RESEARCH

Estimating the effect of adjuvant chemo-therapy for colon-cancer using registry data: a method comparison and validation

Lingjie Shen[1*], Erik Visser[2], Hans de Wilt[4], Henk Verheul[5], Felice van Erning[3], Gijs Geleijnse[3] and Maurits Kaptein[2]

[*]Correspondence: L.SHEN@uvt.nl
[1]Department of Methodology and Statistics, Tilburg University, Warandelaan 2, 5000 LE Tilburg, The Netherlands
Full list of author information is available at the end of the article

Abstract

Background: Although randomized controlled trials (RCT) are the gold standard to estimate treatment effects, they are often criticized in terms of generalizability. Observational data might alleviate this problem by being readily available in large quantities. However, observational data are potentially confounded. In this methodological study we use a large-scale RCT as the gold standard to examine the performance of various statistical methods to control for potential confounding in observational data.

Methods: In this paper we compare three types of methods that allow researchers to correct for such potential confounding: direct methods, inverse probability weighting (IPW) methods and doubly robust (DR) methods. We uniquely compare estimates obtained from the population-wide Netherlands Cancer Registry (NCR) on colon cancer ($n = 52621$) with estimates obtained from a large-scale RCT. As the RCT differs from the observational data both in its sampling mechanism and in its treatment assignment mechanism, we first resample the NCR data to reflect the distribution in RCT data. Next, we correct for potential confounding using three alternative types of methods and consequentially evaluate their estimates to those obtained in the RCT.

Results: We find that while all estimators qualitatively approximate to findings in the RCT, methods that can flexibly model the response (i.e., direct estimation and DR estimation) performed consistently superior to the inverse propensity score method. Subgroup analysis indicates that relatively simple models allow us to properly estimate the treatment effect. However, these simple models do not properly identify heterogeneous treatment effects in stage II colon cancer. Careful sensitivity analysis using more flexible models demonstrates both the uncertainty and the potential heterogeneous treatment effect in stage II cancer and provides robust estimation of treatment effect in stage III cancer.

Conclusions: Our results suggest that both the direct method and the DR method, when executed with care, can be used to reliably estimate treatment effects based on registry data. This methodological validation opens the door to more extensive use of registry data for the estimation of (individual) treatment effects. Additionally, our identification of potentially meaningful subgroups of stage II colon cancer patients who, based on our analysis seem to benefit from chemotherapy, should be further explored.

Keywords: average treatment effect; methods comparison; confounding; guideline

1 Background

Colorectal cancer is the third most commonly diagnosed cancer and the second leading cause of cancer death worldwide in 2018 [1]. The majority of patients are curatively treated with surgical removal of the tumor, surrounding tissues, and lymph nodes leading to excellent long term survival [2]. For stage III colon cancer, Randomized Control Trials (RCTs) [3–5] have shown an improved overall survival when surgery is followed by adjuvant chemotherapy compared to surgery only. However, results from RCTs were not supportive of prescribing adjuvant chemotherapy to stage II patients [6] due to lack of convincing evidence. For example, the meta-analysis by Erlichman and Charles (1999) [7] showed in a pooled analysis a hazard ratio (HR) of 0.83 (90% CI 0.72, 1.07) for disease-free survival and a HR of 0.86 (90 % CI 0.68, 1.07) for overall survival for fluoropyrimidine monotherapy compared to observation. The same finding was also reported [8]: while both RCT and real world data suggest a clinically relevant benefit of adjuvant chemotherapy for stage II colon cancer, the estimates do not reach statistical significance. A borderline significantly improved overall survival was found in stage II cancer in QUASAR trial (Relative Risk 0.82 95% CI 0.7-0.95) and consequently the authors suggest adjuvant chemotherapy for all stage II patients. Without critical pathological and clinical characteristics to identify the heterogeneous treatment effect in subgroups (e.g., emergency resection, tumor size, grade, number of positive lymph nodes), conclusions from the QUASAR trial may however not be applicable to all stage II patients.

These mixed and inconsistent results from current trials may be partly due to the heterogeneity between populations resulting from selection rules of eligible patients. The strict selection rules in RCTs make baseline populations in trials differ from daily clinical practice, which can threaten the external validity of studies. Registry data, without strict patient selection, may truly reveal the clinical relevance and may provide additional evidence regarding the effect of the adjuvant chemotherapy on a population outside the context of trials. Consequently, researchers have recently tuned their interest to such registry data to estimate treatment effects because it covers a large, heterogeneous and country-wide population [8–11].

Estimating the effect of adjuvant chemotherapy using registry data, however, is challenging. First, variations in the type, the dose, and the length of the chemotherapy commonly occur across individuals, which make comparison of treatment effect across studies difficult. For instance, although capecitabine and oxaliplatin (CAPOX) and fluorouracil, leucovorin, and oxaliplatin (FOLFOX) are regarded as current standard treatments regarding their improved survival compared to fluorouracil and leucovorin (FU/LV) [4], the detailed information of the chemotherapy regimen for each individual is not always properly recorded. This hinders the estimation of the long-term effect of CAPOX and FOLFOX using existing registry data. To fully utilize the registry data regardless of the limitations, researchers uses a binary variable to indicate the chemotherapy (FU/LV and FOLFOX) and control, and consequentially compares treatment effect using registry data with the estimation from trials (FU/LV vs. control) [8]; an approach we also adopt in the current study.

Second, unlike RCT data, in which the observed and unobserved confounding variables are balanced by random treatment assignment, the observational registry

data runs the risk of being confounded. More precisely, in observational data, the treatment assignment mechanism is not exactly known, and thus confounding cannot be ruled out [12]. Understanding the performance of statistical methods to correct for such potential confounding is an extremely important methodological question: if we are able to properly control for confounding in registry data we open up a treasure throve of new data to be used to estimate (individual) treatment effects.

Many statistical methods to control for potential confounding have been suggested over the past few decades, most often under the assumptions of positivity and unconfoundedness which we describe in detail in section 3. These methods can roughly be classified into one of three classes [13]:

1 *Direct methods:* Direct methods aim to directly model the outcome as a function of the treatment and potential covariates. Some methods consider the treatment itself as a covariate and fit the outcome directly, while others fit the separate outcome for the treatment and control groups [14]. Using multivariate regression to directly model the associations between the outcome and the treatment while adjusting the covariates is the most commonly applied implementation of the direct method [15]. Other, more modern, implementations use more flexible models like Bayesian additive regression trees (BART) [16–18] or Causal Forests [19].

2 *Inverse probability weighting (IPW) methods:* IPW methods control for potential confounding by estimating, and controlling for, the conditional probability that an individual unit receives the treatment (i.e., the propensity score, PS) [12]. IPW methods estimate the treatment effect by weighting each sample with the inverse PS [20], thereby generating the pseudo-population whose covariates are independent of the treatment assignment (as would be the case in an RCT).

3 *Doubly robust (DR) methods:* DR methods combine a model for the treatment assignment mechanism and a model for the outcome, which can practically yield accurate estimation when one of the two models is good (but not necessarily consistent) [21, 22].

Note that for any of the methods above researchers can make a variety of modeling choices. Historically, parametric models have been commonly used. For instance, in the majority of published studies about the treatment effect estimation in the field of colorectal oncology [8, 23, 24] the direct method, using standard logistic regression, is employed. Although these parametric models are efficient and well-understood, they impose strong parametric assumptions about the relationship between potential confounders and outcomes. Recently, more flexible semi-parametric and non-parametric models have demonstrated good performance [14]. In particular, BART seems well suited to uncover the true treatment effect when the treatment assignment is confounded [14, 25–27].

When comparing direct estimation with IPW and DR methods in simulation studies, an improved performance of direct estimation over IPW methods is found [14], while in other study DR methods outperform both direct estimation and IPW methods [28]. Comparisons between these different classes of methods are however hard to carry out in non-simulated scenarios: often when observational data (registry data) is available there is no ground truth of the estimate of interest available

(i.e., there is no RCT data available). The unavailability of evaluations of different methods to control for confounding on real-world data hinders the adoption of these methods in general.

The case of adjuvant chemotherapy for colon cancer is different, however: here we uniquely have access to a large RCT [29] (QUASAR trial) as an experimental benchmark and we have access to the observational data collected by the NCR. This provides us with the opportunity to evaluate the performance of different (classes of) methods to control for confounding on real-world registry data. To our best knowledge, the QUASAR trial is the best possible gold standard to compare the effect of chemotherapy verses control as it has large samples, long follow up years, and good completion of six months of chemotherapy.

To conclude, observational registry data and RCT data potentially differ in two respects which requires us to be aware of when estimating and comparing the treatment effect [30]:

1 *the sampling mechanisms*, i.e., the mechanism by which subjects are included into studies, can differ substantially. While the absence of strict inclusion criteria is an argument in favor of the use of observational data, the sampling process will ultimately affect the distributions of the covariates and subsequently will affect the estimated average treatment effect. The sampling mechanism may also lead to differences in effect estimation between the RCT and the observational study.
2 *the treatment assignment mechanism* differs: while in the RCT the assignment mechanism is known – and often uniformly random – for observational studies, the assignment mechanism is unknown and confounding is potentially present.

In this study, of which we provide an overview in Figure 1, we attempt to control for both sources of divergence and aim to explore the effects of three classes of correction methods of controlling for the second source of divergence. This study thus helps us understand how these different methods perform on real-world registry data, potentially opening up their use to inform the current treatment guidelines.

In the next section, we first describe the two data sources used in this study. In section 3 we describe the different methods for controlling for confounding that we examine in this paper: direct methods, IPW methods, and DR methods. In section 4 we explain our approach to ensure that, next to controlling for confounding, the sampling mechanism playing in both data sources are comparable: effectively, we estimate the distribution of the main covariates in the QUASAR trial as $\mathcal{D}_{QUASAR'}$, and subsequently subsample the NCR data accordingly ($X_{sub} \sim \mathcal{D}_{QUASAR'}$). Note that we explicitly order these two steps such that the models used to control for confounding are estimated using all the available data ($X_{NCR} \sim \mathcal{D}_{NCR}$). In section 5 we present our results: we compare several estimates obtained using both direct, IPW, and DR methods with the estimates of the average treatment effect (ATE) originating from the RCT, and explore the performance of IPW and DR with different trimming values. The results of the subgroup analysis and sensitivity analysis are also presented. Finally, in section 6 and 7 we discuss our findings and provide recommendations for the analysis of observational data.

2 Overview of data sources

In this study we effectively used two data sources: on one hand we used the estimates resulting from the QUASAR trial [29] as the gold standard. Note that we did not have access to the patient-level data of the QUASAR trial and relied on the estimates and descriptive statistics provided in the original paper. On the other hand, we used registry data from the NCR ($N = 52621$) as our observational dataset of choice. The impact of the discrepancy of the treatments in two data sources on the estimation will be further discussed and interpreted in section 6.

2.1 The gold standard: the QUASAR trial

The QUASAR trial was an RCT designed to test the effects of adjuvant chemotherapy for colorectal cancer patients. Between May 25, 1994, and Dec 24, 2003, $N = 3239$ patients from 150 centres in 19 countries were entered into the trial (2963 [91.5%] with stage II disease, 260 [8%] with stage III, 16 [0.5%] with stage 1, 2291 [71%] with colon cancer, 948 [29%] with rectum or colorectal cancer, median age 63 [IQR 56-68] years), of whom 82% of patients were entered from UK.

In the trial, eligible patients had undergone the resection and had no evidence of distant metastases. Furthermore, they had no definite contraindications to chemotherapy. 1622 patients were randomly assigned to receive chemotherapy consisting of FU/LV. The primary outcome was overall survival. We refer readers to [29] for more details regarding the exact protocol of the QUASAR trial.

2.2 The NCR

The NCR is a registry containing all incidences of cancer in the Netherlands managed by the Netherlands Comprehensive Cancer Organisation (Integraal Kankercentrum Nederland, IKNL). We analyzed a subset of patients who were diagnosed with colon cancer between 2006 and 2015 and who underwent surgical resection [1]. Follow-up was completed until January 31, 2019 . The total study population consisted of pathological stage II and III ($N = 52621$) patients . We excluded 100 patients who had the missing value of treatment and incidence. The primary outcome is overall survival.

2.2.1 Covariates description and data processing

Table 1 presents the distribution of the covariates in the current colon-cancer guideline and some important prognostic covariates as well as the basic demographic covariates, in which we exclude 4586 patients with clinical metastasis. For height and weight, a large number of values were not recorded (82.3 % and 81.3 %) which are imputed with median. The BMI is then computed with height and weight. For categorical covariates, the missing value is replaced with category as 'not recorded'. To explore the impact of covariates included on the estimation, we will consider four groups of covariates for analysis:

[1] We only focus on the colon cancer patients instead of colorectal cancer patients in NCR for two reasons: first in NCR, the treatment regimen for rectum cancer patients are much more complicated than colon cancer; second, reported by [29], the treatment effect of adjuvant chemotherapy didn't differ significantly by tumor site, namely, colon or rectum.

1 *3 covariates:* age, sex, and stage;
2 *5 covariates:* 2 covariates (age, sex) in addition to 3 covariates in the guideline (stage, MSI, high risk [2]);
3 *13 covariates:* 2 covariates (age, sex) in addition to 11 covariates in the guideline (stage, cM, pT, pN, high risk, MSI, number of lymph nodes assessed, colon perforation, lymphomatic invasie, agio invasie, grade);
4 *All covariates in Table 1*

3 Controlling for confounding: Direct, IPW, and DR methods

In this section, we will introduce three groups of methods for controlling for confounding in detail. Note that this controls for confounding; we cover controlling for sampling mechanism in section 4.

To formalize our methods, we first formalize the problem of estimating treatment effects. We use the potential outcome framework by Rubin [see 32, chapter 1]. We have a population with distribution $\mathcal{D}$ and size n, let X denote a vector of covariates with length d sampled according to $\mathcal{D}$, Z denote the binary treatment variable (1=adjuvant chemotherapy, 0=no adjuvant chemotherapy) and Y denote a binary response variable (1 ==death, 0 ==alive). Capital roman letters denote random variable, while lower case letters denote realized values. We use $Y_i(1)$ to denote the potential outcome for subject i under the treatment $Z = 1$ and $Y_i(0)$ denote the potential outcome under the treatment $Z = 0$. The observed outcome under the realized treatment for the subject i can be denoted as

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0). \tag{1}$$

The population average treatment effect (ATE) with covariates $X \sim \mathcal{D}$ can be formulated as:

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \tag{2}$$

However, by definition we cannot learn the true treatment effect because only one of the $Y_i(1)$ and $Y_i(0)$ is observed, which is the fundamental problem of causal inference. To estimate the average treatment effect, assumptions of unconfoundedness

$$\{Y(0), Y(1)\} \perp\!\!\!\perp Z|X \tag{3}$$

and positivity (or overlap)

$$0 < P(Z = 1|X) < 1. \tag{4}$$

[2]risk factors implied in the 2014 Dutch guideline for colon cancer [31]: perforation at diagnosis = yes, pT = 4, lymphatic invasion = yes, angio invasion = EMVI or IMVI, grade = poor to undifferentiated or unknown, the number of lymph nodes assessed less than 10. As long as one of risk factors occurs in stage II patients, patients are identified as 'high risk = high'. For stage III, 'high risk = not applicable'.

are made [33]. In RCTs, the known assignment mechanism causes the treatment assignment Z independent of both observed and unobserved covariates X, consequentially leading to the treatment Z independent of potential outcomes under the assumption of unconfoundedness. In this case, the average treatment effect (ATE) can be formulated as

$$\begin{aligned} \text{ATE} &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[Y_i(1)|Z_i] - \mathbb{E}[Y_i(0)|Z_i] \\ &= \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] \end{aligned} \tag{5}$$

In observational study where the treatment assignment Z is correlated with the covariates X, the ATE can be formulated as [34]:

$$\begin{aligned} \text{ATE} &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}_{X\sim\mathcal{D}}\{\mathbb{E}[Y_i(1)|X_i]\} - \mathbb{E}_{X\sim\mathcal{D}}\{\mathbb{E}[Y_i(1)|X_i]\} \text{ (law of iterated expectation)} \\ &= \mathbb{E}_{X\sim\mathcal{D}}\{\mathbb{E}[Y_i(1)|Z_i, X_i]\} - \mathbb{E}_{X\sim\mathcal{D}}\{\mathbb{E}[Y_i(0)|Z_i, X_i]\} \text{ (unconfoundedness)} \\ &= \mathbb{E}_{X\sim\mathcal{D}}[\mathbb{E}(Y_i|Z_i = 1, X_i)] - \mathbb{E}_{X\sim\mathcal{D}}[\mathbb{E}(Y_i|Z_i = 0, X_i)] \text{ (equation 1)} \end{aligned} \tag{6}$$

Note that the population treatment effect is defined as the average treatment effect on the target population $\mathcal{D}$. To make the estimator of the treatment effect comparable across the QUASAR ($\mathcal{D}_{QUASAR}$) and NCR ($\mathcal{D}_{NCR}$), we need to resample the subjects X_{sub} in NCR according to $\mathcal{D}_{QUASAR}$ ($X_{sub} \sim \mathcal{D}_{QUASAR}$)such that the population in NCR is matched to QUASAR. Although the true $\mathcal{D}_{QUASAR}$ is unknown, provided by the descriptive statistics of the covariates in QUASAR, we can estimate the $\mathcal{D}_{\mathcal{QUASAR}'}$ by assuming the covariates are independent, which will be introduced in section 4.

Now that we have formalized the problem of estimating the treatment effect, we will introduce different methods for modeling the $\mathbb{E}[Y_i(1)]$ and $\mathbb{E}[Y_i(0)]$ in observational study, namely, direct methods, IPW methods, and DR methods. In each case we have multiple implementations: we discuss implementations relying on relatively simple regression models and implementations relying on the flexible, non-parametric BART model. All the models are fitted based on the population (N=52521). Then the estimators of the treatment effect of the target population in NCR can be obtained by averaging the treatment effect of subjects sampled according to $\mathcal{D}_{\mathcal{QUASAR}'}$.

3.1 Direct methods

The direct method, also known as regression adjustment, uses the regression method to directly fit the outcome model $\mathbb{E}[Y|Z, X]$. If the unconfoundedness assumption holds, $\mathbb{E}[Y_i(1)]$ and $\mathbb{E}[Y_i(0)]$ can be obtained by marginalizing the conditional outcome of Y given covariates X across the population under treatment $Z = 1$ and $Z = 0$, respectively, and can be formalized as follows:

$$\mathbb{E}[Y(z)] = \mathbb{E}_{X\sim\mathcal{D}}[\mathbb{E}(Y_i|Z_i = z, X_i)], \text{for } z = 0, 1 \tag{7}$$

which have been proved in equation 6. Therefore, by fitting the model for potential outcomes conditional on the covariates to adjust for the direct and indirect associations between covariates and outcomes, we can obtain the potential outcomes for each subject and subsequentially the ATE of the population. In the following section, we will introduce two methods to fit the model for potential outcomes $\mathbb{E}(Y|Z,X)$ based on observed outcomes, namely, a parametric model of logistic regression (Direct/LReg) and a non-parametric model of BART (Direct/BART).

3.1.1 Logistic regression (LReg)

Logistic regression (LReg) is a commonly used parametric method to control for confounding. When the covariates are strongly predictive of outcomes, we can efficiently identify the relation between treatment and outcome by making use of the information about relations between covariates and outcomes. By specifying the main effect and interaction effect between treatment and each independent covariate, $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ can be expressed as:

$$\begin{aligned}\mathbb{E}[Y(z)] &= \mathbb{E}_{X\sim\mathcal{D}}[\mathbb{E}(Y_i|Z_i=z,X_i)] \\ &= n^{-1}\sum_{i=1}^{n}\frac{1}{1+exp(\hat{\beta}_0+\hat{\boldsymbol{\beta}}_{\mathbf{1}}{}'X_i+\hat{\alpha}z+\hat{\boldsymbol{\beta}}_{\mathbf{2}}{}'X_iz)^{-1}} \quad \text{for} \quad z=1,0\end{aligned} \tag{8}$$

where $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$, $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ and $\hat{\alpha}$ are estimated by maximum likelihood. We perform lasso regularization to prevent overfitting [35].

3.1.2 Bayesian additive regression tree (BART)

BART is a sum-of-trees model where each tree is constrained by a regularization prior to a weak learner, and fitting and inference are obtained by Bayesian Markov Chain Monte Carlo (MCMC) algorithm that generates samples from a posterior [16]. [27] and [36] first promoted the use of BART to obtain the heterogeneous and average treatment effect. [26] also compared the BART with other methods using simulated observational data in healthcare and results show that BART has a low bias. Unlike many other non-parametric models, BART is flexible and it can handle non-linear main effects and multi-way interactions without much input from researchers [37]. We refer the readers to [16] for details. For binary outcomes, $\mathbb{E}(Y|Z,X)$ can be estimated using a probit model and then $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ can be obtained by averaging the outcome conditional on X for $Z=1$ and $Z=0$ across the population.

In our study, we use the default setting in [16] because of BART's robust performance with respect to various hyperparameter settings: the number of trees = 50 (we found that the results show no gain with more number of trees incorporated), the number of posterior draws = 1000, and the prior distribution of σ is inverse gamma distribution $IG(\alpha,\beta)$ where shape parameter(splitting probability) of α is 0.95 and rate parameter (depth penalty) β is 2, respectively.

3.2 Inverse probability weighting estimation (IPW) methods

IPW is one of the PS methods for correcting the mismatch of data distribution produced by the treatment assignment. We define PS as follows:

$$\pi(X)=\mathbb{E}(Z=1|X) \tag{9}$$

We refer readers to [28, 33] for details about the PS. The IPW method is considered to have some methodological advantages because it imitates the same treament assignment mechanism as the RCT. Weighting by the PS effectively generates the pseudopopulation in which the treatment assignment is independent of observed covariates, generating the similar distribution of covariates between two groups. Hence, in this pseudopopulation, we can obtain $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ simply by averaging the outcome in treatment and control groups after weighting by the inverse of PS. Then the $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ with respect to the distribution of $X \sim \mathcal{D}$ in the target population with n subjects can be obtained via

$$\mathbb{E}[Y(1)] = \left[\sum_{i=1}^{n} \frac{\mathbb{1}(Z_i = 1)}{\pi(X_i)}\right]^{-1} \sum_{i=1}^{n} Y_i \frac{\mathbb{1}(Z_i = 1)}{\pi(X_i)} \tag{10}$$

and

$$\mathbb{E}[Y(0)] = \left[\sum_{i=1}^{n} \frac{\mathbb{1}(Z_i = 0)}{1 - \pi(X_i)}\right]^{-1} \sum_{i=1}^{n} Y_i \frac{\mathbb{1}(Z_i = 0)}{1 - \pi(X_i)}. \tag{11}$$

Then the ATE can be estimated by directly comparing the difference between $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$. In this study, we compute the PS using two regression models, namely, LReg with lasso regularization (IPW/LReg), and BART (IPW/BART) with default setting. We trim the weights at 0.1 and 0.9, which is implemented in both IPW and DR methods, and explore the effect of different trimmed values on the ATE estimation with these methods, which we will further illustrate in section 5.

3.3 Doubly robust methods (DR)

The DR estimator combines a model for the potential outcomes as a function of covariates $\mathbb{E}(Y|Z, X)$ and a model for assignment mechanism as a function of covariates $\mathbb{E}(Z|X)$ (i,e, $\pi(X)$). Particularly, we fit a separate outcome model $m_z = \mathbb{E}(Y_i|Z_i = z, X_i)$ for the treatment group $Z = 1$ and control group $Z = 0$. $m_z(X_i)$ is the predicted values for the subject i given X_i and the treatment $Z_i = z$. Then $\mathbb{E}[Y_i(1)]$ and $\mathbb{E}[Y_i(0)]$ with respect to the distribution of $X \sim \mathcal{D}$ in the target population can be estimated via [28]:

$$\begin{aligned}
\mathbb{E}[Y(1)] &= n^{-1} \sum_{i=1}^{n} \left\{ \frac{\mathbb{1}(Z_i = 1)}{\pi(X_i)} [Y_i - m_1(X_i)] + m_1(X_i) \right\} \\
\mathbb{E}[Y(0)] &= n^{-1} \sum_{i=1}^{n} \left\{ \frac{\mathbb{1}(Z_i = 0)}{1 - \pi(X_i)} [Y_i - m_0(X_i)] + m_0(X_i) \right\}
\end{aligned} \tag{12}$$

The name DR is derived from the fact that the estimation remains unbiased when either a model for treatment assignment mechanism or a model for the potential outcomes is correctly specified [22]. Combining a direct estimator and an IPW estimator under the form of a DR estimator leads to lower bias than the direct estimator alone, and lower variance than the IPW estimator alone [38].See [39, p148-149] for more details of the proof of the DR property. In this study, four DR estimators are constructed based on two models for potential outcomes $\mathbb{E}[Y|Z, X]$

fitted by LReg and BART and two models for assignment mechanism $\mathbb{E}[Z|X]$ fitted by LReg and BART. Table 2 shows the details of these four combinations.

3.4 Additional methods

Next to the three groups of methods described above, we also include two other alternatives:

1 To explore the effect of sampling mechanism and treatment assignment on estimation, we roughly model potential outcomes $\mathbb{E}(Y_i|Z_i)$ on the whole population before the subsampling (($X_{NCR} \sim \mathcal{D}_{NCR}$) and the study population after subsampling ($X_{sub} \sim \mathcal{D}_{QUASAR'}$)respectively.

2 Furthermore, in addition to covariates, we also include the PS in the direct method $\mathbb{E}[Y_i|Z_i, X_i, \pi(X_i)]$ (Direct/ps-BART). Both the model for potential outcomes and PS (i.e., $\pi(X_i) = \mathbb{E}(Z_i|X_i)$) are fitted by BART with the default setting. Direct/ps-BART incorporates the estimate of PS in the specification of the outcome model, implicitly inducing a covariate-dependent prior to the regression function, which even outperforms the BART model ([25, 40].

3.5 Overview of methods

Table 2 provides an overview of the methods introduced in section 3.1, 3.2, 3.3 and 3.4. Parametric LReg and non-parametric BART with the default setting introduced in section 3.1.2 are used to fit both models for the potential outcomes $\mathbb{E}[Y_i|Z_i, X_i]$ and the treatment assignment $\mathbb{E}(Z_i|X_i)$.

When fitting the model using LReg, we need to specify the relationship between each variable: the main effect of covariates and the interaction effect between the treatment and each covariate are specified for the model of $\mathbb{E}(Y_i|X_i, Z_i)$ in the direct method; the main effect of all covariates and two-way interaction effects of all covariates are specified for the model of $\mathbb{E}(Y_i|X_i, Z_i = z)$ in DR method and the model of $\mathbb{E}(Z_i|X_i)$ in IPW method. To prevent overfitting, lasso regularization is implemented on all the models fitted by LReg, the loss function is to minimize the mean squared error, and the ten-fold cross-validation is implemented to choose the regularization parameter.

Note that all the models are initially fitted based on the whole population ($X_{NCR} \sim \mathcal{D}_{NCR}$) and the estimator of ATE for comparison with the QUASAR are then obtained from the subsamples ($X_{sub} \sim \mathcal{D}_{QUASAR'}$) from the whole population. We quantify the sampling uncertainty by stratified sampling the study population for 5000 times, which we will introduce in section 4.

4 Study population sampling

In this section, we will select the study population from the whole population ($N = 52621$) to ensure that the subsamples of the NCR ($X_{sub} \sim \mathcal{D}_{QUASAR'}$) is similar to the samples in the QUASAR trial ($X_{QUASAR} \sim \mathcal{D}_{QUASAR}$). We first select the eligible patients ($N = 45539$) by complying with the inclusion criteria of QUASAR. Then we subsample the target population according to the probablity estimated from the joint distribution of $\mathcal{D}_{QUASAR'}$ over important covariates by assuming independence.

4.1 Eligible participates

To select the eligible participates for the study population, we first sought to comply with the same inclusion criteria as the QUASAR trial [29]. From a total of $N = 52621$ patients with colon cancer, 100 patients missed the information of treatment and incidence, 45442 patients were without distant metastases (cM==1), diagnosed stage II or stage III and aged 23 to 86. All of these 45442 remaining patients are included for a two-year follow-up analysis and $n = 34832$ for a five-year follow-up analysis in which patients diagnosed after 2013 are excluded. Figure 2 provides an overview of this first selection step [3].

4.2 Sampling

Next to following the inclusion criteria in the QUASAR trial, we also estimate the distribution $\mathcal{D}_{QUASAR'}$ over covariates by correcting for the distribution employed in the NCR. Effectively, we ensured that our subsample $X_{sub} \sim \mathcal{D}_{QUASAR'}$ resulting from the NCR data corresponds – in terms of the distribution of several important covariates – as closely as possible to the QUASAR trial data. By correcting for the sampling mechanism such that the distribution of covariates are the same in both data sources (i.e., $\mathcal{D}_{QUASAR'}$ and $\mathcal{D}_{QUASAR}$), the estimator obtained by averaging the outcomes across subsamples of NCR ($X_{sub} \sim \mathcal{D}_{QUASAR'}$) can be reliably compared to the QUASAR in further analysis. To do so, we divide the eligible population into 16 subgroups by stage, sex, and age, and sampled from each subgroup according to the joint probability of their cancer stage, their sex, and their age as observed in the QUASAR sample (see Appendix) [4].

Then we implemented this subsampling such that the resulting sample was as large as possible while maintaining the desired joint distribution of the covariates. The stratified subsampling leads to the selection of 7704 and 6378 patients from the NCR for two-year follow-up and five-year follow-up analysis; Table 3 demonstrates that our subsampling procedure effectively ensured the desired joint distribution of the NCR subsample comparable to the original QUASAR sample. Note that, to incorporate the variability induced by our subsampling procedure in our remaining procedure, we create $m = 5000$ stratified (sub)samples.

5 Results

In this section, we compare the performance of the different methods to control for confounding. We examine the ATE in two and five years of follow-up. The main results are presented in Figure 3, where the ATEs in each method are estimated with four groups of covariates. Numerical details are provided in Table 5 and 6. In addition, to assess the magnitude of violation of positivity assumption in IPW

[3]It is worth noting that in our study the models for potential outcomes and assignment mechanism were fitted based on a very large sample ($N = 52521$), utilizing the information from the general population which is beyond the scope of most RCTs. Thus, it is reasonable to, in the future, estimate the treatment effect of patients who are not included in the RCT using these methods.

[4]No other covariate distributions are reported in the paper of QUASAR trial. Note that only the marginals are reported, and we estimate the joint distribution by assuming independence supported by [41].

and DR methods, we visualize the PS distribution between treatment and control groups and explore the variation of IPW and DR to different trim values – an approach to improve the performance in face of the positivity violation – which are shown in Figure 4. We also show the results of subgroup analysis to identify the heterogeous treatment effect and sensitivity analysis to assess the assumption of unconfoundedness.

5.1 Results of the estimation of treatment effect

To explore the impact of different covariates sets used on the estimation, we compare the results of three methods using four sets of covariates, which are shown in Figure 3. We find that the methods using the least covariates are the closest to the ground truth of RCT while the results using more covariates substantially overestimate the ATE.

We then explore the results of the different methods after the two-year follow-up. In the RCT the estimated ATE -0.021 (95% CI -0.039, -0.001). The result is presented in Figure 3 using the black dotted line (point estimate) and the grey band (CI). Using the NCR data without adjusting for any confounders — i.e., the Crude method — we obtain an ATE of -0.069 (95% CI -0.076, -0.062) for $X_{NCR} \sim \mathcal{D}_{NCR}$ and a median ATE of -0.019 (range from -0.041, 0.003) for $X_{sub} \sim \mathcal{D}_{QUASAR'}$. Thus, if neither control for sampling mechanism nor treatment assignment, the Crude method overestimates the treatment effect. This overestimation could be alleviated by controlling for the sampling mechanism, although the estimation still deviates from the RCT and shows large variance produced by the sampling process.

Regarding the impact of the number of covariates controlled for on the ATE estimation, it seems that using sex, age, and stage can properly obtain the ATE as close as possible to the RCT. When high risk and more covariates added (shown in the second, the third and the fourth row in Figure 3), the estimate deviates far from the results from the RCT.

Regarding the direct methods, the direct/BART and direct/ps-BART are very close to the ground truth while the direct/LReg seems to overestimate the ATE. Notably, the variation over different subsamples obtained to correct for the sampling mechanism in direct methods is very small: the direct method seems very robust to this sampling variation. Compared to direct methods, the IPW methods, however, seem to have a huge variation. The DR methods seem to – as expected – combine the best of both worlds: the estimates are close to the ground truth, and they exhibit small sampling variance compared to IPW methods. For each of the DR methods, irrespective of using parametric or non-parametric models, the majority of the point estimates are within the confidence bound of the original ICR. The DR using 13 and all covariates all show the same results, which probably means that unconfoundedness is hold in both cases.

In the five-year follow-up analysis, the ATE in RCT is -0.025 (95% CI -0.059, 0.008). In NCR, ATEs using crude method for $X_{NCR} \sim \mathcal{D}_{NCR}$ and $X_{sub} \sim \mathcal{D}_{QUASAR'}$ are -0.073 (95% CI -0.083, -0.063) and a median of -0.031 (range from -0.065, 0.005), respectively. These results are qualitatively similar to those obtained after two years: the Crude method overestimates the treatment effect, all the DR methods are very close to the ground truth. The majority of the point estimates obtained from the DR

method – whether using a parametric or a non-parametric approach – fall within the confidence bound obtained in the RCT.

To conclude, our results show that a) in general all the estimates are consistent in terms of direction with the results of RCT; b) strength of the association between treatment and outcome varies across statistical methods and covariates included, which is also indicated in [42]. Clearly, without cautious controlling for the confounding and sampling mechanism, results using the crude method will not be reliable in the aspect of both magnitude and directionality of the treatment effect, as reported in Crude methods on population $X_{NCR} \sim \mathcal{D}_{NCR}$ and $X_{sub} \sim \mathcal{D}_{QUASAR'}$. However, when correction methods are used carefully, we do obtain estimates that are close to the ground truth; c) regarding the performance of the four groups of methods, the DR estimations, particularly modeled by the non-parametric BART, seem more robust in obtaining the treatment effect than IPW. Within the direct methods the choice of a model, the non-parametric model, namely, ps-BART, is quite flexible and performs properly.

5.2 The results of the estimation with different trimmed values

Regarding the effect of covariates on the estimation of ATE, we find that the estimate using the least number of covariates is closest to the ground truth from the RCT while estimates from models using more covariates deviate from the ground truth. The right panel of Figure 4 shows the distribution of PS between treatment and control, indicating that adjusting for more covariates leading to more extreme PS. Although IPW estimator is theoretically unbiased, it is sensitive to extreme weights, and hence weights are commonly truncated or bounded to shrink the variance of the estimator.

In our study, weights are trimmed at a fixed level, namely, at 0.1 and 0.99, 0.5 and 0.95, 0.1 and 0.9, and 0.2 and 0.8 [20] and the results are shown in the left panel of Figure 4. We found that across the results, trimming the IPW weights decreases the variance of the estimator but makes the point estimate substantially vary across trimmed values. Besides, we also find that IPW methods benefit from the trimming, which is quite clear when looking at complex models using 5, 13 and all covariates. This impact on the complex model can also be validated by looking at the estimates with only three covariates: in this case, the point estimation doesn't vary across different trim values probably due to the comparable specification of PS model with the ground truth. Overall, in our study, trimming the extreme weight can help the IPW method approach to the ground truth.

5.3 Subgroup analysis

From the previous results, we find that high risk and other pathological characteristics are sensitive to the estimation. To understand the reason that leads to this overestimation, we explore the estimation in the subgroups of stage II and stage III classified by the high-risk level using Direct/ps-BART. The results are shown in Figure 5. We found that by adjusting the high risk the method can identify the heterogeneous treatment effect among the subgroups in stage II cancer. This finding is supported by a) the treatment effect on stage III patients are the same in the model with and without adjusting for high risk covariate ; b) the treatment effect

for stage II colon cancer who are not identified as high risk are the same and is approximately equal to zero in all the models, regardless of high risk being adjusted or not; c) the heterogeneity of treatment effect in stage II cancer is supported by many studies indicating that part of stage II high-risk colon cancer. for example, pT4, can achieve the survival benefit from the adjuvant chemotherapy [11, 43].

Based on these findings, although the models using three covariates can estimate the treatment effect comparable to the RCT, we can't rule out the possibility that the true treatment effect of the NCR data is very close to the estimate from the models using 5, 13 and all covariates, because no pathological information was provided in the RCT to identify the heterogeneous treatment effect of the high risk stage II patients. Overall, the model using three covariates without adjusting for high risk can estimate the treatment effect as close as possible to the ground truth from the RCT while at the cost of inability to identify the heterogeneity among the population, in particular, the subgroup of stage II cancer.

5.4 Sensitivity analysis

The subgroup analysis implicitly indicates that the estimation for stage III patients are robust to various covariates included while stage II patients are quite sensitive to the type and the number of covariates. Hence, we can use sensitivity analysis in subgroups to test for the impact of an unmeasured confounder and assess the extent to which we allow the unconfoundedness assumption to be violated in subgroups.

In our study, a two-parameter sensitivity analysis based on flexible non-parametric methods BART is implemented [44] and the results are presented in Figure 6. It implies that the estimate of stage II colon cancer, particularly those are unknown risk, is sensitive to the unmeasured confounder and the sign of the estimate can even change in the case where, for instance, some high risk factors are not controlled for. This implies that more information about this subgroup should be collected to identify the high risk stage, otherwise the estimation might not truly reveal the treatment effect. However, for stage III patients, the estimation of the treatment effect is quite robust to the unmeasured confounders and seemingly deterministic.

6 Discussion

In this study, we exploited the unique opportunity of having access to large-scale colon cancer registry data to compare three different methods of controlling for potential confounding in observational data, namely, direct method, IPW method, and DR method. Our study, by comparing the resulting estimates to those obtained in a large scale RCT, validated that these methods can be used to obtain the estimates of treatment effects close to the ground truth. This opens up the door to use the large-scale, high-quality registry data regarding colon cancer to further strengthen the (Dutch) guidelines by adding, as a third source of information next to RCT data and expert consensus, the available registry data. It has to be noted, however, that adding this third source requires care: methods using naive estimators — such as the crude method evaluated in this study — can be highly biased. We can only use the registry data when both the sampling process and the assignment process are properly controlled for.

Next to demonstrating the utility of observational data in this specific case, we also find evidence of the advantages of modeling the outcome (as done in the direct

method and the DR method) over solely modeling the treatment assignment process (IPW method). This latter result might be because it is hard to properly estimate this process. Besides, the supreme performance of non-parametric BART over simple parametric LReg is quite promising, which implies its use for further analysis. In the subgroup analysis, we find that when adjusting for the high-risk in models, the heterogeneous treatment effect could be identified, although the estimation might deviate from the result of the RCT. This dilemma of whether or not to include more covariates as to obtain both the accurate and heterogeneous estimation could be solved by access to a variety of data with complete information in each subgroup, which could probably shrink the variation of estimation from the model using all the covariates as shown in Figure 5. The unconfoundedness assumption is also explored using sensitivity analysis and the result is qualitatively reasonable.

In general, the estimates we obtained from the NCR observational data — as can be seen in Figure 3 — are often higher than those in the ground truth Quasar trial. While we have discussed several sources of this discrepancy due to our explored correction methods, there are possible two alternative explanations. On one hand, although we have resampled the registry data to as much as possible match the population in two data sources, there is still uncorrectable heterogeneity between populations at baseline. For example, more stage II patients are identified as high risk in registry data than RCT, which consequentially increases the overall estimation. On the other hand, it is now known that CAPOX and FOLFOX are more effective than FU/LV [4]. The fact that in the NCR data more patients received CAPOX/FOLFOX as opposed to FU/LV, while in the QUASAR trial all patients received FU/LV, might truly increase the average treatment effect of chemotherapy overall in the NCR.

Regarding the subsampling process, we made two assumptions a) the assumption of independence among gender, age and stage when estimating the distribution of population in QUASAR based on the population-level statistics provided in the end of the enrollment; b) the assumption of no change of the estimated distribution over time when subsampling the target population accordingly from the NCR data. These two assumptions, fortunately, can approximately be checked by the paper published during that time. Some studies reported that males and females have a similar distribution of incidence by age from 1971 to 1999 [45, 46]. Therefore, the variation of the distribution of age and gender in colon cancer over time could be ignorable and hence doesn't violate the assumption. However, all these two assumptions are still untestable without patient-level data with covariate information.

7 Conclusion

To conclude, our study shows that it is possible to obtain treatment effects from observational data when analyzed carefully. Without cautious controlling for confounders, the results from observational data may mislead people to make incorrect decisions. Our study also shows that the colon cancer registry data, when handled with care, can potentially be used to inform future guidelines: at the very least the ATE obtained from this data source – with the appropriate controls – provides estimates that are similar to those obtained in an RCT.

List of abbreviations
RCT: randomized controlled trial
IPW: inverse probability weighting
DR: doubly robust
NCR: Netherlands cancer registry
HR: hazard ratio
QUASAR: quick and simple and reliable
CAPOX: capecitabine and oxaliplatin
FOLFOX: fluorouracil, leucovorin, and oxaliplatin
FU/LV: fluorouracil and leucovorin
BART: bayesian additive regression trees
IKNL: Integraal Kankercentrum Nederland
BMI: body mass index
ATE: average treatment effect
LReg: logistic regression

Declarations
Ethics approval and consent to participate
Permission to access the Netherlands Cancer registry (NCR) and the databases of Statistics Netherlands (CBS) was granted by both IKNL (Netherlands Comprehensive Cancer Care Organisation) and the CBS (Statistics Netherlands). According to the Dutch Medical Research Involving Human Subjects Act, this study did not require medical-ethical approval, as was confirmed in writing by the medical ethical committee of the AMC, 10 July 2013. We followed the ethical principles for medical research involving human subjects as laid down in the Declaration of Helsinki and adopted by the World Medical Association (WMA Declaration of Helsinki, 2000).

Consent for publication
Not applicable.

Availability of data and materials
The data that support the findings of this study are available from IKNL and CBS but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Competing interests
The authors declare that they have no competing interests.

Funding
The project was funded by the grant from the China Scolarship Council Program, grant number 201806860047. The funding body had no role in the design of the study and collection, analysis, and interpretation of data, nor in writing the manuscript.

Authors' contributions
SLJ and EV carried out the statistical analyses. SLJ and MCK drafted the manuscript. SLJ participated in the interpretation of the statistical analyses. All authors critically read the drafts of this paper, and read and approved its final version.

Acknowledgements
The authors thank the registration team of the Netherlands Comprehensive Cancer Organisation (IKNL) for the collection of data for the Netherlands Cancer Registry, the Prospective National Colon Rectum Cancer (PLCRC) working group for the collection of Patient Reported Outcome data as well as IKNL staff for scientific advice.

Author details
[1]Department of Methodology and Statistics, Tilburg University, Warandelaan 2, 5000 LE Tilburg, The Netherlands. [2]Jheronimus Academy of Data Science, Sint Janssingel 92, 5211 DA 's-Hertogenbosch, The Netherlands. [3]IKNL (Integraal Kankercentrum Nederland), Godebaldkwartier 419, 3511 DT Utrecht, The Netherlands. [4]Department of Surgery, Radboud University Medical Center, Geert Grooteplein Zuid 10, 6525 GA Nijmegen, The Netherlands. [5]Department of Medical Oncology, Radboud University Medical Center, Geert Grooteplein Zuid 10, 6525 GA Nijmegen, The Netherlands.

References
1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **68**(6), 394–424 (2018)
2. Brouwer, N.P., Heil, T.C., Rikkert, M.G.O., Lemmens, V.E., Rutten, H.J., de Wilt, J.H., van Erning, F.N.: The gap in postoperative outcome between older and younger patients with stage i-iii colorectal cancer has been bridged; results from the netherlands cancer registry. European Journal of Cancer **116**, 1–9 (2019)
3. Labianca, R., Marsoni, S., Pancera, G., Torri, V., Zaniboni, A., Erlichman, C., Pater, J., Shepherd, L., Zee, B., Seitz, J., *et al.*: Efficacy of adjuvant fluorouracil and folinic acid in colon cancer. The Lancet **345**(8955), 939–944 (1995)

4. André, T., Boni, C., Navarro, M., Tabernero, J., Hickish, T., Topham, C., Bonetti, A., Clingan, P., Bridgewater, J., Rivera, F., *et al.*: Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage ii or iii colon cancer in the mosaic trial. J clin oncol **27**(19), 3109–3116 (2009)
5. Yothers, G., O'Connell, M.J., Allegra, C.J., Kuebler, J.P., Colangelo, L.H., Petrelli, N.J., Wolmark, N.: Oxaliplatin as adjuvant therapy for colon cancer: updated results of nsabp c-07 trial, including survival and subset analyses. Journal of clinical oncology **29**(28), 3768 (2011)
6. Gill, S., Loprinzi, C.L., Sargent, D.J., Thomé, S.D., Alberts, S.R., Haller, D.G., Benedetti, J., Francini, G., Shepherd, L.E., Francois Seitz, J., *et al.*: Pooled analysis of fluorouracil-based adjuvant therapy for stage ii and iii colon cancer: who benefits and by how much? Journal of clinical oncology **22**(10), 1797–1806 (2004)
7. Erlichman, C.: Efficacy of adjuvant fluorouracil and folinic acid in b2 colon cancer. Journal of Clinical Oncology **17**(5), 1356–1363 (1999)
8. Jongeneel, G., Klausch, T., van Erning, F.N., Vink, G.R., Koopman, M., Punt, C.J., Greuter, M.J., Coupé, V.M.: Estimating adjuvant treatment effects in stage ii colon cancer: comparing the synthesis of randomized clinical trial data to real world data. International Journal of Cancer (2019)
9. Hugen, N., Verhoeven, R., Radema, S., De Hingh, I., Pruijt, J., Nagtegaal, I., Lemmens, V., De Wilt, J.: Prognosis and value of adjuvant chemotherapy in stage iii mucinous colorectal carcinoma. Annals of oncology **24**(11), 2819–2824 (2013)
10. Hugen, N., Verhoeven, R.H., Lemmens, V.E., van Aart, C.J., Elferink, M.A., Radema, S.A., Nagtegaal, I.D., de Wilt, J.H.: Colorectal signet-ring cell carcinoma: benefit from adjuvant chemotherapy but a poor prognostic factor. International journal of cancer **136**(2), 333–339 (2015)
11. Verhoeff, S., Van Erning, F., Lemmens, V., De Wilt, J., Pruijt, J.: Adjuvant chemotherapy is not associated with improved survival for all high-risk factors in stage ii colon cancer. International journal of cancer **139**(1), 187–193 (2016)
12. Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research **46**(3), 399–424 (2011)
13. Atan, O., Jordan, J., van der Schaar, M.: Deep-treat: Learning optimal personalized treatments from observational data using neural networks. (2018). AAAI
14. Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., *et al.*: Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. Statistical Science **34**(1), 43–68 (2019)
15. Prentice, R.: Use of the logistic model in retrospective studies. Biometrics, 599–606 (1976)
16. Chipman, H.A., George, E.I., McCulloch, R.E., *et al.*: Bart: Bayesian additive regression trees. The Annals of Applied Statistics **4**(1), 266–298 (2010)
17. Pratola, M.T.: Efficient metropolis–hastings proposal mechanisms for bayesian regression tree models. Bayesian Anal. **11**(3), 885–911 (2016). doi:10.1214/16-BA999
18. Mohammadi, R., Pratola, M., Kaptein, M.: Continuous-time birth-death mcmc for bayesian regression tree models. arXiv preprint arXiv:1904.09339 (2019)
19. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association **113**(523), 1228–1242 (2018)
20. Austin, P.C., Stuart, E.A.: 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 **34**(28), 3661–3679 (2015)
21. Dudík, M., Langford, J., Li, L.: Doubly robust policy evaluation and learning. arXiv preprint arXiv:1103.4601 (2011)
22. Bang, H., Robins, J.M.: Doubly robust estimation in missing data and causal inference models. Biometrics **61**(4), 962–973 (2005)
23. Algra, A.M., Rothwell, P.M.: Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. The lancet oncology **13**(5), 518–527 (2012)
24. Nishihara, R., Wu, K., Lochhead, P., Morikawa, T., Liao, X., Qian, Z.R., Inamura, K., Kim, S.A., Kuchiba, A., Yamauchi, M., *et al.*: Long-term colorectal-cancer incidence and mortality after lower endoscopy. New England Journal of Medicine **369**(12), 1095–1105 (2013)
25. Hahn, P.R., Murray, J., Carvalho, C.M.: Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. Confounding, and Heterogeneous Effects (October 5, 2017) (2017)
26. Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., Gallego, B.: Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. Statistics in medicine **37**(23), 3309–3324 (2018)
27. Hill, J.L.: Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics **20**(1), 217–240 (2011)
28. Lunceford, J.K., Davidian, M.: Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine **23**(19), 2937–2960 (2004)
29. Group, Q.C., *et al.*: Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. The Lancet **370**(9604), 2020–2029 (2007)
30. Imbens, G., Menzel, K.: A causal bootstrap. (2018)
31. Oncoline: Dutch Guideline Colorectal Cancer 2014. `https://www.oncoguide.nl/#!/projects/27/tree/199/202`
32. Rubin, D.B.: Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association **100**(469), 322–331 (2005)
33. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
34. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology **66**(5), 688 (1974)
35. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the royal statistical

society: series B (statistical methodology) **67**(2), 301–320 (2005)
36. Green, D.P., Kern, H.L.: Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. Public opinion quarterly **76**(3), 491–511 (2012)
37. Tan, Y.V., Roy, J.: Bayesian additive regression trees and the general bart model. arXiv preprint arXiv:1901.07504 (2019)
38. Bibaut, A., Malenica, I., Vlassis, N., Van Der Laan, M.: More efficient off-policy evaluation through regularized targeted learning. In: International Conference on Machine Learning, pp. 654–663 (2019)
39. Tsiatis, A.: Semiparametric Theory and Missing Data. Springer, ??? (2007)
40. Santos, P.H.F.d., Lopes, H.F.: Tree-based bayesian treatment effect analysis. arXiv preprint arXiv:1808.09507 (2018)
41. Hayne, D., Brown, R., McCormack, M., Quinn, M., Payne, H., Babb, P.: Current trends in colorectal cancer: site, incidence, mortality and survival in england and wales. Clinical oncology **13**(6), 448–452 (2001)
42. Cnossen, M.C., van Essen, T.A., Ceyisakar, I.E., Polinder, S., Andriessen, T.M., van der Naalt, J., Haitsma, I., Horn, J., Franschman, G., Vos, P.E., *et al.*: Adjusting for confounding by indication in observational studies: a case study in traumatic brain injury. Clinical epidemiology **10**, 841 (2018)
43. Babcock, B.D., Aljehani, M.A., Jabo, B., Choi, A.H., Morgan, J.W., Selleck, M.J., Luca, F., Raskin, E., Reeves, M.E., Garberoglio, C.A., *et al.*: High-risk stage ii colon cancer: not all risks are created equal. Annals of surgical oncology **25**(7), 1980–1985 (2018)
44. Dorie, V., Harada, M., Carnegie, N.B., Hill, J.: A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. Statistics in medicine **35**(20), 3453–3470 (2016)
45. Quinn, M., Babb, P., Brock, A., Kirby, L., Jones, J.: Cancer trends in england and wales 1950–1999: Studies on medical and population subjects. The Stationary Office: Norwich, UK (2001)
46. Mistry, M., Parkin, D., Ahmad, A.S., Sasieni, P.: Cancer incidence in the united kingdom: projections to the year 2030. British journal of cancer **105**(11), 1795 (2011)

Figures

Figure 1 the flowchart of this study

Figure 2 The first step of selecting eligible patients from the NCR; patients without follow up, and patients who did not meet the inclusion criteria for the QUASAR trial are removed.

Figure 3 ATE in two (left panel) and five (right panel) years of follow-up. It is worth noting that the treatment effect on colon cancer in the QUASAR trial should be smaller than the dotted line in the plot because in the QUASAR 29% rectum cancer patients were included for whom the treatment effect is higher than colon cancer based on the results reported in the paper.

Figure 4 left panel: ATE of IPW and DR methods with different trimming values in two years follow-up. Right panel: the distribution of propensity score modeled by BART across the treatment and control groups.

Figure 5 Subgroup analysis of stage II (risk=normal, risk=high) and stage III in two and five years follow up.

Figure 6 The sensitivity analysis implemented using treatSens package in stage II and stage III cancer, controlling for gender and age. NS denotes 'Not significant'.

Tables

Table 1 Baseline characteristics of stage II and III patients who underwent surgery between 2006 And 2015

Characteristics	Whole population (n=47935)	Chemotherapy (n=15712)	No chemotherapy (n=32223)
Age	70.6 (11.2)	64.3 (9.8)	73.7 (10.4)
Sex			
male	24383 (50.9)	8336 (53.1)	16047 (49.8)
female	23552 (49.1)	7376 (46.9)	16176 (50.2)
BMI	26.4 (6.3)	26.5 (7.2)	26.4 (5.9)
Number of co-morbidity			
1	9159 (19.1)	3646 (23.2)	5513 (17.1)
$\geq$2	6101 (12.7)	1528 (9.7)	4573 (14.2)
not recorded	32675 (68.2)	10538 (67.1)	22137 (68.7)
Perforation at diagnosis			
no	6095 (12.7)	2248 (14.3)	3847 (11.9)
yes	330 (0.7)	125 (0.8)	205 (0.6)
unknown	508 (1.1)	181 (1.2)	327 (1.0)
not recorded	41002 (85.5)	13158 (83.7)	27844 (86.4)
Location of the tumor			
proximal	27472 (57.3)	8017 (51.0)	19455 (60.4)
distal	19479 (40.6)	7401 (47.1)	12078 (37.5)
other/not otherwise specified(nos)	984 (2.1)	294 (1.9)	690 (2.1)
pT:pathological size of tumor			
T0	3 (0.0)	3 (0.0)	0 (0.0)
T1	515 (1.1)	330 (2.1)	185 (0.6)
T2	1890 (3.9)	1204 (7.7)	686 (2.1)
T3	36970 (77.1)	10423 (66.3)	26547 (82.4)
T4	8557 (17.9)	3752 (23.9)	4805 (14.9)
pN:pathological spread to lymph nodes			
N0	25622 (53.5)	2139 (13.6)	23483 (72.9)
N1	14687 (30.6)	8647 (55.0)	6040 (18.7)
N2	7327 (15.3)	4884 (31.1)	2443 (7.6)
NX	299 (0.6)	42 (0.3)	257 (0.8)
Number of lymph nodes assessed	17 (10.5)	18 (10.9)	17 (10.3)
cM:clinical presence of metastasis			
M0	45790 (95.5)	15157 (96.5)	30633 (95.1)
MX	2145 (4.5)	555 (3.5)	1590 (4.9)
Stage			
II	25801 (53.8)	2114 (13.5)	23687 (73.5)
III	22134 (46.2)	13598 (86.5)	8536 (26.5)
Morphology			
mucinous	6991 (14.6)	2108 (13.4)	4883 (15.2)
adenocarcinoom	39703 (82.8)	13189 (83.9)	26514 (82.3)
signet ring cell	611 (1.3)	238 (1.5)	373 (1.2)
neuroendocrine tumors	211 (0.4)	33 (0.2)	178 (0.6)
other/nos	419 (0.9)	144 (0.9)	275 (0.9)
Grade			
good to moderate	35561 (74.2)	11314 (72.0)	24247 (75.2)
poor to undifferentiated	8078 (16.9)	3035 (19.3)	5043 (15.7)
unknown	4296 (9.0)	1363 (8.7)	2933 (9.1)
Microsatellite instable (MSI) present			
no: microsatellite stable	1205 (2.5)	703 (4.5)	502 (1.6)
yes: microsatellite unstable	328 (0.7)	117 (0.7)	211 (0.7)
unknown	4158 (8.7)	1355 (8.6)	2803 (8.7)
not recorded	42244 (88.1)	13537 (86.2)	28707 (89.1)
Lymphatic invasion			
not found	4213 (8.8)	1351 (8.6)	2862 (8.9)
yes	1079 (2.3)	607 (3.9)	472 (1.5)
suspect	4 (0.0)	2 (0.0)	2 (0.0)
unknown	273 (0.6)	102 (0.6)	171 (0.5)
not recorded	42366 (88.4)	13650 (86.9)	28716 (89.1)
Angio invasion			
not found	2848 (5.9)	961 (6.1)	1887 (5.9)
extramural venous invasion (EMVI)	753 (1.6)	408 (2.6)	345 (1.1)
intramural venous invasion (IMVI)	220 (0.5)	118 (0.8)	102 (0.3)
no EMVI, intramural unknown	2505 (5.2)	843 (5.4)	1662 (5.2)
not applicable	6 (0.0)	3 (0.0)	3 (0)
unknown	667 (1.4)	245 (1.6)	422 (1.3)
not recorded	40936 (85.4)	13134 (83.6)	27802 (86.3)
High risk*			
high	11532 (24.1)	1638 (10.4)	9894 (30.7)
normal	643 (1.3)	12 (0.1)	631 (2.0)
unknown	13626 (28.4)	464 (3.0)	13162 (40.8)
not applicable	22134 (46.2)	13598 (86.5)	8536 (26.5)
Maximum length of stay	9.4 (20.8)	8.3 (31.3)	9.9 (13.0)

Data is mean (sd) or numbers (percentage)

Table 2 Methods description

Methods	**Models**				
	$\mathbb{E}(Y_i\|Z_i)$	$\mathbb{E}(Y_i\|X_i,Z_i)$		$\mathbb{E}(Z_i\|X_i)$	
	LReg	LReg	BART	LReg	BART
Crude	✓	-	-	-	-
Direct/LReg	-	✓	-	-	-
Direct/BART	-	-	✓	-	-
IPW/LReg	-	-	-	✓	-
IPW/BART	-	-	-	-	✓
DR/LReg-LReg	-	✓	-	✓	-
DR/LReg-BART	-	✓	-	-	✓
DR/BART-LReg	-	-	✓	✓	-
DR/BART-BART	-	-	✓	-	✓
Direct/ps-BART	-	-	✓	-	✓

Table 3 Baseline characteristics of the study population of in the NCR after subsampling and QUASAR

Covariates	**the two-year follow-up NCR after subsampling** (n=7704)	**the five-year follow-up NCR after subsampling** (n=6378)	**QUASAR** (n=3239)
Stage			
II	7011(91%)	5803(91%)	2963(91%)
III	693(9%)	575(9%)	260(9%)
Sex			
Male	4699(61%)	3890(61%)	1979(61%)
Female	3005(39%)	2488(39%)	1260(39%)
Age			
<50	847(11%)	702(11%)	370(11%)
50-59	1996(26%))	1651(26%)	855(26%)
60-69	3236(42%)	2679(42%)	1351(42%)
>70	1625(21%)	1632(21%)	663(21%)

Table 4 Stratified sampling profile

index	**stage; sex; age**	**# people in Stratum** (N= 45442)	**P(Stratum)**	**# people in STRATUM** (N=7704)	**P(STRATUM)**
1	II; male; <50	470	0.010	470	0.061
2	II; male; 50-59	1259	0.028	1110	0.144
3	II; male; 60-69	3740	0.082	1795	0.233
4	II; male; 70+	7010	0.154	901	0.117
5	II; female; <50	476	0.010	300	0.039
6	II; female; 50-59	1163	0.026	709	0.092
7	II; female; 60-69	3036	0.067	1148	0.149
8	II; female; 70+	7148	0.157	578	0.075
9	III; male; <50	557	0.012	46	0.006
10	III; male; 50-59	1385	0.030	108	0.014
11	III; male; 60-69	3614	0.080	177	0.023
12	III; male; 70+	5407	0.119	92	0.012
13	III; female; <50	574	0.013	31	0.004
14	III; female; 50-59	1323	0.029	69	0.009
15	III; female; 60-69	2961	0.065	116	0.015
16	III; female; 70+	5319	0.117	54	0.007

P(*Stratum*): the probability of stratum in the population P with sample size 45485
P(*STRATUM*): the probability of stratum in the study population S with sample size N=7704, which is consistent with that of QUASAR
For the first *STRATUM*, $P(STRATUM) = P(stageII \cap male \cap age{<}50) = P(stageII) \times P(male) \times P(age{<}50) = 0.91 \times 0.61 \times 0.11 = 0.061$ (see Table 3 for more details of distributions of these covariates in the QUASAR).

Table 5 Results for 2-year follow-up

covariates	Methods	ATE(range)	overlap(%)
3 cov	Crude($X_{sub} \sim \mathcal{D}_{QUASAR'}$)	-0.019 (-0.041, 0.003)	99.48
	Direct/LReg	-0.029 (-0.03, -0.029)	100
	Direct/BART	-0.02 (-0.02, -0.019)	100
	IPW/LReg	-0.029 (-0.057, 0.006)	89.16
	IPW/BART	-0.034 (-0.06, 0)	76.72
	DR/LReg-LReg	-0.028 (-0.053, 0.008)	92.9
	DR/LReg-BART	-0.028 (-0.051, 0.004)	94.44
	DR/BART-LReg	-0.021 (-0.046, 0.016)	98.72
	DR/BART-BART	-0.022 (-0.045, 0.01)	99.14
	Direct/ps-BART	-0.022 (-0.022, -0.022)	100
5 cov	Crude($X_{sub} \sim \mathcal{D}_{QUASAR'}$)	-0.019 (-0.041, 0.006)	99.36
	Direct/LReg	-0.047 (-0.047, -0.046)	0
	Direct/BART	-0.034 (-0.035, -0.034)	100
	IPW/LReg	-0.04 (-0.068, -0.013)	47.54
	IPW/BART	-0.046 (-0.074, -0.022)	19.02
	DR/LReg-LReg	-0.044 (-0.068, -0.022)	21.92
	DR/LReg-BART	-0.043 (-0.066, -0.023)	25.96
	DR/BART-LReg	-0.04 (-0.064, -0.018)	44.22
	DR/BART-BART	-0.042 (-0.064, -0.022)	34.22
	Direct/ps-BART	-0.041 (-0.041, -0.04)	0
13 cov	Crude($X_{sub} \sim \mathcal{D}_{QUASAR'}$)	-0.019 (-0.041, 0.006)	99.36
	Direct/LReg	-0.055 (-0.057, -0.054)	0
	Direct/BART	-0.056 (-0.057, -0.055)	0
	IPW/LReg	-0.046 (-0.073, -0.02)	20.18
	IPW/BART	-0.053 (-0.08, -0.027)	4.24
	DR/LReg-LReg	-0.05 (-0.074, -0.029)	4.44
	DR/LReg-BART	-0.05 (-0.073, -0.031)	3.02
	DR/BART-LReg	-0.05 (-0.072, -0.03)	4.46
	DR/BART-BART	-0.052 (-0.074, -0.033)	1.14
	Direct/ps-BART	-0.051 (-0.052, -0.05)	0
all cov	Crude($X_{sub} \sim \mathcal{D}_{QUASAR'}$)	-0.019 (-0.041, 0.006)	99.36
	Direct/LReg	-0.053 (-0.055, -0.052)	0
	Direct/BART	-0.045 (-0.046, -0.044)	0
	IPW/LReg	-0.043 (-0.072, -0.017)	31.16
	IPW/BART	-0.049 (-0.075, -0.021)	11.54
	DR/LReg-LReg	-0.046 (-0.069, -0.026)	12.32
	DR/LReg-BART	-0.046 (-0.068, -0.025)	12.46
	DR/BART-LReg	-0.045 (-0.065, -0.026)	17.58
	DR/BART-BART	-0.048 (-0.068, -0.028)	7.3
	Direct/ps-BART	-0.053 (-0.055, -0.052)	0

RCT:-0.021 (95% CI -0.039, -0.002), Crude ($X_{NCR} \sim \mathcal{D}_{NCR}$): -0.069 (95% CI -0.076, -0.062)

Table 6 Results for 5-year follow-up

covariates	Methods	ATE(range)	overlap(%)
3 cov	Crude($X_{sub} \sim \mathcal{D}_{QUASAR'}$)	-0.031 (-0.065, 0.005)	99.76
	Direct/LReg	-0.035 (-0.036, -0.034)	100
	Direct/BART	-0.025 (-0.026, -0.025)	100
	IPW/LReg	-0.05 (-0.095, -0.011)	78.8
	IPW/BART	-0.06 (-0.105, -0.02)	46.7
	DR/LReg-LReg	-0.041 (-0.085, -0.002)	94.8
	DR/LReg-BART	-0.038 (-0.079, -0.003)	98.18
	DR/BART-LReg	-0.034 (-0.078, 0.005)	98.66
	DR/BART-BART	-0.032 (-0.072, 0.003)	99.54
	Direct/ps-BART	-0.021 (-0.022, -0.021)	100
5 cov	Crude($X_{sub} \sim \mathcal{D}_{QUASAR'}$)	-0.031 (-0.07, 0.008)	99.76
	Direct/LReg	-0.065 (-0.066, -0.065)	0
	Direct/BART	-0.053 (-0.053, -0.052)	100
	IPW/LReg	-0.064 (-0.103, -0.026)	33.14
	IPW/BART	-0.075 (-0.114, -0.039)	7.94
	DR/LReg-LReg	-0.065 (-0.097, -0.03)	27.14
	DR/LReg-BART	-0.061 (-0.093, -0.029)	39.66
	DR/BART-LReg	-0.06 (-0.092, -0.026)	44.62
	DR/BART-BART	-0.058 (-0.089, -0.025)	53.8
	Direct/ps-BART	-0.049 (-0.049, -0.049)	100
13 cov	Crude($X_{sub} \sim \mathcal{D}_{QUASAR'}$)	-0.031 (-0.07, 0.008)	99.76
	Direct/LReg	-0.081 (-0.082, -0.078)	0
	Direct/BART	-0.093 (-0.095, -0.092)	0
	IPW/LReg	-0.08 (-0.121, -0.043)	3.92
	IPW/BART	-0.09 (-0.129, -0.054)	0.24
	DR/LReg-LReg	-0.078 (-0.114, -0.049)	1.42
	DR/LReg-BART	-0.077 (-0.111, -0.05)	1.54
	DR/BART-LReg	-0.077 (-0.112, -0.048)	2.2
	DR/BART-BART	-0.077 (-0.111, -0.049)	1.52
	Direct/ps-BART	-0.08 (-0.081, -0.078)	0
all cov	Crude($X_{sub} \sim \mathcal{D}_{QUASAR'}$)	-0.031 (-0.07, 0.008)	99.76
	Direct/LReg	-0.078 (-0.08, -0.076)	0
	Direct/BART	-0.07 (-0.071, -0.069)	0
	IPW/LReg	-0.078 (-0.117, -0.041)	5.66
	IPW/BART	-0.086 (-0.124, -0.052)	0.76
	DR/LReg-LReg	-0.075 (-0.109, -0.046)	3.4
	DR/LReg-BART	-0.073 (-0.105, -0.045)	5.22
	DR/BART-LReg	-0.073 (-0.107, -0.044)	4.82
	DR/BART-BART	-0.073 (-0.106, -0.044)	4.34
	Direct/ps-BART	-0.076 (-0.078, -0.075)	0

RCT:-0.025 (95% CI -0.059, 0.008),Crude ($X_{NCR} \sim \mathcal{D}_{NCR}$): -0.073(95% CI -0.083, -0.063)

Appendix

Details regarding the subsampling procedure

Here we describe our stratified (sub) sampling approach to estimate the sampling distribution of $\mathcal{D}_{QUASAR'}$ by correcting for the sampling mechanisms in the NCR as introduced in section 4 in the context of two-year follow-up analysis. Our aim is by correcting for the sampling mechanism, the subsamples in NCR ($X_{sub} \sim \mathcal{D}_{QUASAR'}$) is similar with QUASAR ($X_{QUASAR} \sim \mathcal{D}_{QUASAR}$) in terms of distributions of all the known covariates.

To do so, first, due to the fact that only three covariates, namely, stage, sex, and age and only their marginals were reported in QUASAR, we can compute the joint probability of each stratum according to the marginal probabilities of these three covariates by assuming independence and the joint probability is given by $P(stage \cap sex \cap age) = P(stage) \times P(sex) \times P(age)$ which is presented in the column $P(STRATUM)$ in Table 4.

Second, subsampling the population in NCR according to the joint probability of each stratum such that the probability of each stratum in the study population after subsampling is the same as QUASAR. Here we subsample the target population for two-year follow-up analysis. Let the eligible population in NCR be *P* with sample size 45442, the study population after sampling be *S* with sample size N, the first stratum in *P* be *Stratum1* with sample size $n1$, the first stratum in the *S* be *STRATUM1* with sample size $N1$. To maximize the N of *S*, presume all the units in *Stratum1* (n1=470) are included in *S*, then the sample size of *S* should be $N = \frac{n1}{P(STRATUM1)} = \frac{470}{0.061} \approx 7704$. Based on this study population ($N = 7704$), the sample size of *STRATUM2* should be $N2 = N \times P(STRATUM2) = 1259$; likewise, the sample size of other *STRATUMs* in the *S* can be obtained in the same way. Under this scenario, all the units in each *STRATUM* in *S* can be sampled from *Stratum* of the population *P* without replacement. Next, we continue another round of stratified sampling in the same way. Different from the first round, this time we presume all the units in *Stratum2* ($n2 = 1259$) are included in the *S* and the sample size of the *S* should be $N = \frac{n2}{P(STRATUM2)} = \frac{1259}{0.144} = 8743$. However, if we use this scenario, then the sample size of the *STRATUM1* will be $N1 = N \times P(STRATUM1) = 8743 \times 0.061 = 533$, which is much larger than the sample size of *Stratum1* ($n1 = 470$). Hence we can't sample the units from *Stratum1* without replacement and we won't use this sampling scenario. Next, we replicate this sampling method for the third time, and include all the samples of *Stratum3* into *S*. Then the sample size of *S* is equal to $N = \frac{n3}{P(STRATUM3)} = \frac{3740}{0.233} = 16051$. On the basis of the sample size N of *S*, we check if we can stratified sample from each $Stratum$ of P without replacement. We replicate this sampling method in total for 16 times in which each *Stratum* is chosen in turn to generate the sample size of *S*. Finally, only one method is adopted: all the units in *Stratum1* are included into *S* to generate the sample size of *S* ($N = 7704$), and subsample from *P* according to $P(STRATUM)$ without replacement on the basis of the sample size of *S*, as details shown in Table 4. After subsampling, the distributions of the stage, sex and age of the study population *S* in the NCR are consistent with the QUASAR, as shown in Table 3 in section 4. In further analysis, to quantify the uncertainty of sampling procedure, we will subsample the *P* using this stratified sampling method for 5000 times.