Rev Esp Cardiol. 2011;64(10):897–903

Enfoque: Métodos contemporáneos en bioestadística (V)

Métodos de puntuación de propensión para crear una distribución equilibrada de las covariables en los estudios observacionales

Cassandra W. Pattanayak[a], Donald B. Rubin[a,*] y Elizabeth R. Zell[b]

[a] *Department of Statistics, Harvard University, Cambridge, Massachusetts, Estados Unidos*
[b] *Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, Estados Unidos*

Historia del artículo:
On-line el 27 de agosto de 2011

Palabras clave:
Puntuaciones de propensión
Estudios observacionales
Equilibrio de covariables

RESUMEN

La asignación aleatoria del tratamiento en los experimentos divide a los pacientes en grupos de tratamiento que están aproximadamente equilibrados en cuanto a las covariables basales. Sin embargo, en los estudios observacionales, en los que la asignación del tratamiento no es aleatoria, los pacientes de los grupos de tratamiento activo y de control difieren a menudo en covariables cruciales que están relacionadas con las variables de respuesta. Estos desequilibrios en las covariables pueden conducir a estimaciones sesgadas del efecto del tratamiento. La puntuación de propensión *(propensity score)* es la probabilidad de que a un paciente con unas características basales específicas se le asigne el tratamiento activo, y no el control. Aunque las puntuaciones de propensión son desconocidas en los estudios observacionales, al parear o subclasificar a los pacientes según las puntuaciones de propensión estimadas, podemos diseñar estudios observacionales que sean análogos a los experimentos aleatorios, con un equilibrio aproximado entre pacientes en cuanto a las covariables observadas. Los diseños de estudios observacionales basados en puntuaciones de propensión estimadas pueden producir estimaciones aproximadamente insesgadas del efecto del tratamiento. Una cuestión crucial es que los diseños de puntuación de propensión deben crearse sin tener acceso a las respuestas, imitando la separación entre el diseño del estudio y el análisis de las respuestas que es propia de los experimentos aleatorios. En este artículo se describen el marco conceptual de las respuestas potenciales para la inferencia causal y las mejores prácticas para el diseño de estudios observacionales con puntuaciones de propensión. Comentamos el uso de puntuaciones de propensión en dos estudios en los que se evaluaron la efectividad y los riesgos de los fármacos antifibrinolíticos durante las cirugías cardiacas.

Propensity Score Methods for Creating Covariate Balance in Observational Studies

Keywords:
Propensity scores
Observational studies
Covariate balance

ABSTRACT

Randomization of treatment assignment in experiments generates treatment groups with approximately balanced baseline covariates. However, in observational studies, where treatment assignment is not random, patients in the active treatment and control groups often differ on crucial covariates that are related to outcomes. These covariate imbalances can lead to biased treatment effect estimates. The propensity score is the probability that a patient with particular baseline characteristics is assigned to active treatment rather than control. Though propensity scores are unknown in observational studies, by matching or subclassifying patients on estimated propensity scores, we can design observational studies that parallel randomized experiments, with approximate balance on observed covariates. Observational study designs based on estimated propensity scores can generate approximately unbiased treatment effect estimates. Critically, propensity score designs should be created without access to outcomes, mirroring the separation of study design and outcome analysis in randomized experiments. This paper describes the potential outcomes framework for causal inference and best practices for designing observational studies with propensity scores. We discuss the use of propensity scores in two studies assessing the effectiveness and risks of antifibrinolytic drugs during cardiac surgery.

Full English text available from: www.revespcardiol.org

INTRODUCCIÓN

En un experimento aleatorio, la asignación aleatoria de los pacientes a los grupos de tratamiento activo o de control conduce a un equilibrio aproximado de estos grupos en cuanto a los parámetros basales, como edad, sexo y antecedentes patológicos. Denominamos «covariables» a estas determinaciones previas al tratamiento. El equilibrio en las covariables que crea la aleatorización permite realizar estimaciones insesgadas del efecto del tratamiento. Sin embargo, a veces los experimentos aleatorios no son factibles por razones éticas, logísticas, económicas o de otro tipo. En estas situaciones, podemos intentar el diseño de estudios

* Autor para correspondencia: Department of Statistics, Harvard University, 1 Oxford Street, 7th Floor, Cambridge, MA 02138, Estados Unidos.
Correo electrónico: rubin@stat.harvard.edu (D.B. Rubin).

0300-8932/$ – see front matter Publicado por Elsevier España, S.L. en nombre de la Sociedad Española de Cardiología.
doi:10.1016/j.recesp.2011.06.008

que sean lo más semejantes a los experimentos aleatorios con el empleo de datos observacionales (es decir, no aleatorios).

Cuando se asigna a los pacientes a los grupos de tratamiento activo o control de manera no aleatorizada, es frecuente que los grupos de tratamiento difieran de manera importante en covariables clave que estén relacionadas con las variables de respuesta. Por ejemplo, si se considera que el tratamiento activo es riesgoso para los pacientes ancianos, en términos generales los pacientes asignados al grupo control serán de mayor edad que los asignados al grupo de tratamiento activo. Una comparación sin más de las respuestas de estos grupos de tratamiento y de control observados conduciría a una estimación sesgada del efecto del tratamiento, debido al desequilibrio existente en cuanto a la edad.

Para generar estimaciones insesgadas del efecto del tratamiento utilizando datos observacionales, los pacientes deben ser agrupados («subclasificados») o pareados de manera que los pacientes tratados y los pacientes de control dentro de cada subclase o pareja estén bien equilibrados respecto a las variables clave observadas. La subclasificación o pareo mediante puntuaciones de propensión estimadas puede crear simultáneamente el equilibrio para muchas covariables observadas, lo cual conducirá a estimaciones insesgadas del efecto del tratamiento[1].

Los métodos de puntuación de propensión han ido apareciendo cada vez más en la literatura cardiológica[2–6]. Dado que las técnicas de puntuaciones de propensión no siempre se aplican de la manera correcta, en este artículo se presenta el marco conceptual en que se basan y se indican las mejores prácticas para su aplicación. En el primer apartado se presentan dos ejemplos que utilizaremos para ilustrar el diseño de estudios observacionales. En el apartado siguiente se explica el marco conceptual para las respuestas potenciales. A continuación presentamos las mejores prácticas para diseñar un estudio observacional utilizando las puntuaciones de propensión, y luego argumentamos que los modelos de regresión habituales no son apropiados para los estudios observacionales. En el último apartado, comentamos los métodos de puntuación de propensión utilizados en los dos ejemplos, seguido de una conclusión.

EJEMPLOS: APROTININA FRENTE A ÁCIDO TRANEXÁMICO

Para ilustrar la cuestión, nos centraremos en dos estudios observacionales en los que se emplearon las puntuaciones de propensión para examinar los efectos del inhibidor de serina proteasa aprotinina durante las cirugías cardiacas. Tanto Karkouti et al[5] como Mangano et al[6] compararon la pérdida hemática y las tasas de acontecimientos adversos de los pacientes tratados con aprotinina con las de los tratados con otros fármacos antifibrinolíticos, entre ellos el ácido tranexámico.

Karkouti et al[5] incluyeron en su análisis a 10.949 pacientes cardiacos del *Toronto General Hospital* que recibieron aprotinina (tratamiento activo) o ácido tranexámico (control) durante la cirugía cardiaca con *bypass* cardiopulmonar entre junio de 1999 y junio de 2004. De estos pacientes, 60 fueron excluidos a causa de su participación en otro estudio y 19 porque no habían recibido ni aprotinina ni ácido tranexámico. De los 10.870 pacientes restantes, 586 recibieron aprotinina y 10.284, ácido tranexámico.

Mangano et al[6] incluyeron en su análisis a 5.436 pacientes cardiacos de 69 hospitales de cuatro continentes, que fueron tratados con cirugía de *bypass* arterial coronario entre noviembre de 1996 y junio de 2000[7]. Los pacientes no recibieron ningún fármaco antifibrinolítico o fueron tratados con aprotinina, ácido tranexámico o ácido aminocaproico. De los pacientes que cumplían los criterios adicionales de elegibilidad, 1.295 recibieron aprotinina y 822 recibieron ácido tranexámico.

MARCO CONCEPTUAL DE LAS RESPUESTAS POTENCIALES PARA LA INFERENCIA CAUSAL

Respuestas potenciales

Limitaremos nuestra exposición a los estudios que comparan dos opciones de tratamiento, aunque el marco conceptual puede ampliarse a más de dos tratamientos. Para cada paciente, existe una respuesta potencial (p. ej., acontecimiento adverso grave o no) que sería observada si el paciente fuera asignado al tratamiento activo y una respuesta potencial que podría observarse si el paciente fuera asignado al grupo control. El problema fundamental de la inferencia causal es que solamente puede observarse una respuesta potencial en cada paciente, puesto que cada paciente es asignado al tratamiento activo o al control, pero no a ambos[8,9]. En consecuencia, la inferencia causal es un problema de faltantes: el objetivo es completar las respuestas potenciales faltantes estimando lo que habría ocurrido a cada paciente si lo hubieran asignado al otro grupo de tratamiento.

Cualquier estimación del efecto del tratamiento asume, implícita o explícitamente, un valor para cada respuesta potencial disponible. El estimador más sencillo y poco elaborado del efecto del tratamiento es la diferencia de las respuestas para los grupos de tratamiento activo y control. Este método supone, implícitamente, que las respuestas potenciales faltantes para el tratamiento activo en los pacientes asignados al control son iguales a la media de las respuestas observadas en el grupo de tratamiento activo, y que las respuestas potenciales faltantes para el control en los pacientes asignados al tratamiento activo son iguales a la media de las respuestas observadas en el grupo control.

El uso de las medias de las respuestas observadas hace que la estimación de los posibles respuestas faltantes esté justificada si el tratamiento se asigna de modo completamente aleatorio. De lo contrario, se debe estimar las respuestas potenciales faltantes de una manera que incluya el proceso de toma de decisión en la asignación del tratamiento activo o el control.

Mecanismo de asignación y puntuaciones de propensión

El mecanismo de asignación es el proceso de decisión que se emplea para asignar a unos pacientes al tratamiento activo y a otros al grupo control. La puntuación de propensión de cada paciente es la probabilidad de que el paciente hubiera sido asignado al tratamiento activo en vez de al control, dadas sus covariables. En un experimento aleatorio, se conoce la puntuación de propensión de cada paciente. Por ejemplo, en un experimento completamente aleatorio en el que la mitad de los pacientes son asignados a cada grupo de tratamiento, la puntuación de propensión de cada paciente es de 1/2. Una comparación simple de las respuestas observadas en los grupos de tratamiento y control no estaría sesgada en este caso.

En un experimento con un diseño de bloques al azar, los pacientes se agrupan en función de unas covariables observadas similares, y la probabilidad de asignación del tratamiento activo puede ser diferente en los pacientes de cada bloque. Por ejemplo, si el tratamiento activo se considera de mayor riesgo para los pacientes ancianos, a los que tengan más de 65 años de edad se les puede asignar el tratamiento activo con una probabilidad de 0,4, y a los de 65 años o menos se les puede asignar el tratamiento activo con una probabilidad de 0,7. Una comparación simple, sin más, de las respuestas observadas en los grupos de tratamiento activo y de control estaría sesgada, puesto que el grupo de tratamiento activo contendría un número desproporcionado de pacientes de menos edad. Para generar estimaciones insesgadas del efecto del tratamiento, deberíamos comparar a pacientes asignados al

tratamiento activo con otros asignados al control dentro de cada grupo de edad. Los pacientes de cada grupo de edad tienen la misma puntuación de propensión. Al estimar los efectos del tratamiento dentro de cada grupo de edad, llenamos implícitamente la respuesta potencial faltante de cada paciente basándonos en las respuestas observadas de los demás pacientes del mismo grupo de edad.

En un estudio observacional, continúa siendo cierto que la agrupación de pacientes con puntuaciones de propensión similares conduce a un efecto insesgado de las estimaciones del tratamiento. Sin embargo, la probabilidad de que cualquier paciente concreto sea asignado al tratamiento activo y no al control, dados los valores de las covariables, no se conoce cuando el tratamiento se asigna de manera no aleatoria. El investigador puede estar razonablemente satisfecho en cuanto a que todas las covariables que podrían afectar a la asignación del tratamiento han sido incluidas en el conjunto de datos. En este caso, denominamos al mecanismo de asignación mecanismo sin confusión, y podemos estimar las puntuaciones de propensión desconocidas en función de esas covariables observadas. Al comparar a pacientes con puntuaciones de propensión estimadas similares, podemos diseñar un estudio observacional que se asemeje a un experimento aleatorio.

En el estudio descrito por Karkouti et al[5], el proceso de toma de decisiones para la asignación de un fármaco antifibrinolítico se basó en las guías conocidas. Se indicó a los médicos del *Toronto General Hospital* que utilizaran aprotinina solamente en un subgrupo de pacientes de alto riesgo y que emplearan ácido tranexámico en los demás casos. Dado que las guías informan las decisiones de tratamiento pero no las determinan, hay un subgrupo de pacientes que podrían haber recibido aprotinina o ácido tranexámico. Este subgrupo de pacientes que tenían cierta probabilidad de recibir uno u otro tratamiento es necesario para el diseño de un estudio observacional que pueda conducir a una estimación insesgada del efecto del tratamiento. Las guías del hospital proporcionan un punto de partida útil para estimar el mecanismo de asignación y las puntuaciones de propensión.

Dado el amplio ámbito geográfico del estudio de Mangano et al[6], es probable que el mecanismo de asignación sea más complicado. El proceso de toma de decisiones respecto al tratamiento puede haber funcionado de manera distinta en cada una de las 69 instituciones participantes en el estudio.

DISEÑO DE UN ESTUDIO OBSERVACIONAL

Identificación del momento de la asignación del tratamiento

El primer paso para el diseño de un estudio observacional es identificar el momento de asignación del tratamiento. En un experimento aleatorio, suele ser fácil identificar el momento en que se asigna aleatoriamente a cada paciente el tratamiento activo o el control lanzando una moneda al aire, abriendo un sobre, con un ordenador, etc. Establecer exactamente el momento de la asignación del tratamiento en un estudio observacional puede ser más difícil. Si el médico opta por el tratamiento activo o por el control en un paciente concreto, el momento de esa decisión es el momento de la asignación del tratamiento. Otra posibilidad es que el paciente se haya autoseleccionado para el grupo de tratamiento activo o de control, y es preciso identificar el momento de esa decisión en relación con las demás medidas observadas.

El momento de la asignación del tratamiento es importante porque nos permite diferenciar las covariables previas al tratamiento («apropiadas») de las determinaciones posteriores a este («inapropiadas»). Las covariables previas al tratamiento, apropiadas, se miden o podrían medirse antes de la asignación del tratamiento. Los antecedentes médicos previos a la decisión de tratar constituyen una covariable apropiada. La edad y el sexo son también covariables, aun cuando se registren, de hecho, después de la decisión de tratar, puesto que ni una ni otra pueden verse afectadas por el tratamiento.

Cualquier otra información obtenida después de la asignación del tratamiento constituye una respuesta. Las respuestas primarias pueden ser la muerte, la pérdida hemática, los acontecimientos adversos, etc. La presión arterial de un paciente el día siguiente al de su autoselección para el grupo de tratamiento activo o de control es otra respuesta y no una covariable, aun en el caso de que el efecto del tratamiento sobre esta determinación de la presión arterial no sea de interés.

Un diseño de estudio observacional debe crear un equilibrio respecto a las covariables previas al tratamiento, ya que, en promedio, en un experimento la aleatorización las equilibraría. Sin embargo, no debemos intentar equilibrar estas determinaciones posteriores al tratamiento, puesto que estas podrían estar influidas por el tratamiento recibido. Esta distinción es crucial: la clasificación errónea de una variable de respuesta que pudiera ser afectada por el tratamiento considerándola una covariable apropiada puede enmascarar el efecto del tratamiento.

Por ejemplo, consideremos un estudio en el que se comparen dos fármacos antifibrinolíticos administrados durante la cirugía cardiaca, en el que la respuesta de interés sea la hemorragia dos días después de la cirugía (día 2). Supongamos que la hemorragia un día después de la cirugía (día 1) predice claramente la hemorragia en el día 2. Si la hemorragia en el día 1 se clasifica erróneamente considerándola una covariable apropiada, agruparemos a los pacientes según la hemorragia hasta el día 1. Dada la intensa correlación existente entre la hemorragia en el día 1 y en el día 2, la agrupación de los pacientes según la hemorragia hasta el día 1 enmascara el efecto real del tratamiento: de los pacientes con hemorragia el día 1, la mayoría tendría hemorragia el día 2, independientemente de la asignación del tratamiento. Entre los pacientes sin hemorragia el día 1, la mayoría no tendría hemorragias el día 2, independientemente de la asignación del tratamiento. Aun en el caso de que haya un efecto importante del tratamiento en comparación con el control, tanto en la hemorragia del día 1 como en la del día 2, el condicionar equivocadamente con respecto a la hemorragia en el día 1 conduce a una estimación de ausencia de efecto porque la hemorragia del día 1 predice la del día 2.

Dado que los fármacos antifibrinolíticos se transmiten durante la cirugía, es probable que las decisiones de tratamiento de los estudios publicados por Karkouti et al[5] y Mangano et al[6] se tomaran antes de la cirugía. Ambos estudios fueron condicionados con respecto a medidas que podrían haberse visto afectadas por el fármaco antifibrinolítico. Varios de los indicadores de medicación considerados en los modelos generados por Mangano et al[6] se clasifican como intraoperatorios[7,10]. El modelo de puntuación de propensión en el estudio de Karkouti et al[5] incluía la duración del *bypass* cardiopulmonar, que podría estar influida por un fármaco aplicado al inicio de la intervención quirúrgica.

Separación de diseño y análisis

El protocolo de aleatorización para un experimento se ha finalizado necesariamente antes de la obtención de las respuestas. Con objeto de que sea semejante a un experimento aleatorio, el diseño de un estudio observacional debiera separarse de modo similar del análisis de las respuestas. Hay que extraer las respuestas del conjunto de datos antes de iniciar el diseño del estudio, en cuanto se ha identificado el momento de la asignación del tratamiento[11,12]. Separar el diseño del estudio observacional del análisis de las respuestas aporta protección contra el sesgo real o sospechado del investigador.

Identificación y priorización de covariables

Antes de diseñar un estudio observacional, y si es posible antes de obtener los datos, los expertos en el campo deben identificar las covariables que podrían predecir la decisión de tratamiento y/o las respuestas. Obsérvese que, para preservar la objetividad, este debate debe realizarse sin tener acceso a datos de respuestas del estudio en cuestión, si bien la literatura previa puede ayudar a orientar la selección de las covariables. Si la decisión de tratamiento puede haber sido influenciada por una covariable que no se ha obtenido o no está disponible por algún otro motivo, será imposible determinar si los grupos de tratamiento están equilibrados respecto a esa covariable, y el conjunto de datos puede no ser útil para abordar la cuestión planteada en el estudio. Un mecanismo de asignación de este tipo es un mecanismo con confusión, dadas las covariables observadas.

Si se dispone de todas las covariables que se cree que son importantes en relación con la decisión de tratamiento y con las respuestas, esas covariables deben dividirse en grupos de prioridad. De manera similar a la de un diseño experimental aleatorio, un diseño de estudio observacional conducirá a un mejor equilibrio respecto a algunas variables que respecto a otras. La priorización de las covariables sirve de guía para comparar diversos diseños observacionales propuestos.

Las covariables clave que a menudo se pasan por alto en los estudios médicos incluyen la fecha de inclusión y el centro clínico. Karkouti et al[5] indicaron una tendencia de la probabilidad de tratamiento a lo largo del tiempo. Sin embargo, como apuntan los autores, la fecha de inclusión no se tuvo en cuenta como covariable. Cuando se tarda cierto tiempo en obtener los datos, los avances médicos y los cambios de las guías pueden afectar a las respuestas de los pacientes, y puede ser importante comparar a pacientes con fechas de inclusión similares.

Los 69 centros distintos que participaron en el estudio de Mangano et al[6] pueden haber diferido en varias formas que probablemente predigan las variables de respuestas, como la formación del personal, los protocolos, el equipo y las influencias culturales. El diseño del estudio podría haber mejorado condicionando con respecto a los múltiples centros clínicos.

Discusión acerca del desequilibrio en una sola covariable

Subclasificación en una sola covariable

La subclasificación de los pacientes respecto a una sola covariable categórica es sencilla. Por ejemplo, si un estudio observacional incluye tanto a varones como a mujeres y se espera que el sexo prediga las respuestas, el efecto del tratamiento activo frente al control puede estimarse por separado en varones y mujeres. Las estimaciones del efecto del tratamiento dentro de cada sexo pueden promediarse para estimar el efecto conjunto del tratamiento en la población. La subclasificación respecto a una sola covariable elimina el sesgo debido a esa covariable: la respuesta potencial faltante que se habría observado con el tratamiento activo en un varón que en realidad ha recibido el control se estima utilizando las respuestas observadas para los varones solamente, en vez de utilizar la muestra completa.

Este enfoque se extiende, de una forma sencilla, al caso de una sola variable continua. Los pacientes podrían subclasificarse, por ejemplo, en función de grupos de edad. Habitualmente, cinco subclases son suficientes para reducir el 90% del sesgo respecto a una única covariable continua[13].

A menudo, algunos pacientes de un grupo de tratamiento son diferentes de cualquiera de los pacientes del otro grupo de tratamiento en cuanto a una covariable clave. Por ejemplo, los pacientes de edad superior a 65 años pueden no haber sido elegibles para el tratamiento activo o uno de los centros clínicos de un estudio multicéntrico puede haber prescrito el tratamiento activo a todos los pacientes. No hay información útil disponible para imputar las respuestas potenciales faltantes de esos pacientes: ¿qué habría ocurrido a los pacientes de más de 65 años si se les hubiera asignado el tratamiento activo? ¿Qué habría ocurrido a los pacientes del centro que utiliza el tratamiento activo en todos los casos si se les hubiera asignado el control? Los pacientes para los que no hay una contrapartida en el otro grupo de tratamiento deben ser excluidos del conjunto de datos, puesto que el estudio no puede estar diseñado para generar estimaciones útiles del efecto del tratamiento en esos casos.

Emparejar respecto a una sola covariable

Muchos estudios observacionales incluyen un grupo relativamente pequeño de pacientes que han recibido el tratamiento activo y un conjunto amplio de pacientes de control que no han recibido el tratamiento activo. Los pacientes del grupo control pueden proceder de una base de datos de vigilancia o de una fuente distinta a la del grupo tratado. Es común que la mayoría de los pacientes de control tengan valores de las covariables muy diferentes de los de los pacientes tratados y no se habrían incluido si los datos se hubieran recogido con la finalidad de abordar esa pregunta concreta de investigación. En esta situación, puede identificarse a un paciente de control coincidente con cada paciente del grupo de tratamiento activo basándose en una covariable importante, con lo que se crea un diseño de pares equiparados que se aproxima a un experimento de pares aleatorios. Los controles candidatos que no son emparejdos pueden ser descartados. El diseño de pares equiparados lleva a estimaciones insesgadas del efecto del tratamiento en los pacientes con valores de la covariable similares a los del grupo de tratamiento activo. La respuesta observada en cada paciente de control equiparado se utiliza para estimar la respuesta potencial faltante de un paciente tratado pareado con él.

Es crucial tomar en cuenta que el diseño de pares equiparados que describimos difiere de un modo fundamental de un estudio de casos y controles (o, para evitar la confusión, un «estudio de casos/no casos»). En un estudio de casos/no casos, un paciente con una respuesta positiva se empareja con un paciente con una respuesta negativa; ambos pacientes pueden haber recibido tratamiento activo o ambos pueden haber recibido el tratamiento de control. Esta equiparación se basa en la observación de la respuesta y no es análogo a algún diseño experimental aleatorio. En el diseño de pares equiparados que describimos, un paciente que ha recibido tratamiento activo es emparejado con un paciente que ha recibido el tratamiento de control. Emparejar a pacientes con tratamiento activo o tratamiento de control no requiere de las variables de respuesta y es análogo a un experimento aleatorio en el que los pares de pacientes con covariables observadas similares son asignados aleatoriamente, uno al tratamiento activo y el otro al control.

Naturalmente, en la mayor parte de los estudios, se prevé que haya más de una covariable que esté relacionada con las variables respuesta. Puede desearse un equilibrio de las covariables en cuanto a edad, sexo, diversos indicadores de los antecedentes patológicos, etc. Emparejar o subclasificar simultáneamente a los pacientes para múltiples covariables rápidamente deja de ser manejable: con cinco grupos de edad, dos sexos y cinco indicadores binarios para los trastornos médicos previos, serían necesarias 320 subclases diferentes. Con cinco indicadores binarios más para otras características demográficas o trastornos médicos previos, serían precisas más de 10.000 subclases. El objetivo de estimar puntuaciones de propensión es simplificar este proceso y crear un equilibrio aproximado para muchas covariables a la vez.

Parear o subclasificar con puntuaciones de propensión estimadas

Aunque las puntuaciones de propensión verdaderas son desconocidas en los estudios observacionales, pueden estimarse al modelar la probabilidad de asignación al tratamiento activo, dadas las covariables observadas, sin tener acceso a las respuestas[1]. Lo más frecuente es que las puntuaciones de propensión se estimen mediante una regresión logística[12]. Los valores ajustados obtenidos de la regresión logística son las puntuaciones de propensión estimadas.

De la misma manera que cada paciente tiene una edad y un sexo, cada paciente tiene también una puntuación de propensión estimada, es decir, un único valor entre 0 y 1 que corresponde a la probabilidad de que a alguien con las covariables del paciente se le hubiera asignado el tratamiento activo y no el control. Al parear o subclasificar a los pacientes con puntuaciones de propensión estimadas similares, puede establecerse un equilibrio aproximado para todas las covariables incluidas en el modelo de puntuación de propensión[1,14,15].

El éxito del modelo de puntuación de propensión y el método de parear o subclasificar debe evaluarse mediante la verificación explícita del equilibrio de covariables en el diseño propuesto. Si los pacientes tratados y los pacientes de control se emparejaran respecto a unas puntuaciones de propensión estimadas similares, podríamos verificar que los pacientes emparejados fueran suficientemente similares en cuanto a edad, historial clínico, etc. Si se clasificara a los pacientes según las puntuaciones de propensión estimadas y se los dividiera en subclases en función de unos valores de corte de la puntuación de propensión estimada, podríamos verificar que los pacientes con tratamiento activo y de control de cada subclase tuvieran valores de covariables similares. Las medias de las covariables observadas deberían ser aproximadamente iguales en los grupos de tratamiento activo y control tras equipararlos o dentro de cada subclase y al promediar las diversas subclases. Las varianzas, rangos, *logs* y cuadrados de las variables continuas deberían estar equilibrados, y las interacciones entre las covariables deberían estar equilibradas también.

Dado que las respuestas se separan del conjunto de datos durante el proceso de diseño, podemos realizar una iteración entre la estimación de la puntuación de propensión creando subclases o pares y verificando el equilibrio de las covariables. Si una determinada covariable no está suficientemente equilibrada tras el primer diseño propuesto, un modelo de puntuación de propensión revisado podría incluir interacciones entre esa covariable y otras covariables, o el *log* o el cuadrado de esa covariable si es continua. La elección de un determinado conjunto de subclases o pares requiere compromisos: algunos diseños propuestos alcanzarán un mejor equilibrio respecto a determinadas variables y un equilibrio menos adecuado respecto a otras. Los grupos de prioridad de covariables deben servir de guía para comparar los posibles diseños de estudio.

Habitualmente bastan cinco subclases de puntuación de propensión basadas en los quintiles de las puntuaciones de propensión estimadas para reducir el 90% del sesgo en todas l as covariables utilizadas en el modelo de puntuación de propensión[14]. Si el tamaño muestral es grande o si algunas covariables no están suficientemente equilibradas, pueden crearse más de cinco subclases. Cuando los tamaños muestrales relativos del grupo de tratamiento activo y el grupo control son tales que hacen que parear sea más apropiado que subclasificar, emparejar a cada uno de los pacientes del grupo de tratamiento activo con el paciente de control que tiene una puntuación de propensión estimada más similar conduce habitualmente a un equilibrio aproximado de las covariables[1,15,16]. Pero si el equilibrio en el diseño propuesto no es satisfactorio, el estudio puede limitarse a los pares de pacientes situados a menos de una determinada distancia máxima entre sí para la puntuación de propensión estimada.

Es importante señalar que un diseño de estudio observacional propuesto no debe evaluarse en función de lo bien que el modelo de puntuación de propensión se ajusta a los datos o lo bien que el modelo de puntuación de propensión describe el proceso real de toma de decisiones. La estimación del modelo de puntuación de propensión es un paso en la dirección de establecer subclases o pares bien equilibrados, y el mejor modelo de puntuación de propensión es aquel que lleva al diseño con un mejor equilibrio de las covariables.

Un diseño de estudio observacional riguroso requiere limitar el estudio a una submuestra bien definida de los datos en la que algunos pacientes hayan recibido el tratamiento activo y otros el control, como en un experimento aleatorio. Si las covariables incluidas en el modelo de puntuación de propensión están intensamente relacionadas con la asignación del tratamiento, algunos pacientes pueden presentar puntuaciones de propensión estimadas extremas que estén fuera de la gama de puntuaciones de propensión estimadas de los pacientes del otro grupo de tratamiento. Esta situación es semejante a la falta de traslape en una única covariable: no hay información para estimar las respuestas potenciales faltantes para los pacientes situados fuera de la gama de traslape de las puntuaciones de propensión estimadas. A menudo es posible determinar los valores de covariables que caracterizan a los pacientes con puntuaciones de propensión extremas (p. ej., tal vez los varones de una edad inferior a determinado umbral hayan recibido siempre tratamiento activo y, por lo tanto, tengan puntuaciones de propensión estimadas muy altas). Los pacientes que cumplen estos criterios deben ser retirados del estudio. La exclusión de pacientes en base a los valores de covariables en lugar de las puntuaciones de propensión estimadas simplifica la posibilidad de generalizar las respuestas del estudio.

Dado que no se dispone de respuestas durante el diseño del estudio observacional, se puede y se debe comunicar los pares o las subclases propuestas a los clínicos y otras partes interesadas para discutirlas y aprobarlas. Cualquier objeción que se plantee al equilibrio en las covariables observadas en el diseño propuesto debe abordarse antes del análisis de las respuestas. Este proceso es similar al de la solicitud de aprobación de un ensayo clínico aleatorio antes de iniciar la inclusión de pacientes: en ausencia de variables de respuesta, la modificación del diseño del estudio no puede sesgar la estimación final del efecto del tratamiento.

Una vez finalizado el diseño, se puede analizar las respuestas. En un diseño de pareo en el que los pacientes asignados al tratamiento activo han sido emparejados con pacientes asignados al control, las respuestas observadas en los grupos emparejados de tratamiento y de control pueden compararse directamente. En un diseño con subclasificación, las respuestas observadas con el tratamiento activo y con el control pueden compararse dentro de cada subclase, y puede obtenerse una estimación general mediante la media ponderada de las estimaciones del efecto del tratamiento en cada subclase.

PELIGROS DE LA REGRESIÓN EN LOS ESTUDIOS OBSERVACIONALES

La regresión, también denominada ajuste de covarianza, se utiliza con frecuencia para abordar el desequilibrio de las covariables en los estudios observacionales. Los investigadores que aplican métodos de regresión incluyen a menudo la variable indicadora del tratamiento y las covariables observadas en un modelo para predecir las respuestas observadas. Sin embargo, si las covariables no están bien equilibradas inicialmente, es probable que este ajuste de regresión se base en supuestos no válidos y a veces puede aumentar el sesgo en vez de reducirlo[1,17,18]. A menos que las respuestas puedan predecirse con exactitud a partir de las

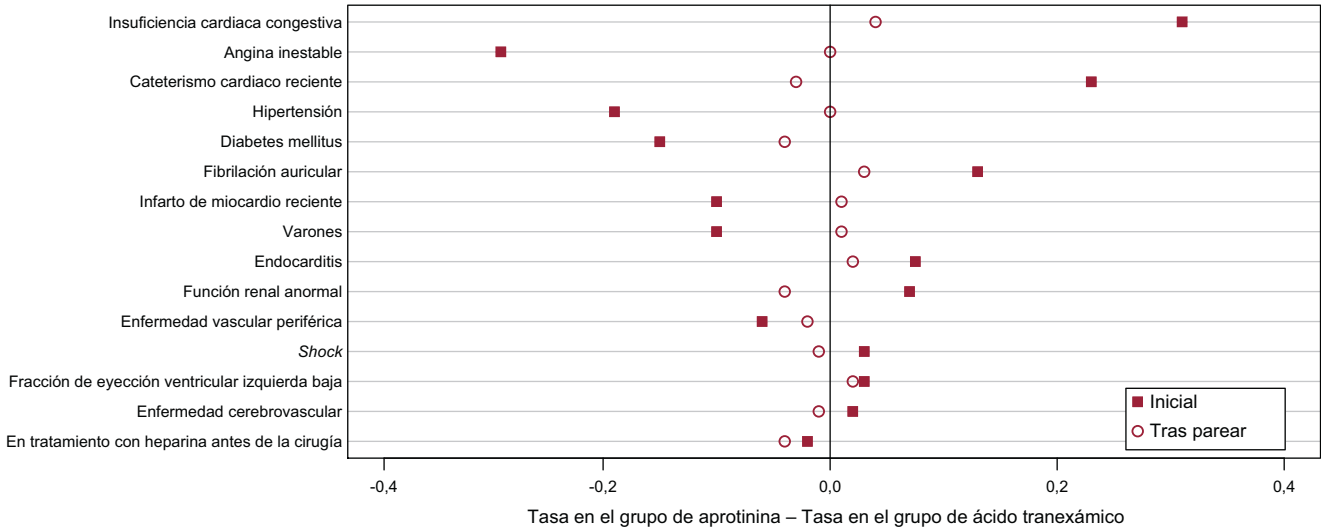

Figura 1. Diferencias en las tasas de covariables binarias relativas al paciente entre los grupos de tratamiento con aprotinina y con ácido tranexámico antes y después de parear en el estudio de Karkouti et al[5].

covariables utilizando líneas rectas y que el efecto del tratamiento sea el mismo en cada paciente, las estimaciones de las respuestas potenciales faltantes que la regresión implica pueden carecer de significado o ser confusas.

Como consecuencia de los importantes supuestos de la modelización, la regresión genera estimaciones del efecto del tratamiento a pesar de que el sentido común indique que la información existente es insuficiente. Por ejemplo, aun en caso de que el paciente de mayor edad que haya recibido tratamiento activo tenga 30 años, el programa informático de regresión extrapolará las respuestas (generalmente basándose en una línea recta) para estimar lo que habría sucedido en un paciente de 80 años del grupo control si hubiera recibido el tratamiento activo.

La regresión conduce a menudo a unos intervalos de confianza relativamente estrechos para el efecto del tratamiento. Aunque los intervalos estrechos son deseables cuando se prevé que el intervalo esté centrado alrededor del efecto real del tratamiento, los ajustes de regresión en los estudios observacionales conducen a menudo a unos intervalos engañosamente pequeños alrededor de un efecto del tratamiento erróneo. Los intervalos estrechos reflejan los supuestos de la modelización (habitualmente no válidas) y no la información aportada por los datos.

Las estimaciones de la regresión son sensibles a los tamaños muestrales relativos de los grupos de tratamiento activo y de control observados. Si el conjunto de los datos incluye un conjunto relativamente pequeño de pacientes tratados y un grupo amplio de controles, el modelo de regresión estará determinado principalmente por la relación entre las variables de respuesta y las covariables observadas en los pacientes de control, aun en caso de que la mayoría de esos pacientes de control no se parezcan en nada a los pacientes que han recibido el tratamiento activo.

El defecto más importante del ajuste de regresión para la inferencia causal en los estudios observacionales es que el diseño del estudio no se separe del análisis de las respuestas. ¿Con qué frecuencia los investigadores aplican solamente un modelo de regresión? Resulta tentador intentar pescar alguna respuesta aplicando un ajuste de varios modelos hasta que aparece la respuesta deseada o esperada. Dado que las respuestas y las covariables no se separan de manera explícita, también es fácil no considerar el momento de la asignación del tratamiento e incluir en el modelo de regresión como predictores variables que en realidad son respuestas.

Los modelos de regresión son apropiados a veces como parte del análisis de la variable respuesta, tras haber finalizado un diseño de casos pareados o subclasificados. Si hay un equilibrio respecto a las covariables observadas, la estimación del efecto del tratamiento está aproximadamente insesgada con o sin regresión, y la regresión puede ser una forma eficaz de producir intervalos más estrechos alrededor de la respuesta correcta.

PUNTUACIONES DE PROPENSIÓN PARA COMPARAR APROTININA CON ÁCIDO TRANEXÁMICO

Pareo en el estudio de Karkouti et al

En el trabajo de Karkouti et al[5], los pacientes que recibieron aprotinina en vez de ácido tranexámico tenían mayor probabilidad de ser mujeres, no presentar antecedentes de angina inestable, hipertensión o diabetes mellitus y tener antecedentes de insuficiencia cardiaca congestiva, cateterismo cardiaco reciente o fibrilación auricular, entre otras covariables. Karkouti et al[5] crearon pares equiparados de pacientes tratados con aprotinina o con ácido tranexámico utilizando puntuaciones de propensión estimadas para establecer un equilibrio respecto a las covariables observadas.

Las puntuaciones de propensión se estimaron con un modelo de regresión logística que predecía el estado de tratamiento a partir de 20 covariables observadas, incluidas varias interacciones (al menos una de estas covariables puede haberse medido tras el tratamiento). Los autores identificaron parejas en el grupo de ácido tranexámico para 449 de los 586 pacientes tratados con aprotinina basándose en los valores similares de las puntuaciones de propensión estimadas, y descartaron a 137 pacientes del grupo de aprotinina que no fueron emparejados, ya que no eran similares a los pacientes tratados con ácido tranexámico respecto a las covariables observadas.

En la figura 1 se muestran las diferencias en las tasas de las covariables binarias, relacionadas con el paciente, entre los grupos de aprotinina y ácido tranexámico, antes y después del pareo. El equilibrio respecto a esas covariables observadas es mucho mejor tras el pareo que antes. Concretamente, la angina inestable en un plazo de 30 días respecto a la cirugía fue menos frecuente en el grupo de aprotinina inicial que en el grupo de ácido tranexámico inicial, con una diferencia de aproximadamente 30 puntos porcentuales. Sin embargo, las tasas de angina inestable en los pacientes del grupo pareado de aprotinina y en los pacientes del grupo pareado de ácido tranexámicoson similares. La insuficiencia cardiaca congestiva fue más frecuente en el grupo de aprotinina

inicial que en el grupo de ácido tranexámico inicial, con una diferencia de aproximadamente 30 puntos porcentuales, pero las tasas de insuficiencia cardiaca congestiva son similares en los grupos pareados de aprotinina y de ácido tranexámico. Obsérvese también que el desequilibrio entre los pacientes tratados con aprotinina y los pacientes tratados con ácido tranexámico en cuanto al uso de heparina antes de la cirugía, en realidad, aumentó a causa del pareo. La elección de un conjunto final de parejas requiere una priorización de las covariables observadas, y podría haberse elegido un conjunto de pares diferente si el desequilibrio en cuanto al uso de heparina hubiera sido considerado inaceptable durante el diseño del estudio.

La posibilidad de generalizar los resultados del estudio se limita a la población de pacientes con los valores de las covariables similares a las de los pacientes pareados. Los pacientes pareados son de mayor edad, tienen mayor probabilidad de presentar hipertensión y angina inestable. Es menos probable que recientemente se les haya practicado un cateterismo cardiaco o que hayan sufrido una endocarditis. Además, los pacientes pareados tienden a tener concentraciones de hemoglobina más altas que las del grupo inicial de pacientes que recibieron aprotinina. Dado que para los pacientes de máximo riesgo que cumplían claramente los criterios del hospital para el uso de aprotinina no hay muchos pacientes con los que se los pueda emparejar en el grupo de ácido tranexámico, los pacientes pareados que fueron tratados con aprotinina tienen un estado de salud algo peor que los del grupo de aprotinina inicial.

Karkouti et al[5] observaron unas tasas similares de transfusiones y de acontecimientos adversos en los pacientes emparejados que fueron tratados con aprotinina y con ácido tranexámico, excepto porque la disfunción renal se produjo con una frecuencia significativamente mayor en los pacientes pareados que fueron tratados con aprotinina que en los pacientes pareados tratados con ácido tranexámico.

Regresión en el estudio de Mangano et al

En el trabajo de Mangano et al[6], los pacientes con antecedentes de insuficiencia cardiaca congestiva, enfermedad pulmonar o valvulopatía, entre otras covariables, parecen haber tenido mayor probabilidad de recibir aprotinina que de ser tratados con ácido tranexámico. En vez de crear parejas o subclases basadas en las puntuaciones de propensión estimadas, Mangano et al[6] ajustaron un modelo de regresión a las respuestas observadas respecto a las puntuaciones de propensión estimadas. Los resultados obtenidos de la regresión de las respuestas observadas sobre la puntuación de propensión estimada, son muy similares a los resultados obtenidos de la regresión directa sobre las covariables incluidas en el modelo de puntuación de propensión[1]. Los estadísticos[17] han propuesto alguna vez este uso de las puntuaciones de propensión estimadas, pero luego ha sido corregido[19]. La regresión respecto a las puntuaciones de propensión estimadas tiene los mismos inconvenientes que el ajuste de regresión comentado anteriormente, y se recomienda en su lugar el uso de las puntuaciones de propensión estimadas para crear parejas o subclases como parte del diseño del estudio en vez del análisis[20].

Mangano et al[6] llegaron a la conclusión de que la aprotinina y el ácido tranexámico resultaban en una pérdida similar de sangre, pero que la aprotinina estaba asociada con un mayor riesgo de insuficiencia renal, infarto de miocardio o insuficiencia cardiaca, así como el de ictus o encefalopatía.

CONCLUSIONES

El emparejamiento o la subclasificación basados en puntuaciones de propensión estimadas pueden conducir a un equilibrio aproximado respecto a las covariables observadas entre los grupos de tratamiento activo y de control en los estudios observacionales. Es crucial tener en cuenta que los estudios observacionales deben ser diseñados sin tener acceso a los datos de las variables de respuesta. Al diseñar estudios observacionales que sean análogos a los experimentos aleatorios, podemos generar estimaciones insesgadas de los efectos del tratamiento, a pesar de la asignación no aleatoria de los pacientes a los grupos de tratamiento.

AGRADECIMIENTOS

Los autores están muy agradecidos a Valeria Espinosa Mateos por su generosa colaboración en la versión española de este artículo.

CONFLICTO DE INTERESES

Ninguno.

BIBLIOGRAFÍA

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
2. Chikwe J, Goldstone AB, Passage J, Anyanwu AC, Seeburger J, Castillo JG, et al. A propensity score-adjusted retrospective comparison of early and mid-term results of mitral valve repair versus replacement in octogenarians. Eur Heart J. 2011;32:618–26.
3. Charlot M, Grove EL, Hansen PR, Olesen JB, Ahlehoff O, Selmer C, et al. Proton pump inhibitor use and risk of adverse cardiovascular events in aspirin treated patients with first time myocardial infarction: nationwide propensity score matched study. BMJ. 2011;342:d2690.
4. Ahmed A, Husain A, Love TE, Gambassi G, Dell'Italia LJ, Francis GS, et al. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. Eur Heart J. 2006;27:1431–9.
5. Karkouti K, Beattie WS, Dattilo KM, McCluskey SA, Ghannam M, Hamdy A, et al. A propensity score case-control comparison of aprotinin and tranexamic acid in high-transfusion-risk cardiac surgery. Transfusion. 2006;46:327–38.
6. Mangano DT, Tudor IC, Dietzel C. The risk associated with aprotinin in cardiac surgery. N Engl J Med. 2006;354:353–65.
7. Mangano DT, Miao Y, Vuylsteke A, Tudor IC, Juneja R, Filipescu D, et al. Mortality associated with aprotinin during 5 years following coronary artery bypass graft surgery. JAMA. 2007;297:471–9.
8. Holland PW. Statistics and causal inference. J Am Stat Assoc. 1986;81:945–60.
9. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688–701.
10. Ischemia Research and Education Foundation Aprotinin and Long Term Mortality, Appendix 1 [citado 8 Jun 2011]. Disponible en: http://www.iref.org/LTFU_Death_Appendices1_to_8.html.
11. Rubin DB. The design versus the analysis of observational studies for causal effects: Parallels with the design or randomized trials. Stat Med. 2007;26:20–36.
12. Rubin DB. For objective causal inference, design trumps analysis. Ann Appl Stat. 2008;2:808–40.
13. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics. 1968;24:295–313.
14. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc. 1984;79:516–24.
15. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat. 1985;39:33–8.
16. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. Biometrics. 1985;41:103–16.
17. D'Agostino RB. Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17:2265–81.
18. Rubin DB. Using multivariate matched sampling and regression adjustment to control bias in observational studies. J Am Stat Assoc. 1979;74:318–28.
19. D'Agostino Jr RB, D'Agostino Sr RB. Estimating treatment effects using observational data. JAMA. 2007;297:314–6.
20. Rubin DB. Matched sampling for causal effects. New York: Cambridge University Press; 2006. p. 167.