ISSN 1669-9106 MEDICINA (Buenos Aires) 0000: 00: 00-00



# SESGO DE INDICACIÓN EN ESTUDIOS OBSERVACIONALES RETROSPECTIVOS: APLICACIONES DEL PUNTAJE DE PROPENSIÓN (*PROPENSITY SCORE*)

**MARIANA VAENA, MARÍA FLORENCIA GRANDE RATTI, IVÁN HUESPE**

Área de Investigación en Medicina Interna, Servicio de Clínica Médica,
Hospital Italiano de Buenos Aires, Buenos Aires, Argentina

**Dirección postal:** Mariana Vaena, Área de Investigación en Medicina Interna, Servicio de Clínica Médica, Hospital Italiano de Buenos Aires, Juan Domingo Perón 4190, 1199 Buenos Aires, Argentina
**E-mail:** mariana.vaena@hospitalitaliano.org.ar
**Recibido:** 28-I-2026
**Aceptado:** 27-VI-2026

## Resumen

Los estudios observacionales analíticos son útiles para evaluar el efecto de intervenciones cuando los ensayos clínicos aleatorizados no son factibles. Sin embargo, la ausencia de aleatorización introduce sesgo de indicación, ya que existen factores externos denominados confundidores, que se asocian a la probabilidad de recibir un tratamiento como también a desarrollar un desenlace o *outcome* (sin estar en la vía causal), y pueden distorsionar los resultados observados.

Entre las estrategias existentes para ajustar por posibles confundidores y emular una aleatorización, el puntaje de propensión (En inglés: *Propensity Score* [PS]) es uno de los métodos más utilizados. Se define como la probabilidad de recibir una intervención según un *set* dado de covariables, las cuales deben seleccionarse basándose en criterios causales y verificando rigurosamente el balance logrado entre grupos mediante diferencias estandarizadas y gráficos.

Puede aplicarse mediante diferentes métodos: ajuste directo, estratificación, emparejamiento (*matching*), o ponderación IPTW (*inverse probability of treatment weighting*). Cada técnica permite obtener distintos tipos de estimadores, condicionales o marginales, según el objetivo del análisis. Aunque estas herramientas no garantizan inferencia causal por sí mismas, constituyen aproximaciones robustas para reducir la confusión en estudios observacionales y por ello es fundamental conocerlas.

**Palabras clave:** puntaje de propensión, sesgo de indicación, estudios observacionales, ponderación por probabilidad inversa

## Abstract

***Indication bias in retrospective observational studies: applications of propensity scores***

Analytical observational studies are useful for evaluating the effects of interventions when randomized controlled trials are not feasible. However, the absence of randomization introduces confounding by indication, as external factors known as confounders may be associated both with the probability of receiving a treatment and with the occurrence of an outcome (without being part of the causal pathway), potentially distorting observed results.

Among the available strategies to adjust for potential confounders and emulate randomization, the Propensity Score (PS) is one of the most widely used methods. It is defined as the probability of receiving an intervention given a specific set of covariates, which should be selected based on causal considerations and followed by rigorous assessment of covariate balance between groups using standardized differences and graphical tools.

The PS can be implemented through several approaches, including direct adjustment, stratification, matching, or inverse probability of treatment weighting (IPTW). Each technique yields different types of estimands (conditional or marginal) depending on the analytic objective. Although these methods do not guarantee causal inference on their own, they represent robust approaches for reducing confounding in observational studies, underscoring the importance of their proper understanding and application.

**Key words:** propensity score, confounding by indication, observational studies; inverse probability weighting

**PUNTOS CLAVE**
**Conocimiento actual**

- Los estudios observacionales permiten evaluar intervenciones cuando los ensayos aleatorizados no son factibles, pero pueden presentar sesgo de indicación. Los métodos basados en el *Propensity Score* se utilizan para equilibrar covariables entre grupos y reducir la confusión, emulando algunas condiciones de la aleatorización.

**Contribución del artículo al conocimiento actual**

- Este artículo revisa los métodos más utilizados del *Propensity Score*, explica cómo seleccionar covariables, cómo diagnosticar el balance y cómo interpretar efectos condicionales y marginales. Detalla las aplicaciones de ajuste, *matching* e IPTW (*inverse probability of treatment weighting*), y ofrece ejemplos prácticos reproducibles con código en Stata y R.

Los estudios observacionales son cada vez más utilizados para estimar el efecto de tratamientos, intervenciones o exposiciones sobre un *outcome*, sobre todo cuando realizar un ensayo clínico aleatorizado (ECA) es inviable. Además, permiten evaluar intervenciones en un contexto de vida real (resultados pragmáticos), analizar población que habitualmente no sería incluida en un ECA (aportando mayor validez externa), y estudiar *outcomes* o enfermedades poco frecuentes que requerirían grandes tamaños muestrales[1,2].

Sin embargo, una ventaja indiscutible de los ECA radica en la aleatorización de la intervención, la cual garantiza que los grupos comparados sean similares respecto de sus características basales[3]. En cambio, en estudios observacionales, la intervención recibida por los sujetos no fue indicada de manera aleatorizada, sino influenciada por factores externos, generando así un sesgo de indicación que podría ocultar o exagerar el verdadero efecto en estudio[4].

***Por ejemplo, los pacientes más graves podrían haber recibido la terapia más intensa, de modo que en una comparación cruda de mortalidad el tratamiento parecerá ineficaz o incluso perjudicial, cuando en realidad el mayor riesgo proviene de la gravedad basal.***

Para estimar asociaciones entre intervenciones/exposiciones y *outcomes* en ausencia de aleatorización, existen estrategias que apuntan a equilibrar las covariables entre grupos, es decir, hacer que los tratados (o quienes recibieron la intervención en estudio) y no tratados (quienes no la recibieron) sean comparables entre sí. Las alternativas metodológicas principales se detallan en la Tabla 1.

### *Advertencia metodológica sobre inferencia causal*

Aunque los métodos basados en el puntaje de propensión (*Propensity Score*, PS) que presentaremos permiten reducir el sesgo de indicación y emular las condiciones de un ensayo clínico aleatorizado, su uso en estudios observacionales no garantiza por sí solo una inferencia causal válida. Para que las estimaciones obtenidas puedan interpretarse como efectos causales deben cumplirse ciertos supuestos, que quedan fuera del alcance de este manuscrito.

En la siguiente dirección web de *github*: https://github.com/marianavaena/propensity-score-ejemplos se encuentran disponibles de manera gratuita archivos de códigos para el *software stata* y R, con explicaciones paso a paso para la aplicación práctica de todos los análisis descriptos en el presente trabajo y sus interpretaciones, así como una base de datos ficticia donde el lector podrá probar el funcionamiento en la construcción del código y los resultados obtenidos.

**¿Qué es el *Propensity Score* y cómo seleccionamos las variables para calcularlo?**

**Tabla 1** | Alternativas metodológicas más frecuentemente usadas de ajuste de covariables en estudios observacionales

| Técnica | ¿Cómo funciona? | Tipo de efecto estimado | Ventajas principales | Limitaciones principales |
|---|---|---|---|---|
| *Matcheo* directo por covariables | Cada paciente del grupo tratado se "empareja" con un paciente no tratado que tenga valores similares en las covariables elegidas | Condicional / marginal sobre la muestra emparejada | Muy fácil de aplicar, intuitivo | Imposible con muchas covariables |
| Ajuste multivariable directo | Se ajusta un modelo clásico (ej. un modelo de regresión) que incluye las covariables como predictores junto al tratamiento/ intervención. | Condicional | Fácil de aplicar | Requiere modelar bien interacciones. Si son muchas covariables, requiere un tamaño de muestra muy grande. |
| *Propensity score* | Representa la probabilidad de que un paciente reciba la intervención, dadas sus covariables basales.<br>Se puede utilizar para:<br>– Ajustar<br>– Ponderar (IPTW)<br>– Emparejar<br>– Estratificar | Marginal o condicional (según el método) | Balancea covariables; útil con muchas variables | Requiere correcta especificación del PS |

*PS: Propensity Score; IPTW: inverse probability of treatment weighting (ponderación por la inversa de probabilidad de recibir tratamiento)*

El PS se define como: "La probabilidad condicional de recibir la intervención en estudio, dadas sus características basales"[5].

Para su cálculo, se deben seleccionar de antemano las variables consideradas confundidoras, teniendo en cuenta que, a diferencia de un modelo predictivo tradicional, el objetivo del modelo de PS no es predecir con la mayor precisión posible quién recibirá el tratamiento, sino controlar la confusión. Por lo tanto, la elección de variables no debe basarse en criterios puramente estadísticos, sino en una comprensión clínica y causal del problema[6].

En términos generales, el modelo de PS debe incluir todas las covariables que estén asociadas tanto con la probabilidad de recibir el tratamiento como con el *outcome*, es decir, los verdaderos confundidores. También se deben incorporar variables que sean fuertes predictores del *outcome*, aunque no estén necesariamente asociadas con la exposición, ya que su inclusión suele reducir la varianza del estimador sin introducir sesgo adicional. En cambio, *no* deben incluirse variables asociadas al tratamiento, pero no al *outcome,* ya que funcionan como instrumentos y aumentan el sesgo y la inestabilidad del estimador. Mucho más importante aún, deben excluirse las variables **afectadas por** el tratamiento (como mediadores o consecuencias tempranas de la intervención), ya que incorporarlas introduce sesgo y rompe la lógica temporal del modelo causal. Una herramienta útil para esta selección

es el uso de diagramas causales o DAGs (*Directed Acyclic Graphs*)[7].

Una vez calculado, el PS puede usarse de varias maneras, entre ellas el ajuste directo por PS, la estratificación, el emparejamiento (*matching*), o el IPTW: *inverse probability of treatment weighting* (ponderación por la inversa de probabilidad de recibir tratamiento)[8]. Estos métodos se comparan en la Tabla 2, y los dos últimos se describen con más detalle ya que son los recomendados para estimar efectos marginales.

**¿Cómo controlar que el *Propensity Score* funcionó?**

Estimar el PS es solo el primer paso. Para que sea útil, después de aplicarlo (ya sea mediante emparejamiento, estratificación o ponderación), los pacientes tratados y no tratados deben parecerse en sus covariables basales. A esto lo llamamos "diagnóstico de balance".

Un grupo está balanceado respecto de otro cuando *la distribución de las covariables confundidoras es comparable entre ambos grupos*. Esto significa *medias similares* en variables continuas y *proporciones similares* en variables categóricas.

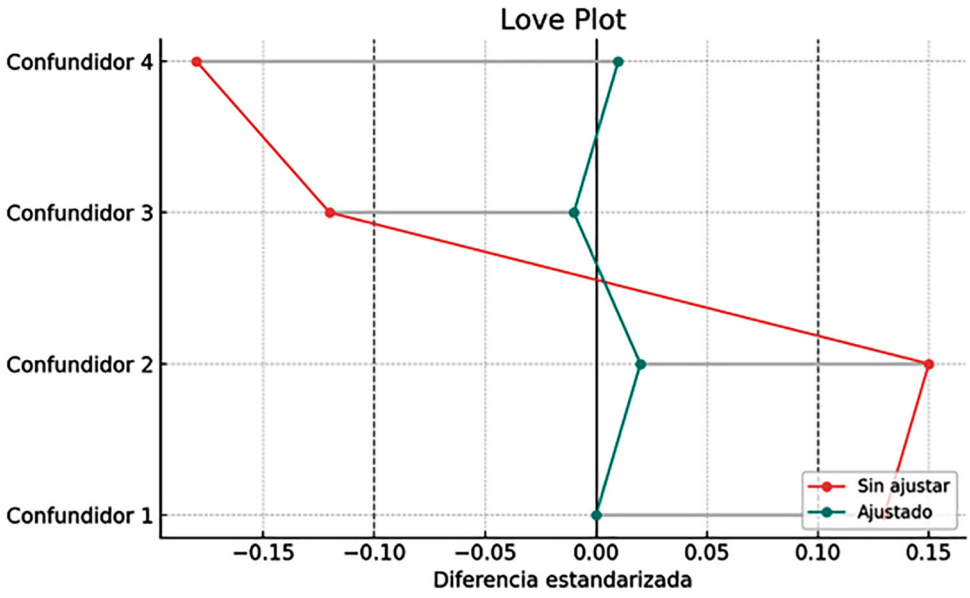
La forma más recomendada para medirlo es usando la diferencia estandarizada (o *Standardized Mean Difference*), la cual se interpreta como una medida de distancia entre grupos en unidades de desvío estándar, considerando <0.1 un balance aceptable, entre 0.1 y 0.2 un desbalance leve pero aceptable en algunos contextos, y > 0.2 desbalance importante, que estaría indicando que el PS no logró balancear correctamente la población[9]. Este diagnóstico suele visualizarse gráficamente mediante un *Love plot*, que permite comparar el balance de múltiples covariables antes y después del ajuste (Fig. 1).

Si luego de la aplicación del PS las covariables persisten desbalanceadas entre grupos, el ajuste realizado no es válido. Puede ocurrir incluso que variables inicialmente equilibradas, presenten desequilibrio luego de la aplicación del

**Tabla 2** | Métodos más comunes de uso de Propensity Score, con detalle de forma de uso, efecto estimado, ventajas y limitaciones de cada método.

| Método de PS | ¿Cómo se usa? | Tipo de efecto estimado | Ventajas principales | Limitaciones principales |
|---|---|---|---|---|
| **Ajuste por PS** | Se incluye el PS como covariable en el modelo del *outcome* | Condicional | Fácil de implementar | No balancea covariables |
| **Estratificación por PS** | Se divide la muestra en quintiles/deciles de PS y se compara el resultado entre tratados y no tratados en cada estrato | Condicional | Útil para visualizar balance y heterogeneidad del efecto. | Requiere que haya pacientes tratados y no tratados en cada estrato, requiere alto tamaño muestral |
| **Matching por PS** | Se emparejan pacientes tratados y no tratados con PS similar | Condicional / marginal sobre cada subgrupo emparejado | Buen balance en covariables emparejadas | Pierde pacientes sin *match*, requiere alto tamaño muestral |
| **IPTW** | Se aplican pesos o ponderaciones = 1/ PS o 1/(1-PS) según tratamiento recibido | Marginal | Emula randomización (genera población nueva con las covariables balanceadas) | Pesos extremos pueden inestabilizar el modelo |

*PS: Propensity Score; IPTW: inverse probability of treatment weighting (ponderación por la inversa de probabilidad de recibir tratamiento*




puntaje de PS (en particular mediante *PS matching*). Este fenómeno se ha descrito como una "paradoja" del uso del PS[10]. Por esta razón, independientemente de la forma de aplicación (emparejamiento, estratificación o ponderación), es importante evaluar sistemáticamente el balance de todas las covariables relevantes antes y después del ajuste.

De todos modos, en los casos en que no se consiga el balance luego del ajuste o si existiese la mencionada paradoja del PS, existen diversas estrategias para mejorar el balance, corregir el modelo y volver a intentarlo. Las causas más frecuentes se detallan en la Tabla 3.

**¿Por qué queremos estimar el efecto marginal o *average treatment effect* (y no el condicional) cuando queremos emular ensayos clínicos?**

Cuando el objetivo de un estudio es evaluar el efecto de una intervención de la misma manera en que lo haría un ensayo clínico, el tipo de efecto que estimamos cambia sustancialmente el mensaje clínico que comunicamos.

*El efecto condicional responde a la pregunta: "¿Cuál es el efecto del tratamiento dado cierto valor de covariables (edad, comorbilidades, gravedad, etc.)?"*, mientras que el efecto marginal responde a una pregunta más general: *"¿Qué pasaría con el outcome promedio (por ejemplo, la mortalidad) si todos los pacientes de la cohorte recibieran el tratamiento, comparado con si ninguno lo recibiera?" (ejemplo en Tabla 4)*. De esta manera, el efecto marginal reproduce la lógica de un ensayo clínico aleatorizado y es útil para la toma de decisiones generalizables[11].

Emparejamiento (*matching*) por *Propensity Score*: utilidad y limitaciones

El emparejamiento (o *matching*) es una de las aplicaciones más difundidas y consiste en emparejar pacientes tratados con no tratados que tengan valores similares de PS, con el objetivo de crear grupos comparables en términos de covariables basales. La definición de qué tan similares deben ser los sujetos para ser considerados comparables, puede establecerse a través del *caliper*, que define el grado máximo de diferencia aceptable en el PS entre los pares emparejados[12].

Si bien este enfoque resulta intuitivo y clínicamente atractivo, ya que permite construir una cohorte emparejada que se asemeja a una población "balanceada" sin necesidad de ponderaciones complejas, presenta limitaciones importantes cuando el objetivo es estimar un efecto marginal comparable al de un ensayo clínico. En

**Tabla 3** | Motivos más frecuentes de desbalance de grupos post-aplicación de *Propensity Score* y posibles soluciones

| Problema común | ¿Qué está pasando? | ¿Qué puedo hacer? |
|---|---|---|
| Modelo de PS muy simple | El modelo utilizado no alcanza a distinguir bien qué pacientes tienen más probabilidad de recibir el tratamiento. El puntaje resultante no logra separar correctamente los grupos y no genera un buen balance. | Agregar más variables relevantes que puedan influir en la indicación del tratamiento. Pensar en edad, comorbilidades, signos vitales, etc. |
| Relación no lineal mal modelada | Algunas variables no influyen de forma lineal. Por ejemplo, la edad puede afectar mucho en jóvenes y ancianos, pero no tanto en edades intermedias. | Como en todo modelo de regresión, se debe analizar la relación entre las variables y el tratamiento, para luego agregar términos cuadráticos, cúbicos o usar funciones más flexibles (ej. splines), en caso de ser necesario. |
| Variables categóricas mal codificadas | Usar variables que tienen categorías (por ejemplo, raza o nacionalidad) como si fueran números continuos, ya que el modelo trata de asignarles un orden o distancia que no tienen. | Codificar las variables categóricas usando i.variable en Stata para que se generen variables separadas (dummies) para cada categoría. |

PS: Propensity Score

**Tabla 4** | Ejemplo para comparar tipos de efectos estimados según el método utilizado y la interpretación correspondiente. En el ejemplo se plantea un estudio hipotético en donde se intenta estimar el efecto de administrar dexametasona en pacientes con insuficiencia respiratoria moderada.

| Tipo de efecto | Interpretación | Método | OR estimado |
|---|---|---|---|
| **Condicional** | *"En pacientes con características clínicas similares de gravedad, edad y comorbilidades, la administración de dexametasona se asoció con una reducción del OR de mortalidad del 42 % (OR = 0.58)..."* | Ajuste multivariado o PS como covariable | 0.58 (p = 0.01) |
| **Marginal (ATE)** | *"En toda la cohorte, si todos recibieran dexametasona vs. si ninguno lo hiciera, el OR de mortalidad se reduciría un 55 % (OR = 0.45)..."* | IPTW | OR: 0.45 (p = 0.01) |

*OR: Odds Ratio; PS: Propensity Score; ATE: average treatment effect; IPTW: inverse probability of treatment weighting (ponderación por la inversa de probabilidad de recibir tratamiento)*

primer lugar, el emparejamiento suele implicar la exclusión de individuos sin un par adecuado, lo que reduce el tamaño muestral y puede afectar la validez externa del estudio. Además, el efecto estimado se restringe a la población emparejada y no necesariamente refleja lo que ocurriría si toda la cohorte recibiera (o no) el tratamiento.

**Ponderación por la inversa de probabilidad de recibir tratamiento (*inverse probability of treatment weighting*, IPTW): cómo funciona y por qué permite estimar el efecto marginal**

El IPTW es una técnica basada en PS, que permite estimar el efecto marginal de una intervención. Para lograr esto, el método asigna a cada individuo un "peso" o una "ponderación" que depende de su probabilidad de recibir la intervención, y si efectivamente la recibió. Estos pesos se calculan así:

- **Pacientes que recibieron la intervención**: IPTW = 1 / PS
- **Pacientes que no recibieron la intervención**: IPTW = 1 / (1 - PS)

Con este cálculo, veremos que tanto los pacientes que recibieron la intervención con alta

probabilidad de recibirla, como los que no la recibieron con baja probabilidad de recibirla tienen un peso o ponderamiento bajo. En cambio, los pacientes que no recibieron la intervención pese a tener alta probabilidad de recibirla, al igual que aquellos que la recibieron pese a tener baja probabilidad tendrán un peso o ponderamiento alto[13].

Al aplicar estos pesos, cada individuo en la muestra contribuye proporcionalmente a lo "inusual" que fue su asignación de tratamiento. La lógica detrás de esto es que, al darle más peso a los casos menos esperables, se contrarrestan los patrones de variables que pudieran estar asociadas a la decisión que existió para la indicación del tratamiento. De este modo, el IPTW redistribuye el peso de cada individuo hasta que la distribución de covariables entre tratados y no tratados sea similar. Así, cualquier diferencia observada en el *outcome* puede atribuirse con mayor confianza al tratamiento y no a diferencias basales.

Sin embargo, una limitación importante del IPTW son los pesos extremadamente grandes, que aparecen cuando el PS está muy cercano a 0 o 1. En estos casos, unos pocos individuos con pesos muy altos pueden dominar el análisis, ya que aumentan de forma marcada la varianza del estimador y lo vuelven inestable. Para mitigar este problema se han propuesto varias estrategias, la más utilizada es el recorte de pesos extremos (por ejemplo, fijando un límite máximo, o recortando en percentiles como el 1–99 o 5–95)[13].

**¿Por qué decimos que el *inverse probability of treatment weighting* estima el efecto marginal?**

Cuando aplicamos el IPTW, estamos modificando el "peso" o la "importancia" de cada individuo en el análisis para que, en conjunto, la muestra represente una población en la que las características basales están distribuidas de manera similar entre quienes recibieron y no recibieron la intervención. Es decir, construimos una "pseudo-población", en la cual cada paciente contribuye más o menos al análisis.

En esta población ponderada, es posible estimar el efecto promedio del tratamiento, ya que no estamos realizando un ajuste explícito por cada covariable confusora, sino que el análisis se lleva a cabo sobre una población artificialmente equilibrada. De esta manera, se obtiene una estimación que refleja lo que ocurriría en promedio si toda la cohorte recibiera (o no recibiera) el tratamiento, sin necesidad de condicionar el análisis a perfiles específicos de pacientes[14].

## Conclusión

Aunque los estudios retrospectivos no permiten asignación aleatoria, existen estrategias estadísticas que nos permiten reducir el sesgo de indicación y estimar efectos comparables a los de un ensayo clínico.

Entre los enfoques disponibles, el PS permite estimar efectos condicionales o marginales según el objetivo del análisis. Sin embargo, su correcta implementación no es automática: requiere definir bien los confundidores, modelar adecuadamente sus relaciones, y verificar rigurosamente el balance logrado entre grupos. La técnica de IPTW, una de las aplicaciones más potentes del PS, permite estimar efectos marginales y es una herramienta muy útil cuando buscamos estimar asociaciones en estudios observacionales, especialmente si estudiamos causalidad emulando un ensayo clínico (*target trial emulation*). Pero como toda herramienta metodológica, requiere buena planificación, comprensión de los supuestos y diagnóstico riguroso del balance.

**Conflicto de intereses:** Ninguno para declarar

## Bibliografía

1. Bosdriesz JR, Stel VS, Van Diepen M, et al. Evidence-based medicine—When observational studies are better than randomized controlled trials. *Nephrology* 2020; 25:737-43.
2. Frakt AB. An observational study goes where randomized clinical trials have not. *JAMA* 2015; 313:1091-2.
3. Cook TD, DeMets DL. Introduction to Statistical Methods for Clinical Trials. CRC Press 2007; P 452.
4. Signorello LB, McLaughlin JK, Lipworth L, Friis S, Sørensen HT, Blot WJ. Confounding by Indication in Epidemiologic Studies of Commonly Used Analgesics. *Am J Ther* 2002; 9:199.

5. Joffe MM, Rosenbaum PR. Invited Commentary: Propensity Scores. *Am J Epidemiol* 1999; 150: 327-33.
6. Yang JY, Webster-Clark M, Lund JL, Sandler RS, Dellon ES, Stürmer T. Propensity score methods to control for confounding in observational cohort studies: a statistical primer and application to endoscopy research. *Gastrointest Endosc* 2019; 90:360-9.
7. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006; 163:12.
8. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* 2011; 46:399.
9. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 2009; 28:3083-107.
10. King G, Nielsen R. Why propensity scores should not be used for matching. *Polit Anal* 2019; 27:435-54.
11. Hernan MA, Robins JM. Causal Inference. *CRC Press* 2019; P 352.
12. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat* 2010; 10:150.
13. Chesnaye NC, Stel VS, Tripepi G, et al. An introduction to inverse probability of treatment weighting in observational research. *Clin Kidney J* 2021; 15:14.
14. Austin PC, Stuart EA, Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* 2015; 34:3661-79.