Propensity score matching (PSM) is an increasingly applied method of ensuring comparability between groups of interest. However, PSM is often applied unconditionally, without precise considerations. The purpose of this study is to provide a nonmathematical guide for clinicians at the stage of designing a PSM-based study. We provide a seed of thought for considering whether applying PSM would be appropriate and, if so, the scope of the list of variables. Although PSM may be simple, its results could vary substantially according to how the propensity score is constructed. Misleading results can be avoided through a critical review of the process of PSM.

Keywords: Propensity score; Propensity score matching; Observational study; Selection bias; Indication bias
score [2, 4, 5]. Therefore, subjects with similar propensity scores are comparable, since the confounding variables are balanced, and an unbiased result could be derived from a comparison of groups consisting of the selected study participants with similar propensity scores.

Still, PSM may leave the wrong impression unless each methodological step is correctly conducted. It sounds obvious that the propensity score should be calculated correctly in order to reach a solid conclusion. As PSM has been widely adopted in the field of coloproctology, we should critically review the methods before believing the results. This article was conducted to review the process of PSM conducted in original articles published in *Annals of Coloproctology*. There were 5 articles using PSM, as described in Table 1 [6-10], and each was reviewed for issues related to comparability between the comparison groups and the appropriateness of applying PSM.

**COMPARABILITY BETWEEN COMPARISON GROUPS**

The premise of PSM is the “exchangeability” between comparison groups; as mentioned before, RCTs produce the most robust evidence due to randomization, and properly constructed PSM could be a good alternative to random allocation in a retrospective observational study.

In a study comparing 2 different diagnostic tests, the design would take intrapatient comparability for granted (i.e., conducting both tests in every study participant so that a paired test could be performed). However, in studies comparing treatments, each patient only receives 1 treatment, meaning that each group involves different subject groups. Patients are typically not allocated to treatment randomly in clinical practice; instead, treatment is assigned based on the clinician’s subjective judgment or the patient’s choice. Consequently, the study participants included in each comparison group would reflect selection bias or indication bias, and retrospective analyses would lead to substantially biased results for treatment (e.g., measures of death or recurrence). Therefore, a mere comparison of “face values” without accounting for all potential confounders may lead to a false conclusion. This ultimately translates to incorrect medical practice, once described as a “scandal of poor medical research” [11].

Most researchers believe that PSM enhances the quality of observational studies by increasing the comparability between each group through a reduction in the extent of the unequal distribution of various clinical factors due to selection bias or indication bias. Nonetheless, unreasonable comparisons should be avoided, and an “exchangeable” comparison group should be selected. In other words “comparing apples and oranges” should be avoided.

In the study by Kataoka et al. [6], the participants were divided into 2 according to the C-reactive protein to albumin ratio (CAR) in a preoperative examination and at postoperative days 1 and 7 using cut-off values of CAR determined by analyzing the receiver operating characteristic curves for the survival rate. However, the CAR has been reported to be associated with colorectal cancer prognosis [12, 13]. As the classification rationale is directly associated with the results, indication bias would occur. Indication bias, 

<table>
<thead>
<tr>
<th>Study</th>
<th>PICO</th>
<th>Description (No. of patients before/after PSM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kataoka et al. [6]</td>
<td>Patients</td>
<td>Colorectal cancer</td>
</tr>
<tr>
<td></td>
<td>Intervention vs. comparison</td>
<td>Patients with a high C-reactive protein to albumin ratio (CAR, n = 75/n = 72) vs. patients with low CAR (n = 245/n = 72)</td>
</tr>
<tr>
<td></td>
<td>Outcome</td>
<td>Relapse-free survival, overall survival</td>
</tr>
<tr>
<td>Hyun et al. [7]</td>
<td>Patients</td>
<td>Clinical T1 rectal cancer</td>
</tr>
<tr>
<td></td>
<td>Intervention vs. comparison</td>
<td>Local excision (n = 106/n = 91) vs. total mesorectal excision (n = 965/n = 91)</td>
</tr>
<tr>
<td></td>
<td>Outcome</td>
<td>Overall survival</td>
</tr>
<tr>
<td>Park et al. [8]</td>
<td>Patients</td>
<td>Colorectal cancer with synchronous resectable liver metastasis</td>
</tr>
<tr>
<td></td>
<td>Intervention vs. comparison</td>
<td>Neoadjuvant chemotherapy followed by operation (n = 66/n = 60) vs. surgical resection only (n = 363/n = 60)</td>
</tr>
<tr>
<td></td>
<td>Outcome</td>
<td>Disease-free survival, overall survival</td>
</tr>
<tr>
<td>Nasir et al. [9]</td>
<td>Patients</td>
<td>Locally advanced rectal tumors were identified as T3B or T4</td>
</tr>
<tr>
<td></td>
<td>Intervention vs. comparison</td>
<td>Locally advanced rectal cancer (LARC, ND/n = 109) vs. non-LARC (NLARC, ND/n = 109)</td>
</tr>
<tr>
<td></td>
<td>Outcome</td>
<td>Short-term outcomes related to the operation (e.g., conversion to open surgery, clinical anastomotic leak, readmission, reoperation, 30-day mortality)</td>
</tr>
<tr>
<td>Yun et al. [10]</td>
<td>Patients</td>
<td>Colorectal cancer</td>
</tr>
<tr>
<td></td>
<td>Intervention vs. comparison</td>
<td>Signet-ring-cell carcinomas (n = 71/n = 71) vs. adenocarcinomas (n = 12,570/n = 142)</td>
</tr>
<tr>
<td></td>
<td>Outcome</td>
<td>Clinical outcome, overall survival, disease-free survival</td>
</tr>
</tbody>
</table>

PSM, propensity score matching; ND, not described.
also referred to as confounding by indication, is a specific type of selection bias that is primarily responsible for the incomparability between groups in retrospective analyses of clinical data [1, 14]. This bias occurs when a patient's condition that determines the selection of any particular treatment is also associated with the outcome of the treatment. Therefore the previously mentioned “exchangeability” is not established between patients with high and low CARs. This violates the “counterfactual assumption” of PSM, making the results unreliable.

In the study by Nasir et al. [9], locally advanced rectal cancer (LARC) and non-LARC (NLARC) patients were compared using PSM. However, the criterion for defining each comparison group was invasion depth, which is not a parameter that we can modify at will. As in the abovementioned study by Kataoka et al. [6], this would involve “comparing apples and oranges.” It would rather be better to compare outcomes according to whether patients underwent laparoscopy in each stratum (LARC or NLARC), and the efficacy of laparoscopy in patients with NLARC should be determined according to differences in the quality of surgical specimens, morbidity, and mortality.

Likewise, in the study by Yun et al. [10], histology—as a non-“exchangeable” parameter—was the criterion used for the classification of comparison groups. As signet-ring-cell carcinoma is a rare subtype compared to adenocarcinoma, it would instead be preferable to select a comparable set of adenocarcinoma by directly matching a set of variables [15], including various underlining characteristics that could affect the prognosis of patients (e.g., age, sex, preoperative carcinoembryonic antigen [CEA] levels, location, operation method, stage, lymphatic invasion, vascular invasion, perineural invasion, and adjuvant treatment). A propensity score is the conditional probability of receiving a specific treatment exposure given a set of covariates, and the key property of a propensity score is exchangeability between comparison groups [16]. If the exchangeability assumption is unclear or inapplicable, it would be helpful to design an RCT with each comparison group.

THE APPROPRIATENESS OF APPLYING PROPENSITY SCORE MATCHING

The process of PSM could be divided into 5 steps, including (1) estimation of the propensity score, (2) evaluation of the propensity score distribution, (3) PSM, (4) assessment of the balance in covariates between groups after PSM, and (5) the main analysis of between-group differences after PSM. Thus, PSM is of no use if the first step is not performed properly [2].

Multivariable logistic regression modeling is commonly used for estimating the propensity score, in which potential confounders for group allocation to be adjusted are included as independent variables (“x” variables), and the group assignment is included as the dependent variable (“y” variable). Unlike randomization, which ensures a random allocation of unmeasured confounders, the propensity score method cannot overcome biases caused by variables that are not included in the model [2, 5, 17].

Thus, it is generally better to include as many potential confounders as possible in the propensity score model as independent variables. Although debate continues regarding variable selection during propensity score model construction, variables that are related to both group assignment and the outcome can satisfy the ignorable group assignment and minimize study bias [2, 5, 17-19].

Propensity scores were estimated by Kataoka et al. [6] based on a logistic regression model addressing the following 11 variables: age as a continuous variable, sex, American Society of Anesthesiologists physical status classification, serum CEA level, serum carbohydrate antigen 19-9 (CA19-9) level, tumor location, tumor size, tumor histotype, lymphatic invasion, venous invasion, and TNM stage as categorical variables. While the variable selection for propensity score estimation seems to be appropriate, the propensity score could be different if the serum CEA level or CA19-9 level is included as a continuous variable in the model. Likewise, age could be included as a categorical variable in a clinically meaningful manner (such as age of <65 or ≥65 years, as used by Yun et al. [10]).

In contrast, Hyun et al. [7] constructed a propensity score with only 5 variables: sex, age, tumor location, tumor size, and T classification. Even Nasir et al. [9] used only tumor height (low vs. middle vs. high) to estimate the propensity score. While PSM is expected to select an experimental study-like dataset mimicking randomization by removing sources of incomparability between groups, a proper selection of independent variables during the propensity score estimation is absolutely crucial for the validity of the propensity score method [2, 20].

Therefore, it is necessary to think about which variables should be selected among various candidates with clinical importance. Moreover, how the selected variables are included could affect the estimation of the propensity score. If categories are used for continuous covariates, clinically meaningful thresholds (e.g., normal vs. abnormal serum CEA levels) are recommended over data-driven classifications (e.g., quartiles). Moreover, when continuous variables are believed to have non-linear associations, adding quadratic or even cubic terms (e.g., age² + age³) or a transformation (e.g., logarithm) could allow more flexible fitting of the data. Interaction terms could also be considered [21]. For this reason, the propensity score could function like a black box, and the result could be different according to how the propensity score is estimated. Therefore, a sensitivity analysis should be conducted to reveal whether the result is robust regardless of the propensity score model. While the c-index (the area under the receiver operating characteristic curve of the logistic regression model) is often used to assess the adequacy of a propensity score model, empirical evidence may be used to augment preexisting knowledge [2, 18, 21, 22].

Additionally, any possible changes in the extent of the imbalance...
of covariates used to construct the propensity score after PSM should be checked thoroughly, and a P-value is not sufficient because it depends on the sample size. Intergroup differences are usually measured with the standardized mean difference (SMD) [23] and an improved balance after PSM compared to that before PSM could be described in terms of the change in the SMD for each variable used to estimate the PS. However, all 5 articles analyzed herein presented P-values to represent the balance in covariates between groups after PSM. Still, as mentioned above, only the observed and included variables for propensity score estimation could be balanced. Therefore, residual confounding is possible, so a multivariable model to compare the selected comparison groups after PSM could be considered.

CONCLUSION

While PSM is a good alternative to randomization for retrospective observational studies, the covariates for propensity score estimation should be selected carefully among those with clinical importance. However, this process is often ignored and applied unconditionally. Therefore, it is necessary to review issues in the PSM process, including whether the comparison group shows comparability or exchangeability, as well as the appropriateness of applying PSM. A step-by-step checklist for each process of PSM could be used for objective and transparent reporting [21, 24, 25], and sensitivity analyses with various propensity score models could be used for objective and transparent reporting [21, 24, 25], and sensitivity analyses with various propensity score models should be conducted actively to reveal whether the results are robust. Well-conducted PSM using a well-estimated propensity score can be a superb surrogate for RCTs using real-world data.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

FUNDING

None.

ORCID

Hwa Jung Kim, https://orcid.org/0000-0003-1916-7014

REFERENCES