

Causal effect of chemotherapy received dose intensity on survival outcome: a retrospective study in osteosarcoma

M. Spreafico*, F. Ieva, M. Fiocco

*Corresponding author: m.spreafico@math.leidenuniv.nl

Supplementary Material

A CTCAE grades specification for *rule-specific* and *generic* toxicities

In both BO03 and BO06 studies, toxic side effects were recorded using the Common Terminology Criteria for Adverse Events Version 3 (CTCAE v3.0) [1], with grades ranging from 0 (none) to 4 (life-threatening). Table S1 reports the CTCAE-grades for *rule-specific* toxicities (i.e., *leucopenia*, *thrombocytopenia*, *oral mucositis*, *ototoxicity*, *cardiotoxicity* and *neurotoxicity*) and *generic* ones (i.e., *nausea/vomiting* and *infections*).

B Assumptions for causal inference through marginal structural models

The four main assumptions for causal inference with Marginal Structural Models (MSMs) through Inverse Probability of Treatment Weighting (IPTW) are here discussed.

Exchangeability (or conditional exchangeability) implies the well-known assumption of *no unmeasured confounding* [2]. It states that exposure allocation is independent of the potential outcomes conditional on confounders [2], that is:

$$T^a \perp\!\!\!\perp A | L.$$

Table S1 Toxicity coding based on Common Terminology Criteria for Adverse Events (CTCAE) v3.0 by [1] 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
<i>Rule-specific</i>					
<i>Leucopenia</i>	$\geq 4.0 \times 10^9 / L$	$[3.0 - 4.0) \times 10^9 / L$	$[2.0 - 3.0) \times 10^9 / L$	$[1.0 - 2.0) \times 10^9 / L$	$< 1.0 \times 10^9 / L$
<i>Thrombocytopenia</i>	$\geq 100 \times 10^9 / L$	$[75 - 100) \times 10^9 / L$	$[50 - 75) \times 10^9 / L$	$[25 - 50) \times 10^9 / L$	$< 25 \times 10^9 / L$
<i>Oral Mucositis</i>	No change	Soreness or erythema	Ulcers: can eat solid	Ulcers: liquid diet only	Alimentation not possible
<i>Cardiac toxicity</i>	No change	Sinus tachycardia	Unifocal PVC ¹ arrhythmia	Multifocal PVC ¹	Ventricular tachycardia
<i>Ototoxicity</i>	No change	Slight hearing loss	Moderate hearing loss	Major hearing loss	Complete hearing loss
<i>Neurological toxicity</i>	None	Paraesthesia	Severe paraesthesia	Intolerable paraesthesia	Paralysis
<i>Generic</i>					
<i>Nausea/Vomiting</i>	None	Nausea	Transient vomiting	Continuative vomiting	Intractable vomiting
<i>Infection</i>	None	Minor infection	Moderate infection	Major infection	Major infection with hypotension

¹PVC = Premature Ventricular Contraction

In the absence of randomization, as is the case in observational studies or RCTs with interventions, it is not possible to test for exchangeability. In such situations, expert knowledge becomes essential for identifying an adequate set of joint predictors of exposure and outcome. These predictors should be chosen in a way that, within their respective levels, any associations between exposure and outcome resulting from shared underlying causes are effectively controlled for [2].

Consistency means that the outcome observed for each individual is the counterfactual outcome under the observed treatment history, that is:

$$T^a = T_i \quad \text{for every individual } i \text{ with } A_i = a.$$

This assumption is violated in the presence of misclassification bias [3] and has two requirements [4]:

- i. the exposure must be properly defined so that the counterfactual outcomes are well-defined (this implies that a specific exposure may be hypothetically assigned to a subject exposed to a different level);
- ii. a link between counterfactuals and observed data is reasonable in the context under study (this means that the equality should be valid for at least some individuals).

Although consistency can not be empirically verified, it is assumed to be plausible in (observational) studies of medical treatments, since it may be possible to change an individual's treatment status [5].

Positivity states that there is a non-zero (i.e., positive) probability of receiving every level of exposure for every combination of covariates that occur among individuals in the population [2]. In the context under analysis, this corresponds to

$$\Pr(A_i = a | \mathbf{L}_i = \mathbf{l}, V_i = v) > 0 \quad \forall a, \mathbf{l}, v.$$

If this assumption is violated, then the weights in IPTW in Manuscript Equation (9) are undefined leading to biased estimates of the causal effect.

If a subject 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 specific exposure. A solution is to restrict the inference to the subset with a positive probability of exposure, whenever possible [5]. 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 [5].

No misspecification of both weighting and outcome models means 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 all statistical models [3], as model misspecification leads to instability in the Cox MSM estimates [6, 7].

Since the presence of estimated stabilized weights with the mean far from one or with extreme values suggest possible violation of positivity or misspecification of the weight model [5], proper model specifications can be checked by exploring the distribution of weights [2]. 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 [8].

C Flowchart BO03/BO06 cohort selection

Figure S1 displays the consort diagram related to the final cohort of 276 patients (114 and 162 from BO03 and BO06, respectively) included in the analyses.

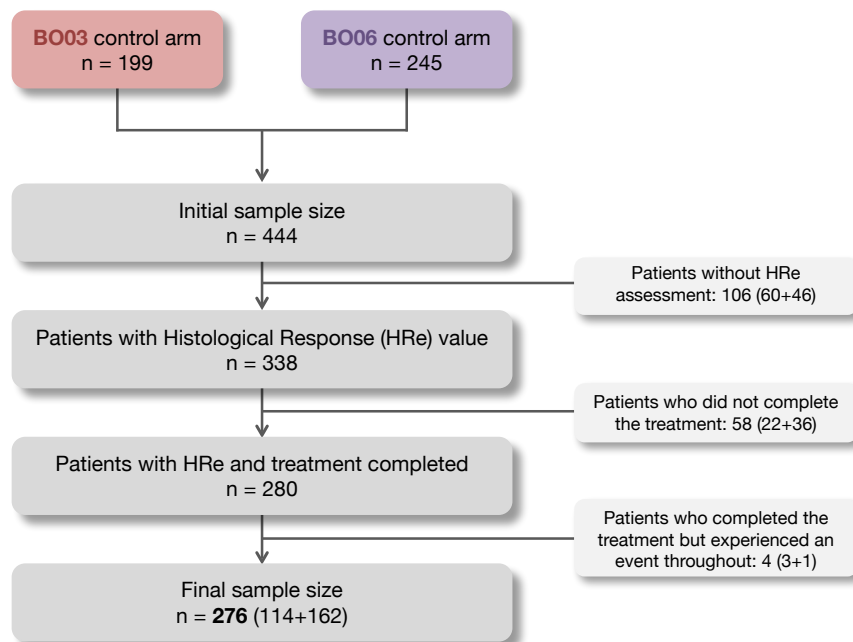


Fig. S1 Flowchart of cohort selection.

D Examining denominator models for IPTW

Different specifications of the subject-specific stabilized weights are investigated in order to check whether and which models best satisfied *positivity* and *no misspecification*. Multinomial logistic regression models are used for both numerators and denominators of SW_i in Manuscript Equation (9).

The numerators $\Pr(A_i|V_i)$, i.e., the probability that a subject i received exposure A_i given his/her histological response V_i , are modelled as:

$$\begin{aligned}\Pr(A_i = a|V_i) &= \frac{\exp(\alpha_{0a} + \alpha_{1a} \cdot \mathbf{GR}_i)}{1 + \sum_{j=1}^2 \exp(\alpha_{0j} + \alpha_{1j} \cdot \mathbf{GR}_i)} & a = 1, 2; \\ \Pr(A_i = 0|V_i) &= \frac{1}{1 + \sum_{j=1}^2 \exp(\alpha_{0j} + \alpha_{1j} \cdot \mathbf{GR}_i)},\end{aligned}$$

where variable $\mathbf{GR}$ is the dummy variable for good responders created from categorical variable $\mathbf{HRe}$ (*poor*; *good*).

The denominator $P(A_i|\mathbf{L}_i, V_i)$ is the probability that the subject received exposure A_i given relative confounders $\mathbf{L}_i$ and effect modifier V_i . Five different options are modelled as:

$$\begin{aligned}\Pr(A_i = a|\mathbf{L}_i, V_i) &= \frac{\exp(\eta_{ia})}{1 + \sum_{j=1}^2 \exp(\eta_{ij})} & a = 1, 2; \\ \Pr(A_i = 0|\mathbf{L}_i, V_i) &= \frac{1}{1 + \sum_{j=1}^2 \exp(\eta_{ij})}.\end{aligned}$$

by considering five different linear predictors η_{ia} in terms of confounding features.

- **IPTW 1:** categorical/binary confounding covariates and binary effect modifier are included as a main effect only and the MOTox scores are assumed linearly related to the log-odds:

$$\begin{aligned}\eta_{ia} &= \gamma_{0a} + \gamma_{1a} \cdot \mathbf{B006}_i + \gamma_{2a} \cdot \mathbf{adolescent}_i + \gamma_{3a} \cdot \mathbf{adult}_i + \gamma_{4a} \cdot \mathbf{male}_i + \\ &\quad + \gamma_{5a} \cdot \mathbf{MOTox}_{i,gen,pre} + \gamma_{6a} \cdot \mathbf{MOTox}_{i,rule,pre} + \\ &\quad + \gamma_{7a} \cdot \mathbf{MOTox}_{i,gen,post} + \gamma_{8a} \cdot \mathbf{MOTox}_{i,rule,post} + \gamma_{9a} \cdot \mathbf{GR}_i.\end{aligned}$$

- **IPTW 2:** same as in IPTW 1 + interaction terms for toxicity confounders linearly related to the log-odds:

$$\begin{aligned}\eta_{ia} &= \gamma_{0a} + \gamma_{1a} \cdot \mathbf{B006}_i + \gamma_{2a} \cdot \mathbf{adolescent}_i + \gamma_{3a} \cdot \mathbf{adult}_i + \gamma_{4a} \cdot \mathbf{male}_i + \\ &\quad + \gamma_{5a} \cdot \mathbf{MOTox}_{i,gen,pre} + \gamma_{6a} \cdot \mathbf{MOTox}_{i,rule,pre} + \\ &\quad + \gamma_{7a} \cdot \mathbf{MOTox}_{i,gen,post} + \gamma_{8a} \cdot \mathbf{MOTox}_{i,rule,post} + \gamma_{9a} \cdot \mathbf{GR}_i + \\ &\quad + \gamma_{10a} \cdot \mathbf{MOTox}_{i,gen,pre} \cdot \mathbf{MOTox}_{i,rule,pre} + \gamma_{11a} \cdot \mathbf{MOTox}_{i,gen,post} \cdot \mathbf{MOTox}_{i,rule,post}.\end{aligned}$$

- **IPTW 3:** same as in IPTW 1 + interaction terms between toxicities and trial linearly related to the log-odds:

$$\begin{aligned}\eta_{ia} = & \gamma_{0a} + \gamma_{1a} \cdot \text{B006}_i + \gamma_{2a} \cdot \text{adolescent}_i + \gamma_{3a} \cdot \text{adult}_i + \gamma_{4a} \cdot \text{male}_i + \\ & + \gamma_{5a} \cdot \text{MOTox}_{i,gen,pre} + \gamma_{6a} \cdot \text{MOTox}_{i,rule,pre} + \\ & + \gamma_{7a} \cdot \text{MOTox}_{i,gen,post} + \gamma_{8a} \cdot \text{MOTox}_{i,rule,post} + \gamma_{9a} \cdot \text{GR}_i + \\ & + \gamma_{10a} \cdot \text{trial}_i \cdot \text{MOTox}_{i,gen,pre} + \gamma_{11a} \cdot \text{trial}_i \cdot \text{MOTox}_{i,rule,pre} + \\ & + \gamma_{12a} \cdot \text{trial}_i \cdot \text{MOTox}_{i,gen,post} + \gamma_{13a} \cdot \text{trial}_i \cdot \text{MOTox}_{i,rule,post}.\end{aligned}$$

This choice is motivated by the statistically different distributions of the MOTox scores in BO03 and BO06 trials (see Manuscript Table 2).

- **IPTW 4:** categorical/binary confounding covariates and binary effect modifier are included as a main effect only; B-spline basis matrix for cubic polynomial splines with three internal knots are used to model the relationship between each continuous MOTox score and the log-odds:

$$\begin{aligned}\eta_{ia} = & \gamma_{0a} + \gamma_{1a} \cdot \text{B006}_i + \gamma_{2a} \cdot \text{adolescent}_i + \gamma_{3a} \cdot \text{adult}_i + \gamma_{4a} \cdot \text{male}_i + \\ & + \gamma_{5a}^T \mathbf{B}(\text{MOTox}_{i,gen,pre}) + \gamma_{6a}^T \mathbf{B}(\text{MOTox}_{i,rule,pre}) + \\ & + \gamma_{7a}^T \mathbf{B}(\text{MOTox}_{i,gen,post}) + \gamma_{8a}^T \mathbf{B}(\text{MOTox}_{i,rule,post}) + \gamma_{9a} \cdot \text{GR}_i.\end{aligned}$$

- **IPTW 5:** same as in IPTW 1 + interaction terms between toxicities and HRe linearly related to the log-odds:

$$\begin{aligned}\eta_{ia} = & \gamma_{0a} + \gamma_{1a} \cdot \text{B006}_i + \gamma_{2a} \cdot \text{adolescent}_i + \gamma_{3a} \cdot \text{adult}_i + \gamma_{4a} \cdot \text{male}_i + \\ & + \gamma_{5a} \cdot \text{MOTox}_{i,gen,pre} + \gamma_{6a} \cdot \text{MOTox}_{i,rule,pre} + \\ & + \gamma_{7a} \cdot \text{MOTox}_{i,gen,post} + \gamma_{8a} \cdot \text{MOTox}_{i,rule,post} + \gamma_{9a} \cdot \text{GR}_i + \\ & + \gamma_{10a} \cdot \text{GR}_i \cdot \text{MOTox}_{i,gen,pre} + \gamma_{11a} \cdot \text{GR}_i \cdot \text{MOTox}_{i,rule,pre} + \\ & + \gamma_{12a} \cdot \text{GR}_i \cdot \text{MOTox}_{i,gen,post} + \gamma_{13a} \cdot \text{GR}_i \cdot \text{MOTox}_{i,rule,post}.\end{aligned}$$

Variable B006 is the dummy variable created from categorical variable `trial` (*BO03*; *BO06*). Variables `adolescent` and `adult` are the dummy variables created from categorical variable `age` (*child*; *adolescent*; *adult*). Variable `male` is the dummy variable created from categorical variable `gender` (*female*; *male*).

References

- [1] US Department of Health and Human Services. Common Terminology Criteria for Adverse Events v3.0 (CTCAE); 2006. <https://www.eortc.be/services/doc/ctc/ctcae3.pdf>.
- [2] Cole SR, Hernán MA. Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology*. 2008;168(6):656–664. <https://doi.org/10.1093/aje/kwn164>.
- [3] Williamson T, Ravani P. Marginal structural models in clinical research: when and how to use them? *Nephrology Dialysis Transplantation*. 2017;32(suppl 2):ii84–ii90. <https://doi.org/10.1093/ndt/gfw341>.
- [4] Hernán M, Robins J. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC; 2020.
- [5] Cole SR, Frangakis CE. The Consistency Statement in Causal Inference. *Epidemiology*. 2009;20(3-5). <https://doi.org/10.1097/EDE.0b013e31818ef366>.
- [6] Karim ME, Gustafson P, Petkau J, Yinshan Zhao AS, Kingwell E, Evans C, et al. Marginal Structural Cox Models for Estimating the Association Between β -Interferon Exposure and Disease Progression in a Multiple Sclerosis Cohort. *American Journal of Epidemiology*. 2014;180(2):160–171. <https://doi.org/10.1093/aje/kwu125>.
- [7] Karim ME, Petkau J, Gustafson P, Tremlett H, Group TBS. On the application of statistical learning approaches to construct inverse probability weights in marginal structural Cox models: Hedging against weight-model misspecification. *Communications in Statistics - Simulation and Computation*. 2017;46(10):7668–7697. <https://doi.org/10.1080/03610918.2016.1248574>.
- [8] Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*. 2015;34(28):3661–3679. <https://doi.org/10.1002/sim.6607>.