A Comparison of Matching and Weighting Methods for Causal Inference Based on Routine Health Insurance Data, or: What to do If an RCT is Impossible

Ein Vergleich von „matching“ und „weighting“-Verfahren zur Kausalanalyse mit Routinedaten von Krankenversicherungen, oder: Was tun wenn ein RCT nicht möglich ist

Authors
Herbert Matschinger, Dirk Heider, Hans-Helmut König

Affiliations
Institut für Gesundheitsökonomie und Versorgungsforschung, Universitätsklinikum Hamburg Eppendorf, Hamburg, Germany

Key words
matching, weighting, Routine data, RCT, entropy balancing

Schlüsselwörter
propensity score, gewichtung, routine daten, RCT, entropy balancing

Bibliography
DOI https://doi.org/10.1055/a-1009-6634
Online-Publikation: 17.2.2020
Gesundheitswesen 2020; 82 (Suppl. 2): S139–S150

ISSN 0949-7013

Correspondence
Dr. Herbert Matschinger
Institut für Gesundheitsökonomie und Versorgungsforschung, Universitätsklinikum Hamburg Eppendorf, Martinistraße 52
20246 Hamburg
Germany
herbertmatschinger@yahoo.de

Die Appendix A-E finden Sie online unter https://doi.org/10.1055/a-1009-6634

ABSTRACT

Due to a multitude of reasons Randomized Control Trials on the basis of so-called “routine data” provided by insurance companies cannot be conducted. Therefore the estimation of “causal effects” for any kind of treatment is hampered since systematic bias due to specific selection processes must be suspected. The basic problem of counterfactual, which is to evaluate the difference between two potential outcomes for the same unit, is discussed. The focus lies on the comparison of the performance of different approaches to control for systematic differences between treatment and control group. These strategies are all based on propensity scores, namely matching or pruning, IPTW (inverse probability treatment weighting) and entropy balancing. Methods to evaluate these strategies are presented. A logit model is employed with 87 predictors to estimate the propensity score or to estimate the entropy balancing weights. All analyses are restricted to estimate the ATT (Average Treatment Effect for the Treated) Exemplary data come from a prospective controlled intervention-study with two measurement occasions. Data contain 35 857 chronically ill insurants with diabetes, congestive heart failure, arteriosclerosis, coronary heart disease or hypertension of one German sickness fund. The intervention group was offered an individual telephone coaching to improve health behavior and slow down disease progression while the control group received treatment as usual. Randomization took place before the insurants’ consent to participate was obtained so assumptions of an RCT are violated. A weighted mixture model (difference-in-difference) as the causal model of interest is employed to estimate treatment effects in terms of costs distinguishing the categories outpatient costs, medication costs, and total costs. It is shown that entropy balancing performs best with respect to balancing treatment and control group at baseline for the first three moments of all 87 predictors. This will result in least biased estimates of the treatment effect.

ZUSAMMENFASSUNG

Aus verschiedensten Gründen kann auf der Basis sogenannter „Routine-Daten“ von Versicherungsgesellschaften ein RCT nicht durchgeführt werden. Daher ist die Schätzung „kausaler“ Effekte unmöglich, da mit systematischer Verzerrung durch spezifische Selektionsprozesse gerechnet werden muss. Die grundlegenden Probleme des „Kontrafaktischen“, also die Beurteilung der Differenz zwischen zwei potentiellen Ergebnissen an derselben Beobachtungseinheit, werden abgehandelt. Der Fokus dieser Studie liegt im Vergleich von methodischen Zugängen die Differenzen zwischen Versuchs- und Kontrollgruppe zu kontrollieren. Alle Methoden basieren auf dem Propensity score, nämlich „Matchig“ bzw. „Pruning“, „Inverse Probability Weighting“ und „Entropy Balancing“. Methoden der Evaluation dieser Strategien werden dargestellt. Zur Balanzierung und/oder Schätzung des Propensity Scores dient ein Logit Modell mit 87 Prädiktoren. Alle Analysen beschränken sich auf die Schätzung des ATT (Average Treatment Effect for the Treated) Als Beispiel dienen Daten aus einer prospektiv kontrollierten Intervention-Studie mit 2 Messzeitpunkten. Die Daten beinhalten 35 857 chronisch kranke Versicherte mit Diabetes, Herzinsuffizienz, Arteriosklerose, Koronarer Herzkrankheit und Hypertonie. Der Interventionsgruppe wurde ein individuelles Telephoncoaching zur Verbesserung des Gesundheitsverhaltens und zur Verlangsamung des Krankheitsfortschrittes angeboten, wohingegen die Kontrollgruppe konventionelle Therapien bekam. Die Randomisierung wurde vor dem Einholen der Teilnahmezustimmung durchgeführt, wodurch die Vorrausetzungen eines RCT verletzt sind. Zur Schätzung des Behandlungseffektes mit Rücksicht auf Kosten wurde ein gewichtetes Mixture Modell (Differenz-in-Differenz) eingesetzt. Dabei wurde zwischen ambulanten Kosten, Medikationskosten und Gesamtkosten differenziert. Es kann gezeigt werden, dass das Verfahren des „Entropy Balancing“ die Verteilung der Prädiktoren zur baseline mit Rücksicht auf die ersten drei Momente am besten balanziert und damit die wohl am wenigsten verzerrten Behandlungseffekte liefert.

Introduction

In recent years more and more studies try to estimate intervention or other types of treatment effects on the basis of so-called “routine data” provided by insurance companies. Due to a multitude of accounts randomization and/or randomized control trials on the basis of that archive data often cannot be conducted. The retrospective character of these data as well as ethical reasons hampers the application of standard control trials and therefore systematic bias due to selection processes must be suspected. The estimation of a “causal effect” therefore relies on the performance of methods primarily used in observational studies. We will focus on the comparison of different approaches all based on the propensity score [1] if a randomized control trial (often seen as a “gold standard” in evaluation research and many other disciplines) is not possible for whatever reasons. All these procedures try to model selection bias by making treatment and control group as similar as possible before the intervention. Several studies have been conducted in order to compare different strategies to estimate a causal effect. These studies compare propensity score weighting and matching for binary outcomes (see for instance [2–5]) for different matching methods [6, 7], as well as for multivalued treatment [8] The studies mentioned above differ considerably with respect to their conclusions. None of these studies included “entropy balancing” as a distinguished approach to achieve balance between treatment and control group(s).

The remainder of this paper is organized as follows. First, we will address basic problems of causal analysis. Next, we will present possible strategies to deal with these problems; all based on propensity concepts, namely propensity score matching, IPTW (inverse probability treatment weighting) and entropy balancing. We will then present methods to evaluate these strategies. Afterwards, we will describe the difference-in-difference model as the causal model of interest. Thereafter, we will apply the described methods to routine health insurance data from an illustrative example study. Finally, we will discuss our findings and draw some conclusions.

Basics of Causal Analyses

The main goal of evaluation studies always is to causally attribute the difference between the study groups to the treatment, intervention or any other type of systematic difference introduced by the research design. “The main question of impact evaluation is one of attribution—isolating the effect of the program from other factors and potential selection bias” ([9] p. 4). Or, evenly important: “Inferences about the effect of treatments involve speculations about the effect one treatment would have on a unit which, in fact, receives some other treatment”. The central problem (“Fundamental Problem of Causal Inference” [10]) to solve is the problem of counterfactual that is to evaluate the difference between two potential outcomes for the same unit. Literature on the topic of causal inference and propensity methods in particular is vast and will not be pursued or reviewed in detail.

The framework of potential outcomes

The data consist of Y the observed outcome, T indicating the treatment status 0 or 1 and X a set of characteristics which are suspected to be related to both T and Y. It is of central importance to look at “potential outcomes” and not directly on realized outcomes. For each individual i two possible outcomes exist:

y_i^1 “Potential outcome” for state 1 (intervention) for the individual i
y_i^0 “Potential outcome” for state 0 (control) for the individual i

The success of the treatment is commonly evaluated by inspection of the differences between these potential outcomes $\delta_i = y_i^1 - y_i^0$ which cannot be directly observed since an individual can only be observed for one status of T and therefore frequently is named “counterfactual” (see for instance [11] and ▶ **Table 1**).

▶ **Table 1** The relation between potential outcome and treatment assignment.

	Potential outcome	
Study group	**Y^0**	**Y^1**
Treatment (T = 1)	counterfactual	observable
Control (T = 0)	observable	counterfactual

The observed response is defined as $Y_i \equiv T_i Y_i^1 + (1 - T_i) Y_i^0$, which implies the so-called "Stable Unit Treatment Value Assignment" – SUTVA [12]. It says that the observed outcome only depends on the potential outcome and the treatment status and not on other individuals from the data. This additionally means that it is assumed that every person of the population has the same probability of being chosen for the treatment group. Since we must expect that differences between potential outcomes are different for each individual we will refer to expectations $E(Y^0)$ and $E(Y^1)$.

According to the research question different forms of treatment effects are to be estimated.

1. ATE (Average Treatment Effect)
   $E(\delta) = E(Y^1 - Y^0) = E(Y^1) - E(Y^0)$ ← the expected difference for individuals sampled from the total population.
2. ATT (Average Treatment Effect for the Treated)
   $E(\delta \mid T = 1) = E(Y^1 - Y^0 \mid T = 1) = E(Y^1 \mid T = 1) - E(Y^0 \mid T = 1)$ ← the expected difference for individuals sampled from the population which actually is exposed to treatment. $E(Y^0 \mid T = 1)$ can never be observed and has to be substituted by a properly preprocessed or selected control group.

We must assume that both the outcome and the treatment assignment depend on a set of covariates X. Randomization will, on expectation, balance both groups with respect to both measured and unmeasured characteristics such making causal inference straightforward. In observational studies and/or non-randomized designs the assignment of individuals might not be independent from individually varying characteristics and data must be preprocessed to control for selection on observable variables [13, 14] and to make both groups as similar as possible.

In order to reduce the high dimensionality of X the so called "propensity score"(called ρ(X) in the following) is used; the probability of being a member of the treatment group conditioning on X:$\rho(X) \equiv P(T = 1 \mid X)$. If this score is known then $X \perp T \mid \rho(X)$ - „Treatment assignment and the observed covariates are conditionally independent given the propensity score" ([15] p. 44 theorem 1). In observational studies this score is unknown and has to be estimated [15, 16]. In most instances this is done by either a logit or a probit model. However, "……conditioning on the propensity score allows one to replicate some of the characteristics of a randomized control trial (RCT) " ([17] p. 2038)."Conditioning on ρ(X) balances the distribution of Y^0 and Y^1 with respect to T"[1], ([18] p. 265). This is possible if certain assumptions hold:

(1) unconfoundedness $(Y^0, Y^1) \perp T \mid X$
Potential outcomes are independent of treatment assignment given a set of observed covariates - X
(2) overlap $0 < P(T = 1 \mid X) < 1$
The probability of receiving a treatment must be positive for all values of X and never equal 0 or 1, so for any X there must be both "treated" and "untreated" subjects.

If both conditions hold then necessarily $P(T = 1 \mid Y^0, Y^1, X) = P(T = 1 \mid X)$ and treatment assignment is "strongly ignorable" in the sense of Rosenbaum & Rubin ([15] p. 45 theorem 3]) and it assumed that $(Y^0, Y^1) \perp T \mid \rho(X)$ - conditioning on the propensity score alone is acceptable.

It is assumed that the propensity score distributions are similar for both groups and sufficient "overlap" is observed. The overlapping regions define the "common support". Observations from the control group outside the common support are inappropriate for comparison particularly for the estimation of the ATT. Nearest neighbor matching with a caliper excludes observation outside the area of common support but for weighting methods this has to be done "by hand". Crump. et.al. [19] for instance advocated to discard all observations outside a particular range [α,1-α] searching for an "optimal" subpopulation. Consequently, assumption 2 is defined as:

for some $c > 0$, $c < P(T = 1 \mid X) < 1 - c$ (see [19] p. 189 assumption 2)

But defining any cut-off (c) may lead to heavy reduction of both groups if one group is much smaller than the other. Most importantly, cut-off criteria based on the propensity score alone are arbitrary and might not be justified easily [20], so we do not further pursue this approach.

So far, if treatment effects are defined for persons sampled from the whole population, the ATE (average treatment effect) will gain external validity. If the effect will be valid only for those exposed to treatment, the ATT (average treatment effect of the treated) should be estimated, and conditional independence is based on weaker assumptions: $(Y^0) \perp T \mid X$ and $P(T = 1 \mid X) < 1$. In most instances only the assumption $E(Y^0 \mid T = 1) - E(Y^0 \mid T = 0) = 0$ (the difference between treatment and control group with respect to the potential outcome WITHOUT treatment) can be substantiated and the estimated effects are of internal validity only. The other assumption $E(Y^1 \mid T = 1) - E(Y^1 \mid T = 0) = 0$ (sometimes called: "absence of differential treatment bias") can rarely be justified which particularly holds for the empirical example outlined below.

Typically, the estimated propensity score is used by matching on the score, stratification or subclassification, covariate adjustment or inverse probability treatment weighting (IPTW) based on the estimated propensity score ([1] chapter 5, 6 & 7, [21]). All these methods tend to yield unbiased estimates only if assumptions 1 & 2 hold. We will focus only on matching and weighting strategies as well as on a relatively new strategy called entropy balancing and apply them in our illustrative example.

Matching, weighting and entropy balancing

Matching

In case of matching control individuals are searched which are similar with respect to a distance measure. Matching is employed to make the multivariate distribution of all covariates X as similar as

1 In the original text D instead of T and P instead of ρ – changed here to achieve notational consistency

possible by selecting appropriate control observation(s) for each treatment observation.

There are, at least, four types of distance measures: exact, Mahalanobis distance and the propensity score or the linear logits predicted by the logit-model. We only will focus on the last one. "Similarity" usually is defined by the standard deviation of the distance measure – the caliper which is defined as proportions of the linear logits [22]. Matching can be done 1:1 or 1:k with and without replacement after each draw. With replacement allows a control subject to be matched to different treatment subjects. The advantage is that the order of the subjects in the control-group has no effect on the matched sets. To obtain optimally similar groups used afterwards in a parametric mixture model the following steps have to be passed ([23] p. 5 section 1.4):

"1. Defining "closeness": the distance measure used to determine whether an individual is a good match for another,
2. Implementing a matching method, given that measure of closeness,
3. Assessing the quality of the resulting matched samples, and perhaps iterating with steps (1) and (2) until well-matched samples result, and
4. Analysis of the outcome and estimation of the treatment effect, given the matching done in Step (3)".

As an example we will use the linear propensity score $D_{ij} = |\text{logit}(e_i) - \text{logit}(e_j)|$ as distance measure which is seen as very effective for reducing bias induced by confounders [24]. Following the advices of Austin [25] a caliper of 0.2 is acceptable."The rationale for matching on the logit of the propensity score is that the logit of the propensity score is more likely to be normally distributed than the propensity score itself" ([25] p. 152). For 1:k matching matched control units are weighted proportional to the number of treatment units they are matched to [26]. This weighting procedure must not be mixed up with the weighting strategies described below.

Weighting

Weighting employs the PS to generate weights for each single observation. Dependent on the research question several kinds of weights are proposed in the literature (see [27] p. 392 Table 1), and to describe and evaluate all possibilities is far beyond the scope of this article. We will employ weights based on **I**nverse **P**robability **T**reatment **W**eights (IPTW). ATE weights for the groups are defined as $\omega(T,X) = \frac{T}{P(X)} + \frac{1-T}{1-P(X)}$ which results in 1/P(X) for the treatment and 1/(1-P(X)) for the control group. Since we want to compare different approaches with entropy balancing we will focus on weighting schemes for which the target population is the population of the treated. Weights for the ATT are defined as $\omega(T,x) = T + \left((1-T)\frac{P(X)}{1-P(X)}\right)$, so weights for the treatment individuals are 1 and subjects of the control-group are weighted by P(X)/(1-P(X)) (comp. [1] p. 244). Whatever weighting is employed it is always necessary to check the common support, which sometimes make discarding necessary.

Problems with matching and weighting

Basically, all approaches are conducted to generate a "pseudo-population" were all observations are conditionally exchangeable ([28] p. 177) and the (potentially) weighted control group provides a surrogate outcome for the counterfactual outcome (see [29] p. 335). One essential drawback of matching and weighting approaches is the "propensity score tautology" [30, 31] since repeated estimation of the propensity score, implementing a matching algorithm, computing weights and evaluating balance not necessarily yields optimal results with respect to balance even if the complexity of the propensity model is very high. Because the generating process for the propensity score is unknown, finding the "correct" model in order to mimic a randomized experiment turns out to be a sometimes never ending story from step 1 to step 3. Furthermore, all matching procedures necessarily yield different (sub)-groups for both treated and untreated subjects, which may result in severe problems too: „Increasing the number of untreated subjects matched to each treated subject will increase the size of the matched sample, probably resulting in estimates of treatment effect with increased precision. However, increasing the number of untreated subjects matched to each treated subject may result in the matching of increasingly dissimilar subjects. This may increase bias in estimating the effect of treatment" ([32] p. 1093).

The matching procedures are sometimes called "pruning" instead, but pruning the data at hand is considered an important disadvantage particularly for some methods like for instance exact matching: "Moreover, exact matching has the disadvantage in many applications of using relatively little of the data. Finding matches is often most severe if X is high dimensional (another effect of the curse of dimensionality) or contains continuous variables. The result may then be a preprocessed data set with very few observations that leads to a parametric analysis with large standard errors."([13] p. 212). Discarding individuals from the intervention group outside the common support may also cause problems in estimating the ATT since the focal group might be changed (comp. [23] p. 13). Unfortunately, most preprocessing methods are prone to result in low balance [30, 33] or: "Even worse, matching may counteract bias reduction for the subsequent treatment effect estimation when improving balance on some covariates decreases balance on other covariates." ([34] p. 26) Summarizing, one could say :"At least given the current state of the literature, only the propensity score tautology is useful in practice. Other theoretical results have no direct bearing on practice" ([13] p. 219]).

Another fundamental critique directly addresses the theorem 1 [15] mentioned above. The motivation was that it is easier to match on one scalar (propensity score) than on the high-dimensional *X*, but: "Balancing on π only is unbiased but inefficient ex ante, leaving researchers with more model dependence, discretion, and bias ex post"([35] p. 13).

Entropy balancing

To avoid the necessity to check the balance again and again and for the other reasons just mentioned a relatively new approach gained increasing attention which is called "entropy balancing" [34].

"Entropy balancing is a preprocessing procedure that allows researchers to create balanced samples for the subsequent

estimation of treatment effects. The preprocessing consists of a reweighting scheme that assigns a scalar weight to each sample unit such that the reweighted groups satisfy a set of balance constraints that are imposed on the sample moments of the covariate distributions. The balance constraints ensure that the reweighted groups match exactly on the specified moments". ([34] p. 30 section 3)

If $(Y^0) \perp T \mid X$ is equal to $(Y^0) \perp T \mid \rho(X)$ than balance on all covariates X can be achieved relying on this single score. Consequently, the counterfactual mean can be written as

$E(Y^{(0)} \mid T = 1 = \int E[Y \mid \rho(X) = \rho, T = 0] f_{p \mid T-1}(\rho) d\rho$ where $f_{p \mid T=1}$ is the distribution of the propensity score in the target population (treatment group). The main goal is to preprocess the control group in such a way that the weighted density $f^*_{X \mid T-0}$ corresponds to $f_{X \mid T=1}$ Entropy balancing tries to achieve covariate balance directly and can be seen as a generalization of the propensity score weighting approach to use a weighted average of the control-group to estimate the counterfactual expectation $E(Y^0 \mid T = 1) = \frac{\sum_{\{i \mid T=0\}} Y_i \omega_i}{\sum_{\{i \mid T=0\}} \omega_i}$ (see [34] p. 30 eq. 1).

For each control unit a weight ω_i is supplied which is obtained by minimizing the loss function: $\min_{\omega_i} H(\omega) = \sum_{i \mid T=0} h(\omega_i)$ For the loss function $h(\omega_i)$ the so-called directed Kullback entropy divergence between ω_i and the base weight q_i is chosen: $\omega_i \log(\omega_i / q_i)$ The base weights are set to $q_i = 1/n_0$ (n_0 = size of the control group). $\omega \log(\omega)$ is also seen as the Shannon entropy metric (comp. [34] p. 31 footnote 9).

The loss function needs balance as well as normalizing constraints:

$$\sum_{(i \mid T=0} \omega_i c_{ri}(X_i) = m_r \text{ with } r \in 1, \ldots\ldots R$$

$$\sum_{(i \mid T=0)} \omega_i = 1 \; \omega \geq 0 \text{ for all i}$$

$c_{ri}(X_i) = m_r$ defines R balance constraints imposed on the covariate (X) moments of the control group. m_r containes the r^{th} order moment of a particular covariate X_j from the treatment group; the moment functions are specified for the control group as $c_{ri}(X_i) = X_{ij}^r$ [34]. Weights have to sum to a constant – usually but not necessarily - one. Furthermore, weights must be constrained to be nonnegative because the distance metric is not defined for negative weights. The derivation of the iterative computation scheme to minimize the loss-function $H(\omega)$ can be found in section 3.2 of Hainmueller [34].

Conventional approaches in a first step try to estimate the weights by means of a logistic regression. A second step then becomes necessary to check whether the weights actually balance the covariate distributions. "Entropy balancing tackles the adjustment problem from the reverse and estimates the weights directly from the imposed balance constraints. Instead of hoping that an accurately estimated logistic score will balance the covariates stochastically, the researcher directly exploits her knowledge about the sample moments and starts by prespecifying a potentially large set of balance constraints that imply that the sample moments in the reweighted control group exactly match the corresponding moments in the treatment group."([34] p.31)

In doing so it exactly matches the covariate moments for the groups to be compared within its optimization problem [31]. The application of the entropy balancing procedure has the potential to improve balance in the covariate distribution with a maximum retention of information. The procedure of entropy balancing provides us with weights for the subjects of the control-group which can be employed subsequently in explanatory models.

Methods to evaluate balance

As said before it is of vital importance that the density of the weighted control group $f^*_{X \mid T=0}$ mirrors the density $f_{X \mid T=1}$. For evaluation of balance in our illustrative example we will calculate the standardized mean differences, variance ratios, skewness ratios for each covariate and presented them as box-plots over all covariates employed to estimate the propensity score (comp. [4] p. 244 Fig 2 and [27] p. 395 Fig. 2). Ideally, these box-plots are a simple line at 0 for differences or 1 for ratios. This method provides an intuitive way to compare a huge amount of different numbers; the existence of outliers, indicating bad balance, can be quickly identified. We want to underpin that even the evaluation of the first 3 moments is not sufficient since imbalances can exist anywhere within a distribution, and distributional equivalence is a key feature within the framework of potential outcomes [36]. Auxiliary, distributional equivalence will be checked by means of weighted Q-Q plots[37]. Since balance is not a problem of inference but rather a problem of the sample only, statistical testing is not warranted [17, 38, 39]. It is shown that, for instance,t-statistic decreases if only more and more control units are dropped which falsely suggests a better balance (comp. [30] p. 496 Fig. 1).

The parametric mixture model (Difference in Difference)

In the 2nd step, we will estimate the causal model of interest (see [40] : 402 pp) using the data from pre- and post-period of our illustrative example. Thereby, either the pruned samples or the weights derived from the 1st step are applied to the causal model. Given a two-period setting where t = 0 before the treatment and t = 1 after the treatment implementation, letting Y_t^T and Y_t^C be the respective outcomes for treatment and control units in time *t*, the DD (Difference in Difference) method will estimate the average treatment impact (using differences as counterfactual):

$$DD = E(Y_1^T - Y_0^T \mid T_1 = 1) - E(Y_1^C - Y_0^C \mid T_1 = 0)$$

$T_1 = 1$ and $T_1 = 0$ denotes the treatment at t = 1 and $T_1 = 0$ denotes no treatment at t = 1 ([9] p.72 eq. 5.1) If $E(Y_1^C - Y_0^C \mid T_1 = 0)$ can be employed as counter factual for $E(Y_1^C - Y_0^C \mid T_1 = 1)$ this can be written as a mixture regression:

$$Y_{it} = \alpha + \beta T_{i1} t + \rho T_{i1} + \gamma t + \varepsilon_{it}$$

The interaction coefficient β is the difference in change between intervention and control group. It represents the DD. ρ and γ pick up the difference between treatment and control at baseline and

change over time for the control group respectively. Conditional expectations of differences between measurement occasions for each group can be written as ([9] p. 73 eq. 5.3a & 5.3b):

$$E(Y_1^T - Y_0^T \mid T_1 = 1) = (\alpha + \beta + \rho + \gamma) - (\alpha + \rho)$$
$$E(Y_1^C - Y_0^C \mid T_1 = 0) = (\alpha + \gamma) - \alpha$$

Subtracting the second equation from the first yields exactly DD. DD (the interaction parameter β) is an unbiased estimator only if the potential source of selection bias is additive and time invariant and ε_{it} is uncorrelated with t, T_{i1} and $T_{i1}t$. The latter is called the parallel trend assumption. It means that, given the "treatment" group would have received no treatment, change between measurement occasions will be the same as in the control group. Comparing the change for the treatment group only will result in DD + γ; estimation of the difference between the 2 groups after treatment only will yield DD + ρ. Therefore, both parameters – γ and ρ – must be part of the model.

Software

On our illustrative example, data management, data analyses and graphical displays were conducted using STATA 15 [41]. The entropy balancing was estimated by ebalance [42] for STATA. Nearest neighbor matching was done using MatchIt [43] for R [44].Covariate balance was estimated by covbal [45], different forms of weights and their balancing performance by means of pbalchk and propwt for STATA developed by Mark Lunt downloaded at http://personalpages.manchester.ac.uk/staff/mark.lunt. The weighted Q-Q-plots were generated by qqplot3 for STATA [37].

Application

Data

We used data from a prospective controlled intervention-study with two measurement occasions. Data contain 35 857 chronically ill insurants with diabetes, congestive heart failure, arteriosclerosis, coronary heart disease or hypertension of one German sickness fund. Insurants were randomized into two groups: the intervention group (IG, N = 18 019) was offered an individual telephone coaching to improve health behavior and slow down disease progression while the control group (CG, N = 17 838) received treatment as usual. For reasons of data protection, randomization took place before the insurants' consent to participate was obtained. Finally, only 4 430 of originally 18 019 insurants randomized to the IG consented to participate. The estimation of treatment effects therefore relies on the performance of methods used in observational studies as outlined above. Treatment effects were analyzed in terms of costs from the perspective of the sickness fund, distinguishing the categories outpatient costs, medication costs, and total costs.

Matching, weighting and entropy balancing

All further analyses are based on the data available for the control group and those who actually participated (Control group = 17 838 Intervention group = 4 430). In the first step we estimated the logit model with 87 predictors which turned out to fit the data fairly well (chi^2 = 22 367 df = 22 180 p = 0.19). Since the set of predictors have no missings the whole set keeps available to estimate the predicted probabilities. The model comprises gender, age, occupational status, disease management program, status of health insurance, level of care, Federal state of residence, baseline values of health care services and costs as well as the 31 constituents of the Elixhauser comorbidity index [46] (for details see **Appendices A-E, Online**). Linear logits were used for matching. From the predicted probabilities the IPT-weights for the ATT were computed. We do not present the parameters of the logit model, since they are of minor interest. All analyses were restricted to estimate the ATT in order to provide a sound comparison with the ATT obtained from entropy balancing. As examples for matching we present results for nearest neighbor 1:1 matching without and 1:4 with replacement, the latter is considered to elicit lowest bias [32]. For both models the linear logit with a caliper of 0.1 standard deviations was used. Discarding observations outside the common support was allowed for both groups. ▶ **Table 2** shows how the two matching algorithms result in different subsamples for which the smaller one (1:1 matching) is not a strict subset of the other (1:4 matching). For both approaches 70 observations from the control group and only 1 observation from the treatment group had to be discarded.

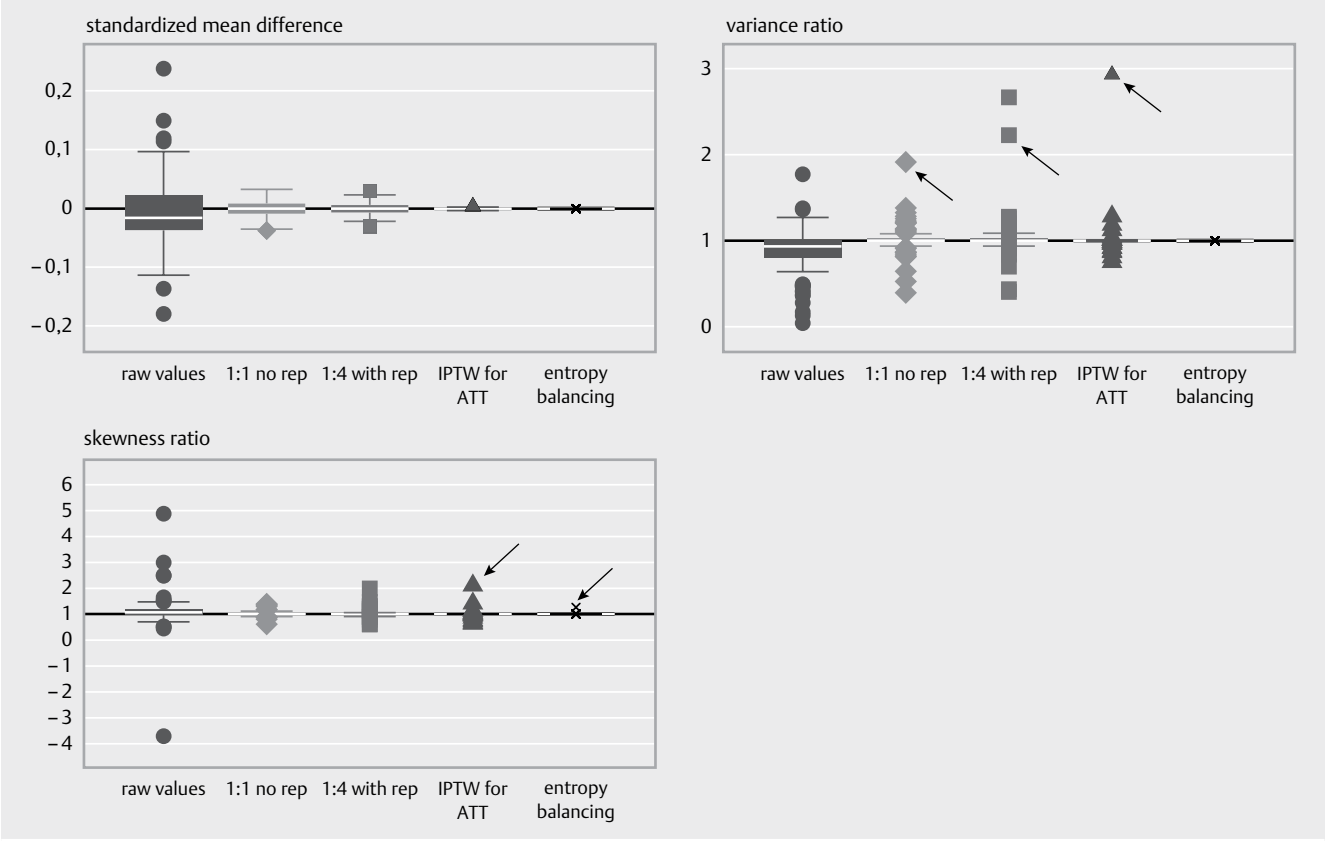
For the weighting approaches the sample size did not change as no observations had to be discarded. Last not least, the same 87 predictors were employed for entropy balancing. To achieve convergence for the iterative procedure a difference of 0.0001 was allowed as the maximum deviation across **all** specified moments. It turned out that, except for medication costs, all 3 moments could be balanced perfectly without any loss of observations. Medication costs could be balanced for mean and variance only.

All four balancing approaches were evaluated with respect to their performance and employed in the mixture models for 3 different cost categories (outpatient costs, medication costs and total costs).

Balancing checks

In the first step we check whether the propensity score shows enough common support. It turns out that even 60 quantiles always contain observations from both the control and treatment group. Looking at the box-plots (▶ **Fig. 1**) it becomes obvious that each of the propensity score based models contributes quite well to the balance of means (upper left panel : standardized mean differences). Unfortunately, this does not hold for variance and skewness ratios. Arrows within each panels point to an outlier for the 1:1 matching model, the 1:4 model with replacement as well as for the ATT weighting, indicating that these balancing procedures do not work acceptably well for the baseline values of this predictor. Looking at the **Appendices** we find that for each method medication costs generate the greatest difference.

Surprisingly, the box-plots for the entropy balancing approach (most right box-plot in each panel) yield a line at zero or one (ratios), with virtually no distribution around. Numerical values show that all confounders could be balanced perfectly for the first 3 moments, except again for medication costs. In the lower left panel (skewness ratios) of ▶ **Fig. 1** we see an x above the 1-line in the box plot for entropy balancing. After entropy balancing, the skewness for medication costs is still 32.429 for treatment and 25.781 for the control group (13.05 for raw data). This is the one and only differ-



▶ **Fig. 1** Box-plots for standardized mean differences, variance ratios and skewness ratios. Within each panel box-plots for raw data; 1:1 without replacement; 1:4 with replacement; IPTW for ATT and entropy balancing.

ence for all the 3 moments to be found for entropy balancing (see **Appendices C-E Online**).

Q-Q-Plots unweighted and weighted

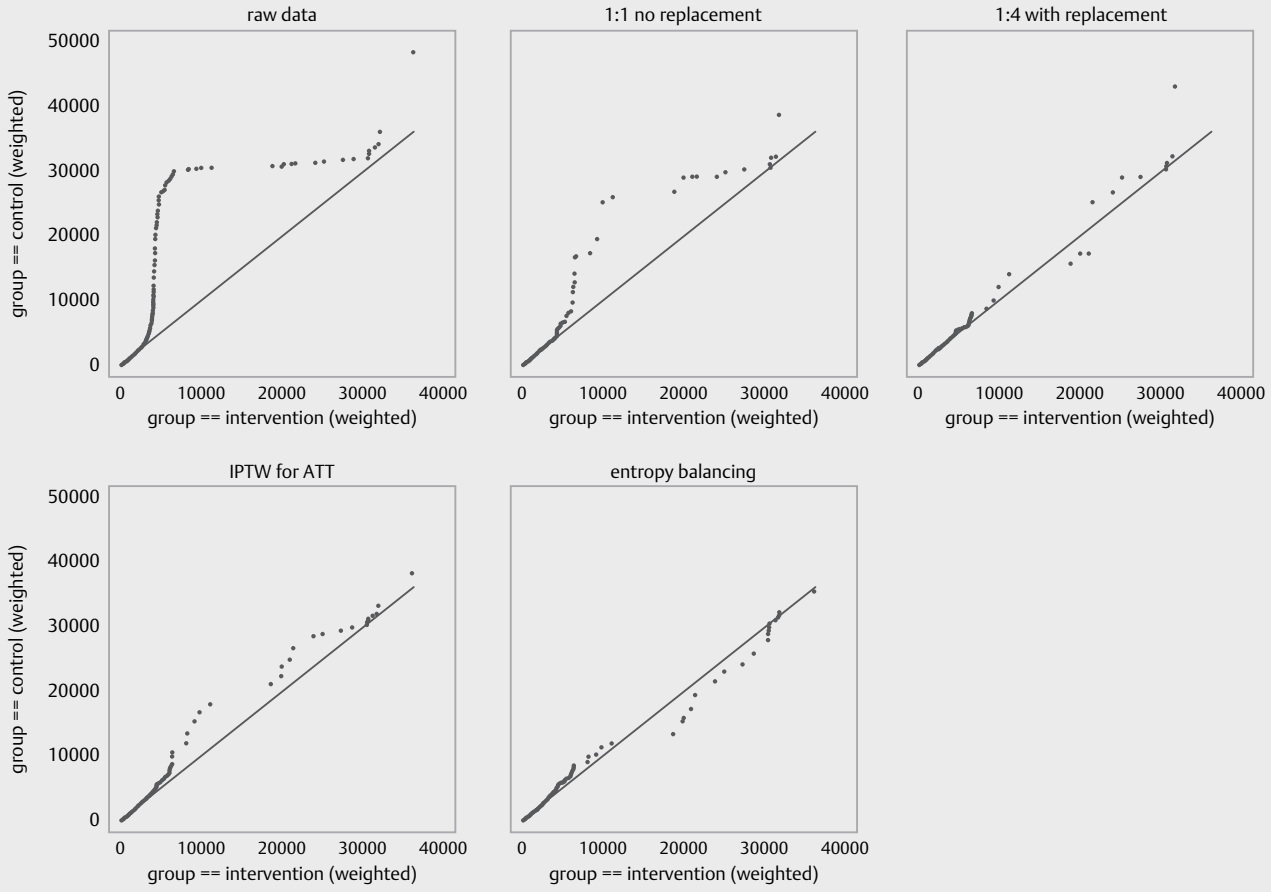
The quantile-quantile plots for continuous variables are only shown for those outcome variables for which the mixture model is presented below although the propensity score model comprises several other continuous variables. Outpatient costs may both serve as an example for an exceptionally good working balance, medication costs for a less perfect balance (skewness), and total costs are presented because they were the primary outcome of the study the data come from.

The Q-Q plots (▶ **Figs. 2** – **4**) clearly demonstrate the superiority of entropy balancing with respect to distributional equivalence at baseline. As already mentioned this does not hold for medication costs, as the skewness still differs between the two groups at baseline. We checked distributional equivalence for all the other continuous variables adopted to estimate the propensity score and found a similar pattern for all these characteristics, too. In order to save space, we pick out three as an example.

Mixture model (pre-post)

For raw data and for each of the 4 propensity score based approaches a mixture model was estimated employing either the pruned samples, or the weights at the individual level (level 2). The first parameter ("treatment") models the mean-difference between treatment- and control-group at baseline. The second parameter (follow up) is the change between measurement occasions for the control group. The interaction parameter portrays the difference in change between treatment and control group. Although the interaction parameter represents the ATT both change parameters should be interpreted as an additive linear combination. All types of costs shown in ▶ **Table 3** decline for the control group, but the interaction parameter which denotes the ATT is positive so the decline is less for the treatment group.

For each example we observe considerably different interaction effects, which clearly show the suspected model dependence estimating the ATT. The two matching models are based on different parts of the original sample so it becomes unclear what kind of a population these groups are representative for. Balance at baseline is acceptable only, or at least best for the entropy balancing approach. Looking at the results for the weighting methods based 22 197 observations (Control group = 17 768 Intervention group = 4 429) after discarding observations outside the common support (see ▶ **Table 2**) it becomes obvious that both the IPT weighting and the entropy balancing yield less biased results, with respect to baseline balance compared to the two matching models. Following the advice of Crump et.al. [19] discarding outside the range of 0.1 – 0.9 results in a control group of 16 589 and an intervention group of size 4377 which means that 1249 obs. from



▶ **Fig. 2** Q-Q plots for Outpatient costs.

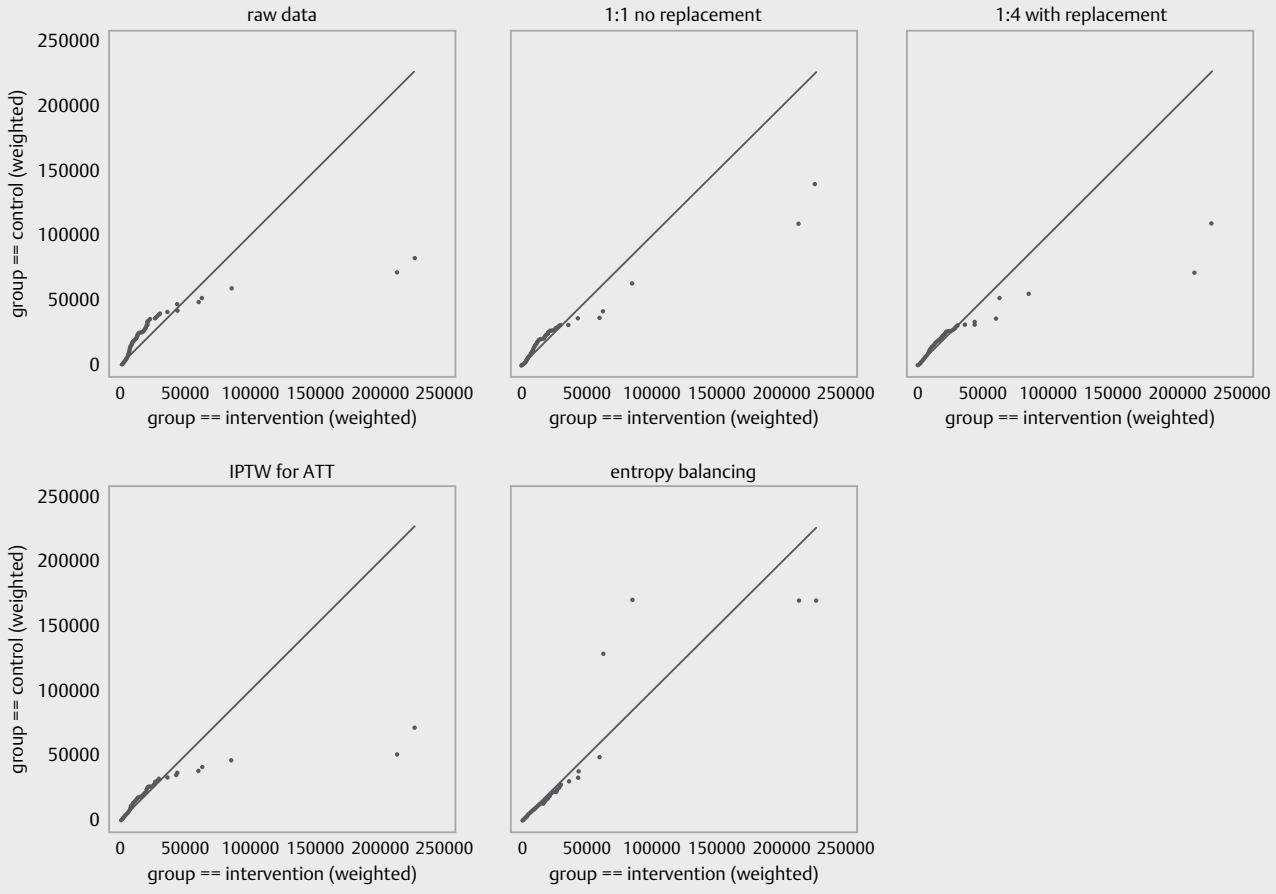
▶ **Table 2** Sample size for the two matching models.

	1:1 without replacement		1:4 with replacement	
	Control	**Treated**	**Control**	**Treated**
All	17838	4430	17838	4430
Matched	3991	3991	9015	4045
Unmatched	13777	438	8753	384
Discarded	70	1	70	1

the control group and 53 obs. from the intervention group need to be discarded. Obviously, the target population is changed implicitly and it is impossible to decide whether this truncation can be neglected. For sake of comparability we decided to keep all observations except those discarded before, also because the distribution of the PS is very similar for both groups.

Discussion

The use of propensity score analyses has become most popular for causal analysis not only in the field of observational studies. These methods are widely employed if it is impossible to conduct a randomized control trial for reasons whatsoever. The strand of literature available provides a virtually unmanageable amount of different approaches for each of which the applied scientist will find their advocates and detractors. This paper was aimed to compare selected methods employing the same set of covariates which were suspected to be connected both with the outcome characteristics at baseline and the treatment assignment. We focused on matching and weighting, but did not consider stratification or direct covariate adjustment. We do not discuss matching on propensity score compared to matching on X and the inefficiency from reducing the high-dimensional space of X [47]. However, it could be clearly shown that entropy balancing is superior at least compared to the other methods since it balances not only for means, but also for variance and skewness. This is in line with findings of other investigators like for instance Marcus [48], who also observed the superiority of entropy balancing. It is not surprising that all of these approaches yield results far away from the so-called "naïve" estimator, for which particularly the interaction effect must be severely biased. This also holds for all the other cost categories (e.g. hospital costs and rehabilitation costs) not presented here. It also was shown that the estimates differ considerably between the 4 models applied to estimate the ATT.



▶ **Fig. 3** Q-Q plots for Medication costs.

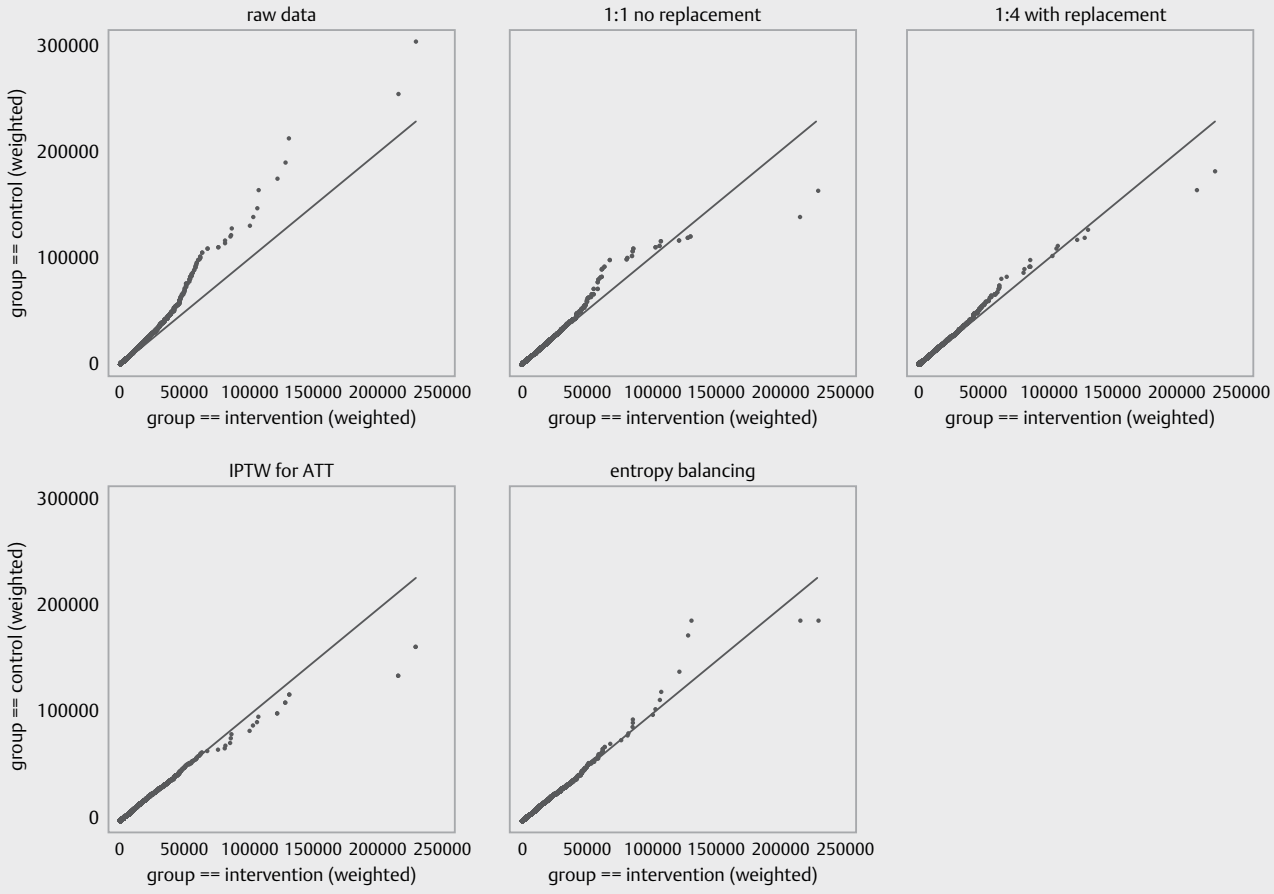
We focused on the ATT because the definition of a target population is considered an important problem of causal inference. The target population is defined both by the population both groups are drawn from, or – more realistically - by the population the actual intervention groups is representative. The latter sometimes is hard to determine. Discarding individuals outside the common support has no considerable effect in our application and all the conclusions from comparing the different approaches are still valid even for the total sample. Of course, this only holds because the amount of individuals subject to discarding is very small compared to the overall N which might not the case for other applications.

Weighting of regression models commonly is employed in order to reduce bias. However, it is well-known that weighting affect the standard error of parameters and we always have to face the trade-off between bias and efficiency [40]. Even though we undoubtedly tend to prefer less bias sacrificing efficiency it is obvious that there is not one and only one way to substitute an RCT. Methods presented above always only control for observed variables and never for unobserved – perhaps unobservable – confounders. Replication of an RCT using propensity scores is always conditional on observed variables, and unobserved variables may still differ considerably between treatment and control group.

Several restrictions should be mentioned. First of all, we only present two models for matching, although there exist much more possibilities [2, 3]. Secondly, we only employed a logit model to estimate the propensity score for the first three approaches presented, although there are several other concepts like for instance Generalized Boosted Models [49] which are based on decision trees. This iterative procedure includes interactions and polynomials and perhaps provides a better propensity score model. To evaluate this is beyond the scope of this article, too. Thirdly, a linear mixture model to predict costs was employed, without controlling for all the confounders again. Since costs have a lower limit of zero several other parameterizations are conceivable. Finally, in our example all the covariates are free of missing values so we do not employ any method to deal with missing values (compare [4, 50] and **Appendix B, Online**). This will be not the case in most instances. Of course, all other problems resulting from the necessity to estimate the unknown propensity score still apply.

Consequences

Most importantly it is recommended to always check balances between treatment and control group both in RCT and in observational studies. Randomization may fail, but there are ways to han-



▶ **Fig. 4** Q-Q plots for Total costs.

▶ **Table 3** DD mixture model to explain costs in raw data and by 4 different matching schemes.

Outpatient costs	raw data	1:1 no rep.	1:4 with rep.	IPTW for ATT	entrop.bal.
treatment	−194.18869 * * *	−13.114089	27.873286	2.307427	−0.00001264
Follow up	−113.74492 * * *	−23.317741	−31.51633 *	−46.79681 * * *	−59.183751 * * *
Group * fup	76.579606 *	−18.217726	−11.101702	11.015181	23.402122
Medical costs					
Treatment	−208.06421 * *	−35.046326	12.542562	−8.3878811	−8.842e-06
Follow up	−263.7955 * * *	−136.92768 * * *	−145.81334 * * *	−197.68052 * * *	−149.26739 * * *
Group * fup	0.17549 * *	23.942093	37.763437	92.89417 *	44.481042
Total costs					
Treatment	−1478.6604 * * *	−386.23622	−297.05235	12.206846	-0.00098796
Follow up	−4723.9841 * * *	−3974.2687 * * *	−3987.0304 * * *	−3855.6075 * * *	−3855.3822 * * *
Group * fup	1212.2154 * * *	424.69666	453.78854	349.50177	349.27648

* p<0.05; * * p<0.01; * * * p<0.001

dle this situation. As results show, to balance with respect to the mean of confounders only is sufficient for dichotomous covariates but not for continuous ones, as distributional equivalence is of vital importance. Entropy balancing seems to be - at the moment – a method which at least in big samples allows for balancing the first three moments which results in very similar distributions for both groups. The parametric model may yield a parameter indicating no difference at baseline between treatment- and control-group but we should not forget that this indicates only differences in means. The interaction parameter as an indicator for the average treatment effect may nevertheless be biased. Inspection of the Q-Q plots and the distribution of moments for **all** covariates of the PS model is absolutely necessary.

Furthermore, one should not restrict the PS model to only a few covariates but rather employ as much as possible regardless whether these variables show any significant effect or not. Inference is obsolete in the framework of constructing a balancing score. It is advisable to dichotomize all the categorical variables taking care of linear dependencies, and to decompose all indices as we have done for the Elixhauser Index. This decomposition instead of using summarizing indices considerably facilitates and improves the balancing procedure.

Nevertheless, it becomes clear that even small deviations from a multivariate balance may result in considerable differences of the estimated parameters in the second step. We purposely show results for the "medical costs", knowing that all the different procedures only yield a questionable balance which results in very different estimations for the ATT.

We want to underpin that there is no "gold-standard" on how to correct for selection bias, as there are always unobserved confounders which may result in hidden bias (compare [51] chap. 4). Last not least, the definition and theoretical justification of the target population is of vital importance and one should keep in mind that searching for the "optimal subpopulation" might implicitly change that target, which will sacrifice the generalizability of treatment effects.

Acknowledgements

We thank two anonymous reviewers for their valuable comments

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Guo S, Fraser M. Propensity Score Analysis. Statistical Methods and Applications. 2. Ed. Los Angeles London New Delhi: SAGE; 2015

[2] Austin P. A comparison of 12 algorithms for matching on the propensity score. Stat Med 2014; 33: 1057–1069

[3] Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. Value Health 2006; 9: 377–385

[4] Harder V, Stuart E, Anthony J. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychol Methods 2010; 15: 234–249

[5] Seeger JD, Bykov K, Bartels DB et al. Propensity score weighting compared to matching in a study of dabigatran and warfarin. Drug Saf 2017; 40: 169–181

[6] Imbens G.. Matching Methods in Practice: Three Examples. IZA Discussion paper Nr. 8049. 2014;

[7] Imbens G, Wooldridge J. Recent developments in the econometrics of program evaluation. J Econ Lit 2009; 47: 5–86

[8] Linden A, Uysal D, Ryan A et al. Estimating causal effects for multivalued treatments: A comparison of approaches. Stat Med 2015; 35: 534–552

[9] Khandker S, Koolwal G, Samad H. Handbook of Impact Evaluation: Quantitative Methods and Practices.Washington DC: The International Bank for Reconstruction and Development / The World Bank; 2010 Available at: doi:https://doi.org/10.1596/978-0-8213-8028-4

[10] Holland PW. Statistics and Causal Inference. J Am Stat Assoc 1986; 81: 945–960

[11] Schafer J, Kang J. Average causal effects from nonrandomized studies: A practical guide and simulated example. Psychol Methods 2008; 13: 279–313

[12] Rubin D. Bayesian inference for causal effects: The role of randomization. Ann Stat 1978; 6: 34–58

[13] Ho D, Imai K, King G et al. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal 2007; 15: 199–236

[14] Stuart E, Lalongo N.. Matching methods for selection of participants for follow-up. Multivar Behav Res 2010; 45: 746–765

[15] Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41–55

[16] Rosenbaum P. Design of Observational Studies. New York: Springer; 2010; Available at: doi:https://doi.org/10.1007/978-1-4419-1213-8

[17] Austin P. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 2008; 27: 2037–2049

[18] Heckman J, Ichimura H, Todd P. Matching as an econometric evaluation estimator. Rev Econ Stud 1998; 65: 261–294

[19] Crump R, Hotz J, Imbens G et al. Dealing with limited overlap in estimation of average treatment effects. Biometrika 2009; 96: 187–199

[20] Traskin M, Small DS. Defining the study population for an observational study to ensure sufficient overlap: A tree approach. Stat Biosci. 2011; 3: 94–118

[21] Rosenbaum PR. model-based direct adjustment. J Am Stat Assoc 1987; 82: 387–394

[22] Rosenbaum P, Rubin D. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 1985; 39: 33–38

[23] Stuart E. Matching methods for causal inference: A review and a look forward. Stat Sci 2010; 25: 1–21

[24] Rubin D. Using propensity scores to help design observational studies: Application to the tobacco litigation. Health Serv Outcome Res Methodol 2001; 2: 169–188

[25] Austin P. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat 2011; 10: 150–161

[26] Austin P. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. Pharmacoepidemiol Drug Saf 2008; 17: 1218–1225

[27] Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J Am Stat Assoc 2018; 113: 390–400

[28] Linden A, Adams JL. Using propensity score-based weighting in the evaluation of health management programme effectiveness. J Eval Clin Pract 2010; 16: 175–179

[29] Emsley R, Lunt M, Pickles A et al. Implementing double-robust estimators of causal effects. Stata J 2008; 8: 334–353

[30] Imai K, King G, Stuart E. Misunderstandings between experimentalists and observationalists about causal inference. J R Stat Soc Ser A Stat Soc 2008; 171: 481–502

[31] Zhao Q, Percival D. Entropy Balancing is Doubly Robust. J Causal Inference 2017; 5: Available at: https://www.degruyter.com/view/j/jci.2017.5.issue-1/jci-2016-0010/jci-2016-0010.xml

[32] Austin P.. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. Am J Epidemiol 2010; 172: 1092–1097

[33] Imai K, Ratkovic M. Covariate balancing propensity score. J R Stat Soc Ser B Stat Methodol 2014; 76: 243–263

[34] Hainmueller J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Polit Anal 2012; 20: 25–46

[35] King G, Nielsen R. Why propensity scores should not be used for matching 2016 Available at: http://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-for-matching

[36] Linden A. Graphical displays for assessing covariate balance in matching studies. J Eval Clin Pract 2015; 21: 242–247

[37] Linden A. qqplot3: Stata module for plotting unweighted and weighted Q-Q plots. Available at http://ideas.repec.org/c/boc/bocode/s457856.html 2014

[38] Austin P. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009; 28: 3083–3107

[39] Hansen BB. The essential role of balance tests in propensity-matched observational studies: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. Stat Med 2008; 27: 2050–2054

[40] Freedman D, Berk R. Weighting regressions by propensity scores. Eval Rev 2008; 32: 392–409

[41] StataCorp LP. Stata Statistical Software: Release 15.College Station, TX: Stata Corporation; 2017. Available at: www.stata.com

[42] Hainmueller J, Xu Y. ebalance: A STATA Package for Entropy Balancing. J Stat Softw 2013; 54: 1–18

[43] Ho D, Imai K, King G et al. MatchIt: Nonparametric preprocessing for parametric causal inference. J Stat Softw 2011; 42: 2–28

[44] R Core Team R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available at: https://www.R-project.org/

[45] Linden A. covbal: Stata module for generating covariate balance statistics. 2016; Available at: http://ideas.repec.org/c/boc/bocode/s458188.html

[46] Quan H, Sundararajan V, Halfon P et al. Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data. Med Care 2005; 43: 1130–1139

[47] Frölich M. On the inefficiency of propensity score matching. AStA Adv Stat Anal 2007; 91: 279–290

[48] Marcus J. The effect of unemployment on the mental health of spouses – Evidence from plant closures in Germany. J Health Econ 2013; 32: 546–558

[49] McCaffrey D, Ridgeway G, Morral A. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 2004; 9: 403–425

[50] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984; 79: 516–524

[51] Rosenbaum P. Observational Studies. 2nd edition. New York: Springer; 2002