In clinical research, the gold standard level of evidence is the randomized controlled trial (RCT). The availability of nonrandomized retrospective data is growing; however, a primary concern of analyzing such data is comparability of the treatment groups with respect to confounding variables. Propensity score matching (PSM) aims to equate treatment groups with respect to measured baseline covariates to achieve a comparison with reduced selection bias. It is a valuable statistical methodology that mimics the RCT, and it may create an “apples to apples” comparison while reducing bias due to confounding. PSM can improve the quality of anesthesia research and broaden the range of research opportunities. PSM is not necessarily a magic bullet for poor-quality data, but rather may allow the researcher to achieve balanced treatment groups similar to a RCT when high-quality observational data are available. PSM may be more appealing than the common approach of including confounders in a regression model because it allows for a more intuitive analysis of a treatment effect between 2 comparable groups.

We present 5 steps that anesthesiologists can use to successfully implement PSM in their research with an example from the 2015 Pediatric National Surgical Quality Improvement Program: a validated, annually updated surgery and anesthesia pediatric database. The first step of PSM is to identify its feasibility with regard to the data at hand and ensure availability of data on any potential confounders. The second step is to obtain the set of propensity scores from a logistic regression model with treatment group as the outcome and the balancing factors as predictors. The third step is to match patients in the 2 treatment groups with similar propensity scores, balancing all factors. The fourth step is to assess the success of the matching with balance diagnostics, graphically or analytically. The fifth step is to apply appropriate statistical methodology using the propensity-matched data to compare outcomes among treatment groups.

PSM is becoming an increasingly more popular statistical methodology in medical research. It often allows for improved evaluation of a treatment effect that may otherwise be invalid due to a lack of balance between the 2 treatment groups with regard to confounding variables. PSM may increase the level of evidence of a study and in turn increases the strength and generalizability of its results. Our step-by-step approach provides a useful strategy for anesthesiologists to implement PSM in their future research. (Anesth Analg XXX;XXX:00–00)
PSM is an intuitive alternative to a multiple regression modeling approach. PSM incorporates the information provided by the baseline factors into 1 propensity score and is used to balance the treatment groups of those factors.

In this statistical primer, we present a 5-step approach for the anesthesia researcher to successfully implement and evaluate PSM to compare 2 treatments using observational data. Herein, we may refer to treated and untreated groups or more generally to treatment groups and use these terms interchangeably, as PSM may be used to compare treatment to standard care, to compare 2 active treatments, or for comparisons of related scenarios. PSM is growing in appeal in the clinical research community and is quickly becoming a highly regarded approach for analyzing nonrandomized, observational data and adjusting for potentially confounding variables. Each step will be practically explained and applied to an anesthesia example. These 5 steps are as follows:

1. Identify that PSM is viable and appropriate
2. Calculate the propensity scores
3. Match subjects on the propensity scores
4. Assess balance diagnostics to determine the quality of the matching
5. Analyze the propensity-matched cohort

METHODS

Step 1: Identify That PSM Is Viable and Appropriate

PSM is a viable option when you have nonrandomized, observational data of high quality that include information on baseline factors, exposures and treatments, and outcomes of interest. Most important is that information on potential confounding variables is included in the data and that these factors that are suspected to be related to the outcome of interest seem to be imbalanced between the 2 treatment groups. The nature of the nonrandomized data is such that perhaps 1 treatment is given to certain kinds of patients more readily in practice.

For our NSQIP example, as noted above, we will exclude the 300 cases where information on weight at surgery is missing, as this is a confounder that we will consider. Similar to analyses using multivariable regression, alternative approaches to missing data are available and should be considered by the investigator based on the extent, reason for, and potential impact of missing data.

Table 1 shows the comparison of baseline factors between our 2 treatment groups. We should be concerned because clear imbalances are seen between the 2 oxygen support groups with respect to all baseline factors (all \( P < .001 \)), and in this population, these factors are also known to be associated with the 30-day mortality, our outcome of interest. We see that the proportion of neonates is higher in the oxygen support group (26.68%) as compared to the no oxygen support group (3.73%). The same is true for inotropic support (7.58% vs 0.29%), ventilator dependence (49.61% vs 1.50%), and occurrence of previous cardiac surgery (17.49% vs 3.12%) comparing the 2 exposure groups. The mean weight at surgery in the oxygen support group is 10.7 kg compared to 31.49 kg in the no oxygen support group. PSM should be considered to treat these imbalances.

Here, we used parametric tests to compare exposure groups with respect to the baseline variables because we have a very large sample size and our baseline factors are not known to be skewed. If, for instance, we were looking at hospital length of stay as a variable to be compared between the groups, we would opt for nonparametric testing because hospital length of stay was generally skewed.

Step 2: Calculate the Propensity Scores

The estimated propensity score can be conceptualized as the patient’s probability of being treated as a function of measured baseline covariates. Thus, it is equal to a probability of treatment given baseline factors. Most studies consider

DATA SOURCE

We will demonstrate each step with an anesthesiology example using the Pediatric 2015 data from the American College of Surgeons National Surgical Quality Improvement Program (NSQIP). The 2015 American College of Surgeons NSQIP Pediatric data feature approximately 84,000 cases for approximately 120 variables ranging from baseline factors to postoperative outcomes. The NSQIP data are audited to ensure high quality and reliability.

The example that we will use with the NSQIP data is the following. We will explore if required oxygen support at the time of surgery is associated with 30-day postoperative mortality. Therefore, oxygen support is the exposure defining “treatment” group, and 30-day mortality is the outcome of interest. The measured baseline factors what we suspect to be confounders are as follows: neonatal status (age <30 days at the time of surgery), intravenous inotropic pharmacologic support at the time of surgery, ventilator dependence within 48 hours before surgery, history of a previous cardiac surgery, and weight at surgery (kg). Because data for weight at surgery are missing on 300 surgeries, these have been excluded from the analysis, bringing the total sample size to 83,756. Note that the 2015 Pediatric NSQIP features a very low 30-day mortality event rate of 0.37% and includes information about all baseline confounding factors, so PSM should be considered.

We used the PSMATCH2 software package in Stata (StataCorp LLC, College Station, TX) to implement PSM in our example. The Stata code used for the clinical example will be included in Supplemental Digital Content, Appendix, http://links.lww.com/AA/C198, for reference. Statistical software other than Stata can be used to perform PSM. SPSS (IBM Corp, Armonk, NY) has a PSM tab under the data tab; SAS/STAT 14.2 (SAS Institute Inc, Cary, NC) has the capability to perform PSM with its PSMATCH procedure; and R (R Foundation for Statistical Computing, Vienna, Austria) features the packages Matching, MatchIt, and Optmatch. We use R for figures in our clinical example.
a binary treatment (treated versus untreated). In that scenario, the propensity score can be estimated from a multi-variable logistic regression model fit with treatment group as the outcome. The covariates in the multivariable model are the baseline factors on which we wish to balance our treatment groups. Note that many variables can be used as covariates in this model, and regardless of how many, we will only obtain a single propensity score.

There are no stringent rules pertaining to the choice of predictors of treatment group. However, it should be noted that the model to obtain the propensity scores with treatment group as the outcome cannot include post-baseline variables that may be modified by treatment.8

A study by Austin8 considered the merits of including different sets of variables in the propensity model, including “all measured baseline covariates, all baseline covariates that are associated with treatment assignment, all covariates that affect the outcome (ie, potential confounders), and all covariates that affect both treatment assignment and the outcome (ie, true confounder).” This built on a study by Austin et al6 who performed a simulation study regarding relative benefits of including the different set of baseline factors. They found that using a propensity score model that includes only true confounders or includes all variables associated with the outcome reduces bias slightly better than when the propensity score model includes all measured variables or only includes the variables associated with treatment selection. Thus, we recommend the anesthesia researcher include variables associated with the outcome(s) of interest that occur temporally before the treatment, a subset of which will be confounders. Had an RCT been performed, we would expect these variables to be balanced at baseline.

One does not need to perform an iterative model building approach for the propensity score model; a covariate can still be included even if its associated P value is not statistically significant. The goal of this model is to accurately produce a fitted probability of treatment group while using variables that predict the outcome or that predict both the treatment and the outcome (a propensity score for each subject using a logistic regression framework).

Once the logistic regression model has been fit, the propensity score for each patient can be obtained as the fitted probability. In our example, the propensity score model includes variables known to be associated with 30-day mortality and that are imbalanced between oxygen support groups: neonatal status, inotropic support, ventilator dependence, previous cardiac surgery, and weight at surgery.

Subjects with similar propensity scores may be matched to one another, so we must see overlap in the distributions of propensity scores between the 2 treatment groups to be able to implement matching.5 Figure 1 below is a mirrored histogram of the propensity scores by treatment group in the prematched “raw” data. We can see that the propensity scores are distributed differently in each treatment group, which indicates that the groups have patients with generally different baseline factors. One requirement of PSM is that there is common support, such that we see some overlap in the mirror histogram of propensity scores to be able to better perform matching in our next step. Because oxygen support is fairly rare in this dataset and therefore has a low probability in the data, notice that most of the probability mass of the distribution of the propensity scores is close to 0, but we do see overlap of propensity scores between the 2 groups of interest.

Step 3: Match Subjects on the Propensity Scores

It should be noted that there are several options for using the propensity scores after they are obtained. We will focus on an intuitive and commonly used option: matching using the propensity score. The other approaches have certain appeals and limitations but have the same goal to balance treatment groups with respect to the measured baseline factors that went in to calculating the propensity scores. The other approaches are stratification by the propensity score, using the propensity score for inverse probability of treatment weighting, and including the propensity score as a regression covariate.8

**Table 1. Prematched Baseline Covariates**

<table>
<thead>
<tr>
<th>Baseline Covariate</th>
<th>No Oxygen Support Group (n = 80,446)</th>
<th>Oxygen Support Group (n = 3310)</th>
<th>Total Cohort (n = 83,756)</th>
<th>P Value</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neonate (age &lt;30 d)</td>
<td>2997  3.73%</td>
<td>883  26.68%</td>
<td>3880  4.63%</td>
<td>&lt;.001</td>
<td>0.675</td>
</tr>
<tr>
<td>Inotropic support</td>
<td>231   0.29%</td>
<td>251   7.58%</td>
<td>482   0.58%</td>
<td>&lt;.001</td>
<td>0.382</td>
</tr>
<tr>
<td>Ventilator dependence</td>
<td>1208  1.50%</td>
<td>1642  49.61%</td>
<td>2850  4.0%</td>
<td>&lt;.001</td>
<td>1.322</td>
</tr>
<tr>
<td>Previous cardiac surgery</td>
<td>2509  3.12%</td>
<td>579   17.49%</td>
<td>3088  3.69%</td>
<td>&lt;.001</td>
<td>0.486</td>
</tr>
<tr>
<td>Weight at surgery (kg)</td>
<td>31.49  25.25</td>
<td>10.7   15.24</td>
<td>30.68  25.26</td>
<td>&lt;.001</td>
<td>0.997</td>
</tr>
</tbody>
</table>

Abbreviation: d, absolute standardized mean difference.
Matching using the propensity score is exactly as it sounds: most commonly, the investigator starts with a group of treated patients and selects untreated patients with similar propensity scores as matches to those treated patients. In this primer, we demonstrate 1-to-1 nearest-neighbor matching. That is, 1 patient from the treated group will get matched with 1 patient from the untreated control group with a similar propensity score.10 However, PSM is flexible enough to allow for 2-to-1 matching, or integer score.8 The third choice is the size of the caliper width. The investigator should consider the number of untreated controls available for matching and potential quality of matches when determining k. Other matching techniques that we will not discuss but that are available as options to use include Mahalanobis metric matching and interval matching.

When matching is done, several choices must be made. The first is between matching with replacement and matching without replacement.13 We will perform matching without replacement, so once an untreated subject is matched to a treated subject, they are not available to be matched again to a different treated subject. When matching without replacement is performed, the estimates may depend on the order that the matching is done. To minimize the influence of possibly systematic orderings of the subjects in the dataset, researchers might consider randomly ordering their data.13 Matching with replacement allows an untreated control to act as the match for possibly multiple-treated patients. This may perform better in the case where many treated patients are excluded when matching without replacement is done. This may occur if there is no good overlap (known as common support) between the propensity scores of the treatment and control patients or if there are more or a similar number of treated patients than untreated controls. Investigators will need to use methods to account for matching with replacement in Step 5. The second is between optimal and greedy matching. With greedy matching, a treated patient is selected, then an untreated patient with the closest propensity score to that of the treated subject is matched to the chosen treated patient. This is done sequentially through the list of treated patients until the list is exhausted of treated patients for whom untreated patients can be found.8 The other option is optimal matching, which we use in our clinical example. Optimal matching minimizes the total within-pair difference of propensity score.8 The third choice is the size of the caliper width. The caliper width is the maximum difference between propensity scores allowed for 2 patients to potentially be matched. Austin14 suggests that when matching patients on the propensity scores within a certain caliper width, the optimal caliper width is equal to 0.2 times the standard deviation of the logit of the propensity scores (the logit of a given propensity score “p” is equal to the natural logarithm of p/[1 − p], also referred to as the log odds).

Statistical software packages will perform the matching since PSM by hand can be tedious for large datasets. Options for type of matching to be performed can be specified.

---

**Step 4: Assess Balance Diagnostics to Determine the Quality of the Matching**

Once the matching is complete, we need to assess the balance of the 2 groups on the baseline factors. If our matching was successful, then the 2 groups will be balanced and therefore we will be able to move on to Step 5 and isolate a true treatment effect from any confounding. Quality of matching refers to the extent that the treatment groups are balanced with respect to baseline factors once matching is done. Since P values comparing the baseline factors between the 2 treatment groups are highly driven by sample size,15,16 we recommend 2 analytic measures of matching quality: the absolute standardized mean difference for each baseline factor17,18 and the reduction in pseudo R² in the logistic regression model for exposure group (the propensity score model).19

The absolute standardized mean difference is a numeric summary that can be calculated for every baseline covariate, whether continuous or binary. It compares each baseline factor between the treatment and the control groups after PSM is completed and uses a pooled standard deviation calculation. An absolute standardized mean difference < 0.1 has been taken to indicate a negligible difference between groups for that covariate.18 For a continuous covariate, the absolute standardized mean difference is defined as follows:

\[ d = \frac{|\bar{x} - \bar{t}|}{\sqrt{\frac{s^2_x + s^2_t}{2}}} \]

where \(\bar{x}\) and \(\bar{t}\) are the sample means and \(s^2_x\) and \(s^2_t\) are the sample variances in the treatment and control groups, respectively. For a binary baseline factor, the absolute standardized mean difference is defined as follows:

\[ d = \frac{|\hat{p}_x - \hat{p}_t|}{\sqrt{\frac{\hat{p}_x(1 - \hat{p}_x) + \hat{p}_t(1 - \hat{p}_t)}{2}}} \]

where \(\hat{p}_x\) and \(\hat{p}_t\) are the prevalences in the treatment and the control groups, respectively.17 Note that all absolute standardized mean differences in our example are < 0.1, reflecting balance between the treatment groups (Table 2).

We can look at the pseudo R² from the logistic regression model for treatment in the unmatched raw data, and then again in the matched data. The pseudo R² is calculated by any statistical software when running a logistic regression. Because pseudo R² is a measure of predictive value of a logistic regression model, a reduction in pseudo R² from the propensity score model in the raw data compared to the same model in the matched data means that the baseline factors are no longer predictive for determining treatment group, as desired.19 In our NSQIP example, the pseudo R² for the propensity score model in the unmatched data is 0.3556 where in the 1-to-1 matched cohort, it is 0.0008. This is a large reduction in pseudo R², so the PSM has successfully reduced the predictive value of neonatal status, inotropic support, ventilator dependence, history of a previous cardiac surgery, and weight at surgery on predicting oxygen support group.
It is an indication that a successful balance has been achieved if the pseudo $R^2$ is reduced and becomes close to 0. There is no rule of thumb for how much of a reduction of pseudo $R^2$ is a reasonable reduction. We wish to see any reduction and for the pseudo $R^2$ in the matched data to be close to 0.

The quality of the PSM can also be assessed graphically. We wish for the postmatching distributions of the baseline covariates to be similar, not only for the means or proportions to be close. For continuous baseline covariates, one can construct box plots by treatment group to look at the distribution of this variable in the matched dataset. In our NSQIP example, we have 1 continuous baseline covariate: weight at surgery (kg). We see in Figure 2 that the distribution of weight at surgery is similar in the oxygen support group and the no oxygen support group. Therefore, PSM has successfully incorporated information about weight at surgery and balanced the treatment groups with respect to this continuous variable.

It is best to look at the absolute standardized mean differences for categorical baseline factors after matching, but to visualize the balance achieved, we can construct bar graphs of the proportions for each variable by treatment group before and after matching as shown in Figure 2.

A useful graph for evaluating the quality of PSM is a plot showing the absolute standardized mean difference before and after matching for all variables in the propensity score model (Figure 3). We can see in Figure 3 that the PSM in our example was successful because all absolute standardized mean differences were >0.10 before matching, and after matching, they are all <0.10.

Balance diagnostics allow the investigator to assess the quality of PSM. The absolute standardized mean difference and reduction in pseudo $R^2$ provide analytic summaries of quality of matching for all (continuous and categorical) baseline factors. Looking at $P$ values alone is not advisable because they are highly driven by sample size. Box plots

<table>
<thead>
<tr>
<th>Table 2. Propensity-Matched Baseline Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Covariate</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Neonate (age &lt;30 d)</td>
</tr>
<tr>
<td>Inotropic support</td>
</tr>
<tr>
<td>Ventilator dependence</td>
</tr>
<tr>
<td>Previous cardiac surgery</td>
</tr>
<tr>
<td>Weight at surgery (kg)</td>
</tr>
</tbody>
</table>

Abbreviation: $d$, absolute standardized mean difference.

![Figure 2](image)

Figure 2. Before (A) and after (B) propensity score matching. The graphical assessment shows postmatching balance on the baseline variables.
and bar plots can display the pre- and postmatching balance of baseline factors between the 2 treatment groups. Once balance diagnostics confirm good quality matching was done, then we can proceed to the final data analysis.

However, good balance is not guaranteed to be observed after PSM. The goal of PSM is to achieve baseline covariate balance between the treatment groups; however, the investigator is only able to balance on baseline factors that are measured. There may be unmeasured confounders, which is an issue with all datasets. If poor balance is observed, the researcher may want to revise their propensity score model, for example, by including interactions or high-order terms and then checking the resulting covariate balance. Another option available to the researcher is to make modifications to the matching algorithm. This can include changing the caliper width or utilizing matching with replacement if matching without replacement was used. The resulting covariate balance can be checked after each modification. Continued evidence of poor balance may suggest that there is insufficient common support and the population being considered may not be appropriate for the hypothesis.

Step 5: Analyze the Propensity-Matched Cohort

Finally, it is time to analyze the data and evaluate a treatment effect with the reduction of confounding bias due to PSM. To analyze the data after PSM, the matched pairs must be taken into account.

For a binary outcome (like we have in our example with 30-day mortality), it is appropriate to use a conditional logistic regression model to assess the effect of the exposure. Conditional logistic regression is needed to take into account the matched pairs (or matched sets) created during PSM. It can be implemented by specifying the “group” as being the matched pair (or matched set) from PSM. For continuous outcomes, to assess the treatment effect on the outcome, we advise using generalized estimating equations with the identity link function and Gaussian family, where the “subject ID” is treated as the matched pair variable. Generalized estimating equation is flexible enough to also be applied under logistic regression for binary outcomes and Poisson regression for count outcomes. Other analytic possibilities that appropriately utilize the propensity-matched data include a paired t test for continuous outcome and McNemar test for a binary outcome. For the analysis of survival or time-to-event outcomes, Austin shows that propensity score methods can be used.

After PSM, from a conditional logistic regression model taking into account propensity-matched pair, we find that oxygen support is a statistically significant predictor of 30-day mortality (odds ratio [OR] = 1.79, \(P = .012\)) (Table 3). Use of traditional multivariable logistic regression in the unmatched data leads to a larger but similar OR (2.77) and a smaller \(P\) value (\(P < .001\)). It is clear that adjustment for measured baseline factors using multivariable logistic regression controls for confounding as does PSM, because a simple logistic regression without any adjustment for baseline factors produces the unadjusted OR of 29.88. Our results using PSM are unchanged regarding statistical significance and direction of the association as compared to using the traditional multivariable regression approach. Because the 2015 Pediatric NSQIP collects data from participating institutions across the country, the findings of studies using its data are generalizable to similar institutions.

DISCUSSION

When observational (nonrandomized) data are being analyzed, there is no way to transform them into RCT data. PSM does not produce a gold standard level of evidence like the RCT, but it is a highly effective methodology to balance treatment groups and leads to an intuitive analysis. There are certain strengths...
and limitations of the PSM approach. PSM balances treatment groups with respect to confounders, making for a more objective analysis. Of course, PSM only allows for adjustment for measured confounders, and this limitation is shared with all multivariable adjustment methods. The propensity score incorporates information on multiple measured baseline factors on which to balance treatment groups, which cannot be done easily when matching by hand. Similar to adjusting for confounders using multiple regression analysis, PSM requires the availability of data on these variables. Researchers need to consider their specific research scenario and data when considering use of PSM. PSM is more desirable when 1-to-1 or k-to-1 matching can be done without discarding too many data points. Note that we did discard many data points due to a “rare exposure” of oxygen support in our NSQIP example to demonstrate 1-to-1 matching, but many-to-1 matching could have been used to keep a larger sample size.

Our piece is an educational primer or general framework; however, there are other aspects of PSM that we did not cover in this overview. These limitations must be acknowledged. We did not cover sample size and power considerations as our focus was developing the 5-step approach for implementing the propensity scores. Sample size and power should be considered and discussed with a biostatistician for all studies. Additionally, our primer does not discuss the issue of adjustment for multiple testing that may arise in certain research studies.

PSM is not a magic bullet for all observational (nonrandomized) data. Although PSM is more appealing when the outcome of interest is rare, multivariable regression adjustment methods are very useful in other scenarios. PSM can add unneeded steps to an analysis that could be done without it. However, PSM is an important procedure to obtain balance of observed relevant covariates and allow more objective group comparisons.

CONCLUSIONS

Our 5-step approach provides a useful model for anesthesiologists to implement PSM in their future research. We have demonstrated an example of 1-to-1 PSM from the 2015 Pediatric NSQIP. PSM is becoming an increasingly more popular methodology,24–27 and when implemented correctly, it increases the level of evidence of a study in turn improving the strength and generalizability of its results.

DISCLOSURES

Name: Steven J. Staffa, MS.

Contribution: This author helped write the manuscript, conduct the data analysis, and review the literature.

Name: David Zurakowski, MS, PhD.

Contribution: This author helped write the manuscript, conduct the data analysis, and review the literature.

This manuscript was handled by: Thomas R. Vetter, MD, MPH.

Table 3. Comparison of Results With and Without PSM

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Odds Ratio</th>
<th>95% CI</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without PSM unadjusted</td>
<td>29.88</td>
<td>(23.79–37.52)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Without PSM adjusted</td>
<td>2.77</td>
<td>(1.95–3.94)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>With PSM</td>
<td>1.79</td>
<td>(1.14–2.82)</td>
<td>.012</td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; PSM, propensity score matching.

REFERENCES

