Poco ruido y muchas nueces

Sí, ya sé que es al revés. Ese precisamente es el problema con tanta nueva tecnología de la información. Hoy día cualquiera puede escribir y hacer público lo que se le pase por la cabeza, llegando a un montón de gente, aunque lo que diga sea una chorrada (y no, yo no me doy por aludido, ¡a mí no me lee ni mi cuñado!). Lo malo es que gran parte de lo que se escribe no vale un bit, por no referirme a ningún tipo de excretas. Hay mucho ruido y pocas nueces, cuando a todos nos gustaría que ocurriese lo contrario.

Lo mismo pasa en medicina cuando necesitamos información para tomar alguna de nuestras decisiones clínicas. Vayamos a la fuente que vayamos, el volumen de información no solo nos desbordará, sino que encima la mayoría no nos servirá para nada. Además, incluso si encontramos un trabajo bien hecho es posible que no sea suficiente para contestar completamente a nuestra pregunta. Por eso nos gustan tanto las revisiones de la literatura que algunas almas generosas publican en las revistas médicas. Nos ahorran el trabajo de revisar un montón de artículos y nos resumen las conclusiones. Estupendo, ¿no?. Pues a veces sí y a veces no.

Las revisiones tienen también sus limitaciones, que debemos saber valorar. Quizás la más habitual, y generalmente más fácil de digerir, sea la que se conoce como revisión narrativa o de autor. Este tipo de revisiones las suele hacer, generalmente, un experto en el tema, que revisa la literatura y analiza lo que encuentra como lo cree conveniente (para eso es experto) y que hace un resumen de síntesis cualitativa con sus conclusiones de experto. Este tipo de revisiones son buenas para hacernos una idea general sobre un tema, pero no suelen servir para responder a preguntas concretas. Además, como no se especifica cómo se hace la búsqueda de la información, no podemos reproducirla ni comprobar que incluya todo lo importante que haya escrito sobre el tema.

El otro tipo de revisión es la llamada revisión sistemática (RS), que se centra en una pregunta concreta, sigue una metodología de búsqueda y selección de la información claramente especificada y realiza un análisis riguroso y crítico de los resultados encontrados. Incluso, si los estudios primarios son parecidos, la RS va más allá de la síntesis cualitativa, realizando también un análisis de síntesis cuantitativa, que tiene el bonito nombre de metanálisis. El prototipo de RS es la realizada por la Colaboración Cochrane, que ha elaborado una metodología específica. Pero, si queréis mi consejo, haced una lectura crítica incluso si la revisión la han hecho ellos.

Y para hacerlo, nada mejor que revisar sistemáticamente nuestros tres

pilares: validez, importancia y aplicabilidad.

En cuanto a la **VALIDEZ**, trataremos de determinar si la revisión nos da unos resultados no sesgados y que responden correctamente a la pregunta planteada. Como siempre, buscaremos unos **criterios primarios** de validez. Si estos no se cumplen pensaremos si es ya la hora de pasear al perro: probablemente aprovechemos mejor el tiempo.

¿Se ha planteado claramente el tema de la revisión?. Toda RS debe tratar de responder a una pregunta concreta que sea relevante desde el punto de vista clínico, y que habitualmente se plantea siguiendo el esquema PICO de una pregunta clínica estructurada. Es preferible que solo haya una pregunta, ya que si hay varias se corre el riesgo de no responder adecuadamente a ninguna. Esta pregunta determinará, además, el tipo de estudios que debe incluir la revisión, por lo que debemos valorar **si se ha incluido el tipo adecuado**. Deben especificarse los criterios de inclusión y exclusión de los trabajos, además de considerarse sus aspectos referentes al ámbito de realización, grupos de estudio, resultados, etc. Diferencias entre los trabajos incluidos en cuanto a los (P)pacientes, la (I)intervención o los (O)resultados hacen que dos RS que se plantean la misma preguntan puedan llegar a conclusiones diferentes.

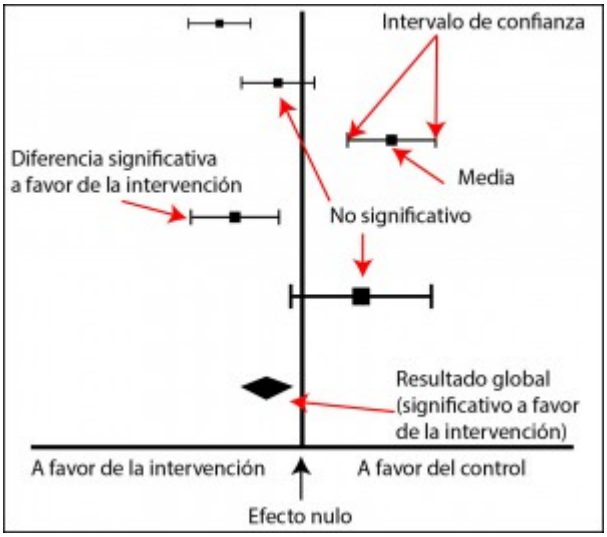
Si se cumple lo anterior, pasaremos a considerar los **criterios secundarios**. **¿Se han incluido los estudios importantes que tienen que ver con el tema?.** Debemos comprobar que se ha realizado una búsqueda global y no sesgada de la literatura. Lo frecuente es hacer la búsqueda electrónica incluyendo las bases de datos más importantes (generalmente PubMed, Embase y la Cochrane Library), pero esta debe completarse con una estrategia de búsqueda en otros medios para buscar otros trabajos (referencias de los artículos encontrados, contacto con investigadores conocidos, industria farmacéutica, registros nacionales e internacionales, etc), incluyendo la denominada literatura gris (tesis, informes, etc), ya que puede haber trabajos importantes no publicados. Y que nadie se extrañe de esto último: está demostrado que los trabajos que obtienen conclusiones negativas tienen más riesgo de no publicarse, por lo que no aparecen en las RS. Debemos comprobar que los autores han descartado la posibilidad de este sesgo de publicación. En general, todo este proceso de selección se suele plasmar en un diagrama de flujo que muestra el devenir de todos los trabajos valorados en la RS.

Es muy importante que **se haya hecho lo suficiente para valorar la calidad de los estudios**, buscando la existencia de posibles sesgos. Además, esto debe hacerse de forma independiente por dos autores y, de forma ideal, sin conocer los autores del trabajo o la revista de publicación. Además, debe quedar registrado el grado de concordancia entre los dos revisores.

Por último, en el caso de que se hayan combinado los resultados de los

estudios para sacar conclusiones comunes (con o sin metanálisis), debemos preguntarnos si **es razonable combinar los resultados de los estudios primarios**. Es fundamental para poder sacar conclusiones de datos combinados que los trabajos sean homogéneos y que las diferencias entre ellos sean debidas únicamente al azar. Aunque cierta variabilidad de los estudios aumenta la validez externa de las conclusiones, no podremos unificar los datos para el análisis si la variabilidad es grande. Hay numerosos métodos para valorar la homogeneidad en los que no vamos a entrar ahora, pero sí que vamos a insistir en la necesidad de que los autores de la revisión lo hayan estudiado de forma adecuada.

En cuanto a la **IMPORTANCIA** de los resultados debemos considerar **cuál es el resultado global de la revisión y si la interpretación se ha hecho de forma juiciosa**. La RS debe proporcionar una estimación global del efecto de la intervención en base a una media ponderada de los artículos de calidad incluidos. Lo más frecuente es que se expresen medidas relativas como el riesgo relativo o la odds ratio, aunque lo ideal es que se complementen con medidas absolutas como la reducción absoluta del riesgo o el número necesario a tratar (NNT). Además, hay que **valorar la precisión de los resultados**, para lo que recurriremos a nuestros queridos intervalos de confianza, que nos darán una idea de la precisión de la estimación de la verdadera magnitud del efecto en la población.



Los resultados de los metanálisis se suelen representar de una manera estandarizada. Se dibuja un gráfico con una línea vertical de efecto nulo (en el uno para riesgo relativo y odds ratio y en el cero para diferencias de medias) y se representa cada estudio como una marca (su resultado) en medio de un segmento (su intervalo de confianza). Los estudios con resultados con significación estadística son los que no cruzan la línea vertical. Generalmente, los estudios más potentes tienen intervalos más estrechos y contribuyen más al resultado global, que se expresa como un diamante cuyos extremos laterales representan su intervalo de confianza. Solo los diamantes que no crucen la línea vertical tendrán significación estadística. Además, cuanto más estrechos, más precisión. Y, por último, cuánto más se alejen de la línea de efecto nulo, más clara será la diferencia entre los tratamientos o las exposiciones comparadas.

Concluiremos la lectura crítica de la RS valorando la **APLICABILIDAD** de los resultados en nuestro medio. Habrá que preguntarse **si podemos aplicar los resultados a nuestros pacientes** y cómo van a influir en la atención que

les prestamos. Tendremos que fijarnos si los estudios primarios de la revisión describen a los participantes y si se parecen a nuestros pacientes. Además, aunque ya hemos dicho que es preferible que la RS se oriente a una pregunta concreta, habrá que ver **si se han considerado todos los resultados relevantes para la toma de decisiones en el problema en estudio**, ya que a veces será conveniente que se considere alguna otra variable secundaria adicional. Y, como siempre, habrá que valorar la **relación beneficios-costes-riesgos**. El que la conclusión de la RS nos parezca válida no quiere decir que tengamos que aplicarla de forma obligada.

Si queréis valorar correctamente una RS sin olvidar ningún aspecto importante os recomiendo que uséis una lista de verificación como la PRISMA o alguna de las herramientas disponibles en Internet, como las parrillas que se pueden descargar de la página de CASPe.

Como veis, no hemos hablado prácticamente nada del metanálisis, con todas sus técnicas estadísticas para valorar homogeneidad y sus modelos de efectos fijos y aleatorios. Y es que el metanálisis es una fiera a la que hay que echar de comer aparte. Pero esa es otra historia…

¿A qué lo atribuye?

Parece que fue ayer. Yo empezaba mis andanzas en los hospitales y tenía mis primeros contactos con El Paciente. Y de enfermedades no es que supiese demasiado, por cierto, pero sabía sin necesidad de pensar en ello cuáles eran las tres preguntas con las que se iniciaba toda buena historia clínica: ¿qué le pasa?, ¿desde cuándo?, ¿a qué lo atribuye?.

Y es que la necesidad de saber el porqué de las cosas es inherente a la naturaleza humana y, por supuesto, tiene gran importancia en medicina. Todo el mundo está loco por establecer relaciones de causa-efecto, por lo que a veces estas relaciones se hacen sin mucho rigor y llega uno a creerse que el culpable de su catarro de verano es el fulano del supermercado, que ha puesto el aire acondicionado muy fuerte. Por eso es de capital importancia que los estudios sobre etiología se realicen y se valoren con rigor. Por eso, y porque cuando hablamos de causa nos referimos también a las que hacen daño, incluidas nuestras propias acciones (lo que la gente culta llama iatrogenia).

Esta es la razón de que los estudios de etiología/daño tengan diseños

similares. El ideal sería el ensayo clínico, y podemos usarlo, por ejemplo, para saber si un tratamiento es la causa de la curación del paciente. Pero cuando estudiamos factores de riesgo o exposiciones nocivas, el principio ético de no maleficencia nos impide aleatorizar las exposiciones, por lo que hemos de recurrir a estudios observacionales como los estudios de cohortes o los estudios de casos y controles, aunque siempre el nivel de evidencia será menor que el de los estudios experimentales.

Para valorar críticamente un trabajo sobre etiología/daño recurriremos a nuestros consabidos pilares: validez, importancia y aplicabilidad.

En primer lugar nos centraremos en la **VALIDEZ** o rigor científico del trabajo, que debe responder a la pregunta sobre si el factor o la intervención que estudiamos fue la causa del efecto adverso o la enfermedad producida.

Como siempre, buscaremos unos **criterios primarios** de validez. Si estos no se cumplen, dejaremos el trabajo y nos dedicaremos a otra cosa más provechosa. Lo primero será determinar si **se han comparado grupos similares en cuanto a otros factores determinantes del efecto diferentes de la exposición estudiada.** La aleatorización de los ensayos clínicos facilita que los grupos sean homogéneos, pero no podemos contar con ella en el caso de estudios observacionales. La homogeneidad de las dos cohortes es fundamental y sin ella el estudio no tendrá validez. Uno siempre se puede defender diciendo que ha estratificado por las diferencias entre los dos grupos o que ha hecho un análisis multivariante para controlar el efecto de las variables confusoras conocidas pero, ¿qué hacemos con las desconocidas?. Lo mismo se aplica a los estudios de casos y controles, mucho más sensibles a sesgos y confusiones.

¿Se han valorado la exposición y el efecto de la misma forma en todos los grupos?. En los ensayos y cohortes debemos comprobar que el efecto ha tenido la misma probabilidad de aparecer y ser detectado en los dos grupos. Por otra parte, en los estudios de casos y controles es muy importante valorar adecuadamente la exposición previa, por lo que debemos investigar si ha habido posibles sesgos de recogida de datos, como el sesgo de memoria (los enfermos suelen acordarse mejor de sus síntomas pasados que los sanos). Por último, debemos considerar si **el seguimiento ha sido lo suficientemente largo y completo**. Las pérdidas durante el estudio, frecuentes en los diseños observacionales, pueden sesgar los resultados.

Si hemos contestado sí a las tres preguntas anteriores, pasamos a considerar los **criterios secundarios** de validez. Los resultados del estudio deben ser evaluados para determinar si **la asociación entre exposición y**

efecto satisface las pruebas de causalidad razonable. Una herramienta que podemos usar son los criterios de Hill, que fue un señor que sugirió utilizar una serie de aspectos para tratar de distinguir el carácter causal o no causal de una asociación. Estos criterios son los siguientes: a) fuerza de la asociación, que es la razón de riesgos de exposición y efecto, que consideraremos en breve; b) consistencia, que es la reproducibilidad en poblaciones o situaciones diferentes; c) especificidad, que quiere decir que una causa produce un único efecto y no múltiples; d) temporalidad: es fundamental que la causa preceda al efecto; e) gradiente biológico: a más intensidad de causa, mayor intensidad de efecto; f) plausibilidad: tiene que tener su lógica según nuestros conocimientos biológicos; g) coherencia, que no entre en conflicto con lo que se sabe de la enfermedad o el efecto; h) evidencia experimental, difícil de obtener muchas veces en humanos por problemas éticos; y, finalmente, i) analogía con otras situaciones conocidas. Aunque estos criterios son ya viejecillos y alguno puede ser irrelevante (evidencia experimental o analogía) o erróneo (especificidad), pueden servirnos de orientación. El criterio de temporalidad sería necesario y se complementaría muy bien con los de gradiente biológico, plausibilidad y coherencia.

Otro aspecto importante es estudiar si, **al margen de la intervención en estudio, se han tratado los dos grupos de forma similar**. En este tipo de estudios en los que el doble ciego brilla por su ausencia es en los que hay más riesgo de sesgo debido a cointervenciones, sobre todo si éstas son tratamientos con un efecto mucho mayor que la exposición en estudio.

En cuanto a la **IMPORTANCIA** de los resultados, debemos considerar la magnitud y la precisión de la asociación entre exposición y efecto.

¿Cuál fue la fuerza de la asociación?. La medida de asociación más habitual es el riesgo relativo (RR), que podremos usar en los ensayos y en los estudios de cohortes. Sin embargo, en los estudios de casos y controles desconocemos la incidencia del efecto (ya se ha producido al realizarse el estudio), por lo que utilizamos la odds ratio (OR). Como ya sabemos, la interpretación de los dos parámetros es similar. Incluso los dos son similares cuando la frecuencia del efecto es muy baja. Sin embargo, cuánto mayor es la magnitud o la frecuencia del efecto, más diferentes son RR y OR, con la peculiaridad de que la OR tiende a sobreestimar la fuerza de la asociación cuando es mayor que 1 y a subestimarla cuando es menor que 1. De todas formas, estos caprichos de la OR excepcionalmente nos modificarán la interpretación cualitativa de los resultados.

Hay que tener en cuenta que en un ensayo es válido cualquier valor de OR o RR cuyo intervalo de confianza no incluya el uno, pero en estudios observacionales hay que ser un poco más exigente. Así, en un estudio de cohortes daremos valor a RR mayores o iguales a tres y, en uno de casos y controles, a OR de cuatro o más.

Otro parámetro muy útil (en ensayos y cohortes) es la diferencia de riesgos o diferencia de incidencias, que es una forma rebuscada de llamar a nuestra conocida reducción absoluta de riesgo (RAR), que nos permite calcular el NNT (o NND, número necesario a dañar), parámetro que mejor nos cuantifica la importancia clínica de la asociación. También, similar a la reducción relativa del riesgo (RRR), contamos con la fracción atribuible en los expuestos, que es el porcentaje de riesgo observado en los expuestos que se debe a la exposición.

Y, **¿cuál es la precisión de los resultados?**. Como ya sabemos, tiraremos de nuestros queridos intervalos de confianza, que nos servirán para determinar la precisión de la estimación del parámetro en la población. Siempre es conveniente disponer de todos estos parámetros, por lo que deben figurar en el estudio o debe ser posible su cálculo a partir de los datos proporcionados por los autores.

Para finalizar, nos fijaremos en la **APLICABILIDAD** de los resultados en nuestra práctica.

¿Son aplicables los resultados a nuestros pacientes?. Buscaremos si hay diferencias que desaconsejen extrapolar los resultados del trabajo a nuestro medio. Además, consideraremos **cuál es la magnitud del riesgo** en nuestros pacientes en función de los resultados del estudio y de las características del paciente en quien queramos aplicarlos. Y, finalmente, teniendo todos estos datos en mente, habrá que pensar en nuestras condiciones de trabajo, las alternativas que tenemos y las preferencias del paciente para decidir si hay que evitar la exposición que se ha estudiado. Por ejemplo, si la magnitud del riesgo es alta y disponemos de una alternativa eficaz la decisión está clara, pero las cosas no siempre serán tan sencillas.

Como siempre, os aconsejo que utilicéis los recursos CASPe para valorar los trabajos, tanto las parrillas adecuadas a cada diseño para hacer la lectura crítica, como las calculadoras para valorar la importancia de los resultados.

Antes de acabar, dejadme aclarar una cosa. Aunque hemos comentado que en las cohortes y ensayos usamos RR y en los casos y controles usamos OR, podemos usar OR en cualquier tipo de estudio (no así RR, para los cuáles hay que conocer la incidencia del efecto). El problema es que son algo menos precisas, por lo que se prefieren los RR y los NNT, cuando es posible

utilizarlos. De todas formas, la OR es cada vez más popular por otro motivo, y es su utilización en los modelos de regresión logística, que nos permiten obtener estimadores ajustados por las diferentes variables de confusión. Pero esa es otra historia…

Doctor, ¿es grave?

Me pregunto cuántas veces habré escuchado esta pregunta o alguna de sus muchas variantes. Porque resulta que siempre estamos pensando en ensayos clínicos y en preguntas sobre diagnóstico y tratamiento, pero pensad si algún paciente os preguntó alguna vez si el tratamiento que le proponíais estaba refrendado por un ensayo clínico aleatorizado y controlado. A mí, al menos, no me ha pasado nunca. Pero sí que a diario me preguntan qué les va a ocurrir en el futuro.

Y de aquí deriva la importancia de los estudios sobre pronóstico. Tened en cuenta que no siempre se puede curar y que, por desgracia, muchas veces lo único que podemos hacer es acompañar y aliviar lo que podamos ante el anuncio de graves secuelas o de la muerte. Pero para esto es fundamental disponer de información de buena calidad sobre el futuro de la enfermedad de nuestro paciente. Esta información nos servirá también para calibrar los esfuerzos terapéuticos en cada situación en función de los riesgos y los beneficios. Y, además, los estudios sobre pronóstico sirven para comparar resultados entre servicios u hospitales diferentes. A nadie se le ocurre decir que un hospital es peor que otro porque su mortalidad es mayor sin comprobar antes que el pronóstico de sus pacientes sea semejante.

Antes de meternos con la lectura crítica de los artículos sobre pronóstico aclaremos la diferencia entre factor de riesgo y factor pronóstico. El factor de riesgo es una característica del ambiente o del sujeto que favorece el desarrollo de la enfermedad, mientras que el factor pronóstico es aquél que, una vez que se produce la enfermedad, influye sobre su evolución. Factor de riesgo y factor pronóstico son cosas diferentes, aunque a veces pueden coincidir. Lo que sí comparten los dos son el mismo diseño de tipo de estudio. Lo ideal sería utilizar ensayos clínicos, pero la mayor parte de las veces no podemos o no es ético aleatorizar los factores pronóstico o de riesgo, por lo que lo habitual es que se usen estudios de cohortes. En los casos que precisan seguimientos muy largos o en los que el efecto que queremos medir es muy raro se pueden usar estudios de casos y controles, pero siempre serán menos potentes por

tener más riesgo de sesgo.

Un estudio de pronóstico nos debe informar de tres aspectos: qué resultado queremos valorar, qué probabilidad hay de que suceda y en qué periodo de tiempo esperamos que pase. Y para valorarlo, como siempre, nos asentaremos sobre nuestros tres pilares: validez, importancia y aplicabilidad.

Para valorar la **VALIDEZ** tendremos primero en cuenta si cumple una serie de **criterios primarios** o de eliminación. Si la respuesta es no, tirad el artículo y mirad a ver qué chorrada nueva han escrito vuestros amigos en Facebook.

¿Está bien definida la muestra de estudio y es representativa de pacientes en un momento similar de la enfermedad?. La muestra, que se suele denominar cohorte incipiente o cohorte de inicio, debe estar formada por un grupo amplio de pacientes en el mismo momento de la enfermedad, idealmente al inicio, y que se sigue de forma prospectiva. Debe estar bien especificado el tipo de pacientes incluidos, los criterios para diagnosticarlos y el método de selección. Además, debemos comprobar **que el seguimiento haya sido lo suficientemente largo y completo** como para observar el evento que estudiamos. Cada participante debe seguirse desde el inicio hasta que sale del estudio, ya sea porque se cure, porque presenta el evento o porque el estudio se acaba. Es muy importante tener en cuenta las pérdidas durante el estudio, muy habituales en diseños con seguimiento largo. El estudio debe proporcionar las características de los pacientes perdidos y los motivos para la pérdida. Si son similares a los que no se pierden, probablemente los resultados sean válidos. Si las pérdidas son de más de un 20% se suele hacer un análisis de sensibilidad utilizando el escenario de "el peor de los casos": consideramos que todas las pérdidas han tenido mal pronóstico y recalculamos los resultados para ver si se modifican, en cuyo caso quedaría invalidado el estudio.

Una vez vistos estos dos aspectos, pasamos a los **criterios secundarios** de validez interna o rigor científico.

¿Se han medido los resultados de forma objetiva y no sesgada?. Debe especificarse con claridad qué se va a medir y cómo antes de iniciar el estudio. Además, lo ideal es que la medición de los resultados se haga de forma ciega para el experimentador, que debe desconocer si el sujeto en cuestión está sometido a alguno de los factores pronósticos para evitar el sesgo de información.

¿Se han ajustado los resultados según todos los valores pronósticos relevantes?. Hay que tener en cuenta todas las variables confusoras y los factores pronósticos que puedan influir en los resultados. En el caso de que se conozcan por estudios previos pueden tenerse en cuenta los factores

conocidos. En caso contrario, los autores determinarán los efectos mediante análisis estratificado de los datos (el método más sencillo) o mediante el análisis multivariante (más potente y complejo), habitualmente mediante un modelo de riesgos proporcionales o de regresión de Cox. Aunque no vamos a entrar ahora en los modelos de regresión, sí que hay dos cosas sencillas que podemos tener en cuenta. La primera, estos modelos necesitan de un número determinado de eventos por cada variable incluida en el modelo, así que desconfiad cuando se analicen muchas variables, sobre todo con muestras pequeñas. La segunda, las variables las decide el autor y son diferentes de un trabajo a otro, por lo que tendremos que valorar si no se ha incluido alguna que pueda ser relevante para el resultado final.

¿Se han validado los resultados en otros grupos de pacientes?. Cuando hacemos grupos de variables y empezamos a comparar unos con otros corremos el riesgo de que el azar nos juegue una mala pasada y nos muestre asociaciones que realmente no existen. Por eso, cuando se describe un factor de riesgo en un grupo (grupo de entrenamiento o derivación), conviene replicar los resultados en un grupo independiente (grupo de validación) para estar seguros de la relación.

A continuación, debemos fijarnos en **cuáles son los resultados** para determinar su **IMPORTANCIA**. Para esto comprobaremos si se proporciona la estimación de la probabilidad de que suceda el desenlace de estudio, la precisión de esta estimación y el riesgo asociado a los factores que modifican el pronóstico.

¿Se especifica la probabilidad del suceso en un periodo de tiempo determinado?. Hay varias formas de presentar el número de sucesos que se producen durante el periodo de seguimiento. La más sencilla sería dar una tasa de incidencia (sucesos/persona/unidad de tiempo) o la frecuencia acumulada en un momento dado. Otra forma es dar la mediana de supervivencia, que no es más que el momento del seguimiento en el cuál el suceso se ha producido en la mitad de la cohorte (recordad que aunque hablemos de supervivencia, el suceso no tiene que ser obligatoriamente la muerte).

Para determinar la probabilidad de que se produzca el suceso en cada periodo y el ritmo al cual se va presentando pueden utilizarse curvas de supervivencia de varios tipos. Las tablas actuariales o de vida se utilizan para muestras grandes, cuando no sabemos el momento exacto del evento y con periodos de tiempo fijos. Sin embargo, probablemente nos encontremos con más frecuencia con las curvas de Kaplan-Meier, que miden mejor la probabilidad del suceso para cada momento concreto con muestras más pequeñas. Con este método se pueden proporcionar los cocientes de riesgos instantáneos en cada momento (los hazard ratios) y la mediana de supervivencia, además de otros parámetros según el modelo de regresión

utilizado.

Para valorar la **precisión de los resultados** buscaremos, como siempre, los intervalos de confianza. Cuanto mayor sea el intervalo, menos precisa será la estimación de la probabilidad del suceso en la población general, que es lo que realmente nos interesa saber. Hay que tener en cuenta que el número de pacientes suele ser menor según pasa el tiempo, por lo que es habitual que las curvas de supervivencia sean más precisas al comienzo que al final del seguimiento. Por último, valoraremos **cuáles son los factores que modifican el pronóstico.** Lo correcto es representar todas las variables que puedan influir sobre el pronóstico con sus correspondientes riesgos relativos, que serán los que nos permitan evaluar la importancia clínica de la asociación.

Por último, tendremos que considerar la **APLICABILIDAD** de los resultados. **¿Son aplicables a mis pacientes?.** Buscaremos las similitudes entre los pacientes del estudio y los nuestros y evaluaremos si las diferencias que encontremos nos permiten extrapolar los resultados a nuestra práctica. Pero además, **¿son útiles los resultados?.** El que sean aplicables no quiere decir que tengamos que ponerlos en práctica obligatoriamente, sino que tendremos que valorar cuidadosamente si nos van a ayudar a decidir qué tratamiento aplicar o a cómo informar a nuestro paciente o a sus familiares.

Como siempre, os recomiendo que uséis alguna plantilla, como las que proporciona CASPe, para realizar la lectura crítica de forma sistemática y no dejar ningún aspecto importante sin valorar.

Ya veis que los trabajos sobre pronóstico tienen mucha miga. Y eso que no hemos comentado prácticamente sobre modelos de regresión y curvas de supervivencia, que muchas veces son el núcleo del estudio estadístico de este tipo de trabajos. Pero esa es otra historia…

El rey a examen

Todos sabemos que el ensayo clínico aleatorizado es el rey de los diseños metodológicos de intervención. Es el tipo de estudio epidemiológico que permite un mejor control de los errores sistemáticos o sesgos, ya que el investigador controla las variables del estudio y los participantes son asignados al azar entre las intervenciones que se comparan. Se entiende entonces que el ensayo clínico, bien de forma directa o como parte de un metaanálisis, constituya la prueba de mejor calidad científica para apoyar

(o no) la eficacia de una intervención y que sea el diseño preferente de los estudios científicos sobre tratamiento.

Claro que esto no quiere decir que cuando veamos que un artículo nos cuenta un ensayo clínico nos podamos relajar y darlo por bueno. El ensayo clínico puede también contener sus trampas y argucias, por lo que, como con cualquier otro tipo de trabajo, será buena práctica realizar una lectura crítica del mismo, basándonos en nuestros tres pilares: validez, importancia y aplicabilidad.

Como siempre, a la hora de estudiar el rigor científico o **VALIDEZ**, nos fijaremos primero en una serie de **criterios primarios** imprescindibles. Si estos no se cumplen, mejor no perder el tiempo con el trabajo y buscar otro más provechoso.

¿Existe un pregunta clínica claramente definida?. Se debe plantear una hipótesis de trabajo con sus correspondientes hipótesis nula y alternativa, a ser posible sobre un tema relevante desde el punto de vista clínico. Es preferible que el estudio trate de responder solo a una pregunta. Cuando se quiere responder a varias suele complicarse el estudio en exceso para acabar no contestando ninguna de forma completa y adecuada.

¿Se realizó la asignación de forma aleatoria?. Para poder afirmar que las diferencias entre los grupos se deben a la intervención es necesario que sean homogéneos. Esto se consigue asignando los pacientes al azar, única forma de controlar las variables confusoras conocidas y, más importante, también las que desconocemos. Si los grupos fueran distintos y atribuyésemos la diferencia únicamente a la intervención podríamos incurrir en un sesgo de confusión. El ensayo debe contener una tabla con la frecuencia de aparición de las variables demográficas y de confusión de ambas muestras para estar seguros de que los grupos son homogéneos. Un error frecuente es buscar las diferencias entre los dos grupos y valorarlas según su p, cuando sabemos que la p no mide homogeneidad. Si los hemos repartido al azar, cualquier diferencia que observemos se deberá obligatoriamente al azar (no necesitaremos una p para saberlo). El tamaño muestral no está pensado para discriminar entre las variables demográficas, por lo que una p no significativa puede indicar simplemente que la muestra es pequeña para verla. Por otro lado, cualquier mínima diferencia puede alcanzar significación estadística si la muestra es lo suficientemente grande. Así que olvidaos de la p: si hay alguna diferencia, lo que hay que hacer es valorar si tiene la relevancia clínica suficiente como para poder haber influido en los resultados.

Hay que considerar también si la **secuencia de aleatorización se hizo de forma correcta.** El método utilizado debe garantizar que todos los componentes de la población seleccionada tengan la misma probabilidad de ser elegidos, por lo que se prefieren las tablas de números aleatorios o

secuencias generadas por ordenador. Y aquí pasa algo muy curioso: resulta que es bien conocido que la aleatorización produce muestras de diferente tamaño, sobre todo si las muestras son pequeñas, motivo por el que a veces se usan muestras aleatorizadas por bloques balanceados en tamaño. Y yo os pregunto, ¿cuántos estudios habéis leído con el mismo número de participantes en las dos ramas y que afirmaban ser aleatorizados?. Desconfiad si veis grupos iguales, sobre todo si son pequeños. Además, la aleatorización debe ser oculta, de forma que no se pueda saber a qué grupo va a pertenecer el siguiente participante. Por eso se prefieren los sistemas centralizados vía telefónica o a través de Internet.

También es importante que **el seguimiento haya sido completo**, de forma que todo participante que entre en el estudio tiene que ser tenido en cuenta al finalizar. Si las pérdidas superan el 20%, se admite que hay que valorar su efecto en los resultados. Lo más habitual suele ser el llamado escenario del peor de los casos: se supone que todas las pérdidas del grupo control han ido bien y todas las del grupo de intervención han ido mal y se repite el análisis para comprobar si las conclusiones se modifican, en cuyo caso la validez del estudio quedaría seriamente comprometida. El último aspecto importante es considerar si los pacientes que no han recibido el tratamiento previamente asignado (siempre hay alguno que no se entera y mete la pata) se han analizado **según la intención de tratamiento**, ya que es la única forma de preservar todos los beneficios que se obtienen con la aleatorización.

Una vez comprobados estos criterios primarios, nos fijaremos en tres **criterios secundarios** que influyen en la validez interna. Habrá que comprobar que los **grupos fueran similares al inicio del estudio** (ya hemos hablado de la tabla con los datos de los dos grupos), que se llevó a cabo el **enmascaramiento** de forma adecuada como forma de control de sesgos y que los **dos grupos fueron manejados y controlados de forma similar** a excepción, claro está, de la intervención en estudio.

Pasaremos a continuación a considerar cuáles son los resultados del estudio para calibrar su **IMPORTANCIA** clínica. Habrá que determinar las variables medidas para ver si el trabajo expresa de forma adecuada la magnitud y la precisión de los resultados. Es importante, una vez más, no conformarnos con que nos inunden con múltiples p llenas de ceros. Recordad que la p solo nos indica la probabilidad de que estemos dando como buenas diferencias que solo existen por azar (o, dicho con elegancia, de cometer un error de tipo 1), pero que significación estadística no tiene porqué ser sinónimo de relevancia clínica.

En el caso de variables continuas como tiempo de supervivencia, peso, tensión arterial, etc, lo habitual será expresar la magnitud de los resultados como diferencia de medias o de medianas, dependiendo de cuál sea

la medida de centralización más adecuada. Sin embargo, en casos de variables dicotómicas (vivo o muerto, sano o enfermo, etc) se utilizarán el riesgo relativo, su reducción relativa y absoluta y el número necesario a tratar (NNT). De todas ellas, la que mejor expresa la eficiencia clínica es siempre el NNT. Cualquier trabajo digno de nuestra atención debe proporcionar estos datos o, en su defecto, la información necesaria para que podamos calcularlos.

Pero para permitir conocer una estimación más real de los resultados en la población necesitamos saber la precisión del estudio, y nada más fácil que recurrir a los intervalos de confianza. Estos intervalos, además de la precisión, nos informan también de la significación estadística. Será estadísticamente significativo si el intervalo del riesgo relativo no incluye el uno y el del NNT el cero. En el caso de que los autores no nos los proporcionen, podemos utilizar una calculadora para obtenerlos, como las disponibles en la web de CASPe.

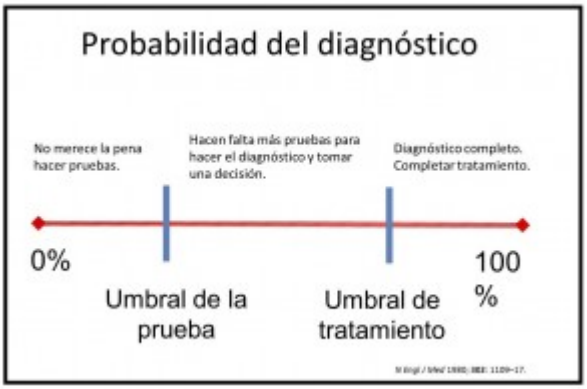
Para finalizar la lectura crítica de un trabajo de tratamiento valoraremos su **APLICABILIDAD**, para lo cual nos tendremos que preguntar si los resultados pueden generalizarse a nuestros pacientes o, dicho de otro modo, si existe alguna diferencia entre nuestros pacientes y los del estudio que impida la generalización de los resultados. Hay que tener en cuenta en este sentido que cuánto más estrictos sean los criterios de inclusión de un estudio, más difícil será generalizar sus resultados, comprometiéndose así su validez externa.

Pero, además, debemos considerar si **se han tenido en cuenta todos los resultados clínicamente importantes**, incluyendo efectos secundarios e indeseables. La variable de resultado medida debe ser importante para el médico y para el paciente. No hay que olvidar que el hecho de que demostrar que la intervención sea eficaz no significa obligatoriamente que sea beneficiosa para nuestros pacientes. Habrá que valorar también los efectos nocivos o molestos y estudiar el **balance beneficios-costes-riesgos**, así como las dificultades que puedan existir para aplicar el tratamiento en nuestro medio, las preferencias del paciente, etc.

Para terminar, recomendaros que utilicéis alguna de las herramientas disponibles para lectura crítica, como las plantillas CASPe, o una lista de verificación, como la CONSORT, para no dejaros ninguno de estos puntos sin considerar. Eso sí, todo lo que hemos hablado se refiere a ensayos clínicos aleatorizados y controlados, ¿Y qué pasa si se trata de ensayos no aleatorizados o de otra clase de estudios cuasiexperimentales?. Pues para eso se siguen otra serie de normas, como pueden ser las de la declaración TREND. Pero esa es otra historia…

Hay que saber lo que se pide

A diario encontramos artículos que nos muestran nuevas pruebas diagnósticas que parecen haber sido diseñadas para solucionar todos nuestros problemas. Pero no debemos caer en la tentación de hacer caso a todo lo que leamos sin recapacitar antes un poco en lo que hemos leído. Al fin y al cabo, si hiciésemos caso a todo lo que leemos estaríamos hinchados de beber Coca-Cola.



Ya sabemos que una prueba diagnóstica no nos va a decir si una persona está o no enferma. Su resultado únicamente nos permitirá aumentar o disminuir la probabilidad de que el individuo esté enfermo o no, de forma que nosotros nos atreveremos a confirmar o descartar el diagnóstico, pero siempre con cierto grado de incertidumbre. Cualquiera tiene cierto riesgo de padecer cualquier enfermedad, que no es más que la prevalencia de la enfermedad en la población general. Pero si, además de pertenecer a la población, uno tiene la desgracia de tener síntomas, esa probabilidad irá aumentando hasta alcanzar un primer umbral en el que se justifique realizar pruebas diagnósticas. La utilidad de la prueba diagnóstica estará en su capacidad para disminuir la probabilidad por debajo de este umbral (y descartar el diagnóstico) o, por el contrario, en aumentarla hasta el umbral en el que se justifique iniciar el tratamiento. Claro que a veces la prueba nos deja a medio camino y tenemos que hacer pruebas adicionales antes de confirmar el diagnóstico con la seguridad suficiente como para comenzar el tratamiento.

Los estudios de pruebas diagnósticas deben proporcionarnos información sobre la capacidad de una prueba para producir los mismos resultados cuando se realiza en condiciones similares (fiabilidad) y sobre la exactitud con la que las mediciones reflejan aquello que miden (validez). Pero, además, deben darnos datos sobre su capacidad discriminatoria (sensibilidad y especificidad), su rendimiento clínico (valor predictivo positivo y valor predictivo negativo) y sobre otros aspectos que nos permitan valorar si nos va a merecer la pena practicarla en nuestros pacientes. Y para comprobar si un estudio nos proporciona la información adecuada tenemos que hacer una lectura crítica basada en nuestros tres pilares: validez, importancia y aplicabilidad.

Comencemos por la VALIDEZ. Lo primero será hacernos unas preguntas

básicas de eliminación o criterios primarios sobre el estudio. Si la respuesta a estas preguntas es no, probablemente lo mejor que podamos hacer es usar el artículo para envolver el bocadillo de media mañana.

¿Se ha comparado la prueba diagnóstica de forma ciega e independiente con un patrón de referencia adecuado?. Hay que revisar que el resultado de la prueba de referencia no se interprete de forma diferente según el resultado de la prueba de estudio, ya que caeríamos en un sesgo de incorporación, que podría invalidar los resultados. Otro problema que puede surgir es que el patrón de referencia tenga muchos resultados poco concluyentes. Si cometemos el error de excluir estos casos dudosos incurriremos en un sesgo de exclusión de indeterminados que, además de sobrestimar la sensibilidad y la especificidad de la prueba, comprometería la validez externa del estudio, que solo sería aplicable a los pacientes con resultado no dudoso.

¿Los pacientes abarcan un espectro similar al que nos vamos a encontrar en nuestra práctica?. Deben estar claros los criterios de inclusión del estudio, en el que deben participar sanos y enfermos con distinta gravedad o evolución de la enfermedad. Como ya sabemos, la prevalencia influye en el rendimiento clínico de la prueba, con lo que si la validamos, por ejemplo, en un centro terciario (estadísticamente la probabilidad de estar enfermo será mayor) puede sobrestimarse su capacidad diagnóstica si va a utilizarse en un centro de Atención Primaria o en población general (en el que la proporción de enfermos será menor).

Llegados a este punto, si creemos que merece la pena seguir leyendo, pasaremos a los criterios secundarios, que son aquellos que aportan un valor añadido al diseño del estudio. Otra pregunta que debemos hacernos es: ¿influyeron los resultados de la prueba de estudio para decidir si se hacía la de referencia?. Hay que comprobar que no se haya producido un sesgo de secuencia o sesgo de verificación diagnóstica, mediante el cual excluimos a los que tienen la prueba negativa. Aunque esto es habitual en la práctica corriente (empezamos por pruebas sencillas y solo hacemos las caras o las invasoras en los casos positivos), el hacerlo en un estudio de pruebas diagnósticas compromete la validez de los resultados. Ambas pruebas deben hacerse de forma independiente y ciega, de forma que la subjetividad del observador no influya en los resultados (sesgo de revisión o sesgo de valoración ciega). Por último, ¿se describe el método con el detalle suficiente para permitir su reproducción?. Debe quedar claro qué se ha considerado normal y anormal y cuáles han sido los criterios para definir la normalidad y la forma de interpretar los resultados de la prueba.

Una vez analizada la validez interna del estudio pasaremos a considerar la IMPORTANCIA de los datos presentados. El objetivo de un estudio de diagnóstico es determinar la capacidad de una prueba para clasificar

correctamente a los individuos según la presencia o ausencia de enfermedad. En realidad, y para ser más exactos, queremos saber cómo aumenta la probabilidad de estar enfermo tras el resultado de la prueba (probabilidad postprueba). Es, por tanto, esencial que el estudio nos informe acerca de la dirección y magnitud de este cambio (preprueba/postprueba), que sabemos depende de las características de la prueba y, en gran medida, de la prevalencia o probabilidad preprueba.

¿Nos presenta el trabajo las razones de verosimilitud o es posible calcularlas a partir de los datos?. Este dato es fundamental, ya que sin él no podemos calcular el impacto clínico de la prueba de estudio. Hay que tener especial precaución con las pruebas de resultado cuantitativo en las que es el investigador el que establece un punto de corte de normalidad. Cuando se utilizan curvas ROC es frecuente desplazar el punto de corte para favorecer la sensibilidad o la especificidad de la prueba, pero tenemos que valorar siempre cómo afecta esta medida a la validez externa del estudio, ya que puede limitar su aplicabilidad a un grupo determinado de pacientes.

¿Son fiables los resultados?. Habrá que determinar si los resultados son reproducibles y cómo pueden verse afectados por variaciones entre diferentes observadores o al repetir la prueba de forma sucesiva. Pero no solo hay que valorar la fiabilidad, sino también cuán precisos son los resultados. El estudio se hace sobre una muestra de pacientes, pero debe proporcionar una estimación de sus valores en la población, por lo que los resultados deben expresarse con sus correspondientes intervalos de confianza.

El tercer pilar de la lectura critica es el de la APLICABILIDAD o validez externa, que nos ayudará a determinar si los resultados son útiles para nuestros pacientes. En este sentido, debemos hacernos tres preguntas. ¿Disponemos de esta prueba y es factible realizarla en nuestros pacientes?. Si no disponemos de la prueba lo único que habremos conseguido leyendo el estudio es aumentar nuestros vastos conocimientos. Pero si disponemos de ella debemos preguntarnos si nuestros pacientes cumplirían los criterios de inclusión y exclusión del estudio y, en caso de que no los cumplan, pensar cómo pueden afectar estas diferencias la aplicabilidad de la prueba.

La segunda pregunta es si conocemos la probabilidad preprueba de nuestros pacientes. Si nuestra prevalencia es muy diferente de la del estudio se puede modificar la utilidad real de la prueba. Una solución puede ser hacer un análisis de sensibilidad valorando cómo se modificarían los resultados del estudio estudiando varios valores de probabilidad pre y postprueba que sean clínicamente razonables.

Por último, deberíamos hacernos la pregunta más importante: ¿la probabilidad postprueba puede hacer cambiar nuestra actitud terapéutica y servir de ayuda para el paciente?. Por ejemplo, si la probabilidad

preprueba es muy baja, probablemente la probabilidad postprueba sea también muy baja y no alcanzará el umbral de justificación terapéutica, con lo que igual no merece la pena gastar dinero y esfuerzos con esa prueba. Por el contrario, si laprobabilidad preprueba es muy alta, en algunos casos merecerá la pena tratar sin hacer ninguna prueba, salvo que el tratamiento sea muy costoso o peligroso. Como siempre, en el medio estará la virtud y será en esas zonas intermedias donde más nos podamos beneficiar del uso de la prueba diagnóstica en cuestión. En cualquier caso, no nos olvidemos nunca de nuestro jefe (me refiero al paciente, no al otro): no hay que contentarse solo con estudiar la eficacia o el coste-efectividad, sino que debemos considerar también los riesgos, molestias y preferencias del paciente, así como las consecuencias que le puede acarrear la realización o no de la prueba diagnóstica.

Si me permitís un consejo, cuando estéis valorando un trabajo sobre pruebas diagnósticas os recomiendo el uso de las plantillas CASPe, que podéis descargaros de su página web. Os ayudarán a hacer la lectura critica de una manera sistemática y sencilla.

Para terminar, comentaros que todo lo dicho hasta ahora vale para los trabajos específicos de pruebas diagnósticas. Sin embargo, la valoración de pruebas diagnósticas puede formar parte de estudios observacionales como los de cohortes o los de casos y controles, que pueden tener alguna peculiaridad en la secuencia de realización y en los criterios de validación de la prueba de estudio y del patrón de referencia, pero esa es otra historia…