La cúspide de la pirámide de la evidencia*

The top of the evidence pyramid

Juan Manuel Senior-Sánchez • Medellín (Colombia)

DOI: https://doi.org/10.36104/amc.2024.3248

Resumen

Un ensayo clínico con asignación aleatoria es la forma más rigurosa para establecer relación causa-efecto, con la menor incertidumbre posible, entre la eficacia de una intervención y un desenlace definido. Sus características más importantes son: la asignación aleatoria a la intervención, el investigador y el sujeto de estudio quienes desconocen cual tratamiento está recibiendo, es decir, están enmascarados; los grupos son tratados en igual forma, excepto por la intervención experimental, los sujetos de estudio son analizados en el mismo grupo al cual fueron asignados en forma aleatoria (análisis por intención a tratar) y el análisis está enfocado sobre la estimación del tamaño de la diferencia en desenlaces predefinidos entre los grupos intervenidos.

Para evaluar la efectividad de las intervenciones, los estudios observacionales deben manejar las variables de confusión que sesgan la asociación al estar relacionadas tanto con la exposición como con la intervención. Estos estudios suelen demostrar correlación, mas no causalidad, aunque los investigadores y los clínicos tienden a interpretarlo erróneamente de esta última forma. La idea de sintetizar el cuerpo de la evidencia en forma matemática a través de metaanálisis es bastante lógica y atractiva, para clasificarla en el más alto nivel. Sin embargo, a pesar de ser bien desarrollados y conducidos, sus resultados están sujetos al diseño de cada estudio original, la heterogeneidad clínica inherente y aspectos metodológicos.

El objetivo del análisis estadístico de un ensayo clínico de asignación aleatoria (ECA) es estimar la magnitud de la diferencia de las intervenciones sobre los desenlaces. Primero, se determina un estimativo puntual que corresponde a la diferencia observada. Luego, se determina el grado de incertidumbre de los datos, usualmente con el uso de los intervalos de confianza del 95% (IC 95%). El tipo de estimativo utilizado depende de la naturaleza del desenlace de interés, básicamente son tres: binario, tiempo al evento y desenlace cuantitativo. **(Acta Med Colomb 2024; 49. DOI: https://doi.org/10.36104/amc.2024.3248).**

Palabras clave: *ensayo clínico, supervivencia, aleatorización, replicabilidad, no inferioridad, hipótesis, Hazard, Kaplan Meier, riesgo, odds.*

Abstract

Randomized clinical trials are the most rigorous way to establish a cause-effect relationship between an intervention's efficacy and the outcomes with the least level of uncertainty. Its key features are that the study subjects are randomly assigned to the intervention, the researcher and the subjects are blinded to the treatment they are receiving, and all groups receive the same care except for the experimental intervention. The intention-to-treat analysis involves analyzing the samples in the same group to which they were randomly assigned. The main goal of this analysis is to find out how much the predetermined outcomes of the intervention groups differed from one another.

To evaluate the effectiveness of the interventions, observational studies must manage confounding variables that bias the association by being related to both the exposure and the intervention. They typically only show correlation, not causality, although investigators and clinicians tend to mistakenly interpret them as showing the latter. The idea of synthesizing the body of evidence mathematically using meta-analysis is very logical and attractive for the highest possible classification. However, despite being well developed and implemented, the results of meta-analyses are subject to the design of each original study, the inherent clinical heterogeneity and methodological aspects.

The objective of the statistical analysis of an RCT is to estimate the magnitude of the difference caused by the interventions on the outcomes. First, a point estimate is determined, which corresponds to the observed difference. Then, the degree of uncertainty of the data must be determined, usually

Dr. Juan Manuel Senior-Sánchez: Cardiólogo intervencionista, especialista en Medicina Crítica y Cuidados Intensivos, Magister en Epidemiología Clínica, Universidad de Antioquia, jefe posgrado cardiología clínica y cardiología intervencionista, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia. Cardiólogo Intervencionista Hospital Universitario San Vicente Fundación sede Medellín y Rionegro y Hospital Alma Mater de Antioquia -HAMA- (Colombia). Correspondencia: Dr. Juan Manuel Senior-Sánchez. Medellín (Colombia).
E-Mail: juan.senior64@gmail.com
*Conferencia Lombana Barreneche dictada el 10 de agosto de 2022, durante el Acto Inaugural del XXVII Congreso Colombiano de Medicina Interna, Bucaramanga, 10-14 de agosto de 2022, Centro de Convenciones NEOMUNDO.
Recibido: 5/V/2023 Aceptado: 21/V/2024

using 95% confidence intervals (95% CI). The type of estimate used depends on the nature of the outcome of interest; there are basically three types: binary, time-to-event and quantitative outcome.
(Acta Med Colomb 2024; 49. DOI: https://doi.org/10.36104/amc.2024.3248).

Keywords: *clinical trial, survival, randomization, replicability, non inferiority, hipothesis, Hazard, Kaplan Meier, risk, odds.*

Ensayos clínicos (parte I)

Introducción

La evidencia empírica resulta del cúmulo de información que puede ser obtenida a través de la observación de diversos fenómenos o de la documentación de ellos, aunque no sea de forma organizada y sistemática. Está relacionada con la evidencia científica, sin embargo, no toda evidencia empírica logra cumplir los estrictos estándares del método científico (1).

El método científico empieza con la escéptica observación de un fenómeno, conocido o desconocido, para analizarlo en su real contexto y no influenciado por creencias o experiencias previas, luego del cual se genera una pregunta de investigación. Posteriormente, se establece el marco conceptual en el que se puede incluir y, a partir de este, generar una hipótesis que debe necesariamente ser probada en un experimento, del cual se analizarán los resultados y se extraerán conclusiones. La replicabilidad de estos permitirá mayor confianza en el resultado (2).

En este punto es importante tener en cuenta tres términos que pueden utilizarse en forma indiferente pero cuyo significado difiere ligeramente: reproducibilidad, replicabilidad y repetibilidad. La *Reproducibilidad* se define como la obtención consistente de los mismos resultados utilizando los mismos datos, pasos computacionales, métodos, códigos y condiciones de análisis, convirtiéndose en sinónimo de reproducibilidad computacional. También se refiere al uso del mismo procedimiento y sistema de medida, bajo las mismas condiciones operativas en un sitio diferente, por lo que se confunde frecuentemente con replicabilidad (grupo diferente, configuración experimental diferente). La *replicabilidad* se refiere a la obtención consistente de resultados similares a través de diferentes estudios que analizan la misma pregunta de investigación con datos propios (grupo diferente, igual configuración experimental). La *repetibilidad* es el término menos utilizado y se refiere a la obtención de los mismos resultados para definir la precisión, realizado por el mismo grupo de investigación, bajo las mismas condiciones (mismo grupo, misma configuración experimental) (3).

El desarrollo de la evidencia científica en el área de la salud se ha dado por la observación sistemática del fenómeno salud-enfermedad por excelsos maestros, quienes lo describieron en sus diversas fases. Esto permitió el desarrollo de técnicas y métodos que nos acercaran a la elusiva verdad con la menor incertidumbre posible, dando lugar a la arquitectura de la investigación científica. De allí surge la necesidad de cambiar el escenario del laboratorio y la experimentación en animales para incluir humanos como sujetos de estudio, con el fin de tomar decisiones con base en la etiología, distribución, diagnóstico, pronóstico y tratamiento, marcando el inicio de la epidemiología clínica (4).

Clasificación de los tipos de investigación clínica

El clínico requiere, por lo tanto, de información que permita establecer la verdadera eficacia de alternativas terapéuticas con la menor incertidumbre posible. La taxonomía de la investigación científica ha llevado a establecer una jerarquía que permite clasificar los diversos tipos de estudios (Figura 1). Describir el tipo de estudio cobra relevancia para evitar errores en la interpretación y el alcance de los resultados (5).

El diseño del estudio debe ser coherente con el tipo de pregunta definida. Los estudios descriptivos están reservados para enfermedades nuevas o desconocidas, que pueden iniciar con un simple reporte de caso o series de casos. Los estudios de corte transversal están dirigidos a determinar prevalencia, puesto que miden al mismo tiempo exposición y desenlace. Los estudios de casos y controles son importantes en enfermedades raras, miran en forma retrospectiva desde el desenlace hacia la exposición, su principal dificultad está en la escogencia de un grupo control adecuado. Los estudios de cohorte llevan una secuencia lógica, desde la exposición al desenlace; son importantes para establecer la evolución natural de la enfermedad, detectar factores de riesgo, para pronóstico y en intervenciones para generar hipótesis y mostrar su comportamiento en el mundo real.

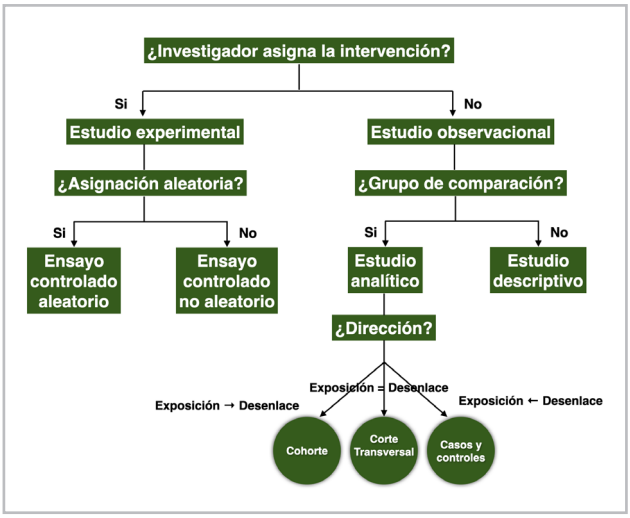

Figura 1. *Clasificación de los tipos de investigación clínica (Tomado y adaptado de: Lancet 2002; 359: 57–61).*

Para evaluar la efectividad de las intervenciones, los estudios observacionales deben manejar las variables de confusión que sesgan la asociación al estar relacionadas tanto con la exposición, en este caso un tratamiento, como con el desenlace establecido. Por lo cual se utilizan métodos de regresión múltiple, pareamiento o estratificación; sin embargo, estas técnicas solo ajustan por confusores observados o conocidos. Se han desarrollado otras técnicas como el análisis del puntaje de propensión (*propensity score*), y el de variables instrumentales, este último tiene la bondad de controlar no solo la confusión residual, sino el sesgo de selección, puesto que un sujeto de estudio puede recibir determinado tratamiento dadas sus características individuales; por ejemplo, presencia de una o más comorbilidades, gravedad de la enfermedad y pronóstico, entre otras (6). En estos casos suele demostrarse correlación, mas no causalidad, aunque los investigadores y los clínicos tienden a interpretarlo erróneamente de esta última forma, cuando es analizado sin el contexto adecuado (7), podría ayudar el análisis de sensibilidad con el nominado *E-value*. Es importante anotar que, ocasionalmente, podemos encontrar causalidad sin una correlación claramente observable.

Un ensayo clínico controlado con asignación aleatoria (ECA) es un experimento para evaluar intervenciones en seres humanos, entendido experimento como una serie de observaciones sistemáticas, bajo condiciones que son controladas por el investigador, de carácter prospectivo y comparativo, puesto que incluye grupo control, que puede ser activo o inactivo como el placebo. En este caso, el investigador controla los factores que pueden influenciar la variabilidad del desenlace, el sesgo de selección, la aplicación inconsistente de una intervención y la evaluación incompleta o sesgada del desenlace. En los estudios no experimentales los sujetos son expuestos a intervenciones por razones que no controla el investigador.

Algunas consideraciones y críticas sobre la validez externa de los ensayos clínicos tradicionales, especialmente en el proceso de asignación aleatoria (8), en consonancia con la disponibilidad de bases de datos gigantes, que muestran el efecto en el mundo real, aunado a la disponibilidad de nuevas herramientas analíticas, han llevado a promulgar al sentido común y a la observación clínica como los métodos preferidos para soportar las decisiones clínicas. Sin embargo, más de cuatro décadas de experiencia de ensayos clínicos bien diseñados y conducidos contradicen esta práctica poco deseable (9). Es importante anotar que se pueden desarrollar ensayos clínicos anidados en cohortes, cuando se dispone de grandes bases de datos.

Por estas razones los ECA (estudio primario) son clasificados en la cúspide de la pirámide de evidencia en conjunto con los metaanálisis de ensayos clínicos (estudio secundario) (Figura 2) (5).

La idea de sintetizar el cuerpo de la evidencia en forma matemática es bastante lógica y atractiva para clasificarla en el más alto nivel. No obstante, a pesar de ser bien desarrollados y conducidos, sus resultados están sujetos al diseño de cada estudio original, la heterogeneidad clínica inherente y aspectos metodológicos (10). Los metaanálisis de estudios observacionales no son equiparables a los realizados con ensayos clínicos con asignación aleatoria. Por tanto, es importante utilizar a las revisiones sistemáticas y metaanálisis como un lente con el cual observar y analizar detenidamente el cuerpo de la evidencia científica (11).

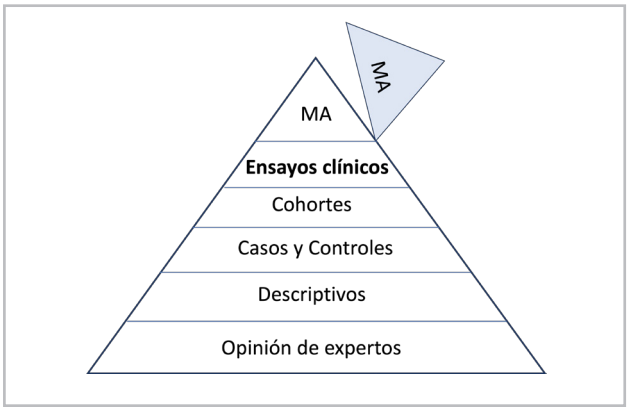

Figura 2. *Pirámide de evidencia científica (MA: metaanálisis).*

Suele abusarse de ellos cuando, al demostrar resultados negativos de mega-ensayos clínicos, diseñados para responder una pregunta de investigación clara, se utilizan para "diluirlos", al mezclarlos con estudios previos que no habían logrado resultados contundentes o para resaltar efectos sobre subgrupos específicos, que solo son generadores de una nueva hipótesis. La combinación de dos o más bases de datos, por los mismos investigadores, sin una estructura y diseño claro, no puede considerase un verdadero metaanálisis, dista mucho del realizado con datos de participantes individuales (puesto que generalmente se hacen con datos agrupados).

Principios éticos

La investigación en seres humanos se rige por principios éticos inviolables para permitir su adecuada participación; son básicamente tres: respeto por las personas, beneficencia y justicia. El *respeto por las personas* se refiere a la autonomía para participar después de deliberar sobre las condiciones y perspectivas del estudio y la protección de las personas con autonomía disminuida. La *beneficencia* se refiere a la obligación de maximizar los beneficios a la luz del conocimiento científico y, por consiguiente, minimizar el daño. La *justicia* es la obligación de tratar a cada persona con lo que se considera moralmente correcto y apropiado, según lo establece las guías CIOMS sobre ética en investigaciones biomédicas (12).

Como lo mencionaba Osler, la medicina es la ciencia de la incertidumbre y el arte de la probabilidad. En este punto es importante introducir el anglicismo *equipoise*, que puede asimilarse al término *incertidumbre*. De acuerdo con este principio ético, los pacientes deben recibir la mejor

alternativa terapéutica disponible en términos de efectividad, de acuerdo con su condición clínica y características individuales. Ante la presencia de opciones diversas, estas deben ser "equiparables"; sin embargo, el ensayo clínico asigna en forma aleatoria el tratamiento sin tener en cuenta esas variables, solo con base en unos criterios de inclusión y exclusión.

Esto plantea un dilema entre el beneficio poblacional y del avance del conocimiento en sí, por los resultados obtenidos y el interés particular de cada paciente. Es importante establecer la plausibilidad biológica con respecto al efecto y la incertidumbre en cuanto a su real eficacia, puesto que no es ético llevar a cabo estudios sin un respaldo sólido experimental que permita predecir con probabilidad aceptable un resultado exitoso, como tampoco si está claro de antemano el beneficio de la terapia (13). Otros aspectos no menos importantes en los ensayos clínicos es que deben sustentar su valor científico y social.

Fases de la investigación clínica

Clásicamente los ensayos se clasifican en fases de I a IV, sin que eso signifique que en algunos casos sean mutuamente excluyentes (Figura 3) (14).

Los estudios fase I están típicamente representados por la evaluación farmacológica de moléculas en humanos sanos. En general evalúan los siguientes aspectos: estimación de la seguridad y tolerabilidad, farmacocinética, farmacodinamia y actividad de la sustancia o su potencial beneficio terapéutico.

Los estudios fase II son usualmente exploratorios terapéuticos iniciales, pueden tener diversidad de diseños, incluir grupos controles o comparaciones con el estado inicial, se incluye una población homogénea con criterios muy estrictos para su monitoreo. Su objetivo más importante radica en la determinación de la dosis y el régimen a administrar en la fase III. Algunos utilizan la subdivisión en fase IIa para establecer dosis y seguridad y IIb para establecer eficacia.

Los estudios fase III son diseñados para demostrar o confirmar la evidencia preliminar obtenida en las fases iniciales, especialmente de la fase II. Incluyen una población más amplia, pueden explorar dosis-respuesta, diferentes estados de la enfermedad, diferentes escenarios y/o combinaciones con otras drogas. Se consideran confirmatorios.

Los estudios fase IV comienzan cuando el medicamento ha sido aprobado para su comercialización, para monitorizar sus efectos cuando es aplicado a una población abierta con potencial beneficio y para la indicación determinada en el registro mercantil.

Diseños comunes de los ensayos clínicos fase III

El diseño de los ensayos clínicos esta básicamente determinado por la solidez de la evidencia previa sobre su posible efecto. Uno de los más socorridos es el de grupos paralelos, en el cual se selecciona en forma aleatoria un grupo de pacientes, para asignarlos a una de dos o más intervenciones. Esta comparación puede darse con grupo control placebo, cuando no hay alternativa terapéutica usual, o contra tratamiento activo en caso de existir terapia estándar; también puede compararse la combinación de terapia estándar más placebo contra terapia estándar más intervención experimental o suspensión de la terapia a evaluar contra su continuación en caso de intervenciones utilizadas por evidencia previa de baja calidad metodológica (15).

En estos protocolos, no es infrecuente utilizar un periodo en que todos los pacientes sean expuestos al efecto de la intervención, antes de ser asignados en forma aleatoria, para evaluar su tolerabilidad, conocido como periodo de *run in*. Esto aumenta la validez interna, pero disminuye la externa, al seleccionar aún más la población incluida en el ensayo. Una situación similar se presenta cuando un número significativo de sujetos que cumplen criterios son excluidos por razones ajenas al protocolo, por ejemplo, la decisión del médico tratante, algunas veces por utilizar la misma intervención

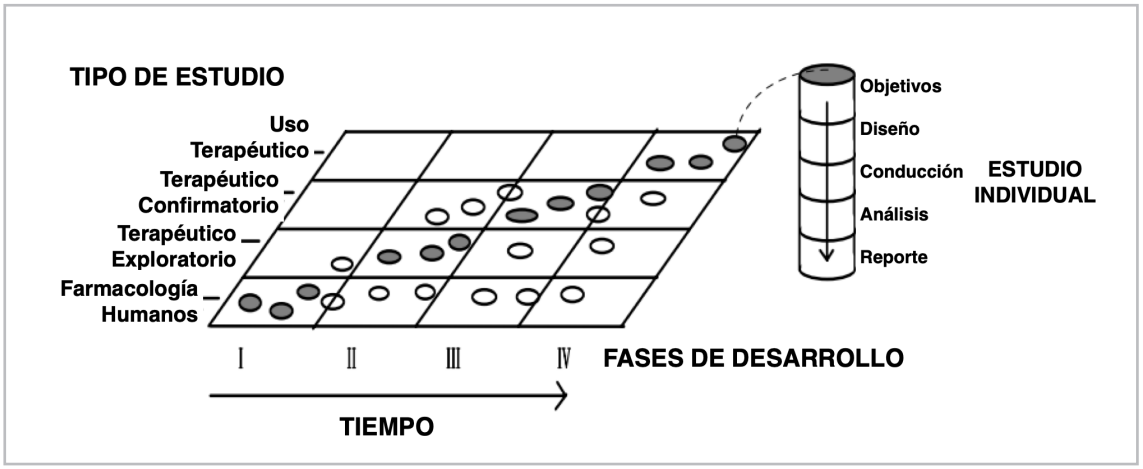

Figura 3. *Fases de los ensayos clínicos (Tomado de: International Harmonised Tripartite Guideline: General Considerations for Clinical Trials).*

fuera del ECA. También se puede utilizar un protocolo falso (*sham*) en intervenciones invasivas, simulando el uso en todos los pacientes para controlar el sesgo inducido por el propio paciente, con evaluación independiente del resultado para evitar el del investigador. De igual forma, puede utilizarse doble simulación (*double dummy*) en intervenciones con diferente régimen y posología (cada 12 horas vs cada 24 horas o administración parenteral vs oral), en el cual se utiliza terapia más placebo para reemplazar o simular la administración de una de las intervenciones, haciéndolas indistinguibles.

Usualmente, los ECA aleatorizan a los sujetos de investigación a uno de dos grupos de intervención; sin embargo, es posible diseñarlo con múltiples brazos, en los cuales se combinan diversos elementos que responden la pregunta de investigación. Estos pueden incluir la comparación de varias intervenciones, la combinación de tratamientos activos, diferentes dosis de una misma intervención, un placebo, una intervención no activa o el tratamiento usual o estándar. Las opciones de análisis son variadas, puesto que se pueden comparar, por ejemplo, tratamiento A_1 vs A_2 vs A_3, si son diferentes dosis de la misma intervención o A vs B vs C si son diferentes moléculas o si uno de los grupos es placebo. El uso de varios brazos aumenta la posibilidad de encontrar eficacia en uno de ellos, mejora el enrolamiento y es menos costoso.

Dentro de esta línea de diseño de ECA están los ensayos factoriales 2X2, en los cuales se pueden comparar dos tratamientos diferentes, no relacionados (sin interacción), al aleatorizar los pacientes dos veces, uno al tratamiento A y el control y otro al tratamiento B y el control. Dos ECA por el precio de uno. Otro diseño socorrido es el de grupos cruzados, en el cual los sujetos incluidos reciben ambas intervenciones, pero el orden en que los reciben difiere, de acuerdo con el grupo al cual fue asignado en forma aleatoria con un periodo de "lavado" antes de cambiar la intervención. Diseños menos utilizados son por grupos o tipo *cluster*, N de 1, adaptativos y los conocidos como pragmáticos. Los más frecuentes son los de grupos cruzados (61.3%) y los de grupos paralelos (24%) (16).

De acuerdo con el propósito del estudio, los ECA se pueden clasificar en: superioridad, no inferioridad o equivalencia. Tradicionalmente, los ECA se han desarrollado para demostrar que una terapia nueva es mejor que la terapia estándar o que el placebo (superioridad). En estos estudios la hipótesis nula (H_0) señala que la terapia nueva es igual (usualmente no es mejor) a la terapia estándar/placebo, mientras que la hipótesis alternativa (H_1), la que se intenta probar, es que la terapia nueva es diferente (usualmente es mejor) a la terapia estándar/placebo (H_0: $\mu_1=\mu_0$ vs. H_1: $\mu_1\neq\mu_0$; μ= media). Si los resultados alcanzan significancia estadística se rechaza la hipótesis nula y se acepta la alternativa (17).

Con mayor frecuencia vemos ECA con la intención de demostrar que un nuevo tratamiento no es inferior al estándar. El nuevo tratamiento posee alguna ventaja sobre la terapia clásica, por ejemplo, menos efectos secundarios, mayor seguridad, más fácil administración o posología, menos invasivo, menos costoso, entre otros, pero conserva un porcentaje importante de su efecto (18), por lo que son poco entendibles cuando se comparan con placebo, con el argumento de evaluar seguridad, estrategia utilizada en la evaluación de algunos antidiabéticos. La hipótesis nula es que la nueva terapia es inferior a la estándar y la alternativa es que es no inferior (H_0: $\mu_1-\mu_0\leq-\delta$ vs. H_1: $\mu_1-\mu_0>-\delta$; δ: delta, con $\delta\geq 0$).

Un aspecto importante y crucial es la determinación del conocido margen de no inferioridad o delta para el desenlace primario. Este margen representa la diferencia más pequeña aceptada entre los tratamientos para ser declarada como no inferior. Se espera que el nuevo tratamiento al menos conserve el 50% del efecto demostrado en ECA previo del grupo intervenido versus el placebo. El margen determina la posibilidad de declarar la intervención como no inferior y el tamaño de la muestra del estudio (inversamente proporcional a la raíz cuadrada del delta escogido). Por lo que debe ponderarse adecuadamente para no cometer errores (19,20) (Figura 4). La reevaluación de algunos de ellos o la comparación con cohortes posteriores ha llevado a determinar que muchos tenían fallas metodológicas importantes, especialmente en la elección del delta (21, 22).

Se puede utilizar una estrategia secuencial preestablecida en la cual, si se confirma no inferioridad de la intervención, se proceda a evaluar superioridad. Aunque el proceso contrario es posible, sus resultados son poco confiables, dado que la base experimental con la que se asumió probar superioridad de la nueva intervención queda en entredicho con lo observado y expresa una forma desesperada de "salvar" el ensayo.

Finalmente, los estudios de equivalencia intentan demostrar que la intervención es al menos similar, no exactamente igual, que la terapia estándar, por lo que también exige un margen pre-especificado de tolerancia. Puede considerarse como la intersección de dos ensayos de no inferioridad (H_0: $\mu_1-\mu_0\leq-\delta$ and $\mu_0-\mu_1\leq-\delta$ vs. H_1: $\mu_1-\mu_0>-\delta$ and $\mu_0-\mu_1>-\delta$) (19).

Características de un ensayo clínico

Un ensayo clínico con asignación aleatoria es la forma más rigurosa para establecer relación causa-efecto, con la menor incertidumbre posible, entre la eficacia de una intervención y un desenlace definido.

Tiene características importantes que lo definen, tales como: 1. Asignación aleatoria a la intervención; 2. El investigador y el sujeto de estudio desconocen cual tratamiento está recibiendo, es decir, están enmascarados (cegados, del inglés *blinding*); 3. Los grupos son tratados en igual forma, excepto por la intervención experimental; 4. Los sujetos de estudio son analizados en el mismo grupo al cual fueron asignados en forma aleatoria (análisis por intención a tratar); y 5. El análisis está enfocado sobre la estimación del tamaño de la diferencia en desenlaces predefinidos entre los grupos intervenidos (23).

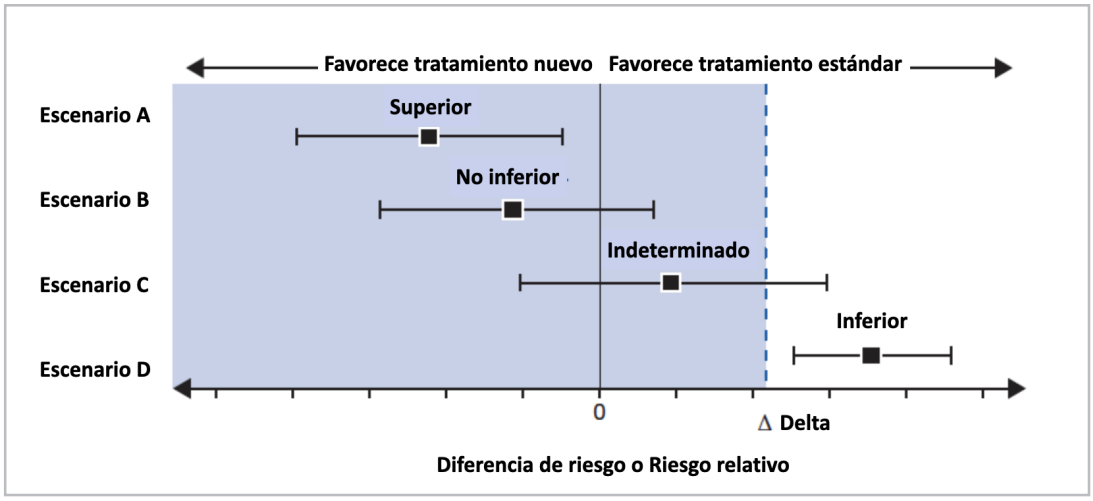

Figura 4. *Ensayos clínicos de no inferioridad.*

La aleatorización tiende a producir grupos de estudio comparables con respecto tanto a los factores de riesgo conocidos como los desconocidos, por lo que suprime el sesgo del investigador en la asignación de los sujetos participantes. Además, garantiza que las pruebas estadísticas tengan una tasa de falsos positivos válida.

Sesgos en los ensayos clínicos

El sesgo (*bias*) es definido como la tendencia sistemática a desviar un estimativo en una dirección diferente al verdadero valor. Esta desviación puede llevar a subestimar o a sobrestimar el verdadero efecto de una intervención. Los diversos tipos de sesgos son: de selección, de realización, de detección, de desgaste (*attrition*), de notificación y de reporte. Existen otros múltiples tipos de sesgos, algunos de los cuales están relacionados directamente con el diseño del estudio, otros no, como el sesgo accidental, en el cual la aleatorización no alcanza un balance adecuado entre los factores de riesgo y/o en las covariables pronósticas (8).

El sesgo de selección se produce cuando sujetos potencialmente elegibles son excluidos, causando diferencia sistemática entre los grupos de estudio. Todos los sujetos del estudio deben tener una probabilidad definida de ser asignados a un grupo especifico de intervención. Esta asignación no debe ser determinada por el investigador ni debe existir un patrón predecible de asignación. Para evitarlo se debe considerar la generación aleatoria de la secuencia de asignación y el ocultamiento de la asignación (24).

La generación de la secuencia se refiere al método utilizado para asignar en forma aleatoria los sujetos de estudio en cada grupo de intervención o tratamiento, logrando así balancear las características basales entre los grupos. El ocultamiento de la asignación se refiere al método utilizado para prevenir que cualquiera sea capaz de predecir la asignación del paciente a un grupo definido de tratamiento.

Existen diversos métodos de asignación aleatoria, clasificadas como fijas y adaptativas. Dentro de las fijas están la aleatorización simple, tal como lanzar una moneda al aire, puede producir desequilibrio de factores importantes; aleatorización por bloques permutados, en la cual se definen un número de bloques para intervención, el orden de las intervenciones en cada bloque es asignado en forma aleatoria y pueden variar para evitar ser predecibles; y estratificada, en la que se asegura un balance en factores clave determinados previamente, en general siempre se incluye el centro de estudio, pero pueden incluirse otras variables, como diabetes mellitus, edad o sexo. Se debe evitar la sobreestratificación por la generación de múltiples estratos (25). Las estrategias adaptativas son menos utilizadas, las más conocidas son la minimización y la de Urn.

El enmascaramiento o *blinding* es importante para manejar los sesgos de realización y detección. Al enmascarar al *participante,* hay menos probabilidad de respuesta sicológica a la intervención, mayor probabilidad de cumplimiento del régimen de estudio, disminuye la búsqueda de intervenciones adicionales o complementarias y se reduce la probabilidad de abandono del estudio sin datos del desenlace. Al enmascarar al *evaluador,* se reduce el sesgo en la evaluación del desenlace de interés; y al enmascarar al *investigador,* es menos probable que transmita sus inclinaciones o actitudes al participante hacia la intervención, menor probabilidad de administrar cointervenciones en forma diferencial, menor probabilidad de suspender la intervención en forma diferencial, de ajustar dosis en forma diferencial y de alentar o desalentar a los participantes para continuar el estudio (26).

En algunos casos es complejo el enmascaramiento, como por ejemplo en cirugías mayores; sin embargo, aun en estos escenarios es posible desarrollar protocolos falsos o simulados (*sham*), en los cuales parte de la intervención es realizada en ambos grupos (ejemplo, incisión quirúrgica) y la evaluación realizada por un investigador que desconoce quien fue sometido a la cirugía completa (26, 27). Están suficientemente discutidos y justificados los aspectos éticos de esta estrategia.

Análisis por intención a tratar

Existen diversas aproximaciones tradicionales para evaluar el efecto del tratamiento en los ECA. La estrategia recomendada es el análisis por intención a tratar (ITT), en la cual cada paciente es analizado en el grupo al cual fue asignado en forma aleatoria, independiente del tratamiento que recibe, ya sea el asignado en el estudio, el del grupo control (cruce) u otro disponible no considerado en la evaluación. Esta estrategia permite un estimativo no sesgado del efecto del tratamiento en toda la población incluida, particularmente es poco afectado por la no adherencia, el cruce entre grupos o potenciales confusores, al respetar el balance obtenido durante la asignación aleatoria (28). No obstante, puede subestimar el efecto real del tratamiento esperado en los sujetos de estudio incluidos que fueron adherentes, en otras palabras, al no utilizarse puede sobreestimarse el efecto, incluso con las estrategias nominadas como ITT modificada (29).

Es importante anotar que, en general, la no adherencia y el cruce en las intervenciones no son fenómenos que se producen en forma aleatoria, tienen la capacidad o no de impactar el desenlace. El análisis por ITT se considera conservador en estudios de superioridad, pero puede ser más liberal en los estudios de no inferioridad y equivalencia, con tendencia a sesgar los resultados, haciendo que los dos tratamientos luzcan similares. Por lo tanto, se recomienda en estos casos análisis por protocolo (30), especialmente cuando los investigadores anticipan no adherencia en un grupo importante de pacientes (>5%) (31, 32).

El análisis alternativo es por protocolo (PP), en el cual se incluyen todos los pacientes que cumplieron con el tratamiento asignado; excluye los pacientes que violaron el protocolo, ya sea porque cruzaron el grupo de intervención o porque nunca lo tomaron. Al no respetar la aleatorización está sujeto a sesgo de selección, a través del análisis de grupos con diferencias en variables pronósticas, asemejándose a un análisis de subgrupos.

Otra alternativa es el análisis por tratamiento o como fue tratado, denominado "*as treated*" en inglés, en el cual se analizan los pacientes en el grupo del tratamiento que reciben, independientemente de a cuál fueron asignados. Este enfoque desvía el efecto en cualquier dirección, puesto que puede sobreestimar el efecto si el desequilibrio es por la creación de un grupo con buen pronóstico en la terapia con mejor efecto o, por el contrario, subestimarlo si el grupo restante es de pobre pronóstico (29).

Validez interna versus validez externa

El ensayo clínico es el diseño metodológico más apropiado para responder una pregunta de investigación en cuanto a la eficacia de una intervención. Es importante tener en cuenta la población objetivo a la cual está dirigida, para determinar con relativa precisión la indicación resultante después de analizar el estudio, la cual será utilizada para establecer políticas de salud pública en el escenario específico en un espectro más amplio (población general). Definir la población objetivo depende de unos criterios de selección, inclusión y exclusión, establecidos, los cuales justifican la pregunta de estudio. Entre más estrictos estos criterios (validez interna), más difícil será, posteriormente, extrapolar los resultados para su uso en la práctica clínica (validez externa). La validez interna expresa la concordancia entre el efecto medido con el verdadero efecto de la población incluida (muestra) en el estudio. Este aspecto define con claridad la representatividad de la muestra seleccionada (33).

No es infrecuente encontrar diferencias marcadas entre la población incluida en los ECA y las seguidas en registros en estudios de cohorte, en el nominado "mundo real"; además, suele observarse limitada representación de grupos específicos, como algunos grupos étnicos, mujeres y/o pacientes de edad avanzada (34), lo que limita su validez externa (Figura 5).

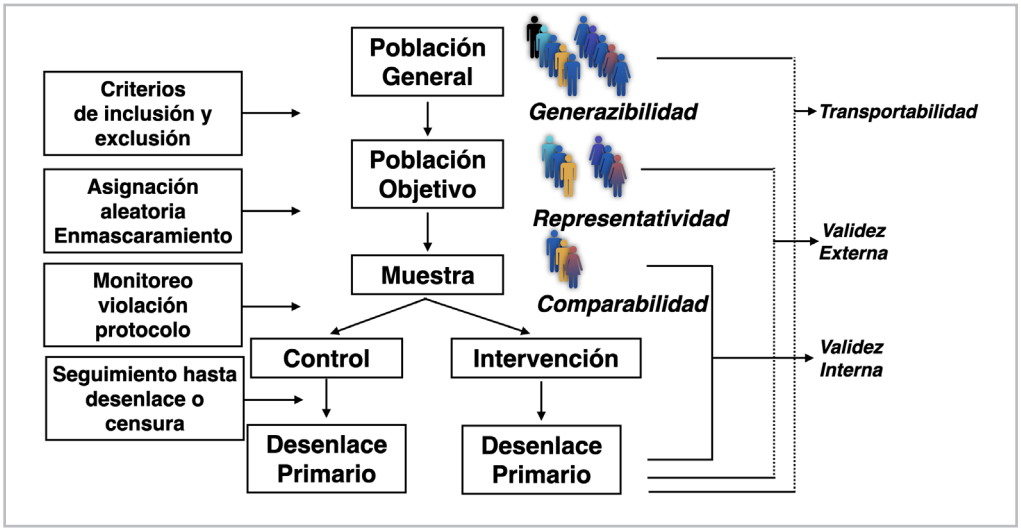

Figura 5. *Representatividad de la población en los ECA.*

Este aspecto no es irrelevante, puesto que algunas características de la población pueden modificar la respuesta a ciertas intervenciones. Tanto factores intrínsecos, como la edad, el género, el peso, el grupo étnico, la carga genética, la biología de la propia enfermedad y las comorbilidades, como factores extrínsecos como el uso de tabaco, alcohol y la dieta, entre otros, pueden modificar la farmacodinamia y farmacocinética. Estos factores producen respuestas diferenciales e individuales que deben explorarse (35).

La razón de los ECA es la de realizar inferencia causal no solo sobre la muestra incluida, sino en una población más amplia, que permita su aplicación clínica. Esto permite extender los resultados obtenidos en la población estudiada a la población objetivo, puesto que generalmente es un subconjunto de esta última (generazibilidad). Además, se puede extender a una población externa, con características diferentes, por ejemplo, geográficas, permitiendo su transportabilidad (36, 37).

Significancia estadística

La significancia estadística se refiere al simple hecho de demostrar que el resultado obtenido dentro de un experimento clínico pueda ser atribuido con alta probabilidad a una causa específica, en este caso como resultado de una intervención médica o quirúrgica. Ha sido erróneamente interpretado e identificado como resultante de una prueba estadística asociada a un valor de P (38).

El valor de P es un objetivo aparentemente claro y fácil, por lo que es ampliamente utilizado, como también ampliamente abusado. No hemos agotado aún la argumentación al mencionar tantos ceros antes del 1 (0.0000X1) para resaltar el peso del resultado de un ECA en las discusiones académicas. Nada más lejos de su real interpretación. El punto de corte $P<0.05$ es interpretado en forma errada como la probabilidad del 5% que el resultado sea explicado por el azar, o, en otros términos, hay 95% de probabilidad de que el resultado sea verdadero.

Entonces surge la pregunta: ¿Qué significa el valor de P? El valor de P es simplemente la probabilidad de que, bajo un modelo estadístico específico el resumen de los datos, como la diferencia del promedio entre dos grupos comparados sea igual o más extremo que los observados (39). Indica la incompatibilidad de los datos con el modelo estadístico especificado. Bajo un modelo construido con ciertos supuestos, un valor de P pequeño indica la incompatibilidad de los datos obtenidos con la hipótesis nula, que siempre plantea la ausencia de efecto o la ausencia de relación entre una variable o factor y un desenlace (por ejemplo, H_0: $\mu_1=\mu_0$; μ= media), con lo que se podría rechazar, para aceptar la hipótesis alternativa (H_1: $\mu_1\neq\mu_0$).

La determinación de la significancia estadística debe empezar por el diseño experimental planteado, si hay concordancia con la pregunta de investigación para responderla, la sustentación científica de la hipótesis a probar, lo adecuado de la conducción del estudio, la obtención de los datos, el monitoreo de estos, los análisis exploratorios realizados bajo el modelo estadístico definido, para finalmente llegar al valor de P (la punta del iceberg). Es frecuente observar el conocido fenómeno de dragado de datos, inferencia selectiva o *hackeo*, que expresa la actitud de los investigadores por encontrar asociaciones sin el debido soporte de evidencia científica que sustente esa hipótesis, por no haber sido planteada previamente en el diseño.

Otro fenómeno que logra impresionar incautos es el de la compra del valor de P. Cualquier efecto o asociación, por débil que sea, puede producir valores muy pequeños de P, simple y llanamente por tamaños muestrales grandes o utilizar una medida de precisión alta. Lo contrario también puede ser cierto, efectos importantes pueden producir valores de P que no impresionan por tamaños muestrales pequeños o el uso de medidas imprecisas. Un valor de P pequeño nos dice que los datos obtenidos son inusuales bajo los supuestos del modelo probado, pero puede ser simplemente el resultado de una hipótesis falsa o porque se violaron los protocolos del estudio (40).

Tener siempre presente que significancia estadística no representa significancia científica o clínica. Por ejemplo, la reducción de 1% del nivel de cualquier variable (glicemia, colesterol total, cifras de presión arterial) puede arrojar significancia estadística con $P<0.05$; sin embargo, le corresponde al clínico interpretarlo a la luz del conocimiento para decidir si el efecto real es relevante, sobre un desenlace subrogado o sobre un desenlace de importancia. Por tanto, el valor de P no mide la importancia del resultado ni el tamaño del efecto (40).

El nivel de significancia estadística también representa la probabilidad del conocido error tipo I, denotado por α. Este error ocurre cuando H_0 es rechazada, pero es cierta. La potencia estadística está relacionada con el conocido error tipo II, como el complemento de 1-β; así, la potencia estadística de una hipótesis representa la habilidad para detectar un tamaño de efecto especificado a un nivel de significancia α, o sea, rechazar H_0 cuando H_A es cierta (Tabla 1). Idealmente, un ECA debería tener una potencia suficiente para aceptar correctamente la H_A cuando es cierta. La mayoría de los ECA escogen una potencia del 80%.

Análisis e interpretación de los resultados

La fase final, no menos importante, es la interpretación, análisis y reporte de los resultados. La presentación debe ser lo más clara y concisa posible para lograr una comunicación asertiva, que le permita al lector la adecuada interpretación de los datos, ponderar los resultados y llegar a conclusiones válidas, de acuerdo con lo demostrado y publicado. Es responsabilidad de los investigadores hacer análisis crítico de los resultados, sin caer en la tentación de redactarlo de tal forma que minimice los riesgos encontrados y luzca más robusto de lo que en realidad es.

Se debe evitar la no infrecuente práctica de girar ("*spin*") los resultados para hacerlos benevolentes en caso de haber

Tabla 1. *Probabilidades asociadas con la prueba de hipótesis.*

Decisión	Situación actual	
	H_0 correcta	H_0 falsa (H_A correcta)
No rechazar H_0	Decisión correcta (1-α)	Error tipo II (β)
Rechazar H_0	Error tipo I (α)	Decisión correcta (1-β= potencia)

H_0= hipótesis nula; H_A=hipótesis alternativa

sido negativos o convertirlo en un estudio positivo al resaltar desenlaces secundarios, subgrupos o análisis pos hoc, sin la debida explicación que son análisis exploratorios que generan una nueva hipótesis (41).

El hecho de estar publicado en una revista de alto impacto, ni el que haya sido desarrollado por investigadores reconocidos en el medio científico, es un argumento que soporte los resultados y la interpretación dada. Las revistas científicas tienen una gran responsabilidad entre manos al publicar un manuscrito; sin embargo, cada año son sometidos millones para considerar su publicación, en miles de revistas, lo que hace particularmente difícil la elección de los mejores. Adicionalmente, la evaluación por pares no está exenta de problemas como la pobre preparación de quienes realizan ese trabajo, muchos sin retribución alguna por parte de las editoriales, conflictos de interés y hasta fraude, sin dejar de lado el sesgo hacia la publicación de estudios positivos, especialmente los patrocinados por la industria (42).

La sofisticación en el desarrollo y la conducción de los ECA ha permitido que la interpretación de los resultados pueda ser manipulada para impresionar al lector desprevenido, especialmente con el uso de técnicas de análisis estadístico más complejas. Ni siquiera la exigencia de la publicación previa del protocolo, en diversos registros o bases de datos, ha logrado trasparencia en este aspecto, encontrando falencias en ellos y discrepancias claras entre lo registrado como protocolo y lo publicado, que lamentablemente incluye el cambio de desenlace primario evaluado, en algunas ocasiones (43, 44, 45).

A pesar de estar claras las recomendaciones, el reporte de ECA en revistas científicas tiende a estar sesgado hacia la exageración de la diferencia de las intervenciones en los resultados. Dentro de los problemas estadísticos más frecuentes están el uso de múltiples desenlaces primarios, uso de desenlaces primarios no relacionados, análisis de subgrupos numerosos o no pre-especificados, uso de medidas repetidas en el tiempo sin estrategia predefinida, uso de más de dos tratamientos sin análisis establecido, uso de numerosas pruebas de significancia estadística, falta de cálculo del tamaño de muestra o modificación de esta sin justificación clara, falta de reglas claras para detener el estudio (detención temprana sobreestima el efecto), puntos de corte diferenciales en las pruebas estadísticas y la selección tendenciosa de los resultados en el resumen (*abstract*), privilegiando incluso desenlaces secundarios, aún con primario negativo (46, 47).

Un aspecto crucial de los ECA está en la escogencia del desenlace adecuado, aquel que tenga la capacidad de captar la eficacia del tratamiento, ya sea clínicamente relevante o subrogado. Ocasionalmente los ECA fallan al "apostarle" al desenlace equivocado, por ejemplo desenlace combinado o mortalidad en lugar de rehospitalización. Se considera desenlace subrogado aquel escogido para medir como substituto de otra variable, especialmente porque puede reducir el tamaño de la muestra o la duración del estudio, al reemplazar uno de rara ocurrencia o con mayor tiempo de presentación, por uno más frecuente o de más rápida presentación (48, 49).

Luego de la obtención de un resultado positivo con significancia estadística, es importante determinar otros aspectos que muestren la robustez del resultado. Desafortunadamente, es frecuente quedarse con el resultado binario de es o no estadísticamente significativo (P<0.05), sin recabar en otros aspectos de suma importancia.

Por lo anterior, se sugiere realizar algunas preguntas:

- **¿Cuál es la magnitud del efecto?**
  La diferencia encontrada entre las intervenciones debe ser clínicamente relevante, lo suficientemente grande para ser considerada importante. Es importante revisar el estimativo utilizado, riesgo relativo o riesgo instantáneo (*hazard ratio*), su intervalo de confianza (IC 95%). También es relevante determinar la diferencia en la tasa de eventos en el seguimiento y el número necesario a tratar (NNT= inverso de la diferencia del riesgo absoluto) (50).

- **¿Es el desenlace primario clínicamente importante?**
  En general, los ECA fase III utilizan desenlaces clínicos como mortalidad, aunque pueden usar algunos subrogados, que generan controversia. El otro punto importante es el uso de desenlaces combinados, los cuales deben al menos tener alguna relación, fisiopatológica o de otro tipo, presentarse con frecuencia parecida y tener impacto similar con la intervención; aunque es muy frecuente que el resultado demuestre mayor efecto sobre uno de ellos, en forma individual, lo que dificulta su interpretación. Tienen la ventaja de reducir el tamaño de muestra requerido, reducen el tiempo de seguimiento y costos, y pueden incluir el beneficio clínico neto, al incorporar eventos adversos (51,52). Es posible utilizar desenlaces coprimarios, no necesariamente combinados, como también establecer un análisis jerárquico, en el que se analizan en forma secuencial los desenlaces definidos con una jerarquía preestablecida hasta perder significancia estadística, con lo que los siguientes se consideran exploratorios.

- **¿Son los resultados consistentes?**
  Al utilizar desenlaces combinados es importante que el impacto sea similar y en la misma dirección de cada uno

de ellos. También analizar el efecto sobre desenlaces secundarios, lo cual le da mayor peso a los resultados (50). En el análisis de subgrupos es importante tener en cuenta ciertos aspectos para no interpretarlo en forma inadecuada, dándole un peso que en realidad no poseen. La primera mirada debe ser para la gráfica que muestra los resultados discriminados por subgrupos, usualmente al final del reporte del estudio, en la cual los estimativos deben ubicarse al mismo lado del obtenido en la muestra total (todos a la derecha o todos a la izquierda), así el IC 95% supere el 1, puede ir acompañado por un P de interacción, ya sea aditiva o multiplicativa (Figura 6) (53).
Luego debemos establecer si la magnitud de la diferencia fue clínicamente importante, si alcanzó significancia estadística, si la hipótesis precede al análisis dada la sustentación teórica para explorar ese subgrupo específico, la diferencia fue sugerida por la evidencia en otros estudios y consistente entre ellos, si el número de subgrupos elegidos fue uno de un pequeño grupo de hipótesis para probar en el estudio (54). En general deben considerarse generadores de hipótesis (55).
Otros aspectos igualmente importantes son evaluar si el tamaño de la muestra es suficiente para ser convincente, dado que se debe ser cauteloso con los ECA pequeños con resultados abrumadores, como también de estimativos sorprendentemente bajos (de eso tan bueno no dan tanto; aunque la muestra sea grande), máxime si no son consistentes con resultados de otros estudios. También si hubo detención temprana del estudio (por eficacia, futilidad, eventos adversos o problemas logísticos), si hay adecuado balance entre eficacia y seguridad y, por último, si se detecta inadecuada conducción del

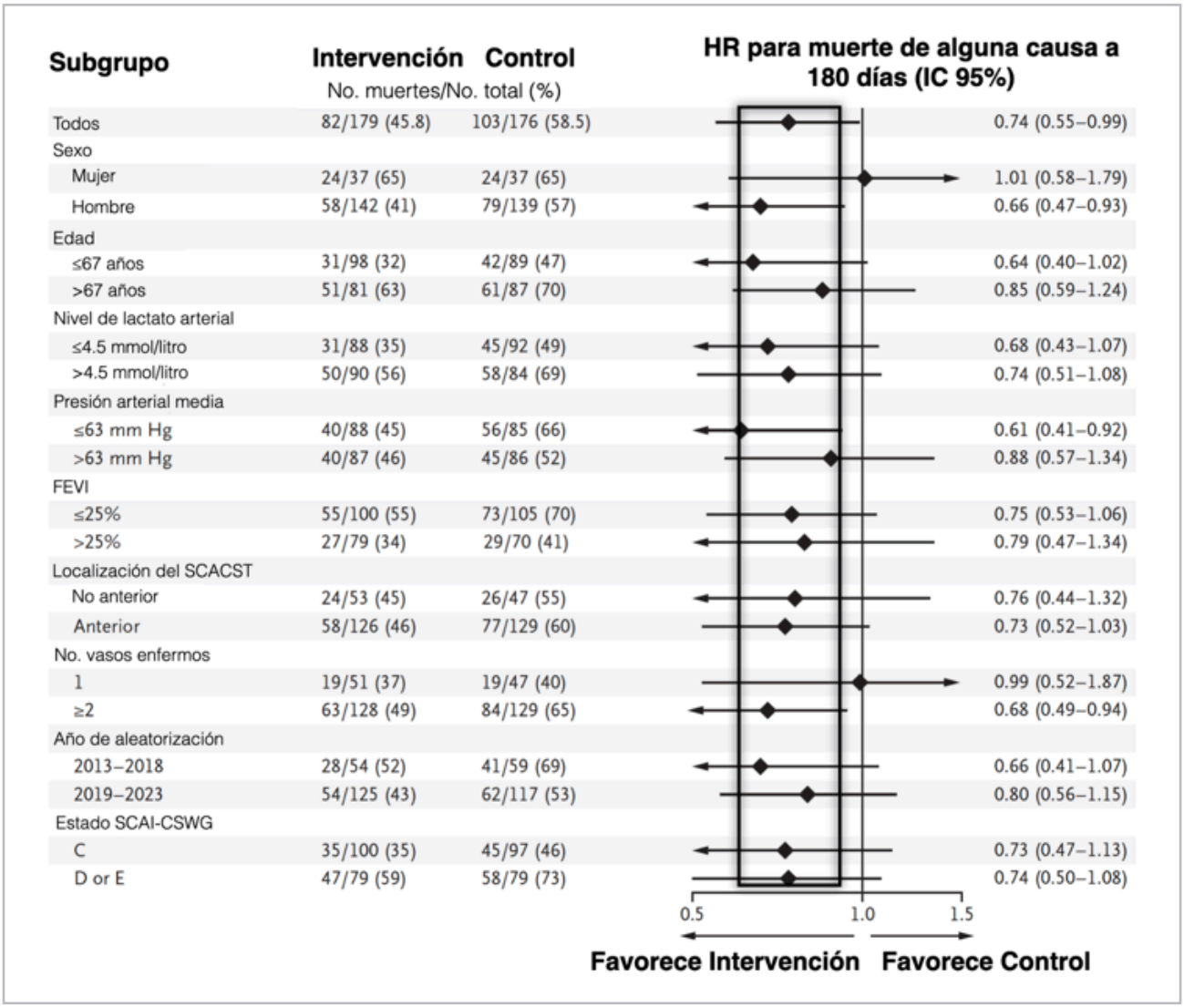

Figura 6. *Resultados por subgrupos en ECA. (FEVI: fracción de eyección del ventrículo izquierdo; SCACST: síndrome coronario agudo con elevación del segmento ST; SCAI-CSWG: Society for Cardiovascular Angiography and Intervention- Cardiogenic Shock Working Group). Fuente: ejemplo tomado de ECA publicado en New England Journal of Medicine*

ECA, recalculo del tamaño de la muestra o del tiempo estimado de enrolamiento sin justificación, calidad del dato, % de pérdidas elevado o de datos faltantes que no pueda superarse con imputación múltiple de datos, falta de adherencia, entre otros.

Existen intervenciones que se han ido posicionando por su perfil innovador o mecanismo atractivo, pero que no han sido probados adecuadamente en ECA, a pesar de lo cual se convirtieron en parte de la terapia estándar. Al realizar un ECA para evaluar esas intervenciones nos sorprendemos si los resultados son negativos, con particular reticencia a aceptarlos e incorporarlos en la práctica clínica diaria. Desaprender es un proceso difícil y complejo.

- **¿Qué hacer cuando un ECA es negativo?**
  Se deben explorar varios aspectos críticos como si al menos se vislumbra un beneficio potencial. Es fundamental evaluar si tuvo baja potencia; si se escogió el desenlace, la población y el régimen terapéutico apropiado; si hubo deficiencias en la conducción de estudio. También es necesario conocer si hubo planteamiento de análisis de no inferioridad, si se tuvo alguna señal positiva en subgrupos o desenlaces secundarios que puedan generar hipótesis; si se pueden explorar análisis alternativos (ajuste por covariables, análisis por protocolo o como fue tratado, análisis de eventos repetidos o de riesgo competitivo) (56, 57).

Estimación del efecto de las intervenciones (parte II)

El objetivo del análisis estadístico de un ECA es estimar la magnitud de la diferencia de las intervenciones sobre los desenlaces, por ejemplo, la diferencia de medias entre dos grupos en los desenlaces escogidos. Primero se determina un estimativo puntual, que corresponde a la diferencia observada, luego se debe determinar el grado de incertidumbre de los datos, usualmente con el uso de los intervalos de confianza del 95% (IC 95%) (58).

El tipo de estimativo utilizado depende de la naturaleza del desenlace de interés. Básicamente son tres: 1. Binario (si o no, vivo o muerto); 2. Tiempo al evento (supervivencia); y 3. Desenlace cuantitativo (ejemplo, cifras de presión arterial) (58).

En enfermedades agudas con tiempos de seguimiento generalmente cortos, se establece la comparación de dos intervenciones en términos binarios como "ausencia" o "presencia" de un evento clínico relevante. La característica del evento clínico depende del escenario y la patología estudiada, puede ir desde mortalidad (vivo o muerto) hasta complicaciones mayores y menores. De esta forma cobra relevancia la comparación de los grupos al finalizar el periodo definido de seguimiento y pierde interés la forma en que se desarrolló el evento determinado, durante el tiempo de observación (58). En estos casos se puede cuantificar el efecto por medidas como la reducción del riesgo absoluto, la reducción del riesgo relativo o por el número necesario a tratar (59).

El riesgo se refiere a la probabilidad de ocurrencia de un evento o un desenlace, en términos estadísticos es la probabilidad del desenlace de interés sobre todos los posibles desenlaces. *Odds* o "suertes" o "momios" se refiere a la probabilidad de ocurrencia de un evento sobre la probabilidad de que no ocurra. Aunque es un término más difícil de entender y es utilizado con frecuencia en el mundo de las apuestas, resulta ser útil cuando se usan modelos de regresión logística para ajustar por variables, por su mayor versatilidad matemática y conversión a Log Odds. Comúnmente, se confunden *odds* y riesgo utilizándolos en forma intercambiable; sin embargo, son conceptos diferentes, aunque son muy similares cuando la tasa de eventos es <10% (Tabla 2) (60).

Al conocer la probabilidad de un evento se puede despejar el *odds*. Por ejemplo, si la probabilidad es 0.2 (segundo escenario en Tabla 2), el *odds* de ocurrencia del evento

Tabla 2. *Odds y riesgo con diferentes tasas de eventos.*

Intervención	Desenlace				
	Muerte (a)	Supervivencia (b)	Total (a+b)	Riesgo [a/(a+b)]	*Odds* (a/b)
Primero	30	70	100	30/100=0.3	30/70=0.43
Segundo	20	80	100	20/100=0.2	20/80=0.25
Tercero	10	90	100	10/100=0.1	10/90=0.1
Cuarto	1	99	100	1/100=0.01	1/99=0.01
Nota: se plantean 4 escenarios con tasa de eventos diferentes; se utiliza N=100 para facilitar el cálculo.					

sería 0.2/0.8=0.25 o la probabilidad dividida por 1 menos la probabilidad 0.2/1-0.2=0.25 (ecuación 1). De esto se deduce que cuando la probabilidad es pequeña, el odds es casi idéntico a esa probabilidad (ejemplo, probabilidad de 0.05). El log *odds* de que ocurra el evento sería: Ln [0.2/0-8]=-1.38 (logaritmo de un cociente es igual al logaritmo del numerador menos el logaritmo del denominador) o Ln (0.25) =-1.38. La probabilidad puede ser calculada como *odds*/1+*odds* (ecuación 2), o sea, 0.25/1.25=0.2 o exp [Ln (odds)]/1 + exp [Ln (odds)]= exp (-1.38) /1 + exp (-1.38) =0.25/1.25. ¡Fácil! (61).

El riesgo relativo (RR) se refiere a la relación entre la tasa de eventos en el grupo de intervención con la tasa de eventos en el grupo control (expuestos/no expuestos). El *odds ratio* (*OR*) o razón de momios es la relación de *odds* del grupo intervenido con el *odds* del grupo control. Tomando los datos de la Tabla 2 como ejemplo, el primer escenario sería el grupo control y el segundo el grupo intervenido, el riesgo relativo sería 0.2/0.3= 0.66 y el odds ratio 0.25/0.43=0.58. Los intervalos de confianza del 95% pueden calcularse por técnica matemática (62). Es posible calcular el RR a partir del OR con la formula RR= OR/ [1-P_0 + (P_0 x OR)], donde P_0 es el riesgo basal de la población estudiada (en el ejemplo expuesto sería P_0=0.25) (63).

Se puede describir la diferencia en los desenlaces entre grupos en términos absolutos o en relativos. En estos casos se utilizan los términos de reducción del riesgo absoluto, que es simplemente la diferencia entre la tasa de eventos entre dos grupos; o la reducción del riesgo relativo que es la diferencia entre la tasa de eventos entre dos grupos expresada como la proporción de la tasa de eventos del grupo control o no tratado, usualmente constante a través de diferente riesgo basal de la población (Tabla 3).

La reducción del riesgo absoluto debe ser interpretado a la luz del riesgo basal, es más pequeño con tasas de eventos más bajas mientras que la reducción del riesgo relativo permanece relativamente constante. Entre más baja la tasa de eventos en el grupo control, mayor la diferencia entre la reducción del riesgo relativo y el riesgo absoluto. Se recomienda reportar ambas medidas y corresponde al clínico definir la relevancia del resultado.

El número necesario a tratar (NNT) o a dañar (NNH) expresa el número de pacientes que deben recibir la intervención para evitar un desenlace; se calcula como el inverso de la reducción del riesgo absoluto x100 (64). En el primer ejemplo expuesto en la Tabla 3, se necesitaría tratar 20 pacientes para evitar una muerte con la intervención evaluada. Se debe tener en cuenta que para desenlaces binarios el NNT depende del tiempo de seguimiento; esto significa que no hay un solo valor del NNT, sino varios que pueden ser calculados en un punto específico luego de iniciar el tratamiento (65). En un ECA, si se dispone del dato del *hazard ratio* (HR), se podría calcular el NNT en un punto específico al conocer la probabilidad de supervivencia del grupo control con la fórmula:

NNT = 1/ {[Sc(t)]^{HR} - Sc(t)} (ecuación 3); donde Sc(t) es la probabilidad de supervivencia del control en el tiempo definido, por lo que la probabilidad de supervivencia del grupo activo sería [Sc(t)]^{HR}. Por ejemplo, si la probabilidad de supervivencia a 2 años es de 0.33 y el HR reportado es 0.72 (IC 95% 0.55-0.92), el NNT a 2 años sería = 1/0.33$^{0.72}$ - 0.33 = 8.32 (ejemplo tomado de la referencia 65).

Tabla 3. *Relación entre reducción de riesgo absoluto y riesgo relativo de acuerdo con el riesgo basal.*

Riesgo de desenlace Mortalidad		RRA	RRR	NNT
Intervención	Control			
10%	5%	5%	5/10 = 50%	20
80%	40%	40%	40/80 = 50%	2.5
2%	1%	1%	1/2 = 50%	100

Tiempo al evento (supervivencia)

En enfermedades crónicas, como neoplasias o cardiovasculares, cobra importancia el tiempo entre la exposición, en este caso la intervención, y el evento de interés definido, usualmente mortalidad. Mortalidad y supervivencia no son términos intercambiables, puesto que la primera es dicotómica, utilizada para comparar dos grupos en un tiempo específico (30 días, un año, cinco años), sin importar el intervalo transcurrido, mientras que la segunda le da importancia al tiempo al evento como variable fundamental. Finalmente, todos tendrán el desenlace, dicho más crudamente, todos los sujetos de estudio morirán, solo depende del tiempo definido para medirlo; por lo tanto, el tiempo hasta el evento cobra relevancia. Aunque se planteó como análisis de supervivencia el desenlace no solo es mortalidad, puede incluir otros desenlaces de interés como recaídas, rehospitalizaciones, progresión, entre muchas otras.

A pesar de la creencia errónea de que todos los pacientes en los ECA son tratados en forma simultánea, los sujetos de estudio van ingresando a medida que se desarrolle el tiempo de reclutamiento, el cual puede incluso ser mayor que el periodo de seguimiento, por lo que no es inusual que algunos pacientes salgan del ensayo antes del ingreso de otros, o que lleguen al final del estudio sin haber presentado el evento de interés (8).

En análisis de supervivencia, se refiere a la variable tiempo como "*tiempo de supervivencia*" y a la presentación del evento como "*falla*", precisamente por su connotación negativa, aunque algunas podrían ser positivas. El tiempo de supervivencia es una variable positiva con sesgo a la derecha, por lo que es imposible utilizar una distribución normal como modelo. Lo ideal es que todos los sujetos incluidos en el estudio puedan ser seguidos hasta el evento para establecer comparaciones. Sin embargo, puede suceder que se llegue al final del estudio, porque se alcanzó el número de desenlaces necesarios (hablamos de tamaño de muestra, pero lo que se calcula es el número de eventos necesarios

para obtener la potencia definida), sin haber presentado el evento, por lo que en realidad desconocemos su tiempo de supervivencia completo (66).

Este fenómeno se conoce como censura (censura administrativa). Existen otras razones para censurar, como la pérdida de seguimiento o retiro del estudio por alguna razón o por presentar un evento diferente al de interés, lo que se conoce como riesgo competitivo; este tipo de censura se conoce como censura a la derecha. La censura a la izquierda es poco común en ensayos y se produce por presentar la *falla* antes de ingresar al estudio (67) (Figura 7).

Distribución de supervivencia

La distribución de supervivencia es generalmente descrita en términos de dos funciones: la función de supervivencia y la función *hazard*. La función de supervivencia S(t) representa la probabilidad que una persona sobrepase un tiempo específico t, o sea, la función S(t) da la probabilidad que la variable aleatoria T supere el tiempo especificado t. Representada por la ecuación

$S(t)=P[T>t]=1- F(t)= \int_t^{\infty} f(x)\, dx$ (ecuación 4).

Es una función monótona y decreciente. Teóricamente como el rango del tiempo t es de 0 al infinito se podría graficar como una curva de pendiente suave (Figura 8).

La función *hazard* h(t) representa el potencial instantáneo (riesgo instantáneo) por unidad de tiempo para que ocurra el evento, dado que el sujeto ha sobrevivido hasta el tiempo t. Está representada por la ecuación 5:

$$h(t)= \lim_{\triangle t \to 0} \frac{P(t \leq T < t+\triangle t \mid T \geq t)}{\triangle t}$$

En términos matemáticos la función *hazard* es una probabilidad condicional, similar al clásico enunciado de probabilidad de A, dado B [P (A|B)]. La probabilidad que un sujeto de estudio supere el tiempo T, el cual estará entre el intervalo de tiempo t y Δ t, dado que el tiempo de supervivencia T es mayor o igual a t (numerador ecuación 5) ¡no tan fácil!; sin embargo, el denominador introduce el concepto de tiempo, en este caso Δ t, lo que la convierte en tasa y por lo tanto su rango no va de 0 a 1, como una probabilidad, sino de 0 a ∞, definido en unidad de tiempo como años o meses (66).

La función *hazard* puede graficarse en forma similar a la función de supervivencia, pero a diferencia de esta última no empieza en 1 y tiende a 0, inicia en cualquier parte ($h(t) \geq 0$) y aumenta en cualquier dirección, hacia arriba o hacia abajo, en el tiempo (no tiene límite superior). De las dos funciones, S(t) y h(t), la función de supervivencia es la más utilizada. La función h(t) es importante porque mide el potencial instantáneo mientras que la función S(t) es una medida acumulativa en el tiempo; además, la primera identifica un modelo específico (exponencial, Weibull o lognormal) y es la forma matemática de modelar la curva de supervivencia (66, 67, 68) (Figura 9).

Estimador de Kaplan Meier

En oncología la estimación de la supervivencia del paciente es aceptado como el principal criterio para evaluar la eficacia de un tratamiento. Se reporta como supervivencia a 1, 3, 5 o 10 o más años. Para lograrlo se debe seguir en forma individual cada paciente por el tiempo determinado, utilizando las conocidas tablas de supervivencia, aunque no es infrecuente que se desconozca la evolución de todos los pacientes de una cohorte, puesto que son seguidos por menor tiempo o porque se pierden del seguimiento, lo que

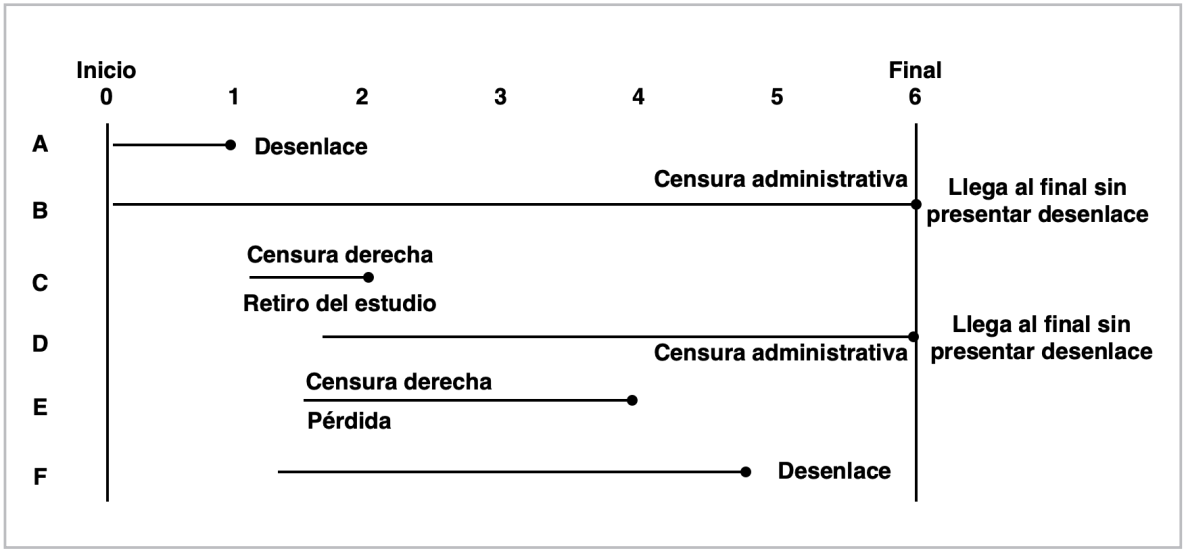

Figura 7. *Tipos de censura en análisis de supervivencia.*

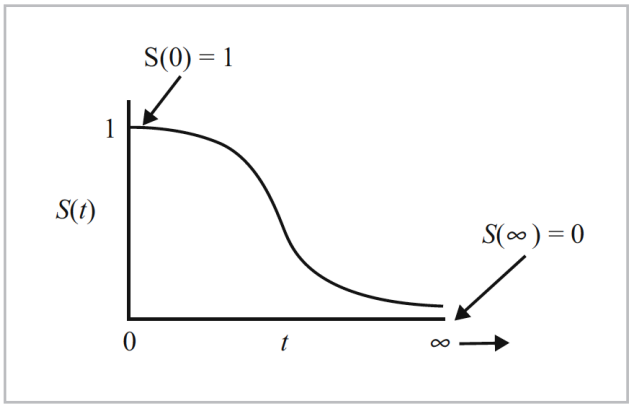

Figura 8. *Gráfica teórica de la función de supervivencia. Fuente: referencia 66*

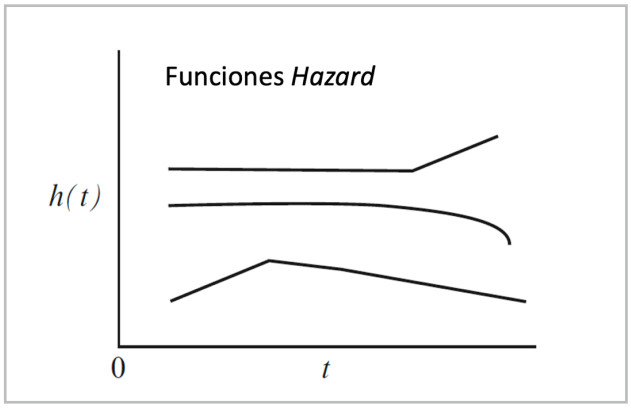

Figura 9. *Gráficas de funciones hazard. Fuente: referencia 66*

genera observaciones con datos incompletos. En 1958 Kaplan y Meier publican su clásico artículo sobre la estimación de tasas de supervivencia con datos incompletos o censurados, el cual se convirtió en el análisis estándar (69), cada sujeto de investigación aporta al análisis un tiempo de seguimiento (hasta presentar el desenlace, perderse del seguimiento o presentar un evento diferente al de interés).

Retomamos el concepto de censura para definir *censura aleatoria* como aquella en la que los sujetos de estudio que son censurados en tiempo t, deben ser representativos de todos los sujetos que permanecen en riesgo al mismo momento con respecto a su experiencia de supervivencia. La *censura independiente* expresa el mismo concepto, pero dentro de cualquier subgrupo de interés. La *censura no informativa* determina que el mecanismo de censura no puede estar relacionado con la distribución del tiempo al evento; por ejemplo, se considera censura informativa cuando un sujeto abandona el estudio por una razón relacionada con el estudio, valga decir evento adverso, lo que viola los supuestos en que se basa el análisis de supervivencia (66, 70).

Aclarados estos conceptos, pasamos a la comparación de dos grupos de sujetos de estudio. Inicialmente tabulamos la experiencia de cada uno de ellos (desenlace o censura), lo cual nos da una idea de su supervivencia en un punto específico. Cada sujeto es caracterizado por tres variables: 1. Su serie de tiempo (t_j); 2. Su estado al finalizar la serie de tiempo ((censura (c_j) o desenlace (d_j)); y 3. El grupo de estudio en el que se encuentra (Tabla 4). Calcula la probabilidad de supervivencia por cada tiempo (t_j) (ecuación 6) (71).

Tabla 4. *Construcción de tabla para análisis de Kaplan-Meier.*

Sujeto	Tiempo (años)	Estado al finalizar periodo	Grupo
1	0	D	I
2	1	C	C
3	2	C	I
4	3	D	C
5	4	D	C
X	5	C	I

D: desenlace; C: censura; Intervención; Control

$$S(t)= \prod \frac{n_j - d_j}{n_j} = 1- d_j/n_j$$

De esa forma, Kaplan Meier calcula la probabilidad de supervivencia por cada tiempo específico (t_j) como una productoria (ecuación 7), que expresa la probabilidad de sobrepasar al tiempo previo en el que se presentó una falla (desenlace) multiplicado por la probabilidad condicional de sobrevivir pasado el tiempo (t_j), dado que al menos se llegó a ese tiempo (t_j). ¡no tan fácil, pero con un ejemplo si!

$$S(t_{(j-i)})=\prod T>t_{(i)} \mid T\geq t_{(i)}$$

Se calcula la supervivencia para cada tiempo t_j; el tiempo transcurrido entre cada desenlace está determinado por un intervalo I_j, que discurre entre el tiempo $t_{(j-1)}$ hasta el tiempo del evento t_j. Para cada tiempo del evento se definen n_j, que es el número de sujetos que llegan vivos (o sin el desenlace, en caso de no ser mortalidad), y dj como el número de sujetos que tienen el desenlace en ese intervalo o cj, si fueron censurados (Tabla 5) (67).

Finalmente, podemos graficar la curva de supervivencia con esos datos en los tiempos t1, t2, t3…t6, marcando las observaciones censuradas con el signo + (Figura 10) (72).

Log rank test (prueba de rangos logarítmicos)

Podemos, de igual forma, graficar las curvas de supervivencia tanto del grupo de intervención como del grupo control, para darnos una idea del comportamiento a través del tiempo y observar diferencias entre ambas, incluso comparar las proporciones de supervivencia en algún tiempo específico, sin embargo, se pierde la comparación total de la experiencia de supervivencia entre los dos grupos

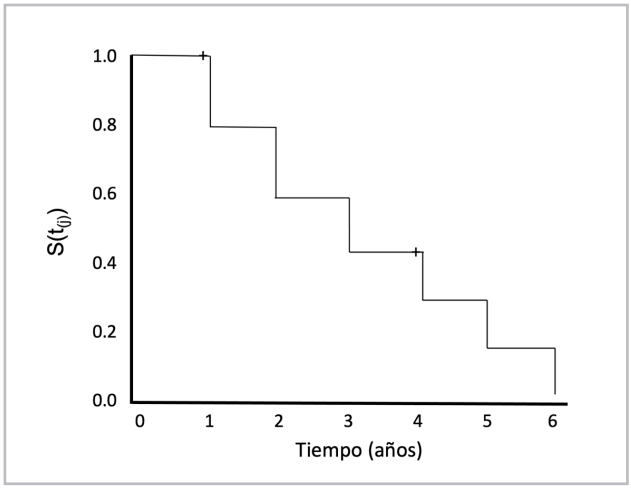

Figura 10. *Curva de supervivencia de 18 sujetos en grupo de intervención en ECA.*

durante el tiempo de seguimiento. Para poder captarlo en toda su magnitud se utiliza el *log rank test* (73). Se prueba la hipótesis nula que no hay diferencia entre las poblaciones (intervención y control) en la probabilidad de muerte en cualquier punto; se calculan por cada intervalo el número de muertes observadas y las esperadas, si en realidad no hubiese diferencias entre ellos ($(O-E)^2/E$) y en la tabla de distribución de Chi^2 se determina la significancia estadística. Se puede también calcular la mediana de supervivencia, que constituye el tiempo en el que la mitad de los pacientes están vivos y la mitad muertos y el promedio de supervivencia por el área bajo las curvas.

El *log rank test* fue propuesto para darle igual peso a todos los tiempos de *falla* en el seguimiento, por lo que asume que la relación de *hazards* (*hazard ratio*) entre los grupos es proporcional a través del tiempo de seguimiento; sin embargo, no es infrecuente que el *hazard ratio* no sea constante, caso en el cual se debe utilizar alguna de estas otras pruebas como la de Gehan-Wilcoxon, Tarone-Ware, Peto-Peto o Fleming-Harrington, entre otros (74).

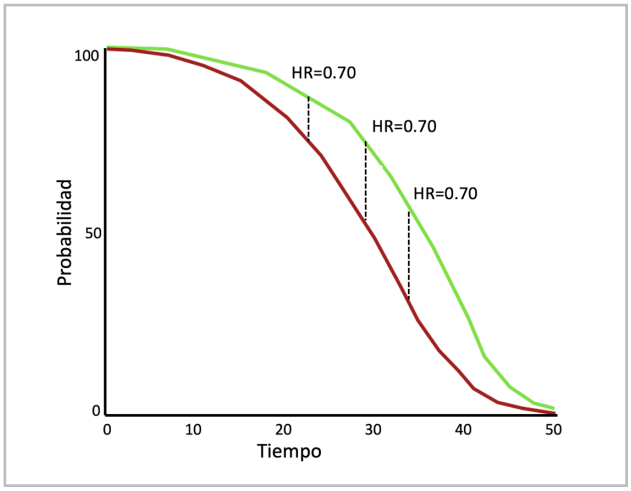

Figura 11. *Riesgo proporcional. Dos grupos en ECA: intervención (rojo) y control (verde). Los hazard pueden cambiar, pero el hazard ratio es constante.*

Modelo de regresión de Cox

La técnica más utilizada para evaluar la relación existente entre variables explicativas y el tiempo de supervivencia fue descrito por Cox en 1972, conocido como modelo de regresión de riesgos proporcionales (75). El modelo describe la relación entre la función *hazard* (riesgo de un desenlace) y un conjunto de covariables o factores. La ecuación (ecuación 8) se escribe así:

$$h(t; X) = h_0(t)e^{\beta X} = h_0(t)\, e^{(\beta 1x1+\beta 2x2+\cdots+\beta pxp)}$$

En la cual h(t; X) es la función *hazard* para el individuo i con valores $X= (x_1,x_2,\ldots x_p)$ en las variables explicativas en el instante t, por lo tanto es la variable a modelar y representa el riesgo de morir en el tiempo t de los sujetos de estudio que tienen un determinado patrón de X de las covariables explicativas; $e^{\beta X}$ es la función exponencial de las p variables explicativas x_i con el correspondiente coeficiente de regresión; $h_0(t)$ es la función *hazard* basal cuando todas las covariables tienen valor de 0 (76).

La función *hazard* basal no es especificada en el modelo, por lo que no se asume que siga un patrón de distribución particular; el efecto del tratamiento y de las covariables es multiplicativo; el *hazard* del evento de interés en un grupo es múltiplo constante del *hazard* del otro; el *hazard ratio* es constante a lo largo del tiempo (Figura 11); hay independencia de los tiempos de supervivencia entre sujetos y estos tienen el mismo riesgo hasta el desenlace o censura (76).

La estimación semiparamétrica en la regresión de Cox se realiza por el método de verosimilitud parcial, aunque

Tabla 5. *Ejemplo del estimador de Kaplan Meier en el grupo de intervención en ECA.*

Grupo de intervención				
t_j (años)	n_j	d_j	c_j	$S(t_{(j)})$
0	18	0	0	1
1	18+	3	1	1 x (1-3/18) =0.834
2	14	3	0	1 x (1-3/18) x (1-3/14) = 0.834 x 0.7857=0.655
3	11	3	0	0.655 x (1-3/11) = 0.655 x 0.727 = 0.476
4	8+	2	1	0.476 x (1-2/8) =0.476 x 0.75 = 0.357
5	5	3	0	0.357 x (1-3/5) = 0.357 x 0.4= 0.142
6	2	1	0	0.142 x (1-1/2) = 0.142 x 0.5 = 0.071

Tiempo en años para facilidad del ejemplo, pero está definido por el momento en que se presenta el evento; t_j: tiempo del evento; n_j: número de sujetos que llegan vivos; d_j: número de sujetos con desenlace; c_j: número de sujetos censurados; $S(t_{(j)})$: probabilidad acumulativa de supervivencia.

cuando se presentan empates (múltiples individuos con el mismo tiempo de *fallo*), puede requerir tiempos de computación considerables, por lo que se utilizan aproximaciones como la propuesta por Breslow, Efron y Cox. El supuesto de riesgo proporcional debe probarse, inicialmente la evaluación de las curvas de Kaplan Meier da una idea de su violación, si se cruzan o una de ellas desciende mientras la otra termina en una meseta. Luego se analizan los llamados residuales de Schoenfeld o por el método de log log Plot para confirmarlo, entre otras pruebas. También puede pasar que el efecto de una covariable cambie en el tiempo, para lo cual se debe plantear un análisis diferente (77, 78).

El HR se calcula por la tasa de *hazard*, tasa de riesgo instantáneo, en cada grupo incluido en el ensayo. La tasa de hazard se define como la tasa de eventos instantáneos condicionados calculado en función del tiempo. Por ejemplo, si a un grupo de 100 pacientes se le somete a un tratamiento y al cabo de un mes fallecen dos, la tasa de *hazard* (*hazard rate*) es 2/100; si en el mes dos fallece un paciente la tasa de *hazard* será 1/98, y así sucesivamente. En este caso la tasa de *hazard* es el número de pacientes que fallecen dividido por el número de pacientes que están vivos al inicio del intervalo de tiempo. Por consiguiente, el HR es la relación de las tasas de *hazard* a la cual experimentan el desenlace los pacientes en los dos grupos (79).

Un error común en la práctica es interpretar el HR como la reducción del riesgo de evento, como un porcentaje igual a 100(1-HR); la reducción del HR significa que el tiempo al evento, por ejemplo, la muerte, se prolonga, no exactamente que se evita. Un HR de dos significa que en cualquier momento el doble de pacientes del grupo intervenido presenta un evento ***proporcionalmente*** al compararlo con el grupo control y un HR de 0.5 indica lo contrario, la mitad de los pacientes en el grupo intervenido tiene un evento en cualquier intervalo de tiempo ***proporcionalmente*** comparado con el grupo control. Un HR=0.6 luce bastante impresionante, pero no representa la reducción del 40% del riesgo de eventos en ningún momento del seguimiento. Para calcular la reducción del riesgo del evento (x) se debe conocer la distribución de supervivencia tanto del grupo intervenido o experimental [$S_e(x)$] como del control [$S_c(x)$] (ecuación 9) (80).

$$x = \frac{S_E(x) - S_C(x)}{1 - S_C(x)}$$

Riesgos no proporcionales

El modelo de riesgos proporcionales asume que el HR es constante (Figura 9). Sin embargo, dada la heterogeneidad de la población incluida en el ECA, es posible que donde se combinan pacientes de alto y bajo riesgo, no se logre un adecuado balance de covariables, lo que dificulta su interpretación, a pesar de demostrar el supuesto. Esto se produce posiblemente por la presencia de covariables no medidas (o desconocidas), que pueden impactar el desenlace (81).

Se asume que esos factores no medidos son multiplicativos en la escala *hazard*, que es la forma de parametrización de la regresión de Cox, por lo que el sesgo incluido (sesgo de selección) aumenta con la magnitud del efecto causal, la heterogeneidad del riesgo basal y el tiempo de seguimiento. Podemos encontrarnos en un escenario donde las curvas divergen temprano, pero tienden a converger en el seguimiento, aunque difícil de explicar, puede ser debido a la disminución gradual de la eficacia del tratamiento o a sesgo de supervivencia. El otro escenario posible es cuando las curvas divergen gradualmente, lo que posiblemente indica que la eficacia del tratamiento aumenta con el tiempo (81). Hay otros posibles escenarios en los que el supuesto de proporcionalidad no se cumple, por lo que se deben utilizar estrategias alternativas para la evaluación.

Tiempo de supervivencia media restringido (TSMR)

Las pruebas de *log rank test* y el modelo de regresión de Cox, el cual permite ajustar por covariables, son robustas en presencia de riesgo no proporcional, en el sentido que retienen algo de la potencia para diferenciar entre dos tratamientos cuyas funciones *hazard* no son proporcionales. No obstante, en casos extremos, como ejemplo clásico cuando las curvas de supervivencia se cruzan, esa potencia se reduce. Por lo anterior, Royston P. y Parmar M. propusieron el análisis de tiempo de supervivencia media restringido (82).

En teoría, el promedio del tiempo de supervivencia puede ser calculado como el área bajo la curva de la función de supervivencia hasta el infinito (Figura 8), siempre que no haya observaciones censuradas; por lo que se utiliza con mayor frecuencia la mediana del tiempo de supervivencia, definida como el tiempo en el que la mitad de los pacientes desarrollan el desenlace de interés, para lo cual se requiere de un número importante de eventos y tiempo de seguimiento para ser estimado. El TSMR es similar al promedio del tiempo de supervivencia, pero como su nombre lo indica, restringido a un tiempo específico (83).

El TSMR puede ser definido como el área bajo la curva de supervivencia de T hasta el tiempo t, es decir, desde el tiempo 0 hasta un momento determinado por el investigador. De igual modo, se puede estimar el tiempo de supervivencia media restringido perdido (TSMRP), como el complemento. El efecto de la intervención se establece al comparar TSMR entre los grupos (intervenido y control). Por ejemplo, definido t=40 meses, si el área bajo la curva es de 35.4, quiere decir que, en promedio, futuros pacientes expuestos a la intervención estarían vivos (si el desenlace es muerte) 35.4 meses de los 40 meses de seguimiento; el TSMRP sería de 4.6 meses. Si al comparar dos grupos el TSMRP es de 4.6 meses en la intervención y 6.7 meses en el control, se interpreta que, en promedio, un paciente en el grupo control estaría vivo 2.1 meses menos (para mayor claridad revisar ejemplo de la referencia 84).

Este análisis permite una interpretación más sencilla e intuitiva, considera toda la distribución de supervivencia

hasta el momento determinado, no requiere cumplir el supuesto de riesgos proporcionales, puede utilizarse como análisis complementario en caso de cumplirlo. Sin embargo, las desventajas pueden ser que la conclusión del TSMR puede variar según el momento especificado para el análisis en el caso de riesgos no proporcionales. Por lo tanto, el intervalo de tiempo debe estar clínicamente motivado y preespecificado en el protocolo del ensayo clínico y el delta de TSMR (diferencia entre grupos) puede parecer un efecto relativamente pequeño en cuanto a los meses (o días) de vida ganado por los años de terapia, posiblemente influenciado por tiempos relativamente cortos de seguimiento (85, 86).

Tiempo de fallo acelerado (TFA)

La literatura es esquiva en la descripción de este modelo y su utilidad, fue descrito en 1966 por Pike MC (87) en el fenómeno de carcinogénesis. Se denomina tiempo de fallo acelerado porque, de similar forma a otros modelos, el término *fallo* representa el desarrollo del evento o desenlace y acelerado porque se supone que el efecto de un covariable es acelerar o desacelerar el curso en una forma constante. El modelo explica la relación entre las probabilidades de supervivencia y unas covariables, estimando una asociación relativa. Proporciona una estimación de las medianas de tiempo del evento, que puede traducirse en la reducción de la duración de la enfermedad (88).

En lugar de estimar *hazard ratio,* el modelo estima el *time ratio* (TR), algo así como la *razón de tiempo*; el TR estima el retraso hasta la ocurrencia de un evento con el tratamiento en comparación con el grupo control. Por ejemplo, un TR de 2 significa que el tiempo hasta que un evento ocurra es el doble de largo en el grupo intervenido con relación al grupo control. Si se tiene un HR de 0.71 y un TR de 1.51, el primero expresa la reducción del 29% del *hazard*, del riesgo instantáneo, de eventos en el grupo intervenido con respecto al control (ojo reducción de *hazard*), mientras que el otro expresa que el tiempo al evento es retrasado por 51% en el grupo intervenido con relación al control. El inverso del TR se conoce como el factor de aceleración (FA), representa la misma dirección del efecto, es decir, un TR=2 es lo mismo que un FA=0.5, indica que el tiempo al evento es dos veces el del control, como se mencionó. Se puede formular en la escala logarítmica (similar a una regresión lineal) como

$Y = \beta_0 + \beta'X + \varepsilon$ (ecuación 10), donde Y=log (T),

ε es un término de error aleatorio que se supone sigue alguna distribución paramétrica y β_0 es el intercepto; existen diversos tipos de modelos de TFA tales como exponencial, Weibull, log-normal, gamma, log-logística, esta última no está restringida al supuesto de riesgos proporcionales (89). Como se pueden ajustar diversas distribuciones de probabilidad, se puede seleccionar el que mejor se ajuste a los datos con el criterio de información de Akaike (AIC).

Análisis de supervivencia restringido a un tiempo puntual (*Milestone análisis -MS-*)

MS es definido como la probabilidad de supervivencia definido por Kaplan-Meier, en un tiempo específico determinado idealmente a priori (probabilidad de 0 a 1 o proporción de 0 a 100%). El método es similar al utilizar regresión logística o al cálculo de OR. Se recomienda que el análisis se realice cuando al menos el último sujeto de la cohorte haya llegado al tiempo definido para análisis. Se utiliza para realizar análisis interinos o para evaluar la "cola" en supervivencia a largo plazo (90, 91, 92). Hay que diferenciarlo del análisis de punto de referencia (*landmark analysis*) descrito por Anderson y colaboradores, para evaluar el sesgo en el análisis de supervivencia cuando una covariable de interés es de hecho una medida del estudio como estado de respuesta (respondedores vs no respondedores); en este último solo se analiza los sujetos que llegan vivos al tiempo de interés (93).

Proporción de victorias o *Win Ratio* (WR)

Los desenlaces compuestos son frecuentemente utilizados en ECA por disminuir el tamaño de muestra requerido, el tiempo de seguimiento y los costos. Generalmente se evalúan con el estimativo de KM, la prueba de *Log Rank* y el modelo de regresión de Cox para ajustar por variables y obtener un HR, en términos de tiempo al evento. Esta aproximación tiene limitaciones inherentes dado que considera a todos los componentes con igual importancia y solo aplica el primer evento, desechando los eventos recurrentes; los eventos fatales son tratados de la misma forma como los no fatales; además, no toma en cuenta desenlaces categóricos o continuos, como calidad de vida, fracción de eyección o prueba de seis minutos, por mencionar algunos (94).

Para superar estas dificultades, se han explorado métodos que incorporan la importancia clínica de los desenlaces, tales como las comparaciones generalizadas por pares (*generalized pairwise comparisons*, GPC) y la estadística de las victorias *(win ratio*, *win odds* y *net clinical benefit*) (95). Fue inicialmente propuesto por Finkelstein and Schoenfeld en 1999 (96), luego propuesto como GPC por Buyse (97) y en 2012 descrito como WR por Pocock y colaboradores (98).

El método de WR tiene tres pasos para la comparación de los eventos con respecto a la intervención versus el control: 1. Se forman pares de pacientes, teniendo en cuenta el riesgo basal; 2. Para cada pareja se analiza el desenlace más importante (ejemplo, muerte de causa cardiovascular), si un paciente tuvo muerte CV, el otro se sigue por más tiempo para definir quien presentó primero el evento (victoria en muerte); si ninguno de los dos falleció, entonces, se determina el segundo desenlace en importancia (ejemplo, hospitalización por falla cardiaca), utilizando la misma estrategia (victoria en hospitalización por falla); el resto son empates o no ganadores.

De esta forma quedan cinco categorías: a) Paciente intervenido (I) tiene muerte CV primero; b) Paciente en

grupo control (C) tiene muerte CV primero; c) Paciente intervenido tiene hospitalización por falla primero, d) Paciente en grupo control tiene hospitalización por falla primero; y d) Ninguno de las alternativas se cumple. Con estos datos se resumen los hallazgos: N_a, N_b, N_c, N_d y N_e, donde $N_b + N_d = N_w$ son las victorias de la intervención nueva; de igual forma, $N_a + N_c = N_L$ son las derrotas para la nueva intervención. El WR es igual a N_w/N_L (98). Se puede hacer análisis sin emparejar, en el cual cada paciente en el grupo intervenido es comparado con cada paciente en el grupo control.

En el ejemplo del artículo de Pocock y colaboradores reanalizan el estudio EMPHASIS-HF, en el cual se comparó el efecto de la eplerenona contra placebo en pacientes con falla cardiaca, NYHA clase II, con fracción de eyección ≤ 35% con mediana de seguimiento de 21 meses. El HR para el desenlace combinado de muerte CV u hospitalización por falla fue de 0.63 (IC 95% 0.54-0.74, p<0.0001), muerte CV 0.76 (IC 95% 0.61-0.94) y hospitalización por falla 0.58 (IC 95% 0.47-0.70). Aunque el resultado es importante, parte del efecto sobre muerte CV se diluye porque las hospitalizaciones tienden a presentarse primero. Al hacer el análisis con WR se mostraron los siguientes resultados: N_a=90, N_b=118, N_c=61, N_d=131 y N_e=964, número total de pares=1364; con esto se obtiene WR para muerte CV 118/90=1.31 y WR para desenlace combinado (118+131) / (90+61) =1.65 (98).

El análisis realizado con WR puede parecer simple y no tan robusto como el análisis clásico de tiempo al evento. Sin embargo, tiene soporte estadístico importante en su forma de cálculo y evidencia clínica en reanálisis de estudios que sugieren que el HR y WR proporcionan un estimativo similar del efecto de una intervención. Aunque son conceptos diferentes, el uso del inverso del HR podría compararse con el WR para darse una idea general del efecto de la intervención (por supuesto no equiparables). En forma simplista, el mensaje centrado en el paciente podría ser que para un WR =1.2 con un IC 95% que no cruce la unidad, el tratamiento nuevo es 20% mejor en reducir la muerte y rehospitalizaciones que el placebo, considerando el desenlace de muerte como prioritario (99).

Un problema importante del WR es cuando ocurren empates (ni victorias ni pérdidas), las cuales son ignoradas, lo que resulta en sobreestimación del efecto. Para esto se ha propuesto el uso de *win odds* (WO), en el cual la mitad de los empates se les suman a las victorias del grupo de tratamiento y del control. El estudio TRILUMINATE evaluó la eficacia de la reparación percutánea de la regurgitación tricuspídea severa con el dispositivo TriClip, con un desenlace jerárquico compuesto por muerte de cualquier causa o requerimiento de cirugía de la válvula tricúspide. Se obtuvieron 11.348 victorias y 7.643 derrotas para un WR calculado de 1.48 (IC 95% 1.06-2.13); sin embargo, hubo 11.634 empates (40% de las parejas), lo que llevaría a calcular un WO de 1.28, mucho menos impactante que el dato inicial (100).

Limitaciones

La revisión fue realizada con el objetivo de dar bases claras dirigida a los clínicos, especialmente internistas y cardiólogos, que se esfuerzan en hacer lectura crítica de la evidencia científica, no para quienes no les interesa lectura crítica y se conforman con leer el resumen o *abstract*. Las notaciones y ecuaciones matemáticas son aproximadas, no inexactas, puesto que está dirigido a clínicos, no a bioestadísticos, ni mucho menos matemáticos, con la intención de hacerlas más simples y el concepto general entendible, por eso espero que me excusen mis puristas profesores bioestadísticos. En algunos casos, puede partir de la lógica de la difícil comprensión de un clínico, adentrado como neófito de la epidemiología clínica, en un intento por traducir extraños saberes a nuestra práctica, pero comunes, por el soporte que dan a la evidencia científica que utilizamos, en el día a día sin percatarnos de ello.

Agradecimientos

Lo extenso tiene que ver con la importancia del tema y al honor aceptado al ser elegido para dictar la conferencia central del Congreso Nacional de Medicina Interna: Conferencia Lombana Barreneche.

Referencias

Parte I

1. **Castillo M.** The scientific method: a need for something better? *AJNR Am J Neuroradiol*. 2013;34(9):1669-71.
2. **Liu L, Jones BF, Uzzi B, Wang D.** Data, measurement and empirical methods in the science of science. *Nat Hum Behav*. 2023;7(7):1046-1058.
3. **Plesser HE.** Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Front Neuroinform. 2018; 18:11(76).
4. **Feistein A.** Clinical epidemiology. I. The populational experiments of nature and of man in human illness. Ann Intern Med. 1968;69(4):807-20.
5. **Grimes DA, Schulz KF.** An overview of clinical research: the lay of the land. Lancet. 2002;359(9300):57-61.
6. **John ER, Abrams KR, Brightling CE, Sheehan NA.** Assessing causal treatment effect estimation when using large observational datasets. *BMC Med Res Methodol*. 2019; 19(1): 207.
7. **Geershman SJ, Ullman TD.** Cuasal implicatures from correlational statements. *PLoS One*. 2023; 18(5): e0286067.
8. **Senn S.** Seven myths of randomisation in clinical trials. *Stat Med*. 2013;32(9):1439-50.
9. **Fanaroff AC, Califf RM, Harrington RA, Granger CB, McMurray JJV, Patel M et al.** Randomized Trials Versus Common Sense and Clinical Observation: JACC Review Topic of the Week. *J Am Coll Cardiol*. 2020;76(5):580-589.
10. **Berlin JA, Golub RM.** Meta-analysis as Evidence Building a Better Pyramid. *JAMA*. 2014; 312(6):603-5.
11. **Murad MH, Asi N, Alsawas M, Alahdab F.** New evidence pyramid. *Evid Based Med*. 2016;21(4):125-7.
12. **Council for International Organizations of Medical Sciences.** International Ethical Guidelines for Health-Related Research Involving Humans. Geneva, Switzerland: Council for International Organizations of Medical Sciences; 2016. [accedido abril 2024]: Disponible en: https://cioms.ch/publications/product/international-ethical-guidelines-for-health-related-research-involving-humans/
13. **Miller FG, Joffe S.** Equipoise and the dilemma of randomized clinical trials. N Engl J Med. 2011;364(5):476-80.
14. **ICH Official website**. International Harmonised Tripartite Guideline: General Considerations for Clinical Trials [Internet]. Accedido marzo 2024. Disponible en https://www.ich.org/
15. **Stanley K**. Design of randomized controlled trials. *Circulation*. 2007;115(9):1164-9.
16. **Martínez-Franco M, Nirta-Perez AR, Donado-Gómez JH.** Tipos de ensayos clínicos con asignación aleatoria publicados en PubMed durante 40 años. *Acta Med Colomb*. 2021; 46(2).

17. **Leung JT, Barnes SL; Lo ST, Leung DY.** Non-inferiority trials in cardiology: what clinicians need to know. *Heart*. 2020;106(2):99-104.
18. **Pocock SJ, Clayton TC, Stone GW.** Challenging Issues in Clinical Trial Design: Part 4 of a 4-Part Series on Statistics for Clinical Trials. *J Am Coll Cardiol*. 2015;66(25):2886-2898.
19. **Wang, B, Wang H, Tu XM, Feng CH.** Comparisons of Superiority, Non-inferiority, and Equivalence Trials. Shanghai Arch Psychiatry 2017;29(6):385-388.
20. **Bikdeli B, Welsh JW, Akram Y, Punnanithinont N, Lee I, Desai N et al**. Noninferiority Designed Cardiovascular Trials in Highest-Impact Journals. *Circulation*. 2019;140(5):379-389.
21. **Kaul S, Diamond GA.** Good enough: a primer on the analysis and interpretation of noninferiority trials. *Ann Intern Med*. 2006;145(1):62-9.
22. **Ellis SG, Kereiakes DJ, Metzger DC, Caputo RP, Rizik DG, Teirstein PS et al.** Everolimus-Eluting Bioresorbable Scaffolds for Coronary Artery Disease. N Engl J Med. 2015;373(20):1905-15.
23. **Sibbald B, Roland M.** Understanding controlled trials: Why are randomised controlled trials important? *BMJ*. 1998;316(7126):201.
24. **Phillips MR, Kaiser P, Thabane L, Bhandari M, Chaudhary V, Wykoff CC, et al.** Risk of bias: why measure it, and how? *Eye (Lond)*. 2022;36(2):346-348.
25. **Pocock SJ, Clayton TC, Stone GW.** Design of Major Randomized Trials: Part 3 of a 4-Part Series on Statistics for Clinical Trials. *J Am Coll Cardiol*. 2015;66(24):2757-2766.
26. **Schulz KF, Grimes DA.** Blinding in randomised trials: hiding who got what. *Lancet*. 2002;359(9307):696-700.
27. **Cobb LA, Thomas GI, Dillard DH, Merendino KA, Bruce RA.** An evaluation of internal-mammary-artery ligation by a double-blind technic. *N Engl J Med*. 1959;260(22):1115-8.
28. **Bhatt DL, Kandzari DE, ONeill WW, DAgostino R, Flack JM, Katzen BT et al.** A controlled trial of renal denervation for resistant hypertension. *N Engl J Med*. 2014;370(15):1393-401.
29. **Walter SD, Guyatt G, Montori VM, Cook R, Prasad K.** A new preference-based analysis for randomized trials can estimate treatment acceptability and effect in compliant patients. *J Clin Epidemiol*. 2006;59(7):685-96.
30. **Abraha I, Cherubini A, Cozzolino F, De Florio R, Luchetta ML, Rimland JM et al.** Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. *BMJ*. 2015; 350:h2445.
31. **DAgostino Sr RB, Massaro JM, Sullivan LM.** Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. *Stat Med*. 2003;22(2):169-86.
32. **Mo Y, Lim Ch, Watson JA, White NJ, Cooper BS.** Non-adherence in non-inferiority trials: pitfalls and recommendations. *BMJ*. 2020; 370: m2215.
33. **Rudolph JE, Zhong Y, Duggal P, Mehta SH, Lau B.** Defining representativeness of study samples in medical and population health research. *BMJ Med*. 2023;2(1):e000399.
34. **Tan YY, Papez V, Chang WH, Mueller SH, Denaxas S, Lai AG.** Comparing clinical trial population representativeness to real-world populations: an external validity analysis encompassing 43 895 trials and 5 685 738 individuals across 989 unique drugs and 286 conditions in England. *Lancet Healthy Longev*. 2022;3(10):e674-e689.
35. **Gross A, Harry AC, Clifton CS, Della Pasqua O.** Clinical trial diversity: An opportunity for improved insight into the determinants of variability in drug response. *Br J Clin Pharmacol*. 2022;88(6):2700-2717.
36. **Degtiar I, Rose S.** A Review of Generalizability and Transportability. *Ann Rev*. 2023;10(1):501-524.
37. **Ling AY, Montez-Rath ME, Carita P, Chandross KJ, Lucats L, Meng Z et al.** An Overview of Current Methods for Real-world Applications to Generalize or Transport Clinical Trial Findings to Target Populations of Interest. *Epidemiology*. 2023;34(5):627-636.
38. **Leek JT, Peng RD.** Statistics: P values are just the tip of the iceberg. *Nature*. 2015;520(7549):612.
39. **Wasserstein RL, Lazar NA.** The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*. 2016;70(2):129-33.
40. **Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole Ch, Goodman SN, Altman DG.** Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016; 31: 337–350.
41. **Chiu K, Grundy Q, Bero L.** 'Spin' in published biomedical literature: A methodological systematic review. *PLoS Biol*. 2017; 15(9): e2002173.
42. **Smith R.** ¿ What are medical journals for and how well do they fulfil those functions? [Internet]. *The BMJ Opinion*. 2016 [accedido abril 2024]. Disponible en: https://blogs.bmj.com/bmj/2016/04/19/richard-smith-what-are-medical-journals-for-and-how-well-do-they-fulfil-those-functions/
43. **Venugopal N, Saberwal G.** A comparative analysis of important public clinical trial registries, and a proposal for an interim ideal one. *PLoS One*. 2021; 16(5): e0251191.
44. **Walker KF, Stevenson G, Thornton JG.** Discrepancies between registration and publication of randomised controlled trials: an observational study. *JRSM Open*. 2014;5(5):2042533313517688.
45. **Chen T, Li CH, Qin R, Wang Y, Yu D, Dodd J et al.** Comparison of Clinical Trial Changes in Primary Outcome and Reported Intervention Effect Size Between Trial Registration and Publication. *JAMA Netw Open*. 2019;2(7):e197242.
46. **Pocock SJ, Hughes MD, Lee RJ.** Statistical Problems in the Reporting of Clinical Trials. *N Engl J Med*. 1987;317(7):426-32.
47. **Pocock SJ, McMurray JJV, Collier TJ.** Statistical Controversies in Reporting of Clinical Trials: Part 2 of a 4-Part Series on Statistics for Clinical Trials. *J Am Coll Cardiol*. 2015;66(23):2648-2662.
48. **Prentice RL.** Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*. 1989;8(4):431-40.
49. **Wittes J, Lakatos E, Probstfield J**. Surrogate endpoints in clinical trials: cardiovascular diseases. *Stat Med*. 1989;8(4):415-25.
50. **Pocock SJ, Stone GW.** The Primary Outcome Is Positive - Is That Good Enough? *N Engl J Med*. 2016;375(10):971-9.
51. **Armstrong PW, Westerhout CM.** Composite End Points in Clinical Research: A Time for Reappraisal. Circulation. 2017;135(23):2299-2307.
52. **Baracaldo-Santamaría D, Feliciano-Alfonso JE, Ramirez-Grueso R, Rojas-Rodríguez LC, Dominguez-Dominguez CA, Calderon-Ospina CA.** Making Sense of Composite Endpoints in Clinical Research. *J Clin Med*. 2023;12(13):4371.
53. **Brankovic M, Kardys I, Steyeberg EW, Lemeshow S, Markovic M, Rizopoulos D, Boersma E.** Understanding of interaction (subgroup) analysis in clinical trials. *Eur J Clin Invest*. 2019;49(8):e13145.
54. **Oxman AD, Guyyatt GH.** A consumer's guide to subgroup analyses. *Ann Intern Med*. 1992;116(1):78-84.
55. **Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Smith GD.** Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess*. 2001;5(33):1-56.
56. **Pocock SJ, Stone GW. The Primary Outcome Fails - What Next?** *N Engl J Med*. 2016;375(9):861-70.
57. **Pocock SJ, Rosello X, Owen R, Collier TJ, Stone GW, Rockhold FW.** Primary and Secondary Outcome Reporting in Randomized Trials: JACC State-of-the-Art Review. *J Am Coll Cardio.l* 2021;78(8):827-839.
58. **Pocock SJ, McMurray JJV, Collier TJ.** Making Sense of Statistics in Clinical Trial Reports: Part 1 of a 4-Part Series on Statistics for Clinical Trials. *J Am Coll Cardiol*. 2015;66(22):2536-49.

Parte II

59. **Ranganathan P, Pramesh CS, Aggarwal R**. Common pitfalls in statistical analysis: Absolute risk reduction, relative risk reduction, and number needed to treat. *Perspect Clin Res*. 2016;7(1):51-3.
60. **Sedgwick P.** Odds and odds ratio. *BMJ*. 2013; 347:f5067
61. **Rotella J.** Probability, log-odds, and odds [Internet]. *Montana.edu*. [Accedido abril 2024]. Disponible en: https://www.montana.edu/rotella/documents/502/Prob_odds_log-odds.pdf acceso marzo 2024.
62. **Morris JA, Gardner MJ.** Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *Br Med J (Clin Res Ed)*. 1988;296(6632):1313-6.
63. **Grant RL.** Converting an odds ratio to a range of plausible relative risks for better communication of research findings. *BMJ*. 2014;348:f7450. doi:
64. **Barrat A, Wyer PC, Hatala R, McGinn T, Dans A, Keitz S et al.** Tips for learners of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. *CMAJ*. 2004;171(4):353-8.
65. **Altman DG, Andersen PK.** Calculating the number needed to treat for trials where the outcome is time to an event. *BMJ*. 1999;319(7223):1492-5.
66. **Kleinbaum DG, Klein M.** Introduction to survival analysis. Survival Analysis, a self learning text. Third edition. *Springer Link*, New York;2012, pg: 1-54.
67. **Clark TG, Bradburn MJ, Love SB, Altman DG.** Survival analysis part I: basic concepts and first analyses. *Br J Cancer*. 2003;89(2):232-8.
68. **Deo SV, Deo V, Sundaram V.** Survival analysis-part 1. *Indian J Thorac Cardiovasc Surg*. 2020;36(6):668-672.
69. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958;53(282):457–481.
70. **Rosello X, González-Del-Hoyo M.** Análisis de supervivencia en investigación cardiovascular: lo esencial. *Rev Esp Cardiol*. 2022; 75(1):67-76.
71. **Rich JT, Neely JG, Paniello RC, Voelker CCJ, Nussenbaum B, Wang EW.** A

practical guide to understanding Kaplan-Meier curves. *Otolaryngol Head Neck Surg*. 2010;143(3):331-6.

72. **Bland JM, Altman DG.** Survival probabilities (the Kaplan-Meier method). *BMJ*. 1998;317(7172):1572.
73. **Bland JM, Altman DG.** The long rank test. *BMJ*. 2004;328(7447):1073. doi: 10.1136/bmj.328.7447.1073.
74. **Karadeniz PG, Ercan I.** Examining Tests for Comparing Survival Curves with Right Censored Data. *Statistics in Transition*. 2017;18(2):311-328.
75. **Cox DR.** Regression models and lifetables. *J R Stat Soc Ser B*. 1972;34(2):187-202.
76. **Patel K, Kay R, Rowell L.** Comparing proportional hazards and accelerated failure time models: an application in influenza. Pharm Stat. 2006;5(3):213-24
77. **Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME, Groothuis-Ousdshoorn CGM.** Time-varying covariates and coefficients in Cox regression models. *Ann Transl Med*. 2018;6(7):121.
78. **Ng'andu NH.** An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Stat Med*. 1997;16(6):611-26.
79. **Kay R.** An explanation of the hazard ratio. *Pharm Stat*. 2004; 3(4):295-7
80. **Sashegyi A, Ferry D.** On the Interpretation of the Hazard Ratio and Communication of Survival Benefit. *Oncologist*. 2017;22(4):484-486.
81. **Stensrud MJ, Aalen JM, Aalen OO, Valberg M.** Limitations of hazard ratios in clinical trials. *Eur Heart J*. 2019;40(17):1378-1383.
82. **Royston P, Parmar MKB.** The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med*. 2011;30(19):2409-21.
83. **Han K, Jung I.** Restricted Mean Survival Time for Survival Analysis: A Quick Guide for Clinical Researchers. *Korean J Radiol*. 2022;23(5):495-499.
84. **Zhao L, Claggett B, Tian L, Uno H, Pfeffer MA, Solomon S, et al.** On the restricted mean survival time curve in survival analysis. *Biometrics*. 2016;72(1):215-21.
85. **Kloecher DE, Davies MJ, Khunti K, Zaccardi F.** Uses and Limitations of the Restricted Mean Survival Time: Illustrative Examples From Cardiovascular Outcomes and Mortality Trials in Type 2 Diabetes. *Ann Intern Med*. 2020;172(8):541-552.
86. **Perego C, Sbolli M, Specchia C, Fiuzat M, McCaw ZR, Metra M et al.** Utility of Restricted Mean Survival Time Analysis for Heart Failure Clinical Trial Evaluation and Interpretation. JACC Heart Fail. 2020;8(12):973-983.
87. **Pike MC**. A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*. 1966;22(1):142-161
88. **Rosello X, González-Del-Hoyo M.** Análisis de supervivencia en investigación cardiovascular (II): metodología estadística en situaciones complejas. *Rev Esp Cardiol*. 2022; 75(1):77-85.
89. **Gregson J, Sharples L, Stone GW, Burman CF, Ohrn F, Pocock S.** Nonproportional Hazards for Time-to-Event Outcomes in Clinical Trials: JACC Review Topic of the Week. *J Am Coll Cardiol*. 2019;74(16):2102-2112.
90. **Chen TT.** Milestone Survival: A Potential Intermediate Endpoint for Immune Checkpoint Inhibitors. *J Natl Cancer Inst*. 2015;107(9):djv156.
91. **Damuzzo V, Agnoletto L, Leonardi L, Chiumente M, Mengato D, Messori A.** Analysis of Survival Curves: Statistical Methods Accounting for the Presence of Long-Term Survivors. *Front Oncol*. 2019;9:453.
92. **Hellmann MD, Kris MG, Rudin CM.** Medians and Milestones in Describing the Path to Cancer Cures: Telling "Tails". *JAMA Oncol*. 2016; 2(2):167-8.
93. **Anderson JR, Cain KC, Gelber RD.** Analysis of survival by tumor response. *J Clin Oncol*. 1983;1(11):710-9.
94. **Verbeeck J, De Backer M, Verwerft J, Salvaggio S, Valgimigli M, Vranckx P et al.** Generalized Pairwise Comparisons to Assess Treatment Effects: JACC Review Topic of the Week. *J Am Coll Cardiol*. 2023;82(13):1360-1372.
95. **Dong G, Huang B, Verbeeck J, Cui Y, Song J, Gamalo-Siebers M, et al.** Win statistics (win ratio, win odds, and net benefit) can complement one another to show the strength of the treatment effect on time-to-event outcomes. *Pharm Stat*. 2023 Jan;22(1):20-33.
96. **Finkelstein D, Schoenfeld DA.** Combining mortality and longitudinal measures in clinical trials. *Stat Med*. 1999;18(11): 1341-1354.
97. **Buyse M**. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med*. 2010;29(30):3245-57.
98. **Pocock SJ, Ariti CA, Collier TJ, Wang D.** The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J*. 2012;33(2):176-82.
99. **Ferreira JP, Jhund PS, Duarte K, Clagget BL, Solomon SD, Pocock S, et al.** Use of the Win Ratio in Cardiovascular Trials. *JACC Heart Fail*. 2020;8(6):441-450.
100. **Ajufo E, Nayak A, Mehra MR.** Fallacies of Using the Win Ratio in Cardiovascular Trials: Challenges and Solutions. *JACC Basic Transl Sci*. 2023;8(6):720-727.