



OPEN Methodological guidance on implementing propensity score matching in observational studies of kidney transplantation

Andrea Gomez-Montero¹, Andrea Garcia-Lopez¹✉, Santiago Cabas¹,
Adrián Alfonso Nieves-Rico² & Fernando Giron-Luque^{1,3}

Observational studies in kidney transplantation often face confounding bias due to the absence of randomization, which can compromise validity and limit generalizability. Propensity score matching (PSM) helps mitigate this bias by mimicking random assignment. This guide outlines the implementation of PSM in kidney transplant research, focusing on methodology, practical considerations, common pitfalls, and reporting standards. We summarize key steps, including selecting covariates related to both treatment and the probability of receiving a transplant, estimating propensity scores, applying appropriate matching techniques, assessing balance, and conducting sensitivity analyses to test robustness. Practical considerations include ensuring sufficient overlap in propensity scores and balancing sample size with matching quality. Common challenges involve omitting relevant covariates, inadequate overlap, suboptimal matching, and loss of statistical power due to reduced sample size. By adhering to rigorous methodological practices and transparent reporting, researchers can improve the credibility and impact of their findings. When carefully implemented, PSM can substantially reduce confounding bias, enhance causal inference, and ultimately support better decision-making in kidney transplantation research. The example provided is illustrative only and does not replace a formal time-to-event analysis accounting for competing risks.

Keywords Kidney transplantation, Propensity score matching, Observational studies, Confounding factors, Statistical analysis

Abbreviations

CBPS	Covariate Balancing Propensity Scores
CRAN	Comprehensive R Archive Network
DAG	Directed Acyclic Graph
GEE	Generalized Estimating Equations
IPTW	Inverse Probability of Treatment Weighting
MSMs	Marginal Structural Models
OR	Odds Ratio
PSM	Propensity Score Matching
SE	Standard Error
SMD	Standardized Mean Difference

Observational studies are inherently vulnerable to various types of bias due to the absence of randomization. Therefore, effectively controlling bias in these studies is essential to enhance the validity of the results, ensure generalizability, and inform decision-making. It is crucial to follow established guidelines, such as the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines, to ensure high-quality reporting in observational studies¹.

Mitigating bias presents a significant challenge for researchers, and several strategies have been proposed to address confounding variables, including the use of matching estimators². Propensity score matching (PSM) is a

¹Department of Research, Colombiana de Trasplantes, Bogotá, Colombia. ²Faculty of Medicine, Universidad de los Andes, Bogotá, Colombia. ³Department of Transplant Surgery, Colombiana de Trasplantes, Bogotá, Colombia. ✉email: aegarcia@colombianadetrasplantes.com

statistical method to reduce confounding bias in observational studies first introduced in 1983 by Rosebaum and Rubin³. This method has gained wide acceptance for addressing confounding in observational studies. When properly implemented, it offers value in enhancing covariate balance between treatment groups and in approximating the conditions of a randomized controlled trial, thereby strengthening causal inference⁴. This strategy combines matching with the propensity score to approximate a quasi-randomized environment⁵. Also, propensity score methods can be used not only to reduce confounding but also to define or approximate specific target populations, allowing researchers to emulate the conditions of different randomized controlled trials within observational data⁶. This methodology has gained widespread application in kidney transplantation studies, where researchers must consider multiple patient covariates⁷.

This method offers a robust analytical framework that minimizes confounding factors, ultimately leading to stronger evidence-based practice^{8–10}. PSM is particularly advantageous in scenarios where randomization is impractical or impossible, making it a powerful tool for observational studies in kidney transplantation¹¹. The purpose of this review is to explain PSM, explore its practical applications in the field of kidney transplantation, and provide a practical example.

Applications of PSM in kidney transplant research

In kidney transplant studies, common confounding variables that can affect outcomes include age, sex, dialysis duration, and comorbidities^{12–15}. The use of PSM can be particularly useful in various scenarios within this field. For example, PSM enables the comparison of outcomes from different therapies in transplanted patients, providing a robust methodology to control for confounding variables and derive more accurate conclusions¹⁶. Moreover, this technique is valuable for evaluating the impact of pretransplant conditions on posttransplant outcomes, helping to identify which preexisting factors may influence patient recovery and survival^{17,18}. Additionally, PSM is used to identify factors associated with different outcomes in transplanted patients, facilitating an understanding of variables that can affect transplant effectiveness and posttransplant quality of life^{7,19–21}. These applications underscore the versatility of PSM in kidney transplant research, allowing researchers to explore a wide range of questions related to treatment efficacy and patient outcomes¹¹.

Methodology: implementing PSM

Covariate selection

The starting point in any PSM approach is selecting the right covariates. These should include variables that are related both to the likelihood of receiving the treatment and to the outcome, since failing to control for such confounders can bias the results^{22,23}. To strengthen this step, it is advisable to draw on existing literature and expert knowledge, which can help ensure that all relevant factors are taken into account¹¹.

However, not all variables should be included. It's crucial to distinguish between confounders, which must be adjusted for, and mediators, which lie along the causal pathway and may distort the estimated treatment effect if included. To navigate these decisions systematically, researchers can use tools such as Directed Acyclic Graphs (DAGs), which offer a clear visual representation of hypothesized relationships and help guide covariate selection in a defensible way²⁴.

Estimating propensity scores

Once the covariates are defined, the next step is to estimate each individual's probability of receiving the treatment. This is commonly done using logistic regression, which models the likelihood of treatment assignment based on the selected covariates. While logistic regression remains the standard, alternatives such as probit models or machine learning techniques (e.g., random forests or gradient boosting) may be better suited to complex scenarios or multi-arm settings. In any case, including interaction terms or nonlinear functions often helps improve model fit and enhances balance across groups. Notably, since the goal is not prediction but covariate balancing, overfitting is not considered a major issue here²³.

More recently, approaches like Covariate Balancing Propensity Scores (CBPS) have gained traction. Unlike traditional methods that separate prediction from balance, CBPS directly integrates both objectives, simultaneously optimizing model fit and covariate balance. This makes it particularly useful when there is concern about model misspecification or limited sample size²⁵.

In line with these balance-focused approaches, recent work by Li et al. has demonstrated that enforcing balance directly through estimation methods—such as CBPS and entropy balancing—can lead to more accurate and less biased treatment effect estimates, particularly when traditional models are misspecified. Their simulation studies showed improvements in bias, variance, and mean squared error, supporting the utility of balance-focused methods²⁶. Similarly, Huan et al. proposed a flexible weighting strategy that chooses between global and local scores based on balance quality. This method has proven effective for multi-site survival analyses and performs comparably to individual-level data pooling²⁷.

Matching

After estimating propensity scores, researchers must choose how to match treated and untreated individuals. The decision should be informed by study goals, sample size, and the distribution of the propensity scores. In the practical exercise, we used 1:1 nearest neighbor matching, which is straightforward and maintains interpretability. Nonetheless, other strategies may be more efficient. For instance, one-to-many (K:1) matching allows each treated subject to be matched with multiple controls, improving precision and reducing standard errors. Evidence from simulations and applied studies suggests that variable-ratio matching often results in lower mean squared error with only a small trade-off in bias²⁸.

Each method brings its own strengths and limitations. Nearest neighbor matching is easy to apply but may yield poor matches in the absence of strong overlap. Caliper matching mitigates this by setting a maximum

acceptable difference in propensity scores between matched individuals, though it may reduce the number of matches. Meanwhile, optimal matching seeks to minimize the overall distance across all matched pairs but can be more computationally intensive. In all cases, it is critical to evaluate post-matching balance and justify the selected strategy in light of the dataset's characteristics^{11,29}.

Assessing balance

Once matching is completed, it is essential to assess whether the groups are now comparable in terms of baseline covariates. The most widely recommended metric is the standardized mean difference (SMD), which is not affected by sample size and provides a directional indication of imbalance. An SMD below 0.1 in absolute value is generally considered acceptable³⁰. Complementary to numerical metrics, visual diagnostics—such as histograms or box plots—can help detect residual imbalances and verify that the matching process worked as intended²². This step is fundamental for establishing the internal validity of the treatment effect estimate¹¹.

Sensitivity analysis

Even when good balance is achieved, it is important to assess whether the results are robust to reasonable changes in the matching procedure. Sensitivity analyses may involve varying the caliper width (e.g., 0.05, 0.1, 0.15) or trying alternative matching methods such as optimal or full matching. By comparing the treatment effect estimates across different configurations, researchers can determine whether findings are consistent or dependent on specific modeling choices²³.

Estimating treatment effects

With covariates balanced and matching complete, the treatment effect can be estimated. This is typically done by regressing the outcome on the treatment variable in the matched sample. Because the matching process already adjusted for confounding, it is not necessary to reintroduce covariates into the regression model. This approach ensures that each treated subject is compared only to a comparable control and allows for clean interpretation of the treatment effect¹².

Practical considerations and common pitfalls

While PSM offers a robust framework for addressing confounding, several practical aspects must be considered to ensure its validity.

First, although selecting appropriate covariates has been addressed earlier, it is worth emphasizing that omitting key confounders or including irrelevant variables can still compromise results—either by introducing residual bias or by increasing variance, particularly in small samples^{11,29}.

Second, adequate overlap in propensity scores between treatment groups is essential. A lack of common support can hinder valid comparisons and may require trimming unmatched individuals, potentially reducing statistical power²². Visualizing score distributions can help diagnose this issue early in the process.

Third, match quality should not be assumed. Post-matching balance diagnostics, such as SMD, remain essential to confirm that the procedure was successful and that the matched groups are truly comparable^{22,25}.

Finally, it is important to recognize that matching inherently reduces the available sample size. In settings with limited data, this may threaten power and precision. When excessive data loss occurs, alternative approaches—such as inverse probability weighting or covariate adjustment—may offer a more efficient solution²³.

Practical example of PSM in kidney transplant research

In this section, we conduct a practical exercise in which PSM is applied to a publicly available dataset focused on kidney transplants. In this example, diabetes status is treated as the exposure variable, and the propensity score is estimated as the probability of having diabetes given relevant covariates (such as age, dialysis time, sex, blood type, and subregion). Matching is performed to balance these covariates between patients with and without diabetes, enabling a comparison of transplant outcomes between these groups.

This example has been intentionally simplified into a binary outcome analysis (transplanted or not) to illustrate the steps involved in implementing PSM. It does not account for the probability of transplantation as a time-to-event process with competing risks (e.g., death or permanent waitlist removal), and continuous variables are dichotomized while certain confounders are excluded to enhance didactic clarity.

The objective is to perform a PSM analysis to examine the impact of a diabetes diagnosis on the likelihood of receiving a transplant among patients on a waitlist. This practical example is conducted in RStudio via R version 4.3.3³¹. By following this example, we aim to demonstrate the step-by-step implementation of PSM, emphasizing the importance of meticulous covariate selection and methodological rigor in observational studies. This exercise provides a detailed and systematic approach to applying PSM, ensuring the robustness and validity of the findings. Furthermore, this example serves as a foundational guide for the application of this methodology, thereby enhancing the overall quality and reliability of research in this field.

Dataset description

The dataset used here is sourced from Kaggle, titled “Waitlist Kidney Brazil”³². It includes patient demographics, clinical factors, and treatment details such as age, time on dialysis, race, sex, underlying disease, diabetes status, blood type, subregion, and transplant outcomes.

Running the analysis directly in R

To execute the full PSM analysis directly, simply copy and paste the script shown in Fig. 1 into RStudio. This script automatically downloads the complete R Markdown file containing the analysis and the kidney transplant waitlist dataset from GitHub and runs the workflow in a fully reproducible way. Both the R Markdown file and

```

# Install and load necessary libraries
pkgs <- c("downloader", "rmarkdown", "MatchIt", "dplyr", "readr", "cobalt", "ggplot2")
for (pkg in pkgs) if (!requireNamespace(pkg, quietly = TRUE)) install.packages(pkg)
lapply(pkgs, library, character.only = TRUE)

# Set chunk options to suppress prompts and messages
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE, ask = FALSE)

# Download R Markdown script from GitHub
markdown_url <-
  "https://raw.githubusercontent.com/CoITrasplantes/PSM/refs/heads/main/psm.Rmd"
markdown_dest <- tempfile(fileext = ".Rmd")
download.file(markdown_url, destfile = markdown_dest)

# Render the R Markdown file to HTML
html_output <- tempfile(fileext = ".html")
rmarkdown::render(markdown_dest, output_format = "html_document", output_file = html_output)

# Wait for the rendering to complete before opening
Sys.sleep(2)

# Ensure that all graphs are generated without user interaction
par(ask = FALSE)

# Display the HTML output in the default browser
browseURL(html_output)

```

Fig. 1. One-click, fully reproducible PSM analysis script. Installs required packages, downloads the R Markdown from GitHub, renders to HTML, and opens the report; figures are generated non-interactively.

dataset, while hosted on GitHub, are also permanently archived and accessible via Zenodo³³. The output appears as an HTML report within your R session, without needing to save files manually. This HTML report is also included as Supplementary File 1.

Step-by-Step implementation of the PSM analysis

Before starting the main analysis, an automatic script prepares all necessary elements to ensure reproducibility. This script installs and loads the required R packages (MatchIt, dplyr, readr, cobalt, ggplot2, and gtsummary) from a stable CRAN mirror, downloads the kidney transplant waitlist dataset from GitHub, standardizes column names, corrects character encoding, and recodes the outcome variable (Transplant_Y_N) as a binary indicator (1 = Yes, 0 = No). It also recategorizes key variables such as age and dialysis duration into clinically meaningful groups and removes incomplete cases, generating a clean dataset (data_filtered) ready for matching. Although dichotomizing continuous variables may introduce limitations, such as masking residual imbalance across the full range of values, this step was implemented strictly for didactic purposes to streamline the example and facilitate replication. It is not intended as a methodological recommendation. The full setup code is included in the R Markdown file, although it is not displayed in the HTML output for clarity.

Step 1: sample size before matching

Initially, the analytic dataset size is assessed using the nrow() function to count complete cases. The baseline sample includes 46,817 observations, providing a reference point for evaluating the impact of subsequent matching on sample size and data retention. In addition, Table 1 summarizes baseline characteristics overall and by diabetes status.

Step 2: unadjusted association between exposure and outcome

Prior to matching, a logistic regression (glm) estimates the crude association between the exposure (e.g., diabetes) and outcome (receiving a transplant). In this unadjusted model, the exposure coefficient reflects the log-odds of the outcome, providing an initial baseline. A statistically significant result here (log-odds = 0.41, $p < 0.001$) indicates a strong, yet potentially confounded relationship between diabetes and transplant probability.

Step 3: propensity score matching procedure

To minimize confounding and enhance the comparability between individuals with and without diabetes, we applied PSM using the matchit() function from the MatchIt package. The approach selected was nearest neighbor matching, supplemented with exact matching on key categorical variables—race, blood type, and subregion—to ensure that pairs were only formed within identical strata. Additionally, a caliper width of 0.2 standard deviations

Characteristic	Overall N = 46,817	No Diabetes N = 9,797	Diabetes N = 37,020
Age at list registration (cat)			
< 50	22,160 (47%)	2,226 (23%)	19,934 (54%)
>=50	24,657 (53%)	7,571 (77%)	17,086 (46%)
Time on dialysis (cat)			
< 24 months	35,125 (75%)	7,561 (77%)	27,564 (74%)
>=24 months	11,692 (25%)	2,236 (23%)	9,456 (26%)
Race			
Amarela	906 (1.9%)	354 (3.6%)	552 (1.5%)
Branca	31,333 (67%)	6,642 (68%)	24,691 (67%)
Negra	5,661 (12%)	1,162 (12%)	4,499 (12%)
Parda	8,917 (19%)	1,639 (17%)	7,278 (20%)
Sex			
Female	18,843 (40%)	3,285 (34%)	15,558 (42%)
Male	27,974 (60%)	6,512 (66%)	21,462 (58%)
Blood type			
A	16,465 (35%)	3,594 (37%)	12,871 (35%)
AB	1,683 (3.6%)	390 (4.0%)	1,293 (3.5%)
B	5,671 (12%)	1,251 (13%)	4,420 (12%)
O	22,998 (49%)	4,562 (47%)	18,436 (50%)
Subregion			
FUNDERP	7,715 (16%)	1,587 (16%)	6,128 (17%)
HCFMUSP	13,230 (28%)	2,779 (28%)	10,451 (28%)
UNICAMP	3,362 (7.2%)	568 (5.8%)	2,794 (7.5%)
UNIFESP	22,510 (48%)	4,863 (50%)	17,647 (48%)
Transplant	13,213 (28%)	2,153 (22%)	11,060 (30%)

Table 1. Baseline characteristics before matching, by diabetes status. Values are n (%) by column.

was imposed, restricting matches to those with closely aligned propensity scores. This rigorous combination of methods maximizes the quality and comparability of the matched pairs, substantially reducing potential systematic bias. Inevitably, some cases remain unmatched due to these strict criteria; however, this is an expected and acceptable trade-off, as it prioritizes analytic rigor and matching quality over sheer sample size. Because sex showed residual imbalance in the baseline match (see Steps 5–6), we also examined tighter calipers (0.05–0.15) and an exact-on-sex specification in sensitivity analyses (Steps 8–9).

Step 4: sample size after matching

Following the matching process, we used the `nrow()` function to determine the number of observations retained. The final matched cohort consisted of 19,504 individuals, reflecting a considered trade-off between preserving statistical power and enhancing covariate balance through stringent matching criteria. Such a reduction in sample size is expected in rigorous propensity score analyses, where the primary focus is on ensuring high-quality matches to reduce potential bias.

Step 5: assessing covariate balance numerically

To determine whether the matching procedure effectively balanced the main covariates, we calculated SMD for each variable before and after matching using the `bal.tab()` function from the `cobalt` package. The most relevant SMD results are summarized in Table 2. After matching, nearly all covariates exhibited SMD values close to zero, reflecting successful reduction of imbalance between the diabetes and non-diabetes groups. The only exception was a minor residual imbalance in the “Male” variable. Specifically, the post-match SMD for sex was 0.156 (above the 0.10 threshold), which motivated the sensitivity checks reported in Steps 8–9.

Step 6: visualizing covariate balance with a love plot

To complement the numeric summary, we generated a Love plot using the `love.plot()` function from the `cobalt` package, which offers a visual representation of the SMDs reported in Table 2 (Fig. 2). In this plot, each covariate appears on the y-axis and SMDs are shown on the x-axis. Red dots represent SMD before matching, and blue dots indicate SMD after matching. The vertical dashed line at 0.1 denotes the threshold for acceptable covariate balance. As shown in Fig. 2, after matching, SMDs for all main covariates—except for a slight imbalance in the “Male” variable—fall well below the threshold, visually confirming the success of the matching procedure. Consistent with Step 5, sex remains slightly above the 0.10 line in the baseline match.

Covariate	SMD Before Matching	SMD After Matching
Age \geq 50	-0.311	0.000
Dialysis \geq 24 Months	0.027	0.041
Male	-0.085	-0.156
Blood Type O	0.032	0.000
Blood Type A	-0.019	0.000
Blood Type B	-0.008	0.000
Blood Type AB	-0.005	0.000
Race: Parada	0.029	0.000
Race: White	-0.011	0.000
Race: Black	0.003	0.000
Race: Amarela	-0.021	0.000
Subregion: UNICAMP	0.017	0.000
Subregion: UNIFESP	-0.020	0.000
Subregion: FUNDERP	0.004	0.000
Subregion: HCFMUSP	-0.001	0.000

Table 2. Standardized mean differences before and after propensity score matching for main covariates. SMD = standardized mean difference.

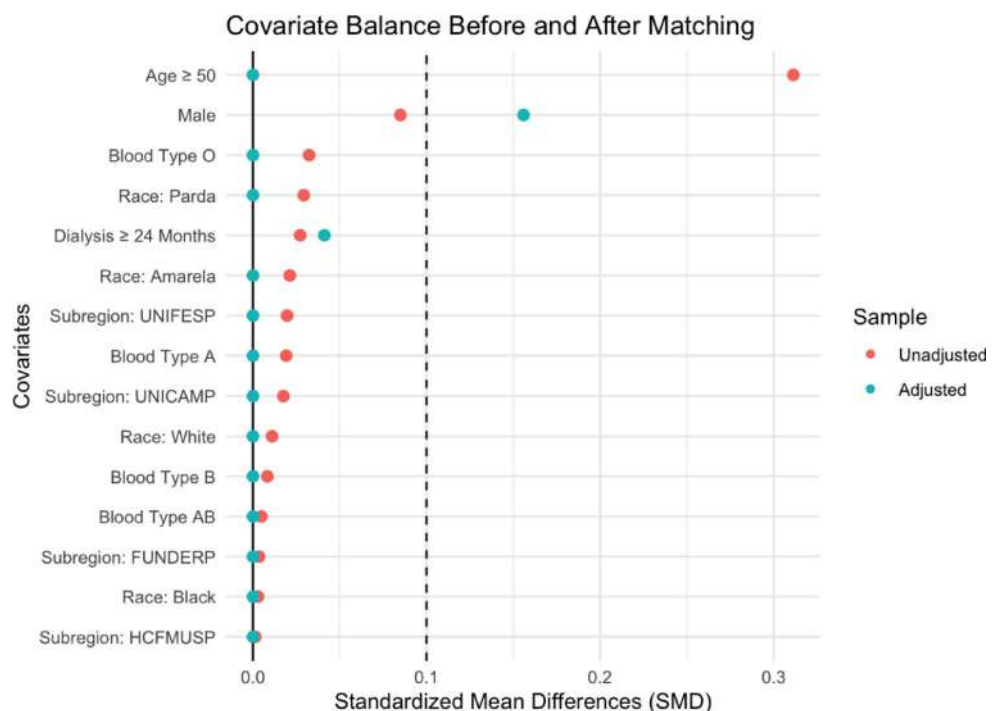


Fig. 2. Love plot of standardized mean differences for key covariates before (red) and after (blue) matching. The vertical dashed line at 0.10 marks the balance threshold.

Step 7: assessing propensity score overlap and distribution

To evaluate whether the PSM procedure achieved adequate overlap and comparability between groups, we examined the distribution of propensity scores using two complementary visual diagnostics generated with the `plot()` function from the `MatchIt` package. First, the jitter plot (Fig. 3) displays the distribution of individual propensity scores among treated (diabetes) and control (non-diabetes) units before and after matching. This plot shows that, after matching, most treated and control observations lie within a common range of propensity scores, supporting the validity of comparisons within the matched sample. Second, the set of histograms (Fig. 4) shows the proportion of subjects at each propensity score interval for both groups, before and after matching. The close alignment of these distributions in the matched samples further demonstrates that the matching process produced analytic groups with similar baseline characteristics. Collectively, these visualizations confirm that the

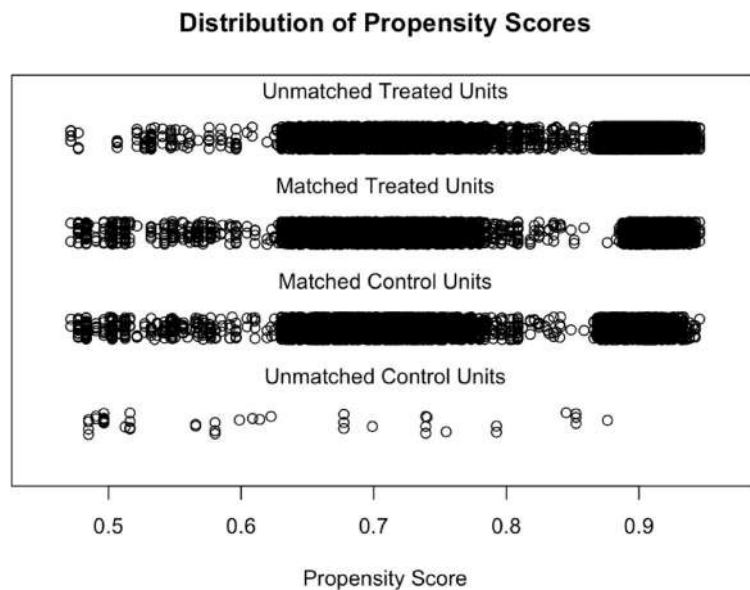


Fig. 3. Jitter plot of propensity-score distributions for treated and control groups before and after matching.

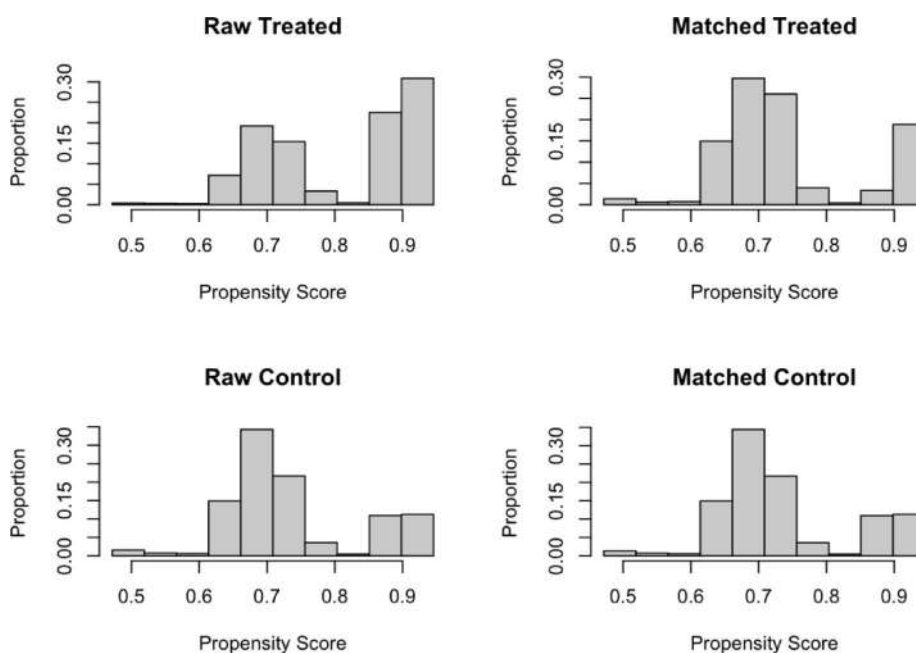


Fig. 4. Histograms of propensity scores for treated and control groups before (left) and after (right) matching. Greater alignment post-matching indicates improved comparability.

matched dataset achieves substantial overlap in propensity scores, which is critical for unbiased estimation of treatment effects in subsequent analyses.

Step 8: sensitivity analysis by caliper width

To test the robustness of findings, a sensitivity analysis is conducted with different caliper widths (0.05, 0.10, 0.15) in the matching process. For each caliper, logistic regression estimates the diabetes effect on transplant likelihood. Results are consistently positive and statistically significant (caliper 0.05: 0.17; caliper 0.10: 0.25; caliper 0.15: 0.27; all $p < 0.001$), indicating stable results despite variations in matching strictness. This stability reinforces confidence in the robustness of the primary findings. Importantly, sex balance improved to $SMD < 0.10$ at calipers 0.05 (-0.067) and 0.10 (-0.078), but not at 0.15 (-0.142).

Step 9: sensitivity analysis by matching method

To check robustness, three nearest-neighbor variants were run at a caliper of 0.10: without replacement, with replacement, and exact-on-sex. In every case, balance on sex improved to an absolute SMD < 0.10. Notably, the exact-on-sex variant achieved perfect balance on sex (SMD = 0.000) while keeping a matched sample size (N) similar to the no-replacement design, and it produced a diabetes coefficient of ≈ 0.216 (standard error, SE, ≈ 0.034 ; $p < 1 \times 10^{-9}$).

Step 10: adjusted association after matching

Finally, logistic regression analysis (glm) was fitted to the matched dataset to estimate the adjusted association between diabetes and the probability of transplant. For inference, dependence within matched sets was addressed using cluster-robust (“sandwich”) standard errors (SEs) at the matched-set level (implemented with the sandwich and lmtest R packages). This is comparable to generalized estimating equations (GEE) with an exchangeable working correlation and is recommended for matched observational data^{34,35}. Under the final specification (nearest-neighbor matching, caliper 0.10, exact-on-sex), the diabetes coefficient was ≈ 0.216 with a cluster-robust SE ≈ 0.033 (z-statistic, $z, \approx 6.49$; p-value, $p, \approx 8.45 \times 10^{-11}$), which corresponds to an odds ratio (OR) of ~ 1.24 . Therefore, compared with the unadjusted analysis, this smaller but still significant effect supports effective confounding control through PSM. If any covariate had remained ≥ 0.10 in SMD after matching, a doubly robust outcome model adjusting for that covariate would have been added, still using cluster-robust SEs.

Reporting guidelines for PSM studies**Essential elements to report***Covariates included and rationale*

It is essential to list all covariates included in the propensity score model and provide a rationale for their inclusion. Covariates should be selected on the basis of theoretical considerations, prior empirical research, or both, ensuring that they are related to both the treatment and the outcome. This comprehensive selection process helps to adequately control for confounding factors^{36,37}.

Matching algorithm and parameters

Describe the matching algorithm used, such as nearest neighbor, caliper matching, or Mahalanobis distance matching. Additionally, specify any parameter set, such as the caliper width or matching ratio. For example, nearest neighbor matching might use a 1:1 or 1:2 matching ratio, and caliper matching might specify a caliper width of 0.1 standard deviations of the logit of the propensity score^{29,37,38}.

Sample sizes before and after matching

The sample sizes of the treatment and control groups are reported both before and after matching. This information is crucial for understanding the extent of data reduction due to matching and the potential impact on statistical power^{29,37}.

Methods and results of balance assessment

The methods used to assess the balance of covariates between the treatment and control groups after matching were outlined. This typically involves SMDs, variance ratios, or graphical methods such as histograms and jitter plots. Reporting the results of these assessments demonstrates the effectiveness of the matching process in achieving covariate balance^{29,36–38}.

Sensitivity analysis

A sensitivity analysis is conducted to check the robustness of the matching results. This includes varying the caliper width and trying different matching methods to ensure that the conclusions are consistent across different specifications. For example, varying the caliper width and using different matching methods can provide insights into the stability of the estimated treatment effects. The results of these analyses show that the findings are robust to different matching parameters and methods.

Statistical analysis of treatment effects

Detail the statistical methods used to estimate treatment effects after matching. This might include regression adjustment, difference-in-differences analysis, or instrumental variable approaches. The results of these analyses, including estimates of treatment effects, confidence intervals, and significance levels³⁸, are presented.

Best practices for transparency and reproducibility*Detailed methodological description*

A comprehensive description of all the steps in the PSM process, including data preprocessing, propensity score estimation, matching procedures, and balance assessment, is provided. This ensures that other researchers can fully understand and replicate the methodology.

Code and data sharing

Share the code used for propensity score estimation and matching, preferably with annotations to explain each step. Where possible, make the dataset or a synthetic version available to allow others to replicate the analysis. This practice enhances transparency and facilitates validation by other researchers.

Documentation of assumptions

Clearly, document all assumptions made during the analysis, such as the ignorability assumption and discuss their potential impact on the study's conclusions. This transparency allows for a better understanding of the limitations and strengths of the study.

Thorough reporting of results

Include detailed tables and figures that show the balance of covariates before and after matching, as well as the estimated treatment effects. Provide supplementary materials if necessary to keep the main text concise. This thorough reporting ensures that the results are clear and interpretable.

Ethical considerations

We discuss any ethical considerations related to the data and analysis, including issues of consent, confidentiality, and potential biases introduced by the matching process. Addressing these issues is crucial for maintaining ethical standards and the integrity of research. It is also recommended to follow established guidelines, such as the STROBE guidelines, to further ensure clarity, transparency, and methodological rigor in reporting observational studies¹.

By following these guidelines, researchers can enhance the transparency, reproducibility, and credibility of PSM studies. Ultimately, this contributes to the robustness of causal inference in observational research.

Complementary causal inference approaches

In addition to PSM, it is advisable to use complementary methods to address limitations inherent to matching alone. One valuable tool in this regard is the use of DAGs, which help make causal assumptions explicit and guide the selection of appropriate covariates for adjustment, reducing the risk of including mediators or colliders that could bias estimates^{39,40}. Furthermore, an approach such as inverse probability of treatment weighting (IPTW) is an alternative to matching that can help preserve sample sizes and hence power; in particular, marginal structural models (MSMs) are well-suited when exposures and confounders vary over time, as is often the case in kidney transplant research^{41,42}. For example, recent studies have demonstrated how IPTW can emulate a target trial comparing transplantation with long-term dialysis, providing more robust effect estimates than matching alone⁴³.

Additionally, doubly robust estimators, which combine outcome regression with propensity-based weighting, offer an extra safeguard by providing valid causal estimates if either the propensity score model or the outcome model is correctly specified^{44,45}. This dual protection makes doubly robust methods a valuable complement to both PSM and IPTW, particularly in observational contexts where some model misspecification is likely. By combining DAGs, IPTW, MSMs, and doubly robust estimation alongside PSM, researchers can strengthen the validity, transparency, and interpretability of causal inferences in kidney transplantation studies.

Conclusion

In conclusion, PSM plays an important role in kidney transplant research by providing a structured approach to control for confounding and strengthen the validity of findings from observational data. It improves comparability between treated and control groups on key baseline variables, allowing researchers to estimate treatment effects more reliably. Nonetheless, its effectiveness depends on appropriate covariate selection, consistent data quality, and careful consideration of potential sample size reductions, which may influence statistical power. To address these challenges, combining PSM with complementary methods such as DAGs, IPTW, and MSMs can help account for complex causal structures and time-varying confounding. Maintaining rigorous methodology, clearly reporting each analytical step, sharing code and assumptions, and conducting sensitivity analyses are essential for ensuring transparent and reproducible results. By applying these principles and adhering to established reporting standards such as STROBE, researchers can contribute to more robust and informative evidence that supports decision-making in kidney transplantation.

Data availability

The dataset analyzed in this article is available on [Kaggle](https://www.kaggle.com/datasets/gustavomodelli/waitlist-kidney-brazil). Additionally, the dataset and the R script used for the analysis can be found in the [Colombiana de Trasplantes' repository](https://github.com/ColTrasplantes/PSM) on GitHub and can be downloaded via [Zenodo](https://doi.org/10.5281/zenodo.17023205).

Received: 25 November 2024; Accepted: 3 December 2025

Published online: 13 December 2025

References

1. Cuschieri, S. *The STROBE guidelines*. Vol. 13, *Saudi Journal of Anaesthesia* S31–S34 (Wolters Kluwer Medknow, 2019). https://doi.org/10.4103/sja.SJA_543_18
2. Corder, N. & Yang, S. Utilizing stratified generalized propensity score matching to approximate blocked randomized designs with multiple treatment levels. *J. Biopharm. Stat.* **32** (3), 373–399. <https://doi.org/10.1080/10543406.2022.2065507> (2022).
3. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *70*. <https://doi.org/10.1093/biomet/70> (1983). <https://academic.oup.com/biomet/article/70/1/41/240879>. Available from
4. Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **46** (3), 399–424. <https://doi.org/10.1080/00273171.2011.568786> (2011).
5. Wang, J. To use or not to use propensity score matching? *Pharm. Stat.* **20** (1), 15–24. <https://doi.org/10.1002/pst.2051> (2021).

6. Thomas, L., Li, F. & Pencina, M. *Using Propensity Score Methods To Create Target Populations in Observational Clinical Research*. Vol. 323, *JAMA - Journal of the American Medical Association* 466–467 (American Medical Association, 2020). <https://doi.org/10.1001/jama.2019.21558>
7. Okumi, M. et al. Preemptive kidney transplantation: a propensity score matched cohort study. *Clin Exp Nephrol.* **21**(6), 1105–12. <https://doi.org/10.1007/s10157-016-1345-x>. Available from: <https://doi.org/10.1007/s10157-016-1345-x> (2017).
8. Kane, L. T. et al. Propensity Score Matching A Statistical Method [Internet]. (2020). <https://doi.org/10.1097/BSD.0000000000000932>. Available from: www.clinicalspinesurgery.com.
9. Nguyen, V. T. et al. Risk of bias in observational studies using routinely collected data of comparative effectiveness research: a meta-research study. *BMC Med.* **19** (1). <https://doi.org/10.1186/s12916-021-02151-w> (2021).
10. Kim, H. Y. et al. Comparison of clinical outcomes between preemptive transplant and transplant after a short period of Dialysis in Living-Donor kidney transplantation: A Propensity-Score-Based analysis. *Ann. Transpl.* **24**, 75–83. <https://doi.org/10.12659/AOT.913126> (2019).
11. Chen, J. W. et al. Best practice guidelines for propensity score methods in medical research: consideration on Theory, Implementation, and reporting. *Rev. Arthrosc. - J. Arthroscopic Relat. Surg.* **38** (2), 632–642. <https://doi.org/10.1016/j.arthro.2021.06.037> (2022).
12. Schold, J. D., Malamon, J. & Kaplan, B. Statistical Confounding in Observational Research and Center Performance Evaluations in Organ Transplantation. *Curr Transplant Rep.* **10**(4), 224–9. <https://doi.org/10.1007/s40472-023-00420-6>. Available from: <https://doi.org/10.1007/s40472-023-00420-6>.
13. Fu, R., Kim, S. J., de Oliveira, C. & Coyte, P. C. An instrumental variable approach confirms that the duration of pretransplant dialysis has a negative impact on the survival of kidney transplant recipients and quantifies the risk. *Kidney Int.* **96**(2), 450–9. Available from: <https://doi.org/10.1016/j.kint.2019.03.007>. Available from: <https://www.sciencedirect.com/science/article/pii/S0085253819303230> (2019).
14. Taber, D. J. et al. The Impact of Diabetes on Ethnic Disparities Seen in Kidney Transplantation. *Ethn Dis.* **23**(2), 238–44. Available from: <https://www.jstor-org.ezproxy.uniandes.edu.co/stable/48667841> (2013).
15. Isaacs, R. B. et al. Racial disparities in renal transplant outcomes. *American Journal of Kidney Diseases.* **34**(4), 706–12. Available from: [https://doi.org/10.1016/S0272-6386\(99\)70397-5](https://doi.org/10.1016/S0272-6386(99)70397-5). Available from: <https://www.sciencedirect.com/science/article/pii/S0272638699703975> (1999).
16. Cristelli, M. P. et al. Efficacy of Convalescent Plasma to Treat Mild to Moderate COVID-19 in Kidney Transplant Patients: A Propensity Score Matching Analysis. Vol. 106, *Transplantation*. Lippincott Williams and Wilkins; E92–4. <https://doi.org/10.1097/TP.0000000000003962> (2022).
17. Lozano-Suárez, N., García-López, A., Gómez-Montero, A. & Girón-Luque, F. Relación Entre La compatibilidad Del HLA y La pérdida Del Injerto En Trasplante renal de Donante cadavérico: Un análisis Por propensity score matching En Colombia. *Revista Colombiana De Cirugía.* **39**, 268–279. <https://doi.org/10.30944/20117582.2491> (2024).
18. Castelli, C. et al. Impact of kidney transplantation in obese candidates: a time-dependent propensity score matching study. *Nephrology Dialysis Transplantation.* **37**(9), 1768–76. <https://doi.org/10.1093/ndt/gfac152>. Available from: <https://doi.org/10.1093/ndt/gfac152> (2022).
19. Sood, M. M. et al. Risk of Major Hemorrhage after Kidney Transplantation. *Am J Nephrol.* **41**(1), 73–80. <https://doi.org/10.1159/000371902>. Available from: <https://doi.org/10.1159/000371902> (2015).
20. Han, S. H., Go, J., Park, S. C. & Yun, S. S. Long-Term Outcome of Kidney Retransplantation in Comparison With First Transplantation: A Propensity Score Matching Analysis. *Transplant Proc* [Internet]. ;51(8):2582–6. Available from: <https://doi.org/10.1016/j.transproceed.2019.03.070>. (2019). Available from: <https://www.sciencedirect.com/science/article/pii/S004113451930288X>
21. Kosoku, A. et al. Sarcopenia as a predictor of mortality in kidney transplant recipients: A 5-year prospective cohort study with propensity score matching. *International Journal of Urology* [Internet]. 2024;n/a(n/a). <https://doi.org/10.1111/iju.15539>. Available from: <https://doi.org/10.1111/iju.15539>.
22. Staffa, S. J. & Zurakowski, D. Five steps to successfully implement and evaluate propensity score matching in clinical research studies. *Anesth. Analg.* **127** (4), 1066–1073. <https://doi.org/10.1213/ANE.0000000000002787> (2018).
23. Ali, M. S. et al. Propensity score methods in health technology assessment: Principles, extended applications, and recent advances. *Front. Pharmacol.* **10** <https://doi.org/10.3389/fphar.2019.00973> (2019).
24. Arif, S. & MacNeil, M. A. Utilizing causal diagrams across quasi-experimental approaches. *Ecosphere* **13** (4). <https://doi.org/10.1029/2022.ecs2.4009> (2022).
25. Imai, K., Ratkovic, M., Stat Soc Series, J. R. & Stat Methodol, B. Covariate Balancing Propensity Score. **76**(1), 243–63. <https://doi.org/10.1111/rssb.12027>. Available from: <https://doi.org/10.1111/rssb.12027> (2014).
26. Li, Y. & Li, L. Propensity score analysis methods with balancing constraints: A Monte Carlo study. *Stat. Methods Med. Res.* **30** (4), 1119–1142. <https://doi.org/10.1177/0962280220983512> (2021).
27. Huang, C., Wei, K., Wang, C., Yu, Y. & Qin, G. Covariate balance-related propensity score weighting in estimating overall hazard ratio with distributed survival data. *BMC Med. Res. Methodol.* **23** (1). <https://doi.org/10.1186/s12874-023-02055-8> (2023).
28. Rassen, J. A. et al. One-to-many propensity score matching in cohort studies. *Pharmacoepidemiol Drug Saf.* **21** (SUPPL.2), 69–80. <https://doi.org/10.1002/pds.3263> (2012).
29. Randolph, J. J., Falbe, K. & Practical Assessment A step-by-step guide to propensity score matching in R. *Res. Evaluation* ;**19**. DOI: <https://doi.org/10.7275/n3pv-tx27> (2014).
30. Austin, P. C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **28** (25), 3083–3107. <https://doi.org/10.1002/sim.3697> (2009).
31. Team, R. C. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from: <https://www.R-project.org/> (2024).
32. Modelli, G. Waitlist Kidney Brazil. Available from: <https://www.kaggle.com/datasets/gustavomodelli/waitlist-kidney-brazil> (2023).
33. ColTrasplantes. ColTrasplantes/PSM: v3.0 . Zenodo. DOI: 10.5281/zenodo.17023205. (2025). Available from: <https://doi.org/10.5281/zenodo.17023205>
34. Austin, P. C. A comparison of variance estimators for logistic regression models estimated using generalized estimating equations (GEE) in the context of observational health services research. *Stat. Med.* **43** (29), 5548–5561. <https://doi.org/10.1002/sim.10260> (2024).
35. Austin, P. C., Kapral, M. K., Vyas, M. V., Fang, J. & Yu, A. Y. X. Using multilevel models and generalized estimating equation models to account for clustering in neurology clinical research. *Neurology* **103** (9). <https://doi.org/10.1212/WNL.000000000000209947> (2024).
36. Prasad, A. et al. Propensity score matching in otolaryngologic literature: A systematic review and critical appraisal. Vol. 15, *PLoS ONE*. Public Library of Science; (2020). <https://doi.org/10.1371/journal.pone.0244423>
37. Yao, X. I. et al. *Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies* 109 (Oxford University Press, 2017). *Journal of the National Cancer Institute*10.1093/jnci/djw323
38. Li, M. Using the propensity score method to estimate causal effects: A review and practical guide. Vol. 16, *Organizational Research Methods*. SAGE Publications Inc.; 188–226. DOI: <https://doi.org/10.1177/1094428112447816> (2013).
39. Suttrop, M. M., Siegerink, B., Jager, K. J., Zoccali, C. & Dekker, F. W. *Graphical Presentation of Confounding in Directed Acyclic Graphs*. 30, 1418–1423 (Oxford University Press, 2015). *Nephrology Dialysis Transplantation*10.1093/ndt/gfu325

40. Digitale, J. C., Martin, J. N. & Glymour, M. M. Tutorial on directed acyclic graphs. *J. Clin. Epidemiol.* **142**, 264–267. <https://doi.org/10.1016/j.jclinepi.2021.08.001> (2022).
41. Williamson, T. & Ravani, P. *Marginal Structural Models in Clinical Research: when and How To Use them?* 32, ii84–90 (Oxford University Press, 2017). Nephrology Dialysis Transplantation [10.1093/ndt/gfw341](https://doi.org/10.1093/ndt/gfw341)
42. Cohen, J. B. et al. Leveraging marginal structural modeling with Cox regression to assess the survival benefit of accepting vs declining kidney allograft offers. *Am. J. Transplant.* **19** (7), 1999–2008. <https://doi.org/10.1111/ajt.15290> (2019).
43. Strohmaier, S. et al. Survival benefit of first Single-Organ deceased donor kidney transplantation compared with Long-term Dialysis across ages in Transplant-Eligible patients with kidney failure. *JAMA Netw. Open.* **5** (10), E2234971. <https://doi.org/10.1001/jamanetworkopen.2022.34971> (2022).
44. Funk, M. J. et al. Doubly robust Estimation of causal effects. *Am. J. Epidemiol.* **173** (7), 761–767. <https://doi.org/10.1093/aje/kwq439> (2011).
45. Gabriel, E. E. et al. Inverse probability of treatment weighting with generalized linear outcome models for doubly robust Estimation. *Stat. Med.* **43** (3), 534–547. <https://doi.org/10.1002/sim.9969> (2024).

Author contributions

Andrea Gomez-Montero: Conceptualization of the article, statistical analysis, and primary manuscript drafting. Andrea Garcia-Lopez: Support in methodological development and critical review of the manuscript. Santiago Cabas: Support in methodological development and critical review of the manuscript. Adrián Alfonso Nieves-Rico: Partial writing and critical review of the manuscript. Fernando Giron-Luque: Overall supervision, support in the study conceptualization, and final review.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-31596-9>.

Correspondence and requests for materials should be addressed to A.G.-L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025