

El control de la confusión en estudios observacionales: el índice de propensión

V. Abraira

Unidad de Bioestadística Clínica. Hospital Ramón y Cajal. Madrid.

El ensayo clínico aleatorizado fue introducido en la investigación médica para evaluar los efectos de los agentes terapéuticos hace ya más de 50 años. La asignación aleatoria de cada paciente, bien al grupo del tratamiento experimental, bien al grupo control, pretende evitar la confusión introducida por el denominado sesgo de selección, o de susceptibilidad. Este sesgo surge siempre que la elección del tratamiento se hace con algún otro criterio. Por ejemplo, en la práctica clínica habitual, los médicos eligen para cada paciente el tratamiento que estiman va a tener un mejor resultado, en función de las características del paciente. Si posteriormente se compararan los resultados de dos tratamientos así asignados, los dos grupos de pacientes serían diferentes en sus características pronósticas basales y, por lo tanto, la posible diferencia en los resultados puede ser causada tanto por estas características como por el distinto efecto terapéutico de los tratamientos¹. El ejemplo más obvio de sesgo de susceptibilidad aparece cuando se compara un tratamiento quirúrgico aplicado a pacientes “operables”, con un tratamiento médico aplicado a los “no operables”.

Por ello, los ensayos aleatorizados se han convertido en el patrón de oro para la evaluación de tratamientos, hasta el punto de que, por ejemplo, un influyente libro de texto, en el capítulo de apreciación crítica de los artículos sobre evaluación de tratamientos, dice: “Si el clínico observa que el estudio no se realizó aleatoriamente, le aconsejamos que deje de leerlo y que pase al siguiente artículo”². Sin embargo, los criterios de inclusión de pacientes en los ensayos aleatorizados, por motivos tanto éticos como pragmáticos, provocan que los pacientes enrolados en estos estudios estén muy seleccionados y sean diferentes de la mayoría de los pacientes que ven los clínicos en su práctica cotidiana³. En los ensayos clínicos no suele haber ancianos, ni niños, ni mujeres embarazadas, ni pacientes con comorbilidades, que sin embargo abundan en las consultas diarias y el médico, entonces, debe aplicar su buen juicio para decidir si su paciente es tan diferente de los que

forman parte del estudio como para que los resultados no le sean aplicables⁴. Seguramente ésta es una de las fuentes de variabilidad de la práctica clínica. Por ejemplo, muchos ensayos aleatorizados han demostrado que los bloqueadores beta y los inhibidores de la enzima conversora de la angiotensina (IECA) son eficaces en el tratamiento de pacientes con enfermedad coronaria y, en consecuencia, son considerados en la actualidad el tratamiento estándar para estos pacientes³, sin embargo Shlipak et al⁵ encontraron que de 20.902 pacientes, mayores de 65 años, con la función ventricular izquierda disminuida, que habían ingresado en distintos hospitales por infarto de miocardio y sobrevivieron, el tratamiento prescrito en el momento del alta hospitalaria fue IECA (44%), bloqueadores beta (13%), ambos (16%) y ninguno (28%). Los autores argumentan como posible explicación de esta variabilidad justamente que este subgrupo de pacientes está muy poco representado en los ensayos clínicos. No obstante el objetivo de la investigación no era estudiar la variabilidad de la prescripción sino comparar los efectos de los tratamientos sobre la supervivencia un año después del alta. Según el libro antes citado, un clínico ocupado, en primera instancia, no debería leer este artículo, si bien la cita continúa: “[...] sólo en caso de no encontrar ningún ensayo aleatorizado debería retroceder y volver a examinarlo”², pero aun en este caso la pregunta es ¿en qué medida son fiables los resultados? o dicho de otra manera ¿qué parte de la posible diferencia en la supervivencia es debida a los tratamientos y qué parte a la diferencia entre las características de los pacientes en los distintos grupos? Por ejemplo, el artículo encuentra que entre los pacientes tratados con IECA había más diabéticos y más pacientes con historia previa de fallo cardíaco, y había menos que hubieran sufrido una revascularización o se les hubiera implantado un *by-pass*; entre los que no recibieron ningún tratamiento había mayor número con insuficiencia renal y con demencia. ¿En qué medida afecta esto a la mortalidad?

Hay tres maneras de controlar el sesgo de selección en estos estudios observacionales. Las tres tienen en común que sólo pueden controlar la confusión provocada por características observadas, a diferencia de la aleatorización que controla por las observadas y también por las no observadas. Son: estratificación, modelos de regresión y el índice de propensión⁶ (*propensity score*).

Correspondencia:
V. Abraira.
Unidad de Bioestadística Clínica.
Hospital Ramón y Cajal. Crta. Colmenar km 9,100.
28034 Madrid.
Correo electrónico: victor.abraigra@hrc.es

La estratificación consiste en hacer las comparaciones en estratos homogéneos con respecto al factor cuyo efecto queremos eliminar. En el ejemplo, para eliminar el efecto confusor de la diabetes, podemos dividir la muestra en dos estratos: pacientes diabéticos y pacientes no diabéticos y estimar el efecto de los tratamientos en cada estrato. El efecto ajustado por la diabetes se calcula como la media de los efectos en ambos estratos. Un ejemplo clásico de estratificación debido a Cochran y adaptado por Rubin⁷ se muestra en la tabla 1, en el que se elimina por estratificación el efecto confusor de la edad al estudiar, en datos observacionales, la mortalidad en relación con el tabaco. Obviamente para que se corrija el sesgo se necesita que en todos los estratos haya suficiente número de pacientes con los distintos tratamientos a comparar (que los tratamientos solapen con los estratos). Si la inmensa mayoría de los diabéticos, y sólo ellos, hubieran recibido IECA, no habría manera de separar el efecto causal del tratamiento, del efecto de la diabetes. Esta limitación se comparte con los otros métodos, pero una ventaja de la estratificación es que, si ocurre, el investigador lo observa inmediatamente. El principal inconveniente de la estratificación es que resulta difícil de realizar cuando se necesita ajustar por muchas variables de confusión, como suele ser el caso. Hay que tener en cuenta también que pequeñas diferencias en muchas variables, que individualmente no necesitarían ser ajustadas, pueden tener un efecto acumulativo importante. Por ejemplo, si en un grupo los pacientes tuvieran ligeramente mayor edad que en el otro, hubiera una ligera mayor proporción de diabéticos y una ligera mayor proporción de insuficiencia renal, finalmente ese grupo puede tener sustancialmente peor pronóstico. Los modelos de regresión⁸ permiten ajustar simultáneamente por muchas variables de un modo conceptualmente similar a la estratificación, son muy fáciles de usar con los paquetes estadísticos, aunque su mayor peligro estriba justamente en su facilidad para ajustar por muchas variables. Ajustando por muchas variables aumenta la probabilidad de que haya falta de solapamiento que comprometa la eficacia del control y ninguno de los resultados del análisis de regresión pone de manifiesto su existencia.

Tabla 1. Comparación de tasas de mortalidad en tres grupos de fumadores a partir de datos observacionales.

	No fumadores	Fumadores de cigarrillos	Fumadores de puro y pipa
Tasa de mortalidad por 1.000 personas-año	13,5	13,5	17,4
Edad media	57,0	53,2	59,7
Tasas ajustadas por estratificación			
Con 2 estratos	13,5	16,6	14,9
Con 3 estratos	13,5	17,7	14,2

Adaptada de Rubin DB⁷.

En el análisis sin ajustar no se observa el efecto de los cigarrillos sobre la mortalidad (los fumadores de cigarrillos tienen la misma tasa de mortalidad que los no fumadores) debido a que, en promedio, los fumadores de cigarrillos son más jóvenes. Estratificando por la edad se elimina este sesgo y se manifiesta el efecto (la tasa de mortalidad de los fumadores de cigarrillos es la más alta).

Puntos clave

- Los estudios observacionales, en los que la asignación a los grupos que se comparan no se hace aleatoriamente, están expuestos a la confusión introducida por el denominado sesgo de selección, o de susceptibilidad.
- El único modo de evitar este sesgo es la asignación aleatoria propia de los estudios experimentales, típicamente los ensayos clínicos.
- No obstante, hay métodos de análisis que permiten controlar este sesgo, entre los que destaca el índice de propensión.

El método del índice de propensión consiste en construir a partir del conjunto de variables de confusión una función de todas ellas que estime la probabilidad (“propensión”) que tienen los pacientes de ser asignados a cada grupo de tratamiento, para posteriormente estratificar por los valores de esta función. De esta forma, en cada estrato los pacientes tienen la misma probabilidad de ser asignados a cada grupo y por tanto se simula la asignación aleatoria. Hay que resaltar otra vez que con este método sólo se tienen en cuenta las variables observadas y por lo tanto puede permanecer confusión residual debida a variables no observadas. El índice de propensión se estima por regresión logística binomial⁸ y por tanto sólo sirve para comparar dos grupos. Cuando como en el artículo de Shlipak et al⁵, hay más de dos grupos se pueden realizar varias comparaciones, en el artículo se hacen tres, comparando cada tratamiento con el grupo no tratado y se encuentra beneficio de los tratamientos en todos los estratos. En el análisis realizado ajustando con modelos de regresión, se encuentra un beneficio adicional por recibir ambos fármacos, que en el análisis con el índice de propensión no aparece.

Las propiedades del método del índice de propensión se han establecido tanto teórica como empíricamente⁶ y son:

– La similitud entre los grupos con respecto a las variables incluidas en el índice es superior si se estratifica por el índice que si se estratifica por esas variables.

– La similitud es mejor que la obtenida por otros métodos de estratificación propuestos.

– La similitud es también superior a la que se conseguiría por asignación aleatoria; como contrapunto a esta llamativa propiedad hay que volver a resaltar la superioridad de la asignación aleatoria respecto a las variables no observadas y también que el criterio para la inclusión de variables en el índice es sólo la relación con la asignación de los grupos, por lo que el índice puede incluir variables no relacionadas con los resultados, y que por lo tanto no son de confusión, reduciendo así su eficacia.

– En los estudios no se usa el verdadero valor del índice sino una estimación, en general las estimaciones no se comportan tan bien como los valores que estiman⁹; sor-

prendentemente el índice de propensión estimado trabaja mejor que su verdadero valor.

BIBLIOGRAFÍA

1. Feinstein AR. Epidemiologic analyses of causation: the unlearned scientific lessons of randomized trials. *J Clin Epidemiol* 1989;42: 481-9.
2. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Medicina Basada en la Evidencia. Cómo ejercer y enseñar la MBE. Madrid: Churchill Livingstone España, 1997.
3. Lauer MS. Medical Therapy for Coronary Artery Disease Works, Even (Especially) in the Real World. *Am J Med* 2001;110:497-8.
4. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use in article about therapy or prevention. *B. What were the results and will they help me in caring for my patients?* Evidence-Based Medicine Working Groups. *JAMA* 1994; 271:59-63.
5. Shlipak MG, Browner WS, Noguchi N, Massie B, Frances CD, McClellan M. Comparison of the Effects of Angiotensin Converting-enzyme Inhibitors and Beta Blockers on Survival in Elderly Patients with Reduced Left Ventricular Function after Myocardial Infarction. *Am J Med* 2001;110:425-33.
6. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epidemiol* 1999;150:327-33.
7. Rubin DB. Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Ann Intern Med* 1997;127:S757-63.
8. Abraira Santos V, Pérez de Vargas A. Métodos multivariantes en Bioestadística. Madrid: Centro de Estudios Ramón Areces, 1996.
9. Abraira V. Estimación: intervalos de confianza. SEMERGEN 2002; 28:84-5.