Biocybernetics and Biomedical Engineering
2009, Volume 29, Number 4, pp. 41–61

Using Propensity Score with Receiver Operating Characteristics (ROC) and Bootstrap to Evaluate Effect Size in Observational Studies

MACIEJ GÓRKIEWICZ*

Jagiellonian University in Krakow, Health Sciences Faculty,
Department of Epidemiology and Population Research, Kraków, Poland

In non-randomised studies, prioritisation of patients who are most likely to benefit from more expensive and more effective treatments usually take place and/or patients select themselves to treatments. Propensity score methods have been considered as means to reduce the effect of selection bias. In this study it was shown that use of receiver operating characteristics (ROC) and area under ROC (AUC) provides an additional insight into analysis of non-randomised studies. The estimates of mean effect obtained with five different techniques were compared and nonparametric bootstrap was recommended as superior tool for propensity score analyses.

K e y w o r d s: effect size, non-randomised study, propensity score, bootstrap, online calculator

1. Introduction

The notion of propensity score was introduced by Rosenbaum and Rubin [1] and defined as the predicted conditional probability of an individual being assigned to a particular treatment in an observational study given his or her baseline measurements. The main motivation declared by Rosenbaum and Rubin [1] was to provide an alternative method for adjusting treatment effects to given individual baseline measurements in the two sample design (e.g. treated versus non-treated group) when treatment assignment is not random, but can be assumed to be independent of expected outcomes. In observational studies, the effect of selection bias can distort

* Correspondence to: Maciej Górkiewicz, Jagiellonian University in Krakow, Health Sciences Faculty, Department of Epidemiology and Population Research, ul. Grzegórzecka 20, 31-531 Kraków, Poland, e-mail: gorkiewicz@poczta.fm
Received 10 October 2008; accepted 28 Juli 2009

results, because prioritisation of patients who are most likely to benefit from diverse ways of therapy usually takes place here. Besides, often the patients select themselves to treatments, e.g. taking into account the necessary costs and expected effects of a therapy. Nevertheless, in the literature one can find some support for high agreement of authors' conclusions in pairs of randomised trials and non-randomised studies with similar settings, population, interventions, and outcomes, see e.g. [2, 3].

In practice, estimation of the propensity score given a known individual treatment and baseline measurements for study participants was carried out either by averaging proportion of the treated observations at clusters of participants that were similar in their baseline features or by logistic regression used there to find a linear combination of the baseline features which best discriminates between treated and non-treated groups, [4, 5]. Then, the estimated individual values of propensity were applied to adjust treatment effect either simply as an additional feature that can be investigated jointly with the remaining ones with the aim of achieving of balance between study groups over all baseline characteristics under consideration, or strictly at frame of the propensity score methods as a sole classification variable with the aim of reducing of selection bias in observational studies for causal effects [4, 5]. There the question arises, in what extent the findings from the propensity score analyses could be considered as an additional support for evidence-based clinical rules [6], or even as some alternative to the randomised controlled trials [7]. With regard to this question, aiming at enlarging of trust in the results of propensity score analysis, in this study the use of two known statistical procedures, that is resampling (bootstrapping) technique and receiver operating characteristics (ROC) method was proposed.

In practice the main reason of investigating of the usefulness of bootstrapping for studies on treatment effect was that these studies tend to generate data that have bounded and skewed distributions, so the standard methods of analysis that assume normality may not be appropriate [8, 9]. In this study the potential benefits of the bootstrap for propensity score analyses in improving effect size reliability were supported with illustrative example. The six different methods were used to estimate the treatment effect basing on the same data from an observational study. Two of six compared methods, the ordinary-*t* confidence intervals, and the meta-analysis are apart from the propensity score approach, although the meta-analysis was applied to the same strata as the compared propensity score procedures. The next three methods of the propensity score analysis, that is basic method [1], modified method [4], and estimation with use of bias-corrected bootstrap were applied to four sub-samples (clusters) of study participants defined by compact strata at their baseline measurements. The last method used the patient's baseline measurements to estimate propensity score with logistic regression [5]. It was showed that the propensity score methodology leads to intermediate width of confidence intervals for estimates, between the smallest ordinary-*t* confidence interval and the largest one for meta-analysis procedure, but the direct use of the bootstrap permits smaller confidence intervals among four

compared methods of the propensity score analyses. It should be pointed out that two ways of use of the bootstrap either to clusters of participants in way showed in this study, or to matching participants with respect to strata at the propensity score in manner showed at [10], both need in principle the same amount of data as the standard procedures [1, 4].

The propensity score analyses need rather intensive and durable cooperation between medical experts and statisticians. The procedures proposed in this paper can make it easier. Generally, in our opinion, it is inadmissible to apply here a typical work-sharing: first clinicians or epidemiologists prepare the data base, and after that statisticians start with their analyses of these data.

The main reason for trying to apply a ROC technique is a presume that if stratified random sampling is practised for almost all participants of a study, at least for weighty share of these, and all influenced predictors for the propensity score are measured at baseline, then the propensity score analysis should lead to estimates comparable with findings from randomised trials, [11], at least with findings from randomised trials with poor concealment of allocation, [12]. Alas, the needed information cannot be usually obtained with literature search only [13, 14], and the domain experts must be recruited to some auxiliary inquiries. First, one can seek advice from medical professionals with aim to disclose the limits for allocation decisions in daily practice, and to reveal other properties of real decision processes [15–17]. Then, in the considered absence of certain external information the various test-retest procedures, see e.g. [18–21], can be used to estimate the underlying propensity to undertake random decisions. In test-retest experiments the use of graphical and audio-video information was suggested, see [22–24]. Nevertheless, the main weakness both of the interviewing experts and the test-retest procedures is that the experts can perceive these investigations as a kind of academic exam without clear relation to their daily professional activity. So, in this study it was proposed to attempt utilize information from usual in clinical practice analyses made after therapy, if available, e.g. analyses of applicability of the acknowledged clinical guidelines, see e.g. [25].

This paper is structured as follows. First, the key terms and concepts of the propensity score methodology are introduced. Then the exclusion procedures, frequently indispensable at the propensity score analyses, are discussed with aim at supporting necessity in getting a supplementary information from medical experts. At the next section, a motivating example from real-life clinical practice is examined with descriptive statistics and with results of several ways of analysis. Then the possible benefits from use of three procedures for getting a supplementary information, that is: structured interview, procedure of repeated arrangements, and the ROC (receiver operating curve) technique, are briefly explained. In discussion the focus is made on necessity of intensive clinicians-statisticians cooperation during the propensity score analyses. The final conclusions are concentrated on possible ways for enlarging of trust in estimates obtained with the propensity score procedures.

2. Key Terms and Concepts

This section is concerned in notions of two core stages of the propensity score method, that is estimation of the propensity score stage, and stage of the estimation of adjusted treatment effect. Then, with reference to these notions the concepts of receiver operating characteristics (ROC) are briefly explained.

Consider a single investigation aimed at comparing of two treatments, say an experimental treatment versus a standard treatment or placebo, with respect to a chosen single outcome. Let S, $s_i \in S$; $i = 1, 2, ..., N$, be considered a sample of N all participants of this study that met defined criteria to be taken into consideration, and then to be not excluded from further analyses aimed at obtaining the propensity score estimates. Let $A \in (0, 1)$ be a dichotomous allocation variable indicating whether an i-th individual got an experimental treatment ($a_i = 1$) or not ($a_i = 0$).

Let $X = (X_1, X_2, ..., X_L)$ be set of L measurable features assumed to affect the allocation variable A. Assume values of A and X known at baseline for each $s_i \in S$; $i = 1, 2, ..., N$; so a sample S divided without remains into two non-overlapping sub-samples $S = S_1 \cup S_0$; $S_1 \cap S_0 = 0$; where: if $a_i = 1$ then $s_i \in S_1$; if $a_i = 0$ then $s_i \in S_0$. The propensity score (PS) was defined with (1) as estimated conditional probability to be allocated to treated sub-sample S_1; given values of features X.

$$PS_i = E(Pr((s_i \in S_1) \mid (x_1, x_2, \ldots, x_L)); \qquad i = 1, 2, ..., N; \tag{1}$$

where: E – symbol of expectation.

Numerous methods applied in practice to estimate PS can be divided into two groups. The first group of methods corresponds to logistic regression approach [5, 26], but the second group of clustering methods [4, 27], bases on notion of similarity (or contrary: dissimilarity, distance) between individuals with respect to their individual values of features X. A basic logistic model implies that expected probability at (1) is equal to $1/(1 + \exp(-L(X))$, where $L(X)$ is a linear combination (2) of variables X, [26].

$$L(X) = b_0 + \Sigma b_1 * X_1.; \qquad 1 = 1, 2, \ldots, L. \tag{2}$$

where: b_0, b_1 – constant coefficients.

All methods from the second group, in spite of seemingly distinct differences between them, can be summarised there jointly with the formula (3):

$$PS_i = \Sigma a_j * w_{ij} / \Sigma w_{ij}; \qquad i, j = 1, 2, ..., N \tag{3}$$

where: w_{ij} – weight of j-th individual with respect to i-th individual; $0 \leq w_{ij} \leq 1$; $w_{ii} = 1$.

In case of usual, non-overlapping clusters, the weights at (3) are dichotomous, that is either $w_{ij} = 1$ or $w_{ij} = 0$; moreover, $w_{ij} = w_{ji}$. For nearest-neighbourhood ap-

proach the weights at (3) are dichotomous too, $w_{ij} = 1$ or $w_{ij} = 0$; but there it is admitted $w_{ij} \neq w_{ji}$. The kernel method implies continuous values of weights, usually symmetrical: $w_{ij} = w_{ji}$. The weights w_{ij} are usually defined basing on analyses of distance matrices or neighbouring graphs, [27].

Generally, logistic regression can lead to non-fractional estimates of propensity score, either $PS = 0$, or $PS = 1$, only in limits, that is if a linear combination (2) tends to infinity, either $L(X) = -\infty$, or $L(X) = +\infty$. Contrary to this, the formula (3) in practice often leads to estimates $0 \leq PS \leq 1$. With aim of better explanation of the formula (3) let us first consider a case of non-overlapping clusters. Let some considered i-th individual belongs to cluster C; $s_i \in C \subset S$. Then for all individuals from this cluster the weights are equal to 1, but for all individuals from outside they are equal to 0: if $(s_i \in C) \wedge (s_j \in C)$ then $w_{ij} = 1$; otherwise $w_{ij} = 0$; $j = 1, 2, ..., N$. In a particular case it is possible that all individuals from some cluster C have the same value of the allocation variable $A = a^*$, either $a^* = 1$, or $a^* = 0$. In such a particular case the formula (3) leads there to estimate $PS_i = a^*$, so either $PS_i = 1$, or $PS_i = 0$. In other case, if the cluster C includes individuals of various values of allocation variable A then the formula (3) must lead there to a fractional estimate of PS. The analogous consideration can be made in a case of the nearest-neighbourhood approach, changing notion of cluster with notion of neighbourhood. The only difference is that for non-overlapping clusters the formula (3) must lead to the same estimates of PS for all members of the same cluster, but members of the same neighbourhood can obtain different estimates (3).

The next stage of the propensity method consists of adjusting treatment effect to the estimated propensity score. It should be pointed out that generally, in result of some exclusions, this adjustment is made using only some sub-sample $U \subset S$.

In the non-overlapping clusters approach it was presumed that the considered clusters differ in their propensity scores, but somewhat surprisingly the estimates obtained with (3) usually aren't there in direct use, [1, 4]. First, let us use the notion of formula (3) to formulate definition for treatment effect D proposed originally at [1]. In the non-overlapping clusters approach each k-th individual $s_k \in U$, $k = 1, 2, \ldots, K$, belongs to only single cluster of individuals, and number m_k of members of this cluster is equal to $m_k = \Sigma w_{kj}$, where sum should be get at $s_k \in U$; $k = 1, 2, \ldots, K$; so share of this cluster at sample U is equal to $P_k = m_k / K$. Let Y be a continuous variable expressing individual outcome, assumed to be a result of therapy or exposition under study. The mean outcomes at individuals with $a = 1$, and with $a = 0$, are given with formula (4) and (5) respectively.

$$\text{Mean}(y_k \mid a = 1) = \Sigma a_j^* \, w_{kj}^* \, y_j \, / \, \Sigma a_j^* \, w_{kj}; j = 1, 2, \ldots, K. \tag{4}$$

$$\text{Mean}(y_k \mid a = 0) = \Sigma (1 - a_j)^* \, w_{kj}^* \, y_j \, / \, \Sigma (1 - a_j)^* \, w_{kj}; j = 1, 2, \ldots, K. \tag{5}$$

It is easy to notice, that homogenous clusters of individuals that have the same value of the allocation variable, either $a = 1$, or $a = 0$, must be excluded from this way of analysis.

The mean effect D_k associated with a single individual $s_k \in U$ is given with the formula (6). Finally, the summary mean effect D for all clusters, weighted with their share $P_k = m_k / K$ as it has been proposed at [1], is given with the formula (7).

$$D_k = (\text{Mean}(y_k \mid a = 1) - \text{Mean}(y_k \mid a = 0)) / \Sigma w_{kj}; j = 1, 2, \ldots, K. \quad (6)$$

$$D = \Sigma(D_k * P_k) = (\text{Mean}(y_k \mid a = 1) - \text{Mean}(y_k \mid a = 0)) / K; k = 1, 2, \ldots, K. \quad (7)$$

It should be pointed out that for all four above formulas (4), (5), (6), and (7), summarisation should be made for all K individuals $s_j \in U, j = 1, 2, \ldots, K$; but generally not at all N individuals from the initial sample $s_j \in S, j = 1, 2, \ldots, N$.

At [4] it was proposed to standardize mean effect D not proportionally to the share of clusters P_k, see formula (7), but proportionally to the inverse of the variance V_k of difference (Mean($y_k \mid a = 1$) – Mean($y_k \mid a = 0$) at the particular clusters, see formulas (4) and (5). To this end for each particular cluster the pooled variance V_k must be estimated, either with resampling methodology [28], suitable to case of frequent exclusions from initial cluster, or with standard methods, see formulas (8), (9), (10).

$$V_k (y_k \mid a = 1) = \Sigma a_j * w_{kj} * (\text{Mean}(y_k \mid a = 1) - y_j)^2 / \Sigma a_j * w_{kj}; j = 1, 2, \ldots, K. \quad (8)$$

$$V_k (y_k \mid a = 0) = \Sigma(1 - a_j) * w_{kj} * (\text{Mean}(y_k \mid a = 0) - y_j)^2 / \Sigma(1 - a_j) * w_{kj}; j = 1, 2, \ldots, K. \quad (9)$$

$$V_k = V_k (y_k \mid a = 1) + V_k (y_k \mid \text{a} = 0); \quad k = 1, 2, \ldots, K. \quad (10)$$

where Mean($y_k \mid a = 1$) and Mean($y_k \mid a = 0$) are computed with (4) and (5); V_k is a pooled variance at k-th cluster.

Finally, following [4], formula (7) changed there to form of (11).

$$D = (\Sigma(D_k / V_k)) / \Sigma (1 / V_k); \quad k = 1, 2, \ldots, K. \quad (11)$$

The formulas (4), (5), (6), and (7), or alternatively (8), (9), (10) and (11), can be applied properly in frame of the nearest-neighbourhood and kernel approaches.

The extremely different approach to estimating a treatment effect consists in one-to-one matching only on the base of individual estimates of the propensity score. The most of higher-level statistical packages such as Stata, SAS, or SPSS, include ready to use procedures of this approach, usually with so named calliper and greedy options [5], [11]. As a base sample for matching the smaller sub-sample S_1 or S_0 is used; where: if $a_i = 1$ then $s_i \in S_1 \subset S$; if $a_j = 0$ then $s_j \in S_0 \subset S$. Without loss of generality let us assume that a sub-sample S_1 was chosen. The procedure for one-to-one matching chooses step-by-step a single individual $s_i \in S_1$, reads his/her PS_i and then looks for couple of this individual among individuals $s_j \in S_0$, according to the criterion of minimal difference $|PS_i - PS_j|$, then both matched individuals, $s_i \in S_1$ and $s_j \in S_0$, are removed from the pool. This process should be repeated until matches are found for all individuals $s_i \in S_1$. The remaining individuals $s_j \in S_0$, are excluded from

the further estimation of the mean effect D. The option calliper gives opportunity to define threshold T_{acc} defining the maximal acceptable difference $|PS_i - PS_j|$, so if none of $s_j \in S_0$ meet the restriction $|PS_i - PS_j| < T_{acc}$ then $s_i \in S_1$ is removed from the further estimation of the mean effect D. The option named greedy gives opportunity to define a sequence of thresholds T, say $T = 0, 0.01, 0.05, 0.10$. Then, step-by-step from the minimal to maximal threshold, the procedure looks for pairs of individuals that meet the restriction $|PS_i - PS_j| < T$. It is easy to notice, that maximal threshold plays there a role of a threshold T_{acc}. Finally, the mean effect D is estimated with any standard procedure for correlated pairs of the continuous variable Y. In the above procedures some dilemma arises if the criterion of minimal difference $|PS_i - PS_j|$ can be met at several pairs of individuals, $s_i \in S_1$ coupled with $s_j \in S_0$. This difficulty can be easy overcome with standard weighting approaches for one-to-many matching, or with bootstrap procedure [10].

The one-to-many matching approach takes into account only individual estimates of PS, but the clustering approach, only features X, nevertheless the both approaches can be united at single procedure in various different ways, [5, 11]. For example, within strata defined with the estimates of PS only, e.g. with a sequence of the above greedy thresholds T, the individuals can be matched according to the criterion of minimal distance defined on the base of features X, e.g. parametric Mahalanobis distance [29] or using non-parametric procedures, like [30]. Then, the propensity score estimates can be used as an additional variable together with the features X with aim of creating the clusters of similar individuals. In result of these and other unifying approaches a great variety of propensity score procedures arose in practice, [5, 11].

Receiver operating characteristics (ROC) are an acknowledged tool to examine the usefulness of a certain measurable variable to differentiate between two specified classes, in the medicine by tradition named a Positive and a Negative class [31]. In this paper a role of classifying variable played the estimated propensity score PS. Then, by definition, let a Positive class includes all individuals that certainly need to be treated with the first treatment under consideration, but a Negative class includes all remaining individuals. So, with respect to this, a sample S under consideration is divided onto two sub-samples, positives C_P versus negatives C_N, $S = C_P \cup C_N$; $C_P \cap C_N = 0$. Then, let for some assumed threshold $PS = T$ it be hypotheses that if for i-th individual $s_i \in S$ his/her propensity score $PS_i \geq T$ then this individual is decided to be included to treated group, $s_i \in S_1$; otherwise $s_i \in S_0$. With respect to both above classifications, two fractions can be recognized inside a sub-sample S_1, a true positive fraction TP that includes all $s_i \in S_1 \cap C_P$ and a false positive fraction FP that includes all $s_i \in S_1 \cap C_N$. The ROC is a two-dimensional curve that displays relationship between true positive ratio (TPR) and false positive ratio (FPR) across all from all possible dichotomous thresholds T at classifying variable under consideration, where TPR (respectively: FPR) represents proportion of a true positive fraction TP (respectively: a false positive fraction FP) with relation to all positives C_P (respectively: to all negatives C_N).

At process of ROC's estimation two main stages appeared, at the first stage for each given value of classifying variable a pair of estimates (TPR^, FPR^) is calculated. Then, at the second stage, the ROC is estimated as continuous convex curve between (TPR=0, FPR=0) and (TPR=1, FPR=1). Estimation of ROC becomes particularly straightforward for piece-wise linear ROCs, [31]. Area under the ROC curve (AUC) obtains here a clear statistical interpretation, in terms of this paper it is an estimate of probability $Pr[PS_i > PS_j]$, where i-th individual is randomly chosen $s_i \in C_P$, but j-th individual is randomly chosen $s_j \in C_N$. So, the closer the area is to AUC = 1.0 the greater the likelihood is that there the non-random allocation ($A = 1$ versus $A = 0$) was applied, otherwise if AUC doesn't differ significantly from AUC = 0.5, then the allocation can be considered as random, as it doesn't differentiate between positive and negative individuals. In the literature one can find tests for hypothesis that AUC = 0.5, [32, 33], and procedures for confidence intervals for receiver operating characteristic curves, [34, 35]. It should be noted that the precise quantitative estimates of the classifying variable (in this case – a propensity score *PS*) aren't necessary to estimate a ROC, and then to provide an analysis of an area under the ROC curve (AUC). It is just sufficient to be able to order the individuals (or: the clusters of individuals) from given sample with respect to their propensity score coefficients, because the values of thresholds were not used explicit in any stage of ROC and AUC calculations.

3. Exclusion Criteria

Generally, in a research practice the exclusions from a data base under investigation were driven by questionable quality of the particular data records. The most frequent reasons to exclude a certain data record from consideration are missing or outstanding values of variables, [29]. Nevertheless, it should be noted, that numerous exclusions are made usually on the beginning of an medical investigation, and during an investigation, in the random clinical trials too. The primary reason is that persons, who are invited to participate in research, simply refuse to participate, [36]. Then, in random clinical trials, some candidates (either patients or controls) are often excluded from randomisation by researcher, basing on economical and ethical considerations, [12, 37]. In the frame of the propensity score methodology the all above kinds of exclusions appear too, but some extra exclusions are aimed here to transform a crude sample from observational study into a sample comparable with samples from random trials. Without these extra exclusions, the propensity score methodology should be considered as a special case of the stratified analysis methodology, aimed to replace to great number of potential confounders with a single variable only. At such intentional process two main stages appear, at the first stage a given crude sample from an observational investigation under study should be cautiously inspected with purpose of separating from this sample a proper initial sample

S for starting with the estimation of propensity score. At the first stage a researcher formulates the exclusion criteria properly to purpose of the analyses, on basis of the own experience and subjective judgments. Usually, they correspond to the criteria that are applied in research practice for creating the samples of individuals in random clinical trials. The second stage arises during the core propensity score analysis.

During the core analyses a sub-sample $U \subset S$ applied finally to adjust a treatment effect arises, in some extend automatically with limited role of a researcher. First, at frame of clustering approach, see formulas (3) – (11), all clusters with estimated propensity $PS = 1$ or $PS = 0$ must be excluded from further analyses. A researcher can establish somewhat more strong criteria, say $PS > 0.9$ or $PS < 0.1$. In frame of a matching approach each individual without proper couple is automatically excluded by the procedure. It seems that two kinds of exclusions can be distinguished, first kind due to limits of the applied matching procedure, say one-to-one instead one-to-many or many-to-many. A second kind of exclusions resulted from factual outstanding, see [29], can be confirmed by researcher on basis of comparisons between exclusions made by several variants of matching procedures.

An auxiliary criterion, developed in this paper, was aimed at excluding the clusters that are characterized with inadmissible great probability of non-random allocation to treatment. Generally, this probability can be independent from the propensity score. It can be estimated with test-retest experiments at domain experts. Finally, it seems, that in a report on the propensity analyses the excluded individuals should be grouped into fractions with respect not only to the stage of analyses but also with regard to the clearly defined reasons and criteria of exclusions [38].

4. Motivating Example

The data for motivating example were acquired from [39]. Variable A indicated whether an individual had been allocated to severe ($A = 1$) cases versus moderate ($A = 0$) cases. Outcome Y was a length of hospital stay. Two confounding variables were considered: patient age (expressed with sample tercile 1st, 2nd or 3rd) and cause of disease [40], i.e.: 1st: vesicle (61 persons), 2nd: alcohol (41 persons), and 3rd: inappropriate diet (40 persons). The method for defining clusters was beyond scope of this study, because it depended mostly on specific knowledge and experience of domain medical experts [40]. The propensity score PS was estimated at each cluster separately as proportion of patients clinically classified as a severe cases (and allocated to special treatment) among the all, severe and moderate, cases. Table 1 shows descriptive statistics of the whole sample of $N = 142$ patients, and separately for each of four clusters. With respect to skew and kurtosis of outcomes $Y_0 = Y \mid A = 0$ and $Y_1 = Y \mid A = 1$ at the whole sample and at each of four cluster there wasn't any severe obstacle to use parametric methods, nevertheless, some non-parametric alternative should be preferred here, [9]. Effect size D for each cluster was separately estimated

using the formula (6). Generally, there are three alternative ways to estimate the summary mean effect *D* for all clusters. First, an optimistic researcher can assume that the data from an observational study under consideration don't differ from data obtained under simple random or stratified random sampling with random assigning to control and treated groups. A pessimistic researcher can consider each cluster as a quite different population what leads to meta-analysis approach. The propensity score methods can be considered as a medium approach between the above extremes. Table 2 (Estimates of a mean effect *D*) compared results of six different techniques applied here. Beyond two above estimates made by an optimistic researcher and by a pessimistic researcher, the next four estimates were obtained using the propensity score adjustment.

Table. 1. Descriptive statistics for clusters in illustrative sample

Cluster	1	2	3	4	Total
PS	60.0%	47.8%	34.6%	29.1%	39.4%
N	15	46	26	55	142
P	10.6%	32.4%	18.3%	38.7%	100.0%
X_1: age (tercile)	2nd or 3rd	3rd	1st	1st or 2nd	all
X_2: cause	2nd	1st or 3rd	2nd	1st or 3rd	all
mean_Y_0 ± SD	9.0 ± 2.8	15.3 ± 9.6	13.5 ± 4.2	13.6 ± 7.3	13.7 ± 7.4
skew_Y_0	1.5	1.3	0.4	0.8	1.3
kurtosis_Y_0	1.9	0.7	-0.1	-0.3	1.5
mean_Y_1 ± SD	56.7 ± 41.1	65.3 ± 43.5	61.7 ± 53.8	69.4 ± 45.9	64.5 ± 44.5
skew_Y_1	2.0	1.1	2.8	1.0	1.4
kurtosis_Y_1	4.9	0.8	8.2	0.4	1.5
effect D± SD	47.7 ± 18.1	50.1 ± 11.8	48.1 ± 21.9	55.8 ± 15.0	50.8 ± 8.0

The first three of estimates at Table 2 were obtained without any additional exclusions from the sample [39], using the formula (7), then the formula (11), and then the bootstrapping to the same four non-overlapping clusters. The last estimate was obtained by matching individuals in one-to-many mode accordingly to their propensity score (as estimated by logistic regression) with restriction on the propensity difference $|PS_i - PS_j| < 0.05$. This restriction led to 24 exclusions (27.9% of 86) from moderate ($A = 0$) cases, and 6 exclusions (10.7% of 56) from severe ($A = 1$) cases, without significant changes in mean outcomes for both groups of participants.

The choice of the technique didn't affect significantly the value of estimate of average effect *D*, $50.8 < D < 51.7$, with a single but distinct exception for estimate $D = 39.8$ obtained with matching approach, but it influenced significantly the estimates of standard deviation of the effect *D*, see Table 2. The propensity score approach led to mediate estimates of standard deviation, $9.4 < SD < 16.0$, between estimates obtained with optimistic ($SD = 8.03$) and pessimistic ($SD = 72.5$) approaches. The choice of the propensity score technique also affected accuracy of the average treatment effect

estimation. Among the procedures addressed to non-overlapping clusters the results obtained with bootstrap were recognised as more accurate and truthful than the estimates based on normality assumption. The choice between the propensity score procedures addressed to non-overlapping clusters versus the matching procedure cannot be decided at this stage of analysis without getting supplementary data from medical experts because of distinct difference in estimates of a mean effect *D*, see Table 2.

Table 2. Estimates of a mean effect D

Method	D	SD_D	95%CI for D
optimistic: Student t for whole sample	50.8	8.03	from 49.5 to 52.1
pessimistic: metaanalysis for clusters	51.0	72.5	from 39.1 to 62.9
propensity score with formula (7)	51.7	16.0	from 49.1 to 54.3
propensity score with formula (11)	50.9	15.5	from 48.4 to 53.4
propensity score by bootstraping	51.7	12.5	from 49.7 to 53.8
propensity score by matching	39.8	9.4	from 37.2 to 42.3

5. Getting Supplementary Data

In this section several designs for getting supplementary data from domain experts were briefly described. All of these designs can be considered as variants of the test-retest trial, that is the scores of the same individuals have been achieved twice, at two different times or at least after a sufficient time delay. In a frame of the two first designs, [18] and [21], a researcher ought to recruit domain experts to test-retest trial, with attempt to imitate decision situations at a clinic level before starting of a therapy. The last two designs, [25], and that one proposed first at [39] and now in this paper, try to utilize real clinical records from usual post-therapy analyses to unusual application of the ROC methodology.

In the procedures of repeated arrangements, [18] and [21], an expert divides the same set of individuals twice into predetermined number of ordered classes, depending only on quality of information available, see [22–24], and, last but not least, on their filling of a seriousness of situation. Ideally, the decision situations and the duties of domain experts in an experiment should be perceived by them as quite similar to their daily practice. The test-retest trial described at [18] was aimed to examine in what extend some defined additional information can modify previous allocation decisions. So, the two estimates of the propensity score, $PS|X$ versus $PS|(X, X')$, at the same sample of individuals were compared, where: X is information available to the domain expert at the test settings, and X' is an additional information, available at the retest settings. The test-retest trial described at [21] was aimed directly at estimation of underlying probability of undertaking of non-random allocation decisions.

At [2, 6] it was proved how the clinicians use to keep in practice the guidelines for artificial nutrition at patients with gastric cancer. In terms of ROC methodology a clinical guideline can be interpreted as diagnostic test under evaluation, but a real clinical decision on a course of nutrition as a manifestation of a true state of a patient. So, at [25], the patients treated with artificial nutrition in full agreement with the acknowledged clinical guidelines were interpreted as true positive cases, but the patients treated in spite of these guidelines as false positive cases. Finally, on basis of data on 915 patients with gastric cancer the area under the receiver operating characteristics (AUC) were estimated equal to AUC = 0.68 for women and AUC = 0.67 for men, both significantly different from AUC = 0.50. In terms of the current study, the treatment decisions made with full agreement with the acknowledged clinical guidelines can be interpreted as the deterministic decisions rather, but the remaining ones as undefined, either random or deterministic decisions. Thus, the results of ROC analysis support here need in extensive exclusions in the propensity score analyses, if used.

In the current study the somewhat another approach to preparing data for the ROC methodology was proposed. The results of initial clinical classification (and allocation to treatment) was interpreted as manifestation of a some (maybe undefined) diagnostic test under evaluation, with the estimated propensity score *PS* as diagnostic variable, but it was assumed that the true state, Positive versus Negative, of a patient at the moment of initial classification can be disclosed only after ending of therapy, on the base of the available clinical records. Thus, if AUC don't differ significantly from AUC = 0.5 then one can conclude that the propensity score (estimated on given patient's features) don't give a reliable basis to predict factual need in a treatment under consideration. With respect to the motivating example, as the Positives were considered the all patients which were recognised as severe cases after ending of therapy, but as true Positives the patients which were recognised as severe cases on the beginning and on the end of therapy too, see Table 3. ROC analysis for exemplary data.

The further routine procedure for estimating of the piece-wise linear ROC was disclosed at Table 3 ROC analysis for exemplary data. The clusters of study participants were ordered with respect to the estimated propensity score *PS*, see Table 1. The numbers of positive (NP) and negative (NN) participants were obtained from the post-therapy analyses made with aim of disclosing the sources of the doubtful initial allocations (if any).The true (false) positive ratios TPR (FPR) for consecutive thresholds at Table 3 were calculated in usual way, [31]. The area AUC for the piece-wise linear ROC was estimated simply with the formula (12), and the squared standard error SE for AUC was estimated with formula (13) from [33]. Knowing SE, one can easy determine significance of the hypothesis: AUC = 0.5.

$$\mathrm{AUC} = 0.5 * \Sigma(\ \mathrm{TPR}(k) + \mathrm{FPR}(k-1)) * (\ \mathrm{TPR}(k) - \mathrm{FPR}(k-1)). \qquad (12)$$

$$\mathrm{SE}^2 = (\mathrm{AUC}(1-\mathrm{AUC}) + (\mathrm{NN}-1)*q_0 + (\mathrm{NP}-1)*q_1) \,/\, \mathrm{NN}*\mathrm{NP} \qquad (13)$$

where: AUC is the estimated area under ROC; $k = 1, 2, 3, 4$ denotes a rank of cluster, see Table 1 or Table 3; for fictive rank = 0 it was assumed $TPR(k=0) = FPR(k=0) = 0$; NP and NN are total numbers of the positive (negative) participants respectively, see Table 3; q_0 and q_1 are estimated by: $q_0 = AUC*AUC*(1 - AUC)/(1 + AUC)$; $q_1 = AUC*(1 - AUC)^2/(2 - AUC)$.

Table 3. ROC analysis for exemplary data

Cluster	1	2	3	4	Total
NP | cluster	9	22	9	16	56
NP | threshold	9	31	40	56	-
NN | cluster	6	24	17	39	86
NN | threshold	6	30	47	86	-
TPR	0.161	0.554	0.714	1	-
FPR	0.07	0.349	0.547	1	-
AUC	0.006	0.10	0.125	0.389	0.619

For the exemplary data in Table 1 and Table 3, it was stated that AUC = 0.619 differed significantly ($p = 0.01$) from AUC = 0.5. Thus, the data under consideration differed significantly from the results of randomised trial what give some reason to prefer the propensity score method to the naïve estimations based on Student-t statistics for the whole sample, see Table 2. From other possible point of view, the value of AUC = 0.619 can be considered as only a little above AUC = 0.5, but not close to AUC = 1, so a great share of exclusions, associated with the matching approach to the estimate propensity score, can be considered as superfluous. It supported a preference to the clustering approach to propensity score estimation, that didn't exclude any cluster from consideration, over the matching approach, that led here to 24 exclusions (27.9% of 86) from moderate cases, and 6 exclusions (10.7% of 56) from severe cases.

Procedure of repeated arrangements bases on notion of indifference classes defined on a choice set [21]. The procedure consists of two sessions. During each session an expert deals with the same set of $N = K*N_k$ objects needed to be distributed into predetermined number K ordered classes, from rank $k = 1$ to rank $= K$, each class of the same number of $N_k = N/K$ objects, $k = 1, 2, ..., K$. During each session an expert can try out some arrangements with aim to select the best one in his/her filling. Between the session there should be settled a proper time delay, necessary to relax and to forget a preceding arrangement. All comparing objects should be properly standardized with aim to make difficult naming and remembering them by expert. The proposed procedure [21] seems to be more reliable from a typical procedure of ordering objects accordingly to known presumed aspect of preference, because of focus on similarity of objects, [22].

At [21], an expert's opinion on each evaluated object took form of width of rank interval (14) with edges equal to lower and higher rank of the classes associated to this object.

$$d_i = |k_{i1} - k_{i2}|, \quad i = 1, 2, ..., N. \tag{14}$$

where: k_{i1}, (k_{i2}) – ranks of I-th object at 1st (2nd) session.

Then, at [21] it was assumed, that an expert can apply either non-random decision mechanism that leads certainly to sure ranks at (14), or a random drawing, with the same chance for each class, equal to $P = 1 / K$. The next assumption, that probability Pc of applying the non-random decision mechanism was stable and the same for all patients at the considered population, creates opportunity to test the null hypothesis that probability $Pc = 0$ versus $Pc > 0$. As a test statistics was proposed difference Δ:

$$\Delta(N, T_{acc}, T_{n-acc}) = N_{acc} - N_{n-acc}. \tag{15}$$

where:
T_{acc}, (T_{n-acc}) – threshold defining the acceptable (non-acceptable) width of rank interval (14);
N_{acc} (N_{n-acc}) – number of acceptable (non-acceptable) width of rank interval (14) in considered N rank intervals.

The thresholds defining the acceptable and non-acceptable width of rank interval (14) should be chosen in each concrete situation basing on practical considerations. In [21] for $K = 9$ classes, the threshold for acceptable difference between classes associated to a same patient was assumed $T_{acc} = 2$, and for the non-acceptable difference $T_{n-acc} = 4$. The critical value of a test statistics (15) can be calculated analytically, for assumed N, T_{acc}, T_{n-acc}, but the simpler way is the Monte Carlo modelling. In [21] the critical value of this statistics (15) for $K = 9$ classes' and $N = 81$ patients, was estimated equal to 22 for usual significance level 0,05. Then, the power of the proposed test was examined for five assumed values of $Pc = 0{,}8$; 0,75; 0,7; 0,65; 0,5. It was stated, that the null hypothesis $Pc = 0$ will be rejected with probabilities, at the same succession: 0,99; 0,96; 0,86; 0,68; 0,16.

The maximum likelihood estimate of probability Pc of non-random allocation can be also obtained, because for any assumed width of rank interval d and probability Pc the conditional probability $P(d|Pc)$ is given with formula (16):

$$P(d|Pc) = (1 - Pc)*(P(d|Pc=0)) + \delta*Pc*(P(d|Pc=0))/(\Sigma\ (\delta*(P(d|Pc=0))). \tag{16}$$

where: coefficient $\delta = 0$ for $d \geq T_{n-acc}$ and $\delta = 1$ otherwise. The chance of observing a given distribution of N values of rank intervals (14) for assumed Pc is equal to the sum of N probabilities $P(d|Pc)$ calculated with (16). Then the maximum likelihood estimate of Pc is that value of Pc, that makes this sum as large as possible. At [21] the maximum likelihood estimate of non-random allocation, equal to $Pc = 0.81$, was reported on basis of the results of test-retest experiment with sample of 81 patients with osteoporosis. This value supports need in extensive exclusions in the propensity score analyses, if used.

6. Discussion

The question, how to draw useful conclusions from observational studies, has a great practical importance. Therefore, from many years up to now, the numerous methodologies were developed, [1–7], [11, 13, 14, 18, 36]. The present study was devoted exclusively to propensity score methodology. Propensity score methods have been considered as means to reduce the effect of selection bias at observational studies, [1, 4, 5, 11, 13, 14, 18]. In the frame of the propensity score methodology two main stages appear. At the first stage the features of participants of an observational study are considered as independent variables, with aim to estimate the relationship between the features and propensity score for each participant, [5, 14]. At the second stage the estimated values of propensity score are used to evaluate unbiased effect size. The great variety of methods applied in practice to estimate propensity can be divided into two groups. The first group of methods corresponds to logistic regression approach [26], but the second group of clustering methods [4, 5], bases on notion of similarity (or contrary: dissimilarity, distance) between individuals with respect to their individual features. Furthermore, the three seemingly different ways of estimation the propensity score in clustering approach (that is: non-overlapping clusters method, nearest- neighbourhood method, and kernel method) were in this paper generalized into a single scheme, following to [27], see formulas (3), (7), (11). The all computation of the propensity score methodology can be executed, without any resorting to complex programming, by any researcher, even with some limited statistical practice. Several ready to use calculators can be located via internet search. Moreover, the most of higher-level statistical packages such as Stata, SAS, or SPSS, included the complete procedures based on so named matching approaches to propensity analyses. The all of higher-level statistical packages included the recognized procedures for defining clusters. The further steps of clustering approach to propensity analyses can be easy implemented within frame of any spreadsheet, basing on easy available literature. With aim to make the last task easier, the three seemingly different ways of estimation the propensity score in clustering approach (that is: non-overlapping clusters method, nearest- neighbourhood method, and kernel method) were in this paper generalized into a single scheme. Moreover, a numerous easy to use free calculators can be located via internet search. Beyond a huge number of free calculators for standard two-sample comparisons and analysis of variance (with their non-parametric counterparts), one can find at WEB some useful tools: for logistic regression [26], for meta-analysis calculations, [41] and [42], and for bootstrapping, see [43, 44], and [45].

The superiority of the resampling (bootstrap) methodology over usual weighted averaging in estimation of the adjusted treatment effects were advocated in literature, [10, 13]. In this study the advantages of bootstrapping was proved again on exemplary data with criterion of the narrower confidence interval for adjusted effects, see Table 2: Estimates of a mean effect *D*. The bootstrap procedure led to width of

confidence interval equal to 53.8 – 49.7 = 4.1; versus 54.3 – 49.1 = 5.2 for classical procedure [10]. 53.4 – 48.4 = 5.0 for modified procedure [4], and 42.3 – 37.2 = 5.1 for matching procedure. Nevertheless, the narrowest was the naïve Student's confidence interval with its width equal to 52.1 – 49.5 = 2.6. Thus, the criterion of the narrower confidence interval for adjusted effects cannot be treated as the decisive criterion for the questions: to prefer use propensity score methodology over standard parametric methodology or not to prefer ? then: which approach to propensity score methodology should be preferred, if any?

Then, in this study it was demonstrated on the exemplary data that the different procedures of propensity score analyses, starting from the same crude data set, can lead to quite different results see Table 2: Estimates of a mean effect *D*. The differences in results can manifest in estimated effect of a therapy (i.e. estimated difference in an outcome between treated and non-treated groups), and/or in the standard errors of estimated effect, and/or in number of individuals excluded from consideration during analyses. Nevertheless, the recommendations how to choose a proper way of analyses within frame of the propensity score methodology are rather obscure and rare in the literature. With aim to overcome this shortage, the total amount of exclusions from crude sample was proposed as assisting criterion to choosing a proper way of the propensity score analyses. Consequently, the focus was made on procedures aimed to get additional information, how great share of exclusions should be preferred. The ideas of three known procedures, [18, 21, 25], were briefly discussed only. The forth procedure, based on unusual use of the receiver operating characteristics (ROC) and area under ROC (AUC) methodology, was explained in detail, because in our best knowledge, it was first proposed at [39] and then in this paper. A role of classifying variable in this procedure played the estimated propensity score. The main idea of this procedure consists on unusual definition of the Positive and Negative classes of individuals. Usually, these classes were defined with respect to baseline recognition (diagnosis) of predicted Positive class versus predicted Negative class, [31–33]. Then the predicted Positive class was divided into two sub-classes, true Positive class versus false Positive class (that is true Negatives among predicted Positives). Contrary to this, in the proposed procedure these classes were defined with respect to factual allocation to a treatment under consideration. Positive class includes all treated individuals, but Negative class all non-treated ones. Then the Positive class was divided into two sub-classes, true Positive class (individuals of certain medical reasons to treatment) versus false Positive class (individuals without medical reasons to treatment). The all further analyses run as usual, [31, 32], maybe with use of resampling (bootstrap) methodology, [34, 35], maybe without assuming infallibility of the clinical post-treatment evaluations, [19, 20].

The procedures, [21, 25], and [39], proposed in this paper as an auxiliary tool for propensity score analyses, generate some general measure of a randomness at an

exemplary sample, like AUC (area under ROC), but they don't generate any direct recommendation on choosing the separate individuals to exclusion from this sample. The estimated randomness can be interpreted, but beyond scope of this paper, in terms of noise to signal, [46]. Let us, contrary to [46], define a noise as random allocation to treatment, and a signal as deterministic allocation. Then one can consider an enough great unintended noise as some surrogate to the intended randomisation at the frame of the randomised trials. From this point of view, the exclusions made in propensity score methodology, can be interpreted as effort aimed to enlarge the noise to signal ratio. It should be noted, that the right allocations to treatment as well as wrong allocation can follows from a variety of causes, [12, 13, 15, 17, 25, 46], so the direct use of the differences in twice repeated classifications to excluding individuals from propensity score analyses should be considered with caution, even in frame of simple statistical models, like [21].

Two important limitations of this paper should be discussed. First, an initial situation for start with analyses in this paper was assumed at moment when use of propensity score methodology has been accepted definitely, so the question is how to carry out it, but not how to validate a preference to propensity score methodology over other competing methodologies. Then, an initial situation was assumed at moment when the observational studies under investigation have been finished previously, so all opportunities to include some extra randomisation into observational study were omitted in this paper. With regard to the first above weak point, it should be noted that the difficulties with drawing conclusions from observational studies have a plentiful literature, see e.g. [3, 6, 7, 36, 37, 38]. Moreover, the usefulness of propensity score methodology and of the competing methodologies were frequently compared, and often the practical recommendations in the matter were proposed, see e.g. [4, 5, 8, 11, 13, 46]. With regard to the second above drawback, it should be noted that in practice the same action, considered as superfluous effort from local or short-time perspective, can be recognized as valuable operation from global or long-time perspective. The actions beyond optimal (from local perspective) necessity frequently arose in frame of grouped psycho-educational interventions, [47]. Let us consider a typical case. Frequently, from practical and economical reasons, the psycho-educational interventions are performed to whole pupils classes or students group. The groups were evolved from some set of candidates basing of predicted share of menaced individuals, see e.g. [48], but the true share can be estimated on the ending of an intervention. With aim to validate a predicting procedure, beyond the groups of the greater predicted share of menaced individuals, some remaining groups can be incorporated to intervention too. It is easy to notice, that a share of menaced individuals can be interpreted here as propensity score, so in this way some clusters with low propensity score arose at the crude data from observational study. Moreover, in this simple way, an observational study acquire some features of the cluster randomised trials, [49].

7. Conclusions

The different procedures of propensity score analyses, starting from the same crude data set, can lead to quite different results. Nevertheless, the recommendations how to argue a preference to propensity score method over some other standard approach, and then how to choose a proper way of analyses within frame of the propensity score methodology are rather obscure and rare in the literature. In our opinion, each report from propensity score analyses should confront results from several propensity score procedures and from some other standard procedures on the same data under consideration.

The crucial aspect of the propensity score methodology is that during all analysis the properly chosen data records can be progressively excluded from further consideration. At such intentional process two main stages were defined, a first stage aimed separating a proper initial sample from a crude sample from observational study under consideration, and a second stage executed in some extend automatically during the core propensity score analysis. At the first stage the exclusion decisions should correspond to the criteria that are applied in research practice for creating the samples of individuals in random clinical trials. The aims and curse of exclusions at the second stage are different for matching and for clustering approaches to propensity analyses. Under matching approach, the proportion between treated and non-treated individuals should be exactly the same at each stratum of individuals with the same (at least: approximately the same) propensity score. The superfluous individuals must be rejected. Under clustering approach, each cluster should include at least some assumed minimal number of treated individuals and of non-treated ones. Cluster that don't satisfy this requirement must be rejected at whole. Nevertheless, at the both stages of exclusions, and under both approaches, a researcher can control the resulting number of individuals excluded from consideration e.g. by proper choice of criterions of similarity between individuals.

In our opinion, the resulting number of individuals excluded from consideration during the propensity score analysis should correspond to proportion of individuals non-randomly allocated to treated and to non-treated groups at a crude sample from an observational study under investigation. Moreover, under the clustering approach, each cluster with a weighty share of non-randomly allocated individuals should be rejected from consideration, independently from other criterions of exclusions. The proportion of individuals non-randomly allocated to treated and to non-treated groups at a sample from an observational study can be estimated on the base of some auxiliary information, obtained from clinicians.

In this paper the procedure of repeated arrangements and the ROC (receiver operating curve) technique were considered as two practical ways aiming to obtain reliable information from experienced clinicians. In a single trial with the procedure of repeated arrangements to ordered classes the results of repeated classifications made by the same expert on the same sample of patients can be obtained. The series

of these trials, with several clusters of patients, and maybe with several experts, gives opportunity to estimate probability of random allocation to the treatment at particular clusters of patients. Contrary to this, the ROC (receiver operating curve) technique utilizes information from usual in clinical practice post-intervention analyses, maybe without additional explanations from medical experts. It should be noted, that both above procedures are aimed to estimate randomness of the treatment decisions with regard to some samples of patients, but not with respect to individual patients.

References

1. Rosenbaum P.R., Rubin D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika 1983, 70, 1, 41–55.
2. Tannen R.L., Weiner M.G., Xie D.: Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. BMJ 2009;338:b81, [Accessed 2009, Jul 14].
3. Furlan A.D., Tomlinson G., Jadad A.A., Bombardier C.: Methodological quality and homogeneity influenced agreement between randomised trials and nonrandomized studies of the same intervention for back pain. J. Clin. Epidemiol. 2008, 61(3), 209–231.
4. Senn S., Graf E., Caputo A.: Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. Statistics in Medicine 2007, 26, 30, 5529–5544.
5. Kurth T., Walker A.M., Glynn R.J., Chan K.A., Gaziano J.M., Berger K., Robins J.M.: Results of Multivariate Logistic Regression, Propensity matching, Propensity Adjustment, and Propensity-based Weighting under Conditions of Nonuniform Effect. Am. J. Epidemiol. 2006, 163, 262–270.
6. Shrier I., Bolvin J.F., Steele R.J., Platt R.W., Furian A., Kakuma R., Brophy J., Rossignol M.: Should Meta-Analyses of Interventions Include Observational Studies in Addition to Randomized Controlled Trials? A Critical Examination of Underlying Principles. Am. J. of Epidemiol. 2007, 166, 1203–1209.
7. West S.G., Duan N., Pequegnat W., Gaist P., Jarlais C., Holtgrave D., Szapocznik J., Fishbein M., Rapkin B., Clatts M., Mullen P.D.: Alternatives to Randomized Controlled Trials. Am. J. Public Helth 2008, 98, 8, 1359–1366.
8. Parker R.I.: Increased Reliability for Single-Case Research Results: Is the Bootstrap the Answer? Behavior Therapy 2006, 17, 326–338.
9. Kelley K.: The Effects of Nonnormal Distributions on Confidence Intervals Around the Standarized Mean Difference: Bootstrapping as an Alternative to Parametric Confidence Intervals. Educational and Psychological Measurement 2005, 65, 51–69.
10. Tu W., Zhou X.H: A Bootstrap Confidence Interval Procedure for the Treatment Effect Using Propensity Score Subclassification. Health Services & Outcomes Research Methodology 2002, 3,135–147.
11. Stürmer T., Joshi M., Glynn R.J., Avorn J., Rothmann K.J., 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 multivariate methods. J. Clinical Epidemiol. 2006, 59, 5, 437–461.
12. Kunz R., Vist G., Oxman A.D.: Randomisation to protect against selection bias in healthcare trials. BMC Med. Res Methodol. 2005, 2, 5(1),10.
13. Deeks J.J., Dinnes J., D'Amico R.A., Sowden A.J., Sakarovitch C., Song F., Petticrew M., Altman D.G.: Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:1–173. [Medline], full text available online: URL_http://www.hta.ac.uk/1117 [Accessed 2009, Jul 14].

14. Brookhart M.A., Schneeweiss S., Rothman K.J., Glynn R.J., Avorn J., Stürmer T.: Variable Selection for Propensity Score Models. Am. J. Epidemiol. 2006, 163, 1149–1156.
15. Redelmeier D.A.: The Cognitive Psychology of Missed Diagnoses. Annals of Internal Medicine 2005, 142(2), 115–120.
16. Roberts R., Goodwin P.: Weight approximations in multi-attribute decision models. J of Multi-Criteria Decision Analysis 2002, 11(6), 292–303.
17. Reid M.C., Lane D.A., Feinstein A.R.: Academic Calculations versus Clinical Judgments: Practicing Physicians' Use of Quantitative Measures of Test Accuracy. Am. J. Medicine 1996, 104(4), 374–380.
18. Stürmer T., Schneeweiss S., Rothmann K.J., Avorn J, Glynn R.J.: Performance of Propensity Score Calibration – A Simulation Study. Am. J. Epidemiol. 2007, 165, 1110–1118.
19. Zou G.: Quantifying responsiveness of quality of life measures without an external criterion. Quality of Life Research 2005, 14(6), 1545–1552.
20. Zhou X.H., Castellucio P., Zhou C.: Nonparametric Estimation of ROC Curves in the Absence of a Gold Standard. Biometrics 2005, 6, 600–609.
21. Górkiewicz M., Ciszek E., Szczygieł A.: Selecting experts and classifying features with procedure of repeated arrangements to classes of similarity, In: J. Wywiał (Ed.), Metoda Reprezentacyjna w Badaniach Ekonomiczno-Społecznych, University of Economics in Katowice 2004, 195–215 [in Polish].
22. Goldstone R.L., Medin D.L., Halberstadt J.: Similarity in Context. 2003, On-line: http:// cognitrn.psych.indiana.edu/rgoldsto/context/context.html [Accessed 2009, Jul 14].
23. Szczygieł A., Ciszek E., Górkiewicz M.: Visual inspection of the osteoporosis functional disability using rescaled standard anatomical pictures. Annales Academiae Medicae Bialostocensis 2005, 50 (Suppl.2), 75–77 URL_ http://www.advms.pl/node/84 [Accessed 2009, Jul 14, by menu options: Supplements / Supplement 2, Vol. 50].
24. Cook A.: Using video to include the experiences of people with dementia in research. Research Policy and Planning 2003 21, 23–32.
25. Walewska E., Ścisło L., Górkiewicz M., Czupryna A., Kłęk S., Szczepanik A.M., Kulig J.: Usefulness of guidelines for nutrition screening at patients with gastric cancer. Ann. Univ. Mariae Curie-Sklodowska, Sect. D, Med. 2005, suppl. 16 (6), 152–156, [in Polish].
26. Pezzullo J.C., Sullivan K.M.: Logistic Regression. available online: URL_ http://statpages.org/logistic.html [Accessed 2009, Jul 14].
27. Filippone M., Camastra F., Masulli F., Rovetta S.: A survey of kernel and spectral methods for clustering. Pattern Recognition 2008, 41(1), 176–190.
28. Mak T.K.: Estimating variances for all sample sizes by the bootstrap. Computational Statistics and Data Analysis 2004, 46, 459–467.
29. Bartkowiak A.: Robust Mahalanobis Distances Obtained Using the 'Multout' and 'Fast-med' Methods. Biocybernetics and Biomedical Engineering 2005, 25(1), 7–21.
30. Hutson A.D.: A semiparametric bootstrap approach to correlated data analysis problems. Computer Methods and Programs in Biomedicine 2004, 73, 129–134.
31. Górkiewicz M.: Multivariable ROCs: for separating planes, voting rules and decision trees. In: L. Bobrowski, J. Doroszewski, C. Kulikowski, N. Victor (Eds.), Statistics and Clinical Practice, Lecture Notes of ICB Seminars, vol. 70: Warsaw 2005, 95–102.
32. Rossa A.: The goodnes-of-fit tests for ROC curves, In: L. Bobrowski, J. Doroszewski, C. Kulikowski, N. Victor (Eds.), Statistics and Clinical Practice, Lecture Notes of ICB Seminars, vol. 70: Warsaw 2005 , 42–47.
33. Hanley J.A., McNeil B.J: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982, 143, 29–36.
34. Hall P., Hyndman R.J., Fan Y.: Nonparametric confidence intervals for receiver operating characteristic curves. Biometrika 2004, 91 (3), 743–750.

35. Mossman D.: Resampling techniques in the analysis of non-normal ROC data. Medical Decision Making 1995, 15, 358–366.
36. Galea S., Tracy M.: Participation Rates in Epidemiologic Studies. Ann. Epidemiol. 2007, 17, 643–653.
37. Owen A.: The ethics of two- and one-sided hypothesis tests for clinical trials. Clinical Ethics 2007, 2(2), 100–102.
38. von Elm E., Altman D.G., Egger M., Pocock S.J., Gotzsche P.C., Vandenbroucke J.P.: The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. PLoS Med. 4 (10), 2007, e296 URL_http://medicine.plosjournals.org/perlserv/ URL_http://www.strobe-statement.org/ [Accessed 2009, Jul 14].
39. Górkiewicz M.: Observational studies: using propensity score with receiver operating characteristics and bootstrap., in: Balcerar-Nicolau H, Bobrowski L, Doroszewski J, Kulikowski C. (eds). Lecture Notes of the VII-th ICB Seminar: Statistics and Clinical Practice, Warszawa 2008: 68–74.
40. Chmiel I., Czupryna A., Górkiewicz M., Brzostek T.: The causes of acute pancreatitis and the range of psychoeducational intervention for convalescens. Studia Medyczne, 2008, 11, 51–56 [in Polish].
41. Basu A.: How to Conduct a Meta-Analysis, 2005, Available On-line: URL_http://www.pitt.edu/~-super1/lecture/lec1171/008.htm.
42. Chang A: Meta-analysis using Mean Difference, 2000. Available On-line: URL_http://department.obg.cuhk.edu.hk/researchsupport/MetaMeans.asp [Accessed 2009, Jul 14, by menu options: Stats toolbox / Statistical tests / Meta-Analysis].
43. Wood M.: The Role of Simulation Approaches in Statistics. J. of Statistics Education 2005, 13, 3 available on-line URL_www.amstat.org/publications/jse/v13n3/wood.html [Accessed 2009, Jul 14].
44. Aksenov S: Confidence Intervals by Bootstrap. Wolfram Research Inc. 2002, available on-line: URL_http://library.wolfram.com/infocenter/MathSource/4272/ [Accessed 2009, Jul 14].
45. Siniksaran R.: BootStrapPackage: A Package of Bootstrap Algorithms for Mean, Simple Linear Regression Models, and Correlation Coefficient. Wolfram Research Inc. 2001, URL_http://library.wolfram.com/infocenter/MathSource/815/ [Accessed 2009, Jul 14.
46. Glasziou P., Chalmers I., Rawlins M., McCulloch P.: When are randomised trials unnecessary? Picking signal from noise. BMJ, 2007 334(7589), 349–351.
47. Chan C.W.: Psychoeducational intervention, a critical review of systematic analyses. Clinical Effectiveness in Nursing, 2005, 9, 101–111.
48. Rosochacka W., Górkiewicz M.: Forgotten duties: universities should be anxious for students' learning styles. Annales Academiae Medicae Bialostocensis 2005, 50 (Suppl.2), 59–60 URL_ http://www.ad-vms.pl/node/84 [Accessed 2009, Jul 14, by menu options: Supplements / Supplement 2, Vol. 50].
49. Eldridge S., Ashby D., Bennett C., Wakelin M., Feder G.: Internal and external validity of cluster randomised trials: systematic review of recent trials. BMJ, 2008, 336(7649), 876–880.