Rev Med Chile 2016; 144: 364-370

¿Qué son las puntuaciones de propensión?

DAGOBERTO OJEDA[1,a], ROCÍO GÓMEZ[2], ÁLVARO BURGOS[2,b]

[1]Clínica Dávila.
[2]Universidad de los Andes.
[a]Magíster Bioestadística.
[b]Becado de Anestesiología.

Fuente de apoyo financiero: Ninguna.

Recibido el 28 de abril de 2015, aceptado el 2 de octubre de 2015.

Correspondencia a:
Dr. Dagoberto Ojeda
Av. El Bosque 701, Dpto. 201, Providencia.
eojedadinamarca@gmail.com

What are the propensity scores?

Medical research is constantly looking for causality. Among study designs, randomized controlled trials are the most reliable way to estimate causal effects but are not always feasible. When this is the case, observational studies must be performed but this type of design unavoidably implies bias. Propensity scores, defined as the probability to receive a treatment conditional to a set of covariables allows to overcome confusion bias when searching for causal effects.

(Rev Med Chile 2016; 144: 364-370)

Key words: *Bias, Propensity Score; Causality; Observational Study.*

Una explicación científica es aquella que identifica causas[1]. Éste es el objetivo de la investigación médica que está en constante búsqueda del motivo que desencadenó un evento[2]. No siempre es fácil discernir entre efecto causal, mera correlación o incluso asociación espuria[3,4]. Aunque no se tengan conocimientos formales de epidemiología, invariablemente existirá una comprensión innata de lo que es un modelo causal. Si no se supiera que el fuego quema o que el cruzar la calle sin mirar podría provocar la muerte, no se habría sobrevivido para estar leyendo este texto[5]. Empíricamente se habla de efecto causal cuando se intenta por ejemplo dilucidar si un analgésico es efectivo en el alivio de una cefalea. Lo ideal sería evaluar las respuestas en un mismo individuo dependiendo de si ingiere o no el medicamento. El contrastar lo que sucede al realizar una intervención con lo que acontece al no intervenir, es lo que se ha llamado modelo contrafáctico o de Resultados Potenciales[4,6], también conocido como modelo de Neyman-Rubin[7] (Figura 1).

En el lenguaje cotidiano, lo contrafáctico está frecuentemente presente al referirnos a escenarios ficticios en relación a un suceso verdadero: "si hubiéramos llegado antes se habría salvado…", es la supuesta respuesta que se habría obtenido de no haber ocurrido el evento real[8].

Como en la realidad no se puede comparar en un mismo individuo y al mismo tiempo lo que sucedería al administrar un tratamiento y al no hacerlo, se recurre a establecer este cotejo en forma colectiva. Vale decir, en un grupo de personas con cefalea se realiza una partición en dos subgrupos, uno en el que se realiza la intervención y otro que no se interviene, y se observa el resultado (alivio del dolor). Si eventualmente en el subgrupo tratado hubiera una mayor proporción de personas con cefalea más dolorosa que en el subgrupo no tratado, podría acontecer que no existiera diferencia en el alivio del dolor con el uso del analgésico. ¿Deberíamos hablar en este caso de ausencia de efecto causal? No, puesto que estamos comparando individuos de características distintas y no lo ocurrido en una misma persona. Lo que ocurre en este ejemplo es que se está introduciendo una tercera variable, además del factor causal y del efecto, que es la intensidad de la cefalea, la que está relacionada tanto con la causa como con el efecto y que confunde la relación entre el analgésico y el alivio del dolor[9,10]. Cuando se investigan relaciones causales en forma natural, es decir, cuando meramente se observa lo que ocurre en grupos de personas expuestas y no expuestas a un determinado factor, será improbable que estos grupos estudiados no difieran en sus características basales, introduciendo este factor de confusión anteriormente descrito.

Este tipo de estudios llamados observacionales, se alejan entonces del principio contrafáctico y

podrían conducir a conclusiones sesgadas provocadas por estas características disímiles a las que se han denominado variables de confusión. La solución a este problema se obtiene identificando estas variables y estableciendo estratos de similares características de manera tal que la comparación del efecto se realiza en grupos similares, lo que se logra estratificando, usando modelos de regresión[11,12], o bien pareando los individuos expuestos o tratados con otros de similares características pero no expuestos o no tratados[13].

La adherencia plena al principio del modelo contrafáctico sólo se logra utilizando estudios experimentales, vale decir cuando el investigador no es un simple observador sino que manipula el factor, tratamiento o exposición mediante asignación aleatoria. El hecho de aleatorizar permitirá explorar lo que ocurre en un grupo de personas que tienen la misma posibilidad de recibir o no el factor en estudio y que tiene características similares, obteniéndose grupos a comparar de características similares o balanceadas que permite llegar a conclusiones libres del sesgo de confusión. Este tipo de estudios son los llamados Ensayos Clínicos Aleatorizados, que constituyen la herramienta más eficaz para identificar efectos causales[5,14-16].

Dado que no siempre es posible efectuar Ensayos Clínicos por motivos éticos, pues no se puede aleatorizar a pacientes con cáncer a no recibir tratamiento, ni se le puede indicar a una persona que fume para evaluar si desarrollará un cáncer en el curso del tiempo, la única herramienta posible en estos contextos para estudiar un efecto causal son los estudios observacionales.

Puntuaciones de propensión

Rosenbaum y Rubin[17], introdujeron el concepto de puntuaciones de propensión en un artículo que trataba justamente de la búsqueda de efectos causales en estudios observacionales.

Las puntuaciones de propensión ilustran sobre la probabilidad de recibir un tratamiento determinado por un grupo de covariables[17]. En un ensayo clínico con dos grupos: uno sometido a un tratamiento y otro de control, la Puntuación de Propensión de todos los pacientes es 0,5. En ausencia de asignación aleatoria del factor de interés (estudio observacional), la probabilidad de recibir un determinado tratamiento es diferente entre los miembros del estudio, esta probabilidad es la puntuación de propensión y es por tanto, un número o puntaje que puede asumir cualquier valor entre 0 y 1. Esta probabilidad se calcula mediante una regresión logística dicotómica cuya variable respuesta es, recibir o no el tratamiento,

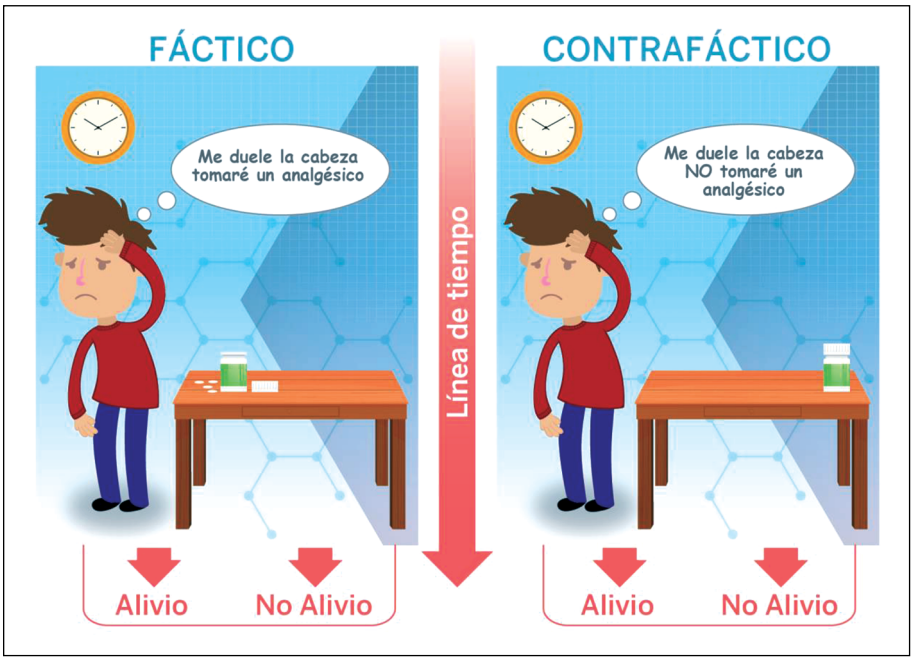

Figura 1. Modelo contrafáctico.

dadas las covariables de que se dispone (características basales de la población en estudio: edad, sexo, etc.). Además de la regresión logística, es posible calcular las puntuaciones de propensión a través de otros métodos estadísticos: análisis discriminante, árboles de clasificación y regresión (CART) y redes neurales[18].

Los valores de puntuación de propensión se estratifican y el efecto causal se compara entre individuos con valores cercanos de puntuación, a este apareamiento es lo que se llama "*Propensity score matching*"[19]. La medición del efecto (Riesgo Relativo, *Odds Ratio*, *Hazard Ratio*) calculado en este subgrupo de poblaciones pareadas constituye el efecto causal. Es imposible que dos individuos tengan el mismo puntaje, de manera que se parean individuos con valores cercanos estableciéndose estratos, la cercanía se calcula en base a métodos estadísticos multivariados (vecino más cercano)[19]. Si no pedimos mucha cercanía parearemos a individuos distintos y se perderá la intercambiabilidad que caracteriza a un modelo contrafáctico. Si pedimos demasiada cercanía nos quedaremos con pocos individuos pareados y el intervalo de confianza será muy ancho por reducción del tamaño muestral. Si no se encuentran individuos con Puntuación de Propensión similares en ambos grupos, los que queden aislados se excluirán del análisis y se producirá un sesgo puesto que no estaremos estimando el efecto en la población objetivo, sino en una población distinta y por tanto los resultados no serán extrapolables a ésta (disminuye la validez externa).

A diferencia de la aleatorización, las puntuaciones de propensión no permiten equilibrar aquellas características o covariables que no fueron registradas u observadas desde un principio[17,20].

En un estudio observacional las puntuaciones de propensión de los pacientes tratados serán mayores que las de los no tratados y esto es lo que causará confusión. Los individuos con la misma Puntuación de Propensión pueden diferir en sus covariables, sin embargo, su probabilidad de recibir el tratamiento es la misma. De esta manera las puntuaciones de propensión reducen las variables de confusión a una sola.

Esta metodología requiere de muestras grandes dado que toda comparación de resultados se realizará sólo en los individuos que puedan ser emparejados y se excluirá a los que no puedan ser apareados, sin embargo, se ha visto un buen desempeño de esta herramienta estadística incluso en muestras tan pequeñas como 40 pacientes[21]. Se necesita además que exista suficiente "sobreposición" de las puntuaciones de propensión entre los "tratados" y "no tratados", así obtendremos dos grupos comparables de tamaño adecuado (se verá esto en un ejemplo más adelante). También se requiere que el número de datos faltantes no sea muy grande, puesto que para el análisis sólo se consideran los pacientes con datos completos.

Es posible implementar la estimación de las puntuaciones de propensión en paquetes estadísticos habituales como STATA, R, SPSS, SAS, siendo los dos primeros especialmente recomendados por Guo y Fraser[22] en la última edición de su libro. En el ejemplo presentado en esta revisión se utilizó el paquete estadístico STATA 13.

Desde su introducción en 1983, el uso de puntuaciones de propensión en la literatura médica ha ido incrementando progresivamente[20]. Algunos autores han planteado que los resultados de los análisis con puntuaciones de propensión no difieren sustancialmente de los obtenidos a través de modelos de regresión[23,24]. Sin embargo, dichos métodos, la regresión lineal por ejemplo, tienen supuestos como exigir independencia entre las covariables que no necesariamente se cumplen al analizar estudios observacionales, lo cual estaría aumentando el sesgo[25]. Además se pierde la efectividad en la regresión cuando el número de covariables es muy numeroso[16,18,26,27], en estos casos se requiere además de tamaños muestrales muy grandes. Las puntuaciones de propensión permiten disminuir la dimensionalidad puesto que el conjunto de covariables quedarán reducidas a una, la Puntuación de Propensión[22,26].

No menos importante es el hecho de que al analizar un estudio observacional, la ausencia de aleatorización provocará un sesgo endógeno debido a la potencial correlación entre alguna variable independiente y el error de la regresión (la diferencia entre los resultados obtenidos con el modelo y los reales), lo que también llevará a resultados sesgados[22,26].

En comparación con el apareamiento, las puntuaciones de propensión también ofrecen la ventaja de disminuir la dimensionalidad al reducir a una sola la variable de confusión. El apareamiento originará un número muy elevado de subclases o estratos si se requiere ajustar por múltiples covariables[25].

Tanto las puntuaciones de propensión como las herramientas estadísticas tradicionales, no son capaces de eliminar las diferencias producidas por variables desconocidas latentes entre los grupos de estudio[28].

No es el objetivo de esta revisión el conferir mayor credibilidad a los resultados obtenidos mediante puntuaciones de propensión, sino simplemente remarcar que en algunos casos las herramientas estadísticas tradicionales, no tienen un buen desempeño y que en cambio las puntuaciones de propensión aproximan los estudios observacionales al modelo contrafáctico y por tanto permiten una identificación menos sesgada de efectos causales.

Ejemplo

Se presenta un ejemplo tomado del ámbito de la anestesiología. El uso profiláctico de dexametasona ha demostrado disminuir las náuseas y vómitos postoperatorios (NVPO)[29]. Se propone realizar un ensayo clínico para investigar este efecto y se decide utilizar el consumo de antieméticos (ondansetrón) en el período postoperatorio como índice de la ocurrencia de NVPO. Los anestesiólogos son reticentes a omitir el uso de dexametasona, puesto que su efecto está lo suficientemente demostrado[30] y, por lo tanto, consideran que no sería ético prescindir de su utilización. Entonces se decide realizar un estudio observacional y se registra en forma retrospectiva lo que ocurrió con una muestra aleatoria de 400 pacientes sometidos a anestesia general (Ojeda D, Burgos A, Gómez R. Factores de riesgo para náuseas y vómitos postoperatorios en anestesia general inhalatoria en adultos. XLII Congreso Chileno de Anestesiología 2014), en los que el uso de dexametasona quedó a criterio del anestesiólogo tratante. Al analizar los datos se observa que estos dos grupos son distintos: El grupo en que se usó dexametasona es de menor edad, hay mayor proporción de mujeres, mayor uso de óxido nitroso y mayor uso de ondansetrón profiláctico (Intraoperatorio), todos ellos son factores de riesgo conocidos para NVPO[30]. Ello hace pensar que se usó dexametasona por tratarse de pacientes con mayor probabilidad de padecer este problema. Dicho de otro modo, la "propensión" a usar dexametasona fue significativamente mayor en mujeres, en pacientes de menor edad, cuando se utilizó óxido nitroso y en los pacientes en que se usó ondansetrón profilácticamente.

Entonces se decide usar puntuaciones de propensión para compensar estas diferencias y obtener dos grupos comparables en los que se podría conocer el efecto causal del tratamiento (dexametasona), en el consumo postoperatorio de ondansetrón. Otra alternativa es ajustar por las variables que diferencian a los grupos, puesto que ellas producen confusión en la relación entre la dexametasona y el consumo de antieméticos postoperatorios, el resultado de esta alternativa lo vemos en la Tabla 1.

Al ajustar por las variables que difieren entre los grupos con y sin tratamiento, se verifica que la dexametasona produce una disminución de 0,27 mg (p-valor > 0,05), no significativa en el consumo de ondansetrón postoperatorio. Los pacientes de sexo femenino tuvieron mayor consumo de ondansetrón postoperatorio y el uso de ondansetrón profiláctico (intraoperatorio) disminuyó significativamente el uso postoperatorio de éste.

Se comparan estos resultados con los obtenidos al utilizar puntuaciones de propensión. Se realiza una regresión logística utilizando como variable dependiente (variable respuesta o *outcome*) el uso o no de dexametasona, y las variables predictores o independientes que son las covariables registradas en la muestra de pacientes estudiados. La Figura 2 muestra la clasificación de los pacientes. En el eje horizontal se observa el valor de las puntuaciones de propensión.

Obsérvese que algunos pacientes no pueden ser pareados, lo que reduce la muestra. Además, los pacientes tratados con dexametasona tienen

Tabla 1. Resultados Regresión Multivariada (ajustando por variables confusoras)

	Consumo ondansetrón postop. (mg)	IC 95%	p-valor
Dexametasona	-0,27	-0,9; 0,3	0,368
Edad	0,01	-0,1; 0,02	0,285
Mujeres	0,7	0,14; 1,3	0,017
N_2O	0,4	-0,8; 1,6	0,493
Ondansetrón*	-0,3	-0,13; -0,52	0,010

mg: Miligramos. IC: Intervalo de confianza. *Uso Intraoperatorio.

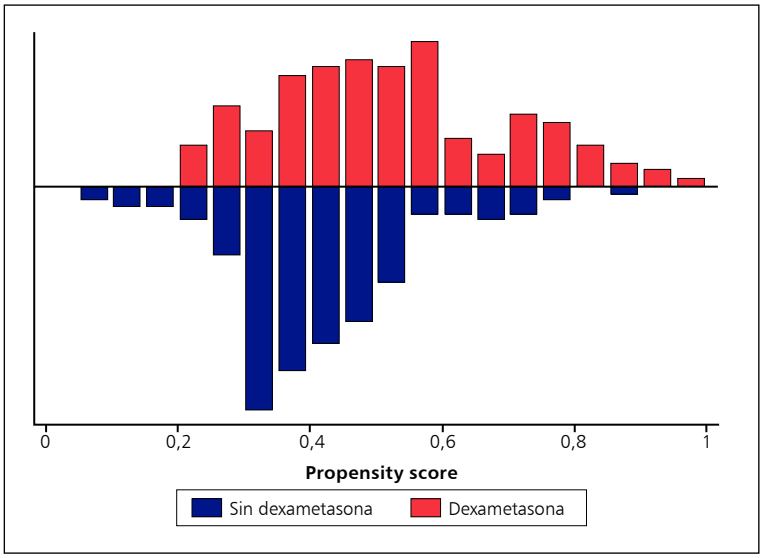

Figura 2. Puntuaciones de propensión de pacientes obtenidos de estudio de factores de riesgo para NVPO. En azul efecto de factores de riesgo de NVPO en ausencia de dexametasona intraoperatoria. En rojo efecto de factores de riesgo de NVPO en presencia de dexametasona intraoperatoria. NVPO: Náuseas y/o vómitos postoperatorios.

Tabla 2. Resultado del uso de puntuaciones de propensión

		Muestra Original			Muestra pareada por puntuación de propensión		
		Dexametasona (-) n = 226	Dexametasona (+) n = 175	p-valor	Dexametasona (-) n = 152	Dexametasona (+) n = 138	p-valor
Edad	(años)	44 ± 16	39 ± 13	< 0.05	42 ± 16	40 ± 14	0,25
Mujeres	n (%)	55 (36)	120 (48)	< 0,05	82 (54)	88 (64)	0,09
N_2O	n (%)	9 (4)	22 (12,5)	< 0,01	8 (5)	15 (11)	0,09
Ondansetrón	(mg)*	1 ± 0,7	1,5 ± 0,9	< 0,01	1,6 ± 0,6	1,7 ± 0,8	0,23

(+): presente; (-): ausente; *Uso Intraoperatorio.

tendencia a tener valores de Propensión de Puntuación más altos.

La Tabla 2 muestra lo sucedido con estas variables que diferían entre los grupos:

El grupo original era de 400 pacientes, el grupo en que se pudo parear sus puntuaciones de propensión es de 290 pacientes. El menor balance en el óxido nitroso es probablemente consecuencia del tamaño muestral que no es lo suficientemente grande, lo que constituye una desventaja de esta metodología, pero finalmente tenemos una muestra constituida por dos grupos perfectamente comparables y cuya única diferencia es el tratamiento en estudio, asimilándose esta situación a un ensayo clínico aleatorizado controlado.

Finalmente se puede investigar el efecto causal de la dexametasona en el consumo postoperatorio de ondansetrón (Tabla 3).

Tabla 3. Efecto del uso de dexametasona sobre NVPO

	Consumo postoperatorio ondansetrón (mg)	IC 95%	p-valor
Dexametasona	-0,7	-1,4; -0,04	0,039

IC: Intervalo de confianza.

En esta nueva muestra constituida por dos grupos de características equilibradas, se puede razonablemente concluir que la dexametasona

causa una disminución significativa (0,7 mg) en el consumo de ondansetrón utilizado para tratar las náuseas y vómitos postoperatorios.

Conclusión

Las puntuaciones de propensión son una herramienta estadística que permite manejar el sesgo de confusión, que inevitablemente surgirá en estudios epidemiológicos observacionales y, por lo tanto, posibilitan obtener una identificación de los efectos causales aproximable a la lograda con los Ensayos Clínicos Aleatorizados.

Referencias

1. Bakker G, Clark L. La explicación. Una introducción a la filosofía de la ciencia. México: F.C.E. 1994.
2. Höfler M. Causal inference based on counterfactuals. BMC Med Res Methodol 2005; 5: 28-40.
3. Rubin D. Estimating causal effects in randomized and non randomized studies. Journal of educational psychology 1974; 66: 688-701.
4. Coughlin S. Causal inference & scientific paradigms in epidemiology. Bentham Science Publishers 2010.
5. Hernan MA, Robins JM. Causal Inference I. Chapman & Hall CRC 2014.
6. Little R, Rubin D. Causal effects in clinical and epidemiological studies via potential Outcomes: Concepts and analytical approaches. Annu Rev Public Health 2000; 21: 121-45.
7. Neyman J. On the application of probability theory to agricultural experiments. Essay of principles. Section 9. Statistical Science 1923 (1990); 5: 465-80.
8. Crocco G, Farinas Del Cerro R, Herzig A. Conditionals: from Philosophy to Computer Science, Oxford Clarendon Press. 1995.
9. Greenland S, Pearl J, Robins J. Causal diagrams for epidemiologic research. Epidemiology 1999; 10: 37-48.
10. Pearl J. Causality. 2nd edition. NY, USA. Cambridge University Press 2009.
11. Harrel FEJ, Lee KL, Califf RM. Regression modeling strategies for improved prognostic prediction. Stat Med 1984; 3: 143-52.
12. Wilkenmayer WC, Kurth T. Propensity scores: help or hype? Nephrol Dial Transplant 2004; 19: 1672-3.
13. Stuart EA, Rubin DB. Matching methods for causal inference: Designing observational studies. Best Practices in Quantitative Methods. J Osborne Thousand Oaks, Ca: Sage Publishing. 2007.
14. Concato J, Shah N, Horwitz R. Randomized controlled trials, observational studies and the hierarchy of research designs. N Engl J Med 2000; 342: 1878-86.
15. Rothman K, Greenland S, Lash TL. Modern Epidemiology. 3d ed. Philadelphia: Lippincot Williams &Wilkins; 2008.
16. Fletcher RH, Fletcher SW, Wagner EH. Clinical Epidemiology: the essentials, 4th ed. Baltimore, USA: Wilkins & Wilkins, 2007.
17. Rosenbaum P, Rubin. The central role of propensity scores in observational studies for causal effects. Biometrika 1983; 70: 41-55.
18. Pattorno E, Grotta A, Belloco R, Schneeweeiss S. Propensity score methodology for confounding control in health care utilization databases. Epidemiology Biostatistics and Public Health 2013; 10: 8940-3.
19. Becker S, Ichino A. Estimation of average treatment effects based on propensity scores. The Stata journal 2002; 2: 358-77.
20. 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.
21. 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 Medical research methodology 2012; 12: 70-80.
22. Guo SY, Fraser MW. Propensity Score Analysis: Statistical Methods and Applications (Advanced Quantitative Techniques in Social Research). 2014.
23. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity scores methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol 2006; 59: 437-47.
24. 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.
25. Pattanayaka C, Rubina D, Zeel E. Métodos de puntuación de propensión para crear una distribución equilibrada de las covariables en los estudios observacionales. Rev Esp Cardiol 2011; 64: 897-903.
26. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity scores when the number of events is low and there are multiple confounders. Am J Epidemiol 2003; 158: 280-7.
27. Arbogast PR, Ray WA. Perfomance of Disease Risk Score, Propensity scores and traditional Multivariable Outcome Regression in the presence of multiple con-

founders. Am J Epidemiol 2011; 174: 613-20.

28. Okoli GN, Sandres RD, Myles P. Demystifying propensity scores. Br J Anaesth 2014; 112: 13-5.
29. De Oliveira GS, Santana Castro-Alves LJ, Ahmad S, Kendal MC, McCarthy RJ. Dexametasone to prevent postoperative nausea and vomiting: An updated meta-analysis of Randomized Controlled Trials. Anesth Analg 2013; 116: 58-74.
30. Apfel CC, Heidrich FM, Jukar-Rao S, Jalota L, Hornuss C, Whelan RP, et al. Evidence-based analysis of Risk factors for postoperative nausea and vomiting. Br J Anaesth 2012; 109: 742-53.