

Puntajes de propensión y estudios observacionales. Aplicaciones en investigación quirúrgica

Carlos Manterola D.^{1,2}, Juan Pablo Holguín C.^{1,3},
María Fernanda García A.^{1,4} y Georgina Muñoz O.^{1,3}

Propensity score and observational studies. Applications in surgical research

The ideal design to estimate the effects of interventions, is a randomized clinical trial (RCT), in which the random allocation (RA) to the intervention allows balancing the observed and unobserved characteristics of the subjects that make up the study groups. In situations where RA cannot be performed, there is an alternative to bias control, which is known as propensity score (PS). This tool is useful for estimating the effect of an intervention or exposure on an outcome in observational studies, especially when it is estimated that confounding and selection biases may exist due to an unintended intervention RA. Is an advanced matching technique for confounding variables such as age, sex, etc., which allows patients of similar age to be compared in the intervention and comparison groups. This can become more complex if it is decided to add many variables to the matching process (age, sex, ethnicity, educational level), since it will be difficult to find exact matches for the subjects under study. PS then solves this dimensionality problem by compressing the relevant factors into a single score, such that patients with similar PS are compared between the intervention and comparison groups. The aim of this manuscript was to generate a study document referring to some issues of the use of PS in surgical research.

Keywords: “propensity score”[Mesh]; “propensity score matching”; “propensity score matching analysis”; “observational study” [Publication Type]; “bias”[Mesh].

Resumen

El diseño ideal para estimar los efectos de intervenciones es un ensayo clínico con asignación aleatoria (AA), en los que esta AA a la intervención, equilibra características observadas y no observadas de los sujetos que componen los grupos en estudio. Si no se puede realizar AA, existe una alternativa para controlar sesgos, conocida como puntajes de propensión (PP). Los PP son útiles para estimar el efecto de una intervención o exposición sobre un resultado en estudios observacionales, especialmente cuando se estima la posibilidad que existan sesgos de confusión y selección debido a una asignación de tratamiento no aleatoria. Son una técnica de emparejamiento avanzada de variables de confusión como edad, sexo, etc., permitiendo que se puedan comparar pacientes de edad similar en grupos de intervención y de comparación. Esto, se puede complejizar si se decide agregar muchas variables al proceso de emparejamiento (edad, sexo, etnia, nivel educacional), puesto que será difícil encontrar coincidencias exactas para los sujetos. Entonces, los PP resuelven este problema de dimensionalidad comprimiendo factores relevantes en un solo puntaje, así, los pacientes con PP similares se comparan entre los grupos de intervención y comparación. El objetivo de este manuscrito fue generar un documento de estudio referente al uso de los PP y su rol en investigación quirúrgica.

Palabras clave: puntaje de propensión; emparejamiento por puntaje de propensión; estudios observacionales; sesgo.

¹Programa de Doctorado en Ciencias Médicas, Universidad de La Frontera, Temuco, Chile.

²Centro de Estudios Morfológicos y Quirúrgicos (CEMyQ), Universidad de La Frontera, Temuco, Chile.

³Universidad del Azuay, Cuenca, Ecuador.

⁴Universidad Central del Ecuador, Ecuador.

Recepción 2021-06-26, aceptado 2021-07-28

Correspondencia a:
Dr. Carlos Manterola D
carlos.manterola@uf Frontera.cl

Introducción

Al comparar dos o más grupos sometidos a intervenciones, se piensa en un proceso de asignación aleatoria (AA), puesto que es el estándar de refe-

rencia. Sin embargo, no siempre es posible asignar intervenciones y exposiciones al azar, representando un potencial riesgo de sesgo de confusión¹.

Para ello, se dispone de alternativas. Una de ellas es conducir estudios con diseños cuasiexperimen-

DOCUMENTOS

tales². Otra, es aplicar el puntaje de propensión o puntuación de propensión (PP) a estudios observacionales, lo que permite controlar la confusión mediante ajuste, emparejamiento o estratificación; incluyendo la PP como covariable en un modelo de regresión multivariable, ponderando la población de estudio mediante el PP o emparejando por puntaje de propensión (EPP)^{3,4}. Esto significa que los métodos de PP pueden controlar más factores de confusión que otros métodos, particularmente en el caso de resultados poco comunes. Una vez aplicada el PP, se deben señalar los factores de confusión considerados y en qué medida se equilibraron entre los grupos de estudio^{3,5}.

Reportes del uso de métodos PP, se vienen entregando en la literatura desde hace un par de décadas⁶⁻⁹, el primero corresponde a Rosenbaum y

Rubin, quienes introdujeron este concepto para estudiar causalidad en estudios observacionales¹⁰. No obstante ello, el término *MeSH* “Propensity Score”, sólo fue introducido como tal en 2010, y hace referencia a comparar dos tratamientos, “caso” y “control”; y la “probabilidad condicional de exposición a un tratamiento dadas las covariables observadas”¹¹.

El EPP es entonces, uno de los métodos de PP, con el que se puede garantizar una distribución uniforme de factores de confusión (como edad y sexo), entre dos grupos a comparar en los que no se puede realizar una AA, incrementando de este modo la comparabilidad entre ellos¹². Por ende, es un método estadístico que se puede aplicar en estudios observacionales, en los que los sujetos a estudio con el mismo PP tendrán, en promedio, una distribución similar de los factores de confusión medidos o considerados (Figuras 1 y 2)^{5,12,13}.

Por otra parte, la cirugía basada en la evidencia se ha establecido como piedra angular de la buena práctica clínica, asociándose a la mejoría de los procesos clínicos y de calidad en la educación quirúrgica. Requiere de cirujanos capacitados para realizar investigación metodológicamente sólida¹⁴⁻¹⁷. Por ello, el conocimiento de metodología de investigación y estadísticas es muy relevante; sobre todo si pensamos que en la mayor parte de los estudios quirúrgicos no se puede ejecutar una asignación aleatoria, por lo que el EPP ofrece una alternativa al análisis multivariado en la evaluación de estudios quirúrgicos observacionales⁶. Desafortunadamente, muchos cirujanos desconocen este importante enfoque analítico que ha ido ganando progresiva importancia en investigación clínica.



Figura 1. Representación gráfica del EPP entre un grupo de expuestos y otro de no expuestos a una intervención.

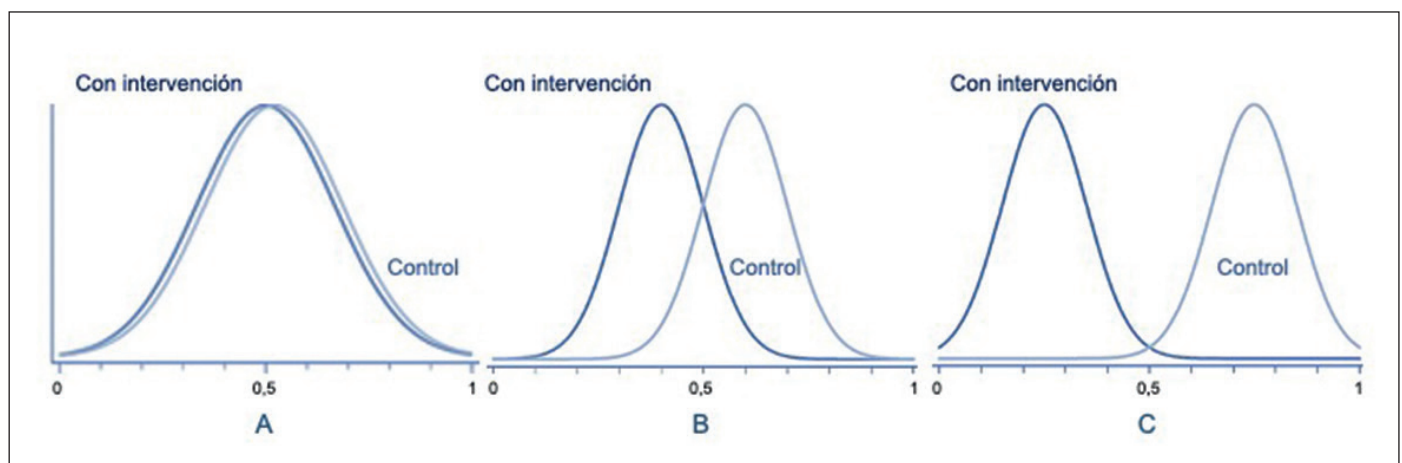


Figura 2. Distribución de la PP para un grupo con intervención y un grupo control. La línea azul representa al grupo experimental y la celeste al grupo control. Observándose desde la opción de una comparabilidad casi perfecta entre los dos subgrupos a estudio (2A), hasta la ausencia de comparabilidad dada en este caso por una inadecuada generación del EPP (2C). **A.** Superposición perfecta. **B.** Superposición parcial. **C.** Sin superposición.

El objetivo de este manuscrito fue generar un documento de estudio referente al uso del EPP en investigación quirúrgica.

¿Cómo funciona?

El PP corresponde a la probabilidad que el paciente sea asignado a intervención o exposición en vez de ser control (considerando sus covariables). Así, en un ensayo clínico con AA, el PP de cada paciente reclutado es conocido, de tal forma que en aquellos escenarios en los que el 50% de los pacientes son asignados a cada grupo en estudio, la PP de cada sujeto es de $\frac{1}{2}$.

En un escenario concreto, por ejemplo, un grupo de pacientes con carcinoma hepatocelular no reseable refractario a quimioembolización (CHCNRRQE)¹⁸, y decidiéramos aplicar “experimentalmente”, una opción de tratamiento emergente con radioembolización transarterial (RETA), con microesferas con itrio-90; tendríamos dos opciones: recurrir a una serie de casos (pacientes con CHCNRRQE tratados con RETA), que aportaría resultados de nivel de evidencia tipo 4 o realizar una comparación con “grupo control” que ya tenemos (CHCNRRQE) y observar los resultados, mejorando el nivel de evidencia de los hallazgos (Figura 3).

Si en el subgrupo tratado con RETA observáramos mayor supervivencia que en el no tratado, podría ser que no existiera diferencias reales en la supervivencia. Esto podría explicarse porque estaríamos comparando subgrupos potencialmente diferentes (edad, características del hepatocarcinoma, etc.), es decir, agregaríamos un efecto de confusión por una o más variables, al estudio de asociación entre la condición de salud en estudio y la intervención en estudio; y si no se tiene esto en consideración, podríamos terminar con una conclusión sesgada por variables de confusión¹.

¿Cómo resolver el problema?

Primero, identificando las potenciales variables de confusión. Sin embargo, los investigadores enfrentan un problema, puesto que la teoría sugiere que las variables que no están relacionadas con la exposición, pero sí con el resultado siempre deben incluirse en el modelo, lo que disminuirá la varianza de un efecto de exposición estimado sin aumentar el sesgo, o incluso minimizando el sesgo no sistemático¹⁹. Es así como en estudios pequeños, incluir variables fuertemente asociadas a la exposición,

pero débilmente con el resultado, puede ser perjudicial para estimar en un sentido de error cuadrático medio²⁰.

Luego, generando estratos similares de forma tal que la comparación del efecto (RETA vs. no RETA) se realice en grupos similares; aplicando modelos de regresión para calcular las PP²¹; y emparejando pacientes con CHCNRRQE tratados con RETA, con otros de similares características pero no tratados con RETA^{5,12}. Posteriormente, los valores de PP se estratifican y el efecto causal (supervivencia global), se compara entre individuos con valores cercanos de puntuación; para finalizar, se puede determinar magnitud del efecto, calculando el RR, el OR o el HR con sus respectivos intervalos de confianza del 95%. Como es muy poco usual que dos pacientes tengan el mismo PP, se empareja los pacientes con CHCNRRQE con y sin RETA, con valores cercanos al PP, estableciéndose estratos (Figura 3).

Es decir, en primer lugar, realizar una regresión logística definiendo la variable dependiente $Y = 1$ (intervención o exposición); $Y = 0$ (control o no exposición), determinando las variables de confusión adecuadas (variables hipotéticamente asociadas con la intervención o exposición y el resultado); y obtener la PP aplicando el concepto de probabilidad predicha (p) o $\log [p / (1 - p)]$. A continuación, generar el emparejamiento de cada sujeto en estudio según PP: emparejamiento por calibre, por estratificación, por “el vecino más cer-

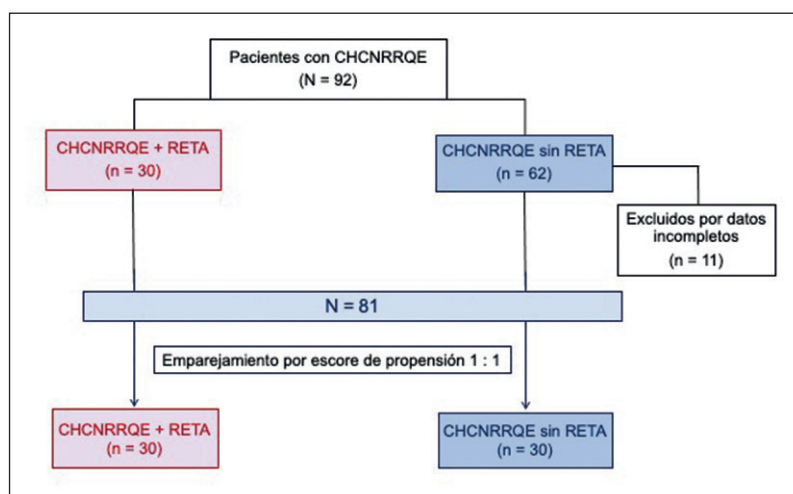


Figura 3. Diagrama de flujo de participantes de un estudio observacional respecto de la utilización de radioembolización transarterial (RETA) con microesferas con itrio-90; en pacientes con carcinoma hepatocelular no reseable y refractario a quimioembolización (CHCNRRQE), aplicando un EPP de 1:1 (18).

DOCUMENTOS

Figura 4. Diagramas de densidad de la gravedad promedio de síntomas basales en un conjunto de datos empíricos, aplicando el método por diferencias en diferencias. **A:** antes y **B:** después de aplicar emparejamiento de Kernel y los pesos lineales locales.

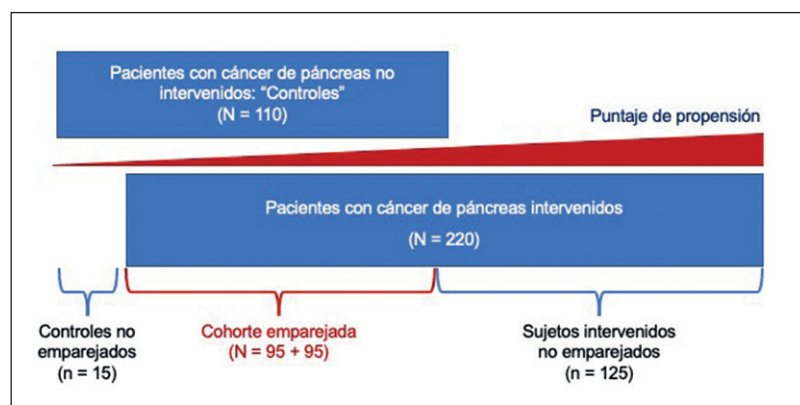
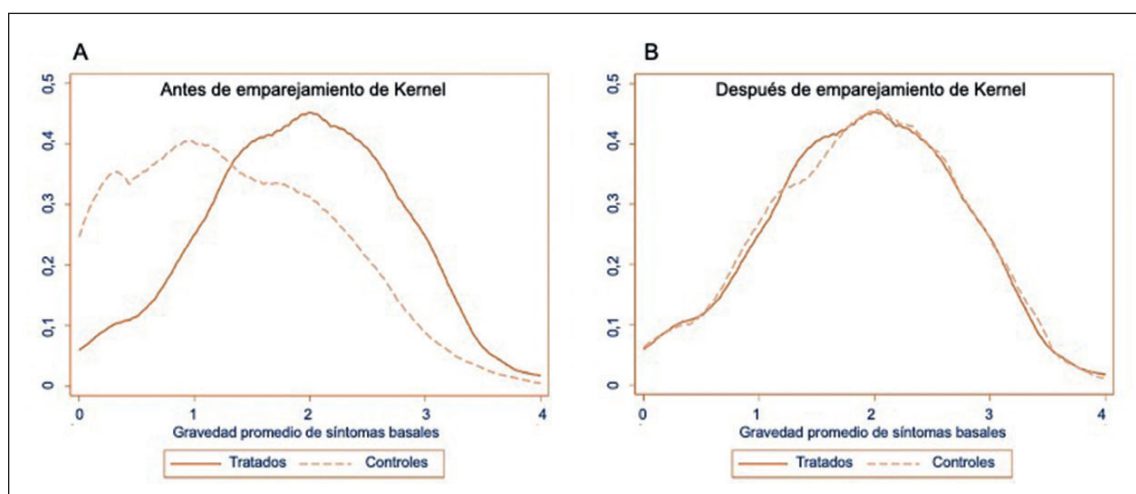


Figura 5. Representación gráfica del EPP. Ciento diez pacientes con cáncer de páncreas no operados (controles), fueron emparejados con 220 pacientes con cáncer de páncreas sometidos a cirugía resectiva. Se estimaron PP a partir de algunas covariables iniciales. Estos PP representan la probabilidad que los sujetos sean tratados dadas las covariables. Luego, las PP se utilizan para encontrar coincidencias de controles y sujetos tratados. Sólo 95 controles se pudieron emparejar con 95 sujetos intervenidos; los 15 restantes no tenían contrapartes con características similares en el grupo operado. De manera análoga, 125 de los pacientes operados quedaron sin emparejar.

cano”, o por diferencias en diferencias aplicando Kernel y pesos lineales locales²² (Figura 4). Luego, verificar que las covariables estén equilibradas entre el grupo de intervención o exposición y el de control de la nueva muestra emparejada. Finalmente, realizar un análisis multivariado sobre la base de la nueva muestra²³⁻²⁶ (Figura 5).

La estimación de las PP se puede realizar en paquetes estadísticos tradicionales, recomendándose el uso de “R” y “STATA”^{22,27-29}. Sin embargo, también hay evidencia del uso de SAS y SPSS^{22,25,26,28,29}.

¿Cómo implementar y evaluar el EPP?

La implementación y evaluación se resume en etapas³⁰:

Primero, identificar la viabilidad de aplicar PP con los datos disponibles y garantizar la disponibilidad de los datos de aquellas variables consideradas como posibles factores de confusión.

Segundo, obtener el conjunto de los PP a partir de un modelo de regresión logística con el grupo de intervención como resultado y las variables de equilibrio como predictores.

Tercero, emparejar a pacientes de los grupos de estudio con PP similares, equilibrando las potenciales variables de confusión identificadas al comienzo del estudio.

La cuarta etapa consiste en evaluar el éxito de la correspondencia con diagnósticos de balanza, de forma gráfica o analítica.

Quinto, aplicar herramientas estadísticas apropiadas, utilizando datos emparejados por propensión para comparar los resultados entre los grupos de intervención en estudio.

Fortalezas y limitaciones de los pp

Fortalezas:

Existe evidencia escasa, respaldando su uso y adecuada eficacia en estudios con poblaciones de 40 a 50 pacientes^{31,32}; donde los resultados de EPP no difieren sustancialmente de los obtenidos a través de modelos de regresión³³. Permite reducir la dimensionalidad si disminuyen las variables de confusión a una o dos³⁴.

Tabla 1. Resumen de ventajas y desventajas del EPP

Fortalezas	Limitaciones
Comparable con modelos de regresión	Sesgos de confusión
Permite reducir la dimensionalidad	Sesgo de censura
Se puede introducir como cofactor de otro análisis multivariado	Afectada por datos faltantes
Eficaz para realizar análisis adicionales en los sujetos emparejados	Requiere muestras de gran tamaño
Útil para estratificar y analizar resultados en grupos más homogéneos	No permite eliminar diferencias producidas por variables desconocidas
	Emparejamiento de poca y gran cercanía afectan el tamaño de la muestra e intervalos de confianza
	Afectada por ausencia de pacientes con PP similares en los grupos en comparación

Reduce tiempos y costes respecto a AA¹⁹.

Destacan indicaciones para su utilización como el hecho que se puede introducir como un cofactor de otro análisis multivariado, por ejemplo, de un modelo de Cox. También es útil para generar criterios de estratificación con el objetivo de analizar otros resultados en grupos *a priori* más homogéneos. Asimismo, es eficaz para análisis adicionales en pacientes emparejados³⁴ (Tabla 1).

Limitaciones

La ausencia de pacientes con PP similares en los grupos en comparación significará excluir a éstos del análisis, lo que puede ser fuente de sesgos al reducir la estimación del efecto en la población objetivo, reduciendo la generalización de resultados a poblaciones similares^{9,35}. También se ha evidenciado que es susceptible de sesgo de censura³⁶.

Es afectada cuando el número de datos faltantes es importante, ya que en el análisis sólo se consideran pacientes con datos completos³⁷.

Se requiere de muestras numéricamente importantes porque la comparación se realizará sólo en pacientes que puedan emparejarse, excluyéndose los que no se puedan emparejar^{8,12}.

No permite eliminar diferencias producidas por variables latentes desconocidas entre los grupos de estudio^{9,35,38}; así como tampoco equilibrar aquellas variables que no fueron registradas desde el comienzo de la investigación^{7,34,37}.

Si el emparejamiento requerido no es de gran cercanía, se emparejarán pacientes distintos. Por otro lado, si el emparejamiento requerido es muy cercano, se producirá una reducción del tamaño muestral debido a que se logrará el emparejamiento de pocos pacientes, determinando un incremento del intervalo de confianza^{7,34,37}.

Se puede generar un importante número de estratos si se ajusta por numerosas covariables⁹, determinando problemas de factibilidad del estudio.

Si no hay superposición entre los grupos, o pacientes con PP que queden fuera de la zona de superposición, no se puede realizar la comparación porque significaría que se trata de poblaciones diferentes^{34,35} (Tabla 1).

Ejemplos

Ejemplo 1: “Ten-Year Outcomes of Percutaneous Radiofrequency Ablation for Colorectal Cancer Liver Metastases in Perivascular vs. Non-Perivascular Locations: A Propensity-Score Matched Study”. Con el objetivo de comparar resultados a largo plazo de ablación por radiofrecuencia percutánea (ARFP) para metástasis hepáticas colorrectales (MHCR) en ubicaciones perivasculares (PV) *versus* no perivasculares (NPV); se ejecutó un estudio retrospectivo de 388 pacientes con MHCR, sometidos a ARFP. Se realizó EPP para los grupos con PV y NPV mediante regresión logística, con valor del calibre de 0,02; emparejando por variables edad, sexo, tamaño tumoral, ubicación primaria, estadio T, metástasis en linfonodos, tiempo hasta metástasis hepática, número de metástasis hepáticas, antecedentes de resección de metástasis hepáticas antes de ARF y metástasis extra-hepáticas, con SPSS 22.0 y R versión 2.15.x. Se compararon tasas de progresión tumoral local acumulativa, SG, recidiva intra/extrahepática y complicaciones entre ambos grupos. Se compararon las tasas de progresión tumoral local acumulativa, SG, recidiva intra/extrahepática y complicaciones entre ambos grupos. Se logró emparejar 104 pacientes en los grupos PV y NPV. Las tasas

DOCUMENTOS

acumuladas de progresión del tumor local a los 6 meses, 1, 3 y 5 años, fueron 8,8%, 14,8%, 18,9% y 18,9% en el grupo PV; y 8,8%, 13,1%, 15,5% y 15,5% en el grupo NPV. La SG a 1, 3, 5 y 10 años fue 91,3%, 45,6%, 23,9% y 18,7% en el grupo PV; y de 88,0%, 47,2%, 27,2% y 22,6% en el grupo NPV. No se detectaron diferencias significativas entre los grupos en la progresión acumulativa del tumor local ($p = 0,567$, índice de riesgo: 1,224) o la supervivencia general ($p = 0,801$, índice de riesgo: 1,047). La tasa de complicaciones mayores fue 1,0% (1/104, $p > 0,999$) en ambos grupos. La recidiva intrahepática (cociente de riesgo: 1,564; $p = 0,002$) se asociaron con una supervivencia global deficiente²⁹.

Ejemplo 2: “*The role of radiotherapy in pulmonary large cell neuroendocrine carcinoma: propensity score matching analysis*”. Con el objetivo de determinar supervivencia de radioterapia (RT) en pacientes con carcinoma neuroendocrino de células grandes de pulmón (CNECGP), se trabajó con el conjunto de datos de vigilancia, epidemiología y resultados finales (SEER) de pacientes con CNECGP. Los PP se estimaron mediante un modelo de regresión logística, que incluyó las covariables que podrían haber afectado la supervivencia de los pacientes, excepto RT (edad al diagnóstico, sexo, raza, estado civil, sitio primario, lateralidad, grado, tamaño del tumor, estadio AJCC, cirugía y quimioterapia). Se utilizó el complemento EPP del software SPSS para determinar PP en cada caso. Posteriormente, se realizó el EPP utilizando correspondencia de vecino más cercano 1: 1 con un calibre de 0,01 para la aceptación de un par emparejado. Se identificaron 1.480 casos elegibles. Tras aplicar PP, se seleccionaron 980 pacientes (490 sin RT y 490 con RT). Los pacientes del grupo RT tuvieron SG y específica a 3 y 5 años significativamente más altas en comparación con los del grupo sin RT (ambos $p < 0,05$). Además, la RT fue un factor pronóstico independiente favorable de SG y específica en el análisis multivariado antes del EPP (SG: HR: 0,840; IC 95%: 0,739-0,954; $p = 0,007$. Supervivencia específica: HR: 0,847; IC 95%: 0,741-0,967; $p = 0,014$); y después del EPP (SG: HR: 0,854; IC 95%: 0,736-0,970; $p = 0,016$. Supervivencia específica: HR: 0,848; IC 95%: 0,735-0,978; $p = 0,023$)²⁵.

Ejemplo 3: “*Acute outcomes and 1-year mortality of intensive care unit-acquired weakness: A cohort study and propensity-matched analysis*”. Con el objetivo de determinar los resultados agudos, mortalidad a un año y costos de la debilidad adquirida en la UCI en pacientes UCI de estancia prolongada (≥ 8 días) y evaluar el impacto de la recuperación de debilidad al alta de la UCI; se recopilaron datos

prospectivos de un ensayo controlado aleatorio. El impacto de la variable “debilidad” en los resultados y los costes se analizó con EPP. Las comparaciones de pacientes con y sin debilidad se realizaron aplicando análisis de riesgo proporcional de Cox, visualizado con gráficos de Kaplan Meier. El 55% de los pacientes de estadía prolongada estaban débiles (227/415). 122 pacientes “débiles” se emparejaron con 122 pacientes no débiles, usando el vecino más cercano, utilizando un caliper de 0,1 de ancho y una proporción 1:1, sin reemplazo con la debilidad como variable dependiente. En comparación con los “no débiles” emparejados, los “débiles” tenían menor probabilidad de destete de ventilación mecánica (HR: 0,709 [0,549; 0,888]; $p = 0,009$), menor probabilidad de salir vivos de la UCI (HR: 0,698 [0,553; 0,861]; $p = 0,008$); y de alta hospitalaria (HR: 0,680 [0,514; 0,871]; $p = 0,007$). El riesgo de muerte a 1 año se incrementó aún más si la debilidad persistió y fue más grave en comparación con la recuperación de la debilidad al alta de la UCI ($p < 0,001$)³⁹.

Conclusiones

En este manuscrito, revisamos las características principales del EPP, fortalezas y limitaciones; desarrollando, finalmente, un ejemplo de la literatura.

EPP es una técnica útil, que salva estudios que de otra manera no se podrían finalizar. Sin embargo, no es la panacea y tampoco es tan simple de usar como parece. Puede ayudar a evaluar la comparabilidad de pacientes en estudios observacionales y permite explicar más factores de confusión potenciales que los enfoques convencionales de ajuste de covariables.

Sin embargo, las conclusiones generadas con su aplicación pueden ser cuestionables y no generalizables, ya que el de confusión de variables no consideradas no se puede corregir. La PP y el enfoque multivariado clásico no reemplazan a la AA, siendo una alternativa en situaciones en las que una AA no se puede realizar.

Responsabilidades éticas

Protección de personas y animales. Los autores declaran que en este manuscrito no se han realizado experimentos en seres humanos ni animales.

Confidencialidad de los datos. Los autores declaran que en este artículo no aparecen datos de pacientes.

Financiación: Ninguna.

Conflictos de interés: Ninguno.

Bibliografía

1. Manterola C, Otzen T. Los sesgos en investigación clínica. *Int J Morphol*. 2015;33:1156-64. doi: 10.4067/S0717-95022015000300056.
2. Manterola C, Otzen T. Estudios Experimentales 2 Parte. Estudios Cuasi-Experimentales. *Int J Morphol*. 2015;33:382-7. doi: 10.4067/S0717-95022015000100060
3. Williamson EJ, Forbes A. Introduction to propensity scores. *Respirology* 2014;19:625-35. doi: 10.1111/resp.12312.
4. Loux TM. Randomization, matching, and propensity scores in the design and analysis of experimental studies with measured baseline covariates. *Stat Med*. 2015;34:558-70. doi: 10.1002/sim.6361.
5. Groenwold RH. Propensity scores in observational research. *Ned Tijdschr Geneesk*. 2013;157:A6179.
6. Adamina M, Guller U, Weber WP, Oertli D. Propensity scores and the surgeon. *Br J Surg*. 2006;93:389-94. doi: 10.1002/bjs.5265.
7. Glynn R, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 2006;98:253-9. doi: 10.1111/j.1742-7843.2006.pto_293.x.
8. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine* 2008;27:2037-49. doi: 10.1002/sim.3150.
9. Pattanayak CW, Rubin DB, Zell ER. Propensity score methods for creating covariate balance in observational studies. *Rev Esp Cardiol*. 2011;64:897-903. doi: 10.1016/j.recesp.2011.06.008.
10. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41-55.
11. National Center for Biotechnology Information (NCBI). MeSH. Propensity Score. Disponible en: <https://www.ncbi.nlm.nih.gov/mesh/?term=propensity+scores>.
12. Pearl J. Understanding propensity scores. *Causality: Models, Reasoning, and Inference* (2nd edition). 2009. Nueva York: Cambridge University Press. ISBN 978-0-521-89560-6.
13. Baek S, Park SH, Won E, Park YR, Kim HJ. Propensity score matching: a conceptual review for radiology researchers. *Korean J Radiol*. 2015;16:286-96. doi: 10.3348/kjr.2015.16.2.286.
14. Manterola C. Medicina basada en la evidencia. Conceptos generales y razones para su aplicación en cirugía. *Rev Chil Cir*. 2002;54:550-4.
15. Manterola C. Medicina basada en la evidencia en la práctica quirúrgica habitual ¿Utopía, moda o inevitable exigencia actual? *Rev Chil Cir*. 2003;55:213-5.
16. Meakins JL. Evidence-based surgery. *Surg Clin North Am*. 2006;86:1-16. doi: 10.1016/j.suc.2005.10.004.
17. Durán Vega HC. Cirugía basada en la evidencia. Una herramienta necesaria. *Cir Cir*. 2015;83:265-70. doi: 10.1016/j.circir.2015.05.012.
18. Klompenhouwer EG, Dresen RC, Verslype C, Laenen A, De Hertogh G, Deroose CM, et al. Safety and Efficacy of Transarterial Radioembolisation in Patients with Intermediate or Advanced Stage Hepatocellular Carcinoma Refractory to Chemoembolisation. *Cardiovasc Intervent Radiol*. 2017;40:1882-90. doi: 10.1007/s00270-017-1739-5.
19. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127:757-63. doi: 10.7326/0003-4819-127-8_part_2-199710151-00064.
20. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149-56. doi: 10.1093/aje/kwj149.
21. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361-87. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
22. Garrido MM, Kelley AS, Paris J, Roza K, Meier DE, Morrison RS, Aldridge MD. Methods for Constructing and Assessing Propensity Scores. *Health Serv Res*. 2014;49:1701-20. doi: 10.1111/1475-6773.12182.
23. Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. *Eur Heart J*. 2011;32:1704-8. doi: 10.1093/eurheartj/ehj1031.
24. Benedetto U, Head SJ, Angelini GD, Blackstone EH. Statistical primer: propensity score matching and its alternatives. *Eur J Cardiothorac Surg*. 2018;53:1112-7. doi: 10.1093/ejcts/ezy167.
25. Cao L, Wu HF, Zhao L, Bai Y, Jiang ZL, Yang WJ, Liu SX. The role of radiotherapy in pulmonary large cell neuroendocrine carcinoma: propensity score matching analysis. *J Radiat Res*. 2020;61:594-601. doi: 10.1093/jrr/rraa036.
26. Navarro JG, Kang I, Hwang HK, Yoon DS, Lee WJ, Kang CM. Oncologic safety of laparoscopic radical cholecystectomy in pT2 gallbladder cancer: A propensity score matching analysis compared to open approach. *Medicine (Baltimore)*. 2020;99:e20039. doi: 10.1097/MD.00000000000020039.
27. Kim H. Propensity Score Analysis in Non-Randomized Experimental Designs: An Overview and a Tutorial Using R Software. *New Dir Child Adolesc Dev*. 2019;2019(167):65-89. doi: 10.1002/cad.20309.
28. Luo Z, Gardiner JC, Bradley CJ. Applying Propensity Score Methods in Medical Research: Pitfalls and Prospects. *Med Care Res Rev*. 2010;67:528-54. doi: 10.1177/1077558710361486.
29. Jiang B, Luo H, Yan K, Zhang Z, Li X, Wu W, et al. Ten-Year Outcomes of Percutaneous Radiofrequency Ablation for Colorectal Cancer Liver Metastases in Perivascular vs. Non-Perivascular Locations: A Propensity-Score Matched Study. *Front Oncol*. 2020;10:553556. doi: 10.3389/fonc.2020.553556.
30. Staffa SJ, Zurakowski D. Five Steps to Successfully Implement and Evaluate Propensity Score Matching in Clinical Research Studies. *Anesth Analg*. 2018;127:1066-73. doi: 10.1213/ANE.0000000000002787.
31. Pirracchio R, Resche-Rigon M, Chevret M. Evaluation of the propensity score methods for estimating marginal Odds ratio in case of small simple size. *BMC Med Res Methodol*. 2012;12:70. doi:

DOCUMENTOS

- 10.1186/1471-2288-12-70.
32. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32:2837-49. doi: 10.1002/sim.570.
 33. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*. 2005;58:550-9. doi: 10.1016/j.jclinepi.2004.10.016.
 34. Filleron T, Kwiatowski F. Propensity score: A credible alternative to randomization? *Bull Cancer*. 2016;103:113-22. doi: 10.1016/j.bulcan.2015.10.012.
 35. Reiffel JA. Propensity Score Matching: The 'Devil is in the Details' Where More May Be Hidden than You Know. *Am J Med*. 2020;133:178-81. doi: 10.1016/j.amjmed.2019.08.055.
 36. Penning de Vries BBL, Groenwold RHH. Cautionary note: propensity score matching does not account for bias due to censoring. *Nephrol Dial Transplant*. 2018;33:914-6. doi: 10.1093/ndt/gfx198.
 37. Streiner DL, Norman GR. The pros and cons of propensity scores. *Chest*. 2012;142:1380-2. doi: 10.1378/chest.12-1920.
 38. Okoli GN, Sanders RD, Myles P. Demystifying propensity scores. *Br J Anaesth*. 2014;112:13-5. doi: 10.1093/bja/aet290.
 39. Hermans G, Van Mechelen H, Clerckx B, Vanhullebusch T, Mesotten D, Wilmer A, et al. Acute outcomes and 1-year mortality of intensive care unit-acquired weakness. A cohort study and propensity-matched analysis. *Am J Respir Crit Care Med*. 2014;190:410-20. doi: 10.1164/rccm.201312-2257OC.