ASPECTS OF THE USE OF PROPENSITY SCORE MATCHING METHODS IN NEUROSURGERY

ASPECTOS DO USO DOS MÉTODOS DE PAREAMENTO DE ESCORE DE PROPENSÃO EM NEUROCIRURGIA

ASPECTOS DEL USO DE LOS MÉTODOS DE EMPAREJAMIENTO DE PUNTAJE DE PROPENSIÓN EN NEUROCIRUGÍA

Aleksandr V. Krutko,[1] Shamil A. Akhmetyanov,[1] KirillYu. Orlov,[2] Victor S. Gladkikh,[3,4] Andrey V. Moskalev[3,4]

1. Novosibirsk Research Institute of Traumatology and Orthopaedics – NRITO, Tsivyan, st, Frunze 17, Novosibirsk, 630112, Russia.
2. Meshalkin Siberian Federal Biomedical Research Center, Novosibirsk, Russia.
3. Biostatistics and Clinical Trials Center, Novosibirsk, Novosibirsk, Russia.
4. Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia.

ABSTRACT

Objective: Observational studies and register data provide researchers with ample opportunities to obtain answers to questions that randomized controlled trials cannot answer for organizational or ethical reasons. One of the most common tools for solving this problem is the use of propensity score matching (PSM) methods. The purposes of our study were to compare various models and algorithms for selecting PSM parameters, using retrospective clinical data, and to compare the results obtained using the PSM method with those of prospective studies. Methods: The results of two studies (randomized prospective and retrospective) conducted at the Novosibirsk Research Institute of Traumatology and Orthopedics were used for comparative analysis. The trials aimed to study the effectiveness and safety of surgical treatment of degenerative dystrophic lesions in the lumbar spine. We compared the results using the recommended PSM parameters (caliper=0.2 and 0.6) the propensity score is the probability of assignment to one treatment conditional on a subject's measured baseline covariates. Propensity-score matching is increasingly being used to estimate the effects of exposures using observational data. In the most common implementation of propensity-score matching, pairs of treated and untreated subjects are formed whose propensity scores differ by at most a pre-specified amount (the caliper widthand the caliper values often used in real-life studies (0.05, 0.1, 0.25, 0.5, and 0.8) with the those obtained in a similar prospective study. Results: After eliminating systematic selection bias, the results of the retrospective and randomized prospective studies were qualitatively comparable. Conclusion: The results of this study provide recommendations for the use of PSM: when evaluating efficacy scores in neurosurgical studies (with a sample size < 150 patients), we recommend matching on the logit of the propensity score using calipers of width equal to 0.6 of the standard deviation of the logit of the propensity score. ***Level of evidence V; Type of study is expert opinion.***

Keywords: Neurosurgery; Spinal Fusion; Spinal Stenosis; Statistical Analysis; Bias.

RESUMO

Objetivos: Estudos observacionais e dados de registro fornecem aos pesquisadores amplas oportunidades de obter respostas às perguntas que os estudos clínicos randomizados não podem responder por razões institucionais ou éticas. Uma das ferramentas mais comuns para resolver esse problema é o uso dos métodos de Propensity Score Matching (PSM, pareamento de escore de propensão). O objetivo do nosso estudo foi comparar vários modelos e algoritmos para a seleção de parâmetros de PSM, usando os dados clínicos retrospectivos e comparar os resultados obtidos com esse método com os de estudos prospectivos. Métodos: Os resultados de dois estudos (randomizado prospectivo e retrospectivo), realizados no Instituto de Pesquisa de Traumatologia e Ortopedia de Novosibirsk, foram utilizados para análise comparativa. Os estudos visaram estudar a eficácia e a segurança do tratamento cirúrgico de lesões distróficas degenerativas na coluna lombar. Comparamos os resultados usando os parâmetros recomendados pelo PSM, isto é calibração (caliper) de 0,2 e 0,6 e os valores de calibração usados com frequência em estudos da vida real (0,05, 0,1, 0,25, 0,5 e 0,8) com os obtidos em um estudo prospectivo semelhante. Resultados: Depois de eliminar o viés sistemático de seleção, os resultados de estudos randomizados prospectivos e retrospectivos foram qualitativamente comparáveis. Conclusões: Os resultados deste estudo fornecem recomendações para o uso do PSM: ao avaliar os escores de eficácia em estudos neurocirúrgicos (com tamanho de amostra < 150 pacientes), recomendamos a correspondência do logit do escore de propensão com calibração de largura de 0,6 do desvio padrão do logit do escore de propensão. ***Nível de evidência V; Opinião do especialista.***

Descritores: *Neurocirurgia; Fusão Vertebral; Estenose Espinal; Análise Estatística; Viés.*

RESUMEN

Objetivos: Los estudios de observación y los datos de registro brindan a los investigadores amplias oportunidades para obtener respuestas a preguntas que los estudios clínicos aleatorizados no pueden responder por razones institucionales o éticas. Una de las herramientas más comunes para resolver este problema es el uso de los métodos de Propensity Score Matching (PSM, emparejamiento de puntaje de

Study conducted at the Novosibirsk Research Institute of Traumatology and Orthopedics – NRITO, Tsivyan, st, Frunze 17, Novosibirsk, 630112, Russia.
Correspondence: Aleksandr V. Krutko. Neurosurgery Department of NRITO. Frunze str., 17, Novosibirsk, Russia. 630091. ortho-ped@mail.ru

http://dx.doi.org/10.1590/S1808-185120201902224291

Received on 06/04/2019 accepted on 07/30/2019

propensión). El objetivo de nuestro estudio fue comparar varios modelos y algoritmos para la selección de parámetros de PSM, utilizando los datos clínicos retrospectivos y comparar los resultados obtenidos con ese método con los de estudios prospectivos. Métodos: *Los resultados de dos estudios (prospectivo aleatorizado y retrospectivo) realizados en el Instituto de Investigación de Traumatología y Ortopedia de Novosibirsk se utilizaron para el análisis comparativo. Los estudios tuvieron como objetivo estudiar la eficacia y seguridad del tratamiento quirúrgico de las lesiones distróficas degenerativas en la columna lumbar. Comparamos los resultados usando los parámetros recomendados por el PSM, esto es, calibración (caliper) de 0,2 y 0,6 y los valores de calibración usados con frecuencia en estudios de la vida real (0,05, 0,1, 0,25, 0,5 y 0,8) con los obtenidos en un estudio prospectivo semejante. Resultados: Después de eliminar el sesgo sistemático de selección, los resultados de estudios prospectivos aleatorizados y retrospectivos fueron cualitativamente comparables. Conclusiones: Los resultados de este estudio proporcionan recomendaciones para el uso del PSM: al evaluar los puntajes de eficacia en estudios neuroquirúrgicos (con tamaño de muestra <150 pacientes), recomendamos la correspondencia del logit del puntaje de propensión con calibración de ancho de 0.6 de la desviación estándar del logit de puntaje de propensión.* ***Nivel de evidencia V; Opinión del especialista.***

Descriptores: *Neurocirugia; Fusión Vertebral; Estenosis Espinal; Análisis Estadístico; Sesgo.*

INTRODUCTION

There has been keen interest in estimating the causal effects of treatment using the observational non-randomized data, as well as registry data processed retrospectively.[1,2] Observational studies are often employed in pharmaceutical and medical research when it is impossible to conduct randomized controlled trials, or when they do not meet ethical requirements. However, a researcher analyzing the data obtained in non-randomized studies generally faces the same problems intrinsic to retrospective studies: incomparability between the study groups with respect to individual clinical parameters.

The reason for this is that the baseline parameters of treated subjects in observational studies and the baseline parameters upon retrospective analysis tend to differ systematically from those of subjects receiving other treatment. The ability to minimize the confounding effect is very important to produce high-quality evidence for informed decision making. Analysis of the retrospective data without adjusting the systematic selection bias in comparison groups may lead to significant distortion or misinterpretation of the results.[4] Propensity score matching is one of the enormously popular techniques employed in medical publications.[5,6] Many researchers find propensity score matching extremely helpful because of its ability to directly compare baseline parameters between treated and untreated subjects in a propensity score matched sample.[7]

Nowadays, the number of publications employing PSM as a tool to analyze and interpret retrospective clinical data is steadily increasing;[8] the number of such publications estimated by Scopus database reached 3,400 in 2017 (Figure 1). In 2007–2017, the percentage of papers written by Russian authors and indexed in international databases was as low as 0.15% of a total of > 15,000 publications.[9,10]

Meanwhile, very few neurosurgical papers were published in 2010-2017. The vast amount of accumulated unique data enable the use of potential of modern statistical analysis methods to make a quantum leap forward and approach "the gold standard" when analyzing the retrospective data. For this very reason, with the vast body of retrospective data that has been accumulated, neurosurgical researchers have shown keen interest in using the PSM method. Even today, analysis of retrospective data is already a required step in planning new study designs, and the importance of this step in trial design and conduct is set to increase. Analysis of the data retrospectively collected by Russian neurosurgeons over the years will enable the quality of future studies to be significantly improved.

In order to plan and conduct a randomized prospective study, researchers typically need to perform a preliminary analysis of the retrospective data, in order to clearly identify the study objectives (primary and secondary endpoints). Most neurosurgical researchers prefer to use their own empirics to identify the criteria for inclusion of patients in the analysis. Complete analysis of baseline characteristics is rarely carried out to analytically solve the problem of identifying specific inclusion criteria.

In order to use the PSM method, we need to build a proper mathematical model to calculate the propensity score and select the parameters of matching algorithm based on clinical expertise and experience in using the PSM in a specific field. Synthetic examples

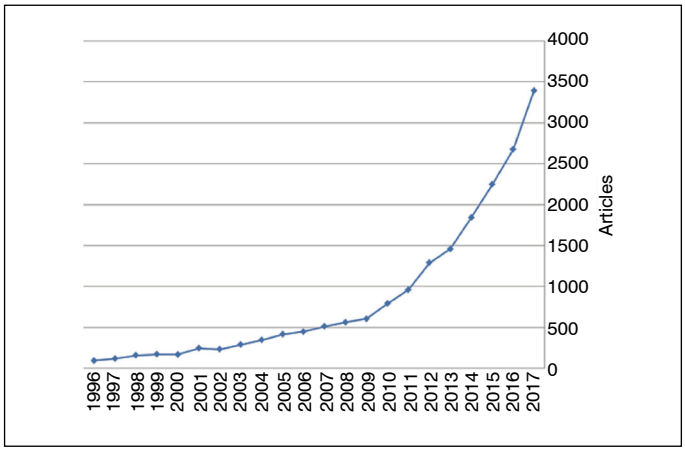

Figure 1. PSM articles per year.

related to bias elimination with pre-determined properties of comparison groups are often used to analyze the behavior of PSM.[11,12] We would like to mention that when using the PSM method, one should strive to achieve accurate results and group comparability that are as close as possible to those of randomized studies. For this reason, the PSM method is known as pseudo-randomization in the Russian-language literature. Therefore, our study aims to compare various models and algorithms for selecting PSM parameters using retrospective clinical data, and to compare the results obtained using the PSM method with those of prospective studies.

METHODS

We used the results of two studies conducted at Neurosurgery Department no. 2 of the Novosibirsk Research Institute of Traumatology and Orthopedics (Novosibirsk, Russia). The first study was a randomized prospective study involving 94 patients operated on for clinically relevant degenerative dystrophic lesions in lumbar spine. The second study was a retrospective analysis of the results of surgeries in 189 patients operated on for clinically relevant degenerative dystrophic lesions in lumbar spine, which was performed using the PSM method.

The PSM method is based on using the propensity score (PS), which is the conditional probability that a subject will be allocated to a treatment group. To build a propensity score model, a researcher identifies clinically relevant variables that need to be balanced as these may potentially have a confounding effect on study results. We selected these variables on the basis of the publication data and our own clinical experience.

Since the Propensity Score is a conditional probability, the PS values belong to the [0; 1] range. The multivariate logistic regression model is most frequently used to build the PS model. The calculated PS values can be used to select comparison groups so that the conditional probability between the treatment and control groups is balanced. Different variations of the PSM method (matching on the

propensity score, stratification/subclassifications, inverse probability of treatment weighting (IPTW), and covariate adjustment) are used for this purpose. None of the known methods is characterized by sufficient universality and applicable to all situations. However, we shall discuss the nearest neighbor caliper matching (CNN) procedure as the main working procedure, according to the available publications.[13]

We selected two criteria for evaluating the adequacy of the model selected to calculate the propensity score for the PSM method, and to select algorithm matching parameters. The first criterion is elimination of systematic selection bias at baseline. The second is qualitative coincidence of the results of PSM analysis of the retrospective data and the results of a randomized prospective study. We compared the results using the recommended matching parameters (caliper=0.2 and 0.6)[11,12] and the caliper values often used in real-life studies (0.05, 0.1, 0.25, 0.5, and 0.8) with the those obtained in a similar prospective study.[11]

The presence of referred and radicular pain syndromes resistant to conservative treatment, either accompanied by neurological deficit or not, was a criterion for selecting patients to receive surgical treatment.

The criteria for inclusion in a randomized study were as follows:
• mono- or polyradicular compression of the spinal cord roots and (or) pseudoclaudication, with a possible combination of pain syndromes.
• a single lumbar functional spinal unit was predominantly affected and caused the clinical symptoms;
• instability, grade I spondylolytic spondylolisthesis, putative large volume of resection of the posterior vertebral support structures disrupting spinal stability and requiring stabilization of a single lumbar functional spinal unit.

The exclusion criteria were as follows:
• polysegmental spinal canal stenosis;
• severe concomitant somatic pathology;
• diabetes mellitus, severe course;
• spondylolytic spondylolisthesis (grade II and higher)
• age $<$ 20 years or $>$ 75 years
• disorders and conditions affecting the development of degenerative changes in lumbar spine (congenital spinal canal stenosis, previous history of spine injuries or tumors, inflammatory lumbar spine disorders, disorders of large joints of the lower limbs, etc.).

The compulsory diagnostic preoperative examination included collecting past medical history, conducting general clinical, neurological and X-ray examination, MRI, and MSCT (in some cases, involving contrasting of the dural sac).

Planning of the level and type of surgery was based on the principle of clinical and morphological matching, according to which the surgery was aimed at eliminating the pathomorphological substrate with clinical manifestations. Some patients underwent minimally invasive surgery of the vertebrogenic pain syndrome. In these cases, decompression and stabilization were performed without skeletization of spine structures. Access to the spine was performed by blunt muscle dissection through skin incisions ~1.5 long, in order to insert pedicle screws in a minimally invasive manner. Incisions 3–4 cm long were made to perform the Wiltse parasagittal approach in order to carry out decompression at one functional spinal unit and to insert pedicle screws. Bilateral resection of hypertrophied overgrowth of cartilage, bone and ligaments was conducted through the Wiltse unilateral parasagittal approach, using a tubular retractor. Application of these methods for microsurgical decompression of neurovascular structures in the spinal canal is a good alternative to bilateral decompression through the interlaminar approach or for conducting decompressive laminectomy.

In another patient group, all the decompression and stabilization interventions were carried out via the conventional posteromedial approach; skeletization of the posterior sections of the vertebral column was performed. These patients underwent the same degree of stabilization and adequate conventional decompression of intracanal neurovascular structures (laminectomy, hemilaminectomy, interlaminectomy, partial and complete facetectomy).

The positions of the puncture needle, pins, cannulated and standard screws, templates and interbody implants in the vertebra were controlled using SXT-1000A (Toshiba Medical Systems Corporation) and Ziehm (Ziehm Imaging GmbH) electron-optical image intensifiers.

Intraoperative injury and the degree of invasiveness were evaluated using a number of parameters:
• time required to perform each stage of surgical intervention (performing the approach, transpedicular fixation, decompression and interbody stabilization);
• the surface area of the surgical wound (sterile polyethylene film was placed on the wound surface and the wound borders were contoured; the film was then put on linear graph paper to measure the wound surface area);
• blood loss volume at each stage of surgical intervention (performing the approach, transpedicular fixation, decompression and interbody stabilization);
• the intensity and dynamics of pain at surgical site during the early postoperative period (up to 14 days) using the VAS score;
• postoperative hospital stay (number of bed days).

Transpedicular fixation was performed using the Legacy, Expedium, Viper, Sextant, and Longitude constructions and instruments.

Porous Ni-Ti implants (Interfix, Capstone, and Concorde), Aesculap instruments, and Quadrant and Pipeline tubular retractors were used for interbody fusion.

The study was performed in accordance with good clinical practice, ensuring that design, implementation, and communication of data were reliable, that patients' rights are protected, and that subject integrity was maintained through the confidentiality of their data. The study was approved by the local ethics committee of the Novosibirsk State Institute of Traumatology and Orthopedics (protocol No. 36 dated October 16, 2008). Written informed consent was obtained from all patients and their parents, in accordance with the Declaration of Helsinki, including consent for their data to be analyzed and reported.

Calculations were performed using the R Statistical Package (http://www.r-project.org). The descriptive statistics are presented as absolute frequencies or median values with the IQR specified. Either the Mann-Whitney U-test, Pearson's χ^2 test, or Exact Fisher Test and non-parametric Kruskal–Wallis analysis of variance by ranks and median multiple comparisons were used, depending on the type of data being processed. The Kendall rank correlation coefficient was calculated to determine possible correlations.

All the reported p-values were based on two-tailed tests for significance; p-values $<$ 0.05 were regarded as statistically significant. Analysis was conducted using the STATISTICA 7.0 software (StatSoft, USA) and RStudio software version 0.99.484 (Free Software Foundation, Inc., USA) with R packages version 3.2.2 (The R Foundation for Statistical Computing, Austria).

RESULTS

In the retrospective study, the treatment group (group I) included 63 patients subjected to minimally invasive (including transcutaneous) surgery: 31 males (49.0%) and 32 females (51.0%). The comparison group (group II) consisted of 126 patients: 44 males (35.0%) and 82 females (65.0%). The preoperative VAS (spine) score was 7.0 (6.0; 7.3) in group I and 7.0 (6.0; 7.0) in group II. The preoperative VAS (leg) score was 7.0 (6.0; 7.0) in group I and 7.0 (6.0; 7.0) in group II. The preoperative Oswestry Disability Index score was 58 (50; 65) in group I and 58 (48.5; 64) in group II. No significant intergroup differences were revealed for all three parameters. The levels of the lesion in group I were distributed as follows: L2–L3, 9.5%; L3–L4, 19.1%; L4–L5, 68.3%; and L5–S1, 3.2%. In group II, the levels of the lesion were distributed as follows: L2–L3, 2.4%; L3–L4, 10.32%; L4–L5, 40.5%; and L5–S1, 45.2%.

In the randomized study, the treatment group (group I) included 55 patients subjected to minimally invasive (including transcutaneous) surgery: 18 males (33.0%) and 37 females (67.0%) aged 23–70 years. The comparison group (group II) consisted of 39

patients, including 15 males (38.0%) and 24 females (62.0%) aged 23–70 years. The preoperative VAS (spine) score was 7.0 (6.0; 7.7) in group I and 7.0 (6.0; 7.0) in group II. The preoperative VAS (leg) score was 7.0 (5.0; 8.0) in group I and 6.0 (4.0; 8.0) in group II. The preoperative Oswestry Disability Index score was 60 (48; 72) in group I and 58 (48.5; 68) in group II. No significant intergroup differences were revealed for all three parameters. The levels of the lesion in group I were distributed as follows: L2–L3, 4.7%; L3–L4, 14.1%; L4–L5, 60.9%; and L5–S1, 20.3%. In group II, the levels of the lesion were distributed as follows: L2–L3, 0.0%; L3–L4, 4.4%; L4–L5, 71.1%; and L5–S1, 23.5%.

Postoperative peridural fibrosis detected in both groups subjected to surgical interventions had no effect on the surgical procedure. The dural sac and spinal cord nerve roots were isolated from unaltered tissues. In the group of patients treated by open surgery via the transforaminal approach, radiculolysis was a simpler procedure, since the unaltered dura mater lay immediately underneath the articular processes.

Measuring the wound surface area using the procedure described above showed that the mean size of surgical wound, after using open transpedicular fixation, was more than ten times greater than the area after performing transpedicular fixation through the Wiltse parasagittal approach. The wound surface area was not taken into account when percutaneous transpedicular systems were inserted.

The mean time required to perform a minimally invasive surgery was shorter than that of an open surgery. However, this variation was statistically insignificant both in the prospective and retrospective studies.

Stepwise comparison of the open and minimally invasive surgical techniques showed smaller intraoperative blood loss at all intervention stages (performing the approach, placing a transpedicular system, and decompression + interbody fusion) in subjects undergoing minimally invasive surgery. A significant reduction in blood loss was observed at the stages of performing the approach and placing the transpedicular system. Shorter total surgery times were also observed in the group of patients treated by minimally invasive surgery, mainly due to the shorter time required to perform the approach, although placement of the transpedicular system took longer in this group.

Both in the prospective and retrospective studies, the VAS score for pain intensity at surgical site in the early postoperative period was decreased in both groups of patients subjected to surgical interventions. However, in the prospective study, the pain intensity score in the group treated by minimally invasive surgery was statistically significantly lower than in the groups subjected to open surgery.

A retrospective data analysis revealed significant heterogeneity among the study groups with respect to surgery level ($p<0.001$, χ^2 test) and concomitant pathologies ($p=0.023$, χ^2 test) (Table 1). Kendall rank correlation coefficients was used to reveal any possible confounding effect on the outcomes, enabling us to detect any correlation between heterogeneous variables and the outcomes of surgical interventions according to the VAS and ODI scores. Therefore, we needed to minimize the systematic selection bias, while still having a sufficiently large sample size to ensure statistical validity of the results. Selection of the model for using PSM is based on clinical experience, while there always is certain variability (freedom) in choosing a set of independent variables to calculate PS. We sought to identify the most efficient parameters for the PSM method.

Table 1. Baseline characteristics for the retrospective data.

	Minimally invasive surgery	Open surgery	p-value
Age (years)	52 (45 : 57)	49 (39 : 57)	0.207
Sex: Female	32 [50.79 %]	82 [65.08 %]	0.083
Sex: Male	31 [49.21 %]	44 [34.92 %]	
Level: L1-L2	0 [0 %]	2 [1.59 %]	<0.001*
Level: L2-L3	6 [9.52 %]	3 [2.38 %]	
Level: L3-L4	12 [19.05 %]	13 [10.32 %]	
Level: L4-L5	43 [68.26 %]	51 [40.47 %]	
Level: L5-S1	2 [3.17 %]	57 [45.24 %]	
IHD: 0	31 [49.21 %]	84 [66.67 %]	0.031*
IHD: 1	32 [50.79 %]	42 [33.33 %]	

In order to choose the optimal PSM model to minimize systematic selection bias while leaving a sufficient number of patients, a heuristic algorithm needs to be built using the empirical clinical data within a reasonable time. We found several qualitatively equivalent models (sets of variables) in which the systematic selection bias for clinically relevant parameters occurring as the comparison groups were chosen has been eliminated. When planning our study design, we selected two criteria for evaluating the adequacy of model selection for calculating the propensity score in the PSM method and for choosing variables for algorithm matching. The first criterion was elimination of systematic selection bias at baseline. The second criterion implied qualitative coincidence between the results of PSM analysis of retrospective data and those of the randomized prospective study.

We analyzed the statistical significance of the results depending on sample size to select several variants of variables for the matching algorithm. According to the results of the prospective study, in order to reveal type I error (assuming that normal data distribution is used), the sample size must be at least 27. A heuristic search identified 3 variants for group matching depending on matching algorithm parameters described below. Using the PSM method is always about weighing sample size with the quality of minimizing systematic selection bias (Figure 2). On one hand, the bias can be minimized more efficiently for a smaller sample size. On the other, the ability to reveal differences in selected subgroups decreases as the sample size decreases, preventing one from obtaining statistically significant results. When the sample size consists of several hundred subjects, we always have to find the golden mean between the two opposite tendencies. Therefore, it is no wonder that the visualized possible variants usually look as follows (the value of matching parameters is plotted on the X axis and model completeness (amount of independent parameters of model), on the Y axis).

The results shown in Table 2 and Figure 3 correspond to these cases. Hence, the problem of searching for the optimal set of variables that will meet all the objectives is a particular problem that must be solved using biostatistical evaluation. To better illustrate the behavior of PSM, we provide a graphical representation of the results for all three variants.

We can see that in the first case that the distributions of propensity score differed in the selected subgroups. In the second and

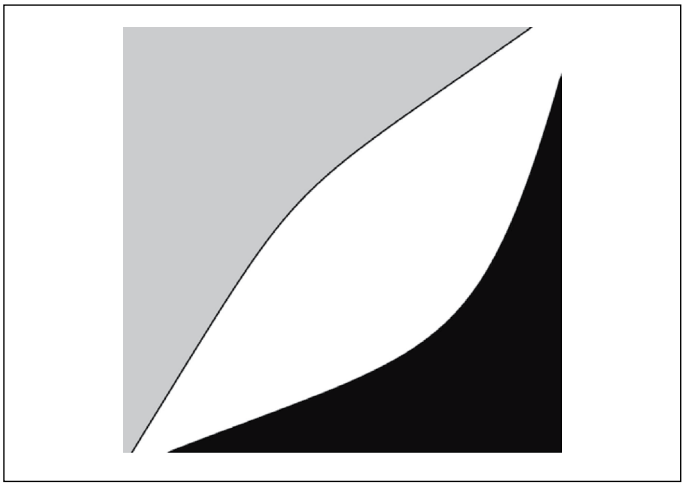

Figure 2. Zone 1 (black): bias was not eliminated using the PSM method; zone 2 (white): bias was eliminated and the sample size enabled us to obtain statistically valid results; zone 3 (gray): bias was eliminated; the size of comparison groups did not enable us to obtain statistically significant intergroup differences..

third cases, the distributions of propensity score were similar but the sample size was noticeably larger in the second case. These diagrams enable us to visually estimate the quality of PSM.

The second model for selecting parameters for the PSM method enabled us to adequately eliminate baseline bias. The groups were matched with respect to age, level of the surgery, and presence of concomitant pathology (ischemic heart disease, IHD). The outcomes according to the VAS and ODI scores qualitatively coincided, showing superiority of the outcomes in the group of subjects treated by minimally invasive surgery to the outcomes in patients treated by open surgery (Table 3).

Hence, the use of PSM and selection of the optimal statistical characteristics and variables that take into account patients' key clinical characteristics enables us to minimize selection bias and obtain comparable results to those of a similar randomized clinical trial.

DISCUSSION

The objectives of our study were to compare different models and methods for variable selection for PSM when analyzing retrospective clinical data, and to compare the results of using the PSM method with those of prospective studies. In the general variant of the PSM method, the objective was to select patient groups that would be pairwise matched with respect to their clinical characteristics.

In order to solve this problem, it is necessary to identify a set of these clinical characteristics (hereinafter referred to as comparison parameters) and to mathematically determine the minimum permissible caliper for the patients with respect to these characteristics. The latter parameter is determined in different ways, depending on type of PSM algorithm used; in general cases, we will refer to it as "caliper".

Table 2. Significance levels for baseline characteristics for different variants of the PSM method.

Name	PSM1 (caliper = 1.0)	PSM2 (caliper = 0.6)	PSM3 (caliper = 0.02)
Number of patients in comparison groups	44	40	13
Baseline			
Age (years)	0.257	0.813	0.609
Levels	0.792	0.709	0.832
Sex	0.009*	0.095	0.111
Pathology: IHD	0.666	>0.999	>0.999
Results			
ODI (24m)	0.006	0.007	0.076
VAS-spine (post-)	0.526	0.858	0.768
VAS-leg (post-)	0.910	0.976	0.811

Table 3. Comparison of the outcomes of the prospective study and the optimal PSM.

	Minimally invasive surgery	Open surgery	p-value
ODI(12m)			
prospective	13 (10 : 20)	18 (12 : 24)	0.014*
PSM2	12 (10 : 20)	18 (14 : 26)	0.006*
VAS-spine(after)			
prospective	2.25 (2 : 4)	2.5 (2 : 4)	0.517
PSM(2)	2 (2 : 4)	2.5 (2 : 3)	0.526
VAS-leg(after)			
prospective	1 (0 : 2)	2 (0 : 2.75)	0.733
PSM(2)	1 (0 : 2)	1 (0 : 2)	0.910

Ideally, we would like to choose all the known clinical characteristics at baseline and from the patient's medical history as parameters for comparison, as this will make the design maximally similar to that of a randomized study. However, one of the key practical problems associated with building models is that the number of patients with all matched characteristics decreases significantly as the number of parameters is increased. This problem can be solved by increasing the permissible propensity score radius; however, in this case, selection makes no sense, since patients become incomparable. In other words, this PSM model does not eliminate systematic selection bias.

We used the heuristic algorithms to select a model to calculate the propensity score and select matching algorithm parameters. Application of the PSM method significantly affected the results of the retrospective data analysis. The results of retrospective and randomized prospective studies qualitatively coincided after systematic selection bias was eliminated.

Recent reviews of the propensity score matching method in medical publications have demonstrated that a wide choice of calipers has been used in specific applications. The choice of caliper usually appeared to have been ad hoc, and not based on substantive theory. We would like to mention that previous experience of using the PSM method is frequently not taken into account when choosing parameters, because the PSM method has not been sufficiently described in these studies. Indeed, very few studies have focused on the relative performance of different calipers for propensity score matching. The CNN algorithm was based on matching on the logit of the propensity score method based on fixed caliper widths on the propensity-score scale (0.05, 0.1, 0.25, 0.5, and 0.8). These caliper widths were chosen because they were the ones most frequently used in practice in medical publications. Other researchers have matched on the propensity score using calipers of width 0.005, 0.01, 0.02, 0.03, and 0.1 on the propensity score scale.[11] Nearest neighbor

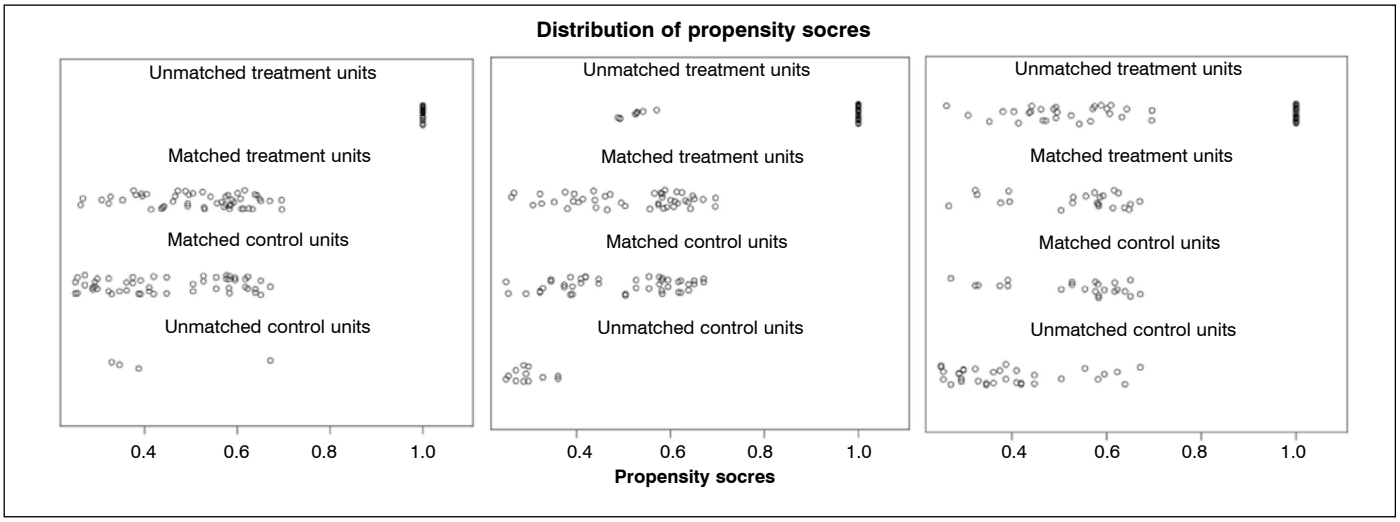

Figure 3. Distribution of propensity scores in the comparison groups. (A)=PSM1, (B)=PSM2, (C)=PSM3.

matching within fixed caliper widths attempts to match each treated subject to the nearest untreated subject (on either the propensity score scale or on the logit of the propensity score scale) within a specified caliper width: matched treated and untreated subjects can only have propensity scores (or logit of the propensity score) that differ by, at most, a fixed, pre-specified amount (the caliper width).

Some researchers have shown that matching on a normally distributed confounding variable using calipers of widths 0.2 and 0.6 eliminates approximately 90 and 99% of the bias, respectively, due to this confounding variable (Cochran and Rubin, 1973).

In our opinion, using the PSM method for retrospective data enables us to obtain results close to those results obtained by randomized studies. We believe that the development and application of mathematical methods for eliminating bias in retrospective data analysis is one of the key tools for designing successful prospective studies.

Comparison of the results of prospective and retrospective studies demonstrated that selection bias can be qualitatively minimized in retrospective data analyses, even for small sample sizes. This enables us to obtain clinically substantiated and valid results by processing the retrospective data, which can be extrapolated and taken into account when planning and conducting prospective studies. The PSM method holds great promise for eliminating selection bias not only in retrospective studies, but also in processing the results of observational and cohort studies. One of the benefits of the PSM method with caliper is that the level of bias balancing can be selected (from 90% to 99.9% in analysis of synthetic tests), making it possible to vary sample size and determine intergroup variation by statistical analysis.[14]

CONCLUSIONS

The recommendations made for the propensity score matching method can now be summarized based on the results of this study. When evaluating the ODI or VAS scores in neurosurgical studies (with a sample size < 150 patients), we recommend matching on the logit of the propensity score using calipers of width 0.6 of the standard deviation of the logit of the propensity score. In many earlier studies, the focus has been placed on the effect of caliper width on bias correcting. We recommend using CNN 0.2-0.6.

All authors declare no potential conflict of interest related to this article.

CONTRIBUTION OF THE AUTHORS: Each author made significant individual contributions to this manuscript. AVK: drafting of the entire research project, intellectual concept, revision and performing the surgeries, statistical analysis, review the manuscript. SAA performed the literature search and review of the manuscript, and contributed to the intellectual concept of the study. KYO performed the surgery, followed up the patients and gathered clinical data. VSG and AVM evaluated the data from the statistical analysis.

REFERENCES

1. Austevoll IM, Gjestad R, Brox JI, Solberg TK, Storheim K, Rekeland F, et al. The effectiveness of decompression alone compared with additional fusion for lumbar spinal stenosis with degenerative spondylolisthesis: a pragmatic comparative non-inferiority observational study from the Norwegian Registry for Spine Surgery. Eur Spine J. 2017;26(2):404–13.
2. Glassman SD, Carreon LY, Ghogawala Z, Foley KT, McGirt MJ, Asher AL. Benefit of Transforaminal Lumbar Interbody Fusion vs Posterolateral Spinal Fusion in Lumbar Spine Disorders: A Propensity-Matched Analysis from the National Neurosurgical Quality and Outcomes Database Registry. Neurosurgery. 2016;79(3):397–405.
3. Zweig T, Enke J, Mannion AF, Sobottke R, Melloh M, Freeman BJC, et al. Is the duration of pre-operative conservative treatment associated with the clinical outcome following surgical decompression for lumbar spinal stenosis? A study based on the Spine Tango Registry. Eur Spine J. 2017;26(2):488–500.
4. Concato J, Shan N, Horwitz R. Randomized , controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342(25):1887–92.
5. Graf E. The propensity score in the analysis of therapeutic studies. Biometrical J. 1997;39(3):297–307.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
7. Brookhart MA, Wyss R, Layton JB, Stürmer T. Propensity score methods for confounding control in nonexperimental research. Circ Cardiovasc Qual Outcomes. 2013;6(5):604–11.
8. Hartnett ME, Tinkham N, Paynter L, Geisen P, Koch G, Cohen KL. 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. 2006;59(5):437–47.
9. Russo GI, Kurbatov D, Sansalone S, Lepetukhin A, Dubsky S, Sitkin I, et al. Prostatic Arterial Embolization vs Open Prostatectomy: A 1-Year Matched-pair Analysis of Functional Outcomes and Morbidities. Urology. 2015;86(2):343–8.
10. Caus T, Sirota D, Nader J, Lyashenko M, Chernyavsky A. Associated bare stenting of distal aorta with a Djumbodis® system versus conventional surgery in type A aortic dissection. Ann Cardiothorac Surg. 2016;5(4):336–45.
11. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and monte carlo simulations. Biometrical J. 2009;51(1):171–84.
12. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011;10(2):150–61.
13. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med. 2014;33(6):1057–69.
14. 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. 2007;26(4):734–53.