Principios de *Estadística en Salud*

1. Principios

El primer video esboza 5 aspectos clave para leer críticamente.

1.1. Determinismo frente a variabilidad

La estadística sobraría si no hubiera variabilidad. La estadística modela las fuentes de variación, distinguiendo entre previsibles y aleatorias, para cuantificar certezas y dudas. Por ejemplo, la predicción del tiempo no puede ser exacta; y por ello, la pregunta adecuada sería: "el parte meteorológico ¿cuánto reduce la duda sobre el tiempo que hará mañana?".

Ahora bien, no percibamos la variabilidad como maligna: incluso yo pararía los penaltis a Messi si él los lanzara siempre por el mismo sitio. Y la vida sería aburrida si pudiéramos adivinar siempre a nuestros seres queridos.

En resumen, la estadística cuantifica el grado de certeza de sus afirmaciones.

1.2. Objetivos clínicos

Para situar cada estudio en su contexto, veamos qué preguntas del paciente debemos responder: desde las más descriptivas "*¿qué me pasa?*" o "*¿qué me pasará?*" hasta las más ambiciosas "*¿por qué me pasa?*", o especialmente: "*¿Puede Vd. hacer algo para cambiar mi futuro?*". Corresponden a los objetivos clásicos de diagnóstico, pronóstico, etiología e intervención.

Científicamente diríamos que el diagnóstico pretende clasificar a los pacientes en grupos internamente homogéneos y externamente heterogéneos: los pacientes de un mismo grupo serán similares entre sí y diferentes de los de otros grupos. El pronóstico pretende avanzar el futuro, es decir, reducir la incertidumbre sobre la evolución. La etiología pregunta por el pasado que nos ha traído hasta aquí, mientras que la intervención quiere mejorar el futuro.

Notemos que el diagnóstico es una pregunta puntual en el tiempo, sobre el presente, mientras que las otras tres son, o bien sobre el pasado, o bien sobre el futuro, y requieren un lapso de tiempo. Pero mientras diagnóstico y pronóstico se basan en la mera relación entre las variables estudiadas; etiología e intervención requieren, además, que dicha relación sea causal. Finalmente, veremos que la intervención requiere una causa asignable.

1.3. Guías de publicación según objetivos

Existe una guía de publicación para reportar los resultados de cada tipo de estudio: STARD para diagnóstico, TRIPOD para pronóstico, STROBE para etiología, o CONSORT para intervención. Además, la publicación del protocolo del ensayo se guía por SPIRIT; y las revisiones sistemáticas que acumulan la evidencia disponible, por PRISMA. La red EQUATOR las promueve y facilita encontrarlas.

1.4. ¿Investigación o Desarrollo?

Cada estudio pertenece a una fase de investigación o desarrollo: Ioannidis resaltó el contraste entre las probabilidades razonables de que se repliquen los resultados de estudios confirmatorios, un 85%, con las de estudios masivos de exploración de datos, que cifraba en un 1 por mil. Incluso Colon debió replicar su viaje antes de abrir una vía que conectara ambos mundos.

La I aporta novedades prometedoras, mientras que la D documenta propiedades de propuestas concretas antes de ofrecerlas a los pacientes. Colón descubrió casualmente América en su primer viaje, pero debió replicarlo antes de hacer realidad la conexión entre ambos mundos.

Así, los análisis exploratorios sugieren nuevas hipótesis. Y los confirmatorios, contrastan hipótesis previas. Ambos son correctos y necesarios. Pero no se engañe: no crea que su estudio confirma cuando tan sólo sugiere. A fin de cuentas, el mayor mérito de Colón estuvo en su primer viaje. Y el de Fleming, en interpretar bien un resultado inesperado. Y replicarlo, claro. Conviene que otros reproduzcan mis resultados en mis propios datos y los repliquen con otros datos.

1.5. Ciencia reproducible y replicable

Finalmente, NO por el hecho de estar publicado será cierto. La frase "la literatura científica médica sobre-estima beneficios, subestima peligros y tiene un impacto negativo en la atención al paciente" es, ni más ni menos, que de un grupo selecto de editores y científicos médicos.

Que la investigación en Salud sea sub-óptima es escandaloso. La falta de transparencia afecta a la seguridad. Publicar resultados que luego no se replican es despilfarrar recursos. Materiales y humanos, ya que el sacrificio de los voluntarios ha sido en vano. Las mejores revistas apelan a la transparencia. NIH y Nature tienen el remedio, pero antes de conocerlo debemos revisar otros principios.

Y con esto termino. Bueno no, recuerde que los objetivos preceden a los datos.

2. <u>Intervención: CONSORT (1)</u>

El segundo vídeo es el primero de los 2 dedicados a CONSORT, la guía para Ensayos Clínicos.

2.1. Tipos de ensayos clínicos según la fase de I+D

En el continuo desde la I hasta la D destaca el ensayo clínico de fase 3 sobre el que pivota la decisión de autorizar la intervención. La guía de publicación CONSORT informa los aspectos clave para mejorar la transparencia de su informe.

2.2. Referencia

La referencia marca la opción alternativa que, en general, deseamos mejorar. Por tanto, debe definirse con precisión. Puede ser la pauta definida en la guía de práctica clínica. O simplemente, el tratamiento habitual o "*standard of care*". Para permitir el enmascaramiento, suele acompañarse de un simulador de la intervención en estudio, quizás un placebo. Así, la referencia puede incluir el mejor tratamiento conocido más

un placebo inocuo. Al inicio, dividimos los pacientes en 2 grupos y a cada uno le asignamos una intervención. Conocer cómo son los voluntarios incluidos en el estudio permitirá después valorar si los resultados aplican a "mis" pacientes.

2.3. Descriptiva inicial

El documento explicativo de CONSORT proporciona este ejemplo. Veamos qué significan estos números, según resuman variables numéricas, como el colesterol LDL; o categóricas, como la historia de angina previa. El dato más relevante es el denominador: ¿cuántos voluntarios aportan información? En las variables cuantitativas, el resumen habitual es el promedio también llamado media. Como los valores concretos se alejan de la media, hay que decir cuánto, lo que valora la desviación típica. Una regla del dedo gordo, útil en variables simétricas, multiplica esta distancia por 2 y la suma y la resta de la media para ver dónde se sitúan estos pacientes. Es decir, 3 más, menos 2, da 5 y 1: el 95% de los casos incluidos tenían valores de LDL entre 1 y 5, aproximadamente. Finalmente, dicotomías como historia previa de angina se resumen reportando el número y porcentaje de síes.

En resumen, la tabla de características iniciales permite ver cómo son los pacientes reclutados y si abarcan todo el espectro dentro de los criterios de elegibilidad.

Además, el lector atento habrá observado cómo se parecen entre sí ambas muestras. ¿Milagro? No, técnica.

2.4. Comparabilidad inicial

Asignar al azar los casos a los grupos les otorga comparabilidad inicial. Por definición. Si se ha respetado el proceso y éste es correcto, no hay discusión: ambas muestras provienen de la misma población. Hay pequeñas diferencias, quizás por azar, claro. Pero éste azar la estadística lo sabe cuantificar y controlar. Por supuesto, esta comparabilidad inicial hay que mantenerla hasta el final mediante un seguimiento completo e idéntico en ambos grupos: no vale ni eliminar a los que no toleran la intervención, ni dar tratamientos adicionales en el grupo de referencia.

2.5. Efecto

Así, después de este seguimiento idéntico y completo, la única diferencia entre los grupos es la intervención, lo que permite comparar los grupos para cuantificar su efecto.

Interpretemos primero el efecto de una respuesta numérica de la que disponemos de esta cantidad de información. Esta tabla del documento CONSORT explicativo es un buen ejemplo porque muestra tanto la situación final como la inicial.

Ahora podemos escoger entre diferentes medidas del efecto del ejercicio. En el dolor en reposo, los tratados tienen 1.18 puntos menos de dolor. Pero como los controles partían con una ventaja de 0.11, la nueva estimación del efecto vale 1.29. El análisis pre-especificado por los autores proponía ajustar también por edad y antigüedad de los síntomas, método que proporciona el mismo valor. Todos ellos son válidos y estiman el mismo efecto de la intervención, con pequeñas oscilaciones por el azar de la asignación. Es preferible el ajustado porque suele ser más preciso.

Los intervalos de confianza ayudan a decidir si la intervención tiene algún efecto. En la primera variable, no hay evidencia de efecto del ejercicio. En cambio, los datos sí que apoyan que el ejercicio mejore el dolor en reposo.

Veamos ahora la interpretación de una respuesta dicotómica con sólo 2 posibles valores, si alcanza o no, el criterio de curación. Las proporciones de curados en tratados y controles son 87 y 23%, por lo que la intervención sube la proporción de curados un 63%: si un paciente opta por la intervención su probabilidad de alcanzar el objetivo sube un 63% Por su parte, el intervalo de confianza aclara que, por lo menos este aumento de la probabilidad vale 0.44. Observen que a esta diferencia de proporciones o probabilidades lleva la etiqueta "diferencia de riesgos", ya que la epidemiología, que suele estudiar eventos negativos introdujo esta medida en la investigación clínica.

Existen 2 formas de comparar dos números, haciendo restas o divisiones. En este ejemplo, se han estudiado 3 subgrupos de pacientes según su gravedad inicial: alta, media y baja Siendo más numerosos los últimos En todos los subgrupos, la intervención reduce la frecuencia del evento negativo a la cuarta parte, es decir, el tratamiento multiplica por un cuarto la probabilidad de sufrir el evento. Esta oportuna homogeneidad del efecto permite combinar todas las estimaciones en una sola, más fácil de comunicar: la intervención reduce los eventos a la cuarta parte sea cuál sea la gravedad inicial. Y más precisa, con menor incertidumbre.

Pero, ¿qué pasa si en lugar del cociente hacemos la diferencia? Ahora los efectos cambian: en los graves, la frecuencia del evento baja un 27%, por un 3% en los leves. Disminuir 27 eventos en 100 tratados implica que por cada 4 tratados, evitaremos un evento. Pero bajar 3 eventos de 100, requerirá tratar a 33, cifra que podría no justificar los eventos adversos de la intervención

Por esta razón, CONSORT pide reportar: la descriptiva de cada grupo por separado, el cociente de probabilidades y su diferencia. En resumen, no se deje impresionar por expresiones del tipo "la intervención divide por 2 la frecuencia de eventos". Interprete estas medidas del efecto junto con los datos en comparación.

3. Intervención: CONSORT (2)

Bienvenid@ al tercer vídeo, segundo dedicado a CONSORT.

3.1. Objetivo frente a Hipótesis

Conviene distinguir entre el objetivo, que será siempre subjetivo, pues depende de la persona; y la hipótesis, a la que una buena transparencia hará reproducible y replicable, objetivos básicos de la Ciencia y la Técnica. En un ensayo confirmatorio sobre el que pivota la decisión de autorizar la intervención el objetivo es garantizar que el 90% de los fármacos eficaces llegan a los pacientes; y a la vez, que el 95% de los no eficaces no lo hacen. Citamos en el primer vídeo referencias que cuestionan si una investigación no reproducible o no replicable tiene valor social o soporte ético.

3.2. Opciones para el control (o ajuste)

El diseño de experimentos controla las fuentes de variación durante su planificación. Además, el análisis permite ajustarlas. Eso sí, desviarnos del plan previsto en el protocolo una vez vistos los resultados implica renunciar a confirmar la hipótesis. Por ejemplo, definir bloques homogéneos de pacientes dentro de los cuales comparamos las intervenciones puede mejorar las propiedades estadísticas. En cambio, ampliar los criterios de elegibilidad de pacientes es muy simple, pero reduce la población de posibles beneficiarios de la intervención. La informática permite emplear asignaciones dinámicas, en las que las probabilidades de ser ubicado en un grupo dependen del equilibrio previo, de forma que se minimizan las diferencias entre los grupos en comparación.

3.3. Azar

Históricamente, reconocemos a Avicena como padre de la Medicina experimental; a Lindl, como introductor del control en las comparaciones entre tratamientos; y a Hill, por proponer el azar como la forma más ética de distribuir la nueva intervención entre los candidatos potenciales. Elevando al mismo tiempo el valor científico de la comparación.

¿Por qué es tan importante el azar? ¿Por qué tiene ventajas introducir entropía? Especialmente porque las fórmulas estadísticas permiten cuantificarlo. En cambio, la Estadística habitual no ofrece una base para cuantificar los errores no aleatorios. ¿Es representativa una muestra? Muy fácil, si es al azar, sí que lo es.

Observe en este gráfico que la dispersión de los ensayos aleatorizados es menor que la de estudios con controles externos concurrentes. Simplemente, cabe esperar menores diferencias entre tratados y controles del mismo centro que si fueran de centros diferentes. El reto es que las fórmulas estadísticas cuantifican la oscilación del azar. Analizar estos estudios como si fueran aleatorizados infra-estima su variabilidad, que tiene más resultados extremos, y conduce a un sesgo impredecible.

Sesgo que el ajuste estadístico no contrarresta.

El Ensayo Clínico usa el azar para repartir entre los grupos, no para extraer los casos de la población. Por ello, los grupos son comparables. Vd. dirá con razón que los dos grupos diferirán por azar. Perfecto, medidas como el error estándar o los intervalos de confianza cuantificarán este azar. Así, el azar explica las diferencias iniciales. Pero las finales pueden ser explicadas también por la intervención. Para rechazar la hipótesis intervención ineficaz, Fisher propuso descartar el azar cuando las diferencias muestrales cayeran en una zona de pequeña probabilidad. Neyman-Pearson fueron más allá y nos enseñaron cómo acotar los riesgos de decisiones erróneas.

Algunas variables son tan importantes que no conviene dejarlas en mano del azar. Si la gravedad inicial es tan importante que podría comprometer la interpretación de los resultados, no nos la jugaremos al azar. Recurriremos al control o al ajuste.

Pero vigilemos: la logística que el control necesita podría hacer irrealizable al estudio. Y no seamos ludópatas: no dejemos en manos del azar aquellas variables importantes que, si quedaran desequilibradas, dificultarían la interpretación de los resultados.

3.4. Control del seguimiento

Sin lugar a dudas, el primer criterio de calidad es el nivel de su seguimiento. El reciente y próximo estudio REVASCAT ha obtenido la evolución a los 3 meses de 206 pacientes con Ictus. Mérito de los investigadores, que siguieron y persiguieron a cada paciente; de los monitores, que promovieron la calidad de los datos; y de los diseñadores, que definieron una respuesta principal que incorpora la muerte entre sus posibles valores.

Si los pacientes se adhirieran sin resquicios al consejo médico y si los investigadores siguieran fielmente el protocolo, no sería necesario distinguir entre eficacia y efectividad, ni definir subconjuntos de datos para "intentar salvar los muebles". Dos avisos: a nivel teórico, sólo la comparación de los grupos tal y como se asignaron está protegida por el azar; y a nivel práctico, los desvíos del protocolo comprometen aconsejar esa intervención en ese entorno.

3.5. Riesgo de sesgo... y su prevención

Repasemos, para terminar, las amenazas a las que se enfrentan los investigadores. Un buen clínico querrá dar su mejor intervención a los más graves. Quizás incluso cambiando de grupo, lo que se conoce como "cross-in". Así, los grupos perderían su comparabilidad. Si un investigador no cree que ambas intervenciones son éticamente asignables a los pacientes, debería renunciar al estudio. Asignar al azar y ocultar la intervención hasta finalizar el reclutamiento lo evitan. Este clínico no convencido de la ética del estudio podría querer dar intervenciones adicionales o de rescate a los controles. Lo que prevenimos enmascarando el tratamiento, quizás, con un placebo. Hemos defendido ya la importancia de un seguimiento completo. Por supuesto, el paciente o el clínico pueden decidir interrumpir la intervención, pero eso no implica excluirlo del seguimiento, ya que conviene conocer en cuántos ha pasado y cómo han evolucionado. Podemos prevenir ser más generosos al valorar la respuesta en uno de los grupos mediante el enmascaramiento o el uso de procesos de medida independientes del investigador. Finalmente conviene no caer en la tentación de reportar aquel resultado estadístico que más nos conviene. Lo garantizamos publicando el plan de análisis antes de tener acceso a los datos. Finalmente, recuerde que debe especificar a quién ha pretendido cegar, al paciente, a quien lo recluta, o lo trata, o lo sigue, o lo evalúa… Y con esto terminamos. Bueno, no: recuerde que no es lo mismo ser ciego que estar cegado.

4. <u>Protocolo</u>

Le doy la bienvenida al cuarto vídeo de Estadística en salud, donde esbozaremos 5 aspectos clave de SPIRIT, la guía para escribir protocolos.

4.1. Ciencia y Técnica:

Ciencia y Técnica van de la mano, pero no son lo mismo. Mientras la Ciencia quiere saber, la Técnica quiere hacer. En Ciencia, Fisher propuso el valor P como medida de la evidencia en contra de cierta hipótesis nula que si se rechaza, prevalece la alternativa que se quiere demostrar. Su reto es que no rechazar la nula no implica haber demostrado que sea cierta.

Para Popper, si hemos intentado refutar una hipótesis y no lo hemos conseguido, quizás no sea definitivamente cierta, pero puede ser aceptada tentativamente, al menos por el momento. Para poder hacer en la Técnica, Neyman y Pearson proponen limitar los riesgos de cometer acciones erróneas.

Usemos el tabaco como ejemplo. Los estudios de Doll y Hill de los años 50 generaron décadas de discusión sobre el nivel de evidencia científica alcanzado. Afortunadamente, Greenland resaltó que, con la evidencia disponible, la acción más *sensata* era actuar contra el tabaco. En un entorno de decisión, *sensata* significa, que las implicaciones de no luchar contra el tabaco si tuviera estos efectos son mucho peores que las de hacerlo si no los tuviera. Y aún más, considerando que es mucho más verosímil que sí que las tenga.

4.2. Tamaño de un ensayo decisorio

Estamos ya en condiciones de abordar el tema del tamaño muestral en un ensayo decisorio que actúa como pivote para una acción u otra. Los 2 errores que debemos controlar son: autorizar una intervención con cero eficacia; y no aprobar una con cierta eficacia delta. La figura muestra la distribución del test estadístico escogido en estos dos simples escenarios alternativos, que el efecto sea nulo, cero, o que sea exactamente delta. Así, el resultado del estudio puede tomar valores a lo largo del eje horizontal de abscisas. Como hay cierto solapamiento entre ambas distribuciones, la decisión no puede ser perfecta. La línea central marca el umbral que guiará la decisión: resultados a su derecha, actuaremos como si procedieran de la distribución con efecto delta y autorizaremos la intervención. A su izquierda, en cambio, no, porque actuaremos como si procedieran del escenario hipotético de efecto nulo. De esta forma, en la mayoría de las ocasiones, adoptaremos la decisión correcta. O si prefiere, controlará los riesgos de tomar decisiones erróneas.

Así, una proporción alfa de ensayos sobre intervenciones no eficaces terminará con autorización. Y similarmente, una proporción beta de los realizados sobre intervenciones con el efecto delta, sin autorización. Digamos que el complementario de beta es la potencia o probabilidad de autorizar una intervención de efecto delta. Aumentar el tamaño muestral 'n' llevará a curvas más puntiagudas, con menor solapamiento. Y también, en el caso de variables numéricas, mayor dispersión sigma, mayor solapamiento. Ésta es la horrible fórmula que no debe recordar. Pero si saber que para calcular la 'n' precisa 4 conceptos. De forma coherente, cuanto menor sea el efecto delta de interés, mayor será la 'n'. También cuanto menores sean los riesgos alfa y beta que está dispuesto a asumir. Y cuanto mayor sea la variabilidad sigma de la respuesta.

Ahora ya está en condiciones de interpretar el cálculo del tamaño para un estudio confirmatorio con respuesta numérica. Podrá identificar el efecto delta, la variabilidad sigma de la respuesta, y los riesgos alfa y beta que está dispuesto a asumir. Toda una lección de griego. Atento: es un detalle importante, las letras griegas representan constantes, valores conocidos de antemano. No olvide que está en un estudio confirmatorio, y que estudios previos le han aportado evidencia para delta y sigma. Bueno, también hay una letra latina, 'n', para representar el resultado del cálculo.

Por tanto, empleando el potencial del diseño de experimentos, puede trabajar sigma y delta para conseguir la potencia necesaria sin aumentar el número de voluntarios.

4.3. Variable respuesta

Lo que nos lleva a hablar sobre la respuesta. Distinguimos entre la variable principal, que servirá para tomar la decisión; y las variables secundarias, que utilizaremos para saber más, para aumentar el conocimiento científico.

El primer requisito de la respuesta principal es que sea pertinente, que refleje un objetivo de interés sanitario, como pueden ser la cantidad y la calidad de vida.

En ocasiones, recurrimos a respuestas subrogadas o intermedias para acortar o abaratar el estudio. Por ejemplo, para disminuir los eventos cardiovasculares, los primeros antihipertensivos tuvieron como variable subrogada precisamente a la presión arterial, que podía obtenerse mucho antes en el tiempo y requería un menor tamaño y seguimiento del estudio. ¿Y por qué algunos de ellos fueron luego desautorizados? Pues porque no cumplían el criterio de Prentice y tenían, además, otros efectos no deseados en la auténtica respuesta, los eventos cardiovasculares.

Además, la respuesta debe ser fiable, en el sentido de que sus variaciones dependan tan solo de las oscilaciones de su objeto de medida; y no del azar del momento o del evaluador. Así, la nota de un examen será fiable cuando dependa de los conocimientos del estudiante y no de su estado puntual de gracia o del humor del profesor. Una respuesta principal fiable conduce a un menor tamaño muestral.

4.4. Responsabilidades en un ensayo

Vistos estos aspectos técnicos, Vd. ya está en condiciones de usar SPIRIT. El objetivo final será mejorar la replicabilidad y el impacto de sus resultados. A corto plazo, conocer SPIRIT le ayudará a redactar el protocolo y a reducir el número de enmiendas. Menudo regalo.

Resaltemos, antes de terminar, dos aspectos de SPIRIT. En el documento explicativo encontrará ejemplos ilustrativos de cómo les gusta a las revistas que haga Vd. el diseño. Por ejemplo, cuáles son las responsabilidades y contribuciones concretas del investigador principal, del comité ejecutivo, del comité de vigilancia no enmascarado, del gestor de datos, de los investigadores principales de los centros, del promotor y del financiador.

4.5. Difusión

Y especificar las intenciones de comunicar los resultados a los voluntarios, los profesionales, o el público en general, puntualizando cualquier limitación a su difusión. Y aclarando si permitirá el acceso al protocolo completo, a la base anonimizada de datos de pacientes, y al código estadístico empleado. Las buenas revistas seleccionarán aquellos estudios que mejores estándares tengan. Y por supuesto, también los comités que deban autorizar el estudio.

Y con esto terminamos. Bueno, no: conviene recordar que mientras antes la Ciencia y la Técnica se conformaban con re-escribir las **leyes** del Universo, hoy tan sólo ajustan **modelos** que lo representan y reproducen con mayor o menor precisión. Pero más importante, permiten mejorar nuestras condiciones.

5. Revisiones: PRISMA

Bienvenid@ al quinto vídeo de Estadística en salud, donde esbozaremos 5 aspectos clave de PRISMA, la guía de revisiones.

5.1. La pregunta

Una revisión sistemática quiere recopilar toda la información de calidad sobre, por ejemplo, los efectos de una intervención. Además, si procede, el meta-análisis la resume en una estimación única del efecto. PRISMA propone la regla nemotécnica PICOS para recordar los 5 componentes que definen la pregunta de salud: P por Población objetivo o Participantes que se beneficiarían; I por la intervención en estudio; C por su comparador; O por el *Outcome*, la variable respuesta evaluada; y S por *Study*, el tipo de estudio, en general, ensayo aleatorizado.

5.2. Evidencia y Decisión

Por tanto, a diferencia del ensayo confirmatorio, el objetivo no es tomar una decisión tipo sí/no, sino agregar la evidencia científica. La fase III es como un control de calidad final. Pero una vez ya tenemos varios ensayos sobre la misma intervención, aparece otro objetivo: ¿qué sabemos sobre el efecto de la intervención? Por supuesto, dos objetivos tan distintos, no descansan en la misma metodología estadística.

El criterio de valor P por encima o debajo de 0.05 encaja bien para tomar una decisión. Pero el intervalo de confianza sobre una medida del efecto de la intervención cuadra mejor para resumir qué es lo que sabemos.

5.3. Efecto homogéneo

Retomemos el gráfico que mostraba el papel de la asignación al azar y el buen seguimiento para proporcionar la comparabilidad de los grupos. Añadamos ahora que estos voluntarios comparten unos criterios de elegibilidad que permiten asumir, de entrada, un efecto razonablemente homogéneo en todos ellos. Esta premisa de efecto constante se puede validar luego observando, por ejemplo, dispersiones similares en la respuesta numérica a los dos tratamientos, o efectos similares en subgrupos pronósticos. Que el efecto sea constante permite aprender en unos pacientes y aplicarlo en otros. Y legitima, así, la Medicina basada en la evidencia, que descansa en protocolos uniformes para las recomendaciones clínicas. Ahora bien, los efectos observados en pacientes de estas características atendidos en el entorno del ensayo ¿aplican también a otros en otras condiciones? Un estudio aislado, debe descansar en criterios sustantivos para discutir su transportabilidad. Pero una revisión sistemática puede observar y cuantificar la heterogeneidad del efecto.

5.4. Oscilación de la estimación del efecto

Este ejemplo del documento explicativo de PRISMA usa como medida del efecto el cociente de probabilidades o riesgo relativo. Su estimación conjunta en todos los estudios vale 2.3, con un intervalo de incertidumbre estrecho, que no incluye el valor 1 de igual probabilidad bajo ambas intervenciones. Las estimaciones por intervalo de los estudios individuales oscilan mucho, como cabe esperar por la buena o mala suerte de la asignación al azar, especialmente en estas pequeñas muestras con pocos eventos.

Sin embargo, todas las estimaciones concuerdan con la global. De hecho, la estimación de la heterogeneidad del efecto con el I^2 vale 0%, indicando que, en estos estudios, el efecto no oscila. Toda la variación en las estimaciones de los estudios es explicable por la imprecisión debida a la asignación al azar. Vamos a estudiar este I^2.

5.5. Cuantificar la heterogeneidad

I^2 es el cociente entre la heterogeneidad del efecto y la suma de ambas variabilidades. Oscilará entre 0 y 1. O, si prefiere, entre 0 y 100%. Mayor I^2, mayor estimación puntual de la heterogeneidad. Veamos unos ejemplos.

Aquí tenemos 5 meta-análisis seleccionados por Higgins y Thompson para mostrar cómo varía I^2. ¿Qué le parece? ¿Tienen los 5 análisis la misma heterogeneidad? La del primero es 0. Y va subiendo progresivamente hasta alcanzar el 98%, cifra absurda, ¿cómo pueden divergir tanto unos estudios con tanta precisión? Habrá que mirar con atención el segundo estudio del último meta-análisis: ¿cómo es posible que ofrezca resultados tan contradictorios? Estos gráficos ayudan a distinguir entre patatas y manzanas.

El gráfico de l'Abbé muestra simultáneamente la proporción de eventos adversos en el grupo tratado, en vertical; y en el control, en horizontal. La línea discontinua marca el empate o identidad entre las proporciones para ambos tratamientos. Algunos estudios tienen mayor proporción de eventos, quizás por mayor duración, pacientes más graves, o criterios más laxos. En otros, menos. Insistimos en que cada estudio tiene su propio protocolo para garantizar su comparabilidad interna; pero que estos protocolos difieren entre estudios. En la mayor parte de ensayos, ha habido menos eventos en el grupo tratado. Bien. El área de los círculos es proporcional a la precisión del ensayo. Así, los puntos pequeños se alejan más, resultado de su mayor oscilación aleatoria. Visualmente, podríamos resumir el efecto por una línea paralela a la recta identidad. Que sea paralela indica que la magnitud del efecto es siempre la misma, tanto si el estudio presentó muchos o pocos eventos. Este gráfico muestra que todo va bien: (1) como resultado de protocolos diferentes, algunos estudios muestran más eventos que otros, ningún problema, ya que el efecto parece ser siempre el mismo; y (2) las distancias al efecto promedio se explican por la precisión de cada estudio. En resumen, apunta a que todo son manzanas y no hay ninguna patata.

En cambio, este gráfico resalta el estudio 14. Aquí tenemos un meta-análisis con 33 estudios. Y lo repetimos 33 veces, eliminando cada vez un ensayo. Todos los meta-análisis que incluyen el estudio 14 excluyen el 1 de no efecto. En cambio, si se excluye, dejan de serlo. En resumen, este metaanálisis no es robusto, ya que "descansa" en un único estudio: los resultados son sensibles a la inclusión del estudio 14. Aquí tenemos una patata. Quizás podrida. Estudiemos ahora posibles fuentes de heterogeneidad.

El método estadístico podría introducirla. ¿Recuerda aquel ejemplo de proporciones de eventos adversos con efectos idénticos si los dividíamos y muy diferentes si los restábamos? Conviene controlar esta fuente de heterogeneidad. La segunda procede de mala metodología. Una de las lecciones de las revisiones sistemáticas es observar, por ejemplo, que los estudios que, o no enmascaran, o no consiguen cegar, suelen dar

estimaciones optimistas del efecto. Podrían entrar en este grupo tanto el ejemplo de la I^2 del 98%, como el estudio 14 de gráfico de sensibilidad.

Revisiones sistemáticas Cochrane han encontrado estimaciones más optimistas del efecto en aquellos estudios que: (1) no documentan el sistema para generar la secuencia aleatoria; (2) no ocultan el tratamiento asignado hasta después de completar el reparto en los grupos; (3) no enmascaran; (4) pierden casos; (5) cambian análisis o variables tras ver los resultados; y (6) el cajón de sastre que nunca conviene olvidar. José Antonio propone SABIOS como regla nemotécnica.

Además del sesgo dentro de cada estudio previo, existe la posibilidad de sesgo en el meta-análisis, lo que PRISMA llama sesgo entre o a lo largo de los estudios, debidos principalmente a la información ausente, o bien estudios completos que no se han publicado ni pueden localizarse, o bien estudios que han perdido o han ocultado parte de su información.

Este gráfico debe tener forma de embudo invertido si todo va bien. En horizontal presenta la magnitud del efecto observado. Y en vertical, su precisión, figurando en la base los menos precisos, por lo general más abundantes. Parte del principio de que los estudios grandes, en la parte superior, es más fácil localizarlos; pero que los pequeños podrían publicarse sólo si apuntan en la dirección deseada. Por eso, na vez más, este meta-análisis, tiene la forma deseada, tranquilizadora. En cambio, PRISMA ofrece este ejemplo que alerta sobre un claro sesgo de publicación, ya que falta una cola del embudo. Y así terminamos con la heterogeneidad metodológica.

En las fuentes clínicas de heterogeneidad, conviene distinguir entre las aleatorias o anónimas y aquellas que, por tener características con nombre, identificables, permiten usar en el futuro el efecto estimado en ese subgrupo. ¿Pero qué hacer en el primer caso, por ejemplo, si existe variación del efecto entre intervencionistas, anónimos o aleatorios en el sentido de que no son los que usaremos en otros entornos?

La solución estadística consiste en cuantificar la variación entre ellos y considerarla en la interpretación. En este ejemplo, el Intervalo considera ambas incertidumbres, además de la habitual aleatoria, ha introducido la derivada de no saber el efecto concreto de *mi* intervencionista. Así, la estimación puntual es que mis síntomas de demencia mejorarán un tercio, pero el $IC_{95\%}$ que considera σ^2 y τ^2, dice que "*mi*" efecto puede ser cualquiera entre 1/5 y 1/2.

5.6. Recomendaciones

En resumen, asegure la homogeneidad metodológica de los estudios en el protocolo de la revisión siguiendo los consejos de PRISMA, Cochrane y GRADE. Y para la posible heterogeneidad anónima, use el modelo de efectos aleatorios.

Antes de abandonar la heterogeneidad, conviene decir que en el caso de un solo ensayo clínico, la interpretación de diferencias del efecto entre subgrupos es tan delicada, que incluso las mejores revistas discrepan en sus recomendaciones. Aunque no en lo fundamental: la significación de un subgrupo no salva un ensayo sin resultados globales concluyentes.

Y con esto terminamos los vídeos dedicados al ensayo clínico. Bueno, no: recuerde que no todos los objetivos de salud requieren un diseño experimental.

6. Etiología: STROBE (1)

Bienvenid@ al sexto vídeo, primero dedicado a STROBE, la guía para estudios epidemiológicos y en él veremos 3 pinceladas cruciales.

6.1. Asociación frente a causalidad

Conviene resaltar que relación y causalidad son conceptos distintos. En un primer ejemplo, usamos el interruptor para apagar y encender la luz. Mediante la primera variable puedo *cambiar* la segunda. Además, si el sistema funciona, sin margen para el error. En un segundo ejemplo, interviniendo en las horas de estudio puedo influir en la nota, aunque ahora existe cierta variabilidad restante en la nota. En un tercer ejemplo, en un centro de enseñanza primaria, un maestro puede utilizar la altura de un niño para aproximarse a su nivel de desarrollo mental y al curso al que va. Pero alargando artificialmente las piernas no conseguirá pasarlo de curso.

En este último ejemplo, un observador hábil puede usar el consumo de internet de un país como *chivato* de su mortalidad por cáncer. No olvide que ambas variables se dan en países desarrollados. Pero prohibir internet no tendría ningún efecto en el cáncer. En resumen, si hay asociación, siempre puede usar una variable para anticipar el valor de la otra. Y con mayor precisión cuanto más intensa sea su relación. Pero para poder intervenir sobre la primera para cambiar el valor de la segunda necesita que la relación sea causal.

Esta magnífica figura de Hernán aclara la diferencia. En azul oscuro tenemos los tratados; y en claro, los no tratados. Hay relación si aquellos tratados tienen una respuesta diferente a los no tratados. Y causalidad cuando si tratáramos a toda la población obtendríamos resultados distintos que si no lo hiciéramos. Note que ambos resultados no se pueden observar a la vez y en las mismas condiciones, lo que en lógica se llama problema fundamental de la inferencia causal. Ya vimos que el ensayo clínico recurría a muestras aleatorias. Aquí, lo importante es diferenciar entre asociación, que puede servir para predecir el futuro y causalidad que, además, permite cambiarlo.

6.2. Odds ratio

Supongamos que disponemos de voluntarios tanto en atención primaria como en el hospital de referencia. En ambos casos, su evolución puede ser o positiva o negativa. Por la razón que sea, quizás pacientes más jóvenes, o más incipientes, en primaria la evolución suele ir bien con más frecuencia que mal: con un cociente de 2 a 1, al que llamaríamos por su término inglés, odd, que permite conocer el premio o momio de una apuesta.

Supongamos que 120 ludópatas apuestan por la buena y 60 por la mala evolución. En caso de ganar, los 120 primeros se repartirían 60 euros, por lo que cobrarían medio euro: su momio sería de un medio. En cambio, los segundos tendrían un momio de 2. Por otro lado, en el hospital, con más comorbilidades, estas apuestas estarían 1 a 2, un medio. Así, dividiendo las odds de la primera fila, 2, por las de la segunda, un medio, obtenemos un cociente de apuestas, un odds ratio de 4. Bueno, recuerde que si da 2

pasteles a media persona, a una persona entera le tocarían 4. Como el intervalo excluye el valor 1 de no relación, el centro o condición Z es un predictor de la respuesta Y.

Así, siempre estaremos tristes por saber que alguien está enfermo, pero apostaremos por una mala evolución 4 veces más en el hospital. Por supuesto, al no ser una relación causal no diremos: "evita ir al hospital".

6.3. Efectos confundidos

Eso sí, tenemos un marrón, ya que los pacientes del hospital no son comparables a los pacientes de atención primaria. Por eso, para probar una intervención X, la balanceamos o equilibramos: asignamos los voluntarios con la misma razón, pongamos 1 a 1, tanto en primaria como en el hospital, resultando en un odds ratio de 1, no relación, al que no hacemos intervalo porque no tenemos incertidumbre sobre su valor, ya que hemos decidido asignar 1 a 1 en ambos centros. Veamos la ventaja estudiando cómo se distribuyen en su evolución Y los 90 casos asignados en cada centro a cada intervención. En primaria teníamos 120 y 60 que, en este ejemplo, se distribuyen por igual en ambas intervenciones, 60 a 30, resultando en una odds de 2, que es igual para tratados y controles, lo que implica un OR igual a 1 de no efecto de X en Y. Formalmente: la estimación puntual del efecto es un OR de 1, no efecto. Y como mucho, la intervención o bien multiplicaría por 2, o bien dividiría por 2 la apuesta por una buena evolución. Si nos movemos al hospital, veremos que ahora la odd general de un medio se repite en ambas intervenciones, también con un OR de 1, la no relación; y la misma incertidumbre. Si combinamos las tablas de primaria y hospital, sumando los casos de cada celda, 60 y 30 suman 90 en las 4 celdas, llevando a esta nueva tabla global que, sin ajustar por centro, tampoco muestra efecto. Ésta es la gran ventaja del diseño balanceado: se llega a la misma estimación ajustando que sin ajustar. Recuerde la broma que dice que un ensayo clínico bien diseñado, ejecutado y analizado no necesita discusión, ya que los resultados lo dicen todo. Note también que este análisis ajustado es ineficiente, ya que subdivide los casos en 2 subgrupos, resultando en mayor incertidumbre en sus intervalos.

Veamos ahora qué podría suceder en el entorno observacional. Partimos, como antes, de pacientes diferentes en primaria y hospital. Pero ahora los clínicos y los pacientes pueden elegir. Pongamos que los 180 voluntarios de primaria tienden a recibir la intervención, con una odd de 5 a 1; mientras que los 180 del hospital es justo al revés, 1 a 5. Por las razones que sea, quizás distinta actitud o de los clínicos o de los pacientes. No importan las razones, el caso es que ahora hay un OR de 25, indicando fuerte relación entre la condición previa Z y la intervención X en esta muestra. No nos importa la población y por tanto no hacemos intervalos: en la muestra que tenemos, Z y X van juntas. Una vez más miremos qué sucede en cada centro por separado: en primaria, los 120 y 60 casos se reparten también 2 a 1: 100 frente a 50 y 20 frente a 10, por lo que el efecto estimado de X en Y es otra vez nulo. Y lo mismo en el hospital. Veamos ahora qué sucede cuando combinamos ambos centros: 100 más 10 dan 110; y 20 más 50, 70. Ahora parece que la primera fila tiene tendencia a ir bien, 110 a 70, aproximadamente 3 a 2; pero la segunda fila a mal, 70 a 110, es decir, 2 a 3. Con un OR de 2 y medio y un intervalo que excluye el 1: con 95% de confianza los tratados tienen una odd que es 1.6 y 3.8 veces mayor que los controles. Difícil situación, ya que

tenemos 2 estimaciones diferentes del efecto, resultado de que X y Z vayan juntas en esta muestra. Observe que la primera fila incluye los tratados, de acuerdo, pero también, 150 de los 180 casos provenían de primaria, con mejor pronóstico. Y al revés en la segunda fila, 150 de 180 provenían del hospital. Es decir, entre ambas filas hay 2 diferencias simultáneas: el tratamiento X y el centro Z. La interpretación de la discusión es ahora imprescindible: si, por cuestiones teóricas, decidiéramos que el centro es una variable por la que no hubiera que ajustar, llegaríamos a un efecto significativo de 2.5. En cambio, si acertadamente argumentamos que los centros son condiciones previas por la que debemos ajustar, interpretamos que la intervención X no tiene efecto en la respuesta Y.

En resumen, para que cambien, en cualquier sentido, los efectos estimados al ajustar por la condición Z, ésta debe estar relacionada tanto con la intervención X como con la respuesta Y. Si ignora Z, aparecerá una relación entre X e Y, que siempre sería más fácil controlar con un buen diseño que con un buen análisis.

Como muy bien recuerda STROBE, la guía para estudios observacionales apoyada por las mejores revistas. Recordemos, para terminar este punto sobre efectos confundidos que la relación entre X e Y es real, existe: conocido el valor de X podemos anticipar, en parte, el de Y. Pero no es causal. Hay asociación, pero no hay causalidad ni opciones para intervenir. Como la asociación es real, STROBE evita hablar de sesgo de confusión: la relación existe, pero no es causal.

Ante un gran incendio, el alcalde puede predecir el nivel de daños a partir del número de bomberos enviados, pero haría mal en impedir que salieran los bomberos para evitar los daños.

7. Etiología: STROBE (2)

Bienvenid@ al séptimo vídeo, segundo dedicado a STROBE, y en él veremos especialmente el sesgo de selección y lo enfrentaremos a la confusión de efectos.

Pero antes de empezar, una pregunta sencilla. ¿Qué le parece? ¿Es obligatorio o es optativo usar estos términos tan populares? ¿O incluso es mejor evitarlos? Desgraciadamente, prospectivo y retrospectivo tienen tantas acepciones técnicas que son ambiguos, por lo que la guía apoyada por los editores de las mejores revistas no los aconseja. Veamos algunos significados

7.1. ¿Causas o Efectos?

Ya dijimos en el primer video que hay preguntas sobre sobre y efectos, que miran hacia adelante; y sobre causas, que miran hacia atrás. No es lo mismo preguntarse "¿se me irá el dolor si me tomo un analgésico?" que "¿el dolor se fue porque tomé un analgésico?".

Un magnífico ejemplo es el asma preolímpica de Barcelona. Algunos ciudadanos las boicoteaban: padecían asma por causa desconocida en barrios y días concretos. El doctor Joan Clos, responsable sanitario de Barcelona, pidió a los epidemiólogos Jordi Sunyer y Josep María Antó afrontar el reto, quienes pusieron banderitas en el mapa de Barcelona allí donde se daban los casos. Y vieron que el día que había descarga de soja en el puerto y el viento soplaba hacia el Putxet, ahí aparecía el asma. Y de forma

consistente en otros días y barrios. Contentos, informaron al gestor sanitario. Pero éste, en lugar de compartir su ilusión les dijo: "genial, conocemos de qué fallecen. Ahora además sabrán el motivo para no darnos las olimpiadas". ¿Sobre qué podían intervenir? Ni sobre el viento ni sobre la soja. Así, su siguiente pregunta fue sobre el efecto de reparar los silos y poner una lona protectora en la descarga. La respuesta facilitó los juegos de Barcelona de 1992.

7.2. Tipos de diseños según variables "fijas":

Es importante tener claro qué números hemos decidido nosotros y cuáles son el resultado observado. Por diseño, podemos decidir estudiar a mil casos y recoger la información de una variable inicial y una respuesta. O bien, 500 pacientes con una intervención A, 500 con la B; y observar la variable respuesta en ambos grupos. Finalmente, podemos decidir estudiar a 500 voluntarios con la respuesta en estudio, 500 sin ella y recuperar su exposición previa a cierto fenómeno de interés en ambos grupos. Así, los estudios de casos y controles condicionan o fijan la respuesta; los ensayos clínicos, la intervención; y los estudios transversales y los longitudinales de cohortes, sólo fijan el tamaño global y ambas variables son un resultado del estudio.

7.3. Experimentar frente a Observar

Es quizás el momento para distinguir entre experimentar y observar. En la primera, asignamos a las unidades el valor de la causa en estudio; en la segunda, "observamos", miramos el proceso. En ambos, seguimos a los voluntarios a lo largo del tiempo hasta observar la respuesta de interés. Pero en uno yo *hago*, *yo asigno* el valor, mientras que en el otro, *veo* como llegan ya con su valor. Si *hago*, puedo recoger la respuesta en casos que sólo se diferencien por la variable en estudio. Si *veo*, pueden diferir entre sí en más de una característica, ofreciendo más de una interpretación. Por ejemplo, si aprueban más los alumnos de mi clase podría ser porque me explico mejor, pero también porque pongo exámenes más fáciles o porque sus padres contribuyen más a su educación. Además, el experimento facilita observar si, al aconsejar cierta intervención, pacientes y clínicos actúan según lo previsto. La Inteligencia Artificial ha añadido el operador "DO", hacer, frente al "SEE", ver, para facilitar la distinta interpretación de la relación en ambas situaciones.

7.4. Sesgo de selección

Cuentan que, como muestra del poderío de los dioses, las sacerdotisas griegas enseñaban a los visitantes los magníficos presentes de marineros que, tras rezar a los dioses, se habían salvado de terribles tormentas. Y decían: "¿Qué mayor evidencia quieren del poderío de los dioses?" A lo que un visitante escéptico respondió: "Bueno, me falta saber cuántos marineros rezaron y no se salvaron". Por eso el sesgo de selección es también conocido como sesgo del superviviente. Es tan frecuente que recibe muchos nombres y debemos siempre vigilar su presencia. Si este vídeo sólo lo termina de ver uno de los muchos que lo empezaron, aunque luego veamos un "me gusta", haremos mal en quedar contentos.

Sus amigas preguntaron a Jordan Ellemberg: "oye, es que los hombres con que quedo o son guapos o son simpáticos, pero ambas cosas, no". Y les contestó: "sesgo de

selección, ya que los feos antipáticos no han sobrevivido hasta la primera cita". Supongamos que los candidatos estén divididos por igual en las 4 celdas: el OR entre belleza y simpatía valdría 1. Pero si pocos feos antipáticos llegan a la cita, la primera fila tendría una odd de 1, pero la segunda de 5 a 1, resultando un OR de un quinto, que cuantifica la relación negativa comentada: los feos que llegan a la cita tienden a ser más simpáticos.

Veamos ahora un ejemplo con una respuesta numérica. Suponga que una universidad promueve a sus profesores si destacan en 2 dimensiones: investigación y docencia. Suponga además, por simplicidad, que su distribución es la de la figura, con casos por todos lados; y que ambos aspectos son independientes: saber si alguien es bueno en investigación no informa sobre su capacidad docente. Estos casos tendrían un 10 en ambos ejes. Estos, un 7.5. Y estos, un 5. Así, esta línea distinguiría entre los muy buenos, arriba a la derecha, con una suma de ambas puntuaciones de 20; y los regulares, abajo a la izquierda, con una suma de tan solo 10. Esta universidad podría usar estos puntos de corte para distinguir entre catedráticos, profesores, lectores, y…

Aquí podemos ver los 4 grupos resultantes. Y que en todos ellos, hay una relación negativa, más fuerte en los grupos centrales: o se es bueno en docencia o en investigación, pero no en ambas cosas… Una vez más, esta relación negativa es el resultado de un proceso de selección por una variable posterior, el nivel de progreso académico alcanzado.

Veamos un posible ejemplo clínico. Supongamos que tanto los lípidos altos L como cierto gen G provocan eventos cardiovasculares E. Supongamos también que este gen no tiene ningún efecto sobre los lípidos, como muestra esta tabla. Ahora bien, los casos que presenten eventos E irán al hospital, subtabla de la izquierda, donde vemos una odd aproximada de 2 a 1 en la primera fila y de 4 a 1 en la segunda, resultando en un OR de 0.4, marcando una relación negativa: el gen previene de los lípidos altos.Y lo mismo en aquellos que no van al hospital. Una vez más, seleccionar por una variable posterior hace aparecer una relación que no es real: sesgo de selección.

7.5. Gráficos dirigidos no cíclicos

Los diagramas con dirección y sin ciclos, DAG por sus siglas inglesas, son otra gran herramienta aportada por la Inteligencia Artificial, Judea Pearl entre otros, que ayudan a entender y sedimentar todos estos términos. En un DAG las flechas están dirigidas, es decir, tienen inicio en una variable y terminan en otra. Y no forman ciclos. En este ejemplo, la presión inicial determina la final e influye en el tratamiento, que a su vez, afecta a la presión final. Pero no hay bucles, ya que la presión final no afecta ni a la inicial ni al tratamiento.

Los DAGs representan relaciones causales. Fumar es causa común de llevar tabaco y tener cáncer. Pero llevar tabaco encima no tiene efecto causal en el cáncer. En cambio, por el reto de los efectos confundidos, observaríamos relación entre llevar tabaco y tener cáncer: aquellos que lleven tabaco encima tendrán mayores odds de tener un cáncer.

Ahora bien, condicionar o ajustar por el hecho de fumar, hace que desaparezca la relación entre llevar tabaco y cáncer. Aquí llamamos a este condicionar "bloquear" y lo

representamos por un cuadrado alrededor de la variable fumar para indicar que se ha fijado.

Este otro ejemplo representa que 2 variables tienen efecto en la misma respuesta. Por ejemplo, tanto un Gen como un Entorno pueden tener efecto en un cáncer, pero ser entre ellos independientes.

Ahora bien, si bloqueáramos el efecto común, observaríamos una falsa relación originada por el sesgo de selección.

Y con esto terminamos los vídeos dedicados a STROBE. Bueno, no, volvamos a la confusión: una variable sin relación causal puede ser útil para predecir el futuro.

8. Diagnóstico y pronóstico: STARD

Le doy la bienvenida al octavo vídeo de Estadística en Salud dedicado a STARD. Ya dijimos al empezar que el diagnóstico pregunta por el presente, pero el pronóstico, por el futuro. Aunque STARD habla de diagnóstico y TRIPOD de pronóstico, ambas aplican a los dos objetivos clínicos. Sus gran diferencia es que STARD aborda cómo cuantificar las propiedades de un indicador ya definido, pero TRIPOD también de su construcción.

8.1. Variables

En el planteamiento más simple, el diagnóstico enfrenta 2 variables, cada una con dos posibles valores. Por un lado, asumiremos que conocemos sin error la situación real del paciente, bien enfermo, bien sano. Por otro lado, tenemos una prueba diagnóstica, o test, o signo, o síntoma, al que llamaremos indicador y que puede tomar un valor o positivo o negativo. Por supuesto, pueden aparecer errores.

Dentro del círculo azul representamos a los enfermos, y fuera, a los sanos. Además, dentro del círculo rosa a los que dan positivo; y fuera, a los negativos. Así, estas 2 dicotomías generan 4 grupos: (1) el de los enfermos que dan positivo, bien; (2) el de los sanos que dan negativo; ¡bien!; (3) el de los enfermos que dan negativo, malo; y (4) el de los sanos que dan positivo, malo también.

Y aquí tenemos una situación mejor, con mayor coincidencia, menos errores y, por tanto, mejores propiedades diagnósticas.

8.2. Diseños

Veamos ahora de qué formas podríamos recoger los datos. En un primer diseño, dispondríamos de 2 poblaciones objetivo, la de enfermos y la de sanos, de las que obtendríamos, por azar, dos muestras de voluntarios y en cada una de ellas determinaríamos el valor del indicador, positivo o negativo. En un segundo tipo de diseño, obtendríamos una sola muestra de la población, y dentro de ella, recogeríamos 2 variables: el estado real del voluntario y el valor del indicador. En resumen, podríamos disponer de 2 muestras y 1 variable, o bien de 1 muestra y 2 variables. En forma de tabla, el primer diseño podría tener estos resultados si hubiéramos decidido estudiar 100 voluntarios de cada población. En el segundo diseño, en cambio, habríamos fijado la 'n' de la única población en, por ejemplo, 100 casos. Recuerde que estos valores de 100, no son variables, no son resultados del estudio que aporten información sobre la población, ya que están decididos de antemano. Veremos que el

segundo diseño, con menos valores prefijados, contestará más preguntas que el primero. Ahora hay que empezar con un concepto teórico muy útil, la probabilidad condicionada, que distingue entre lo que sabemos y lo que nos preguntamos. Imagine esta tabla sencilla de 1 sola muestra con 2 dicotomías: el género en filas y la longitud del pelo en columnas. No es lo mismo la probabilidad de que una chica lleve el pelo largo, que la de que alguien con el pelo largo resulte ser una chica. En el primer caso, sabemos que es chica; y nos preguntamos qué proporción lleva el pelo largo. Como de 60 chicas, 30 lo llevan largo, es un 50%. En el segundo caso, sabiendo que lleva el pelo largo, ¿qué probabilidad hay de que sea chica? Como de los 40 que llevan el pelo largo, 30 son chicas, es un 75%. Ha cambiado el denominador. Por eso hay que estar siempre tan atento al denominador. Observe la simbología de P de Largo dado Chica: lo que va detrás de la barra es la condición que sabemos y lo que va delante, la pregunta que hacemos. Ahora ya estamos en condiciones de estudiar las propiedades de un indicador diagnóstico.

8.3. Sensibilidad y Especificidad

Miremos primero la tabla del diseño de 1 muestra y 2 variables. Sensibilidad mira, en la fila de enfermos, a los que dan positivo: P de más dado E, 9 sobre 10, un 90%. Y especificidad, en la de sanos, a los negativos, P de menos dado S: un 70%. Sensibilidad y Especificidad son probabilidades naturales, ya que, dada la condición preguntan por la consecuencia: los enfermos deben dar positivo y los sanos, negativo. Ambas son porcentajes sobre el total de la fila. Veamos ahora los porcentajes sobre el total de columna.

8.4. Valores Predictivos (VP)

Los Valores Predictivos indican la credibilidad de un resultado, sea positivo o negativo. UVE PE MAS es, del total de positivos, la proporción de enfermos, un 25% . Y UVE PE MENOS, del total de negativos, los sanos, un 98%. Estos datos imaginarios indican que tenemos gran confianza de que un negativo está sano, pero poca de que un positivo esté enfermo. ¿Qué ha pasado? Pues que afortunadamente sólo un 10% de la población atendida está enferma, por lo que la segunda fila pesa mucho más que la primera. Veamos también que el indicador ha aportado información, ya que hemos pasado de una expectativa previa de enfermo del 10% al 25% posterior tras un resultado positivo. Y del 90 al 98% para uno negativo.

Repitamos estos 4 cálculos si tuviéramos un diseño con 2 muestras. Supongamos que se trata del mismo indicador con idénticas sensibilidad y especificidad. Y que las 2 muestras con 100 y 100 casos decididos por nosotros, reproducen exactamente. UVE PE MAS ofrece un resultado diferente. Y lo mismo UVE PE MENOS. Estos 2 valores sólo aplicarían a una situación en la que las filas pesaran lo mismo, una prevalencia del 50%. El reto es, precisamente, que este diseño no aporta información sobre la prevalencia, que dependerá de su entorno concreto.

¿Cómo incorporar esta información adicional? La prevalencia informa de la expectativa previa de la enfermedad E. Si le añadimos la información de la prueba, obtenemos la expectativa posterior. Con las definiciones anteriores más el resultado de la prueba, la fórmula de Bayes proporciona los valores predictivos. Recurriremos a la informática

para calcularla. Vaya a esta dirección para obtener la capacidad predictiva de un resultado de forma cómoda.

Vea a su izquierda los botones que Vd. puede regular: prevalencia, sensibilidad y especificidad. En la derecha tiene un cuadrante que representa, en las columnas, de izquierda a derecha, a enfermos y sanos; y en las filas, a positivos encima y a negativos debajo. Así, si Vd. aumenta la prevalencia, verá cómo crece la columna de enfermos; si aumenta la sensibilidad, cómo crece la primera fila de la primera columna; y si aumenta la especificidad, cómo crece la segunda fila de la segunda columna.

Por ejemplo, supongamos que las mamografías tienen sensibilidades y especificidades casi óptimas del 99%. Pero, afortunadamente, en la población estudiada, la proporción de casos está en menos del 1%. Así, el valor predictivo de un positivo es sólo 0.333: entre los positivos, sólo 1 de cada 3 tendrá la condición en estudio. Así, la información de un resultado positivo en la prueba ha aumentado la expectativa de enfermos desde un 1% a un 33%. Es un paso en la dirección correcta, pero no definitivo.

En cambio, si mira el VP —, verá que es muy bueno, prácticamente 1: el resultado negativo de la prueba ha aumentado las expectativas de sano desde un 99% a casi un 100%.

8.5. Riesgos de sesgo

Veamos finalmente, algunos riesgos de sesgo. Recuperemos el gráfico inicial sobre los 2 tipos de estudios para situar a qué nivel podrían aparecer.

Muchos estudios no disponen de un mecanismo de azar para extraer la muestra, lo que resulta en un sesgo impredecible, ya que los cálculos estadísticos habituales, como el error estándar o el intervalo de confianza, sólo consideran la oscilación aleatoria. Si las definiciones de Enfermo y Sano son tan extremas que facilitan su discriminación, pero se alejan del problema real, caeremos en el sesgo del reto reducido, sin utilidad práctica posterior. Si los valores de una de las 2 variables, el estado real o el indicador, condicionan la inclusión en el estudio, tendremos un sesgo de selección, usualmente llamado de verificación si figura de forma explícita en el protocolo, o del espectro, si se aplica después. Como el diagnóstico es actual, no futuro, ambas variables se recogen en un mismo momento del tiempo, y el seguimiento no es necesario. Pero sí que requiere recoger datos de todos los pacientes, aunque tengan valores no deseados de las pruebas, o bien porque insuficiente material los hace no interpretables; o porque una condición concomitante invalida los resultados y los hace no determinables; o porque son intermedios, sin clara positividad o negatividad; o porque están fuera de los límites de plausibilidad. En resumen, cualquier tipo de dato ausente introduce riesgo de sesgo por desgaste de la muestra. Finalmente, la recogida de cada variable debe ser idéntica siempre, quizás enmascarada, para evitar el sesgo de información si los resultados de una condicionan los de la otra.

Y con esto terminamos el vídeo dedicado a STARD. Bueno, no: recuerde que no todos los objetivos de salud requieren un diseño experimental.

9. Diagnóstico y pronóstico TRIPOD (Vídeo 9)

Le doy la bienvenida al noveno y último vídeo de Estadística en Salud dedicado a TRIPOD, la guía para construir y validar escalas predictivas combinando la información de varias variables.

9.1. Modelo predictivo

El modelo predictivo usual es una fórmula, quizás larga, que expresa la probabilidad de un evento futuro en función de una serie de variables. A partir de unos datos, desarrollamos el modelo predictivo que, aplicado a las condiciones individuales, proporciona su riesgo predicho que, junto a las preferencias del paciente y los recursos disponibles, lleva a la decisión clínica.

9.2. HR: Hazard ratio o razón de tasas

Puede ser una regresión logística basada en odds ratios o una de supervivencia basada en Hazard ratios, la última medida de asociación de que veremos. Si el seguimiento de los casos es variable, a partir de cierto momento sólo sabemos que el paciente llegó vivo, pero no cuanto más resistió. Tenemos, por tanto información parcial, censurada.

Una tasa mide la velocidad de aparición de eventos. En este ejemplo, la velocidad en el grupo tratado era 0.86, un 14% inferior. Pero si yo pregunto, "Dr. ¿Cuánto tiempo más viviré?", podemos aproximarlo al revés: si no sigo el tratamiento, viviré un 14% menos.

También puede interpretar el Hazard Ratio como un riesgo relativo, ya que se cumple que los 3 cocientes, de momios, de tasas y de proporciones son similares. Tanto más cuanto más corto es el seguimiento, menos frecuente el evento y menor el cociente. Y la divergencia siempre es en este sentido.

Volvamos al modelo predictivo basado en odds ratios. El ejemplo muestra, por ejemplo, que síntomas previos de asma tiene un OR igual a 2, indicando que la odd de sensibilizarse es 2 veces mayor en estos casos. Más abajo, indica que un caso sensibilizado será mayor que uno sin sensibilización en el 75% de las ocasiones. ¿Qué importa más: el 2 que cuantifica el posible efecto de una intervención o el 0.75 que resume la capacidad del modelo para avanzar el futuro? Para contestar a esta pregunta, recuerde que ahora queremos predecir el futuro, no cambiarlo.

Así, el nuevo objetivo es clasificar bien a los casos en grupos con diferente pronóstico.

9.3. Capacidad Predictiva

A partir de unos datos, desarrollamos el modelo predictivo que, aplicado a las condiciones individuales, proporciona su riesgo predicho que, junto a las preferencias del paciente y los recursos disponibles, lleva a la decisión clínica. Ahora bien, ¿hasta qué punto acierta la predicción? Esta es la pregunta más importante.

TRIPOD proporciona esta informativa figura. Empecemos por su base. La línea representa la probabilidad predicha por el modelo de padecer coronariopatía, 0 a la izquierda, y 1 a la derecha. Además, muestra la frecuencia de casos que realmente la tuvieron y no la tuvieron. Por ejemplo, parece que para una probabilidad predicha de 0.2, realmente hubo aproximadamente un 20% de casos que sí la tuvieron. A nivel de

grupo, la predicción es correcta. Pero a nivel de casos es más incómoda: aunque lo más probable era que no la tuvieran, y así fue para el 80% de ellos, en un 20%, aparecieron. Es la paradoja estadística: regularidad del grupo frente a variabilidad de individuo. Entonces, ¿qué hacemos en un caso concreto? Pues seguir la sabiduría popular: "deseando lo mejor, pero preparados para lo peor". Aunque, como ya hemos dicho antes, a partir de qué probabilidad de lluvia nos encerramos en casa, depende de las preferencias personales, esa dimensión que los psicólogos llaman de amor o aversión al riesgo.

9.4. Sobreajuste

Un riesgo del modelado es adaptarse excesivamente a los datos actuales. Esta figura de Thomas Gerds muestra una línea verde que hace una clasificación perfecta, aunque la línea negra posiblemente tenga más posibilidades de replicar en el futuro su capacidad predictiva. El riesgo de sobreajuste es mayor cuanto menor es la información disponible y cuanto mayor es la selección automática de variables o de sus umbrales. Que conviene complementar por una decisión razonada y documentada. Este sobreajuste exige una re-estimación de la capacidad predictiva, sea en nuevos datos, sea con más sofisticados procesos estadísticos como el remuestreo o la validación cruzada. Por eso, TRIPOD cubre varios tipos de estudios, según obtengan, valoren o mejoren el modelo.

9.5. Aprender, Valorar y Actualizar

La muestra de aprendizaje, que aporta el modelo, recibe el nombre development, D. La que sirve para valorar la capacidad predictiva, descontando el sobre-ajuste, el de Validation, V. De esta forma, el estudio tipo 1 sólo tiene una muestra para desarrollar el modelo; si bien el subtipo 1b valora la capacidad predictiva mediante remuestreo. El tipo 2 divide la muestra en 2, una para aprender y otra para valorar, distinguiendo según la división se haga, o no, al azar. Note que nuestros datos futuros en los que aplicaremos el modelo no serán una muestra aleatoria, por lo que aplicar criterios geográficos o temporales aproxima más los resultados obtenidos a los futuros. El estudio de tipo 3 tiene ya desde el principio 2 muestras diferentes, mientras que el tipo 4 dedica la muestra a valorar un modelo previo. Si en alguno de estos estudios la valoración de la capacidad predictiva fuera muy pobre, convendría actualizar el modelo, lo que aconseja una nueva valoración futura.

9.6. Discriminar y Calibrar

La clásica "bondad del ajuste" incluyo 2 conceptos: primero, la capacidad de predecir el futuro, de discriminar; y segundo, un buen calibrado en el sentido de que los valores predichos se correspondan con los valores medios observados. Cada figura muestra, en su eje horizontal de las abscisas el valor predicho; y en el vertical de ordenadas, el valor observado. En la primera fila, a la izquierda, vemos una muy buena correspondencia entre lo predicho y lo observado, muy menor a la derecha. En la segunda fila, vemos a la izquierda que la media de los valores observados coincide con el valor predicho, indicando buena calibración. A diferencia de la figura de la derecha. Note que esta última figura, a pesar de tener pobre calibración sería capaz de una buena predicción, por lo que discriminaría bien.

La medida más popular de discriminación es el área bajo la curva ROC que valora el grado de separación entre las distribuciones de casos con y sin el evento. Cuanto más separadas estén, mayor será su valor. No entraremos más a fondo. Puede encontrar en youtube un vídeo nuestro que la explica. Digamos aquí que equivale al estadístico C de concordancia, que cuantifica la probabilidad de que un enfermo tenga un valor mayor en la escala que un sano.

Una medida de discriminación propuesta por Cox es la distancia entre casos y controles del valor medio de las predicciones. En la primera figura de la izquierda, aquellos que no presentaron el evento tienen una media de todas sus predicciones de 0.4; y los que sí lo presentaron de 0.7, por lo que su diferencia es 0.3. En la figura central, ha aumentado a 0.34, pero en la de la derecha ha disminuido a 0.24.

Veamos ahora un ejemplo de calibrado. Y bueno. Esta figura muestra los pacientes en 5 columnas según su nivel de riesgo, empezando por los pacientes con entre 0 y 15 puntos. Sí, lo ha visto bien, este subgrupo tiene casi 70000 casos. La primera fila proporciona el promedio del riesgo predicho para todos ellos: un 0.24%. La segunda fila muestra la proporción de casos que han tenido el evento. También un 0.24%. Y la tercera, la misma proporción pero para la nueva muestra que valora la capacidad predicativa. Una vez más, un 0.24%. Un calibrado perfecto para este grupo. En el siguiente grupo de riesgo, lo primero que vemos es que sube el riesgo a 1.2%. Y el calibrado sigue siendo bueno, pero no perfecto. Y de forma parecida en los otros 3 grupos.

Recuperemos este gráfico que vimos al inicio. En abscisas, la probabilidad predicha para la muestra de 1241 casos dividida en 10 grupos de unos 124 casos. En ordenadas, la proporción observada de eventos en esos 10 grupos. Si todos los puntos medios estuvieran en la diagonal, el calibrado sería perfecto. Casi. Hay 2 subgrupos que presentan más eventos de los predichos. Esta figura permite, por tanto, detectar en qué subgrupos y en qué dirección se producen los desvíos. Además, los intervalos muestran la incertidumbre sobre el promedio de riesgo observado en cada grupo.

¿Y los valores de P? Pues no, que aquí tampoco. Mejor usar gráficos.

Estamos llegando al final. Así, ya tenemos las condiciones para comentar los 2 grandes remedios de NIH, mayor financiador mundial de investigación biomédica, para aumentar el retorno de la inversión en la investigación en salud. Están hartos de financiar investigaciones que se publican pero no se replican. Para ellos, hay 1 causa: pobre formación metodológica. Por lo que propone 2 remedios: formación metodológica de quienes reciban recursos y uso de guías para seleccionar los proyectos que financien.

Recuerde que conviene cuantificar el grado de certeza.