CHAPTER 5

IMPROVEMENT OF PROPENSITY SCORE METHODS

5.1 THE USE OF THE OVERLAPPING COEFFICIENT IN PROPENSITY SCORE METHODS

Edwin P. Martens[a,b], Wiebe R. Pestman[b], Anthonius de Boer[a], Svetlana V. Belitser[a] and Olaf H. Klungel[a]

[a] *Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*
[b] *Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*

Provisionally accepted by American Journal of Epidemiology

ABSTRACT

Propensity score methods focus on balancing confounders between groups to estimate an adjusted treatment or exposure effect. However, there is a lack of attention in actually measuring and reporting balance and also in using balance for model selection. We propose to use the overlapping coefficient in propensity score methods. First to report achieved balance on covariates and second to use it as an aid in selecting a propensity score model.

We demonstrated how the overlapping coefficient can be estimated in practical settings and performed simulation studies to estimate the association between the weighted average overlapping coefficient and the amount of bias. For various incidence rates and strengths of treatment effect we found an inverse relationship between the overlapping coefficient and bias, strongly increasing with sample size. For samples of 400 observations Pearson's correlation was only -0.10, while for $6,000$ observations -0.83 was found. Mainly for large samples the overlapping coefficient can be used as a model selection tool because its value is predictive for the amount of bias. For smaller data sets other methods can better be used to help selecting propensity score models, although an absolute quantification of balance will not be given by these methods.

Keywords: Confounding; Propensity scores; Observational studies; Measures for balance; Overlapping coefficient

INTRODUCTION

A commonly used statistical method to assess treatment effects in observational studies, is the method of propensity scores (PS).[1,2] PS methods focus on creating balance on covariates between treatment groups by first creating a PS model to estimate the conditional probability to be treated given the covariates (the propensity score). In the second step an adjusted treatment effect is estimated, using the propensity score as matching variable, as stratification variable, as continuous covariate or inverse probability weight. Traditionally, in the literature on PS methods there has been more attention for the second step (how to use the PS) than for the first (creating balance with the PS model). This is confirmed in recent literature reviews, where a lack of attention to building the PS model has been noticed.[3–5] Building such a PS model involves the (theoretical) selection of potential confounders and possibly transformations of these variables, higher-order terms or interactions with other covariates to include in the model. Because the objective of the PS model is to balance treatment groups on covariates and not to find the best estimates of coefficients, a check on balance on important prognostic covariates is important.[1,6] Unlike prediction models the selection of variables for a PS model (in which treatment is the dependent variable) is more complex: both the relationship with treatment and outcome has to be taken into account. In a recent study on variable selection it was confirmed that the PS model should contain variables related to both treatment and outcome and that it is better *not* to include variables that are only related to treatment because this will increase the standard error of the estimate.[7] Any stepwise regression method to build the PS model only selects on significance of the relationship with treatment, but does not use information on the strength of the relationship with the outcome. A strong relationship between treatment and covariates is not necessary for having a good PS model or good balance.[6,8] Such an example would be a PS model in a randomized trial: there will be a weak association between treatment and covariates, but still good balance exists.

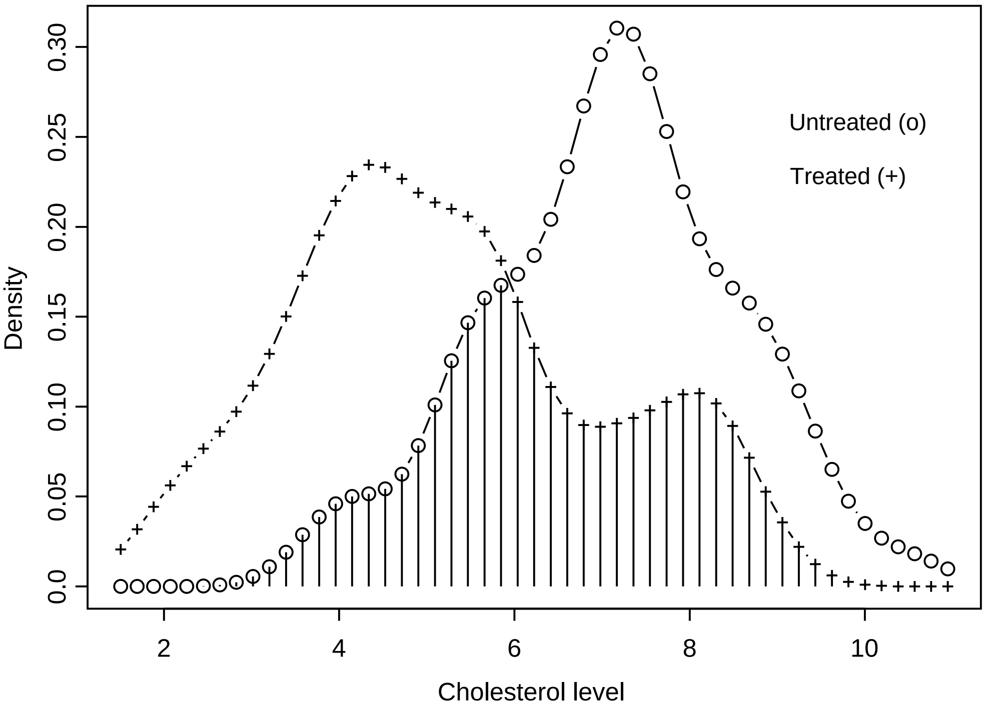
In applications of PS modeling the balance on covariates that has been reached has not been reported frequently or systematically.[4,5,9] When information on balance is given, this is mostly done by performing significance tests within strata of the PS to assure that the mean (or proportion) on covariates do not differ significantly between treatment groups. An early method proposed by Rosenbaum & Rubin to check the balance per covariate using the F-statistic from analysis of variance (ANOVA) is not often used.[10] More frequently the c-statistic (area under the receiver operating curve) is reported, but does not give the information needed: also a low value of the c-statistic can indicate good balance on important covariates (for instance in a randomized trial).

In this paper we propose to use a measure for balance in PS methods, known in the literature as the *overlapping coefficient* (OVL)[11,12] or proportion of similar responses.[13] This measure directly quantifies the overlap between a covariate distribution of the treated and the untreated. Its value gives information on the amount of balance that has been reached in a certain PS model. In the next section we give the definition of the non-parametric OVL. In the third section we show how this measure can be used in PS methods to check and report the amount of balance. In the fourth section we perform a simulation study in which the OVL has been used as a model selection tool and compare it with other approaches.

NON-PARAMETRIC OVERLAPPING COEFFICIENT

The concept of overlap of probability distributions fits the objectives of PS analysis. Without assuming any prediction model, one seeks to create balance on covariates. To understand the meaning of balance in this context, one can look at randomized studies. The randomization process guarantees that the *whole distribution* of all covariates is 'on average' similar between treatment groups. Any departure from similarity of distributions between treatment groups should be measured. In general, a statistical test on differences of group means is only a limited way of collecting information on the similarity of whole distributions. The question whether covariate distributions between treatment groups are similar (see Figure 5.1), can bet-

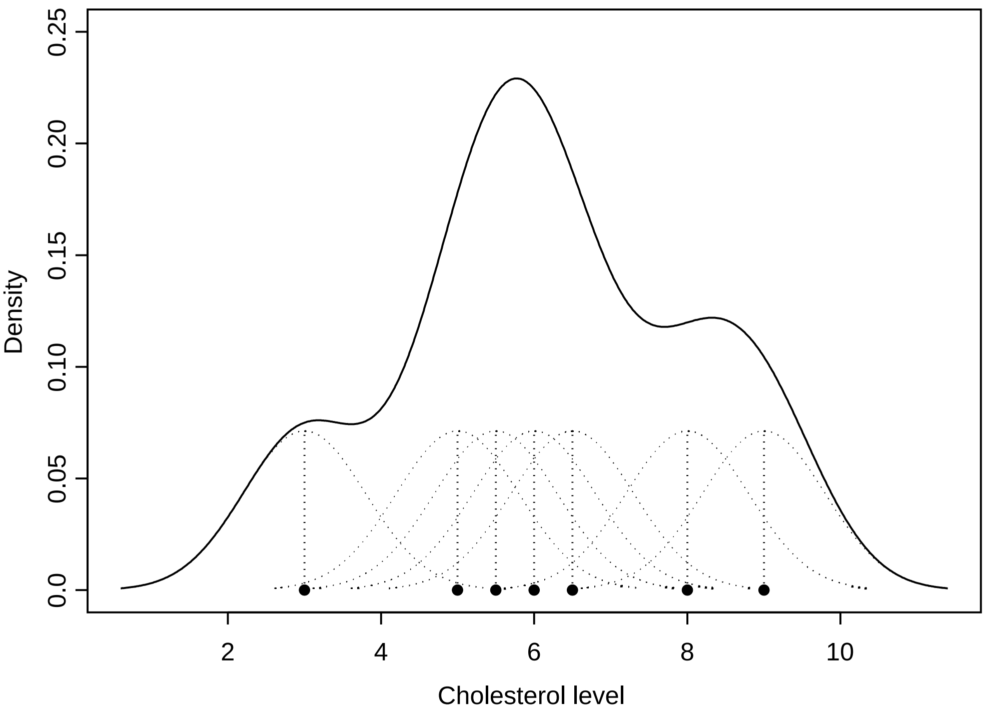
Figure 5.1: Illustration of the concept of overlap for the covariate cholesterol level in two random samples, one from a Gamma distribution (treated group, $n = 50$, $\mu = 6$ and $\lambda = 1$) and one from a normal distribution (untreated group, $n = 50$, $\mu = 7$ and $\sigma = 1.5$), using kernel density estimation



ter be answered by a measure for balance, the overlapping coefficient: it measures the amount of overlap in two distributions in a direct way and has a clear interpretation. The OVL is an estimate of that part of the distribution that overlaps with the other distribution.

In our situation the interest is in an estimate of the overlap between the covariable distributions of treated and untreated individuals. To estimate this overlap we first need to estimate the density of both distributions. It is not reasonable to assume any known theoretical distribution of covariates within subclasses of the propensity score. Therefore, we will estimate the densities in a non-parametrical way[14,15] by using kernel density estimation.[16,17] This can be seen as an alternative for making a histogram of the data with n observations. A kernel density is the sum of n density functions K, located at each observation with a chosen bandwidth h. With larger bandwidths the density function will be more smooth. There are different methods to find an optimal bandwidth. In Figure 5.2 kernel density estimation is illustrated for a small sample of 7 observations, using the normal density function for the kernel and a bandwidth determined by the normal reference rule method.[16]

Figure 5.2: Illustration of kernel density estimation in a sample of 7 cholesterol levels ($3, 5, 5.5, 6, 6.5, 8$ and 9) using the normal density kernel and the normal reference rule bandwidth



When for both treatment groups the density functions $\hat{f}(x|t = 0)$ and $\hat{f}(x|t = 1)$ are estimated, the OVL is the proportion of the density that overlaps with the other. Numerically we calculated this proportion with Simpson's rule using a 101 grid.[14] In Appendices B and C the S-Plus and SAS codes are given to implement the estimation of the OVL in practical settings.

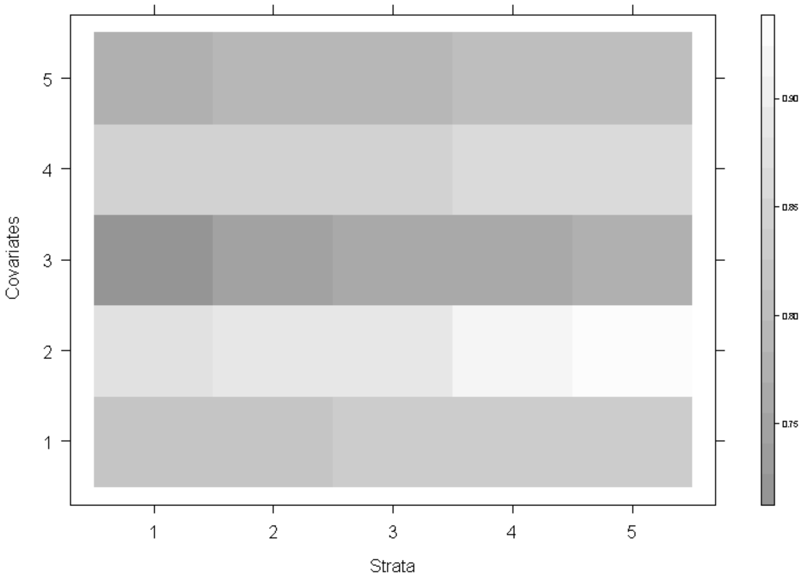
$$\widehat{OVL} = \int_{-\infty}^{\infty} \min\big\{\hat{f}(x|t=0), \hat{f}(x|t=1)\big\}dx \tag{5.1}$$

The influence on the OVL estimate of choosing other functions for the kernel, like Epanechnikov's kernel or fourth-order kernel,[18] other bandwidth methods or other grids is quite small.[14] It should be noted that in case of perfect overlap of both treatment groups in the population, the expectation of the OVL in a sample will be less than its maximum of 1. The variance of the OVL estimator can best be approximated by bootstrap methods, because even the derived formulas for normal distributions are in general too optimistic.[12]

OVERLAPPING COEFFICIENT TO CHECK AND REPORT BALANCE

In a PS analysis it is common practice to divide the sample in five groups based on the quintiles of the propensity score (strata). The OVL can be used to quantify the balance within strata on each of the prognostic factors, which gives a *distribution of estimated OVLs*. This information can be summarized in a cross table or in a graph like Figure 5.3.

Figure 5.3: Visualization of estimated overlapping coefficients for 5 normally distributed covariates within 5 strata of the propensity score, in a simulated data set of $n = 400$



The information as presented in this figure (where dark cells indicate low OVLs) can be used in several ways. First, one will be alerted on covariates or strata in which the balance is comparatively low, which could be improved for instance by including higher-order terms or interactions with other covariates. Second, the average OVL per covariate could be compared with the crude balance that exists on that covariate in the data set to see what *improvement in balance* has been reached by using PS stratification. Third, the information can give an answer to the important question whether the overall balance is *acceptable* in order to continue the analysis and estimate an adjusted treatment effect using the specified PS model. One way to do this, is to compare the whole distribution of estimated OVLs with a *reference distribution of OVLs*. A quite natural reference is the distribution of OVLs under the null hypothesis of similar covariate distributions between treatment groups, which is in fact the expected balance in a randomized trial. In Table 5.2 in Appendix A only the first decile of this distribution is given as a reference for various distributions and sample sizes. Fortunately, the OVL distribution is only slightly dependent on the shape of the covariate distribution. As an example we can use the data underlying Figure 5.3. The first decile estimated on these data equals 0.83 with an average number of observations of 40 per group and stratum. When this value is compared to the closest relevant figure in Table 5.2 (0.82, $n = 50$, normal distributions), it can be concluded that the balance after propensity score modeling is approximately comparable with the balance that would have been found if groups were in fact similar. Although one should strive for even better balance when possible, it gives at least an indication whether any acceptable balance has been reached.

A fourth way to use the balance information is to calculate a summary measure for the distribution of estimated OVLs on a wide range of possible PS models for helping to select a good PS model, a model with high overlap. To use this overall measure of balance for model selection there should exist a strong association with bias. In the next paragraph we give the results of a simulation study in which we investigated the strength of the relationship between a summary measure of balance and bias in estimating a treatment effect.

OVERLAPPING COEFFICIENT TO SELECT A PS MODEL

For a variety of PS models a distribution of estimated OVLs can be estimated which gives information on the amount of balance for that particular model. By relating a summary measure of this distribution to the amount of bias, we assessed the ability of the OVL to serve as a model selection tool in PS analysis. We also explored three other methods that could be used for model selection in PS analysis, but which are hardly used in practical settings. First, we used the *p*-values from *t*-tests, performed on all covariates within strata of the PS to detect a difference between treated and untreated individuals. Second, we used the method described in Rosenbaum & Rubin,[10] who regressed each of the covariates on treatment alone and on both treatment and PS, in order to compare both models with the *F*-statistic from ANOVAs. Third, we calculated the c-statistic for any PS model. Although its absolute value among different data sets is not indicative for the amount of balance, its relative value within data sets could be used to choose among various PS models.

METHODS

We simulated a population of $100,000$ individuals, a dichotomous outcome y, a dichotomous treatment t and 10 normally distributed covariates of which five were prognostic for the outcome ($x_1 - x_5$) and five were not ($x_6 - x_{10}$). First we simulated the distribution of the outcome ($\pi_y = 0.30$) and treatment ($\pi_t = 0.50$) and their relationship by means of the odds ratio ($OR_{ty} = 2.0$), making sure that all covariates were perfectly unrelated to treatment and moderately related to outcome ($x_1 - x_5, OR_{xy} = 1.3$) or not related to outcome ($x_6 - x_{10}, OR_{xy} = 1.0$). This enabled us to know the true marginal treatment effect without assuming any true outcome model (in the population no confounding). When sampling from this population, sample-specific confounding will appear which is in general small. To create stronger confounding in samples, which is common in observational studies, we gave to individuals different sampling probabilities that are related to treatment and to covariates x_1, x_2, x_4, x_6 and x_7. This resulted in unadjusted treatment effects OR_{unadj} between 1.5 and 4.1 with a mean of 2.55. An adjusted treatment effect was estimated with PS stratification, dividing the propensity score into five strata. When averaged over simulations, bias was defined as the percentage difference between the average adjusted treatment effect and the true treatment effect by means of the odds ratio.

To summarize the distribution of OVLs, and similarly the distributions of p-values from t-tests and ANOVAs, we used the first decile (10%), the first quintile (20%) and the median of these distributions. Thereby, we calculated for the OVL distribution an unweighted and a weighted average. The weighted average takes into account the estimated association between covariate i and outcome y ($\widehat{OR}_{x_i y}$) and is defined as:

$$\widehat{OVL}_w = \frac{1}{JI} \sum_{i=1}^{I} \sum_{j=1}^{J} w_i \widehat{OVL}_{ij} \tag{5.2}$$

where I is the number of covariates, J the number of strata, $\widehat{OVL}_{ij}$ the estimated OVL for covariate i in stratum j and

$$w_i = 1 + \widehat{OR}_{x_i y} - \frac{1}{I} \sum_{k=1}^{I} \widehat{OR}_{x_k y} \tag{5.3}$$

The weighted average OVL quantifies the idea that it is more important for strong prognostic factors to be balanced than for covariates that are weakly or not related with outcome.

To evaluate the relationship between bias and measure for balance we created 20 different PS models within each simulated data set, ranging from only three covariates to all 10 covariates with interactions. The simulations were done with various sample sizes ($n = 400, 800, 1,200, 1,600, 2,000, 4,000$ and $6,000$). Because results between simulations were fairly similar, a comparatively small number of 100 simulations per sample size was sufficient for a reliable estimate of the correlation.

For the final conclusion on the strength of the association between the measure for balance and bias, we used Pearson's correlation coefficients within simulations and sample sizes. We also performed an overall analysis on the results, i.e. a linear mixed-effects model (S-Plus function `lme`) with bias as the dependent variable, measure for balance as the fixed effect and simulation number as the random effect (random intercept only). As measure for model improvement we used Akaike's Information Criterion (AIC).

RESULTS

There exists an inverse relationship between the summary measures for the OVL distribution and the percentage of bias when estimating a treatment effect: the higher the measure for balance, the smaller the difference between estimated and true effect. The strength of this relationship is dependent on the chosen summary measure and on sample size. For example, for $n = 400$ Pearson's correlation was $R = -0.11$, while for $n = 6,000$ a correlation of $R = -0.83$ was found (see Table 5.1). The unweighted average and the percentile measures for the OVL showed somewhat smaller correlations (only first quintile shown in Table 5.1 column 2) than the weighted average OVL.

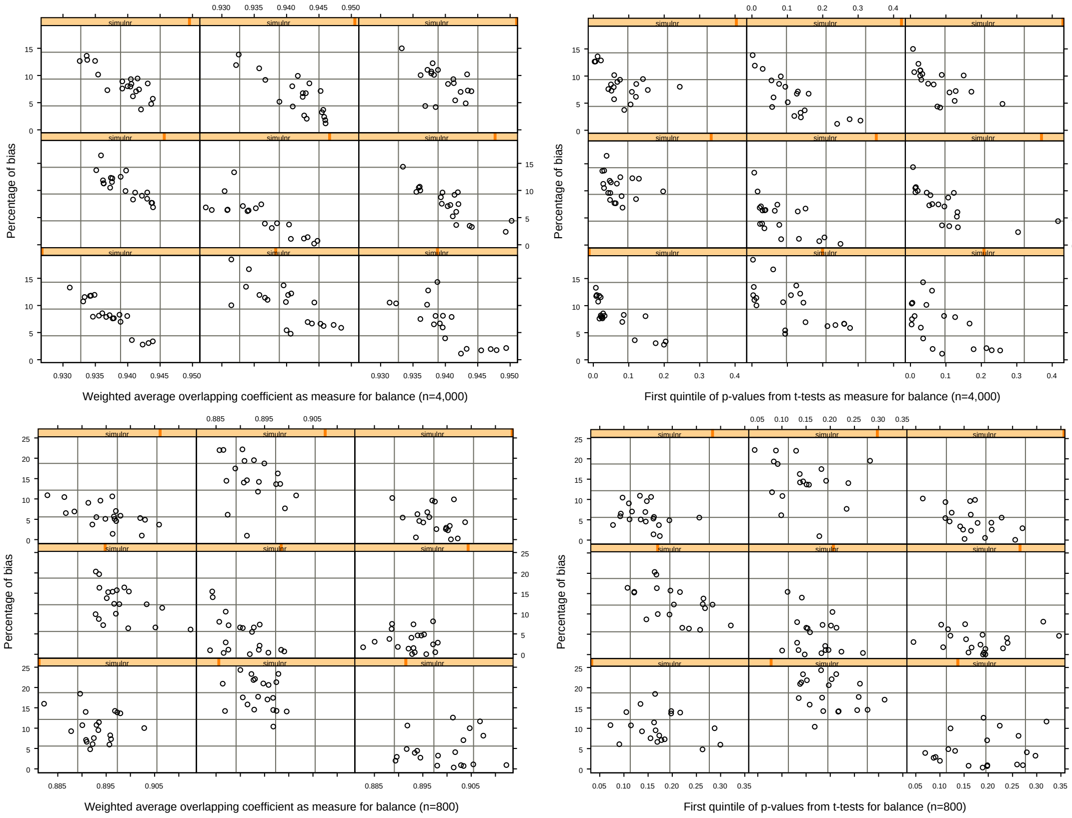
Table 5.1: Average Pearson's correlations (standard deviation) between bias and various summary measures of balance: weighted average and 1^{st} quintile (p_{20}) from OVL distribution, 1^{st} quintile from *p*-value distribution from *t*-tests, 1^{st} quintile from *p*-value distribution from ANOVAs and the *c*-statistic, 100 simulations per sample size

	OVL	OVL	*t*-tests	ANOVAs	
	weighted average	p_{20}	p_{20}	p_{20}	*c*-statistic
n=400	-0.11 (0.24)	-0.06 (0.22)	-0.28 (0.25)	-0.28 (0.31)	-0.26 (0.29)
n=800	-0.30 (0.29)	-0.17 (0.31)	-0.34 (0.28)	-0.27 (0.28)	-0.35 (0.25)
n=1,200	-0.38 (0.28)	-0.28 (0.28)	-0.42 (0.26)	-0.28 (0.33)	-0.39 (0.20)
n=1,600	-0.56 (0.23)	-0.42 (0.25)	-0.55 (0.23)	-0.40 (0.27)	-0.42 (0.25)
n=2,000	-0.59 (0.20)	-0.44 (0.24)	-0.59 (0.16)	-0.43 (0.24)	-0.34 (0.20)
n=4,000	-0.81 (0.09)	-0.68 (0.14)	-0.59 (0.18)	-0.32 (0.22)	-0.37 (0.22)
n=6,000	-0.83 (0.09)	-0.70 (0.15)	-0.58 (0.17)	-0.42 (0.22)	-0.44 (0.18)

In general, it can be concluded that other measures show higher correlations for sample sizes below 800 and lower correlations for sample sizes exceeding $1,600$. The correlations for the summary measure of the *p*-value distribution from ANOVAs range from -0.27 to -0.43 (first quintile shown in column 4), whereas the summary measures based on the distribution of *p*-values from *t*-tests are somewhat higher, ranging from -0.28 to -0.59 (first quintile shown in column 3). For sample sizes of 400 observations this measure is more predictive for bias than the OVL measure, while for samples between 800 and 1, 600 the differences between methods are small. We also performed linear mixed-effects models on the simulation results and found a similar pattern when the AICs were compared (results not shown).

As an illustration for these results, the relationship between bias and measure for balance is captured in Figure 5.4 for a random selection of nine samples of 800 and 4, 000. For samples of 4, 000 observations the predictive power for the weighted OVL (left upper panel) is larger than for the first quintile of the *p*-value distribution from *t*-tests (right upper panel). For sample sizes of 800 the fit is worse for both methods and slightly better for the *p*-values (right lower panel) than for the weighted OVL measure (left lower panel).

Figure 5.4: Association between average percentage of bias and weighted average OVL (left panels) and first quintile of *p*-value distribution from *t*-tests (right panels), within 9 random chosen samples of $n = 4,000$ (upper panels) and $n = 800$ (lower panels)



DISCUSSION

In observational studies that use propensity score analysis to estimate the effect of treatment or any exposure, we propose to focus more on the stage of creating the PS model by using a measure for balance, the overlapping coefficient. In the first place this measure for balance can be used to quantify balance in an absolute sense. Second, it can be used to judge whether the balance created on covariates with propensity score modeling was successful and sufficient to continue the analysis and estimate an adjusted treatment effect. Third, due to its inverse association with bias this measure can also be a help for model selection. The weighted average OVL calculated on the set of available covariates show strongest association with bias for larger data sets. For smaller data sets the p-values from significance tests and the c-statistic have higher predictive power for the bias than the OVL measure.

A disadvantage of the OVL is that it is estimated per covariate and per stratum. For small sample sizes and a large number of covariates this implies a great number of calculations with a small number of observations per stratum which can make the estimated overlap in densities unreliable. This effect will be partly diminished because the focus is on the whole distribution of OVLs. Also when the propensity score is divided in a larger number of strata, although five is common practice, estimation of OVLs becomes less reliable.

We focussed on the use of the OVL in propensity score stratification. When matching on the propensity score is chosen as method to estimate treatment effects, one could similarly estimate OVLs and compare models by using strata before matching takes place. After the matching one could also estimate OVLs on all covariates between both matched sets, but because this does not take into account the dependency of the data, it is not recommended.

We presented the results for a true population treatment effect of $OR_{ty} = 2.0$ and an incidence rate $\pi_y = 0.30$. The results were fairly robust against small changes in these values ($1.5 < OR_{ty} < 3.0$ and $0.10 < \pi_y < 0.40$). In all situations an increasing predictive power for the OVL measures was seen with increasing sample size. Compared to the other methods, the OVL measures performed best with larger data sets.

We summarized the distribution of p-values and OVLs over covariates and strata by using different methods: first decile, first quintile and median. Because the first quintile was most predictive, we only gave these figures in Table 5.1.

To know the true treatment effect in simulation studies it is common to first simulate the covariates, then the treatment variable as a function of the covariates (the propensity score model) and finally the outcome variable as a function of treatment and covariates (the outcome model). Our setting on the other hand was less restrictive because we did not need to specify any model for the outcome nor did we specify any propensity score model. Furthermore, specification of the outcome model using logistic regression can lead to a divergence in the type of effect to be estimated[19–22] with possibly misleading conclusions when comparison with propensity score methods is the main objective.[7,23,24] Finally, using the true propensity score model if it were known, is recommended, because a sample-specific propensity score model generates on average less bias than the true propensity score model.[10,25–27]

We propagate that in studies using propensity score methods, more attention should be paid to creating, measuring and reporting the balance that has been reached by the chosen propensity score model. The use of the overlapping coefficient could be a great help, also as a tool to select a PS model among a variety of possible models.

APPENDIX A: THE OVL DISTRIBUTION UNDER THE NULL

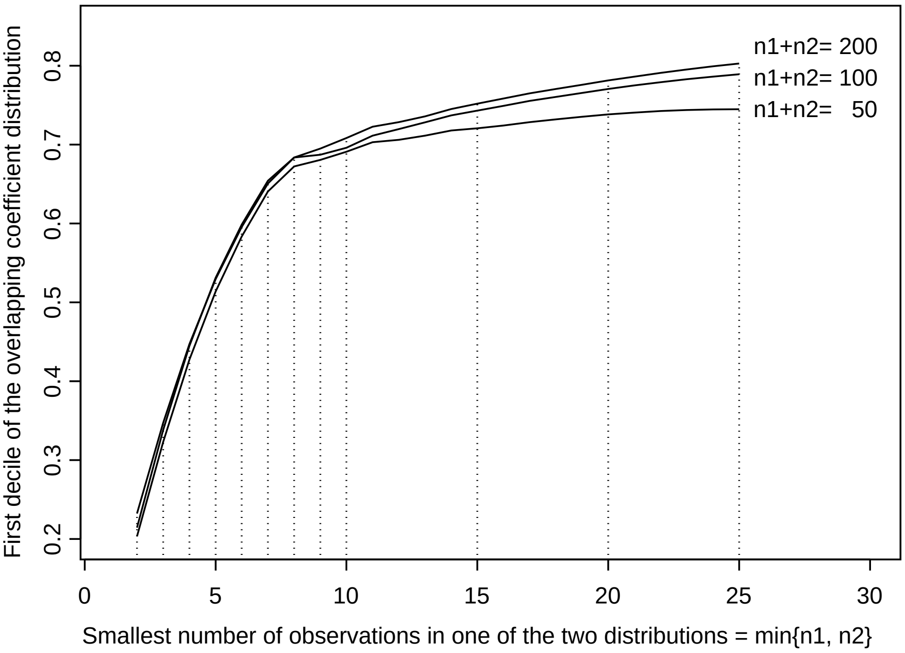
Under the null hypothesis that two samples come from populations with similar covariate distributions, we determined the OVL distribution by simulating $1,000$ samples with an equal number of observations from both populations. In Table 5.2 the first decile (10%) of this OVL distribution is given.

Table 5.2: First decile of the expected distribution of OVLs when both covariate distributions are similar for normal, chi-square, uniform, gamma and exponential distributions and various number of observations (n_i=number of observations in distribution i)

	normal	chi-square	uniform	gamma	exponential
$n_1 = n_2$	μ=0, σ=1	df=2	min=0, max=1	λ=3, μ=1	rate=2
25	0.74	0.70	0.78	0.74	0.71
50	0.82	0.79	0.83	0.81	0.79
100	0.87	0.85	0.88	0.86	0.84
200	0.90	0.89	0.91	0.90	0.89
400	0.93	0.92	0.93	0.92	0.92
800	0.94	0.94	0.95	0.94	0.94
1,600	0.96	0.96	0.96	0.96	0.96
3,200	0.97	0.97	0.97	0.97	0.97

When the first decile of the OVL distribution calculated on own data is higher than the tabulated value, this indicates that the balance is at least as good as could be expected when both groups are similar. Because the underlying distribution of covariates within strata is usually unknown, it is convenient that the values in Table 5.2 are quite similar among different distributions. Note that in this table the number of observations concern the numbers per treatment group within strata and are assumed to be equal ($n_1 = n_2$). When the number of observations is not the same for both groups, the expected OVLs will be lower. Below eight observations in one of the groups (irrespective of the number of observations in the other group), the left tail of the OVL distribution quickly reaches low values. This can be seen in Figure 5.5 for normal distributions. For instance, in case of similarity of groups with sample sizes of $n_1 = 4$ and $n_2 = 46$ in 10% of the cases an OVL will be found lower than 0.45. With such a low number of observations estimation of the overlapping coefficient is questionable.

Figure 5.5: First decile of overlapping coefficient distribution, calculated in 500 simulations with unequal sample sizes in both normal distributions, for $n_1 + n_2 = 50, 100$ and 200



APPENDIX B: S-PLUS CODE FOR ESTIMATING THE OVL

Calculating the OVL involves estimation of two density functions evaluated at the same x-values and then calculating the overlap. The function `ovl` needs two input vectors of observations on the covariate for both groups `(group0` and `group1)`. We used the S-Plus built-in function `density` using the normal density rule `bandwidth.nrd`. For calculation of the overlap we used Simpsons rule on a grid of 101. A plot of the two densities and the overlap is optional (`plot=T`).

```
# S-Plus Function to calculate the non-parametric overlapping coefficient
#                 (plus optional figure)
ovl <- function(group0, group1,plot=F){
    wd1     <- bandwidth.nrd(group1)
    wd0     <- bandwidth.nrd(group0)
    from    <- min(group1,group0) - 0.75 * mean(c(wd1,wd0))
    to      <- max(group1,group0) + 0.75 * mean(c(wd1,wd0))
    d1      <- density(group1, n = 101, width=wd1, from=from, to=to)
    d0      <- density(group0, n = 101, width=wd0, from=from, to=to)
    dmin    <- pmin(d1$y,d0$y)
    ovl     <- ((d1$x[(n<-length(d1$x))]-d1$x[1])/(3*(n-1)))*
               (4*sum(dmin[seq(2,n,by=2)])+2*sum(dmin[seq(3,n-1,by=2)])
                    +dmin[1]+dmin[n])
    if(plot){
        maxy    <- max(d0$y, d1$y)
        minx    <- min(d0$x)
        plot(d1, type="l", lty=1, ylim=c(0,maxy), ylab="Density",xlab="")
        lines(d0, lty=3)
        lines(d1$x, dmin, type="h")
        text(minx, maxy, "  OVL =")
        text(minx+0.085*(max(d1$x)-minx), maxy, round(ovl,3))
        }
    round(ovl,3)
    }

# Example
treated     <- rnorm(100,10,3)
untreated   <- rnorm(100,15,5)
ovl(group0=untreated, group1=treated, plot=T)
```

APPENDIX C: SAS CODE FOR ESTIMATING THE OVL

```
%macro OVL(group0, group1);
  proc univariate data=&group0 noprint;
    var var;
    output out=res0 n=n0 mean=mean0 var=var0 min=min0 max=max0 q1=var0_q1 q3=var0_q3;
  run;
  proc univariate data=&group1 noprint;
    var var;
    output out=res1 n=n1 mean=mean1 var=var1 min=min1 max=max1 q1=var1_q1 q3=var1_q3;
  run;
  data res;
    merge res0 res1;
    bandwidth0= 4*1.06 * min(sqrt(var0), (var0_q3 - var0_q1)/1.34) * n0**(-1/5);
    bandwidth1= 4*1.06 * min(sqrt(var1), (var1_q3 - var1_q1)/1.34) * n1**(-1/5);
    from = min(min0,min1) - 0.75 * mean(bandwidth0,bandwidth1);
    to   = max(max0,max1) + 0.75 * mean(bandwidth0,bandwidth1);
    call symput('from',from);
    call symput('to',to); run;
  PROC KDE data=&group0 ;
    univar var (ngrid=101 gridl=&from gridu=&to) /method=SNR out=dens0; run;
  PROC KDE data=&group1 ;
    univar var ( ngrid=101 gridl=&from gridu=&to) /method=SNR out=dens1; run;
  data dens;
    merge dens0(rename=(var=var0 value=val0 density=dens0 count=n0))
          dens1(rename=(var=var1 value=val1 density=dens1 count=n1)); run;
  data ovl;
    set   dens nobs=last;
    retain two_sums 0 x_first dens_min_first; keep ovl;
    dens_min=min(dens0,dens1);
    if _n_>=2 AND mod(_n_,2)=0 then two_sums = two_sums + 4 * dens_min;
    if _n_>=3 AND mod((_n_-1),2)=0 then two_sums = two_sums + 2 * dens_min;
    if _n_=1 then do;
        x_first = val0;
        dens_min_first = dens_min;
        end;
    if _n_=last then do;
        ovl = ((val0-x_first)/(3*(_n_-1)))*(two_sums + dens_min_first + dens_min);
        output;
    end; run;
  proc print data=ovl; run;
%mend OVL;

%OVL(untreated0, treated1);
```

REFERENCES

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[2] D'Agostino, RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, 17:2265–2281, 1998.

[3] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiol Drug Saf*, 14(4):227–238, 2005.

[4] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*, 58:550–559, 2005.

[5] Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*, 59:437–447, 2006.

[6] Rubin DB. On principles for modeling propensity scores in medical research (Editorial). *Pharmacoepidemiol Drug Saf*, 13:855–857, 2004.

[7] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection in propensity score models. *Am J Epidemiol*, 163:1149–1156, 2006.

[8] Rubin DB, Thomes N. Matching using estimated propensity score: relating theory to practice. *Biometrics*, 52:249–264, 1996.

[9] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*, 13(12):841–853, 2004.

[10] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *JAMA*, 387:516–524, 1984.

[11] Bradley EL. *Overlapping coefficient, in: Encyclopedia of Statistical Sciences, vol. 6*. Wiley, New York, 1985.

[12] Inman HF, Bradley EL. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Commun Statist -Theory Meth*, 18(10):3851–3874, 1989.

[13] Rom DM, Hwang E. Testing for individual and population equivalence based on the proportion of similar responses. *Stat Med*, 15:1489–1505, 1996.

[14] Stine RA, Heyse JF. Non-parametric estimates of overlap. *Stat Med*, 20:215–236, 2001.

[15] Cleveland WS, Devlin SJ, Grosse E. Regression by local fitting. *J Econometr*, 37:87–114, 1988.

[16] Silverman BW. *Density estimation for statistics and data analysis*. Chapman and Hall: London, 1986.

[17] Wand MP, Jones MC. *Kernel Smooting*. Chapman and Hall, 1995.

[18] Mammitzsch V Gasser T, Müller H-G. Kernels for nonparametric curve estimation. *J Royal Stat Soc, Series B*, 47:238–252, 1985.

[19] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431444, 1984.

[20] Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Statist -Theory Meth*, 20(8):2609–2631, 1991.

[21] Hauck WW, Neuhaus JM, Kalbfleisch JD, *et al.* A consequence of omitted covariates when estimating odds ratios. *J Clin Epidemiol*, 44:77–81, 1991.

[22] Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*, 125:761–768, 1987.

[23] Austin PC, Grootendorst P, Normand ST, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med*, 26:754–768, 2007.

[24] Martens EP, Pestman WR, Klungel OH. Letter to the editor: 'Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study, by Austin PC, Grootendorst P, Normand ST, Anderson GM'. *Stat Med*, 26:3208–3210, 2007.

[25] Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495, 1992.

[26] Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epideiol*, 150:327–333, 1999.

[27] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*, 158:280–287, 2003.

5.2 Measuring balance in propensity score methods

Edwin P. Martens[a,b], Wiebe R. Pestman[b], Anthonius de Boer[a], Svetlana V. Belitser[a] and Olaf H. Klungel[a]

[a] *Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*
[b] *Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*

Submitted for publication

ABSTRACT

Propensity score methods focus on balancing confounders between groups to estimate an adjusted treatment or exposure effect. However, there is a lack of attention in actually measuring, reporting and using the information on balance, for instance for model selection. We propose to use a measure for balance in propensity score methods and describe three such measures: the overlapping coefficient, the Kolmogorov-Smirnov distance and the Lévy metric.

We performed simulation studies to estimate the association between these measures for balance and the amount of bias. For all three measures we found an inverse relationship with bias increasing with sample sizes. The simulations further suggest that the predictive power for the overlapping coefficient was highest: for samples of 800 observations the average Pearson's correlation was -0.23, while for $2,000$ observations -0.63 was found. Mainly for large samples the overlapping coefficient can be used as a model selection tool because its value is predictive for the amount of bias. The mean squared error for these balancing strategies is quite similar among these methods, for the overlapping coefficient ranging from 0.031 for $n = 2,000$ to 0.197 for $n = 400$. This is much smaller than when a standard PS model including all covariates is applied (0.076 to 0.302). We conclude that these measures for balance are useful to report the amount of balance reached in any propensity score analysis and can be a help in selecting the final propensity score model.

Keywords: Confounding; Propensity scores; Observational studies; Measures for balance; Overlapping coefficient; Kolmogorov-Smirnov distance; Lévy metric

INTRODUCTION

A commonly used statistical method to assess treatment effects in observational studies, is the method of propensity scores (PS).[1,2] PS methods focus on creating balance on covariates between treatment groups by first creating a PS model to estimate the conditional probability to be treated given the covariates (the propensity score). In the second step an adjusted treatment effect is estimated, using the propensity score as matching variable, as stratification variable, as continuous covariate or inverse probability weight. In the literature in which PS methods are applied there is a lack of attention to building the PS model.[3–5] Building such a PS model involves the selection (and transformations) of variables and possibly interactions or higher-order terms to include in the model and a check whether the chosen model creates balance on the important prognostic covariates. Unlike prediction models the selection of variables for a PS model (in which treatment is the dependent variable) is more complex: both the relationship with treatment and outcome has to be taken into account. That means that model-building strategies like stepwise regression are not very useful in deciding whether a certain PS model is acceptable or not. Any stepwise regression method to build the PS model only selects on significance of the relationship *with treatment*, but this does not use information on the strength of the relationship *with outcome*. A strong relationship between treatment and covariates is not necessary for having a good PS model.[6,7] What should be important in PS models is the *balance on prognostic covariates* that has been reached by using the chosen PS model.

When information on balance is given, this is mostly done by performing significance tests within strata of the covariates to assure that the mean (or proportion) on covariates do not differ significantly between treatment groups. An early method proposed by Rosenbaum & Rubin to check the balance per covariate using the F-statistic from analysis of variance (ANOVA) is not often used.[8] More frequently the c-statistic (area under the receiver operating curve) is reported, but does not give the information needed: also a low value of the c-statistic can indicate good balance on prognostic factors (for instance in a randomized trial). We propose to use an overall weighted measure for balance to report the amount of balance reached and to select the final PS model.

In this paper we describe three measures for balance that can be used in PS methods: the *overlapping coefficient* (OVL),[9,10] also known as the *proportion of similar responses*,[11] the *Kolmogorov-Smirnov distance* (D)[12,13] and the *Lévy metric* (L).[14,15] In the next section we will define these measures. In the third section we give the results of a simulation study in which these measures are compared on their relationship with bias and on their ability in correctly estimating the treatment effect compared to a standard PS model.

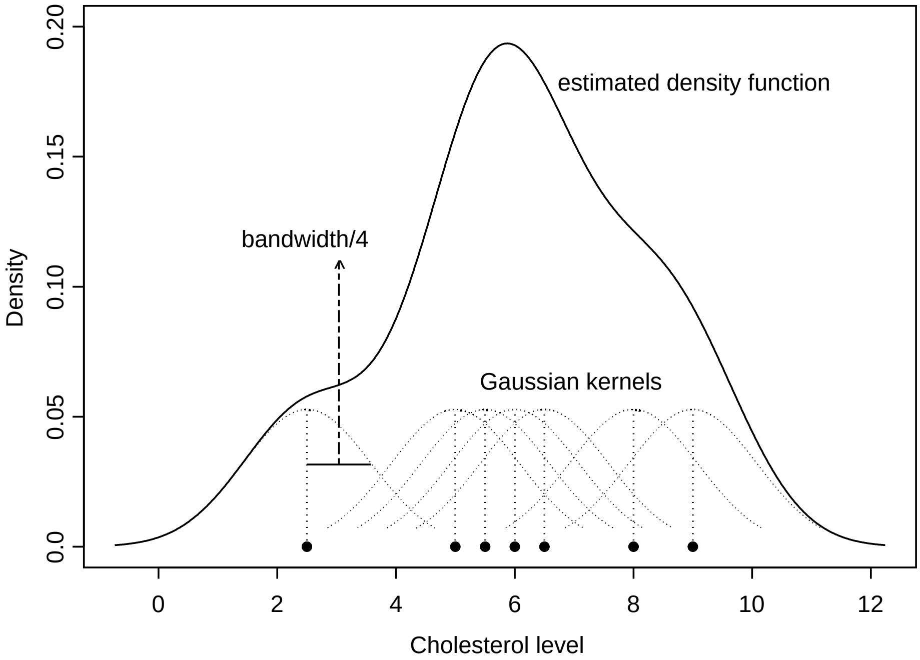
THREE MEASURES FOR BALANCE

The objective of PS methods is to create balance on the covariates that confound the relationship between outcome and treatment in observational studies. In randomized experiments this balance implies that the *whole distribution* of all covariates is 'on average' similar between treatment groups, not only the mean of the distribution or other summary measures. Whether or not covariate distributions of treatment groups are similar can best be approached by actually measuring the balance instead of testing whether the means of both distributions are significantly different. Any departure from similarity on prognostic factors could cause a difference in the outcome not caused by treatment. We will discuss three possible measures for balance.

NON-PARAMETRIC OVERLAPPING COEFFICIENT

The overlapping coefficient measures the amount of overlap in two distributions and is an estimate of that part of the distribution that overlaps with the other distribution. To estimate the overlapping coefficient we first need to estimate the density of both distributions. Because it is not reasonable to assume any known theoretical distribution of covariates within subclasses of the propensity score, we will estimate the densities in a non-parametrical way[16,17] by using kernel density estimation.[18,19] This can be seen as an alternative for making a histogram of the

Figure 5.6: Illustration of kernel density estimation in a sample of 7 cholesterol levels ($2.5, 5, 5.5, 6$, $6.5, 8$ and 9) using the normal density function for the kernel and the normal reference rule bandwidth

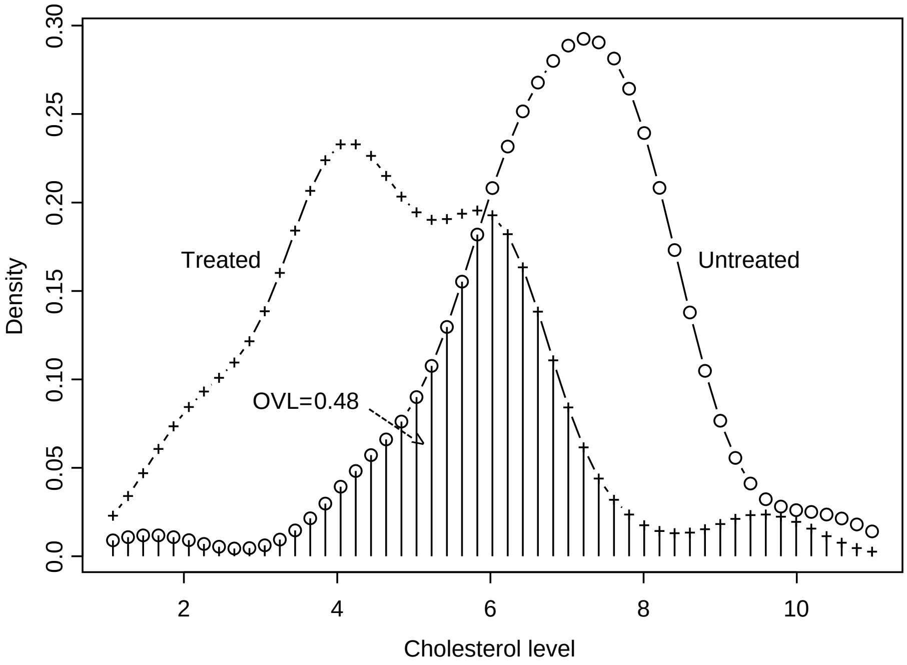


data with n observations. A kernel density is the sum of n density functions K, located at each observation with a chosen bandwidth h. Increasing the bandwidth will make the density function more smooth. In Figure 5.6 kernel density estimation is illustrated for a small sample of 7 observations, using the normal density function for the kernel and a bandwidth determined by the normal reference rule method.[18] When for both treatment groups ($t = 0$ is untreated, $t = 1$ is treated) the density functions $\hat{f}(x|t = 0)$ and $\hat{f}(x|t = 1)$ are estimated, the OVL is the proportion of one density that overlaps with the other. Numerically we calculated this proportion with Simpson's rule using a 101 grid.[16]

$$\widehat{OVL} = \int_{-\infty}^{\infty} \min\big\{\hat{f}(x|t = 0), \hat{f}(x|t = 1)\big\}dx \tag{5.4}$$

The influence on the OVL estimate of choosing other functions for the kernel, like Epanechnikov's kernel or fourth-order kernel,[20] other bandwidth methods or other grids is quite small.[16] Note that in case of perfect overlap of both treatment groups in the population, the expectation of the OVL in a sample will be less than 1. The variance of the OVL estimator can best be approximated by bootstrap methods, because even the derived formulas for normal distributions are in general too optimistic.[10] In Figure 5.7 the overlapping coefficient is illustrated for a

Figure 5.7: Illustration of the overlapping coefficient for cholesterol level in two random samples, treated group drawn from a Gamma distribution ($n = 50$, $\mu = 6$ and $\lambda = 1$) and untreated group from a normal distribution ($n = 50$, $\mu = 7$ and $\sigma = 1.5$), using kernel density estimation



ted group of 50 observations drawn from a Gamma distribution with $\mu = 6$ and $\lambda = 1$ and an untreated group of 50 observations drawn from a normal distribution with $\mu = 7$ and $\sigma = 1.5$. The OVL is calculated at 0.48. In Appendix A the S-Plus code for the OVL is given.

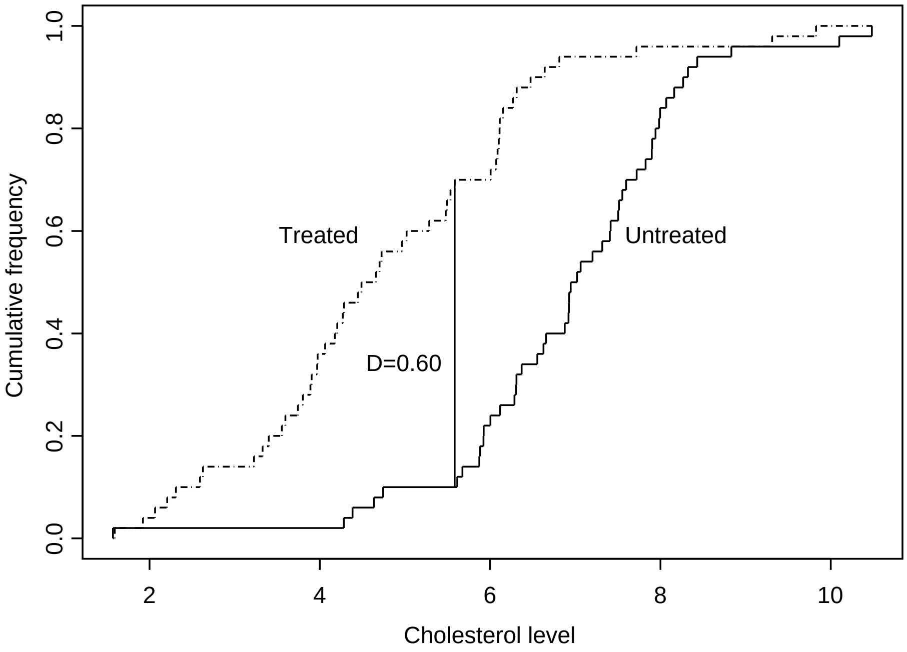
KOLMOGOROV-SMIRNOV DISTANCE

The Kolmogorov-Smirnov distance D can be described as the maximum of all vertical distances between two cumulative distribution functions, expressed as relative frequencies. The minimum distance of 0 will be reached when both distributions are exactly similar. The larger this measure, the less similar distributions are, with a maximum of 1. This distance is also used for the difference between an empirical and a known theoretical distribution. The Kolmogorov-Smirnov distance D between untreated and treated individuals is defined as

$$\hat{D} = \max\left\{\left|\hat{F}(x|t=0) - \hat{F}(x|t=1)\right|\right\} \tag{5.5}$$

where $\hat{F}(x|t = 0)$ is the estimated cumulative distribution function for untreated individuals and $\hat{F}(x|t = 1)$ for treated individuals. An illustration of D as a measure for balance is given in Figure 5.8 for the same data as used for Figure 5.7. In Appendix A the S-Plus code for the Kolmogorov-Smirnov distance is given.

Figure 5.8: Illustration of the Kolmogorov-Smirnov distance D for cholesterol level in two random samples, treated group drawn from a Gamma distribution ($n = 50$, $\mu = 6$ and $\lambda = 1$) and untreated group from a normal distribution ($n = 50$, $\mu = 7$ and $\sigma = 1.5$)



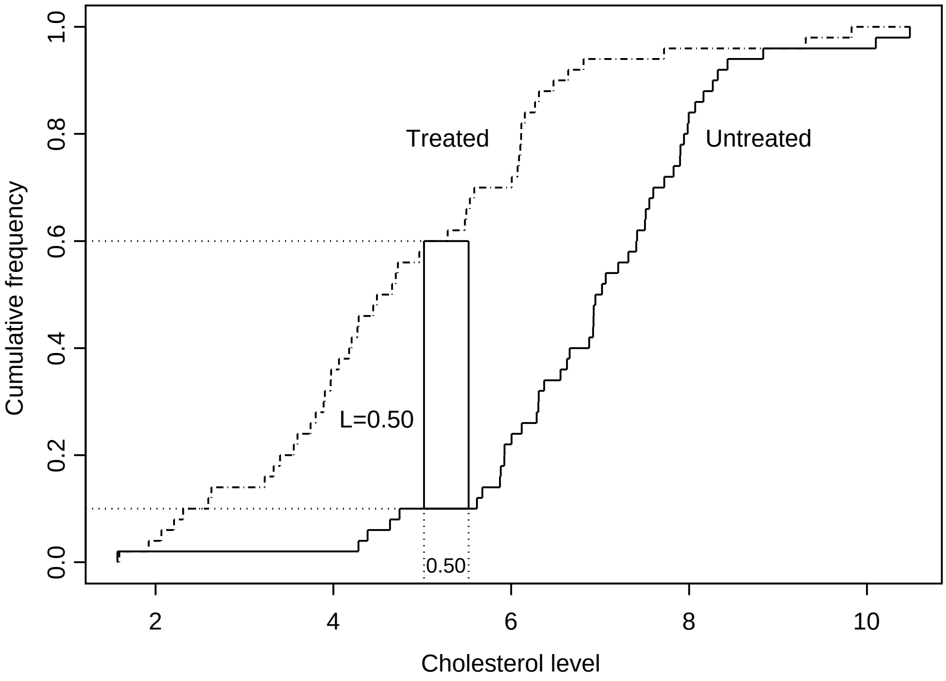
LÉVY METRIC

The Lévy metric L can be considered as a variant on the Kolmogorov-Smirnov distance that takes into account both horizontal and vertical distances between two cumulative distribution functions. The Lévy metric L is defined as

$$\hat{L} = \min\left\{\epsilon > 0 \middle| \hat{F}(x-\epsilon|t=0) - \epsilon \leq \hat{F}(x|t=1) \leq \hat{F}(x+\epsilon|t=0) + \epsilon \text{ for all x in } \mathbb{R}\right\} \quad (5.6)$$

where $\hat{F}(x|t=0)$ is the estimated cumulative distribution function for untreated individuals and $\hat{F}(x|t=1)$ for treated individuals. Intuitively this measure can be understood as follows: if one inscribes squares between the two curves with sides parallel to the coordinate axes, then the side-length of the largest such square is equal to L. An illustration of L as a measure for balance is given in Figure 5.9 where the same data have been used as in Figures 5.7 and 5.8. From this figure it becomes clear that this distance measure is sensitive for the unit of measurement of the covariate. When different covariates are involved one should therefore use some kind of standardization of the covariates before these measures can be compared. In Appendix A the S-Plus code for the Lévy metric is given.

Figure 5.9: Illustration of the Lévy metric L for cholesterol level in two random samples, treated group drawn from a Gamma distribution ($n = 50$, $\mu = 6$ and $\lambda = 1$) and untreated group from a normal distribution ($n = 50$, $\mu = 7$ and $\sigma = 1.5$)



BALANCE MEASURES AS MODEL SELECTION TOOLS

In the first place the described measures for balance can be used to quantify and report the amount of balance reached in any PS analysis, something that is not often done when PS methods are applied.[3–5] Because theory on PS states that within strata of the propensity score distributions of covariates tend to be similar,[8] insight in the actual balance can be given by reporting this balance per covariate and stratum.

Another way to use the information on balance is when a selection among several PS models has to be made. The best PS model can be defined as the model that estimates a treatment effect as close as possible to the true treatment effect in the population. In practical settings this PS model is unknown and a selection of variables and additional terms (for instance interactions and/or quadratic terms) must be made for choosing the final PS model. Standard variable selection methods like forward or backward regression can not be used for PS models, because first the association between outcome and covariates is not taken into account and second, *balance* is the final objective of a PS model and not significance. When balance on prognostic covariates in one model is better than in the other, the one with the best balance should be preferred because in theory the estimated treatment effect has been better adjusted for imbalance of covariates.

In the previous paragraph we generally described three measures that quantify either the degree of overlap or the distance between cumulative distribution functions. To use these measures for model selection we calculated these for all covariates within strata of the propensity score, where the strata were based on the quintiles of the PS.[8] To get an overall measure for balance for every fitted PS model, we calculated a *weighted average* of the measures per covariate and stratum, with weights equal to the strength of the association between covariate and the outcome (on the log-odds scale). These weights express the idea that balance on strong prognostic factors is more important in estimating an adjusted treatment effect than on factors that are only weakly related to the outcome. This implies that a *high* value on the weighted average overlapping coefficient and *low* values on the weighted average Kolmogorov-Smirnov distance and the Lévy metric indicate good balance on prognostic factors. To find out to what degree this balance is related to bias, we performed a simulation study to compare these measures for balance on their ability to select PS models.

METHODS

For our simulations we used the framework of Austin,[21] also extensively described in Austin *et al.*[22,23] First nine normally distributed covariates $x_1 - x_9$ were simulated of which six were related to treatment t according to the following logistic model, including four interactions and two quadratic terms:

$$\text{logit}(p_{i,t}) = \beta_{0,t} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_4 + \beta_4 x_5 + \beta_5 x_7 + \beta_6 x_8 + \\ \beta_7 x_2 x_4 + \beta_8 x_2 x_7 + \beta_9 x_7 x_8 + \beta_{10} x_4 x_5 + \beta_{11} x_1^2 + \beta_{12} x_7^2 \quad (5.7)$$

and where treatment was simulated by a Bernouilli distribution with $\pi = p_{i,t}$. The outcome y was simulated by the following logistic model, including covariates $x_1 - x_6$ and treatment t, including four interactions and two quadratic terms:

$$\text{logit}(p_{i,y}) = \alpha_{0,y} + \beta_t t + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + \alpha_6 x_6 + \\ \alpha_7 x_2 x_4 + \alpha_8 x_3 x_5 + \alpha_9 x_3 x_6 + \alpha_{10} x_4 x_5 + \alpha_{11} x_1^2 + \alpha_{12} x_6^2 \quad (5.8)$$

Compared to the model of Austin we added four interactions and two quadratic terms to both the treatment model and the outcome model. The dichotomous outcome was generated by a Bernouilli distribution with $\pi = p_{i,y}$.
From equations 5.7 and 5.8 it can be deduced that:
the true confounding factors were $x_1, x_2, x_4, x_5, x_2x_4, x_4x_5$ and x_1^2,
the factors only related to treatment were x_7, x_8, x_2x_7, x_7x_8 and x_7^2,
the factors only related to outcome were x_3, x_6, x_3x_5, x_3x_6 and x_6^2.

The strength of the associations was:
for β_7, β_{10}, α_7, α_{10} equal to log(1.2),
for β_1, β_3, β_5, α_1, α_2, α_3, β_8, β_{11}, α_8, α_{11} equal to log(1.4),
for β_9, β_{12}, α_9, α_{12} equal to log(1.6)
for β_2, β_4, β_6, α_4, α_5, α_6 equal to log(2.0).

To assure that half of the subjects were treated and that an event occurred ($y = 1$) for approximately 25% of the untreated individuals, $\beta_{0,t}$ was set to -0.65 and $\alpha_{0,y}$ to -1.8.

An important feature of these simulations is that only a *conditional* effect β_t can be inserted as the true treatment effect, while with PS methods we aim to estimate a *marginal* treatment effect. For the difference between marginal and conditional treatment effects in logistic regression analysis, we refer to the literature.[24–28] Because we wanted to restrict the simulations to a true marginal treatment effect of $OR_{ty} = 2.0$, we had to find the corresponding conditional effect in this setting. We used the iterative process described in Austin[21] to calculate the true conditional treatment effect β_t to be equal to 0.8958, which equals an OR of 2.449. For our simulations we varied sample size ($n = 400, 800, 1,200, 1,600$ and $2,000$).

From the large number of possible PS models we sampled for every simulated data set at random 40 models, calculated the PS, stratified on the PS and calculated for the overlapping coefficient, the Kolmogorov-Smirnov distance and the Lévy metric the weighted average for all these PS models. After calculating the relative bias as $\widehat{OR}_{ty}/OR_{ty} - 1$, we first used Pearson's correlation coefficient within simulations to determine the strength of the association between the measure for balance and bias. We also performed an overall analysis on the results, i.e. a linear mixed-effects model (S-Plus function `lme`) with bias as the dependent variable, measure for balance as the fixed effect and simulation number as the random effect (random intercept only). As measure for model improvement we used Akaike's Information Criterion (AIC).

In the second part we concentrated on the selection of only one PS model, i.e. the model that gave the best balance according to the measure for balance. We used the mean squared error because it directly combines average bias and the spread of the estimator, defined as:

$$\frac{1}{S}\sum_{s=1}^{S}(\widehat{OR}_{ty} - OR_{ty})^2 \tag{5.9}$$

where S is the number of simulations, $\widehat{OR}_{ty}$ the estimated treatment effect and OR_{ty} the true marginal treatment effect which was set in our simulations to 2.0. We compared the mean squared error among the three measures for balance, but also used three other PS models as a reference that include:

1. *all covariates* that are related to either treatment or outcome ($x_1 - x_8$),
2. all *true confounding factors* (x_1, x_2, x_4, x_5, interactions x_2x_4, x_4x_5, quadratic term x_1^2),
3. *all prognostic factors* as given in formula 5.8.

The last two PS models are in practical settings unknown and are used in this simulation only as theoretical models. On the other hand, knowing the true propensity score model is in general not very interesting: including factors that are only related to treatment can be disadvantageous in practice and an estimated propensity score performs better than the true propensity score.[8,29–32]

RESULTS

For all sample sizes the average Pearson's correlation coefficient between the weighted overlapping coefficient and bias is higher than for the other measures, except for sample sizes of 400 observations for which neither of the measures is rather predictive for the amount of bias (Table 5.3). For example, for a sample size of $1,600$ the average correlation for the weighted OVL was -0.61, for the weighted Kolmogorov-Smirnov distance 0.40 and for the weighted Lévy metric 0.46. As a comparison we used the *c*-statistic, for which much smaller correlations were found (at most -0.27). We also checked the correlations for the method proposed by Rosenbaum & Rubin using *F*-statistics from ANOVAs,[8] which were quite similar as for the *c*-statistic (ranging from -0.07 to -0.23). Apart from averaging correlations among simulations, we also performed an overall linear mixed-effects model and calculated AICs. We have chosen to present only the results for the correlations because it gave similar results and has a more direct interpretation of association than the AIC.

Table 5.3: Average Pearson's correlations between bias and various summary measures of balance: weighted average overlapping coefficient, weighted average Kolmogorov-Smirnov distance and weighted average Lévy metric, the *c*-statistic, 200 simulations per sample size

	overlapping coefficient	Kolmogorov- Smirnov distance	Lévy metric	*c*-statistic
n= 400	-0.06	-0.06	-0.04	-0.09
n= 800	-0.23	0.08	0.10	-0.12
n=1,200	-0.39	0.16	0.20	-0.24
n=1,600	-0.61	0.39	0.43	-0.27
n=2,000	-0.63	0.40	0.46	-0.24

In Table 5.4 the results are given when these measures for balance are used to select the model that has best balance according to that measure (columns 1-3). For the overlapping coefficient the mean squared error is 0.031 when the number of observations is $2,000$ and 0.197 for a sample size of 400. The differences between the other two measures for balance are quite small. When we compare the results of the fixed model strategy of including all covariates, either related to outcome or treatment (column 4) with the OVL measure, the mean squared error is considerably larger, ranging from 0.076 to 0.302. This PS model is commonly chosen when PS methods are adopted in practice.[33] The model that contains all confounding factors (column 5) has a slightly higher mean squared error (around 20%) than for the OVL measure, while for the fixed model strategy containing all prognostic factors (column 6) is comparable to the OVL measure. The conclusion is that the measures for balance have lower mean squared error than a commonly used PS model and are slightly less or comparable to models that contain the true factors. Because the true confounding and true prognostic factors are usually unknown in practice, it means that these methods are capable of selecting PS models that are at least as good as when the true confounding and true prognostic factors were known.

Table 5.4: Mean squared error of different methods: weighted average overlapping coefficient, weighted average Kolmogorov-Smirnov distance, weighted average Lévy metric, covariates x_1 to x_8, all confounding factors, all prognostic factors, 200 simulations per sample size

	overlapping coefficient	Kolmogorov- Smirnov	Lévy metric	covariates $x_1 - x_8$	confounding factors	prognostic factors
n= 400	0.197	0.200	0.228	0.302	0.208	0.187
n= 800	0.076	0.085	0.083	0.161	0.103	0.089
n=1,200	0.047	0.061	0.058	0.119	0.065	0.057
n=1,600	0.035	0.038	0.038	0.094	0.051	0.045
n=2,000	0.031	0.035	0.034	0.076	0.036	0.032

When the PS model has been chosen by the *c*-statistic or the F-statistic approach the mean squared error was approximately 30% larger (ranging from 0.05 to 0.25).

Another frequently adopted approach to adjust for confounding that is not based on PS modeling is a multivariable logistic regression analysis. For a model that included treatment and all prognostic factors, the mean squared error ranged from 0.099 for $n = 2,000$ to 0.470 for $n = 400$, which is considerably larger than when PS methods are applied. This should be no surprise, because with logistic regression analysis the conditional treatment effect (=$exp(\beta_t) = 2.449$) is estimated, which is in general not the effect of interest and an overestimation of the marginal treatment effect.[24–28]

DISCUSSION

In observational studies that use propensity score analysis to estimate the effect of treatment or any exposure, we propose to focus more on the stage of creating the PS model by directly quantifying the amount of balance. Examples of such measures are the overlapping coefficient, the Kolmogorov-Smirnov distance and the Lévy metric. These measures can be used to report the amount of balance and can be useful for selecting the final PS model. For all three measures we showed an inverse association with bias, which was strongest for the weighted average overlapping coefficient. This association was stronger for larger than for smaller samples, for the overlapping coefficient $R = -0.06$ for n=400 and $R = -0.63$ for $n = 2,000$. Selecting the PS model with the overlapping coefficient seems to be most effective, because the mean squared error for this method was in general smallest (ranging from 0.031 to 0.197). The differences with the Kolmogorov-Smirnov distance and the Lévy metric were only minor. The PS model that contained all covariates had a considerable larger mean squared error, while the PS model that contained all true, but usually unknown, confounding factors had somewhat larger mean squared error.

The use of these measures should not replace the common sense of epidemiologists who should select, observe and measure those covariates that are potentially confounding factors. When faced with a choice of functional form or possible interactions, it can be worthwhile to use one of these measures to select the final PS model in order to have best balance and probably least bias.

Some remarks can be made about the choice among the three presented measures. First, it seems that the Lévy metric and the Kolmogorov-Smirnov distance give quite similar results, which makes the latter to be preferred because no standardization is needed. The choice between the overlapping coefficient and the Kolmogorov-Smirnov distance is more difficult to make. The overlapping coefficient has a clearer interpretation and performed best in these simulations. On the other hand, for its calculation a bandwidth and a kernel has to be chosen which may influence the estimated value.

A disadvantage of the proposed methods is that these measures must be estimated per covariate and per stratum. For small sample sizes and a large number of covariates this implies a great number of calculations with a small number of observations per stratum which can make the estimated overlap in densities unreliable. Previously we showed that estimation of the overlapping coefficient is not valid when there are less than 8 observations in one of the distributions.[34]

We focused on the use of the OVL in propensity score stratification. When matching on the propensity score is chosen as method to estimate treatment effects, one could similarly calculate OVLs and compare models by using strata before matching takes place. After the matching one could also calculate OVLs on all covariates between both matched sets, but because this does not take into account the dependency of the data, it is not recommended.

We propagate that in studies using propensity score methods, more attention should be paid to creating, measuring and reporting the balance that has been reached by the chosen propensity score model. The use of the overlapping coefficient and the Kolmogorov-Smirnov distance could be a great help, also as a tool to select a PS model among a variety of possible models.

APPENDIX A: S-PLUS CODE FOR OVERLAPPING COEFFICIENT, KOLMOGOROV-SMIRNOV DISTANCE AND LÉVY METRIC

OVERLAPPING COEFFICIENT

Calculating the OVL involves estimation of two density functions evaluated at the same x-values and then calculating the overlap. The function `ovl` needs two input vectors of observations on the covariate for both groups `(group0` and `group1)`.We used the S-Plus built-in function `density` using the normal density rule `bandwidth.nrd`.For calculation of the overlap we used Simpsons rule on a grid of 101. A plot of the two densities and the overlap is optional (`plot=T`).

```
# S-Plus Function to calculate the non-parametric overlapping coefficient
#                 (plus optional figure)
ovl <- function(group0, group1,plot=F){
    wd1     <- bandwidth.nrd(group1)
    wd0     <- bandwidth.nrd(group0)
    from    <- min(group1,group0) - 0.75 * mean(c(wd1,wd0))
    to      <- max(group1,group0) + 0.75 * mean(c(wd1,wd0))
    d1      <- density(group1, n = 101, width=wd1, from=from, to=to)
    d0      <- density(group0, n = 101, width=wd0, from=from, to=to)
    dmin    <- pmin(d1$y,d0$y)
    ovl     <- ((d1$x[(n<-length(d1$x))]-d1$x[1])/(3*(n-1)))*
               (4*sum(dmin[seq(2,n,by=2)])+2*sum(dmin[seq(3,n-1,by=2)])
                    +dmin[1]+dmin[n])
    if(plot){
        maxy    <- max(d0$y, d1$y)
        minx    <- min(d0$x)
        plot(d1, type="l", lty=1, ylim=c(0,maxy), ylab="Density",xlab="")
        lines(d0, lty=3)
        lines(d1$x, dmin, type="h")
        text(minx, maxy, "  OVL =")
        text(minx+0.085*(max(d1$x)-minx), maxy, round(ovl,3))
        }
    round(ovl,3)
    }

# Example
treated     <- rnorm(100,10,3)
untreated   <- rnorm(100,15,5)
ovl(group0=untreated, group1=treated, plot=T)
```

KOLMOGOROV-SMIRNOV DISTANCE

Within S-Plus the Kolmogorov-Smirnov distance can be simply extracted from the object generated by the function `ks.gof` by `ks.gof(group0,group1)$stat`.A function that optionally plots the two cumulative densities is given below (`plot=T`).

```
# S-Plus function to calculate the Kolmogorov-Smirnov distance using
# function ‘ks2’ from within function ‘ks.gof’ (plus optional figure)
ksdist <- function(group0, group1, plot=F){
          n0      <- length(group0)
          n1      <- length(group1)
          total   <- sort(unique(c(group0, group1)))
          ma0     <- match(group0, total)
          ma1     <- match(group1, total)
          F0      <- cumsum(tabulate(ma0, length(total)))/n0
          F1      <- cumsum(tabulate(ma1, length(total)))/n1
          diff    <- abs(F0-F1)
          ks      <- max(diff)
          if(plot){
             x.ks    <- order(ks-diff)[1]
             plot(F1, type="l", lty=1, ylab="Cumulative density", xlab="")
             lines(F0, lty=3)
             lines(c(x.ks, x.ks), c(F0[x.ks],F1[x.ks]), lty=2)
             text(0.08*(n0+n1), 1, "K-S distance =")
             text(0.20*(n0+n1), 1, ks)
             }
             ks
         }

# Example
treated     <- rnorm(100,10,3)
untreated   <- rnorm(100,15,5)
ksdist(group0=untreated, group1=treated, plot=T)
```

LÉVY METRIC

The Lévy metric can be calculated using the next two functions `mecdf` and `Levy`.

```
# S-Plus functions to calculate the L\'evy metric
mecdf <- function(group0,group1) {
           n0      <- length(group0)
           n1      <- length(group1)
           total   <- sort(unique(c(group0, group1)))
           ma0     <- match(group0, total)
           ma1     <- match(group1, total)
           F0      <- cumsum(tabulate(ma0, length(total)))/n0
           F1      <- cumsum(tabulate(ma1, length(total)))/n1
           min     <- min(F1-F0)
           max     <- max(F1-F0)
           m       <- c(min,max)
           return(m)
           }
Levy <- function(u,v){
         f <- function(s,u,v){
             t <- mecdf(u,v-s)+s
             return(t[1])
             }
         g <- function(s,u,v){
             t <- mecdf(u,v+s)-s
             return(t[2])
         }
         a <- min(c(u,v))
         b <- max(c(u,v))
         c <- b-a
         z1 <- uniroot(f,low=-c,up=c,tol=0.00000001,u=u,v=v)
         z2 <- uniroot(g,low=-c,up=c,tol=0.00000001,u=u,v=v)
         z <- max(z1$root,z2$root)
         return(z)
         }

# Example
treated     <- rnorm(100,10,3)
untreated   <- rnorm(100,15,5)
# mecdf(untreated,treated)
Levy(group0=untreated, group1=treated)
```

REFERENCES

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[2] D'Agostino, RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, 17:2265–2281, 1998.

[3] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiol Drug Saf*, 14(4):227–238, 2005.

[4] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*, 58:550–559, 2005.

[5] Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*, 59:437–447, 2006.

[6] Rubin DB, Thomes N. Matching using estimated propensity score: relating theory to practice. *Biometrics*, 52:249–264, 1996.

[7] Rubin DB. On principles for modeling propensity scores in medical research (Editorial). *Pharmacoepidemiol Drug Saf*, 13:855–857, 2004.

[8] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *JAMA*, 387:516–524, 1984.

[9] Bradley EL. *Overlapping coefficient, in: Encyclopedia of Statistical Sciences, vol. 6*. Wiley, New York, 1985.

[10] Inman HF, Bradley EL. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Commun Statist -Theory Meth*, 18(10):3851–3874, 1989.

[11] Rom DM, Hwang E. Testing for individual and population equivalence based on the proportion of similar responses. *Stat Med*, 15:1489–1505, 1996.

[12] Stephens MA. Use of the Kolmogorov-Smirnov, Cramer-von Mises and related statistics without extensive tables. *J Royal Stat Soc Series B*, 32:115–122, 1970.

[13] Pestman WR. *Mathematical Statistics*. Walter de Gruyter, Berlin, New York, 1998.

[14] Lévy P. *Théorie de l'addition des variables aléatoires*. Gauthier-Villars, 1937.

[15] Zolotarev VM. Estimates of the difference between distributions in the Lévy metric. *Proc Steklov Inst Math*, 112:232240, 1973.

[16] Stine RA, Heyse JF. Non-parametric estimates of overlap. *Stat Med*, 20:215–236, 2001.

[17] Cleveland WS, Devlin SJ, Grosse E. Regression by local fitting. *J Econometr*, 37:87–114, 1988.

[18] Silverman BW. *Density estimation for statistics and data analysis*. Chapman and Hall: London, 1986.

[19] Wand MP, Jones MC. *Kernel Smooting*. Chapman and Hall, 1995.

[20] Mammitzsch V Gasser T, Müller H-G. Kernels for nonparametric curve estimation. *J Royal Stat Soc, Series B*, 47:238–252, 1985.

[21] Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*, 2007. On line: DOI: 10.1002/sim.2781.

[22] Austin PC, Grootendorst P, Normand ST, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med*, 26:754–768, 2007.

[23] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*, 26:734–753, 2007.

[24] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431444, 1984.

[25] Gail MH. The effect of pooling across strata in perfectly balanced studies. *Biometrics*, 44:151163, 1988.

[26] Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Statist -Theory Meth*, 20(8):2609–2631, 1991.

[27] Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Control Clin Trials*, 19:249–256, 1998.

[28] Martens EP, de Boer A, Pestman WR, Belitser SV, Klungel OH. An important advantage of propensity score methods compared to logistic regression analysis. *in review*.

[29] Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495, 1992.

[30] Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epideiol*, 150:327–333, 1999.

[31] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*, 158:280–287, 2003.

[32] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection in propensity score models. *Am J Epidemiol*, 163:1149–1156, 2006.

[33] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*, 13(12):841–853, 2004.

[34] Martens EP, de Boer A, Pestman WR, Belitser SV, Klungel OH. The use of the overlapping coefficient in propensity score methods. *in review*.