SCIENTIFIC REPORTS
nature research

OPEN

Nearest Neighbour Propensity Score Matching and Bootstrapping for Estimating Binary Patient Response in Oncology: A Monte Carlo Simulation

Tine Geldof[1,2], Dusan Popovic[3], Nancy Van Damme[4], Isabelle Huys[1] & Walter Van Dyck[2*]

Nearest Neighbour (NN) propensity score (PS) matching methods are commonly used in pharmacoepidemiology to estimate treatment response using observational data. Unfortunately, there is limited evidence on the optimal approach for accurately estimating binary treatment response and, more so, to estimate its variance. Bootstrapping, although commonly used to accurately estimate variance, is rarely used together with PS matching. In this Monte Carlo simulation-based study, we examined the performance of bootstrapping used in conjunction with PS matching, as opposed to different NN matching techniques, on a simulated dataset exhibiting varying levels of real world complexity. Thus, an experimental design was set up that independently varied the proportion of patients treated, the proportion of outcomes censored and the amount of PS matches used. Simulation results were externally validated on a real observational dataset obtained from the Belgian Cancer Registry. We found all investigated PS methods to be stable and concordant, with k-NN matching to be optimally dealing with the censoring problem, typically present in chronic cancer-related datasets, whilst being the least computationally expensive. In contrast, bootstrapping used in conjunction with PS matching, being the most computationally expensive, only showed superior results in small patient populations with long-term largely unobserved treatment effects.

Estimating treatment effects in real-world clinical practice becomes increasingly important in domains like oncology, due to the high complexity of cancers and to recent developments of targeted medicines and immune-oncology drugs[1]. However, cancer registries, which are often used in epidemiological studies while monitoring changes in disease prevalence and investigating differences in incidence rates, only collect data at the time of diagnosis and, as such, do not contain any longitudinal information[2]. Therefore, estimating patient responses to a treatment, which could be based on tumour growth and/or toxic events, becomes very difficult. In those cases, patient-level response to a treatment can be based on overall survival (OS) as identified in clinical trials. Patient-level treatment effect can then be derived using the Propensity Score (PS) modelling technique proposed by Rosenbaum *et al.* (1983) for estimating average treatment effects[3,4], a common method used in pharmacoepidemiology. In this technique, the PS is the likelihood of a patient being assigned to a treatment (treatment status $Z = 1$ for treated vs. $Z = 0$ for control patients), conditional on observed covariates $\boldsymbol{X}$ prior to the application of the treatment. It forms the basis for matching treated and control patients who have a similar PS value, that is, are nearest neighbours[3,4]. Henceforward, changes in patients' individual Survival Gain (SG) (i.e. the difference between OS for treated and matched control patient), for each treated patient forms an indication for the response of the patient to the treatment.

[1]KU Leuven, Department of Pharmaceutical and Pharmacological Sciences, Research Centre for Pharmaceutical Care and Pharmaco-economics O&N II, 3001, Leuven, Belgium. [2]Vlerick Business School, Healthcare Management Centre, Reep 1, 9000, Ghent, Belgium. [3]KU Leuven, Department of Electrical Engineering, Stadius Centre for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, 3001, Leuven, Belgium. [4]Belgian Cancer Registry, Koningsstraat 215, 1210, Brussels, Belgium. *email: walter.vandyck@vlerick.com

Yet, caution should be taken when using Nearest Neighbour (NN) PS matching. First, one assumes that the OS of both treated and matched control patient(s) are exact and hence that the time of death is observed. In reality, patients get lost to follow up or are still alive during data collection, censoring the actual survival time and hence causing a problem when estimating the SG. This is especially true for chronic cancers like colorectal cancer featuring a median OS extending to a couple of years with treatment[5–8]. Secondly, there is still some controversy in the literature as to how variance and standard error of treatment effects should be calculated[9–11]. Providing a prediction error seems to be challenging as the variance will be high for one-by-k NN PS matching, on the one hand due to the small sample sizes when k is very low (approaching one) and on the other hand due to high patient heterogeneity when k is very high, that is, approaching the entire patient population.

Bootstrap-based methods, relying on random resampling, have been proposed in a few simulation studies and found to be effective for estimating variance of average patient treatment effects (ATE)[9–13]. Although rarely used in conjunction with PS matching, the technique is well-known for its ability to accurately measure variances of estimations in analytical difficult cases such as small datasets[14], and shows therefore promise for use in uncertainty surrounding (individual) treatment effects.

The main goal of this research is to assert which technique is the best for reducing the censoring problem and estimating the variance of the estimated SG in the context of predicting binary patient-level treatment response. Simulation-based approaches are well suited for this purpose, because these simulate true values which can then be compared to estimates generated from different PS matching approaches under varying conditions. Therefore, we examine the performance of three different state-of-the-art techniques on simulated datasets with different levels of heterogeneity. In particular, we compare k-NN, weighted k-NN and complex bootstrapping as described by Austin (2014) in series of Monte Carlo simulations. Finally, we further validate our findings on four case studies of metastatic colorectal patients treated with targeted medicines.

Counterintuitively, we found complex bootstrapping not to outperform k-NN or weighted k-NN methods when estimating survival gain variance in highly heterogeneous patient populations. However, from the aflibercept case featuring a small amount of patients assigned to the treatment with highly censored survival times, we did observe the bootstrap method to have favourable estimations. As expected, although computationally being the most expensive, bootstrapping outperformed other methods estimating variance in fairly small datasets.

Background: Propensity Score Matching and Boostrapping

This section presents the PS matching technique for estimating treatment effect and describes how different greedy NN algorithms[14] and the bootstrapping method[9–13] can be used to mitigate the censoring problem and to estimate uncertainty on individual treatment effects. Each of the matching algorithms uses matching with replacement, so that each control unit can be matched to multiple treated units.

PS NN matching and treatment effect. In NN PS matching, each treated patient is matched to one or more patients from the control group based on the closest PS $= Pr(Z_i = 1|X_i)$ value[3,4,15]. In principle, any regression technique can be used to develop the propensity model as long as it provides reasonable fit to the data. It is not necessary that the chosen technique produces calibrated probabilities as units are matched on a score[16]. Optimal selection of variables X_i is based on observed variables which affect the outcome of interest, because this is associated with better PS estimations[17].

Let T and C be the set of N^T treated and N^C control patients respectively and OS_i^T and OS_i^C the observed continuous outcomes of the treated and control units, respectively. Denote by $C(i)$ the set of N_i^C control units matched (using NN based on PS scores) to the treated unit $i \in T$. Define the weights $w_{ij} = \frac{1}{N_i^C}$ if $j \in C(i)$ and $w_{ij} = 0$ otherwise. We define the subject-specific treatment effect (STE) for the treated, derived from the ATE[13,18], as the estimation of the subject-specific survival gain SG_i^s:

$$\widehat{SG_i^s} = OS_i^T - \sum_{j \in C(i)} w_{ij} OS_j^C \tag{1}$$

With variance

$$\widehat{V_i^s} = var(\widehat{SG_i^s}) = var(OS_i^T) + \sum_{j \in C(i)} {w_{ij}}^2 var\left(OS_j^C\right)$$

Following oncology guidelines and depending on disease severity, patients can be labelled with 'response' whenever their SG is longer than a threshold of λ months.

One-by-one matching. The most common implementation of PS matching in practice is one-by-one matching, in which pairs of treated and control units are formed. Using one-by-one nearest neighbour PS matching ($N_i^C = 1$), one treated unit $i \in T$ is matched to one control unit $j \in C$. When the OS of treated, control or both are censored, the estimated SG^s will be highly uncertain (see Supplementary Material). Hence, for those matched pairs where censoring is problematic, the binary response-label based on the estimated SG^s becomes highly uncertain. For those cases, the SG cannot be assessed if any of the following conditions apply:

- when $j \in C(i)$ and $i \in T$ are censored
- when $j \in C(i)$ is censored and $\widehat{SG_i^s} \geq \lambda$
- when $i \in T$ censored and $\widehat{SG_i^s} < \lambda$

Denote $\kappa_{ij} = NA$ when i and j are censored or when j is censored and $OS_i^T - OS_j^C \geq \lambda$ and $\kappa_{ij} = 1$ otherwise. Define $\rho_i = 1$ when i is observed or when i is censored but $OS_i^T - \kappa_{ij} OS_j^C \geq \lambda$ and $\rho_i = NA$ otherwise. Formula (1) becomes:

$$\widehat{SG_i^s} = \rho_i OS_i^T - \kappa_{ij} OS_j^C \tag{2}$$

$$\widehat{V_i^s} = var(\widehat{SG_i^s}) = 0 \tag{3}$$

with OS_i^T assumed to be a constant. We can make a crude estimation of the SG_i^s variance by taking the variance of the entire control set C, i.e. $V_i^s = var(OS^C)$.

One-by-k matching. Using one-by-k nearest neighbour PS matching ($N_i^C = k = 50$), one treated unit $i \in T$ is matched to k nearest control units. Labelling for matched units subject to the censoring problem cannot be estimated if any of the following conditions are satisfied:

- when $\forall\ j \in C(i)$: j and $i \in T$ are censored
- when $\forall\ j \in C(i)$: j is censored and $\widehat{SG_i^s} \geq \lambda$
- when $i \in T$ censored and $\widehat{SG_i^s} < \lambda$

When none of these conditions are met the response label of treated unit $i \in T$ can be estimated. However, if $\exists\ j \in C(i)$: j is censored, j cannot contribute to the estimation of this label.

Define $\delta_{j,C(i)} = 1$ when j is observed or when $\forall\ j \in C(i)$: j is censored and $\delta_{j,C(i)} = 0$ otherwise, $\kappa_{C(i)} = NA$ when $\forall\ j \in C(i)$: i and j are censored, $\kappa_{C(i)} = 1$ when $\exists\ j \in C(i)$: j is observed or when $\forall\ j \in C(i)$: j is censored but $OS_i^T - \sum_{j \in C(i)} w_{ij} OS_j^C < \lambda$. Denote $\rho_i = 1$ when i is observed or when i is censored but $OS_i^T - \sum_{j \in C(i)} \kappa_{C(i)} \delta_j w_{ij} OS_j^C \geq \lambda$, given that the summation can be calculated under the conditions given by the definition of $\kappa_{C(i)}$, and $\rho_i = NA$ otherwise.

$$\widehat{SG_i^s} = \rho_i OS_i^T - \sum_{j \in C(i)} \kappa_{C(i)} \delta_{j,C(i)} w_{ij} OS_j^C \tag{4}$$

$$\widehat{V_i^s} = \sum_{j \in C(i)} {\kappa_{C(i)}}^2 {\delta_{j,C(i)}}^2 {w_{ij}}^2 var\left(OS_j^C\right)_{C(i)} \tag{5}$$

One-by-k weighted matching. In formula (2–5) all the k nearest neighbour units $j \in C$ included in the calculation (i.e. for which censoring is not a problem) have weights $w_{ij} = \frac{1}{N_i^C}$ and all others weight zero, meaning that all matched control units have equal contribution to the calculated mean. This can be generalized to weighted NN PS matching, where the contribution of $j \in C(i)$ to the mean depends on how similar the PSs are of subjects i and j, i.e. on the distance $d_{ij} = PS_i - PS_j$ (with minimal and maximal values equal to 0 and 1 respectively). Using an exponential distance function, the previous defined weights can be generalized to $w_{ij} = \frac{exp(-\alpha d_{ij})}{\sum_{j \in C(i)} exp(-\alpha d_{ij})}$ if $j \in C(i)$ and $w_{ij} = 0$ otherwise[19]. The value of α is set to 5 to ensure weights close to zero for maximal distances while having large enough differences in weights for small distances. Matched control units with equal PS as the treated unit contribute to the mean with a weight equal to one, while matched control units with distance approaching one contribute only with a weight approaching zero.

Matching through bootstrapping. Using the complex bootstrap method as described by Austin (2014), b bootstrap samples are drawn from the original control group with sample size equal to the control group[9]. In each of the bootstrap samples, the PS model is estimated, and one-by-one PS matching is performed for creating matched pairs, forming the set $C(i)$ of control units ($N_i^C = b$) matched to the treated unit $i \in T$. In this k-NN bootstrapping method, one treated patient can be matched multiple times to the same control patient, i.e. $j \in C(i)$ can occur multiple times in $C(i)$, lowering the heterogeneity of the matched sets. The estimation of the subject-specific gain in survival $\widehat{SG_i^s}$ and its variance $\widehat{V_i^s}$ can be calculated as given by Eqs. (4) and (5) in one-by-k matching[14].

Material and Methods

We used simulated datasets with three levels of patient heterogeneity to examine the performance of the different matching techniques over a series of Monte Carlo simulations. There, performance was evaluated based on their ability to estimate the individual SG under these three scenarios. In this section, we describe the design of the datasets and the Monte Carlo simulations. The results were externally validated by examination of case studies for treated metastatic colorectal patients.

Simulated data generation. Data was simulated in R following the data-generating process described by Austin (2014), generating 1000 patients with 10 baseline covariates $X_1 - X_{10}$, of which seven affecting treatment selection ($X_1 - X_7$) and OS outcome ($X_3 - X_{10}$)[9]. Very weak, weak, moderate, strong and very strong effects of the covariates on treatment selection and OS outcome is introduced by the regression coefficients $\alpha_{VW} = log\ log\ (1, 25)$, $\alpha_W = log\ log(1, 5)$, $\alpha_M = log\ log(2)$, $\alpha_S = log\ log(4)$ and $\alpha_{VS} = log(8)$. PSs $p_i = Pr(Z_i = 1|X_i)$ were determined using logistic regression, following:

Product	PS NN techn.	ATE ($\overline{SG^s}$, months)	$\hat{V}$	$\overline{V^s}$	$var(V^s)$	C
Bevacizumab *(2784 patients)*	1:5	8.00	567.51	310.21	2.45e+5	4.6%
	weighted 1:5	7.97	566.04	312.30	2.51e+5	4.6%
	5 bootstrap	*7.39*	*600.79*	*282.97*	*2.35e+5*	*5.2%*
Cetuximab *(845 patients)*	*1:5*	*7.38*	*477.96*	*261.86*	*1.54e+5*	*1.8%*
	weighted 1:5	7.31	482.54	264.88	1.56e+5	1.8%
	5 bootstrap	6.33	534.84	293.23	2.18e+5	2.5%
Panitumumab *(308 patients)*	*1:5*	*11.25*	*359.27*	*282.25*	*2.48e+5*	*1.6%*
	weighted 1:5	10.96	329.42	309.92	2.60e+5	1.6%
	5 bootstrap	9.71	453.27	276.24	2.75e+5	1.6%
Aflibercept *(31 patients)*	1:5	2.28	323.71	424.96	3.16e+5	23%
	weighted 1:5	2.56	320.35	428.04	3.32e+5	23%
	5 bootstrap	*3.18*	*340.35*	*337.58*	*2.50e+5*	*19%*

Table 1. Outcomes for each treatment resulting from the different NN PS matching techniques (k = 5).

$$
\begin{aligned}
logit(p_i)_L &= \alpha_{0,treat,L} + \alpha_{VW}x_1 + \alpha_W x_2 + \alpha_{VW}x_3 + \alpha_W x_4 + \alpha_{VW}x_5 + \alpha_W x_6 + \alpha_M x_7,\\
logit(p_i)_M &= \alpha_{0,treat,M} + \alpha_W x_1 + \alpha_M x_2 + \alpha_S x_3 + \alpha_W x_4 + \alpha_M x_5 + \alpha_S x_6 + \alpha_{VS}x_7,\\
logit(p_i)_H &= \alpha_{0,treat,H} + \alpha_M x_1 + \alpha_S x_2 + \alpha_M x_3 + \alpha_S x_4 + \alpha_M x_5 + \alpha_S x_6 + \alpha_{VS}x_7
\end{aligned}
$$

for low, medium and high heterogeneity respectively. Treatment status was generated from a Bernoulli distribution on the subject-specific PS p_i: $Z_i \sim Be(p_i)$, through which the intercept $\alpha_{0,treat}$ indirectly affects the proportion of patients treated in the simulation. The OS outcome was generated as described by Bender (2005) and Austin (2014)[9,12], that is, based on the linear predictor

$$
\begin{aligned}
LP_L &= \beta_{treat,L}Z + \alpha_W x_4 + \alpha_{VW}x_5 + \alpha_W x_6 + \alpha_M x_7 + \alpha_{VW}x_8 + \alpha_W x_9 + \alpha_{VW}x_{10},\\
LP_M &= \beta_{treat,M}Z + \alpha_W x_4 + \alpha_M x_5 + \alpha_S x_6 + \alpha_{VS}x_7 + \alpha_W x_8 + \alpha_M x_9 + \alpha_S x_{10},\\
LP_H &= \beta_{treat,H}Z + \alpha_S x_4 + \alpha_M x_5 + \alpha_S x_6 + \alpha_{VS}x_7 + \alpha_M x_8 + \alpha_S x_9 + \alpha_M x_{10}
\end{aligned}
$$

for low, medium and high heterogeneity respectively, using the formula $OS = \left(\frac{-log(u)}{\lambda e^{LP}}\right)^{\frac{1}{2}}$, with u a random number from the uniform distribution and λ equal to 0.00002. The conditional hazard ratio $exp(\beta_{treat})$ was fixed to 0.8. The true SG for the treated SG_i was generated from the OS outcome as produced by the linear predictors for $Z_i = 1$ and 0: $SG_i|(Z_i = 1) = OS_i|(Z_i = 1) - OS_i|(Z_i = 0)$. From this, the corresponding average true SG, i.e. the "*true ATE*" $\overline{SG}$, and the variance of the true SG, i.e. the "*true variance*" V, is calculated. The censoring status of the subjects' survival was drawn from a Binominal distribution given the probability of being censored p^*: $c_i \sim Binom(p^*)$.

Case study data. Patients were collected from the Belgian Cancer Registry (BCR), a population based cancer registry. We used ICD-10 codes (C18 up to and including C20) to select 10426 metastatic colorectal patients (stadium IV carcinoma) diagnosed between 2006 and 2014 with vital status information updated until July 1, 2017 (Table 1). Patients were classified in five groups according to their targeted treatment assignment: 2784, 845, 308 and 31 patients received bevacizumab, cetuximab, panitumumab, and aflibercept respectively. 6458 patients were not treated with the targeted medicine and were classified as the control group (irrespective of radiotherapeutic and/or chemotherapeutic treatments). Of these five groups, 26% (731), 15% (127), 11% (35), 52% (16) and 15% (965) had censored survival, respectively.

OS, the RCT's primary endpoint, was used as the main indication of treatment effect. Selected variables were taken from the full standard set of variables nationally collected by the BCR and Inter Mutualistic Agency, which were further limited for relevance by BCR oncologists.

The data set consisted of (a) the patient's OS and censoring status; (b) (historical) treatment paths i.e. radiotherapeutic and/or chemotherapeutic treatment and treatment with the four targeted medicines; and (c) patient and tumour characteristics, i.e. age, sex, tumour differentiation grade, topography, tumour location (left/right), total amount of tumours, WHO performance score at diagnosis and TNM classification. Multiple imputation was used for handling missing data for PS-relevant variables assuming data was missing completely at random[20–23].

Analysis on simulated data. Using the data-generating process described above, three types of heterogeneity were simulated by using the regression coefficients denoting very weak to very strong impact on treatment selection and survival, which were iterated 1000 times. For each of these heterogeneity types (low, medium and high), three factors were varied: the proportion of patients treated (given no censoring), the proportion of outcomes censored (given 20% of patient treated) and the number of nearest neighbours used in matching (given 20% of patient treated and 20% of outcomes censored). (See Supplementary Materials for more information). For all these scenarios and datasets, the PS is estimated using a logistic regression model[3,4], with selected observed variables being those affecting the survival time ($X_3 - X_{10}$)[16].The three PS NN matching techniques (k-NN,

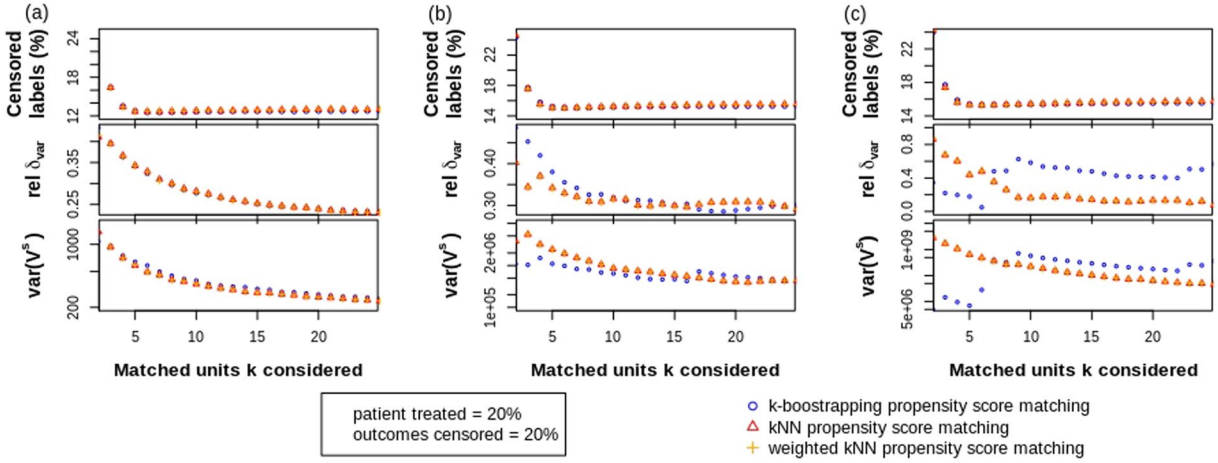

Figure 1. Monte Carlo simulation results in function of the number of NN matched (given 20% of patient treated and 20% of outcomes censored) for (**a**) low heterogeneity; **(b)** medium heterogeneity and (**c**) high heterogeneity.

weighted k-NN and complex bootstrapping described above) are performed to estimate the STE, i.e. the estimated subject-specific gain in survival $\widehat{SG_i^s}$, and STE variance $\widehat{V_i^s}$ for each treated unit $i \in T$, given by formula (4 and 5). These are then investigated for the different PS NN methods by calculating their means over all units, $\overline{SG^s}$ the mean STE) and $\overline{V^s}$ (the mean STE variance), and comparing the latter with the simulated "*true variance*" V. We propose the PS NN method with the smallest relative difference $\frac{\delta_{var}}{V} = (V - \overline{V^s}))/V$ the best estimator of the true variance. Lastly, the variance of $\widehat{V^s}$, i.e. $var(V^s)$, is compared between the PS NN methods, as well as the proportion of patients subject to the label-censoring problem as defined by the rules of formula (4)-(5). These analyses were carried out across the 1000 iterations of the Monte Carlo simulation conducted in R. Therefore, results of these analysis are reported as averaged values over the iterations.

Simulation Results

The following section describes the results of the label-censoring problem and the variance estimations for the three PS NN methods on the simulated datasets with low, medium and high heterogeneity.

Impact of number of units matched. The relative difference δ_{var} between the true variance V and the mean estimated STE variance $\overline{V^s}$ together with the resulting variance of the STE variance $var(V^s)$ and the proportion of labels censored are reported in Fig. 1 for varying amount of NN units k considered during matching. The three panels show the different levels of heterogeneity.

As expected, the amount of predicted labels that are censored decrease with increasing amount of matched units k considered during the three PS NN matching methods for all heterogeneity sets, this at a similar pace until $k = 5$. Hence all methods perform equally well for solving the label-censoring problem, regardless of heterogeneity.

Similarly, no difference is found between the methods for estimating variance in low heterogeneous groups unless for computational complexity using bootstrapping. However, for increasing heterogeneity, we observe the bootstrap method to have less accurate predictions of variances, showing higher relative differences δ_{var} although lower variances of $\widehat{V^s}$ for small k. Hence, for high heterogeneity the bootstrap method would be inferior to the k-NN matching methods based on both accuracy and computational complexity, while for low heterogeneity the bootstrap method is inferior on computational complexity alone.

Impact of proportion of outcomes censored. The relative difference δ_{var} between the true variance V and the mean estimated STE variance $\overline{V^s}$ and the resulting variance of the STE variance $var(V^s)$ are reported in Fig. 2 for varying amount of outcomes OS censored. The three panels show the different levels of heterogeneity.

We see the relative difference δ_{var} to be quite unaffected by the proportion of OS outcomes censored for all heterogeneity sets. Hence, the estimation of $\overline{V^s}$ remains constant, even though increased outcomes censored means less units $j \in C(i)$ contribute to the estimation of V_i^s for every unit $i \in T$. However, we can verify that this has an effect on the accuracy of this estimation, because the variances of $\widehat{V^s}$ have an increasing trend for low heterogeneity. This trend disappears for higher heterogeneity because of both δ_{var} and especially $var(V^s)$ fluctuate. For all levels of heterogeneity, the bootstrap method would be inferior to the k-NN matching method based on computational complexity.

Impact of proportion of patients treated. The relative difference δ_{var} between the true variance V and the mean estimated STE variance $\overline{V^s}$ and the resulting variance of the STE variance $var(V^s)$ are reported in Fig. 3 for varying amounts of patients treated in the population. The three panels show the different levels of heterogeneity.

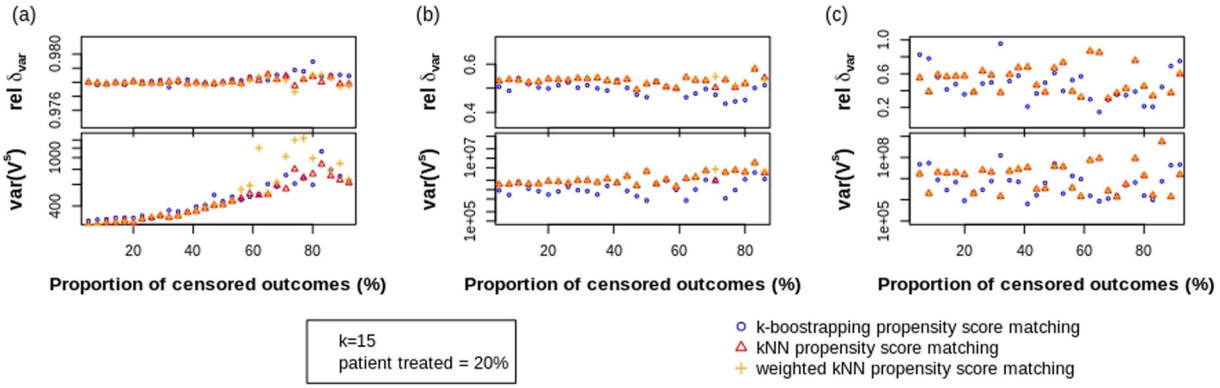

Figure 2. Simulation results in function of the proportion of OS outcomes censored (given k = 15 NN used in matching and 20% of patients treated) for (**a**) low heterogeneity; (**b**) medium heterogeneity and (**c**) high heterogeneity.

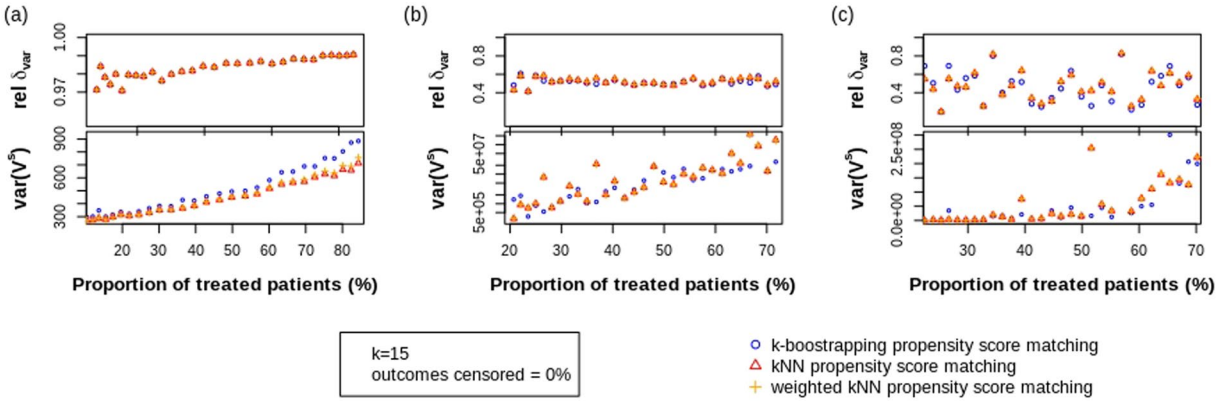

Figure 3. Simulation results in function of the proportion of patients treated (given k = 15 NN used in matching and 0% of outcomes censored) for (**a**) low heterogeneity; (**b**) medium heterogeneity and (**c**) high heterogeneity. The error bars have been omitted for clarity.

As expected, δ_{var} increases and becomes more uncertain ($\widehat{V^s}$ variance increases) for increasing proportion of treated patients in all heterogeneity sets, as the control group C (the pool to which treated units are matched) becomes smaller. No difference is found between the different methods, except for a slightly higher $\widehat{V^s}$ variance for low heterogeneity. However, as for each proportion of treated units a different linear predictor was simulated, affecting the OS outcome of the dataset, we see δ_{var} and $\widehat{V^s}$ variance fluctuates, especially for high heterogeneity. Therefore, results for δ_{var} for the different PS NN methods are inconclusive.

Case Study

In this section, the performance of the three different PS NN techniques are examined on a case study of metastatic colorectal cancer patients treated with bevacizumab, cetuximab, panitumumab or aflibercept as a targeted medicine. Numerical results of the one-by-one, one-by-25 (weighted) and 25-bootstrap PS NN matching techniques are depicted in Fig. 4 and Table 1.

The results show that the three methods are stable and concordant. Only for the aflibercept case, with a small treated population (31) and high amount of survival censoring (52% or 16 out of 31), we observe a difference between the k-NN techniques and the bootstrap method. Specifically, the censoring problem reduces dramatically with increasing k with lower estimated $\overline{V^s}$ and $\widehat{V^s}$ variance for the bootstrap method as opposed to the k-NN techniques.

Discussion

Bootstrapping, a method commonly used to accurately estimate variance, is rarely used together with PS matching. In this Monte Carlo simulation-based study, we examined the performance of the complex bootstrap method, as described by Austin (2014), to estimate binary treatment response and variance in the domain of oncology. Specifically, the subject-specific survival gain (that is, the individual treatment effect) and its variance together with its ability to mitigate the problem of label-censoring, obtained from the individual treatment effect as a binary treatment response label, were the main factors under investigation. The Monte Carlo study was based

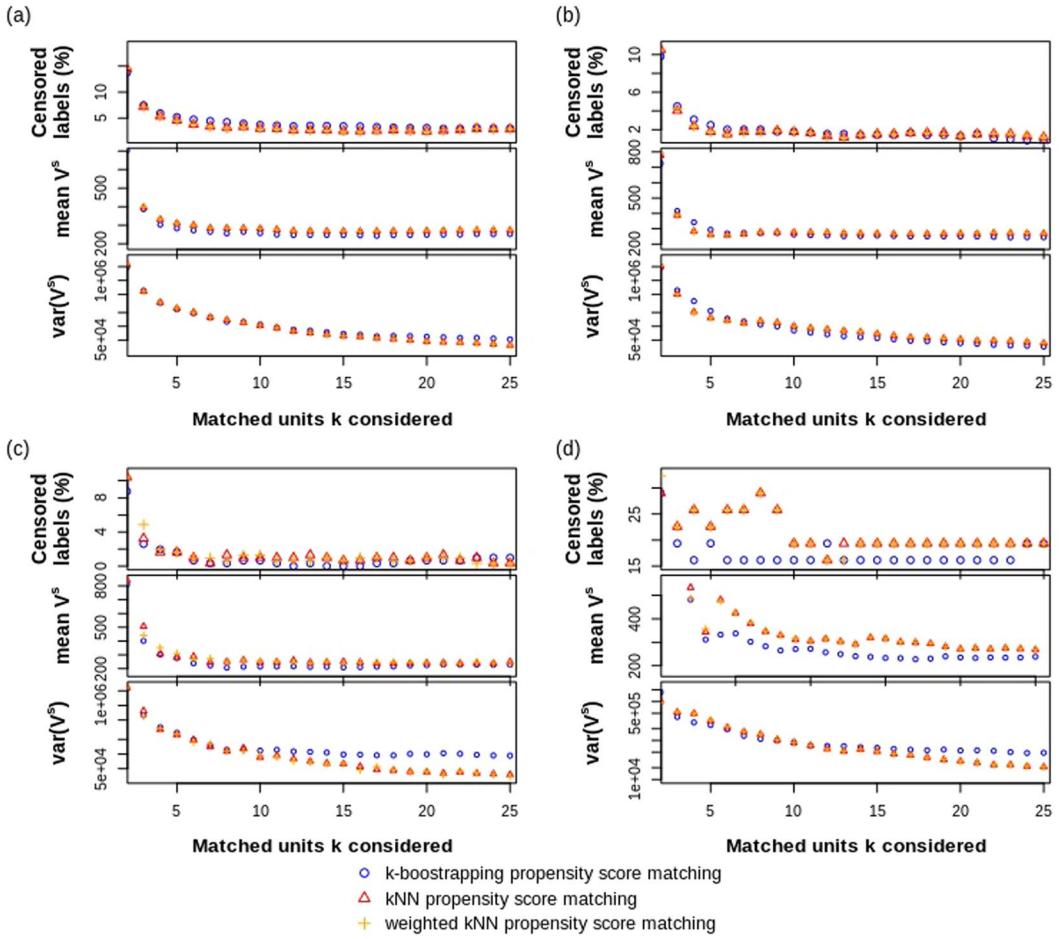

Figure 4. Case study comparison of (weighted) k-NN and bootstrap matching for (**a**) bevacizumab (2784 treated patients), (**b**) cetuximab (845 treated patients), (**c**) panitumumab (308 treated patients), and (**d**) aflibercept (31 treated patients). Shown are the number of SG outcomes being censored (top), the mean subject-specific treatment-effect (STE) variance (middle) and the variance of the STE variance (bottom).

on simulated datasets containing 1000 patients with varying levels of heterogeneity found in real world patient populations. Counterintuitively, we found that the estimation of survival gain variance in patient populations with a high patient heterogeneity does not benefit from using the complex bootstrapping method instead of (weighted) k-NN. Indeed, we expected the relevant matches to be small for increasing k and increasing heterogeneity, implying that k-NN PS matching would contain a large set of irrelevant matches for large k, as opposed to the complex bootstrapping method, which always matches a treated patient to the one closest control patient. As a consequence, while also taking into account the computational complexity we found the bootstrap method to not show to favourable results even for high heterogeneous patient populations. Additionally, no major differences were found between the k-NN and weighted k-NN method, because the resulting weights were approximately equal to one in most cases of the simulated data. While it can be argued that this behaviour would change if one chooses a value of the parameter of the exponential distance function that is better suited to the data at hand, we note that this parameter cannot be tuned in practice because, as opposed to that in simulation study, the real variance is unknown before estimation.

Applying these methods to four colorectal cancer treatments with varying amount of patients treated and unobserved outcomes, we found all three PS methods to be stable and concordant. From the analysis, we can conclude that the computationally cheapest method, being k-NN PS matching, should be used in most of the cases. However, for the aflibercept case, where a small amount of patients are assigned to the treatment while the majority of survival times are censored, we did observe the bootstrap method to have favourable estimations. This result was expected because bootstrapping is a statistical method often used for estimating variance in fairly small datasets[13].

Note that some concerns may arise when using bootstrapping in conjunction with PS matching in observational studies. First, one specified PS model was used for each resampled control group in the analysis of this study, which may be inappropriate in high heterogeneous patient populations. However, identifying the best fitted model for each bootstrapped sample would be highly unpractical. Second, for comparison reasons, a low number

of bootstrap samples was used equal to the number of matched control units in k-NN. Although this amount should provide a decent estimate[13], a higher amount would be recommended in observational studies[13,23,24].

Overall, given these findings, we suggest that the complex bootstrap method, while being computationally more expensive, should not be used for estimating subject-specific survival gain in large cohorts of treated and non-treated patients. However, this most computationally expensive method might show to be necessary when considering small patient populations with long-term and largely unobserved treatment effects.

Data availability

The case study data that support the findings of this study are available from the Belgian Cancer Registry but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Belgian Cancer Registry.

Received: 19 February 2019; Accepted: 9 December 2019;
Published online: 22 January 2020

References

1. Burock, S., Meunier, F. & Lacombe, D. How can innovative forms of clinical research contribute to deliver affordable cancer care in an evolving health care environment? *Eur. J. Cancer.* **49**, 2777–2783 (2013).
2. Parkin, D. The role of cancer registries in cancer control. *Int. J. Clin. Oncol.* **13**(2), 102–111 (2008).
3. Rosenbaum, P. R. & Rubin, D. B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika.* **70**, 41–55 (1983).
4. Rosenbaum, P. R. & Rubin, D. B. Reducing bias in observational studies using sub classification on the propensity score. *J. Am. Stat. Assoc.* **79**, 516–24 (1984).
5. Van Cutsem, E. *et al*. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *N. Engl. J. Med.* **360**(14), 1408–17 (2009).
6. Douillard, J. Y. *et al*. Randomized, phase III trial of panitumumab with infusional fluorouracil, leucovorin, and oxaliplatin (FOLFOX4) versus FOLFOX4 alone as first-line treatment in patients with previously untreated metastatic colorectal cancer: the PRIME study. *J. Clin. Oncol.* **28**(31), 4697–705 (2010).
7. Van Cutsem, E. *et al*. Addition of aflibercept to fluorouracil, leucovorin, and irinotecan improves survival in a phase III randomized trial in patients with metastatic colorectal cancer previously treated with an oxaliplatin-based regimen. *J. Clin. Oncol.* **30**(28), 3499–506 (2012).
8. Tabernero, J. *et al*. Aflibercept versus placebo in combination with fluorouracil, leucovorin and irinotecan in the treatment of previously treated metastatic colorectal cancer: prespecified subgroup analyses from the VELOUR trial. *Eur. J. Cancer.* **50**(2), 320–31 (2014).
9. Austin, P. C. & Small, D. S. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Stat. Med.* **33**, 4306–4319 (2014).
10. Hill, J. & Reiter, J. P. Interval estimation for treatment effects using propensity score matching. *Stat. Med.* **25**, 2230–2256 (2006).
11. Colson, K. *et al*. Optimizing matching and analysis combinations for estimating causal effects. *Sci. Rep.* **6**, 23222 (2016).
12. Austin, P. C. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat. Med.* **35**, 5642–5655 (2016).
13. Samuels, L. R. & Robert, A. G. Bagged one-to-one matching for efficient and robust treatment effect estimation. *Stat. Med.* **37**, 4353–4373 (2018).
14. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman & Hall: New York (1993).
15. Rubin, D. B. Matching to remove bias in observational studies. *Biometrics* **29**, 159–183 (1973).
16. Stuart, E. A. Matching methods for causal inference: a review and a look forward. *Stat Sci* **24**, 1–21 (2010).
17. Austin, P. C., Grootendorst, P. & Anderson, G. M. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine* **26**, 734–753 (2007).
18. Becker, S. O. & Ichino, Z. A. Estimation of average treatment effects based on propensity scores. *Stata Journal, 4th Quarter.* **2**, 358–3770 (2002).
19. Dudani, S. A. The distance-weighted k-nearest neighbor rule. *IEEE Trans. Syst. Man Cybern. Syst.* **6**, 325–327 (1976).
20. Olinsky, A., Chen, S. & Harlow, L. The comparative efficacy of imputation methods for missing data in structural equation modeling. *Eur J Oper Res.* **151**, 53–79 (2003).
21. Steyerberg, E. W. & van Veen, M. Imputation is beneficial for handling missing data in predictive models. *J Clin Epidemiol.* **60**(9), 979 (2007).
22. Sterne, J. A. C. *et al*. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ.* **338**, b2393 (2009).
23. Barakat, M. S. The effect of imputing missing clinical attribute values on training lung cancer survival prediction model performance. *Health Inf Sci Syst* **5**, 16 (2017).
24. Wilcox, R. R. Fundamentals of modern statistical methods: Substantially improving power and accuracy. 155 (Springer 2010).

Acknowledgements

This study has been supported by the Vlerick Business School Academic Research Fund and benefitted from the contribution of Yves Moreau, Department of Electrical Engineering (ESAT) STADIUS Centre for Dynamical Systems, Signal Processing and Data Analytics Department, University of Leuven, Belgium. We would also like to thank the Belgian Cancer Registry for providing us with unique access to historical observational data and their research assistance. Financial support for this study was provided entirely by a grant from the Vlerick Business School. The funding agreement ensured the authors' independence in designing the study, interpreting the data and publishing the report.

Author contributions

T.G., D.P. and W.V.D. participated in the design of the research, interpretation of the results and editing of the manuscript. T.G., I.H. and N.V.D. performed the national Cancer Registry data retrieval and validation following FAIR principles. T.G. and W.V.D. performed the research and analysis and made substantial contributions to the writing of the manuscript. All authors read, amended and approved the final version for publication.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-57799-w.

Correspondence and requests for materials should be addressed to W.V.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.