10. Fuentes de error

Un marco sistemático para identificar las potenciales fuentes y el impacto de la distorsión en estudios observacionales, con la intención de mantener la validez

Ya hemos considerado muchas fuentes de error en los estudios epidemiológicos: sobrevida selectiva, recuerdo selectivo, clasificación incorrecta de los sujetos con respecto a su enfermedad y/o estado de exposición. Dada la limitada oportunidad para controles experimentales, el error, particularmente el "sesgo", es una preocupación de primordial importancia para los epidemiólogos (y nuestros críticos!) además de la base principal para dudar de o discutir los resultados de las investigaciones epidemiológicas.

Exactitud es un término general que denota la ausencia de error de todo tipo. En un marco conceptual moderno (Rothman y Greenland), el objetivo general de un estudio epidemiológico es la precisión en la medición de un parámetro, como la razón de densidad de incidencias que relaciona una exposición con un daño. Fuentes de error en la medición son clasificadas como aleatorias o sistemáticas (Rothman, pág.78.)

Rothman define el ***error aleatorio*** como "aquella parte de nuestra experiencia que no podemos predecir" (pág. 78.) Desde el punto de vista estadístico, el error aleatorio también puede ser considerado como la variabilidad del muestreo. Aún cuando no está involucrado un procedimiento de muestreo formal, como por ejemplo, una única medición de presión sanguínea en un sólo individuo o como una observación del verdadero valor más una observación de un proceso aleatorio representando factores del instrumento y de la situación particular. Lo inverso del error aleatorio es la ***precisión***, que es por lo tanto un atributo deseable de la medición y de la estimación.

El error sistemático, o ***sesgo***, es la diferencia entre un valor observado y el verdadero valor debido a todas las causas menos la variabilidad del muestreo (Mausner y Bahn, 1ª. ed., pág. 139.) El error sistemático puede surgir de innumerables fuentes, incluyendo factores involucrados en la selección o reclutamiento de la población de estudio y los factores involucrados en la definición y medición de las variables de estudio. Lo inverso del sesgo es la ***validez***, también un atributo deseable.

Estos términos – "error sistemático", "sesgo", "validez" – son utilizados en varias disciplinas y en distintos contextos, con significado similar pero no idéntico. En estadística, "sesgo" se refiere a la diferencia entre el valor promedio de un estimador, calculado con múltiples muestras al azar, y el verdadero valor del parámetro que busca estimar. En psicometría, "validez" se refiere habitualmente al grado en que el instrumento de medición mide el constructo que se supone que mide. La distinción entre error sistemático y aleatorio se encuentra en muchas disciplinas, pero como veremos estos dos tipos de error no están totalmente separados. Volveremos al tema de terminología más adelante, en la sección sobre "Conceptos y terminología".

Precisión

La presencia de la variación aleatoria debe ser siempre tenida en mente al diseñar estudios e interpretar datos. Hablando en general, los números pequeños llevan a estimaciones imprecisas. Por lo tanto, pequeñas diferencias basadas en números pequeños deben ser tomadas con precaución dado que estas diferencias tienen tanta probabilidad de ser el producto de la variación aleatoria como de algo interpretable.

Estimaciones de medidas de razón (p.ej. el riesgo relativo) basadas en escasos datos son muy susceptibles a la inestabilidad. Por ejemplo, un riesgo relativo de 5.0 basado en la ocurrencia de 4 casos entre los no expuestos se hace el doble de importante (10.0) si dos casos no expuestos no son captados por las dificultades de muestreo, medición, datos que faltan, u otras razones. Si faltan tres, el riesgo relativo sería 20.

Ejemplo para ilustrar el concepto de precisión

Consideremos los datos siguientes, de la tabla 4 de Hulka y cols., Controles "alternativos" en un estudio caso-control de cáncer de endometrio y estrógenos exógenos, ("Alternative" controls in a case-control study of endometrial cancer and exogenous estrogen), *Am J Epidemiol* 112:376-387, 1980:

Efecto de la duración de estrógenos sobre el riesgo relativo [ajustado por edad] usando tres grupos control entre mujeres blancas, Carolina del Norte, 1970-76

Duración del uso	No. de casos	Controles legrado		Controles ginecológicos		Controles comunitarios	
		No.	RR	No.	RR	No.	RR
Ningún uso	125	136		118		172	
Menos de 6 meses	8	13	0.7	12	0.7	20	0.8
6 meses - 3.5 años	9	14	0.7	9	0.9	21	0.7
3.5 años - 6.5 años	9	16	0.8	1		7	1.7
6.5 años - 9.5 años	9	11	1.2	2	3.8	5	2.5
Más de 9.5 años	19	10	2.0	2	5.1	4	5.5

Primero, recordemos que vimos en el tema anterior que, dado que estos datos provienen de un estudio caso-control, los "riesgos relativos" de la tabla son odds ratio. Dado que la enfermedad es rara, sin embargo, los odds ratio, las razones de riesgo, y las razones de densidad de incidencia serán todos aproximadamente iguales. También sobre la base del capítulo anterior deberíamos poder reformular los datos anteriores como una serie de tablas 2 x 2 si quisiéramos por algún motivo. Dicha reformulación haría más fácil ver como se calculan las estimaciones crudas del riesgo relativo (OR) de los datos en la tabla. (Señalemos que los OR en la tabla están ajustados por edad, a través de

un procedimiento de modelado matemático llamado regresión logística múltiple, de manera que nuestros OR crudos serían diferentes en algunos casos.)

Observemos que las estimaciones del riesgo relativo para los controles ginecológicos y comunitarios en las categorías de mayor duración de uso de estrógenos se basan en muy pocos controles. Por ejemplo, si dos de los controles comunitarios clasificados como con una duración de uso "3.5 años-6.5 años" fueran en realidad "6.5 años - 9.5 años", entonces las estimaciones de riesgo relativo se invertirían y dejaría de aparecer la imagen consistente de dosis respuesta. [Por el momento estamos ignorando el ajuste por edad, aunque para estos dos grupos particulares de duración en los controles comunitarios los OR ajustados son iguales a si hubiesen sido calculados a partir de los datos de la tabla.] De manera similar, los RR mayores de 5 para los sujetos con uso durante mayor tiempo se basan en 2 y 4 controles en los grupos de controles ginecológicos y comunitarios, respectivamente. Por otro lado, el hecho de que resultados similares se encontraron en dos grupos control fortalece la evaluación de que una relación dosis respuesta realmente existe, y no es un hallazgo casual.

Cuantificando el grado de precisión o imprecisión – intervalos de confianza

Las técnicas estadísticas, como errores estándares e intervalos de confianza, se utilizan para cuantificar el nivel de precisión o imprecisión de las estimaciones; también hay reglas generales (p.ej., ver Alvan Feinstein, *J Chron Dis* 1987; 40:189-192.) Un ***intervalo de confianza*** da un amplio rango de valores que se espera incluya el verdadero valor del parámetro estimado. Cuanto más estrecho el intervalo de confianza, más precisa la estimación.

Por ejemplo, supongamos que estamos estimando el riesgo relativo para cáncer de endometrio en mujeres que han utilizado estrógenoterapia de reemplazo durante 3.5 años a 6.5 años (en comparación con mujeres que no han tomado estrógenos) y que el "verdadero" (pero desconocido) riesgo relativo es 1.5. Supongamos también que las estimaciones que obtenemos de los datos de los controles comunitarios de la tabla anterior no son sesgadas, aunque sí reflejan el error aleatorio. Hulka y cols. calcularon el valor ajustado por edad como 1.7 (el valor crudo es muy similar: 1.77.) El 1.7 es una estimación puntual y provee nuestra mejor estimación del verdadero (pero desconocido) riesgo relativo de 1.5.

No esperamos que nuestra estimación sea exactamente el valor correcto, por eso calculamos también una estimación de intervalo, o intervalo de confianza como indicador de cuantos datos había disponibles para la estimación. Supongamos que el intervalo de confianza del 95% es (0.6, 4.7.) La interpretación sería que 1.7 es la mejor estimación puntual del verdadero riesgo relativo (desconocido) y que hay una "confianza del 95% de que el verdadero riesgo relativo se encuentre entre 0.6 y 4.7". "Confianza" no significa lo mismo que "probabilidad". En este caso "95% de confianza" significa que obtuvimos los límites de confianza 0.6 y 4.7 a través de un procedimiento que produce un intervalo que contendrá el verdadero valor en 95% de las instancias en que lo usemos y no contendrá el verdadero valor en el 5% restante de las instancias. Hablando informalmente, un intervalo de confianza de 95% de 0.6-4.7 significa que el valor observado de 1.7 es "compatible", en la usanza convencional, con riesgos relativos verdaderos que se encuentren entre 0.6 y 4.7 inclusive.

Otra forma de describir el significado de "compatible" es el siguiente. Los límites 0.6 y 4.7 se obtienen de la estimación puntual de 1.7 y el error estándar estimado de esa estimación. El error estándar estimado es una función de la magnitud de los números (i.e., de la cantidad de datos) sobre la cual se basa la estimación puntual y que es por lo tanto una medida de su imprecisión. Un intervalo de confianza del 95% (0.6,4.7) significa que si nuestro estudio hubiera dado una estimación puntual en cualquier punto de ese intervalo, el intervalo de confianza del 95% en torno a esa estimación puntual contendría el valor de 1.7. En ese sentido el valor observado de 1.7 es compatible con los verdaderos riesgos relativos entre 0.6 y 4.7.

De hecho, la tabla original de Hulka y cols. incluía intervalos de confianza para los odds ratio. El hecho de prestarle atención a los intervalos de confianza o a la escasez de datos es un aspecto importante de la interpretación de resultados.

Disminuyendo la variación aleatoria (aumentando la precisión)

Los intervalos de confianza y otros procedimientos para evaluar las posibilidades de variación aleatoria en una investigación no aumentan la precisión, simplemente la cuantifican. Las estrategias principales para disminuir la influencia del error aleatorio son:

1. Aumentar el tamaño muestral – una muestra mayor, manteniendo todo lo demás igual, proveerá estimaciones de parámetros poblacionales más precisos;
2. Mejorar los procedimientos de muestreo – una estrategia de muestreo más refinada, p.ej. muestreo al azar estratificado combinado con técnicas analíticas apropiadas a menudo puede disminuir la variabilidad de muestreo en comparación con muestreo al azar simple;
3. Disminuir la variabilidad de la medición usando protocolos de medición estrictos, mejores instrumentos, o promedios de múltiples medidas.
4. Usar métodos analíticos más estadísticamente eficientes, i.e., en el grado de precisión que se puede obtener de un tamaño muestral dado;

Sesgo

El sesgo, por definición, no es afectado por el tamaño muestral. Más bien el sesgo depende del reclutamiento y mantenimiento de los pacientes en el estudio y sobre la medición. [Una definición técnica de "sesgo" en su usanza epidemiológica (basado en Kleinbaum, Kupper, and Morgenstern) es cuánto una estimación difiere del verdadero valor del parámetro estimado, aunque el tamaño muestral se aumente al punto en que la variación aleatoria es insignificante. Esta definición se basa en el concepto estadístico de consistencia; en estadística, un estimador es ***consistente*** si su valor se acerca continuamente al valor del parámetro que estima a medida que aumenta el tamaño muestral.

Conceptos y terminología

En el área de sesgo y validez, como en tantas otras áreas que se dan en muchas disciplinas, la terminología puede ser una fuente significativa de confusión. Dichos peligros se hacen particularmente aparentes cuando los términos también se usan en un sentido no técnico en la conversación común. Una fuente adicional de confusión para la terminología referente a la validez es la superposición de conceptos. Por ejemplo, las mediciones son un ingrediente para los estudios, pero los estudios también pueden ser considerados como procedimientos de medición aplicados a poblaciones o asociaciones. De manera que los mismos términos pueden ser usados para las mediciones individuales y para estudios enteros, aunque el significado cambia con el contexto.

Validez Interna

Los epidemiólogos distinguen entre la validez interna y la validez externa. ***Validez interna*** se refiere a la ausencia de error sistemático que hace que los resultados del estudio (estimaciones de parámetros) difieran de los verdaderos valores como están definidos en los objetivos de estudio. El error sistemático puede resultar de mediciones inexactas de las variables de estudio, falta de uniformidad en el reclutamiento o retención de los participantes del estudio, o comparación de grupos que difieren en características importantes pero desconocidas.. Así la validez interna tiene que ver con el sesgo en las estimaciones para la población blanco especificada en los objetivos de estudio.

Validez externa

Validez Externa se refiere a la medida en que los hallazgos del estudio puedan aplicarse a poblaciones distintas a la investigada. El poder de generalizar a poblaciones más allá de la población blanco para la cual se diseñó el estudio y/o más allá de las circunstancias implícitas en el estudio es un tema de inferencia científica, más que un problema técnico o estadístico (ver Rothman y Greenland.) Por lo tanto la validez externa es mejor considerarla con relación a la inferencia causal y la interpretación de los resultados del estudio. (Rothman y Greenland consideran que "validez externa" es un término inapropiado, prefiriendo distinguir entre validez y generalizibilidad.)

La validez se relaciona con una medida específica

Dado que diferentes tipos de errores afectan hallazgos específicos de maneras diferentes, la validez debe ser discutida con respecto a una medida o medidas específicas. Un estudio enfocado a probar una hipótesis etiológica, típicamente busca estimar una medida de fuerza de asociación como la razón o diferencia de incidencias en distintos grupos. La falta de validez interna en este contexto significa inexactitud (sesgo) en estas estimaciones. De hecho, Kleinbaum, Kupper, y Morgenstern (*Epidemiologic Research*, cap. 10) definen validez (interna) y sesgo en términos de la distorsión sistemática de la "medida de efecto". Un estudio puede dar una medida válida (no sesgada) de efecto a pesar de los errores sistemáticos en los datos si los errores casualmente se contrarrestan unos a otros con respecto a la medida de efecto. Sin embargo, un estudio sin errores sistemáticos puede dar una estimación de la medida de efecto sesgada (por ejemplo, debido a la variación aleatoria en una medida importante – ver apéndice.) Mucho de lo que se ha escrito en metodología sobre sesgo tiene que ver con la distorsión de las medidas de efecto.

No todos los estudios tienen como objetivo la estimación de una medida de efecto, y aún estudios que sí lo tienen también informan estimaciones de otros parámetros (p.ej., tasas de incidencia, prevalencias, medias.) Así, aún si la medida de efecto se estima con exactitud, se debe tener en cuenta la posibilidad y la magnitud del sesgo en otras medidas.

Validez de la medición

En la perspectiva de Rothman y Greenland, la medición es el propósito de todos los estudios, de manera que el concepto de validez de la medición es lo mismo que el de validez. Sin embargo, la validez de ***las mediciones*** que se llevan a cabo al realizar una investigación presenta problemas propios y es abordada en otra categoría de literatura metodológica. La validez de la medición (debo confesar que este término es mi manera de diferenciar este tipo de validez) tiene que ver con el hecho de evitar el error de medición o detección de un factor (p.ej., presión sanguínea, tasa de fumadores, alcoholismo, infección por VIH.) La literatura sociológica y sicológica trata en forma importante con la validez de la medición, particularmente en relación con los datos recolectados a través de los cuestionarios y las entrevistas. La sicología cognitiva estudia los procesos de pensamiento por los cuales los participantes del estudio decodifican los puntos del cuestionario y recuperan la información guardada en la memoria (p.ej., Warnecke *y cols.*, 1997a; Warnecke *y cols.*, 1997b.) La psicometría estudia los aspectos estadísticos de los instrumentos de medición sicológicos (Nunnally, 1994.) Estas disciplinas son especialmente pertinentes para los epidemiólogos interesados en mediciones sofisticadas de medidas auto-reportadas.

Dirección del sesgo – "para qué lado es para arriba"

Los conceptos y la terminología pueden complicar también las descripciones de la dirección en que el sesgo distorsiona una medida de efecto. Las fuentes de confusión son: (1) una asociación puede ser positiva (RR >1.0) o inversa (RR < 1.0, también llamado "negativa"), (2) una fuente de sesgo puede hacer que una medida de efecto aumente de magnitud, disminuya en magnitud, se acerque a 1.0 desde por encima o por debajo, y se aleje del 1.0 en cualquiera de las direcciones, y (3) es fácil perder de vista si la medida de asociación que estamos tratando es la observada en el trabajo o la "verdadera" que existe en la población blanco. [Intenta graficar algunos riesgos relativos a medida que lees los siguientes dos párrafos.]

Describiendo la dirección del sesgo – ejemplo:

Supongamos que las personas "agresivas" tienen más probabilidad de sobrevivir un infarto agudo de miocardio (IAM) que las personas no agresivas. Un estudio caso-control de IAM que recluta sus casos entre pacientes hospitalizadas (vivos) con IAM tendrá por lo tanto una sobre-representación de personas agresivas, dado que proporcionalmente sobrevivirán más, durante más tiempo, como para ser incluidas en el estudio. Si esta es la única fuente de error sistemático, entonces esperamos que el riesgo relativo (RR) observado será mayor que el verdadero riesgo relativo de la incidencia del IAM (dado que el verdadero riesgo relativo incluiría las víctimas que fallecieron antes de que pudieran ser incluidos en el trabajo.) La dirección del sesgo es en sentido positivo (a valores mayores de RR), no importando si el verdadero RR es mayor que 1.0 (i.e., las personas agresivas también tienen más probabilidad de IAM) o menor que 1.0 (las personas agresivas tienen menos probabilidad.)

En contraste con lo anterior, el error aleatorio uniforme en la medición de la agresividad, independiente de otras variables, típicamente modifica el RR observado "hacia el valor nulo" (más cerca de 1.0 que el verdadero RR.) El sesgo hacia el valor nulo puede producir un RR observado menor (si el verdadero RR es mayor que 1.0) o hacia un RR observado mayor (si el verdadero RR es menor que 1.0), pero no un RR que se aleja del valor nulo más que el verdadero RR. Por otro lado, el sesgo de la mayor sobrevida de los casos de IAM con carácter agresivo en el estudio caso-control hipotético anterior se acercará más a 1.0 sólo si el RR es menor a 1.0 y se alejará del 1.0 sólo si el RR es mayor que 1.0

Por estas razones necesitamos cuatro términos para caracterizar los efectos potenciales de las fuentes de sesgo:

- "***Sesgo positivo***" – la medida de efecto observada es un número mayor que la verdadera medida de efecto (si se pudiera conocer);
- "***Sesgo negativo***" – la medida de efecto observada es un número menor que la verdadera medida de efecto (si se pudiera conocer);
- "***Hacia el valor nulo***" – la medida de efecto observada es más cercana a 1.0 que la verdadera medida de efecto (si se pudiera conocer);
- "***Se aleja del valor nulo***" – la medida de efecto observada está más lejos de 1.0 que la verdadera medida de efecto (si se pudiera conocer);

Otra forma de describir la dirección del sesgo es decir que la medida de efecto observada sobreestima (subestima) la verdadera medida. Con estas frases, sin embargo, se necesita disponer de más información, dado que "sobreestima" puede ser entendido como mayor en valor numérico o mayor en fuerza (más lejos del valor nulo.)

Con la intención de mantener una comunicación precisa, trataremos de adherir a la usanza anterior, que no parece ser estándar en la profesión. Sin embargo, la terminología es sólo una fuente de confusión. Consideremos la expresión planteada desde hace mucho tiempo de que la clasificación errónea no diferencial (discutida más adelante) de una exposición o una variable de enfermedad dicotómica, en ausencia de confusión (ver el próximo capítulo) siempre produce un sesgo "hacia el valor nulo". Esta expresión es cierta siempre y cuando la clasificación errónea no diferencial (independiente) no es peor de lo que resultaría de la clasificación de cada observación según el resultado de tirar una moneda. Sin embargo la clasificación errónea no diferencial extrema (en el caso extremo, la clasificación errónea de todos los participantes), puede sesgar la medida del efecto más allá y luego alejándose del valor nulo.

Tipos de sesgo

Los estudiantes de epidemiología a menudo desean un catálogo de los tipos de sesgo para poder reconocerlos en los estudios publicados. David Sackett (Bias in analytic research. *J Chron Dis* 32:51-63, 1979) una vez intentó desarrollar uno. Nueve ejemplos que él describe son:

1. Sesgo de prevalencia - incidencia (Neyman)

   Este es el término usado por Sackett para, entre otras cosas, la sobrevida selectiva. También incluye los fenómenos de retorno a la normalidad de signos de eventos clínicos previos (p.ej. IAM "silente" que no deja posteriormente evidencia electrocardiográfica clara) y/o cambio del factor de riesgo después de iniciarse un proceso fisiopatológico (p.ej., un Tipo A puede cambiar su comportamiento después de tener un IAM) de manera que los estudios basados en prevalencia producirán una imagen distorsionada de lo que ocurre en términos de incidencia.

2. Sesgo de tasa de admisión (Berkson)

   Cuando los casos y/o los controles se reclutan a partir de los pacientes hospitalizados, las características de ambos grupos serán influidas por las tasas de ingreso hospitalario.

3. Sesgo de desenmascaramiento (señal de detección)

   Dado que por necesidad, una enfermedad debe ser detectada para ser contada, a veces se cree equivocadamente que los factores que influyen en la detección de la enfermedad pueden influir en la ocurrencia de la enfermedad. Esta situación tiene mayores probabilidades de ocurrir cuando el proceso de la detección de la enfermedad ocurre fuera del estudio (p.ej., en un estudio caso-control), cuando la enfermedad tiene una fase oculta, o asintomática, y cuando la exposición lleva a síntomas que inducen al individuo a solicitar atención médica.

4. Sesgo de no-respuesta

   Los que no responden en una encuesta a menudo son diferentes a los que responden en relación a temas importantes. De igual manera, los voluntarios a menudo son diferentes de lo no voluntarios, los que responden tardíamente de los que responden oportunamente, y los que abandonan el estudio de aquellos que lo completan.

5. Sesgo de pertenencia

   La pertenencia a un grupo puede implicar un nivel de salud que difiere de la de todos los demás en la población general. Por ejemplo, la observación de que la actividad física vigorosa protege contra la ECC se creía inicialmente que era resultado del hecho de que las personas en mejores condiciones físicas (con un riesgo de ECC inherente menor) tendrían más probabilidad de participar en actividades vigorosas. Otro ejemplo sería si las personas que participaran en un programa de promoción de salud posteriormente realizan más cambios beneficiosos en su estilo de vida que los que no participaron, no debido al propio programa sino a la motivación de los participantes y su disposición al cambio.

6. Sesgo de sospecha diagnóstica

   El proceso diagnóstico incluye muchas oportunidades para formar opiniones. Si el conocimiento de la exposición influye sobre la intensidad y el resultado del proceso diagnóstico, los casos expuestos tienen una mayor (o menor) probabilidad de ser diagnosticados, y por lo tanto contados como casos.

7. Sesgo de sospecha de exposición

   El conocimiento de la situación de enfermedad puede influir sobre la intensidad y los resultados de una búsqueda de la exposición y la causa putativa.

8. Sesgo de recuerdo

   El recuerdo puede ser diferente en los casos y los controles tanto en magnitud como en exactitud (recuerdo selectivo.) Los casos pueden ser interrogados más intensamente que los controles.

9. Sesgo de información familiar

   Dentro de una familia, el flujo de información sobre las exposiciones y las enfermedades es estimulado por, y dirigido a, el miembro de la familia que desarrolla la enfermedad. Así la persona que desarrolla una artritis reumatoidea tiene más probabilidad de conocer un pariente con historia de artritis que sus hermanos sanos.

El apéndice del artículo de Sackett muestra su catálogo completo de sesgos.

Clasificando las fuentes de sesgo

A pesar de la iniciativa de David Sackett, no existe aún un catálogo completo de sesgos. En vez, siguiendo los trabajos de Olli Miettinen de la década de los 70, los epidemiólogos generalmente se refieren a tres tipos de sesgos:

1. ***Sesgo de selección*** – distorsión que resulta de los procesos por los cuales los sujetos son seleccionados para ser parte de la población de estudio.
2. ***Sesgo de información*** (también llamado ***sesgo de clasificación errónea***) – distorsión que resulta de imprecisiones en la medición de las características de los sujetos, y por lo tanto su incorrecta clasificación.
3. ***Sesgo de confusión*** – distorsión en la interpretación de los hallazgos debida al hecho de no tomar en cuenta los efectos de otros factores de riesgo más que los de la exposición de interés.

El sesgo de confusión es algo diferente de las otras dos formas de sesgo por el hecho de que los datos recolectados por la investigación pueden ser en sí mismos correctos; el problema surge de atribuir equivocadamente los efectos observados (o su ausencia), i.e., un efecto aparente es atribuido a la exposición de interés, cuando en realidad debería ser atribuido a algún otro factor. Discutiremos el fenómeno de confusión en el próximo capítulo.

Por supuesto, como en tantas otras áreas de la epidemiología, las divisiones entre las clases son sólo relativas, no absolutas!

Sesgo de Selección

Si dejamos de lado los problemas del error aleatorio del muestreo (i.e., suponemos que todas las muestras son suficientemente grandes para que la variación aleatoria debida al muestreo sea insignificante), podemos ver que, si los procesos por los cuales se seleccionan los sujetos favorecen o pasan por alto ciertos tipos de sujetos, la población de estudio que obtenemos no será representativa de la población para la cual estamos tratando de obtener estimaciones. Por ejemplo, si estamos estudiando las características de las personas con diabetes y obtenemos todos nuestros sujetos de pacientes de hospital, las características de esta población de estudio mostrarán una estimación distorsionada o sesgada de las características de los diabéticos en general.

En estudios caso-control, las situaciones que pueden producir sesgo incluyen:

- La exposición tiene alguna influencia sobre el proceso de determinación de los casos ("sesgo de detección"): la prevalencia de la exposición en los casos estará sesgada;
- La sobrevida selectiva o migración selectiva – la prevalencia de la exposición en los casos prevalentes puede estar sesgada en comparación con aquella en los casos incidentes;
- La exposición tiene alguna influencia sobre el proceso por el cual se seleccionan los controles (p.ej., uso de pacientes con bronquitis crónica como controles para un estudio de cáncer de pulmón y el hábito de fumar): la prevalencia de la exposición en los controles será diferente a la de la población de origen.

En estudios de cohorte, la fuente primaria de sesgo de selección es generalmente una atrición o pérdida al seguimiento diferencial. Ejemplo (hipotético):

Cohorte completa:

	Tipo A	Tipo B	
ECC	40	20	
ECC	160	180	
Total	200	200	RR=2.0

Cohorte observada:*

	Tipo A	Tipo B	
ECC	32	18	
ECC	144	162	
Total	176	180	RR=1.82

*basado en una pérdida del 10% de los sujetos, salvo que los sujetos de tipo A que desarrollan ECC se supone que sufrieron pérdidas a una velocidad del 20%. Si todos los sujetos, incluyendo el grupo ECC/Tipo A hubieran sufrido una velocidad de pérdidas del 10%, la incidencia en cada grupo de tipo de comportamiento, y por lo tanto la razón de riesgos, no estaría distorsionada.

Marco conceptual

[Basado en Kleinbaum, Kupper y Morgenstern, *Epidemiologic Research* y el artículo del *Am J Epidemiol* sobre sesgo de selección (ver bibliografía)].

Población externa: la población de interés última, pero que no estamos tratando de estudiar directamente – p.ej., podríamos querer estudiar la relación entre hipertensión y accidente vascular encefálico isquémico en general, pero estudiamos sólo sujetos en Carolina del Norte, reconociendo que la generalización a otras áreas requeriría la consideración de las diferencias entre Carolina del Norte y esas otras áreas. No nos preocuparemos ahora, en este capítulo, de las posibilidades de generalizar los resultados.

Población blanco (objetivo): la población para la cual tenemos intención de hacer las estimaciones.

Población real: la población a la cual corresponden en realidad nuestras estimaciones. Esta población puede no ser obvia o aún imposible de conocer.

Población de estudio: el grupo de participantes de los cuales hemos recolectado datos. En el marco de Kleinbaum, Kupper, y Morgenstern, la población de estudio se considera una muestra no sesgada de la población real, difiriendo de ella sólo por el error de variabilidad no sistemático del muestreo.

La población de estudio es un subconjunto de la población real. El sesgo es la discrepancia entre la población real y la población objetivo. La posibilidad de generalizar trata con la inferencia de la población objetivo a la población externa (ver antes.)

Al pensar sobre el sesgo de selección y su efecto potencial sobre los resultados del estudio, encontramos que es útil considerar las probabilidades según las cuales las personas en la población blanco pueden acceder a la población real. Estas probabilidades se llaman probabilidades de selección (poblacionales.)

Para simplificar, consideremos una clasificación dicotómica de enfermedad y de exposición, y digamos que la tabla de doble entrada en la población blanco y en la población es como sigue:

	Exp	$\overline{\text{Exp}}$
Enf	A	B
$\overline{\text{Enf}}$	C	D

Blanco

	Exp	$\overline{\text{Exp}}$
Enf	A^o	B^o
$\overline{\text{Enf}}$	C^o	D^o

Real

Podemos definir cuatro probabilidades de selección:

alfa (α) = (A^o/A) la probabilidad de que una persona en la celda A (en la población blanco) será seleccionada para la población real de la cual la población de estudio es una muestra al azar.

beta (β) = (B^o/B) la probabilidad de que una persona en la celda B (en la población blanco) será seleccionada para la población real

gamma (γ) = (C^o/C) la probabilidad de que una persona en la celda C (en la población blanco) será seleccionada para la población real

delta (δ) = (D^o/D) la probabilidad de que una persona en la celda D (en la población blanco) será seleccionada para la población real

Ejemplo: supongamos que existe una sobrevida selectiva, de manera que los fumadores de cigarrillos que sufren un IAM tienen más probabilidades de morir antes de llegar al hospital. Por lo tanto un estudio caso-control de IAM y el hábito de fumar, utilizando pacientes hospitalizados por IAM tendrá un alfa menor que beta (los casos expuestos estarán menos disponibles para el estudio que los casos no expuestos.) Este sesgo producirá una distorsión en el odds ratio que subestimará una verdadera asociación entre el hábito de fumar y el IAM (i.e., un sesgo negativo.)

La tarea de esta lección incluye un ejercicio que solicita que se aplique este marco conceptual a un tema de sesgo de detección que involucra el cáncer de endometrio y los estrógenos. El tema básico

es que el uso de estrógenos puede llevar a sangrados uterinos, lo cual resultaría en que las mujeres solicitarían atención médica y se les realizaría un legrado. Si hubiera un cáncer de endometrio oculto (asintomático), el legrado permitiría detectarlo. Según el escenario del sesgo de detección, las mujeres con cáncer de endometrio oculto tienen mayores probabilidades de recibir atención médica si son usuarias de estrógenos, creando una situación de sesgo de detección.

Este escenario fue vigorosamente discutido, dado que depende de la existencia de un reservorio de tamaño considerable de casos de cáncer de endometrio asintomáticos, lo cual ha sido descartado en la actualidad. Sin embargo, el tema del cáncer de endometrio y el uso de estrógenos permite abundantes ilustraciones de conceptos relacionados con el sesgo de selección y sesgo de información. Volveremos a ver este caso de estudio pronto. (Señalemos que aunque el sesgo en los estudios caso-control atrae la mayor parte del interés teórico, todos los diseños de estudio son vulnerables.)

Recursos — minimizar las pérdidas al seguimiento, obtener poblaciones de estudio representativas, anticipar las fuentes de sesgo y evitarlas. A veces los factores asociados con el sesgo de selección pueden ser medidos, en cuyo caso el análisis de los datos puede intentar tomar estos factores en cuenta. La lógica en la interpretación de los datos puede organizar la evidencia a favor o en contra del sesgo de selección como responsable de una asociación observada. Pero, si lo puedes evitar, eso sería lo mejor!

Ejemplo del caso de los estrógenos y el cáncer de endometrio

Durante la década de los 70, los estudios caso-control informaron una fuerte asociación (OR de aproximadamente 10) entre el cáncer de endometrio y el uso de estrógenos en la postmenopausia. La asociación era biológicamente plausible, dado que el endometrio uterino es un tejido sensible a los estrógenos. Además, las tasas de cáncer endometrial se estaban elevando en las áreas geográficas donde el uso de estrógenos posmenopáusicos estaba aumentando más rápidamente.

Sin embargo, los críticos de los estudios caso-control también estaban aumentando. Por un lado, los estudios caso-control que habían informado una asociación entre cáncer de mama y la medicación antihipertensiva, reserpina, habían recibido amplia difusión, pero la asociación fue luego desechada. También los críticos del diseño caso-control se habían hecho escuchar (notoriamente Alvan Feinstein, que denominó el diseño el estudio "trohoc" ["cohort" escrito al revés].) El *Journal of Chronic Disease* (ahora conocido como *Journal of Clinical Epidemiology*) auspició una reunión para los epidemiólogos más importantes para discutir la validez del diseño (actas publicadas en Michel A. Ibrahim and Walter O. Spitzer. *El estudio caso-control: consenso y controversias* [*The case-control study: consensus and controversy*]. Pergamon, Nueva York, 1979.)

Más o menos en la misma época, Barbara Hulka, Carol J.R. Hogue, y Bernard G. Greenberg (fallecido) (todos en la Escuela de Salud Pública de la Universidad de Carolina del Norte en ese momento) publicaron una revisión exhaustiva de los problemas metodológicos involucrados en la asociación de estrógenos y cáncer de endometrio (Temas metodológicos en los estudios epidemiológicos del cáncer de endometrio y estrógenos exógenos [Methodologic issues in

epidemiologic studies of endometrial cancer and exogenous estrogen] *Amer J Epidemiol* 1978; 107:267-276.) El diseño caso-control es particularmente susceptible al sesgo de selección, porque la enfermedad ya ha ocurrido, la validez del estudio depende fundamentalmente de la selección de casos y controles. La revisión de Hulka y cols. estableció los siguientes puntos (más material de esta revisión se presenta en el apéndice de este capítulo):

1. Determinación de casos

   Los casos proveen una estimación de la exposición a los estrógenos en mujeres que desarrollan cáncer de endometrio. Esta estimación de la prevalencia de exposición entre los casos puede ser expresada en términos probabilísticos como Pr(Exp|E) – la probabilidad de exposición condicionada al hecho de tener la enfermedad.

   Los casos en el estudio, por lo tanto deben ser representativos de todas las personas con la misma descripción (i.e., en cuanto a edad, geografía, diagnóstico secundario) que las personas que desarrollan la enfermedad con respecto al estado de exposición. En el caso del cáncer de endometrio, dos temas particulares son:

   a. La heterogeneidad de los casos (estadío, grado, tipo histológico) puede reflejar etiologías o relaciones con la exposición subyacente diferentes.
   b. Las fuentes para los casos y para los procesos diagnósticos pueden tener implicancias en el estado de exposición (p.ej., los casos de hospitales rurales pueden haber tenido menos acceso a los estrógenos en la postmenopausia.)

2. Selección de los controles

   Los controles permiten una estimación de la prevalencia de exposición en la población de donde surgieron los casos (llamada ahora "base del estudio".) Esta prevalencia puede ser expresada en términos probabilísticos como Pr(Exp) - la probabilidad de que una persona seleccionada al azar de la población base del estudio esté expuesta a los estrógenos exógenos. Por lo tanto los controles deben ser representativos de la población base del estudio con respecto al estado de exposición, de manera que la prevalencia del uso de estrógenos en los controles (en términos probabilísticos, Pr(Exp|no E) estima la exposición en la población base del estudio con precisión. Además los controles deben poder proveer los datos de la exposición y otros con una precisión equivalente a la que se obtiene de los casos (este punto tiene que ver con el sesgo de información y será discutido más adelante en este capítulo.)

   Por lo tanto los controles deben ser similares a los casos en términos de:

   a. Fuentes de datos, de manera que la oportunidad de investigar sobre uso previo de estrógenos es equivalente a la de los casos;
   b. Otros determinantes de la enfermedad que no pueden ser controlados explícitamente

Pero los controles no deben ser demasiado similares a los casos con respecto a factores no determinantes de la enfermedad.

Sobreapareamiento y la selección de controles

Esta última calificación está dirigida al tema del sesgo de detección planteado por Feinstein (ver anteriormente.) La recomendación de Ralph Horwitz y Feinstein para disminuir el sesgo de detección es seleccionar controles entre las mujeres que habían recibido el mismo procedimiento diagnóstico que los casos (legrado), asegurando así que las pacientes control no tenían una enfermedad oculta y haciéndolas más parecidas a los casos. La respuesta de Hulka y cols. fue que dicho procedimiento de selección para los controles constituye un sobreapareamiento

El concepto de sobreapareamiento y los "controles alternativos" propuestos por Horwitz y Feinstein (NEJM, 1978) se refieren a la relación del sesgo de selección con la selección del grupo control en un estudio caso-control, que es la razón por la cual el tema estrógenos-cáncer de endometrio es un tema tan excelente para comprender la selección de los controles.

Los controles en un experimento

En un verdadero experimento, en que un grupo recibe un tratamiento y el otro sirve como grupo control, la situación óptima es, generalmente, que los grupos de tratamiento y de control sean lo más idénticos posibles en el momento de administrar el tratamiento y ser sometidos a ambientes lo más similares posibles aparte del tratamiento. Si la aleatorización de un gran número de participantes no es factible, el grupo control es apareado con el experimental para conseguir la similitud máxima posible en todo lo que pueda afectar el desarrollo del daño.

A generaciones anteriores de epidemiólogos se les enseñaba que, por analogía, el grupo control en un estudio caso-control debe ser similar al grupo de casos en todas sus características menos la enfermedad (y su situación de exposición, que es lo que el estudio va a tratar de estimar.) De esa manera, las diferencias en exposición pueden ser atribuidas a los efectos de la exposición sobre el riesgo de enfermedad, el único punto de diferencia. Con ese objetivo, los controles a menudo han sido apareados a los casos para aumentar la similitud de los grupos.

Analogías entre los diseños experimental y caso-control

Sin embargo, la analogía del grupo control en un estudio caso-control y el grupo control en un experimento es imperfecta. En un experimento, la exposición es introducida en uno de dos grupos que, con un poco de suerte, son equivalentes, y los resultados se desarrollan a posteriori. El grupo control se selecciona para tener un riesgo equivalente para el daño en ausencia de la exposición. En un estudio caso-control, las exposiciones existen en una población, y los daños se desarrollan. La equivalencia que se requiere para una comparación válida es entre los expuestos y los no expuestos. El grupo de casos – los integrantes de la población que han desarrollado el daño – no se ubican en una posición correspondiente con respecto al proceso de la enfermedad como lo está el grupo expuestos en un verdadero experimento. El primero es un grupo de personas que desarrollan el daño; el segundo es un grupo en riesgo de sufrir el daño.

El análogo experimental correcto del grupo de casos en un estudio caso-control es el grupo de participantes que desarrollan el daño durante el experimento. En ambos diseños, los casos surgen de

una población de personas tanto expuestas (o "experimentales") y no expuestas (o "controles".) De igual manera, el análogo correcto para el grupo control en un estudio caso-control es una muestra aleatoria de todos los participantes del experimento en algún momento después del comienzo de la exposición. Si un estudio caso control se lleva a cabo usando los casos que surgieron en ese experimento, entonces el grupo control debe servir para estimar la proporción de exposición en esa población de estudio.

Sesgo de apareamiento y de selección

El forzar al grupo control a ser similar al grupo de casos, sea a través del apareamiento o usando una fuente de reclutamiento para los controles similar a la de reclutamiento de los casos, en general hará que el grupo control se parezca menos a la población de estudio y puede por lo tanto introducir un sesgo de selección. La introducción o no de un sesgo de selección, dependerá de los métodos de análisis utilizados y si los factores de apareamiento están relacionados con la prevalencia de la exposición. Si las características no están relacionadas con la exposición no habrá sesgo de selección para esa exposición, dado que tanto los grupos control apareados como los no apareados presumiblemente darán la misma estimación de la prevalencia de la exposición. Si las características son factores de riesgo para la enfermedad, entonces aunque el apareamiento pueda introducir un sesgo de selección, este sesgo puede ser eliminado controlando para los factores de apareamiento en el análisis (piensa en cada factor de apareamiento como que identifica un subconjunto tanto en los casos como en la población de estudio, de manera que el estudio global puede ser considerado como un conjunto de estudios caso-control, separados, paralelos, cada uno válido en sí mismo.)

Sobreapareamiento

Sin embargo, si las características están relacionadas con la exposición y no son factores de riesgo para la enfermedad, obligando a los controles a ser más como los casos distorsionará tanto la prevalencia de exposición en los controles (haciéndolos más parecidos a los casos y menos como la población en estudio) como el odds ratio que relaciona la exposición y la enfermedad. Este escenario se denomina ***sobreapareamiento***. Si los factores de apareamiento se controlan en el análisis (que habitualmente no es apropiado para factores salvo los factores de riesgo para el daño), el OR estimado será correcto pero menos preciso (i.e., un intervalo de confianza más amplio.)

Un estudio caso-control hipotético:

Supongamos que quieres hacer un estudio caso-control de cáncer de endometrio incidente y estrógenos exógenos. Coordinas para ser avisado cuando se diagnostica un cáncer endometrial en una mujer de 50-70 años de edad, empleada pública permanente, a tiempo completo (o ya retirada y pensionada) o jubilada en un área de varios estados. Supongamos que todas tienen cobertura de atención médica; 100,000 están inscriptas en planes de pago-por-acto, y 50,000 están inscriptas en atención integral ("managed care") (y nadie se cambia!.) Esta es la población de estudio.

Durante 5 años de seguimiento, se desarrollan 200 casos de cáncer de endometrio, dando una incidencia acumulada de cáncer de endometrio de 133 por 100, 000 (0.00133.) De los 200 casos, 175 estuvieron expuestos a los estrógenos, y 25 no estuvieron (estos números fueron derivados suponiendo una incidencia acumulada de 200 por 100,000 (en mujeres expuestas a los estrógenos y

40 por 100,000 (0,0004) en mujeres sin exposición, pero claro, si tú conocieras estas incidencias, no estarías haciendo la investigación.)

Supongamos que un porcentaje mucho mayor (75%) de las mujeres en los planes de pago-por-acto están tomando estrógenos exógenos que las mujeres en atención integral (25%.) Sin embargo, tú no conoces esto tampoco, porque los registros de indicaciones médicas de las organizaciones que están involucradas no están informatizados (que es la razón por la cual has recurrido a un estudio caso-control en vez de seguir a las 150,000 mujeres como una cohorte.)

Para tus controles, primero eliges una muestra al azar (y afortunadamente, exactamente representativa) de 600 mujeres del archivo principal de 150,000 empleadas del estado y jubiladas. Tus datos tienen entonces el siguiente aspecto:

Estudio de cáncer de endometrio y estrógenos de la Región del Sur (ECEES)

	Estrógenos	No estrógenos	Total
Cáncer de endometrio	175	25	200
Controles	350	250	600
Total	525	275	800

OR = 5.0
intervalo de confianza del 95% : (3.19, 7.84)*

*(ver capítulo sobre Análisis e Interpretación de Datos)

(Hasta ahora, va todo bien, dado que la razón de incidencias acumuladas, teórica, sería 0.002/0.0004 = 5.0.)

Sin embargo, tú estás preocupado, dado que sospechas que la indicación de estrógenos es muy distinta en los dos tipos de planes de atención de salud. Tu sospecha se hace mayor por el hecho de que 160/200=80% de los casos están en el plan de pago-por-acto, en comparación con sólo dos tercios de los controles seleccionados por muestreo al azar (400/600) (y 100,000/150,000 en la población de estudio.) De manera que aunque no tengas fundamentos para creer que el sistema de atención de salud de una mujer afecta su riesgo de detección de cáncer de endometrio, decides hacer que tu grupo control se parezca más al grupo de casos con respecto a su pertenencia a sistema de salud (i.e., sobreapareas.)

Dado que 80% de los casos están en el sistema de pago-por-acto y 20% en atención integral, tú utilizas un muestreo al azar estratificado para obtener la distribución deseada en los controles. Para 600 controles, eso significa que 480 (80% de 600) del grupo de pago-por-acto y 120 (20% de 600) del grupo de pago-por-acto. Dado que (sin que tú lo sepas), 75% de las mujeres en los servicios de pago-por-acto toman estrógenos, igual que 25% de las mujeres en el plan de atención integral, tu

grupo control tendrá 390 mujeres que toman estrógenos – 360 mujeres expuestas (75% × 480) del grupo de pago-por-acto y 30 mujeres expuestas (25% × 120) en atención integral. Así que ahora tus datos tendrán el siguiente aspecto:

Estudio de cáncer de endometrio y estrógenos de la Región del Sur (ECEES)
Grupo control APAREADO

	Estrógenos	No Estrógenos	Total
Cáncer de Endometrio	175	25	200
Controles	390	210	600
Total	565	235	800

OR = 3.8
Intervalo de confianza del 95%: (2.40, 5.92)

El odds ratio en esta tabla es 3.8, de manera que tu grupo control apareado ha producido de hecho un sesgo de selección. Por suerte pasa un amigo y te recuerda que cuando usas un grupo control apareado necesitas controlar para el factor de apareamiento en tu análisis. Por lo tanto debes tratarlo como si hubieras llevado a cabo dos estudios separados, uno entre las mujeres que reciben atención con el sistema de pago-por-acto médico y el otro entre las mujeres que reciben atención integral (esto se conoce como un "análisis estratificado" y será discutido en el capítulo sobre Multicausalidad – Enfoques de Análisis.) Tus dos tablas (y la tabla combinada de los totales para casos y controles) son:

Estudio de cáncer de endometrio y estrógenos de la Región del Sur (ECEES)
Grupo control APAREADO, análisis ESTRATIFICADO

	Pago-por-acto			Atención Integral			Ambos
	Estrógeno	No Estrógeno	Total	Estrógeno	No estrógeno	Total	Total total
Casos de cáncer	150	10	160	25	15	40	200
Controles	360	120	480	30	90	120	600
Total	510	130	640	55	105	160	800

OR	5.00	5.00
95% IC:	(2.55, 9.30)	(2.33, 10.71)

Análisis estratificado*
(en ambas tablas): OR=5.0 95% IC: (3.02, 8.73)

* ver el capítulo sobre Análisis multivariado.

Cada uno de los estudios caso-control tiene ahora un OR = 5.0. El grupo control dentro de cada plan de sistema de atención era una muestra al azar simple. El sesgo de selección en el grupo control apareado se mantuvo sólo para el grupo total en comparación con la población de estudio como un conjunto. Sin embargo, la estimación de intervalo del OR para el análisis estratificado (la última tabla) es más amplia que el intervalo de confianza del OR en el análisis no apareado, indicando una estimación menos precisa.

Sesgo de selección en estudios de cohortes

El sesgo de selección se considera generalmente un mayor peligro en los estudios caso-control que en los de cohortes. La razón es que en los estudios de cohortes el investigador habitualmente conoce cuántos participantes y cuáles fueron perdidos al seguimiento, de manera que puede evaluar la potencial importancia del sesgo. El investigador puede también examinar las características de inicio de los participantes que luego abandonan buscando indicaciones de que las pérdidas se distribuyen uniformemente y por lo tanto la ocurrencia de un sesgo de selección resulte menos probable.

Atrición de la cohorte poblacional

Hay, sin embargo, un tipo de atrición(pérdida) que afecta tanto los estudios de cohortes como caso-control pero que no se ve y es difícil de categorizar. El problema se relaciona con la representatividad de las personas seleccionadas para el estudio. Para simplificar explicamos la situación en relación con los estudios de cohortes, pero dado que un estudio caso-control es simplemente un método eficiente para estudiar los mismos fenómenos que un estudio de cohortes, el problema es efectivamente el mismo en ellos.

Una cohorte consiste en personas que están vivas en un punto o un período en el tiempo calendario o en relación con algún evento que comparten (p.ej., graduación de la secundaria, unirse a una fuerza de trabajo, someterse a un procedimiento quirúrgico) y que luego son seguidas hacia delante en el tiempo. La atención del investigador está dirigida fundamentalmente a lo que ocurre después que se ha formado la cohorte, pero es concebible que la mortalidad y la migración que ocurren antes de ese punto pudiesen haber influido ya en quien está disponible para ser reclutado y por lo tanto influirá en lo que será observado. Si estos factores de selección tempranos están relacionados con la exposición bajo estudio pueden disminuir un efecto observado.

Algunos ejemplos:

> Si una cohorte de personas infectadas con VIH es reclutada por medio de la incorporación de personas seropositivas identificadas a través de una encuesta serológica, aquellos que han progresado más rápidamente al SIDA estarán sub-representadas igual que las personas involucradas con comportamientos de riesgo (p.ej. uso de drogas intravenosas) que se asocian con una alta mortalidad. La progresión al SIDA en dicha cohorte aparecerá diferente que lo que se observaría si las personas fueran reclutadas en el momento de la infección inicial con VIH.

Un estudio del efecto de la hipertensión en una cohorte de participantes de la tercera edad no puede reclutar personas a quienes la hipertensión causó la muerte antes de cumplir la edad para participar en la cohorte. Si aquellos que murieron antes tenían características que los hacía más vulnerables a daño en los órganos blanco de la hipertensión, el estudio de cohorte puede observar menos morbilidad y mortalidad asociado con la hipertensión que lo que se observaría si el estudio hubiera incluido participantes más jóvenes.

Aún un estudio de cohortes en una población de recién nacidos puede reclutar sólo niños de concepciones que resultan en nacimientos vivos. Si el humo de tabaco ambiental aumenta la tasa de pérdidas fetales tempranas, posiblemente no detectadas, pueden haber diferencias entre los fetos que mueren y aquellos que sobreviven hasta el nacimiento. Si los fetos que sobreviven son más resistentes a los daños por el humo ambiental, entonces un estudio de cohortes de los efectos dañinos del humo ambiental sobre los niños puede presentar un efecto más débil porque los casos expuestos más susceptibles nunca fueron reclutados para el estudio.

Suponiendo que las poblaciones blanco son definidas como las personas de 70 años o más (en el estudio de hipertensión) o recién nacidos (en el estudio del tabaco ambiental), la validez interna como la hemos definido no parecería estar afectada. Pero los hallazgos de los estudios pueden de igual manera ser engañosos. Si la medicación que disminuye el colesterol prolonga la vida libre de enfermedad en los hipertensos, entonces más hipertensos que usen esa medicación sobrevivirán hasta la edad de 70 para entrar en la cohorte. Si estos hipertensos tienen una mayor tasa de desarrollo de daño de los órganos blanco, las tasas observadas de morbilidad y mortalidad asociadas con la hipertensión será más alta, las personas que toman medicación para disminuir el colesterol, ahora tienen una mayor morbilidad y mortalidad, y el efecto más fuerte de la hipertensión se verá asociado con la medicación que reduce el colesterol. De igual manera, un factor que reduzca las pérdidas de fetos en los embarazos expuestos al humo ambiental aumentará la proporción de niños susceptibles al humo ambiental que podrán ser incluidos en el estudio de cohortes y se asociará con una mayor morbimortalidad infantil.

Este problema parecería acercarse más a una falta de validez externa (posibilidad de generalizar a través del tiempo o el lugar), pero se parece mucho a la sobrevida selectiva como se ve en un estudio de corte transversal o un estudio caso-control (p.ej., el ejemplo del estudio caso-control de la agresividad y el IAM, presentado anteriormente.) Así las pérdidas previas al establecimiento de la cohorte necesitan ser consideradas con atención para que el investigador no sea engañado por factores selectivos que actúan en una etapa temprana.

Sesgo de selección debido a datos faltantes

Otra causa potencial del sesgo de selección en estudios de todo tipo es la ausencia de datos para una variable necesaria en el análisis. El sesgo debido a falta de datos habitualmente es un tema considerado en el análisis, pero su efecto es similar a un sesgo de selección y su prevención requiere evitar las diferencias sistemáticas en proporciones de datos faltantes.

El problema puede ser particularmente severo en los análisis que involucran un gran número de variables. Por ejemplo, los procedimientos de regresión a menudo excluyen íntegramente una observación si le falta un valor para cualquiera de las variables en la regresión. Esta práctica (llamada "eliminación por pasos" ["listwise deletion"]) puede excluir gran porcentaje de observaciones e inducir un sesgo de selección, aún cuando sólo exista un 5% o 10% de valores faltantes para una variable cualquiera. Los procedimientos de imputación pueden a menudo evitar la exclusión de observaciones, y dependiendo de los procesos que llevan a que falten los datos (el "mecanismo" por el cual falta los datos) pueden llevar a análisis menos sesgados o no sesgados. También hay procedimientos analíticos que pueden disminuir el sesgo de no-respuesta (incapacidad para reclutar participantes) y/o atrición (pérdida de participantes después del reclutamiento.)

Sesgo de Información

El sesgo de información se refiere a la distorsión sistemática de las estimaciones resultando de una inexactitud en la medición o clasificación de las variables de estudio (la clasificación errónea es una subcategoría de sesgo de información cuando la variable tiene sólo un pequeño número de valores posibles.) Por ejemplo, una enfermedad puede estar presente pero no ser reconocida, una presión sanguínea puede ser mal tomada o mal registrada, el recuerdo de una exposición previa puede ser defectuoso, o en los casos extremos, los datos pueden simplemente haber sido fabricados por sujetos o personal de investigación poco dispuestos a cooperar. Típicas fuentes del sesgo de información / clasificación errónea son:

1. La variación entre los observadores y entre los instrumentos – o la variación a través del tiempo por el mismo observador o instrumento;
2. Variación en la característica subyacente (p.ej. presión sanguínea) – y la variación que no ha sido adecuadamente tratada por los métodos de estudio;
3. Malentendido de las preguntas por el sujeto que es entrevistado o que completa un cuestionario – o incapacidad o falta de voluntad de dar la respuesta correcta; o el recuerdo selectivo;
4. Datos incompletos o imprecisamente registrados.

Revisión sistemática:

El sesgo de información puede ocurrir con respecto a la enfermedad, la exposición, u otras variables relevantes. A veces, el sesgo de información puede ser medido, como cuando existen dos métodos de medición, uno considerado más exacto que el otro. A veces, se puede suponer que el sesgo de información existe pero no puede ser directamente evaluado.

Por ejemplo, si hay una verdadera relación causal entre los estrógenos y el cáncer de endometrio, i.e., un proceso biológico por el cual las moléculas de estrógenos inician o promueven el crecimiento celular canceroso en el endometrio, este proceso fisiopatológico presumiblemente se relaciona con ciertas especies moleculares específicas, operando durante un cierto período de tiempo, y resultando en ciertos tipos de cáncer endometrial. En la medida que el cáncer de endometrio es una entidad heterogénea, y la forma relacionada con los estrógenos es sólo un subtipo, entonces la asociación observada entre el cáncer de endometrio y el estrógeno es diluida, por decirlo así, por la combinación en un sólo grupo de casos, de cánceres causados por estrógenos y cánceres resultantes de otros mecanismos. El enmascaramiento de la relación también ocurre combinando en un grupo de exposición mujeres cuya exposición causó el cáncer, mujeres cuya exposición a los estrógenos ocurrió sólo antes o después del período de tiempo relevante en términos de la historia natural del cáncer de endometrio, y las mujeres que fueron expuestas a una forma no patógena de estrógeno, dosis no patógena, o forma de administración no patógena (si existiera.)

Otro ejemplo es el estudio de los efectos sobre la salud de la exposición al plomo. El índice tradicional de la absorción, la plombemia, refleja sólo la exposición reciente, porque la vida media del plomo en sangre es sólo de aproximadamente 36 días (ver Landrigan, 1994.) Por lo tanto puede haber poca relación entre una única determinación de plombemia y el contenido corporal de plomo. Estudios pioneros por Herbert Needleman, empleando mediciones de plomo en los dientes de leche permitieron demostrar la relación entre una exposición baja al plomo y los trastornos cognitivos y del comportamiento en niños. Hoy en día, la aparición del análisis de florescencia de rayos X K de plomo en los huesos, donde la vida media es del orden de los 25 años, puede ser una nueva herramienta importante en los estudios epidemiológicos sobre la exposición al plomo (Kosnett MJ y cols, 1994.)

Por estas razones, los diseños y las ejecuciones rigurosos de estudio emplean:

1. Verificación del diagnóstico de caso, empleando procedimientos como revisiones múltiples independientes de muestras de tejidos, rayos x, y otros datos diagnósticos;
2. Definición de subgrupos homogéneos, con análisis separado de los datos de cada uno;
3. Fuentes múltiples de datos en cuanto a la exposición (y otras variables relevantes), permitiendo que se corroboren unas a otras;
4. Caracterización precisa de la exposición actual, con respecto a tipo, período de tiempo, dosis, etc.

Lamentablemente, las restricciones de la realidad imponen compromisos. Por ejemplo, los datos de hace 20 años pueden ser los más relevantes en términos del modelo causal, pero los datos de hace dos años pueden ser más accesibles y precisos. Al usar los datos más recientes, uno tiene que suponer que la exposición reciente también está relacionada con la enfermedad, aunque talvez no tan estrechamente como la exposición previa.

[Para más sobre este tema ver Hulka, Hogue, y Greenberg, "Methodologic issues in epidemiologic studies of endometrial cancer and exogenous estrogen", y Kenneth J. Rothman, Induction and latent periods, *Am J Epidemiol* 114:253-259, 1981 (el trabajo de Rothman plantea el problema de la oportunidad de la exposición.)]

Una consideración planteada por lo anterior es la importancia del desarrollo de hipótesis específicas antes del estudio. Dichas hipótesis, si pueden ser elaboradas, fortalecen tanto el diseño como la interpretación del estudio. El diseño se ve fortalecido porque las hipótesis guían al investigador en la selección de las variables relevantes y sus características (momento de la ocurrencia, etc.) para obtener los datos. Puede no ser posible obtener la información pero por lo menos las hipótesis guían la búsqueda. Las hipótesis también dan lineamientos sobre qué relaciones analizar y cómo construir variables para el análisis (p.ej., qué subcategorías relacionar con qué formas de la exposición.) Hipótesis específicas – basadas en teorías y conocimientos existentes – pueden también apoyar los hallazgos.

Términos y conceptos básicos

Confiabilidad (de una medida o una clasificación) tiene que ver con la reproducibilidad de una medida – a través del tiempo, con distintos instrumentos de medida, hechas por distintos observadores. Si una medida es confiable, no necesariamente es exacta. Pero si una medida no es confiable, entonces los valores de los datos tienen un importante componente aleatorio. Este componente aleatorio disminuye el contenido de información de la variable, la fuerza de las asociaciones que la involucran, y su efectividad para controlar el fenómeno de confusión (que se discutirá en el próximo capítulo.) El concepto de confiabilidad es relevante cuando dos o más medidas de similar autoridad son comparadas.

Validez (de una medida o una clasificación) es el punto al cual una medida mide lo que tiene que medir. La evaluación de la validez, por lo tanto, implica la disponibilidad de un método de medición que pueda ser considerado patrón (a menudo conocido como el "estándar de oro".) Dado que una medida es más confiable, nuestro interés cambia, y del acuerdo entre las medidas de evaluación pasa a la evaluación de la menos autoritaria. Por ejemplo, si la presión arterial promedio de varias tomas medidas con un esfingomanómetro de mercurio con un cero aleatorio en una personas acostada es nuestro estándar para la "verdadera" presión arterial, una toma casual en una persona sentada será sistemáticamente inexacta, dado que tenderá a ser más alta. Aunque examinaremos la coincidencia entre las presiones sanguíneas en posición acostada y la toma casual, nuestro interés es sobre la precisión de la segunda con respecto al "patrón oro".

Relación entre confiabilidad y validez

"Validez" se usa como término general para exactitud o precisión. El procedimiento para evaluar la precisión de un instrumento de medición es a menudo llamado validación. En muchas situaciones, sin embargo, no conocemos el resultado correcto, de manera que lo mejor que podemos hacer es comparar mediciones que se supone que son igualmente precisas. En estas situaciones, la concordancia entre las medidas se denomina "confiabilidad".

En este sentido, la confiabilidad es una subcategoría de la validez. Sin embargo, la confiabilidad (reproducibilidad, consistencia) puede estar presente sin validez (dos profesores pueden estar de acuerdo pero ambos equivocados!.) También, un procedimiento de medición puede ser válido en el sentido de que en promedio da el valor correcto, aunque cada medición incluye una gran cantidad de variación aleatoria (p.ej., recuerdo de 24 horas de la dieta para consumo de colesterol.) A veces se dice que "una medida que no es confiable no puede ser válida". Si esta aseveración es verdadera o no, depende del aspecto de la validez considerada. Más comúnmente, el error aleatorio (falta de confiabilidad) y el sesgo (falta de validez) se consideran como componentes independientes del error total.

La psicometría es la sub-disciplina de la sicología que trata sobre la evaluación de los ítems de los cuestionarios y las escalas. "Validación" como es usada en psicometría engloba tanto la confiabilidad (consistencia) como la validez. Sin embargo, debido a la escasez de clasificaciones y medidas que puedan ser consideradas acreditadas, mucha de la validación psicométrica tiene que ver con la

evaluación de la confiabilidad. Las situaciones comunes en que es importante examinar la confiabilidad son las comparaciones del desempeño de dos evaluadores o entrevistadores de igual envergadura (confiabilidad inter-evaluador), de resultados de mediciones repetidas de una característica que se cree que es estable (confiabilidad prueba-re-prueba), y puntajes de ítems equivalentes que conforman una escala (confiabilidad inter-ítem – a menudo llamado "consistencia interna".)

Evaluación de la confiabilidad

La validación involucra la medición de la concordancia entre dos o más medidas o clasificaciones. Concordancia no es idéntico a "asociación", más bien es un caso especial – un caso en que ambas medidas aumentan en la misma dirección y tienen la misma escala. (Un ejemplo obvio de una asociación que no implica concordancia es una asociación inversa.)

Concordancia porcentual

Para las variables categóricas, una medida simple de confiabilidad es el porcentaje de instancias en que dos instrumentos de medida concuerdan. Supongamos que se le dan 100 electrocardiogramas (ECG) a dos expertos para codificar independientemente como "anormal" o "normal", y que los dos están de acuerdo en 90 casos (30 que ambos llaman "anormales" y 60 que ambos consideran "normales".) Pero esta concordancia del 90% no es tan buena como parece, dado que le da "crédito" a un acuerdo que esperaríamos ocurriera por simple azar. ¿Qué pasaría si los codificadores, prefiriendo jugar al golf, les dejaran los ECG a sus secretarias con instrucciones de codificar cada ECG independientemente tirando un dado y designándolo como "anormal" si sale un 6? Para el desafortunado investigador, cuando verificara la confiabilidad de la codificación, encontraría que los codificadores parecerían tener una concordancia de 72% (3 anormales y 69 normales.)

Variables cualitativas - Kappa

Kappa es una medida de confiabilidad que ajusta la concordancia, por la que se espera que ocurriría por azar (introducida por Cohen en 1960.) Para una variable cualitativa sin un orden inherente (p.ej., diagnóstico inicial de dolor precordial como ¿angina? ¿reflujo gastroesofágico? ¿muscular?), kappa se calcula como:

$$K = \frac{p_o - p_c}{1 - p_c}$$

p_o = proporción observada de concordancia

p_c = proporción de concordancia esperada por azar

La proporción de concordancia esperada por azar se calcula usando los porcentajes marginales, el mismo procedimiento utilizado para el cálculo del chi cuadrado para asociación.

Supongamos que una organización de atención de salud está investigando la confiabilidad de las pruebas diagnósticas indicadas por los médicos a los pacientes con dolor precordial. Las evaluaciones iniciales de dos médicos y el orden de las pruebas diagnósticas indicadas se comparan para 100 pacientes consecutivos que se presentan con dolor precordial simple en su primera consulta a la organización.

Comparación de los diagnósticos del médico A y el médico B de 100 pacientes que presentan dolor precordial

		Médico B			
Médico A		¿Angina?	¿Reflujo?	¿Muscular?	Total
	¿Angina?	12	1	1	14
	¿Reflujo?	2	36	4	42
	¿Muscular?	2	8	34	44
	Total	16	45	39	100

Dado que los médicos concuerdan en el diagnóstico inicial en 12 + 36 + 34 pacientes, su porcentaje de concordancia es 82/100 = 82%. Sin embargo, basado en los marginales esperamos bastante concordancia sólo por azar. La proporción de concordancia por azar se calcula a partir de la distribución marginal como sigue:

Proporción de concordancia esperada =

	(Proporción ¿Angina? Médico A) × (Proporción ¿Angina? Médico B)	14/100 × 16/100
+	(Proporción ¿Reflujo? Médico A) × (Proporción ¿Reflujo? Médico B)	42/100 × 45/100
+	(Proporción ¿muscular? Médico A) × (Proporción ¿muscular? Médico B)	44/100 × 39/100

$$= 14/100 \times 16/100 + 42/100 \times 45/100 + 44/100 \times 39/100$$

$$= 0.0224 + 0.189 + 0.1716 = 0.383$$

El valor de Kappa para esta tabla es por lo tanto: $K = \dfrac{0.82 - 0.383}{1 - 0.383}$

Para evaluar la concordancia entre variables ordinales con pocas categorías, se usan versiones ponderadas de Kappa para asignar pesos variables a distintos grados de desacuerdo. Una discusión sobre Kappa se encuentra en el texto de Joseph Fleiss *Statistical Methods for Rates and Proportions*. La segunda edición sugiere adjetivos para caracterizar los valores de Kappa.

Variables continuas

Para medidas continuas y variables ordinales con muchas categorías, la forma de presentar los datos es mediante un gráfico de puntos, más que una tabla de contingencia. La concordancia perfecta significa que todos los pares de mediciones caen sobre una línea recta con una pendiente de 1 y un intercepto de 0 (i.e., la línea pasa por el origen.) Los indicadores más directos del nivel de concordancia entre las dos mediciones son el coeficiente de regresión y el intercepto de la línea recta que se ajusta mejor a los pares de medidas. Cuanto más se acerca el coeficiente de regresión a 1.0 (pendiente) y el intercepto de la regresión a 0, y cuanto más estrechos sus intervalos de confianza, mejor el nivel de concordancia.

Un índice común de concordancia es el coeficiente de correlación. El coeficiente de correlación producto-momento de Pearson (r) evalúa en qué medida los pares de observaciones de las dos mediciones caen sobre una línea recta. El coeficiente de correlación (de rango) de Spearman, rho, usado para variables ordinales, evalúa hasta que punto los pares de observaciones tienen el mismo orden para los dos instrumentos de medición (el más bajo para el primer instrumento está cerca de las mediciones más bajas del segundo instrumento, la décima menor del primer instrumento se acerca a la décima menor para el segundo, y así sucesivamente.)

Sin embargo, los coeficientes de correlación ignoran la localización y las escalas. Por lo tanto, si las lecturas de un termómetro están siempre exactamente dos grados por debajo de las lecturas de un segundo termómetro, la concordancia no es por cierto perfecta, sin embargo el coeficiente de correlación entre sus lecturas será de 1.0 (para una recta perfecta con pendiente 1.0, pero que no pasa por el origen.) Si las lecturas del primer termómetro son siempre el doble de las del segundo, la correlación también será 1.0 (para una recta a través del origen, pero con una pendiente de 2.0.) Por lo tanto un coeficiente de correlación aislado no es una evaluación adecuada de la concordancia. Debe estar acompañado de una comparación de indicadores de posición (media, mediana) y de escala (desvío estándar u otra medida de dispersión) para las lecturas de dos medidas.

Confiabilidad de una escala [opcional]

Una medida de confiabilidad que se usa ampliamente en psicometría es el coeficiente alfa de Cronbach. Los fundamentos conceptuales del coeficiente alfa (ver Nunnally, *Psychometric Theory*) pueden ser planteados como sigue.

Supongamos que tenemos una serie de preguntas en un cuestionario y que cada una intenta medir el mismo constructo no observable (una "variable latente".) El valor de respuesta de cualquier pregunta individual reflejará el valor de esa variable latente pero también algo de error, que se supone es aleatorio, independiente de todo lo demás, y simétricamente distribuido con un promedio de cero. Bajo estos supuestos el promedio de los valores de respuesta para el conjunto de preguntas dará una mejor medida de la variable latente que la que se consigue con cualquier pregunta aislada (igual que el valor promedio de un conjunto de tomas de presión arterial dará un valor más exacto (preciso) que cualquiera de las medidas

individuales.) Los componentes aleatorios en las respuestas a las preguntas deberían compensarse, de manera que el promedio es una medida más precisa de la variable latente.

En un escenario como el descrito, el coeficiente alfa evalúa cuánto de las puntuaciones de la escala es reflejo de los valores de la variable latente y cuánto es reflejo del error de medición. Cuanto mayor "la varianza compartida por los ítems" (cuantos más ítems individuales en la escala coinciden unos con otros) y cuanto mayor el número de ítems, mayor el valor de alfa. Planteado más concretamente, el coeficiente alfa es la proporción de la varianza total de los puntajes de la escala que representa la varianza de los valores de la variable latente (siendo el resto la varianza de los errores aleatorios de cada ítem.)Valores de alfa de 0.80 son considerados adecuados para calcular correlaciones y ajustar modelos de regresión, y un tamaño muestral de 400 observaciones es considerado suficiente para estimar alfa (ver Nunally.)

Los obstáculos para cumplir con este escenario ideal incluyen la probabilidad de que los ítems no son perfectamente equivalentes, que las respuestas de algunas personas a algunos ítems en la escala afectan sus respuestas a otros ítems, (de manera que los errores en las respuestas a los ítems no son independientes), y que haya factores además de la variable latente contribuyendo con variación no aleatoria en las respuestas a los ítems (introduciendo así un error sistemático, i.e., sesgo.) Señalemos que el coeficiente alfa no trata el sesgo, sólo se ocupa de la variabilidad aleatoria.

Evaluación de la validez – sensibilidad y especificidad

Como señalamos antes, la evaluación de la validez se dirige a la evaluación de un instrumento de medición o evaluador en su comparación con un instrumento o evaluador con autoridad. Para la detección de una característica o una condición, los epidemiólogos generalmente emplean los conceptos de sensibilidad y especificidad que fueron introducidos en un capítulo anterior. Usando las palabras "caso" ("no caso") para denominar respectivamente, las personas que tienen (no tienen) la condición o característica (p.ej., una enfermedad, una exposición, un gen) que se desea medir, la sensibilidad y la especificidad son, respectivamente, las probabilidades de clasificar correctamente los casos y los no casos.

Sensibilidad es la capacidad de detectar un caso.

Especificidad es la capacidad de detectar un no caso.

Ejemplo:

Si un procedimiento identifica correctamente 81 de las 90 personas con la enfermedad, condición o característica, la sensibilidad del procedimiento es:

Se = 81/90 = 0.9 = 90%

Si el mismo procedimiento identifica correctamente 70 de las 80 personas sin la enfermedad, condición o característica, entonces la especificidad del procedimiento es:

$$Esp = 70/80 = 0.875 = 88\%$$

En términos de probabilidad,

$$Se = Pr(D'|D) \qquad Esp = PR(\overline{D}'|\overline{D})$$

donde D = caso, $\overline{D}$ = no caso, D' = "clasificado como 'caso'", y $\overline{D}$' = "clasificado como 'no caso'".

La inversa de la sensibilidad y la especificidad son los "falsos negativos" y los "falsos positivos". Algunos autores prefieren evitar estos términos, por la potencial confusión sobre si "negativo" y "positivo" se refieren a la prueba (de acuerdo a la definición en el *Diccionario de Epidemiología* de John Last) o a la verdadera condición. Sin embargo, los términos son de uso habitual, y nosotros seguiremos la definición del Diccionario, por la cual "falso negativo" es un resultado de la prueba negativo en una persona con la característica (i.e., una prueba negativa equivocada) y un "falso positivo" es un resultado de la prueba positiva equivocado.

La sensibilidad y la especificidad como las hemos definido anteriormente sufren de las mismas limitaciones que hemos señalado para la concordancia porcentual, es decir que sus cálculos no tienen en cuenta la concordancia esperada en forma aleatoria. Hasta un proceso aleatorio clasificará algunos casos y no casos correctamente. Han sido publicados métodos para subsanar esta limitación (Roger Marshall, "Misclassification of exposure in case-control studies", *Epidemiology* 1994;5:309-314), pero aún no son usados ampliamente.

Impacto de la clasificación errónea

El impacto de la clasificación errónea sobre las estimaciones de tasas, proporciones, y medidas de efecto depende de las circunstancias. Consideremos el siguiente ejemplo para una enfermedad rara. Supongamos una cohorte de 1,000 participantes, de los cuales 60 desarrollan ECC durante un intervalo de cuatro años.

Si la sensibilidad de nuestro método diagnóstico es sólo de 0.80 (80%), entonces detectaremos sólo 48 casos (48/60, i.e., 80% de 60.) Habrán 12 falsos negativos.

Si la especificidad de nuestros métodos diagnósticos es de 0.90 (o 90%), entonces clasificaremos incorrectamente 94 de los 940 sujetos que no desarrollaron ECC (90% de los 940 no casos serán correctamente clasificados como tales, dejando 94 (940 menos 846) no casos que se clasificarán incorrectamente como "casos". Estos 94 sujetos serán falsos positivos.

Así, observaremos (o creeremos observar) 142 "casos" de ECC – 48 que de hecho tienen ECC y 94 que en realidad no la tienen. Es de resaltar que la mayor parte de los "casos" no tienen la enfermedad! Este ejemplo ilustra el dilema de los falsos positivos cuando se estudia una enfermedad rara. Los falsos positivos y sus características "diluirán" o distorsionarán las características de cualquier grupo de "casos" que podríamos juntar. De ahí el énfasis que se hace sobre la posibilidad de evitar los falsos positivos a través de la verificación de casos, usando métodos como la confirmación anátomo-patológica.

Supongamos que los participantes de esta cohorte son "expuestos", y existe otra cohorte similar compuesta por 1,000 participantes que no están "expuestos". Supongamos que la precisión diagnóstica no es influida por la situación de exposición, por lo que esperamos que los resultados para las dos cohortes sean como sigue:

Escenario hipotético para mostrar el efecto de sesgo de clasificación errónea sobre las medidas de asociación

Verdadero			Observado									
(Se=1.0, Esp=1.0)			Se=0.8, Esp =1.0			Se=1.0, Esp =0.9			Se=0.8, Esp =0.9			
	Exp	$\overline{\text{Exp}}$		Exp	$\overline{\text{Exp}}$		Exp	$\overline{\text{Exp}}$		Exp	$\overline{\text{Exp}}$
Enf	60	30	Enf	48	24	Enf	154	127	Enf	142	121
$\overline{\text{Enf}}$	940	970	$\overline{\text{Enf}}$	952	976	$\overline{\text{Enf}}$	846	873	$\overline{\text{Enf}}$	858	879
	1,000	1,000		1,000	1,000		1,000	1,000		1,000	1,000
RR	2.0			2.0			1.21			1.17	
DR	0.03			0.024			0.027			0.021	

A partir de este ejemplo, podemos ver que:

1. Aún altos niveles de sensibilidad y especificidad no impiden el sesgo;
2. Diferentes medidas epidemiológicas son afectadas de diferente manera;
3. El RR no necesariamente es afectado por la sensibilidad imperfecta si la especificidad es suficientemente alta; sin embargo la diferencia de tasas sí se afectará.
4. El RR será afectado por una especificidad imperfecta para la detección de enfermedades raras aún cuando la sensibilidad es alta; sin embargo, la diferencia de tasas puede ser afectada sólo levemente.
5. La especificidad es de máxima importancia para el estudio de una enfermedad rara, dado que es más fácil tener más pruebas falso positivo que verdaderos casos, identificados o no;

6. El sesgo en la clasificación de una enfermedad dicotómica típicamente disimula una verdadera asociación, si la clasificación errónea es la misma para los grupos expuestos y no expuestos.

[Intenta crear una planilla para ver como distintos niveles de sensibilidad y especificidad cambian el RR y la diferencia de tasas. Las fórmulas adecuadas se encuentran en el apéndice.]

Tipos de clasificación errónea

El ejemplo anterior trata de la clasificación errónea de una enfermedad, clasificación errónea que es independiente del estado de exposición. La clasificación errónea no diferencial de una variable de exposición dicotómica, i.e., la clasificación errónea que ocurre independientemente del estado de enfermedad – sesgará las medidas de efecto de razón hacia el valor nulo 1.0. Esto también ocurrirá con la clasificación errónea no diferencial de una variable de enfermedad dicotómica o de tanto una variable de enfermedad dicotómica y una variable de exposición dicotómica simultáneamente.

Sin embargo, la clasificación errónea diferencial, en que los errores en la medición de una variable pueden variar según el valor de otra variable, puede producir sesgos en cualquier dirección. Escenarios comunes para la clasificación errónea diferencial son el recuerdo selectivo de exposiciones pasadas o la detección selectiva de la enfermedad basada en el conocimiento de la historia de exposición del paciente. También, cuando la variable erróneamente clasificada tiene más de dos niveles, aún la clasificación errónea no diferencial puede producir un sesgo en cualquier sentido (como este último punto recién se ha enfatizado en los últimos años y como tradicionalmente la enseñanza de epidemiología ha hecho hincapié en variables de enfermedad y de exposición dicotómicas, no es raro escuchar la aseveración de que "el sesgo de la clasificación errónea no diferencial es hacia el valor nulo" sin mencionar las excepciones.)

Además, el error de medición en las otras variables involucradas en el análisis produce un sesgo en una dirección que depende de la relación entre las variables. Por ejemplo, si estamos realizando un ajuste por edad y tenemos un sesgo en la medición de la edad, el ajuste por edad no va a quitar totalmente el efecto de la edad. Una situación de este tipo se conoce como un sesgo de información en la medición de una covariable y es discutido en el texto de Rothman y Greenland.

Dirección y alcance del sesgo

La importancia de poder discernir la dirección del sesgo y, si es posible, evaluar su magnitud, es de permitir la interpretación de los datos observados. Por ejemplo, si se observa una asociación positiva entre dos factores y se puede demostrar que la dirección de la clasificación errónea es hacia el valor nulo, entonces dicho sesgo no puede ser responsable del hallazgo de una asociación positiva. De igual forma, si el sesgo de clasificación errónea tiene dirección positiva, entonces la falta de hallazgo de una asociación no puede deberse a ese sesgo. Además existen técnicas para corregir los errores de medición en varios procedimientos analíticos. Sin embargo, estos procedimientos a menudo requieren alguna estimación externa de sensibilidad y especificidad.

Donde las categorías de sesgo pierden nitidez

Anteriormente mencionamos que los límites entre el error aleatorio y el error sistemático además de los límites entre los tres tipos de sesgos a veces son borrosos. A continuación presentamos situaciones que son un desafío para clasificar.

Falsos negativos en la detección de casos para un estudio caso-control

Si no se detectan los casos por sesgo de información, esas personas no serán contadas como casos en un estudio caso-control. Si esta falta de sensibilidad de alguna manera se relaciona con la situación de exposición (p.ej., mayor detección de cáncer de endometrio entre mujeres que toman estrógenos – el tema del sesgo de detección), el grupo de casos no será representativo de la población de casos.

Desde el punto de vista del estudio caso-control este tipo de sesgo será clasificado como un sesgo de selección, dado que se manifiesta a través de la probabilidad de selección diferencial para los casos. Pero el mecanismo de error en este escenario fue la clasificación errónea de casos. Es más, si algunas mujeres con cáncer de endometrio asintomático llegaran a ser seleccionadas como controles, su presencia en el grupo control se clasifica presumiblemente como sesgo de información (clasificación errónea), dado que en esta situación los sujetos pertenecen al estudio, pero deberían estar en el grupo de casos.

La variabilidad en el parámetro que se mide puede producir tanto un error aleatorio (imprecisión de la medición) como un sesgo de información en la medida de efecto

La presión sanguínea, por ejemplo, varía minuto a minuto, de manera que cada medición de la presión sanguínea refleja cierto grado de variabilidad aleatoria (error aleatorio.) Si la presión sanguínea se mide una sola vez, y luego en los siguientes 5 años se registra la incidencia de una enfermedad u otra medida de resultado, la asociación observada entre la presión sanguínea y el resultado subestimará cualquier asociación verdadera.

La razón de esto es que entre los sujetos que fueron clasificados como valores "altos" en su medición inicial estarán incluidos algunos que tenían valores "altos" sólo por casualidad. Entre los sujetos clasificados como con valores "bajos" se incluirán algunos que tenían valores "bajos" por azar. El error resultante en la medición de la exposición disimulará el contraste entre los resultados de los grupos en comparación con lo que se obtendría si nuestro grupo con valores "altos" contuviera solo aquellos que verdaderamente tienen valores "altos" y el grupo de los bajos contuviera sólo aquellos que verdaderamente tenían valores "bajos".

Suponiendo que la variabilidad aleatoria es independiente de los resultados del estudio, entonces el resultado es una clasificación errónea no diferencial, y la asociación observada será más débil que la "verdadera" asociación. Así el error aleatorio puede producir un error sistemático, o sesgo.

La presión arterial (y otros parámetros fisiológicos) también varía con un patrón diurno, siendo menor en la mañana y elevándose durante el día. Si no se toma en cuenta la variación diurna se generarán todo tipo de errores. Por ejemplo, si las presiones sanguíneas son medidas en los participantes del estudio al azar durante el día (i.e., a cada sujeto se le mide la presión arterial una vez, pero esa medición puede ocurrir en cualquier momento del día), la variabilidad diurna agrega un componente de error aleatorio a aquel debido a la variabilidad de momento a momento. Por lo tanto, las estimaciones de las medias grupales y sus diferencias serán más imprecisas que si las mediciones hubieran sido hechas el mismo momento en el día.

Si por alguna razón los sujetos en una categoría (p.ej. oficinistas) son examinados en la mañana y los participantes de otra categoría (p.ej., amas de casa) son examinados en la tarde, va a haber una diferencia sistemática entre las presiones arteriales promedio de los sujetos en las distintas categorías, una diferencia sistemática que surge de la variación sistemática de la presión arterial combinada con una diferencia sistemática en el momento de la medición. El error sistemático resultante puede llevar a un sesgo de selección o un sesgo de información dependiendo de la naturaleza del estudio.

Regresión a la media

Un fenómeno muy conocido que demuestra como la variabilidad aleatoria puede producir un error sistemático es la regresión a la media (Davis CE: El efecto de la regresión a la media en los estudios epidemiológicos y clínicos [The effect of regression to the mean in epidemiologic and clinical studies] *Am J Epidemiol* 104:493-498, 1976.) Cuando una variable continua, como la presión arterial o el colesterol sérico, tiene cierto grado de variabilidad aleatoria asociada (con el fenómeno mismo o con su medición), cada medición puede considerarse basada en el "verdadero valor" para el sujeto más o menos un factor de ruido aleatorio Si la distribución de la variable aleatoria es simétrica con un promedio de cero, el valor promedio de una serie de lecturas diferentes se acercará al "verdadero valor". Si el componente aleatorio es importante, sin embargo, cualquier medida se puede encontrar sustancialmente por encima o por debajo del promedio.

En semejante situación, una variable para la cual una medición dada cae en extremo superior o inferior de la distribución de esa variable, tenderá a estar más cerca del centro de la distribución en una medición posterior. Por ejemplo, en el Estudio de Prevalencia del Lipids Research Clinics, las poblaciones fueron tamizadas con respecto a los niveles de colesterol y triglicéridos, y aquellos con niveles elevados fueron invitados a retornar para una evaluación adicional. Si, digamos, 15% de los sujetos tamizados fueran invitados a volver, se puede esperar (y es lo que ocurrió) que muchos de esos sujetos no tenían niveles elevados cuando se midieron de nuevo.

La razón de esta "regresión" es que el grupo de sujetos en el 15% superior de la distribución de los lípidos en su visita de tamizaje consistía en sujetos cuyos valores de lípidos fueron altos por un componente aleatorio positivo de magnitud importante además de los sujetos cuyos niveles de lípidos realmente eran altos. Al medirlos de nuevo, el componente aleatorio será, en promedio, menor o negativo, de manera que los sujetos sin niveles verdaderamente altos de lípidos caerán por debajo del punto de corte, igual que otros sujetos con verdaderos niveles altos pero que en esta medición sufren un componente aleatorio negativo de gran magnitud

Si por valor extremo queremos decir uno que es "extraordinariamente alto", esto implica que habitualmente debería ser menor. Lo opuesto es verdad para los valores bajos. Por lo tanto, el valor promedio de colesterol en una población no seleccionada no tenderá a "regresar hacia la media", dado que en un proceso aleatorio los aumentos y disminuciones se compensarán. Pero si seleccionamos una parte de la población basado en que sus mediciones iniciales son altas (y/o bajas), esa población seleccionada tenderá a "regresar" hacia el promedio poblacional.

En la regresión hacia la media, tenemos una situación en que la variabilidad aleatoria puede producir una distorsión sistemática, en el sentido de que el promedio de los niveles de colesterol (o de las presiones arteriales) de los sujetos con valores "elevados" sobrestima su "verdadero promedio" (suponiendo que "verdadero" se define como el promedio de varias mediciones.) Si esta distorsión produce un sesgo de selección o un sesgo de información dependerá del proceso real del estudio.

Supongamos que los sujetos de "alto riesgo" (con valores de colesterol y presión arterial elevados, u otros factores de riesgo cardiovascular) son ingresados a un programa de "salud" y sus niveles de riesgo son medidos varios meses después, probablemente encontremos una disminución en los niveles medidos, más allá de los efectos del programa, simplemente debido a la regresión a la media. Este proceso es una de las razones de la importancia de un grupo control adjudicado aleatoriamente, que se espera que presente la misma regresión.

[Según John R. Nesselroade, Stephen M. Stigler, y Paul B. Baltes, la regresión a la medio no es un proceso ubicuo sino que depende de las características del modelo o proceso subyacente. Se puede hallar un tratamiento profundo, pero fundamentalmente estadístico, del tema en "Regression toward the mean and the study of change," Psychological Bulletin 88(3):622-637, 1980.]

Apéndice 1

Fórmulas para ver los efectos de varios niveles de sensibilidad y especificidad sobre el RR y la diferencia de riesgos:

Si a, b, c, d son los VERDADEROS valores de las celdas de la tabla de contingencia, entonces el RR observado y la diferencia de riesgos (DR) observada en presencia de Sensibilidad (S) y Especificidad (Esp) para medir la enfermedad están dados por:

$$\text{RR observado} = \frac{[(S)a + (1\text{-}Esp)c]/n_1}{[(S)b + (1\text{-}Esp)d]/n_0}$$

$$\text{DR observedo} = \frac{(S)a + (1\text{-}Esp)c}{n_1} - \frac{(S)b + (1\text{-}Esp)d}{n_0}$$

$$= S\left(\frac{a}{n_1} - \frac{b}{n_0}\right) + (1 - Esp)\left(\frac{c}{n_1} - \frac{d}{n_0}\right)$$

Apéndice 2

Más sobre la preocupación sobre como evitar los diagnósticos falsos positivos de la enfermedad en estudios caso-control que estudian enfermedades raras (p.ej., cáncer de endometrio y estrógenos) –la importancia de la verificación del estado de enfermedad: [Esta es la versión simplificada de la presentación del artículo de Hulka, Hogue, y Greenberg que se encuentra en la bibliografía.]

La estrategia del caso-control apunta a estimar la probabilidad de exposición en los casos y en los no casos, siendo lo ideal para estos últimos la población general de la cual surgieron los casos. La clasificación errónea de la enfermedad lleva a la contaminación de estas probabilidades estimadas. En particular, los falsos positivos "diluyen" los casos:

La probabilidad observada de exposición en los sujetos clasificados como "casos" es igual a:

1. la probabilidad de exposición en los casos verdaderos
2. más una distorsión igual a la proporción de "casos" falso positivos multiplicado por la diferencia en probabilidad de exposición entre los verdaderos casos y los falsos positivos.

Algebraicamente,

$$\Pr(\text{Exp}|\text{Enf}') = \quad \text{— la prevalencia de exposición } \textbf{observada} \text{ en los "casos"}$$

$$\Pr(\text{Exp}|\text{Enf}) \quad \text{— la } \textbf{verdadera} \text{ prevalencia de exposición en los casos}$$

$$+ \Pr(\overline{\text{Enf}}|\text{Enf}') \, [\Pr(\text{Exp}|\overline{\text{Enf}}) - \Pr(\text{Exp}|\text{Enf})] \quad \text{— el } \textbf{sesgo}$$

donde Exp = exposición, Enf = un **verdadero** caso, $\overline{\text{Enf}}$ = un **verdadero** no caso,
y Enf' = **cualquier** sujeto que se clasifica como "caso" (correcta o incorrectamente)

De manera que $\Pr(\overline{\text{Enf}}|\text{Enf}')$ es la probabilidad de que alguien que se **denomina** un "caso" de hecho sea un **no-caso**.

y $\Pr(\text{Exp}|\overline{\text{Enf}})$ es la probabilidad de que un **verdadero no caso** sea expuesto.

En forma correspondiente, la probabilidad observada de exposición en los sujetos clasificados como **"no casos"** es igual a:

1. la probabilidad de exposición en los verdaderos **no casos**
2. más una distorsión igual a la proporción de falsos negativos entre las personas clasificadas como "no casos" multiplicado por la diferencia de probabilidad entre los verdaderos no casos y los falsos negativos.

Algebraicamente,

$\Pr(Exp|\overline{Enf}') =$ — la prevalencia de exposición **observada** en los "no casos"

$\Pr(Exp|\overline{Enf})$ — la **verdadera** prevalencia de exposición en los no casos

$+ \Pr(\overline{Enf}|Enf') [\Pr(Exp|Enf) - \Pr(\overline{Exp}|Enf)]$ — el **sesgo**

donde $\overline{Enf}'$ = cualquier sujeto clasificado como un "no caso" (correcta o incorrectamente)

Ejemplo numérico:

Si:

la probabilidad de exposición en los verdaderos casos = 0.4,

la probabilidad de exposición en los verdaderos no casos = 0.2,

la probabilidad de exposición en los falsos positivos = 0.2 (i.e., los falsos positivos son en realidad igual que cualquier otro no caso)

por lo tanto en una muestra de sujetos clasificados como "casos" en que un tercio está equivocadamente clasificado como tal (i.e., falsos positivos) esperamos observar una probabilidad de exposición de:

$$\Pr(Exp|Enf') = 0.4 + (1/3)\ [0.2\text{-}0.4] = 0.4 - (1/3)[0.2] = 0.333$$

o en forma equivalente,

$$\Pr(Exp|Enf') = (2/3)(0.4) + (1/3)(0.2) = 0.333$$

(i.e., la prevalencia de la exposición es un promedio ponderado de la prevalencia de exposición en los casos correctamente clasificados y la prevalencia de exposición en los falsos positivos.)

Dado que la verdadera probabilidad de exposición en los casos es 0.4, los resultados observados están sesgados hacia abajo. Dado que la proporción de falsos negativos en el grupo control (sujetos enfermos clasificados como "no casos") en general será pequeña si la enfermedad es rara, la estimación de la probabilidad de exposición en los no casos generalmente no estará sesgada.

El verdadero OR es 2.67 [{.4/(1-.4)}/{.2/(1-.2)}]; el OR observado es 2.0 [{.333/(1-.333)}/{.2/(1-.2)}]. La discrepancia sería aún mayor si las verdaderas probabilidades de exposición fueran más diferentes.

Bibliografía

Henneckens and Buring. *Epidemiology in Medicine*. Rothman and Greenland, 1998. Rothman. *Modern Epidemiology*. Chapters 7, 8. Kleinbaum, Kupper and Morgenstern. *Epidemiologic Research: Principles and Quantitative Methods*. Chapter 10, Introduction to Validity.

Armstrong BK, White E, Saracci Rodolfo. Principles of exposure measurement in epidemiology. NY, Oxford, 1992, 351 pp., $59.95. Key reference (reviewed in Am J Epidemiol, April 15, 1994.)

Brenner, Hermann and David A. Savitz. The effects of sensitivity and specificity of case selection on validity, sample size, precision, and power in hospital-based case-control studies. Am J Epidemiol 1990; 132:181-192.

Cohen, Bruce B.; Robet Pokras, M. Sue Meads, William Mark Krushat. How will diagnosis-related groups affect epidemiologic research? Am J Epidemiol 1987; 126:1-9.

Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol 1990; 132:746-8; and correspondence in 1991;134:441-2 and 135(12):1429-1431

Feinleib, Manning. Biases and weak associations. Preventive Medicine 1987; 16:150-164 (from Workshop on Guidelines to the Epidemiology of Weak Associations)

Feinstein AR, Horwitz RI. Double standards, scientific methods, and epidemiologic research. New Engl J Med 1982; 307:1611-1617.

Feinstein, Alvan R.; Stephen D. Walter, Ralph I. Horwitz. An analysis of Berkson's Bias in case-control studies. J Chron Dis 1986; 39:495-504.

Flanders, W. Dana; Harland Austin. Possibility of selection bias in matched case-control studies using friend controls. *Am J Epidemiol* 1986; 124:150-153.

Flanders, W. Dana; Coleen A. Boyle, John R. Boring. Bias associated with differential hospitalization rates in incident case-control studies. *J Clin Epidemiol* 1989; 42:395-402 (deals with Berkson's bias for incident case control studies - not a major work)

Flegal, Katherine M., Cavell Brownie, Jere D. Haas. The effects of exposure misclassification on estimates of relative risk. Am J Epidemiol 1986; 123:736-51.

Flegal, Katherine M.; Penelope M. Keyl, and F. Javier Nieto. Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol 1991; 134(10):1233-44.

Gregorio, David L.; James R. Marshall, Maria Zielezny. Fluctuations in odds ratios due to variance differences in case-control studies. *Am J Epidemiol* 1985; 121:767-74.

Horwitz, Ralph I. Comparison of epidemiologic data from multiple sources. J Chron Dis 1986; 39:889-896.

Horwitz, Ralph I. and Alvan R. Feinstein. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. N Engl J Med 1978;299:1089-1094.

Hulka, Barbara S., Carol J.R. Hogue, and Bernard G. Greenberg. Methodologic issues in epidemiologic studies of endometrial cancer and exogenous estrogen. Amer J Epidemiol 1978; 107:267-276.

Hulka, BS, Grimson RC, Greenberg BG, et al. "Alternative" controls in a case-control study of endometrial cancer and exogenous estrogen. Am J Epidemiol 1980;112:376-387.

Hutchison, George B. and Kenneth J. Rothman. Correcting a bias? Editorial. N Engl J Med 1978;299:1129-1130.

Kosnett JM, Becker CE, Osterich JD, Kelly TJ, Pasta DJ. Factors influencing bone lead concentration in a suburban community assessed by noninvasive K X-ray fluorescence. JAMA 1994;271:197-203

Landrigan, Philip J. Direct measurement of lead in bone: a promising biomarker. Editorial. JAMA 1994;271:239-240.

Maclure, Malcolm; Walter C. Willett. Misinterpretation and misuse of the Kappa statistic. *Am J Epidemiol* 1987; 126:161-169.

Marshall, James R. and Saxon Graham. Use of dual responses to increase validity of case-control studies. *J Chron Dis* 1984;37:107-114. (also commentary by Stephen D. Walter, authors' reply, and Walter's reply in that issue.)

Neugebauer, Richard and Stephen Ng. Differential recall as a source of bias in epidemiologic research. *J Clin Epidemiol* 1990; 43(12):1337-41.

Nunnally, Jum C. and Ira H. Bernstein. *Psychometric theory*. New York: McGraw-Hill, 1994.

Roberts, Robin S., Walter O. Spitzer, Terry Delmore, David L. Sackett. An empirical demonstration of Berkson's bias. *J Chron Dis* 1978; 31:119-128.

Sackett, D.L.: Bias in analytic research. *J Chron Dis* 32:51-63, 1979. (and comment.) In Ibrahim: *The Case-Control Study*.

Schatzkin, Arthur; Eric Slud. Competing risks bias arising from an omitted risk factor. *Am J Epidemiol* 1989; 129:850-6.

Sosenko, Jay M.; Laurence B. Gardner. Attribute frequency and misclassification bias. *J Chron Dis* 1987; 40:203-207.

Walker, Alexander M. Comparing imperfect measures of exposure. *Am J Epidemiol* 1985; 121:783-79

Warnecke RB, Johnson TP, Chavez N, Sudman S, O'Rourke DP, Lacey L, Horm J. Improving question wording in surveys of culturally diverse populations *Ann Epidemiol* 1997; 7:334-342.

Warnecke RB, Sudman S, Johnson TP, O'Rourke D, Davis AM, Jobe JB. Cognitive aspects of recalling and reporting health-related events: Pap smears, clinical breast exams, and mammograms. *Am J Epidemiol* 1997; 13(3):305-315.

White, Emily. The effect of misclassification of disease status in follow-up studies: implications for selecting disease classification criteria. *Am J Epidemiol* 1986; 124:816-825.