre}}$:

$$\begin{aligned}\log \frac{\Pr\left(A_{i,pre}^1 = a \middle| \boldsymbol{L}_{i,pre}^1\right)}{\Pr\left(A_{i,pre}^1 = 0 \middle| \boldsymbol{L}_{i,pre}^1\right)} = {} & \gamma_{a0} + \gamma_{a1} \cdot \mathbb{1}_{(\texttt{trial}_i = BO06)} + \gamma_{a2} \cdot \mathbb{1}_{(\texttt{age}_i = adolescent)} + \\ & \gamma_{a3} \cdot \mathbb{1}_{(\texttt{age}_i = adult)} + \gamma_{a4} \cdot \mathbb{1}_{(\texttt{gender}_i = male)} + \\ & \gamma_{a5} \cdot \texttt{MOTox}_{i,pre}^{(gen)} + \gamma_{a6} \cdot \texttt{MOTox}_{i,pre}^{(rule)}\end{aligned}$$

where confounding covariates are

$$\boldsymbol{L}^1_{i,pre} = \left(\boldsymbol{L}^{base}_i,\, \boldsymbol{L}^{tox}_{i,pre}\right) = \left(\texttt{trial}_i,\, \texttt{age}_i,\, \texttt{gender}_i,\, \texttt{MOTox}^{(gen)}_{i,pre},\, \texttt{MOTox}^{(rule)}_{i,pre}\right);$$

- multinomial logistic regression model for denominator of post-operative RDI level $SW_i^{A^1_{post}}$:

$$\begin{aligned}\log \frac{\Pr\left(A^1_{i,post} = a \middle| \bar{\boldsymbol{L}}^1_{i,post}, A^1_{i,pre}\right)}{\Pr\left(A^1_{i,post} = 0 \middle| \bar{\boldsymbol{L}}^1_{i,post}, A^1_{i,pre}\right)} = {} & \gamma_{a0} + \gamma_{a1} \cdot \mathbb{1}_{(\texttt{trial}_i = BO06)} + \gamma_{a2} \cdot \mathbb{1}_{(\texttt{age}_i = adolescent)} + \\ & \gamma_{a3} \cdot \mathbb{1}_{(\texttt{age}_i = adult)} + \gamma_{a4} \cdot \mathbb{1}_{(\texttt{gender}_i = male)} + \\ & \gamma_{a5} \cdot \texttt{MOTox}^{(gen)}_{i,pre} + \gamma_{a6} \cdot \texttt{MOTox}^{(rule)}_{i,pre} + \\ & \gamma_{a7} \cdot \mathbb{1}_{(A^1_{i,pre}=1)} + \gamma_{a8} \cdot \mathbb{1}_{(A^1_{i,pre}=2)} + \\ & \gamma_{a9} \cdot \texttt{MOTox}^{(gen)}_{i,post} + \gamma_{a10} \cdot \texttt{MOTox}^{(rule)}_{i,post}.\end{aligned}$$

where $A^1_{i,pre}$ is the pre-operative RDI level and confounding covariates are

$$\begin{aligned}\bar{\boldsymbol{L}}^1_{i,post} &= \left(\boldsymbol{L}^{base}_i,\, \boldsymbol{L}^{tox}_{i,pre},\, \boldsymbol{L}^{tox}_{i,post}\right) \\ &= \left(\texttt{trial}_i,\, \texttt{age}_i,\, \texttt{gender}_i,\, \texttt{MOTox}^{(gen)}_{i,pre},\, \texttt{MOTox}^{(rule)}_{i,pre},\, \texttt{MOTox}^{(gen)}_{i,post},\, \texttt{MOTox}^{(rule)}_{i,post}\right).\end{aligned}$$

CHAPTER 7