Over-fitting of Propensity Score Models-does it matter ?

Wilfrid Kouokam Lowe

Over-fitting of Propensity Score Models-does it matter?

Wilfrid KOUOKAM LOWE

Master de Mathématiques
et Applications spécialité Statistique
Université de Strasbourg

From January 21st to July 21st 2013

Internship supervisor
Robert W. PLATT

Academic supervisor
Armelle GUILLOU

Host Organization
Department of Epidemiology, Biostatistics and Occupational Health
McGill University
1020 Pine Avenue West
H3A 1A2 Montreal, Quebec
Canada

Abstract

Background

Confounding is an important challenge in non-randomized studies. It is particularly present in observational studies such as pharmacoepidemiological investigations because medications are given on knowledge of the patient's condition. Among the methods available to control for confounding, propensity scores are frequently used. A propensity score represents the probability to be treated conditional on observed covariates. The high-dimensional propensity score (hd-PS) is one important application of the propensity score. It is an algorithm for semi-automated confounding control in healthcare databases. Applications of the hd-PS in literature show that generally acknowledged restrictions on the number of variables included in a prediction model are often not considered. Therefore, in many instances, results and conclusions based on over-fitted propensity score models were reported.

Objective

Our objective was first to assess the impact of over-fitting of propensity score models when estimating true underlying probabilities to receive treatment and second, how inaccuracies in these estimates translate to erroneous estimates of treatment effects. In particular it was questioned if inaccurate estimation of the propensity score due to over-fitted model leads to considerable bias or inflation of variance in estimating the treatment effect on a typically binary outcome.

Methods

Comprehensive simulation studies on the impact of over-fitting propensity score models in a logistic-regression framework were conducted.

The degree of over-fitting of a propensity score model is indicated by the ratio between the number of individuals (respectively number of treated or untreated) and the number of covariates in the considered prediction model.

Assuming a prevalence of treatment of 0.5, the number of individuals per covariate were set to $5, 10, 20, 50$ and 100. Within the simulation study, the estimated propensity score for an individual was compared with its true underlying probability to receive treatment, as given by the individual's observed covariates and predefined regression coefficients. The treatment effect was estimated by ordinary logistic regression considering deciles of the estimated propensity score as adjustment variables, matching on the propensity score as well as by marginal structural modelling with weighting by the inverse of the propensity score. The performance in estimating treatment effects was evaluated by the bias, the standard error (SE) and the root mean squared error (RMSE) of the estimator.

Results

There is a considerably imprecision in estimation of the propensity score if the ratio between the number of treated (or untreated) and the number of covariates in the propensity score is low (<=10). Over-fitting of the propensity score model revealed in no measurable bias in the treatment effect estimation. However, due to a substantially inflated variance of the estimator, a considerably high mean squared error (MSE) in estimation of treatment effects was observed when using over-fitted propensities scores for adjustment purpose.

Conclusion

Over-fitting of propensity scores should be avoided in order to facilitate reliable estimates of treatment effects. Researchers should be cautious when the number of predictors is high relative to the number of treated or untreated individuals in the propensity score model.

Résumé

Contexte

Le bias de confusion est un défi majeur dans les études non-randomisées. Il est particulièrement présent dans les études observationnelles telles que les études pharmaco-épidémiologiques (études de l'utilisation et des effets des médicaments sur une grande population) parceque les médicaments sont prescrits connaissant l'état de santé du patient. Parmis les méthodes disponibles pour controler ces facteurs de confusion, on retrouve le score de propension. Un score de propension représente la probabilité pour un individu de recevoir un traitement conditionnellement aux covariables observées. Le score de propension de grande dimension (SPGD) est une application importante du score de propension. C'est un algorithme semi-automatisé de controle de confusion des bases de données de santé. Les applications du score de propension dans la littérature montrent que généralement, les restrictions concernant le nombre de variables à inclure dans le modèle de prédiction ne sont pas souvent prises en compte. Par conséquent, dans de nombreux cas, les résultats et les conclusions fondées sur des modèles de score de propension trop ajustés ont été signalés.

Objectif

Notre objectif était d'abord d'évaluer l'impact du sur-ajustement des modèles de score de propension lors de l'estimation des vraies probabilités sous-jacentes de recevoir un traitement et ensuite comment l'inexactitude de ces estimations se traduit par des estimations erronées des effets traitements. En particulier, il est question de savoir si l'estimation inexacte du score de propension dû à un modèle sur-ajusté conduit à un biais considérable ou à une augmentation de la variance dans l'estimation de l'effet traitement sur une variable réponse typiquement binaire.

Méthode

Des études de simulations sur l'impact des modèles de score de propension sur-ajustées dans le cadre d'une régression logistique ont été realisées.

Le degré de sur-ajustement du modèle du score de propension est indiqué par le ratio entre le nombre d'individus (respectivement le nombre de patients traités ou non traités) et le nombre de covariables dans le modèle de prédiction considéré.

En supposant une prévalence du traitement égal à 0.5, le nombre d'individus par covariable était de 5, 10, 20, 50 et 100. Dans l'étude de simulation, le score de propension estimé pour un individu a été comparé à sa vraie probabilité sous-jacente de recevoir le traitement étant donné ses covariables observés et ses coefficients de régression prédéfinis. L'effet traitement a été estimée par régression logistique ordinaire considérant le score de propension rangé en déciles. L'effet traitement a aussi été estimée par un modèle structural marginal avec pondération par l'inverse du score de propension et par matching sur le score de propension estimé. Dans l'estimation des effets traitement la performance a été évaluée par le biais, l'erreur standard (SE) et l'erreur quadratique moyenne (RMSE) de l'estimateur.

Résultats

Il y'a une imprécision considérable dans l'estimation du score de propension si le rapport entre le nombre de sujets traités (ou non traités) et le nombre de covariables dans le modèle de score de propension est faible (<=10). Le sur-ajustement du modèle du score de propension a révélé un biais négligeable dans l'estimation de l'effet traitement. Toutefois, en raison d'une augmentation considérable de la variance de l'estimateur, une erreur quadratique moyenne très élevée a été observée dans l'estimation des effets traitement lors de l'utilisation des scores de propension sur-ajustés.

Conclusion

Le surajustement des scores de propension doit être évité afin de faciliter l'estimation fiable de l'effet traitement. Les chercheurs doivent faire preuve de prudence lorsque le nombre de prédicteurs est élevé par rapport au nombre de sujets traités ou non traités dans le modèle du score de propension

Contents

List of graphics

List of abbreviations

PS: Propensity Score

hd-PS: High-Dimensional Propensity Score

SPGD: “Score de Propension de Grande Dimension”

LDI: Lady Davis Institute for Medical Research

EM-algorithm: Expectation maximization algorithm

ROC curve: Receiver Operating Characteristics curve

AUC: Area Under Curve

SD: Standard Deviation

MSE: Mean Squared Error

RMSE: Root Mean Square Error

GLM: Generalized Linear Model

IPTW: Inverse Probability of Treatment Weighting

Foreword

This report was realised within the framework of my "Master de Sciences, Technologies, Santé, Mention Mathématiques et Applications Spécialité Statistique" under the supervision of Dr. Robert W. Platt and Dr. Tibor Schuster. An abstract on this topic has been selected for an oral presentation at the 34th Annual Conference of the ISCB (International Society for Clinical Biostatistics) on 25 – 29 August 2013 in Munich, Germany. A manuscript will be submitted to a peer-review by the end of August 2013.

Host organization presentation

In order to fulfill the requirements of my statistics Master program at Strasbourg University, I did my final internship (stage de fin d'études) at the Department of Epidemiology, Biostatistics and Occupational Health of McGill University. My office was located in the Centre for Clinical Epidemiology of The Lady Davis Institute for Medical Research (LDI). McGill University is located in Montreal, Quebec, Canada and was founded in 1821 by James McGill, a prominent Scottish merchant. Eight years after it was officially established, "McGill College" began holding classes in conjunction with the Montreal Medical Institution.

The LDI is the research arm of Montreal's Jewish General Hospital, a teaching hospital of McGill University. Founded in 1969, the LDI is one of Canada's leading health research institutes. Important discoveries which have contributed to the health and well-being of patients in Quebec, Canada, and around the world, have been made by LDI researchers in the areas of HIV/AIDS, aging, cancer, vascular disease, epidemiology, and psychosocial science. At the time of my internship, about 200 researchers worked at the LDI. About 20% were primarily lab-based, 15% investigated psychosocial aspects of disease or do research in epidemiology, while the majority were principally involved in clinical research and other types of investigations.

The Centre for Clinical Epidemiology and Community Studies was founded in 1991 by Drs. Lucien Abenhaim and Samuel Orkin Freedman. Under the guidance of Dr. Abenhaim (Director from 1991 to 1999) the centre has established itself as a leader in pharmacoepidemiology, randomized clinical trials, neuroepidemiology, health services research, cardiovascular epidemiology, epidemiology of thromboembolic disorders, geriatrics, and emergency medicine. Dr. Samy Suissa was appointed the Director of the Centre in June 2010.

Acknowledgments

I would like to acknowledge the following people for their support and assistance during this internship.

First, I would like to thank my immediate supervisor Dr. Robert Platt for giving me the opportunity to do an internship in his team. It was a great experience and an opportunity for me to work in Canada. It also helped me to improve my English and my interest in biostatistics and to have specific plan for my future career.

I also would like to thank my colleague Dr. Tibor Schuster who agreed to serve as one of my internship advisors and for his many contributions to my work, for his patience and for being my brave collaborator during all my internship. I also would like to thank all the people and colleagues who worked in the Centre for Clinical Epidemiology with me for openness they created an enjoyable working environment. I would like to thank Ludovic Trinquart and Raphael Porcher for their advice and their availability.

Furthermore, I would like to thank also my best friend Michelle Liendze, my roommate and my friend Vanessa Nanfah, my brother Guy Kouokam, my aunt Cecile Waffo and all my friends for their attention and encouragement during my stay in Canada. The families I most wish to thank are Guennec family, Lemoine family, Maheu family, De Laval family and Kouokam family for whom my graduate studies in France would not have been possible and who held faith in me and pushed me to succeed!

I would like to thank the Alsace Region for scholarship grant of mobility. Finally I thank all my teachers of the Department of Statistics of the University of Strasbourg for their lessons and advice during my Master. This internship would not have been possible without their lessons.

Chapter 1: Introduction

1.1 Context and objectives

The propensity score-based analysis is an advanced method of dealing with confounding in non-randomized studies. Confounding of treatment effect estimates occurs when a factor is associated with the treatment assignment and at the same time is related to outcome. The propensity score for an individual is the probability of receiving the experimental treatment (as compared to the control treatment), conditional on the individual's covariate values. The propensity score can be used to balance the covariate distributions between the interesting comparison groups of a study (treated/untreated or exposed/unexposed), and thus reduce bias potentially caused by the considered covariates (D'Agostino 1998).

Popular propensity score approaches are either adjustment for the propensity score in context of multivariable modelling, matching on the propensity score or using the inverse of an individual's propensity score as weight within the framework of marginal structural models (Austin and Mamdani 2006). Marginal structural models are a relatively new class of causal models in which the model parameters are estimated through inverse-probability-of-treatment weighting approach. These models allow for appropriate adjustment for confounding (Hernán, Brumback, and Robins 2000).

In literature, over-fitting can occur in application of the propensity score (Cepeda et al. 2003; Stürmer et al. 2005). It occurs in studies where there are too few patients relative to the number of covariates. This is a particular problem when analysing large data bases as for example when using high dimensional propensity score to automate confounding control in distributed medical product safety surveillance systems (Rassen and Schneeweiss 2012). The high dimensional propensity score (hd-PS) is an algorithm that identifies a large number of potential confounders in claims databases, eliminates covariates with very low prevalence and minimal potential for causing bias, and then uses propensity score techniques to adjust for a large number of target covariates (Schneeweiss S. 2009). There are some problems which occur with the hd-PS. Since the algorithm recodes every single count variable into three binary variables, the number of estimated parameters in the propensity score model considerably increases. The choice between selected covariates within the hd-PS in literature varies between 200 and 500 variables (Rassen et al. 2011). However, as shown in previous simulation studies (Peduzzi et al. 1996), the ratio of number of events (treated or untreated) to estimated parameters in the logistic regression model should exceed 10 in order to achieve reliable effect estimates (with acceptable standard errors). These requirements, however, were not satisfied in a number of previously published studies (Rassen et al. 2011; Rassen, Avorn, and Schneeweiss 2010; Cepeda et al. 2003).

Our objective was therefore first to assess the impact of over-fitting of propensity score models when estimating true underlying probabilities to receive treatment and second, how inaccuracies in these estimates translate to erroneous estimates of treatment effects. In particular it was questioned if inaccurate estimation of the propensity score due to over-fitted model leads to considerable bias or inflation of variance in estimating the treatment effect on a typically binary outcome.

In chapter 2, we present the propensity score and methods how this score can be used to address confounding in common treatment effect studies. In chapter 3, we describe the design of an extensive Monte-Carlo simulations study to investigate to which extend over-fitting of propensity score models leads to inaccurate estimated propensity scores and their consequences on the estimation of treatment effect estimate. In chapter 4, we describe the results of the effect of over-fitting on correct estimation of the propensity score and performance evaluation of treatment effect estimates.

Chapter 2: Propensity score

2.1 Propensity score model

Non-randomized studies on treatment or exposure effects refer to a diverse group of studies that evaluate interventions or exposures which were allocated to individuals by unknown, but certainly non-random, assignment processes. In randomized experiments, the results in the two treatment groups (treated and untreated) may often be directly compared because their units are likely to be similar, whereas in non-randomized experiments, such direct comparisons may be misleading because the units exposed to one treatment generally differ systematically from the units exposed to the other treatment (Rosenbaum 1983). Multivariable analysis is the most commonly used analytic method for dealing with confounding in non-randomized designs. Other advanced methods of dealing with confounding in non-randomized studies include propensity score analysis and instrumental variable analysis.

The use of propensity scores is a two-stage process: the development of the score model and the use of the score for adjusting for covariate differences. Propensity scores combine a large number of possible confounders into a single variable. In Epidemiology, a confounder refers to a variable that is prognostically linked to the outcome of interest and is unevenly distributed between the study groups: treated/non-treated or exposed/unexposed (figure 1). To calculate the propensity score first one has to identify those variables that influence study group membership (e.g., demographic characteristics, disease severity). The variables that influence group membership are then entered into a logistic model estimating the likelihood of being in a particular group. A logistic model will yield a propensity score for each subject ranging from 0 to 1. The score is the estimated probability of being in one of the both groups (generally it is the treated group), conditional on a weighted score of that subject's values on the set of variables used to create the score. Rosenbaum and Rubin (Rosenbaum 1983) introduced the propensity score for subject $i(i = 1, \dots, N)$ as the conditional probability of assignment to a particular treatment $(E_i = 1)$ versus control $(E_i = 0)$ given a vector of p observed covariates, $\boldsymbol{x}_i = x_{i1}, \dots, x_{ip}$:

$$e(\boldsymbol{x}_i) = pr(E_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i)$$

To understand the strength of propensity scores, one may consider two subjects with identical propensity scores, one who received the intervention and one who did not. If one assumes that the propensity score is based on all those factors that would affect the likelihood of receiving an intervention, then one can consider the assignment of the cases to be essentially random (Rubin 1997). Even if the propensity score is based on all of the factors known to influence group assignment, there remains the possibility that there are unknown confounders. If one could include these unknown confounders, they would change the propensity score. Therefore, two subjects with the same propensity score may not have the same likelihood of being in either group if the propensity score model is misspecified (D'Agostino 1998).

Since the propensity score corresponds to a conditional probability, different approaches can be used for estimation. The most commonly used statistical method for estimating the propensity score is binary logistic regression.

2.1.1 Logistic regression model

The logistic regression model is a model in which the response variable y is dichotomous and coded 0 and 1. In this model, the expected value is simply the conditional probability π that y takes the value 1 for an individual with covariate realizations $X_1 = x_1, \dots, X_p = x_p$.

This could be modelled as an ordinary linear model:

$$\pi = E\big(y|\, X_1 = x_1, \dots, X_p = x_p\big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Here it is assumed that $X_1, \dots, X_p$ are independent. As the predicted probability must satisfy $0 \leq \pi \leq 1$, the probability $\pi$is replaced by the logit transformation of the probability, $\ln(\pi/(1-\pi))$. The observed values of y do not follow a normal distribution with mean π, but rather a Bernoulli or binomial distribution. The model becomes: $logit(\pi) = ln\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ where:

- x_j is the jth explanatory variable ($j = 1, 2, \dots, p$);
- π is the estimated value of the true probability that the variable takes the value 1;
- β_0 is the estimated constant term;
- $\beta_1, \beta_2, \dots, \beta_p$ are the estimated logistic regression coefficients.

The logit of the probability is simply the log of the odds of the event of interest. In a logistic regression model, the parameter β_j associated with explanatory variable x_j is commonly interpreted after transforming to $\exp(\beta_j)$, whichis the relative change of the odds that $y = 1$ when x_j increases by 1, conditional on the other explanatory variables remaining the same. Therefore, $\exp(\beta_j)$ refers to the odds ratio.

For subject i, the estimated propensity score is given by:

$$e(\boldsymbol{x}_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})}$$

where $\hat{\beta}_0, \dots, \hat{\beta}_p$ are the estimated of the regression coefficients. When there are missing values in covariates, one can use the EM algorithm to estimate propensity scores.

2.1.2 The EM-algorithm

The EM (Expectation-Maximization) -algorithm is an alternative procedure for computing the maximum likelihood estimator when only a subset of the data is available. The first proper theoretical study of the algorithm was done by Dempster, Laird, and Rubin (A.P. Dempster N.M Laird 1977).

Let $X = (X_1, \dots, X_n)$ be a sample with conditional density $f_{X|\vartheta}(X|\theta)$ given $\vartheta = \theta$. The log-likelihood function is given by $l(\theta; X) = \log f_{X|\vartheta}(X|\theta)$. It is assumed that the data X consists of observed variables $Y = (Y_1, \dots, Y_k)$ and unobserved (missing or latent variables) $Z = (Z_1, \dots, Z_{n-k})$ so that $X = (Y, Z)$. With this notion the log-likelihood function for the observed data Y is $l_{obs}(\theta; Y) = \log \int f_{X|\vartheta}(Y, z|\theta)\nu_z(dz)$. The problem here is that when maximizing the likelihood a complete integral is required. One might not be able to find a closed form expression for $l_{obs}(\theta; Y)$. To maximize $l_{obs}(\theta; Y)$ with respect to θ the idea is to do an iterative procedure where each iteration has two steps, called the E-step and the M-step. Let $\theta^{(i)}$ denote the estimate of ϑ after the ith step. Then the two steps in the (i+1)th iteration are:

E-step: Compute $Q\left(\theta\middle|\theta^{(i)}\right) = E_{\theta^{(i)}}[l(\theta; X)|Y]$.

M-step: Maximize $Q\left(\theta\middle|\theta^{(i)}\right)$ with respect to θ and put $\theta^{(i+1)} = \text{argmax}\, Q(\theta|\theta^{(i)})$. This procedure is iterated until it converges. Given the description just showed above, a reasonable convergence test would be to check if the increase of $l(\theta; X)$ between successive iterations is smaller than some tolerance parameter, and to declare convergence if the EM-algorithm is not considerably improving $l(\theta; X)$ anymore.

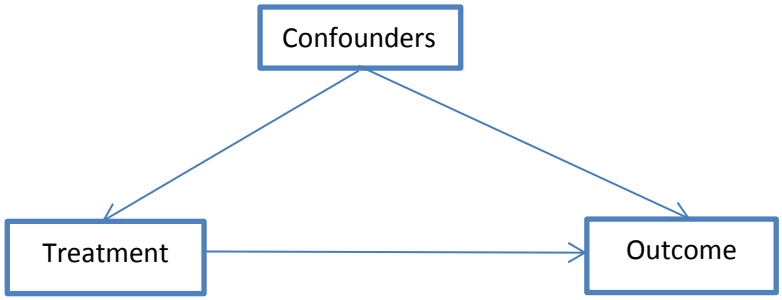
Generally, the EM-algorithm works best when the fraction of missing information is small and the dimensionality of the data is not too large.

2.1.3 Variables in the propensity score

Once the propensity score is created, it can be used to adjust for covariate differences in four different ways: matching, stratification, as a covariate in a multivariable analysis, and as a method for weighting observations (Rosenbaum 1983; Rubin 1997; D'Agostino 1998). Two important questions arise with regard to how to calculate propensity scores: what variables to include in a propensity score and how to assess the adequacy of propensity scores.

It might seem sufficient to include only those variables that are associated with both the treatment and the outcome. After all, if a variable is not a confounder then its inclusion in the propensity score is not likely to improve the adjustment for differences between the two groups (Katz 2010). Missing a confounder is a much more relevant problem than including a variable that is associated with group membership but is not with the outcome.

Figure 1: Graph representing confounding variables.



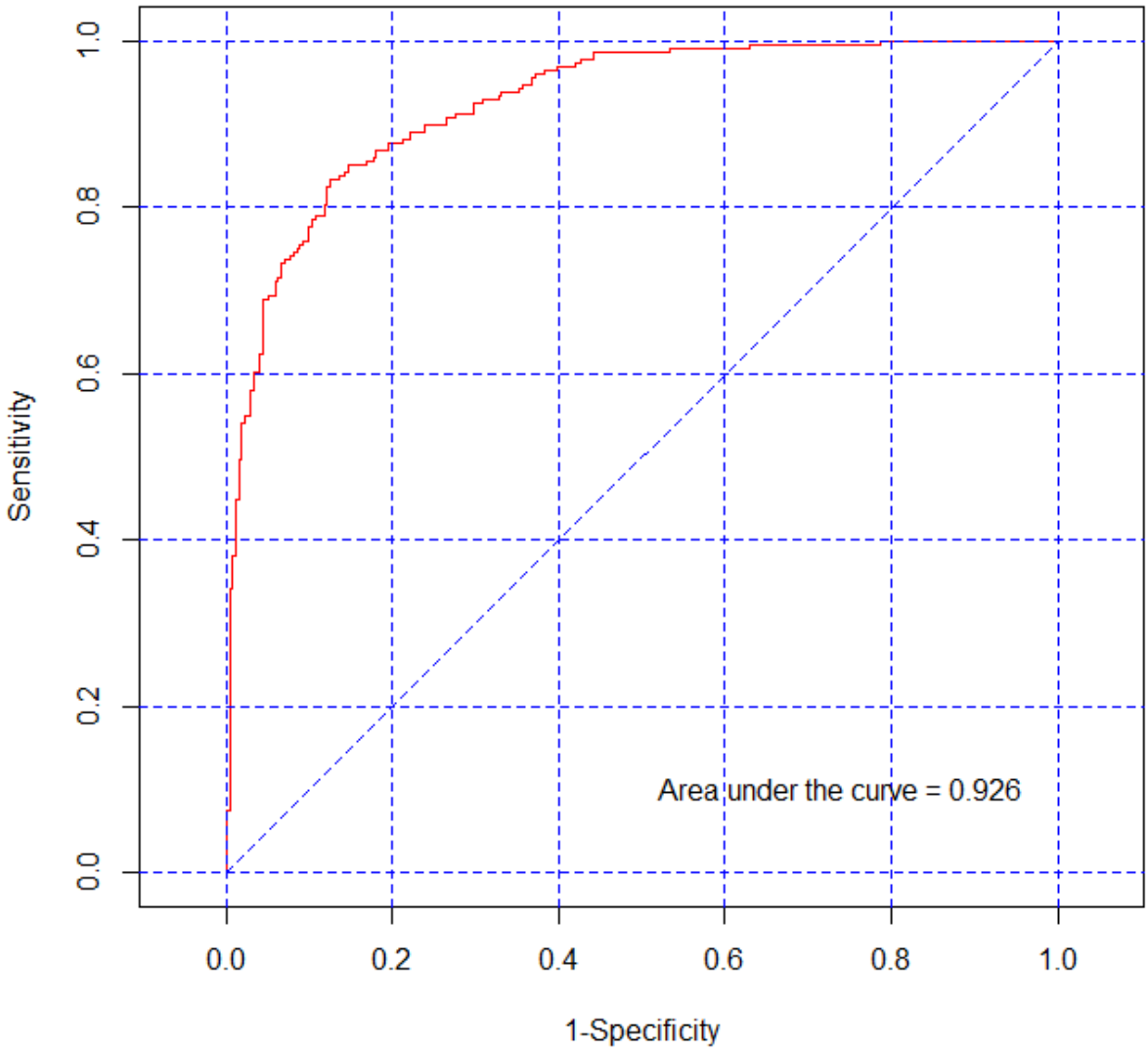
In figure 1, boxes represent variables and arrows represent directed effects. If one can remove the confounder to the treatment arrow, one can remove the effect of confounding. Patients with equivalent probabilities of treatment will not induce a confounder to treatment association.

2.1.4 Adequacy of the propensity score: the ROC curve

Once propensity score is calculated using those variables that differ between the groups, one desired property of the score is the ability of the propensity score to differentiate those who received the intervention from those who did not. To assess how well the propensity score differentiates those who receive the intervention from those who do not, one can use the area under the receiver operating characteristics (ROC) curve AUC (Petrie and Sabin 2009), which is also referred as the concordance index (c-index) (Katz 2006). An AUC value of 0.5 indicates that the model does not distinguish any better than chance. The maximum is one. The higher the c-index, the better the model is at distinguishing the two groups. A ROC curve can be constructed by plotting the sensitivity of the propensity score (in predicting who received the intervention) on the y-axis (true positive rate) and (1-specificity) on the x-axis (false positive rate).

Figure 2 shows an example of a single ROC curve from a propensity score model. In this model one have a sample with 500 individuals. There are 100 predictors and no interactions and the area under the curve represents the probability that a randomly selected treated individual will have a higher propensity score than a randomly selected untreated individual. If the propensity score was perfect, the minimum propensity score of the treated would be higher than the maximum propensity score of the untreated. If the test was worthless, giving no better prediction than a random classification, one would expect that half of the treated would have a higher propensity score value than half of the untreated (and vice versa), as indicated by an AUC value of 0.5.

Figure 2: An example of a ROC curve.



2.2 Propensity score confounder adjustment methods

In the following, four common propensity score methods for confounder adjustment are described.

2.2.1 Matching using propensity score

The concept of propensity score matching was first introduced by Rosenbaum and Rubin (Rosenbaum 1983). Propensity score matching refers to the pairing of treatment and controls units with similar values on the propensity score, and possibly other covariates, and the discarding of all unmatched units (Rubin 2001). It is primarily used to compare two groups of subjects but can be applied to analyses of more than two groups. The most common implementation of propensity score matching is pair matching or 1:1 matching in which matched pairs of treated and untreated subjects are formed (Austin 2012). The major advantage of matching on propensity scores is that compared to other propensity score methods it generally provides the least biased estimates of treatment effect (Austin 2007). A disadvantage of matching by propensity score is that the matched sample may be smaller and not representative of the study population, thereby decreasing generalizability of results.

2.2.2 Stratifying by propensity score

Since it is not always possible to find (near) exact matches on propensity score between cases and controls, one may instead stratify cases and controls by propensity scores. This technique consists of grouping subjects into strata determined by similar propensity score. Once the strata are defined, treated and non-treated subjects who are in the same stratum are compared directly. Cochran notes that as the number of covariates increases, the number of strata grows exponentially (Cochran 1965). For instance, if all covariates were dichotomous categorical variables, then there would be 2^k subclasses for $k$covariates. In terms of how many strata to create, five strata have been shown to eliminate about 90% of the bias in unadjusted analyses (Rubin 1997). An analysis comparing five strata to three or 10 equal-sized strata found similar results (Landrum M.B. 2001). A weakness of stratification is that you may have residual confounding. In other words, within each stratum of propensity score there may be important differences on baseline characteristics.

2.2.3 Covariate adjustment using propensity score

Rather than using the propensity score to match cases or to create strata the score can be entered as an independent variable in a multivariable model. In other words, the propensity score would be an independent variable for which the value for each subject would be that subject's propensity score. Rosenbaum and Rubin proposed covariate adjustment using the propensity score in the context of estimating linear treatment effects for continuous data (Rosenbaum 1983). In our study, we used this approach to estimate the regression coefficient associated with the treatment.

2.2.4 Inverse probability of treatment weighting (IPTW)

Another way to adjust for baseline differences is to use propensity score to weight the multivariable analysis. This strategy uses the propensity score to assign individual weights to all observations resulting in an altered composition of the study population (Robins, Hernán, and Brumback 2000). The weight of a person receiving the intervention is the inverse of the propensity score; the weight for the control is the

inverse of one minus the propensity score. Let E denote the treatment assignment ($E = 1$ denoting treatment, $E = 0$ denoting no treatment). For a given propensity score e the weights are defined to be $w = \frac{E}{e} + \frac{1-E}{1-e}$. For each subject, it is equal to the inverse of the probability of receiving the treatment that the subject received. Some analysts have found that using propensity scores to weight observations is less subject to misspecification of the regression models used to predict average causal effect than using propensity-based stratification (Lunceford and Davidian 2004). A problem with weighting observations is that when the estimated propensity score is near zero or one, the weights may be extreme and unrealistic (Rubin D.N 2001).

Chapter 3: Methods

Within the internship project, extensive Monte Carlo simulation studies were conducted to investigate the impact of over-fitting on different criteria i.e. accuracy in estimation of latent treatment allocation probabilities and impact on estimation of treatment effects. The criteria on which the performance is evaluated were bias, standard error (SE) and root mean squared error (RMSE) of the respective estimator.

All analyses were performed by using R software (R version 2.15.2 (2012-10-26), R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.)

3.1 Effect of over-fitting on correct estimation of the propensity score

The data simulation was conducted as follows: A total of 100 baseline covariates $X_1, \dots, X_{100}$ were generated assuming to be independently normally distributed with mean 0. The variance of the normal distributed covariates was specified in a way, that extreme propensity score values (below 0.05 or above 0.95), would not exceed a prevalence of 5%. To accomplish this, the corresponding regression coefficients β_i were sampled from a beta distribution with the two shape parameters a=0.5 and b=2. The sign of each regression coefficient was assigned by random sampling from a Bernoulli distribution with parameter p=0.5.This configuration was chosen in order to facilitate simulation of realistic treatment prediction scenarios which do not reveal too many strong predictor variables potentially leading to extreme propensity score values close to 0 or 1.

A logistic regression model was used to specify the true probability of receiving treatment P_E for the simulated values of the linear predictor:

$$P_E = \frac{\exp(\beta_0+\beta_1 X_1+\cdots+\beta_{100}X_{100})}{1+\exp(\beta_0+\beta_1 X_1+\cdots+\beta_{100}X_{100})}$$

The intercept β_0 was chosen to be 0 in order to achieve stochastically balanced numbers of exposed and unexposed individuals in the simulation. For each individual, the exposure status E was sampled from a Binomial distribution with parameter P_E. In order to initiate different magnitudes of over-fitted propensity score models, the number of individuals per covariate were set to $5, 10, 20, 50$ and 100. Within each of these datasets, the sample-specific propensity score $\hat{P}_{E,n}$ was estimated via logistic regression on the binary outcome Y_n including all 100 covariates. This procedure, data simulation, sampling and propensity score estimation, was repeated 1000 times for each of the five sample size scenarios.

3.2 Over-fitting and related bias, variance and MSE of treatment effect estimates

In order to evaluate the effect of over-fitting on bias and variance of treatment effect estimates, we considered the common case of a binary outcome variable Y. The effect of treatment on Y was specified by different log odds ratio values $\beta_E \in \{0,0.1,0.25\}$. The outcome variable Y was sampled from a binomial distribution with parameter P_Y derived based on another logistic regression model containing the simulated covariates and the previously generated exposure variable:

$$P_Y = \frac{\exp(\beta_0^* + \beta_1^* X_1 + \cdots + \beta_{100}^* X_{100} + \beta_E E)}{1 + \exp(\beta_0^* + \beta_1^* X_1 + \cdots + \beta_{100}^* X_{100} + \beta_E E)}$$

Likewise the treatment predicting effects, the regression coefficients β_i^* were sampled from a beta distribution with the two shape parameters a=0.5 and b=2. The sign of each regression coefficient was again assigned by random sampling from a Bernoulli distribution with parameter p=0.5. Here again, the intercept was chosen to be 0 leading to an expected outcome prevalence of 0.5. By choosing this simulation setting we assured that predictor variables represent real confounders i.e. being associated with both the treatment and the outcome.

To estimate the interesting treatment effect β_E, three approaches were considered: A) multivariable logistic regression model (GLM) considering deciles of the estimated propensity score as adjustment variable next to the treatment variable (we used propensity score deciles because it is recommended in literature (Rassen et al. 2011)), B) weighted logistic regression model for the estimation of the marginal exposure effect using the estimated propensity score to generate individual weights $w = \frac{E}{\hat{P}_{E,n}} + \frac{1-E}{1-\hat{P}_{E,n}}$ (inverse probability of treatment weighting: IPTW) and C) matching on the estimated propensity score. In the case of multivariable logistic regression model, the dependent variable was the outcome, and the independent variables were the exposure variable and the propensity score recoded into deciles. For the weighted logistic regression model, the dependent variable was the outcome and the independent variable was the exposure. In the case of matching, one used a greedy-matching algorithm to match subjects using caliper that were defined to have a maximum width of 0.2 standard deviations of the estimated propensity score. The caliper wide was chosen according to recommendations in literature (Austin and Mamdani 2006). For each generated data sample in the simulation, the above two logistic regression models (GLM and IPTW) were fitted, matching was applied and the resulting effect estimate $\hat{\beta}_E$ were retrieved. The bias, defined as the expected value of the difference between the effect estimate $\hat{\beta}_E$ and the true effect β_E, was calculated by taking the mean of this difference over all 1000 simulations runs. The empirical standard error of the estimator was calculated by calculating the standard deviation of $\hat{\beta}_E$ over all 1000 simulations runs. To evaluate the precision of effect estimates, we estimated the standard error using the square root of the sample variance of the treatment effect estimates across all 1000 simulation runs. The root mean square error (RMSE) was calculated by taking the square root of the mean of the squared errors over all the simulation runs. It was calculated to summarize estimation performance.

3.3 Truncation

As we use propensity score to define individual weights within a weighted multivariable analysis, extreme weights may be generated when propensity score is near zero or one. To overcome the issue of instable weights, we followed to approaches:

A) Elimination of extreme values of the estimated propensity score by quantile truncation of the tails of the observed propensity score distribution.

B) Elimination of extreme values of the estimated propensity score by restricting the scale of the propensity score to less extreme values than 0 or 1. All observed propensity score values which exceed the respective restriction criteria, were truncated.

In general, it is expected that a higher truncation level leads to a decrease in variance of the estimator but on the other hand to an increment of bias. Therefore, in order to find the best threshold for truncation which minimizes the MSE of the IPTW estimator, six levels of truncation were considered when applying truncation approach A): $0.01, 0.02, 0.05, 0.075, 0.10$ and 0.15

For truncation approach B) the following min/max values for the propensity score were considered as truncation boundaries: $0.01/0.99$, $0.02/0.98$, $0.05/0.95$, $0.075/0.925$, $0.1/0.9$, $0.15/0.85$, $0.2/0.8$ and $0.25/0.75$.

Chapter 4: Results

4.1 Effect of over-fitting on correct estimation of the propensity score

Figures 3 and 4 represent the effect of over-fitting on correct estimation of the propensity score. They show how estimation of propensity score is influenced by over-fitting. Figure 3 represents the density curves of the true underlying probability of receiving treatment P_E and the estimated propensity score $\hat{P}_{E,n}$ for different degrees of over-fitting. As can be seen from the figure 3, if the number of individuals per covariate is low (<=20, corresponding to a number of treated <=10), the density curve of the estimated propensity score deviates considerably from the curve of the true underlying probability of receiving treatment. In particular, when the number of individuals per covariate equals 5, there were many individuals who got assigned a propensity score close to 0 or 1.

Figure 3: Density curves of the true underlying probability of receiving treatment P_E and the estimated propensity score $\hat{P}_{E,n}$ for different degrees of over-fitting.

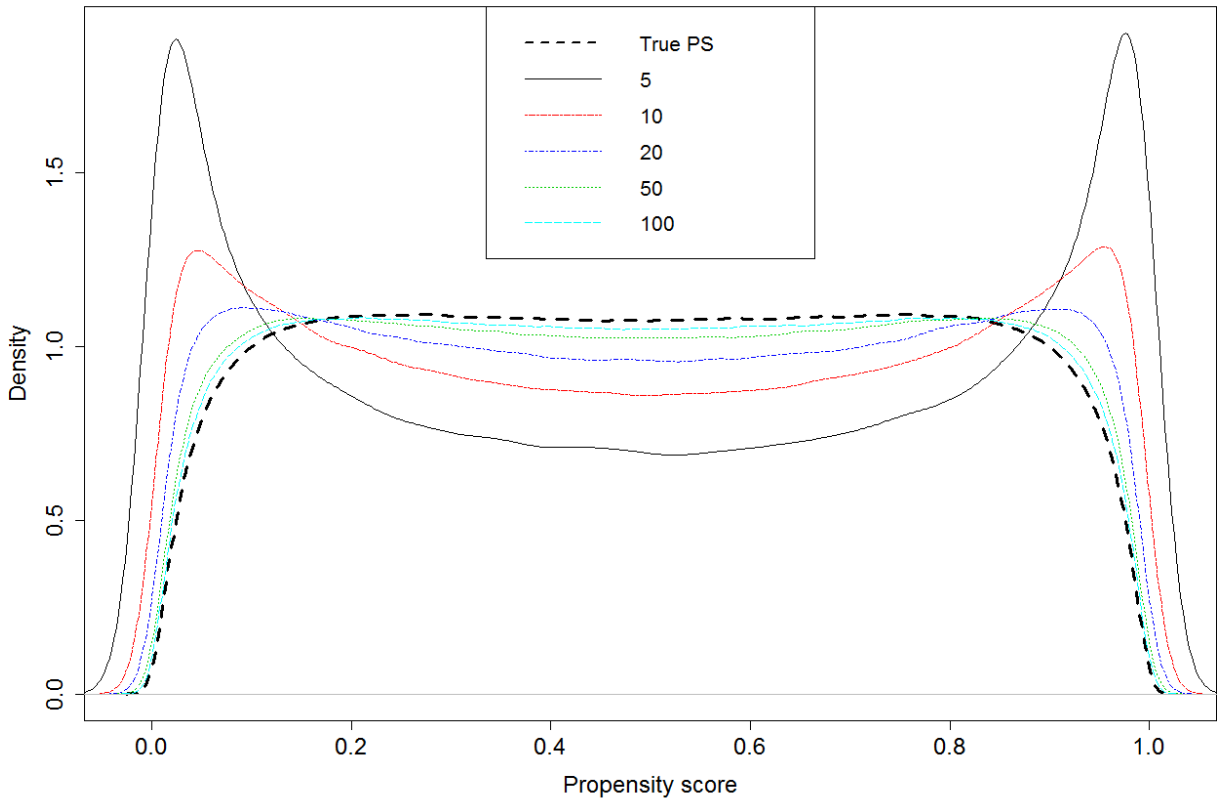


Figure 4 shows the agreement of estimated propensity score and true propensity score conditional to different degrees of over-fitting.

Figure 4: Agreement of the estimated propensity score and true propensity score conditional to different degrees of over-fitting.

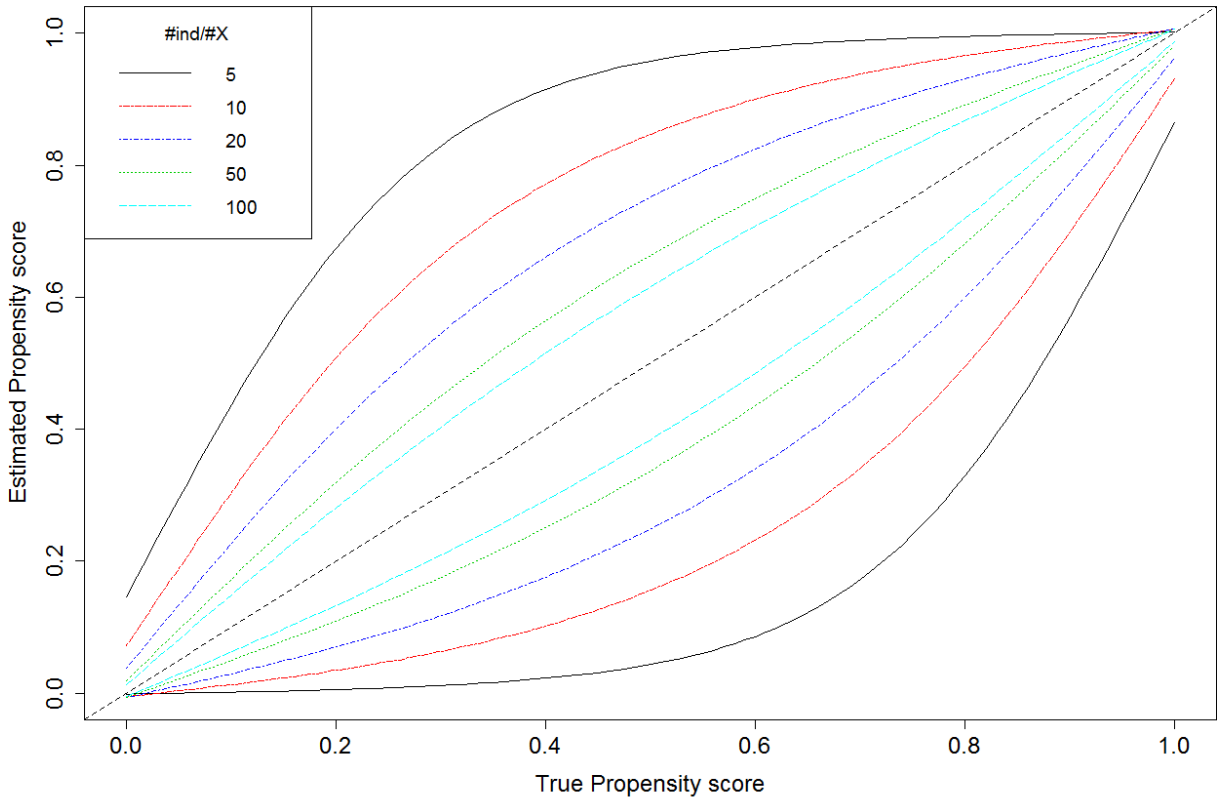
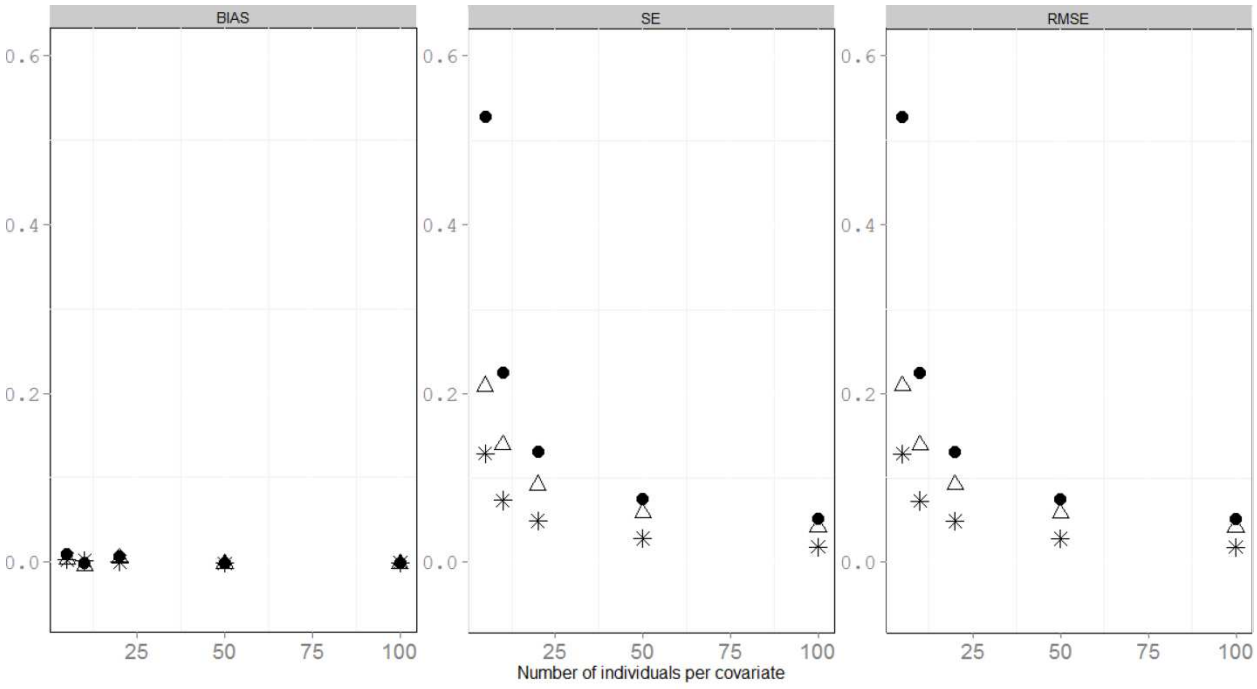


Figure 4 displays for each number of individuals per covariate (degree of over-fitting) and for a given value of the true propensity score the 95% distribution interval of the respectively estimated propensity scores. These intervals were obtained by the 0.025 and 0.975 quantiles of the distribution of estimated propensity scores over all 1000 simulation runs. When the number of individuals per covariate in the propensity score is low (<=20, number of treated <=10), the 95% distribution intervals are large. The dashed black diagonal line represents the ideal value (perfect agreement) of the estimated propensity score for a given value of the true propensity score.

4.2 Over-fitting and related bias, variance and MSE of treatment effect estimates

Figure 5 represents the effect of over-fitting on a parameter estimation. It shows bias, root mean squared error and standard error in estimating the treatment effect for each number of individuals per covariate resulting from both IPTW and ordinary multivariable adjustment approaches and from matching. The black dots represent the estimate of the treatment effect by the IPTW method, the triangles are for the GLM approach and the stars are for matching. One can see that the bias is negligible whatever the degree of over-fitting. On the other hand, when the number of individuals per covariate is low (<=20, number of treated <= 10) the SE increases considerably. As the bias is close to 0, the RMSE and the SE have approximately the same values.

Figure 5: Bias, SE and RMSE of treatment effect estimated by both GLM and IPTW methods and by matching, for $\beta_E = 0$.



4.3 Truncation

Figure 6 shows bias, SE and RMSE of treatment effect estimated by IPTW method using different number of individuals per covariate and for different level of truncation boundaries (truncation approach A). This figure is drawn only for $\beta_E = 0$. One can see that the bias is negligible so the SE and the RMSE have the same value. The dataset which correspond to the number of individuals per

covariate 5, was excluded because the model did not converge. As expected, with increasing number of individuals per covariate, the standard error of the estimator decreased.

Figure 6: Bias, SE and RMSE of treatment effect estimated by IPTW method using different number of individuals per covariate and for different level of truncation (approach A)

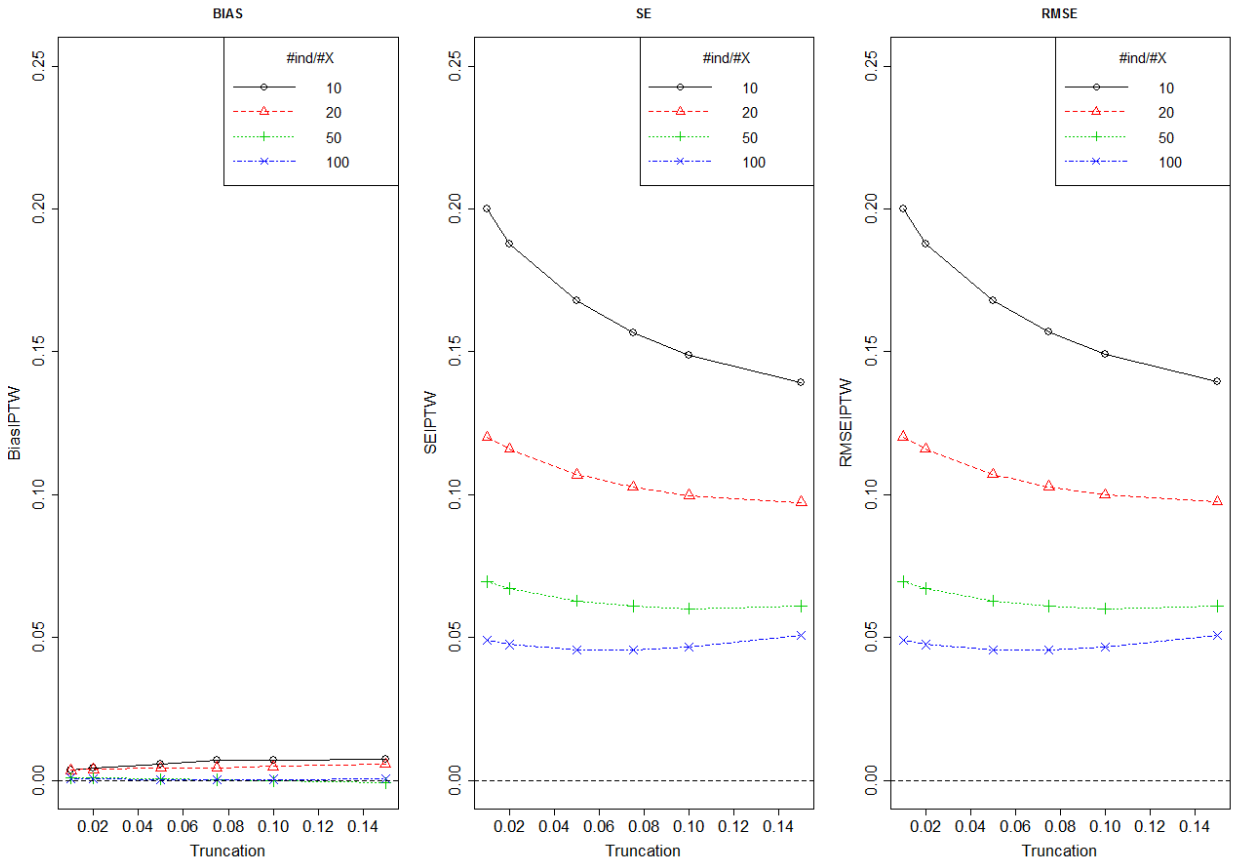


Figure 7 shows bias, SE and RMSE of treatment effect estimated by GLM method using different number of individuals per covariate and for different level of truncation (truncation approach A). This figure is drawn only for $\beta_E = 0$. One can see that the bias is negligible so the SE and the RMSE have the same value. Here again, the dataset which correspond to the number of individuals per covariate 5, was excluded because the model did not converge. The bias, the SE and the RMSE of the treatment effect estimated by the GLM approach are constant whatever the level of truncation.

Figure 7: Bias, SE and RMSE of treatment effect estimated by GLM method using different number of individuals per covariate and for different level of truncation (approach A)

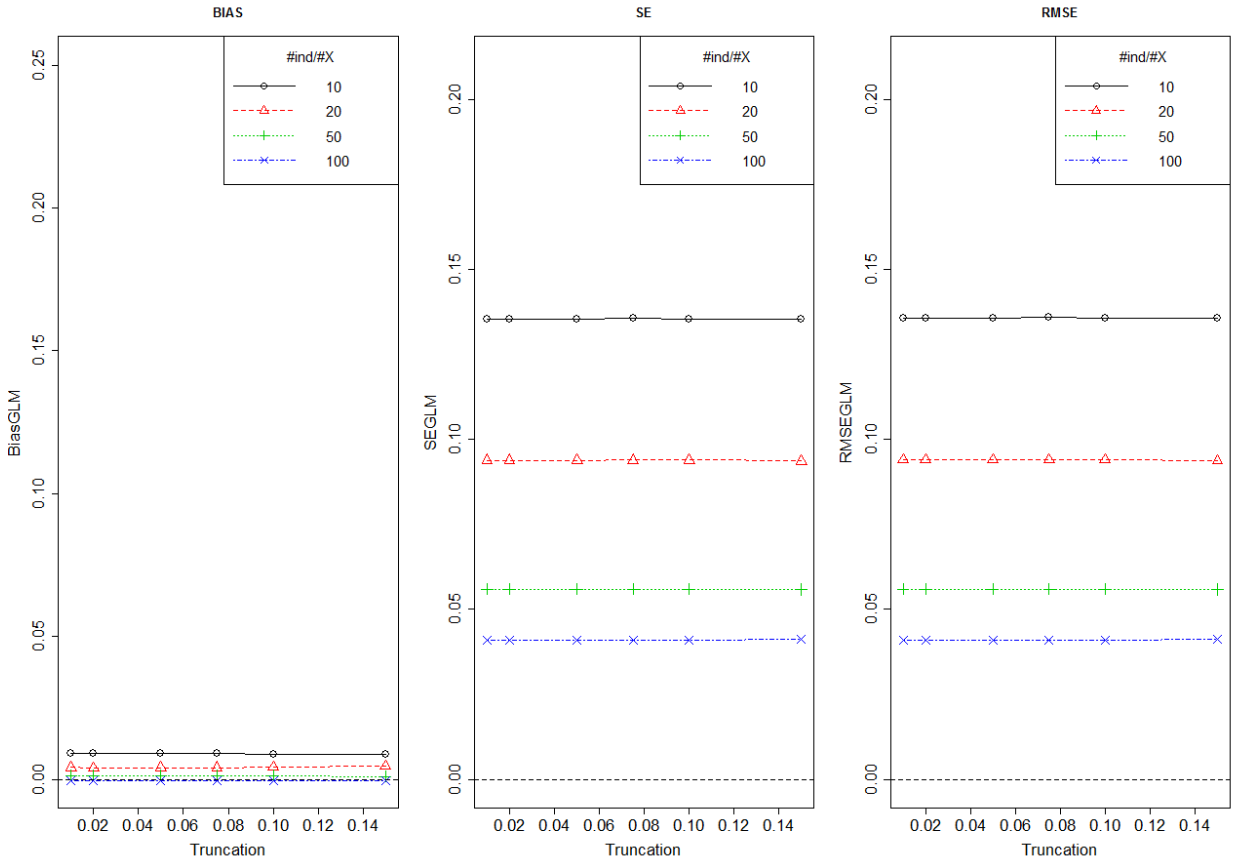


Figure 8 shows bias, SE and RMSE of treatment effect estimated by IPTW method using different number of individuals per covariate and for different level of truncation boundaries (truncation approach B). This figure is drawn only for $\beta_E = 0$. One can see that the bias is negligible so the SE and the RMSE have the same value. The dataset which correspond to the number of individuals per covariate 5, was excluded because the model did not converge. As expected, with increasing number of individuals per covariate, the standard error of the estimator decreased. The MSE curves show an approximately quadratic shape with increasing level of truncation. The minimum MSE occurred for truncation levels between 0.05 and 0.10 suggesting the best possible stabilisation of weights by truncation.

Figure 8: Bias, SE and RMSE of treatment effect estimated by IPTW method using different number of individuals per covariate and for different level of truncation (approach B)

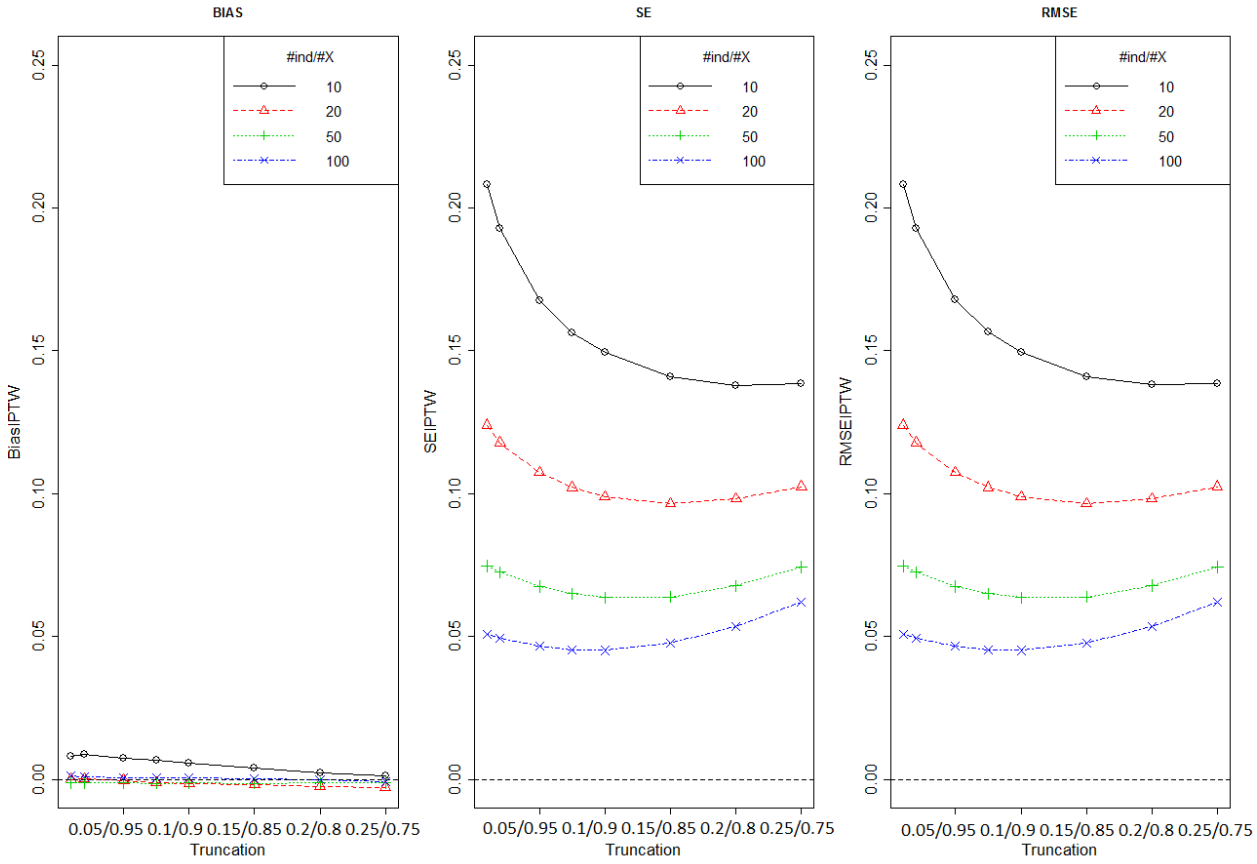
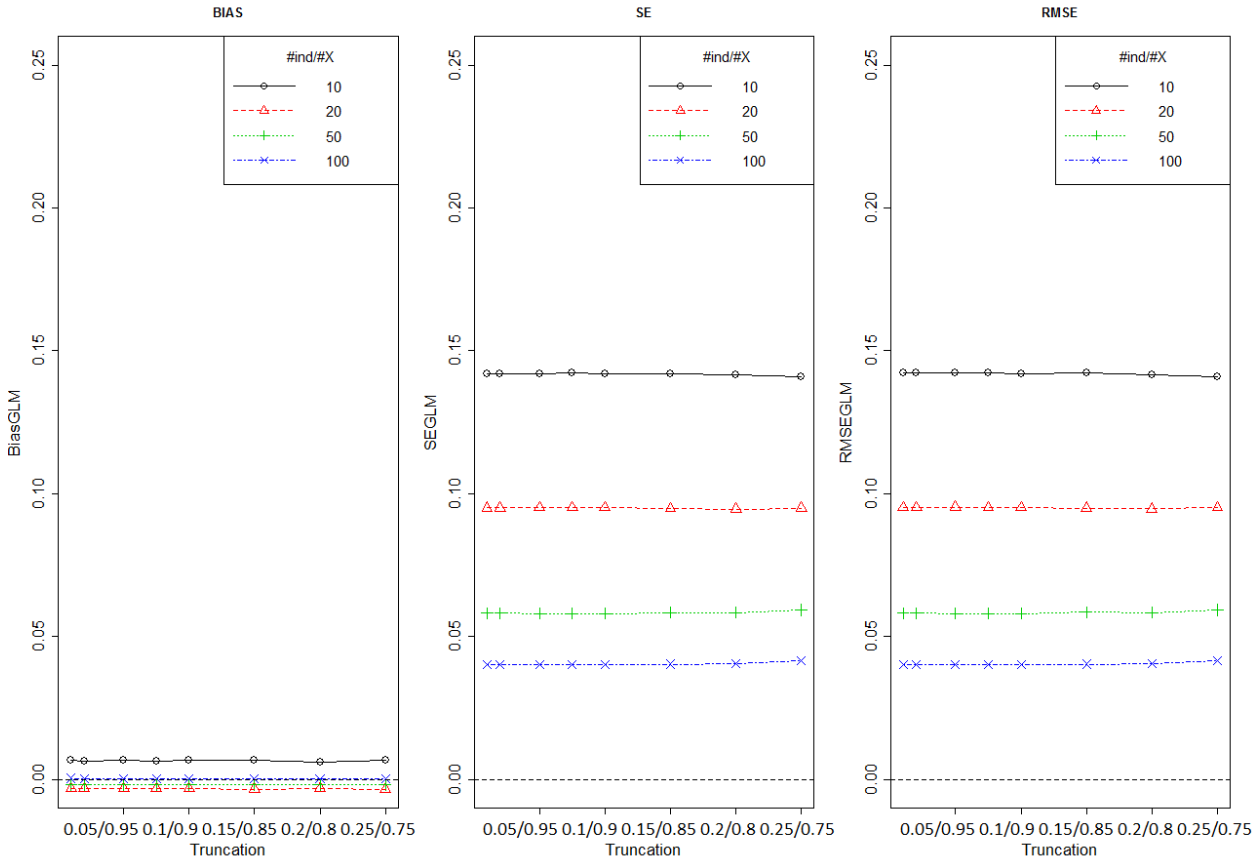


Figure 9 shows bias, SE and RMSE of treatment effect estimated by GLM method using different number of individuals per covariate and for different level of truncation (truncation approach B). This figure is drawn only for $\beta_E = 0$. One can see that the bias is negligible so the SE and the RMSE have the same value. Here again, the dataset which correspond to the number of individuals per covariate 5, was excluded because the model did not converge. The bias, the SE and the RMSE of the treatment effect estimated by the GLM approach are constant whatever the level of truncation.

Figure 9: Bias, SD and RMSE of treatment effect estimated by GLM method using different number of individuals per covariate and for different level of truncation (approach B)



Chapter 5: Discussion

5.1 Key results and Limitations

We conducted an extensive series of Monte Carlo simulations first, to assess the impact of over-fitting of propensity score models when estimating true underlying probabilities to receive treatment and second, to see how inaccuracies in these estimates translate to erroneous estimates of treatment effects. There is considerable imprecision in estimation of the propensity score if the ratio between the number of individuals and the number of covariates in the propensity score is low (<=20, number of treated individuals per covariate <=10).

There are certain limitations to our study. First, our findings were based on simulations which can never capture the whole space of possible data scenarios as they occur in reality. However, since we were mainly interested in the statistical properties of different estimators for the propensity score and treatment impact on an outcome, real data application would have been not useful due to a lack of knowledge of the respective true underlying effects. In our study we only investigated truncation as one possibility to address unstable weights as induced by extreme propensity score values. An other possible option would have been the utilization of stabilized weights. For this purpose the numerator of the weights can be modified, which decreases variability (Robins, Hernán, and Brumback 2000). Stabilized weights might reduce the estimation problem induced by over-fitted propensity score models. In our study, the number of covariate is fixed and the number of individuals varies and it has a direct influence on the RMSE. It would be a good idea to fix the number of individuals and to vary the number of covariates.

5.2 Conclusion and perspectives

The simulations demonstrate the problems that can occur when the propensity score model contains few individuals relative to the number of independent variables being evaluated. They revealed that over-fitting of propensity scores should be avoided in order to facilitate reliable estimates of treatment effects. While methods exist to address this, e.g., truncation or stabilization, researchers should be cautious when the number of predictors is high relative to the number of treated or untreated individuals in the propensity score model. According to our study, we can suggest that at least 25 treated or untreated individuals per covariate in the propensity score model are desirable to maintain valid propensity score estimates and to avoid overly inflated standard errors of treatment effect estimates.

References

A.P. Dempster N.M Laird. 1977. “Maximum Likelihood from Incomplete Data via the EM Algorithm.” *Journal of the Royal Statistical Society* 39. B.

Austin, Peter C. 2007. “The Performance of Different Propensity Score Methods for Estimating Marginal Odds Ratios.” *Statistics in Medicine* 26 (16) (July 20): 3078–3094. doi:10.1002/sim.2781.

———. 2012. “The Performance of Different Propensity Score Methods for Estimating Marginal Hazard Ratios.” *Statistics in Medicine* (December 12). doi:10.1002/sim.5705.

Austin, Peter C, and Muhammad M Mamdani. 2006. “A Comparison of Propensity Score Methods: a Case-study Estimating the Effectiveness of post-AMI Statin Use.” *Statistics in Medicine* 25 (12) (June 30): 2084–2106. doi:10.1002/sim.2328.

Cepeda, M Soledad, Ray Boston, John T Farrar, and Brian L Strom. 2003. “Comparison of Logistic Regression Versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders.” *American Journal of Epidemiology* 158 (3) (August 1): 280–287.

Cochran. 1965. “The Planning of Observational Studies of Human Populations.”

D’Agostino, R B, Jr. 1998. “Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-randomized Control Group.” *Statistics in Medicine* 17 (19) (October 15): 2265–2281.

Hernán, M A, B Brumback, and J M Robins. 2000. “Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-positive Men.” *Epidemiology (Cambridge, Mass.)* 11 (5) (September): 561–570.

Katz, Mitchell H. 2006. *Multivariable Analysis: a Practical Guide for Clinicians*. 2nd ed. Cambridge ; New York: Cambridge University Press.

———. 2010. *Evaluating Clinical and Public Health Interventions: a Practical Guide to Study Design and Statistics*. Cambridge: New York : Cambridge University Press.

Landrum M.B. 2001. “Causal Effect of Ambulatory Specialty Care on Mortality Following Myocardial Infarction: A Comparison of Propensity and Instrumental Variable Analyses.”

Lunceford, Jared K, and Marie Davidian. 2004. “Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: a Comparative Study.” *Statistics in Medicine* 23 (19) (October 15): 2937–2960. doi:10.1002/sim.1903.

Peduzzi, P, J Concato, E Kemper, T R Holford, and A R Feinstein. 1996. “A Simulation Study of the Number of Events Per Variable in Logistic Regression Analysis.” *Journal of Clinical Epidemiology* 49 (12) (December): 1373–1379.

Petrie, Aviva, and Caroline Sabin. 2009. *Medical Statistics at a Glance*. Chichester, UK; Hoboken, NJ: Wiley-Blackwell.

Rassen, Jeremy A, Jerry Avorn, and Sebastian Schneeweiss. 2010. “Multivariate-adjusted Pharmacoepidemiologic Analyses of Confidential Information Pooled from Multiple Health Care Utilization Databases.” *Pharmacoepidemiology and Drug Safety* 19 (8) (August): 848–857. doi:10.1002/pds.1867.

Rassen, Jeremy A, Robert J Glynn, M Alan Brookhart, and Sebastian Schneeweiss. 2011. “Covariate Selection in High-dimensional Propensity Score Analyses of Treatment Effects in Small Samples.” *American Journal of Epidemiology* 173 (12) (June 15): 1404–1413. doi:10.1093/aje/kwr001.

Rassen, Jeremy A, and Sebastian Schneeweiss. 2012. “Using High-dimensional Propensity Scores to Automate Confounding Control in a Distributed Medical Product Safety Surveillance System.” *Pharmacoepidemiology and Drug Safety* 21 Suppl 1 (January): 41–49. doi:10.1002/pds.2328.

Robins, J M, M A Hernán, and B Brumback. 2000. “Marginal Structural Models and Causal Inference in Epidemiology.” *Epidemiology (Cambridge, Mass.)* 11 (5) (September): 550–560.

Rosenbaum, Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70 (April).

Rubin, D B. 1997. “Estimating Causal Effects from Large Data Sets Using Propensity Scores.” *Annals of Internal Medicine* 127 (8 Pt 2) (October 15): 757–763.

Rubin D.N. 2001. “Using Propensity Scores to Help Design Observationnal Studies: Application to the Tobacco Litigation.”

Schneeweiss S. 2009. “High-dimensional Propensity Score Adjustment in Studies of Treatment Effetcs Using Health Care Claims Data.”

Stürmer, Til, Sebastian Schneeweiss, M Alan Brookhart, Kenneth J Rothman, Jerry Avorn, and Robert J Glynn. 2005. “Analytic Strategies to Adjust Confounding Using Exposure Propensity Scores and Disease Risk Scores: Nonsteroidal Antiinflammatory Drugs and Short-term Mortality in the Elderly.” *American Journal of Epidemiology* 161 (9) (May 1): 891–898. doi:10.1093/aje/kwi106.

Annexes

Annexe 1: Simulations when the regression coefficients were on the joined interval $[-0.5; -0.1] \cup [+0.1; +0.5]$

We simulated data for a setting in which there were 100 baseline covariates $X_1, \ldots, X_{100}$ which were independently standard normal distributed. In order to initiate real confounding situations, the regression coefficients β_i were sampled from a uniform distribution on the joined interval $[-0.5; -0.1] \cup [+0.1; +0.5]$. We allowed covariates to have a strong effect of treatment selection or outcome. Confounding effects are small if the regression coefficients are close to 0. We use a logistic regression model to specify the true probability of receiving treatment P_E for the simulated values of the linear predictor:

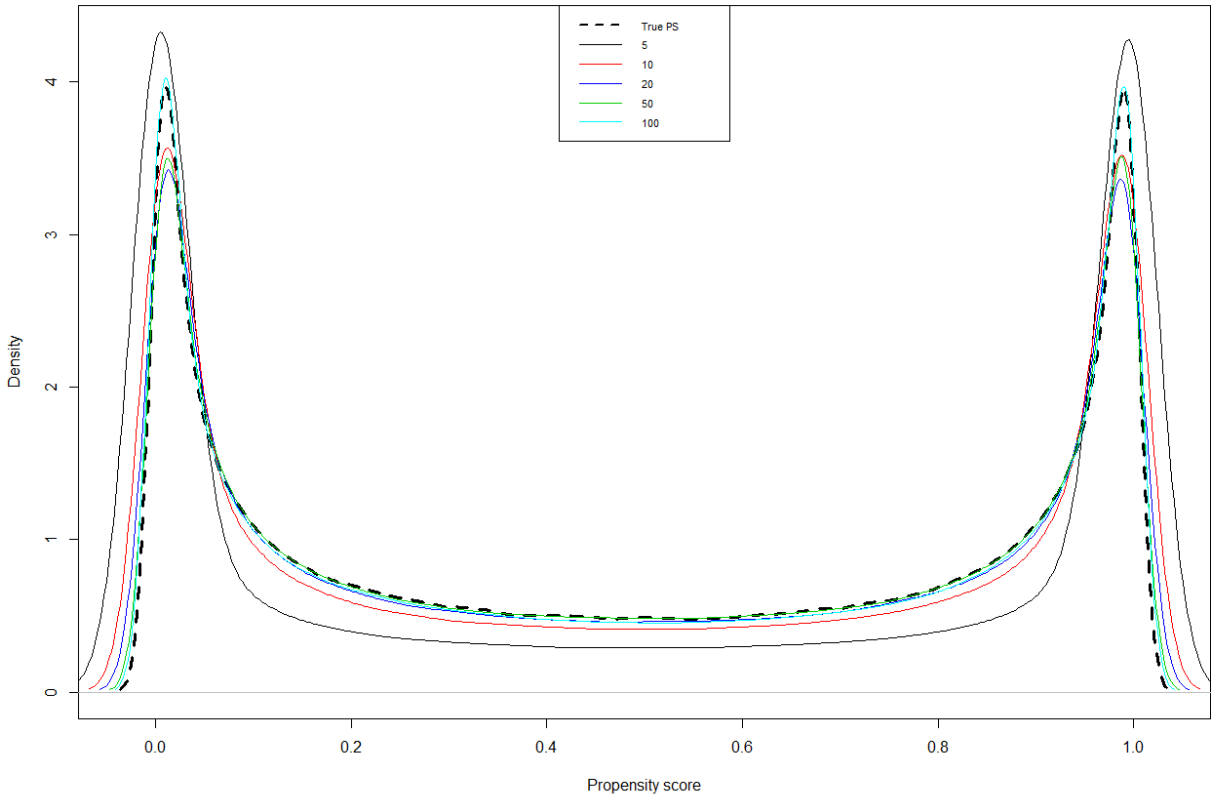
$$P_E = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{100} X_{100})}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{100} X_{100})}$$

The intercept β_0 was chosen to be 0 in order to achieve stochastically balanced numbers of exposed and unexposed individuals in the simulation. The prevalence will be 50 percent because X_i has expectation to be 0. For each individual, the exposure status E was sampled from a Binomial distribution with parameter P_E. E is a function of the X_i (when we estimate the propensity score with a logistic regression, the outcome variable is the exposure E). In order to initiate different magnitudes of over-fitted propensity score models, the number of individuals per covariate were set to $5, 10, 20, 50$ and 100. Within each of these datasets, the sample-specific propensity score $\hat{P}_{E,n}$ was estimated via logistic regression on the binary outcome Y_n including all 100 covariates. This procedure, data simulation, sampling and propensity score estimation, was repeated 1000 times for each of the five sample size scenarios.

Annexe 2: Results when the regression coefficients were on the joined interval $[-0.5; -0.1] \cup [+0.1; +0.5]$

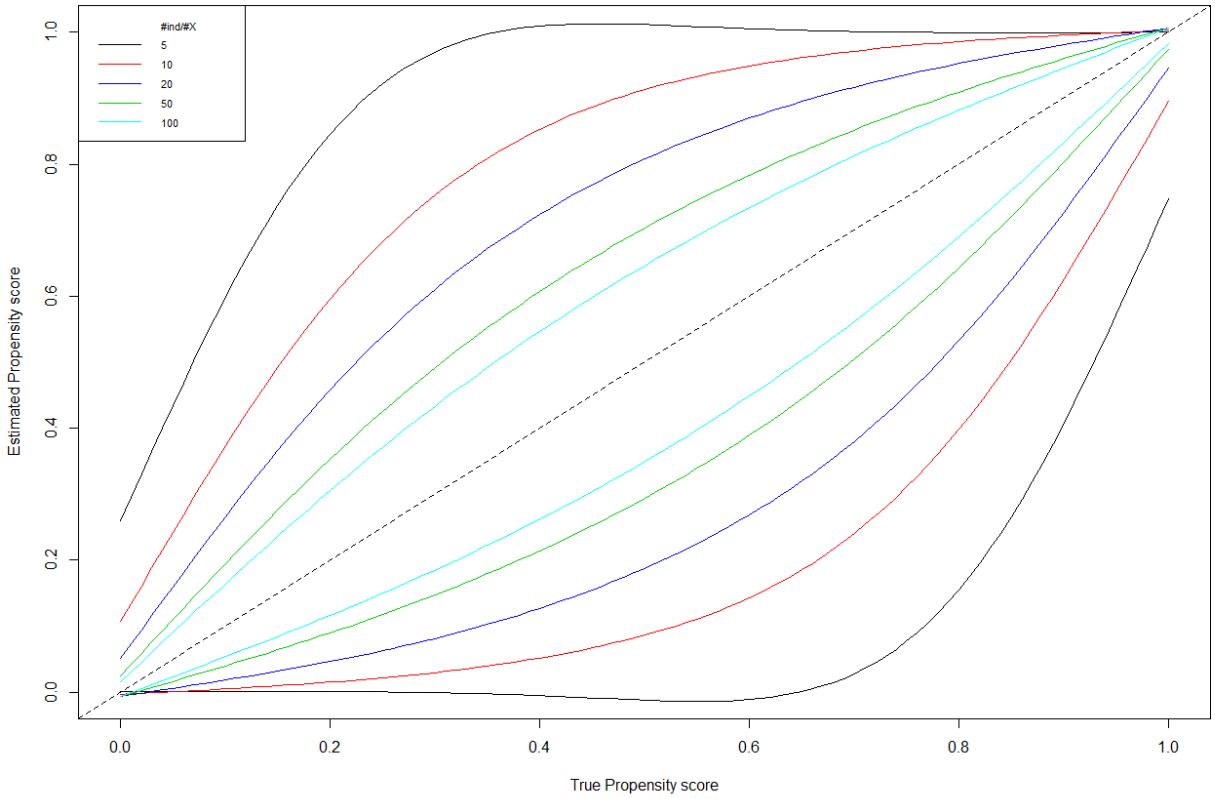
Figure 10 shows the density curves of the true propensity score and estimated propensity score for different degrees of over-fitting. As one can see that when the number of individuals per covariate is low (<=20), the density curve of the estimated propensity score deviates from the density curve of the true propensity score.

Figure 10: Density curves of the true underlying probability of receiving treatment P_E and the estimated propensity score $\widehat{P}_{E,n}$ for different degrees of over-fitting.



We can see that for each density curve even for the true propensity score, there are many individuals who have extreme propensity score values close to 0 or 1. This scenario is not good for our study because later, we use propensity score to weight a multivariable analysis in order to estimate the effect of treatment. And in this case, many unrealistically extreme weights would be generated. That is why we have decided to sample the regression coefficients from a beta distribution with the two shape parameters a=0.5 and b=2.

Figure 11: Agreement of the estimated propensity score and true propensity score conditional to different degrees of over-fitting.



One can see that for each number of individuals per covariate and for a given value of the true propensity score, one have the 95% distribution interval in which should fit the corresponding values of the estimated propensity score. These intervals are non-parametric and they were obtained by the quantiles of the propensity score. When the number of individuals per covariate in the propensity score is low (<=20), the gap between the two lines of the confidence interval is large. The dash line at the center of the graph represents the perfect value of the estimated propensity score for a given value of the true propensity score.

Annexe 3: Simulations when the regression coefficients were on the joined interval $[-0.05;\ -0.01] \cup [+0.01;\ +0.05]$

We simulated data for a setting in which there were 100 baseline covariates $X_1, \ldots, X_{100}$ which were independently standard normal distributed. In order to initiate real confounding situations, the regression coefficients β_i were sampled from a uniform distribution on the joined interval $[-0.05; -0.01] \cup [+0.01; +0.05]$. We allowed covariates to have a weak effect of treatment selection or outcome. As the regression coefficients are close to 0, confounding effects are small. We use a

logistic regression model to specify the true probability of receiving treatment P_E for the simulated values of the linear predictor:

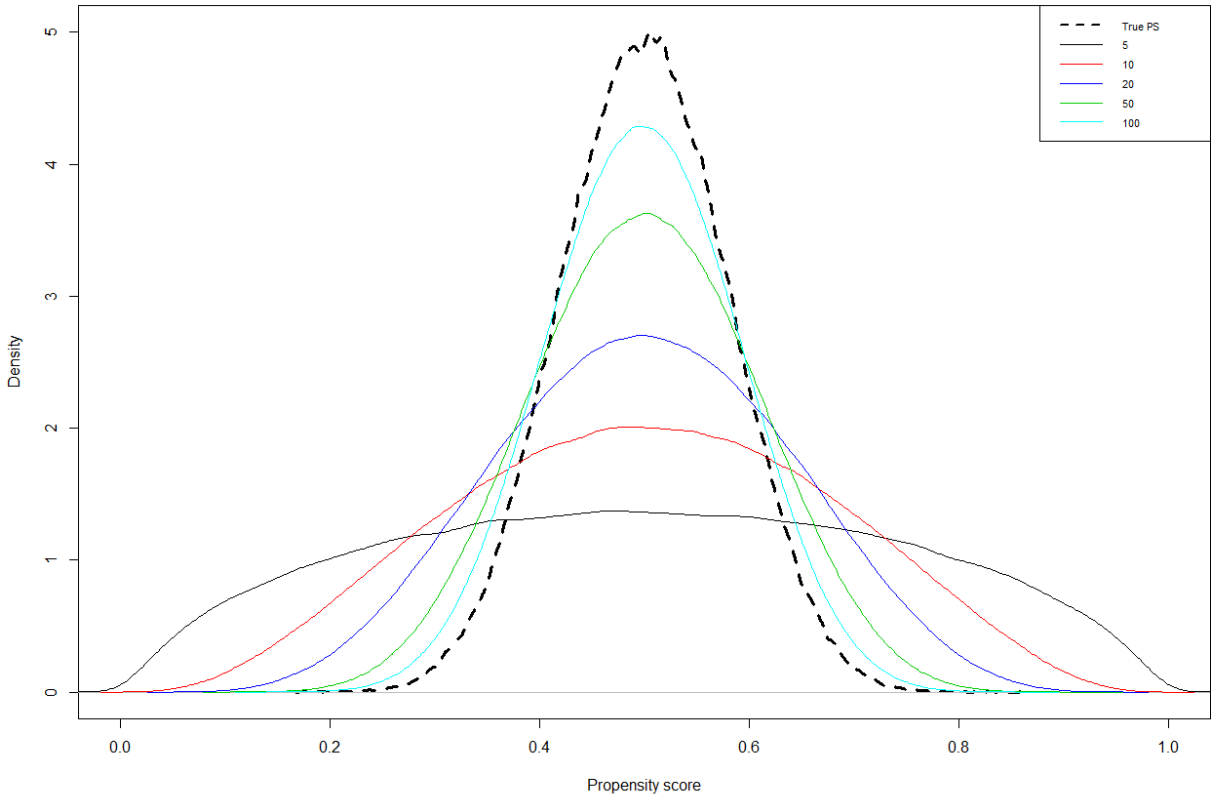
$$P_E = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{100} X_{100})}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{100} X_{100})}$$

The intercept β_0 was chosen to be 0 in order to achieve stochastically balanced numbers of exposed and unexposed individuals in the simulation. The prevalence will be 50 percent because X_i has expectation to be 0. For each individual, the exposure status E was sampled from a Binomial distribution with parameter P_E. E is a function of the X_i (when we estimate the propensity score with a logistic regression, the outcome variable is the exposure E). In order to initiate different magnitudes of over-fitted propensity score models, the number of individuals per covariate were set to $5, 10, 20, 50$ and 100. Within each of these datasets, the sample-specific propensity score $\hat{P}_{E,n}$ was estimated via logistic regression on the binary outcome Y_n including all 100 covariates. This procedure, data simulation, sampling and propensity score estimation, was repeated 1000 times for each of the five sample size scenarios.

Annexe 4: Results when the regression coefficients were on the joined interval $[-0.05;\ -0.01] \cup [+0.01;\ +0.05]$

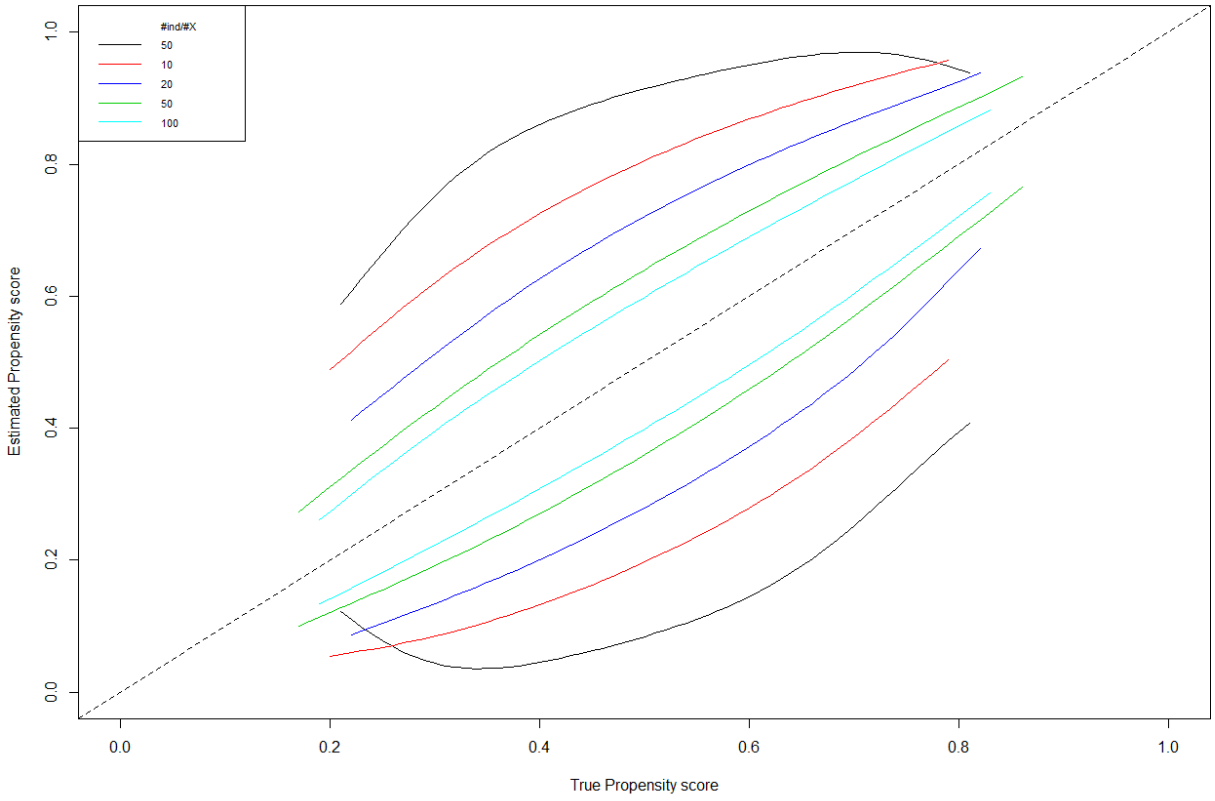
Figure 12 shows the density curves of the true propensity score and estimated propensity score for different degrees of over-fitting. As one can see that when the number of individuals per covariate is low (<=20), the density curve of the estimated propensity score deviates considerably from the density curve of the true propensity score.

Figure 12: Density curves of the true underlying probability of receiving treatment P_E and the estimated propensity score $\widehat{P}_{E,n}$ for different degrees of over-fitting.



One can see that for each density curve even for the true propensity score, there are many individuals who have propensity score values between 0.3 and 0.7. In this case, covariates have a weak effect of treatment selection or outcome. This situation is not realistic. That is the second raison why we have decided to sample the regression coefficients from a beta distribution with the two shape parameters a=0.5 and b=2.

Figure 13: Agreement of the estimated propensity score and true propensity score conditional to different degrees of over-fitting.



One can see that for each number of individuals per covariate and for a given value of the true propensity score, one have the 95% distribution interval in which should fit the corresponding values of the estimated propensity score. These intervals are non-parametric and they were obtained by the quantiles of the propensity score. When the number of individuals per covariate in the propensity score is low (<=20), the gap between the two lines of the confidence interval is large. The dash line at the center of the graph represents the perfect value of the estimated propensity score for a given value of the true propensity score.