

El mito de la evidencia clínica: sesgos, heurísticos, falsedades y errores metodológicos

The myth of clinical evidence: biases, heuristics, falsehoods and methodological errors

Jesús Martín-Fernández^{a-d} y Elena Polentinos-Castro^{b-e}

^a Departamento de Especialidades Médicas y Salud Pública. Facultad de Ciencias de la Salud. Universidad Rey Juan Carlos. Madrid (España).

^b Unidad Docente de Atención Familiar y Comunitaria Oeste. Gerencia Asistencial de Atención Primaria, Servicio Madrileño de Salud. Madrid (España).

^c Red de Investigación en Cronicidad, Atención Primaria y Promoción de la Salud RICAPPS-(RICORS). ISCIII (España).

^d Instituto de Investigación Sanitaria Gregorio Marañón (IISGM). Madrid (España).

^e Unidad de Investigación. Gerencia Asistencial de Atención Primaria. Servicio Madrileño de Salud. Madrid (España).

CORREO ELECTRÓNICO:

Elena Polentinos-Castro.
epolentinos@gmail.com

RESUMEN

La incorporación de la evidencia científica a la toma de decisiones clínicas afronta dificultades relevantes aun cuando el marco metodológico de la investigación clínica sea robusto. En la práctica, muchos estudios publicados ofrecen información de utilidad limitada para la atención sanitaria debido a problemas en el diseño, escasa orientación a resultados relevantes para los pacientes o una insuficiente contextualización de estos. Entre los problemas más frecuentes, se encuentran los sesgos y otros errores metodológicos que pueden aparecer en el diseño, en la recogida de datos y en el análisis o la difusión de resultados, y que tienden a producir estimaciones distorsionadas del efecto. A estas limitaciones se añaden prácticas ampliamente aceptadas, aunque metodológicamente problemáticas, como la selección selectiva de resultados, la falta de registro previo de los estudios, la minimización de limitaciones o la exageración de conclusiones, sin olvidar la confusión entre significación estadística y relevancia clínica, que puede impulsar recomendaciones basadas en beneficios absolutos pequeños asociados a costes elevados. En la práctica clínica, el uso de heurísticos (atajos mentales) puede resultar inevitable para la toma de decisiones rápidas, pero, en la incorporación de la evidencia, estos heurísticos pueden consolidar sobreestimaciones y generalizaciones inapropiadas si no se reconocen explícitamente. Finalmente se aborda el impacto del fraude y la mala conducta científica, sobre la validez de la evidencia, y la capacidad para contaminar revisiones sistemáticas, metanálisis y guías de práctica clínica, durante años. En conclusión, la práctica clínica basada en evidencia exige lectura crítica sistemática, transparencia metodológica y una cultura de integridad científica, junto con conciencia de los sesgos cognitivos que influyen en cómo se interpreta y aplica la investigación en condiciones clínicas reales.

Palabras clave: medicina basada en la evidencia, sesgos, heurísticos, fraude, difusión de la información, relevancia clínica, toma de decisiones clínicas.

ABSTRACT

The incorporation of scientific evidence into clinical decision-making faces relevant difficulties even when the methodological framework of clinical research is robust. In practice, many published studies offer information of limited usefulness for healthcare, due to problems in design, limited orientation toward outcomes relevant to patients or insufficient contextualization of these. Among the most common problems are bias and other methodological errors that may appear in the design, in data collection, in analysis, or in the dissemination of results and that tend to lead to distorted estimates of effect. In addition to these limitations, there are widely accepted but methodologically problematic practices, such as selective reporting of results, failure to register studies in advance, minimization of limitations or exaggeration of conclusions, not to mention confusion between statistical significance and clinical relevance, which can lead to recommendations based on small absolute benefits associated with high costs.

In clinical practice, the use of heuristics (mental shortcuts) may be unavoidable for rapid decision-making. However, in the incorporation of evidence, these heuristics may consolidate overestimations and inappropriate generalizations if not explicitly recognized. Finally, the impact of fraud and scientific misconduct on the validity of the evidence is tackled as well as their capacity to contaminate systematic reviews, meta-analyses and clinical practice guidelines for years.



El contenido de la Revista Clínica de Medicina de Familia está sujeto a las condiciones de la licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0

In conclusion, evidence-based clinical practice requires systematic critical reading, methodological transparency and a culture of scientific integrity, together with awareness of the cognitive bias that influence how research is interpreted and applied in actual clinical conditions.

Keywords: Bias; Clinical Decision-Making; Clinical Relevance; Evidence-Based Medicine; Fraud; Heuristics; Information Dissemination.

La aplicación de las evidencias obtenidas en la investigación clínica se enfrenta a diversas dificultades. El paradigma de la investigación basada en la epidemiología clínica, que abarca estudios experimentales y observacionales, está destinado a responder preguntas sobre tratamiento, prevención, diagnóstico o pronóstico. Dicho paradigma es conceptualmente sólido y ampliamente aceptado, pero en la práctica se ve debilitado por diseños inadecuados, escasa orientación al paciente, poca ganancia informativa y falta de contextualización. La abrumadora cantidad de estudios publicados, muchos de ellos irrelevantes o metodológicamente deficientes, dificulta identificar la evidencia verdaderamente útil y supone un desperdicio de recursos¹.

En este artículo examinaremos de forma breve cómo el camino entre las conclusiones aportadas por la literatura científica a preguntas pertinentes y la toma de decisiones clínicas se ve condicionado por sesgos de publicación y fallos metodológicos, por artificios interpretativos o por conductas fraudulentas que erosionan su fiabilidad^{2,3}. En un entorno clínico marcado por la incertidumbre, en el que las y los profesionales debemos recurrir a decisiones rápidas, resulta imprescindible adoptar una postura crítica que reconozca estas limitaciones. A partir de esta reflexión, exploraremos ejemplos y estrategias prácticas para detectar debilidades en la evidencia y fortalecer la toma de decisiones en la consulta.

Para ello, agruparemos la problemática de la aplicación de la evidencia en la práctica en tres epígrafes: sesgos y otros errores metodológicos, heurísticos y fraudes.

SESGOS Y OTROS ERRORES METODOLÓGICOS

Partimos de la definición de sesgo como la combinación de errores que pueden tener lugar durante el diseño, la recogida de datos, el análisis o la difusión de resultados y que tienden a producir resultados que no se deberían haber producido. Estos sesgos pueden haberse cometido de forma más o menos inadvertida, en relación con la potencia del estudio, con errores derivados del uso inapropiado de test estadísticos^{4,5}, cuando los tamaños de efecto son pequeños o cuando la significación estadística se convierte en el resultado que obtener⁶.

Existen otras deficiencias metodológicas que, aun cuando no impliquen una intención deliberada de distorsión, resultan inaceptables desde una perspectiva científica, como el sesgo de publicación, la falta de discusión de explicaciones alternativas para los hallazgos, selección injustificada de las variables resultado, no pre-registrar los estudios, minimizar u omitir las limitaciones del estudio o exagerar la importancia e impacto de los resultados⁷. Se explican a

continuación las implicaciones de dichos sesgos a través de algunos ejemplos.

Irrelevancia clínica y responsabilidad social

Este error sería de los más importantes a la hora de implantar nuevas tecnologías como nuevos medicamentos, y sucede cuando se atribuye importancia clínica a diferencias estadísticamente significativas sin considerar su relevancia clínica, lo que afectaría a las recomendaciones incorporadas en guías o protocolos, entre otros^{8,9}. Por ello, no solo deben implicarse en su detección los clínicos, sino también todas aquellas agencias públicas y de evaluación de tecnologías encargadas de la financiación de estas a cargo de los presupuestos públicos. Un ejemplo de este tipo de sesgo lo encontramos en la interpretación de los resultados del estudio CAPRIE, cuyos autores concluyeron que el medicamento clopidogrel era más eficaz que el ácido acetilsalicílico para la prevención secundaria de eventos vasculares graves en pacientes con enfermedad cardiovascular previa (infarto reciente, ictus isquémico en los 6 meses anteriores y enfermedad arterial periférica)¹⁰. En el grupo de clopidogrel, el riesgo de eventos cardiovasculares fue de 5,32%, comparado con 5,83% en el del ácido acetilsalicílico ($p = 0,043$), con una reducción del riesgo relativo del 8,7% (intervalo de confianza [IC] 95%: 0,3-16,5) y una reducción absoluta de 0,51% (o 51 eventos por cada 10.000 personas-año). Esto, en términos clínicos, y teniendo en cuenta la baja tasa de eventos que ocurrió en las dos ramas, parece una diferencia con una relevancia clínica bastante marginal, más teniendo en cuenta que no hubo diferencias en términos de seguridad. Pero los autores no dudaron en presentar una conclusión que de forma taxativa señalaba que el clopidogrel era más eficaz que el ácido acetilsalicílico. Otros resultados que no se presentaron en las conclusiones fueron que no todos los subgrupos de personas con enfermedades cardiovasculares parecían beneficiarse de la misma forma en dicha reducción, sino que, por ejemplo, los que tenían infarto reciente parecía que les fue mejor con el ácido acetilsalicílico, y en los que fueron incluidos con antecedente de ictus no hubo diferencias estadísticamente significativas¹¹. Por último, los organismos que deciden la financiación pública deberían valorar si esa mínima diferencia justificaba el incremento del coste que suponía adoptar la preferencia por el clopidogrel, valorando el coste de oportunidad que puede suponer socialmente.

Mitos y equívocos sobre el tamaño de la muestra

La determinación del tamaño muestral, además de sus implicaciones en la factibilidad del estudio, se encuentra estrechamente vinculada a consideraciones éticas, como la exposición innecesaria de participantes a condiciones experimentales cuando su número supera el requerido para alcanzar conclusiones válidas, así como la posible sobreinterpretación de los resultados obtenidos a partir de tamaños muestrales elevados¹². Sin pretender profundizar en aspectos estadísticos, queremos reseñar algunos sesgos que se producen alrededor del tamaño muestral. Cuando calculamos el tamaño de la muestra, realizamos suposiciones sobre medias y desviaciones estándar esperadas o sobre los riesgos de un evento en diferentes grupos (expuestos/no expuestos, rama control/intervención), o bien sobre los tamaños de efecto previstos. Estas suposiciones suelen basarse en estudios previos o en estudios piloto preliminares, a menudo con un tamaño muestral arbitrario. En

este contexto, pueden introducirse sesgos al seleccionar dichos estudios, al no considerar de forma adecuada el tipo de diseño (por ejemplo, en estudios multicéntricos o por conglomerados), o al estimar efectos excesivamente optimistas en favor de la propia hipótesis, algo bastante habitual. Otro sesgo frecuente aparece en los estudios con múltiples objetivos. El tamaño muestral suele calcularse en función del resultado principal, por lo que puede resultar sobredimensionado o insuficiente para los resultados secundarios. En consecuencia, estos últimos deberían interpretarse solo como exploratorios, incluso cuando la p sea $<0,05$, ya que, al contrastar múltiples hipótesis, algunas pueden alcanzar significación estadística por puro azar¹³. Otro sesgo proviene de la creencia de que un tamaño muestral suficientemente grande ofrece resultados válidos, lo que puede dar lugar a convicciones irreversibles. No es que deba desecharse dicho resultado, pero hemos de estar abiertos a nuevos datos o estudios pequeños que pueden ser válidos y útiles¹⁴, y considerar que el tamaño muestral debe tener en cuenta siempre el contexto y los objetivos del estudio⁸.

Por último, en muestras muy grandes, incluso cercanas al tamaño poblacional, la probabilidad de que no existan diferencias entre grupos es casi nula. En estos casos, un tamaño muestral lo bastante grande hace que incluso diferencias mínimas o asociaciones triviales resulten estadísticamente significativas, sin que ello implique necesariamente relevancia clínica o práctica¹⁵.

Generalizaciones inaceptables e indicaciones fuera de ficha técnica (por similitud o difusas)

Este problema surge cuando los resultados de un estudio se aplican a poblaciones o contextos distintos de aquellos para los que fue diseñado y en los que no han sido evaluados. Una manifestación frecuente es la extrapolación indiscriminada de resultados globales a subgrupos sin evidencia específica, como ocurre al prescribir tratamientos en personas mayores cuando los estudios se han realizado en adultos jóvenes, o al indicarlos en mujeres embarazadas sin respaldo empírico. Esto ocurrió con el uso de antidepresivos inhibidores selectivos de la recaptación de serotonina en adolescentes, ilustrado por un ensayo publicado en 2001 que incluyó a 275 adolescentes con depresión y sin antecedentes de conducta suicida, en el que se compararon paroxetina, placebo e imipramina, el fármaco utilizado habitualmente en ese momento, tomando como variable de resultado la mejoría en una escala de depresión¹⁶. La compañía que fabricaba la paroxetina había ocultado datos en cuanto a efectos como comportamiento suicida u hostil que fueron más frecuentes en el grupo de la paroxetina y que se habían registrado como labilidad emocional, entre otros datos fraudulentos¹⁷⁻¹⁹. Años después, la paroxetina se había prescrito a millones de adolescentes en Estados Unidos y Canadá sin valorar si, además de la depresión, tenían antecedentes de conductas suicidas²⁰.

En cuanto a la ampliación de indicaciones por similitud o confusión, un ejemplo sería el uso de la gabapentina, cuya autorización inicial fue para la epilepsia y el dolor neuropático por herpes zoster²¹ y que se prescribió por la similitud de síntomas a otras neuralgias, radiculopatías y dolores inespecíficos como el dolor de espalda^{22,23}. Esto volvió a repetirse cuando salió su hermana, la pregabalina, con uso aprobado para indicaciones particularmente difusas como la fibromialgia, la ansiedad o el dolor neuropático, lo que condujo a

que fuese además prescrita de forma habitual para la lumbociática o el dolor de espalda, sin indicación en ficha técnica, y sin estudios de eficacia^{24,25}.

Sesgos en la difusión de resultados

El sesgo de publicación, también conocido como «el iceberg de los estudios negativos»⁸, hace referencia a la distorsión que se produce cuando los estudios con resultados negativos no se publican, lo que puede alterar la percepción de la eficacia de las intervenciones y conducir a conclusiones erróneas o a prácticas clínicas basadas en evidencia incompleta. Paralelamente, la presión por publicar ha provocado un aumento exponencial en el número de revistas científicas, lo que ha dado lugar a prácticas editoriales cuestionables y criterios de calidad deficientes, como la ausencia de revisiones por pares rigurosas, comprometiendo así la integridad de la investigación publicada. Asimismo, han surgido revistas cuyo objetivo comercial prima sobre el científico. Estas solicitan activamente manuscritos y cobran tarifas de publicación a los autores sin ofrecer procesos sólidos de revisión o edición, motivo por el cual se les ha denominado «depredadoras»²⁶.

Errores en la citación científica y riesgos emergentes con la inteligencia artificial

En las últimas décadas, la gestión de la información médica ha cambiado considerablemente, pero el sistema de referencias en las publicaciones médicas sigue siendo casi el mismo desde principios del siglo xx. Algunas investigaciones indican que hasta un 20% de las referencias bibliográficas son incorrectas, y muchos autores, de forma involuntaria, amplifican los errores de trabajos anteriores al citar fuentes no primarias²⁷. Esto provoca la difusión de afirmaciones y resultados incorrectos, lo que contribuye a la construcción de conocimientos inexactos. A ello se suma el uso acrítico de la información proporcionada por la inteligencia artificial, que en ocasiones incluso puede generar contenidos ficticios o inventar referencias cuando se le solicita. Estos errores subrayan la importancia de una metodología rigurosa y ética en la investigación clínica para garantizar resultados válidos y aplicables.

HEURÍSTICOS

En la práctica médica, las decisiones terapéuticas están frecuentemente condicionadas por heurísticos, atajos mentales que agilizan el razonamiento clínico, pero que, cuando se aplican sin la debida reflexión, pueden generar sesgos y errores. Entre los más habituales, se encuentran los heurísticos de anclaje, disponibilidad, confirmación y representatividad, junto con fenómenos como la sobreconfianza, el efecto arrastre, la familiaridad y el afecto. Todos ellos simplifican la toma de decisiones en contextos de presión y tiempo limitado, pero pueden llevar a mantener diagnósticos iniciales, pese a datos contradictorios, priorizar experiencias previas sobre la evidencia, aplicar protocolos sin considerar particularidades del paciente o filtrar la información que respalda la hipótesis inicial²⁸. Reconocer estos mecanismos mentales es esencial para promover un juicio clínico más equilibrado y centrado en el paciente, así como para mejorar la calidad de las decisiones en todos los niveles de gestión clínica.

Vamos a analizar en un ejemplo el camino recorrido de los datos proporcionados por la evidencia clínica a las recomendaciones de una guía de práctica clínica de gran aceptación con el fin de intentar identificar algún «atajo mental» o heurístico.

En las recomendaciones de las guías europeas de 2021 y 2023 sobre el manejo de la insuficiencia cardíaca con fracción de eyección preservada (ICFEP), se observa un cambio notable en el grado de interpretación favorable de la evidencia disponible. La guía de 2021 adoptaba una postura prudente, señalando que ningún tratamiento había demostrado de manera convincente reducir mortalidad o morbilidad en ICFEP y recordaba que ninguno de los grandes ensayos clínicos disponibles había alcanzado sus objetivos primarios²⁹. En contraste, la guía de 2023 introduce una recomendación clara a favor de los inhibidores del cotransportador sodio-glucosa tipo 2 (iSGLT2), en concreto empagliflozina y dapagliflozina, con clase de recomendación I y nivel de evidencia A, lo que implica un elevado consenso y un gran respaldo científico³⁰.

Revisaremos la evidencia que respalda ambas posturas, no para cuestionar la recomendación en sí, sino para comprender cómo se ha ido configurando la valoración de los datos y si existen diferencias en el grado de interpretación clínica positiva o en la forma de analizar resultados que, en algunos casos, son complejos o heterogéneos. Para ello, presentamos brevemente los hallazgos de dos ensayos relevantes: DELIVER, que evalúa dapagliflozina, y TOPCAT, centrado en espironolactona.

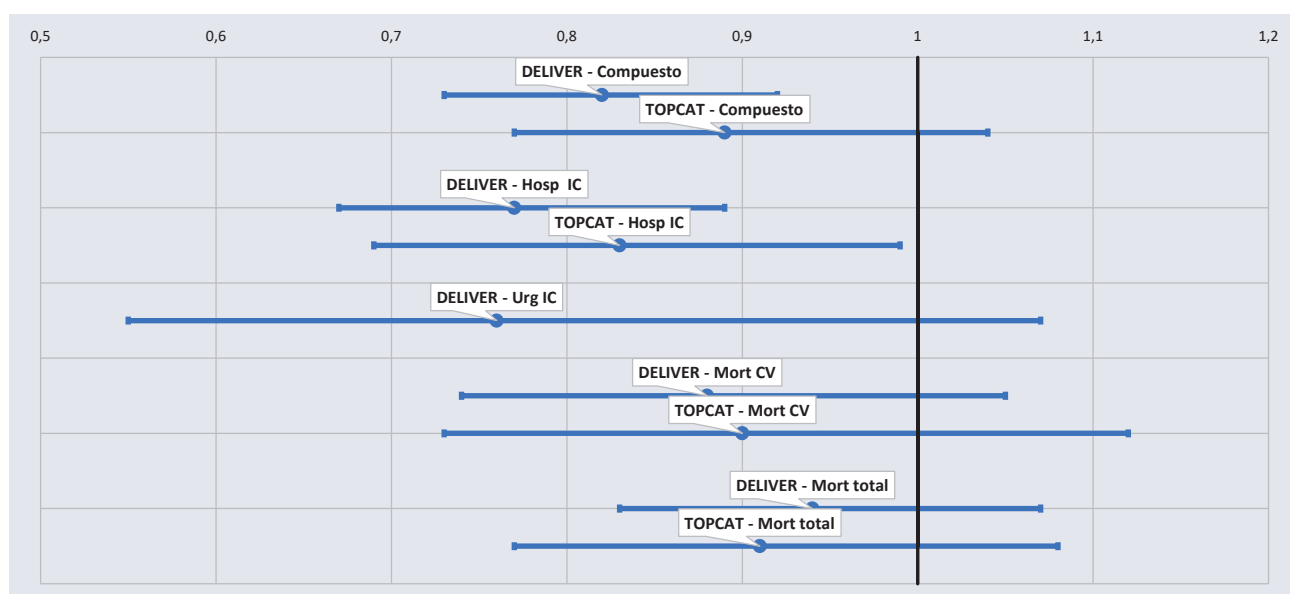
El ensayo DELIVER fue un estudio aleatorizado y doble ciego que incluyó 6.263 pacientes con insuficiencia cardíaca y fracción de eyección >40%, comparando dapagliflozina frente a placebo añadidos al tratamiento estándar. Su objetivo primario fue una variable combinada de empeoramiento de la insuficiencia cardíaca y muerte cardiovascular, evaluada tras una mediana de 2,3 años de segui-

miento. El estudio mostró una reducción relativa del riesgo en el objetivo primario (*hazard ratio* [HR]: 0,82), debida principalmente a una menor tasa de descompensaciones o ingresos por insuficiencia cardíaca. No se encontraron diferencias significativas en mortalidad cardiovascular ni en mortalidad total, de modo que el beneficio observado fue esencialmente la reducción de episodios de empeoramiento clínico (con ingreso)³¹. A pesar de lo limitado de estos resultados, el estudio señalaba: «Los resultados del ensayo DELIVER pueden orientar futuras guías y proporcionar más evidencia para su uso más amplio en la práctica clínica»³¹. En 2023, las guías europeas de insuficiencia cardíaca incorporaron la recomendación de usar iSGLT2, basándose en la reducción del criterio combinado observado en EMPEROR-Preserved³², DELIVER³¹ y un metanálisis³³; sin embargo, se precisa que el beneficio se limita a una «menor tasa de hospitalizaciones por insuficiencia cardíaca, sin haberse demostrado reducción significativa de la mortalidad cardiovascular»³⁰.

Compararemos ahora con la recomendación derivada de otros estudios, en concreto los utilizados en la guía de 2021 sobre la espironolactona con el estudio TOPCAT, para ver que quizá haya «atajos mentales» que guían la toma de decisiones.

El ensayo TOPCAT³⁴, aleatorizado y doble ciego, incluyó 3.445 pacientes con ICFEP (fracción de eyección del ventrículo izquierdo [FEVI] ≥45%) tratados con espironolactona o placebo durante una media de 3,3 años. El objetivo primario —muerte cardiovascular, paro cardíaco revertido u hospitalización por insuficiencia cardíaca— no alcanzó significación estadística (HR: 0,89; *p* = 0,14). Sí se observó una reducción modesta pero estadísticamente significativa en las hospitalizaciones por insuficiencia cardíaca, mientras que no hubo diferencias en mortalidad total ni en hospitalizaciones globales. En conjunto, los resultados individuales son menos distintos de los de DELIVER de lo que podría suponerse, como se muestra en la **figura 1**.

Figura 1. Comparación de los resultados de los estudios DELIVER y TOPCAT



Hosp IC: hospitalizaciones por insuficiencia cardíaca (IC); mort CV: mortalidad cardiovascular; mort total: mortalidad total; urg-IC: visitas a urgencias por IC. Se muestra la HR y el IC 95%.
Compuesto: end point compuesto en los estudios DELIVER y TOPCAT.
Datos extraídos de: Solomon SD³¹ y Pitt B³⁴.

Existen diferencias relevantes entre los estudios DELIVER y TOPCAT que van más allá de los resultados clínicos. El estudio DELIVER, financiado por una empresa farmacéutica, fue seguido rápidamente por un metaanálisis publicado en The Lancet que integró sus resultados con los de otros ensayos, concluyendo que los inhibidores de SGLT2 reducen muerte cardiovascular y hospitalizaciones por insuficiencia cardíaca en un amplio rango de pacientes³³. Aunque dicho metaanálisis se presentó como preespecificado, su registro en PROSPERO ocurrió tardíamente y con modificaciones posteriores, lo que genera dudas sobre ese extremo³⁵. En contraste, el estudio TOPCAT fue financiado por el National Heart, Lung, and Blood Institute, siguiendo un modelo de investigación financiada de forma pública e independiente de la industria farmacéutica.

Sin querer contradecir las recomendaciones de las guías europeas de 2023, podemos identificar al menos dos tipos de heurísticos en la traslación de los hallazgos a las guías.

Primero, la posible presencia de un heurístico de representatividad, mediante el cual se extrapolan los resultados de un ensayo a todos los pacientes asumiendo que el grupo estudiado es representativo de la población diana. En el DELIVER, aunque se preseleccionaron más de 10.000 pacientes, solo 6.263 fueron incluidos, conformando una muestra muy específica (mayores de 70 años, mayoritariamente caucásicos, predominantemente clase NYHA II, con fracción de eyección media del 54%, alta prevalencia de hipertensión y obesidad, y filtrado glomerular (FG) medio de 61 ml/min/1,73 m²). Los beneficios fueron más claros en ciertos subgrupos, como caucásicos, obesos, hipertensos y pacientes con FG <60 ml/min/1,73 m² (figura 2). No obstante, las guías clínicas aplican estas conclusiones de forma amplia a todos los pacientes con insuficiencia cardíaca y fracción de eyección preservada, incluso a aquellos con características muy distintas a las de la población incluida.

El razonamiento clínico puede verse influido, a su vez, por un heurístico de arrastre, que lleva a adoptar automáticamente las recomendaciones de ensayos influyentes como DELIVER cuando se difunden en revistas prestigiosas y guías internacionales. Esto amplifica el efecto del heurístico de representatividad: primero se extrapolan los resultados de una muestra específica a toda la población y luego la fuerza del consenso y de las guías consolida esa extrapolación, promoviendo que se aplique de forma sistemática sin valorar si los pacientes reales se parecen a los del estudio.

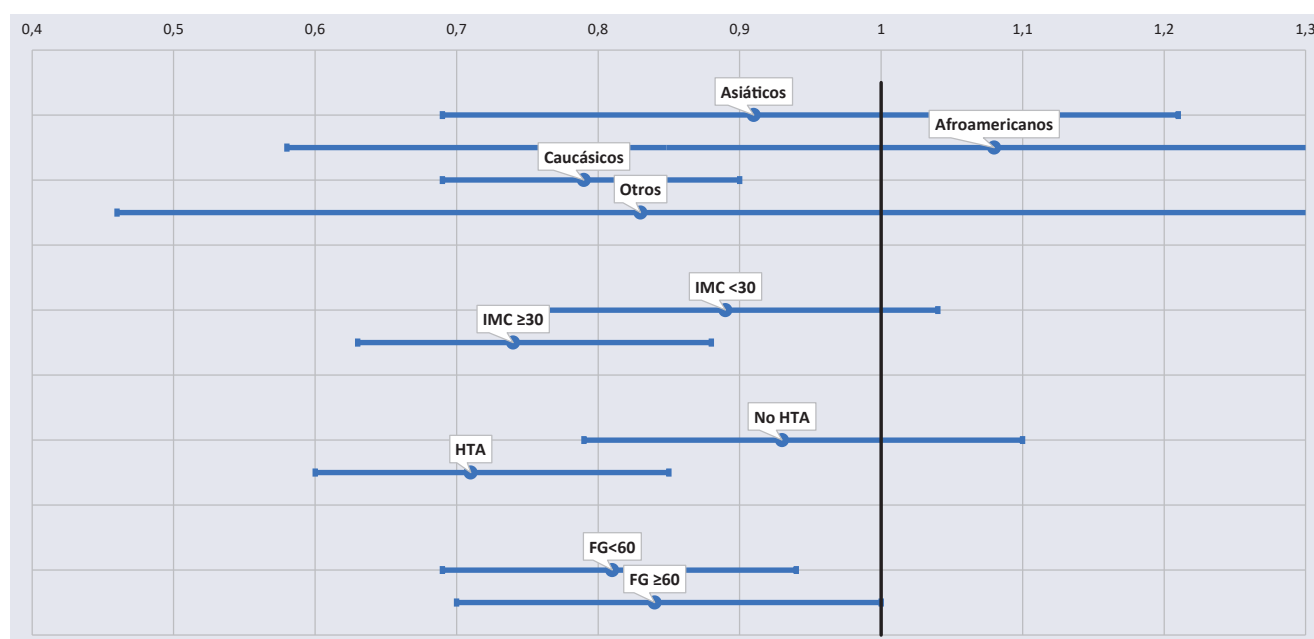
FRAUDE

El fraude en la investigación clínica constituye una amenaza significativa para la integridad científica y la seguridad de las y los pacientes, ya que puede distorsionar los resultados de los estudios y generar riesgos reales en la práctica médica. A pesar de los controles éticos y metodológicos existentes, los casos de mala conducta continúan apareciendo.

Diversas instituciones han definido el fraude o la mala conducta científica. El US Public Health Service lo describe como fabricación, falsificación o plagio, y excluye los errores honestos o diferencias de interpretación³⁶. El Medical Research Council (MRC) amplía esta definición al incluir la fabricación de datos o documentos, la falsificación mediante manipulación o selección indebida de información, el plagio de ideas o trabajos sin reconocimiento y la tergiversación de datos, resultados, intereses, experiencia o autoría, así como las violaciones del deber de cuidado y el manejo inadecuado de denuncias³⁷.

Por su parte, organismos como el COPE (Committee on Publication Ethics) y el ICMJE (International Committee of Medical Journal Editors) adoptan una visión más amplia que considera la mala con-

Figura 2. Análisis por subgrupos del estudio DELIVER



Se muestra la HR y el IC 95%.

FG: filtrado glomerular (< o ≥60 ml/min/1,73 m²); HTA: hipertensión arterial (en realidad HTA >128 mmHg); IMC: índice de masa corporal; Mort CV: mortalidad cardiovascular.

Datos extraídos de: Solomon SD³¹.

ducta como un continuo que va desde el error honesto hasta el fraude deliberado. Además, incluyen en su definición prácticas como la manipulación de imágenes, el ocultamiento de conflictos de intereses o la no publicación de resultados de ensayos clínicos³⁸. En conjunto, estas perspectivas muestran que la mala conducta científica abarca un espectro diverso que excede la definición clásica de fraude.

La magnitud del fraude

La verdadera magnitud del fraude científico es difícil de determinar, ya que quienes lo cometen rara vez lo admiten, los casos detectados representan solo una parte del total y los patrocinadores pueden tener incentivos para ocultarlo. A ello se suman problemas metodológicos, como la falta de consenso sobre qué conductas incluir o qué tipo de estudios evaluar. Las estimaciones varían ampliamente: mientras auditorías formales sitúan el fraude entre 0,01% y 0,4%, encuestas a científicos indican que hasta un 20% ha observado o sospecha mala conducta³⁹.

Un indicador indirecto es el número de artículos retractados, aunque este método tiene limitaciones. El crecimiento masivo de la producción científica dificulta el análisis, y algunos estudios muestran que las retractaciones por fraude han aumentado en las últimas décadas. Un análisis de 24,5 millones de publicaciones (1999-2022) observó que las tasas de retractación subieron hasta 2019 y variaban según disciplina, país y tipo de revista; durante la pandemia de la COVID-19, el aumento de publicaciones no se acompañó de mayores tasas de retractación en este campo. Las revistas de menor impacto concentran más retractaciones, lo que sugiere controles más estrictos en las de mayor prestigio⁴⁰.

Retraction Watch, una web especializada, registra actualmente más de 63.000 artículos retractados⁴¹, lo que supone el 0,02% de los 39 millones de referencias de la base Pubmed. Datos recientes muestran un aumento de retractaciones en España, siendo la duplicación el motivo más frecuente, mientras que en biomedicina predominan la fabricación y falsificación de datos⁴².

¿Cómo detectar el fraude?

Una forma común de detectar el fraude, al menos en un ensayo clínico es la revisión de datos individuales. En los ensayos clínicos aleatorizados, la asignación al azar debe distribuir de forma equilibrada todas las características de los participantes, de modo que cualquier diferencia basal pueda atribuirse al propio azar; cuando aparecen desequilibrios improbables o patrones incompatibles con una distribución aleatoria, se cuestiona la aleatorización y la validez del estudio. Una de las herramientas más utilizadas para detectar estas anomalías es el análisis estadístico propuesto por Carlisle, basado en examinar valores *p* de variables basales para identificar patrones extremadamente improbables. En 2012, J.B. Carlisle, un anestesiólogo, que fue editor de la revista *Anaesthesia*, publicó un análisis crítico sobre los ensayos clínicos de un solo autor, realizados entre 1991 y 2011. Carlisle analizó 168 trabajos de dicho investigador, comparando variables con las distribuciones esperadas. Encontró que el 85% de las variables (28 de 33) presentaban inconsistencias con probabilidades extremadamente bajas de ocurrencia por azar. Estos hallazgos presentaban patrones estadísticos

muy improbables, lo que sugería la posibilidad de manipulación o fabricación de datos⁴³.

Posteriormente, Carlisle aplicó un análisis similar a miles de ensayos clínicos aleatorizados de múltiples autores, publicados en revistas como *Anaesthesia*, *Journal of Anaesthesia*, *European Journal of Anaesthesiology*, *JAMA* y *New England Journal of Medicine*. Utilizó los valores *p* derivados de las comparaciones de las medias basales de los grupos para detectar anomalías, y encontró un exceso de valores difíciles de explicar por el azar, lo que indicaba patrones improbables que podrían reflejar manipulación o fraude. Estos patrones eran mucho más frecuentes en ensayos retractados que en los no retractados. Aunque algunas anomalías podrían explicarse por errores no intencionales o metodologías deficientes, los patrones fueron consistentes en todas las revistas analizadas. Este enfoque estadístico permite identificar ensayos potencialmente fraudulentos, facilitando su corrección o retractación⁴⁴.

Además, la revisión de datos individuales, empleada por el propio Carlisle en 2021, multiplicó la capacidad de identificar fraude y reveló que una proporción significativa de ensayos contenía datos falsificados⁴⁵.

Aunque el método de Carlisle presenta limitaciones, como la ausencia de corrección por comparaciones múltiples, lo que incrementa la probabilidad de falsos positivos⁴⁶, su enfoque ha sido reconocido por epidemiólogos de referencia, quienes destacan que detectar datos falsos en una parte del ensayo obliga a cuestionar la integridad del conjunto, incluidos procesos clave como la aleatorización, el cegamiento o la recogida de datos. Por ello, instan a financiadores, reguladores y sistemas académicos a promover la apertura y el rigor en la investigación mediante incentivos y sanciones para reducir la circulación de estudios fraudulentos⁴⁷.

Algunos ejemplos de fraude

Cada año, la Food and Drug Administration (FDA) inspecciona cientos de centros de ensayos clínicos y detecta desviaciones graves de las buenas prácticas. Clasifica el resultado de las inspecciones según su gravedad, siendo la «official action indicated» la más severa, que en 2013 representó el 2% de 644 inspecciones. Sin embargo, la falta de un mecanismo sistemático para difundir estos hallazgos hace que incluso irregularidades documentadas puedan pasar inadvertidas en la literatura científica. Una revisión realizada entre 1998 y 2013, encontró que 57 ensayos inspeccionados presentaban problemas importantes: falsificación (39%), mala notificación de eventos adversos (25%), violaciones de protocolo (74%), registros inexactos (61%) y fallos en la protección de los pacientes (53%), pero solo el 4% de los artículos que reportaban estos estudios mencionaron estos hechos, y ninguno fue corregido o retractado⁴⁸.

Tres episodios que describimos a continuación ilustran cómo la falta de transparencia y la manipulación de datos pueden afectar a la salud pública.

El caso del rofecoxib mostró cómo la empresa que lo patentó promovió rofecoxib como un antiinflamatorio no esteroideo (AINE) más seguro pese a señales tempranas de aumento del riesgo de eventos cardiovasculares. En el ensayo VIGOR, que comparaba ro-

fecoxib con naproxeno, a pesar de que en los criterios de exclusión constaban los eventos cardiovasculares previos, se vio un incremento de los infartos de miocardio que se justificó por un efecto protector del naproxeno, que no se había descrito hasta el momento⁴⁹. Pronto se supo que se omitieron eventos cardiovasculares en la publicación⁵⁰. El ensayo APPROVe confirmó en 2004 un riesgo duplicado de eventos cardiovasculares, y metanálisis posteriores demostraron que la toxicidad cardiovascular era evidente desde el año 2000⁵¹. Se estima que entre 89.000 y 139.000 estadounidenses sufrieron complicaciones cardiovasculares asociadas al rofecoxib, y hasta 88.000 infartos pudieron estar directamente relacionados con su uso, desencadenando demandas y multas millonarias⁵⁰. El caso marcó un punto de inflexión en la vigilancia de seguridad y en la exigencia de transparencia en la investigación clínica.

Los ensayos RECORD evaluaron la eficacia y seguridad de rivaroxabán, un anticoagulante oral, en la prevención del tromboembolismo venoso (TEV) tras cirugía ortopédica mayor, en comparación con enoxaparina. El RECORD 4, un ensayo fase III en pacientes sometidos a artroplastia total de rodilla, mostró que rivaroxabán oral 10 mg una vez al día durante 10-14 días era superior a enoxaparina subcutánea 30 mg cada 12 horas en la prevención del TEV (6,9% vs. 10,1%; $p = 0,01$), con tasas de sangrado mayor aparentemente similares (0,7% vs. 0,3%; $p = 0,11$)⁵². Pronto se señalaron discrepancias llamativas entre las tasas de sangrado reportadas y las observadas en otros ensayos, atribuidas a redefiniciones de «sangrado mayor» y a escasa potencia estadística⁵³. En 2011, la FDA consideró inaceptables ocho de los 16 centros inspeccionados y cuestionó la fiabilidad del ensayo debido a eliminación de registros, falsificación de datos y fallos en la aleatorización⁴⁷. Aun así, durante años estos problemas no aparecieron en las publicaciones. En 2022, el presidente del comité del programa RECORD reconoció que la FDA no aceptó RECORD 4 como evidencia pivotal y que existían aleatorizaciones incorrectas y subregistro de eventos graves, aunque omitió que estas anomalías ya se conocían desde 2011 y no se hacía ninguna reflexión sobre el conflicto de interés que había en este retraso⁵⁴.

Finalmente, el escándalo de valsartán evidenció cómo la manipulación de datos puede mantenerse oculta durante años. Entre 2002 y 2012, ensayos japoneses financiados indirectamente por la empresa fabricante informaron beneficios cardiovasculares del valsartán no atribuibles al control de la presión arterial^{55,56}.

Investigaciones posteriores revelaron que empleados de dicha empresa participaron en el tratamiento de los datos sin declarar su conflicto de interés y manipularon los resultados para favorecer al fármaco. Esto condujo a retractaciones múltiples, dimisiones académicas y pérdida de confianza pública. Aunque un tribunal reconoció la manipulación intencionada de los datos, determinó que no había delito, pues la publicación científica no constituye «publicidad» según la legislación japonesa⁵⁷.

Estos casos muestran cómo la integridad de los ensayos clínicos puede verse comprometida por fallos regulatorios, conflictos de intereses y manipulación activa de datos. También subrayan la necesidad de transparencia, acceso a datos individuales y vigilancia reforzada para proteger a los pacientes y garantizar la fiabilidad de la evidencia biomédica.

Las consecuencias del fraude

Las consecuencias del fraude científico son difíciles de mensurar. Cuando el fraude pasa inadvertido, la práctica clínica se sustenta sobre datos falsos y el perjuicio para el paciente es claro. Pero incluso cuando se descubre, su impacto puede persistir. Así lo muestra el estudio Delivery-I, que evaluó cómo los ensayos clínicos aleatorizados retractados afectan a las revisiones sistemáticas, metanálisis y guías clínicas. Se identificaron 1.330 ensayos clínicos aleatorizados (ECA) retractados hasta noviembre de 2024, y se buscaron 847 revisiones sistemáticas que los habían incluido en un total de 3.902 metanálisis. Al eliminar los estudios retractados, el 8,4% cambió la dirección del efecto, el 16% la significación estadística y el 3,9% ambas; en un 15,7% la magnitud del efecto varió más del 50%, sobre todo en metanálisis pequeños. Además, 68 revisiones afectadas habían sido utilizadas como base para 157 guías clínicas⁵⁸. Estos hallazgos demuestran que la investigación fraudulenta erosiona de manera sustancial el ecosistema de la evidencia, alterando síntesis de investigación, recomendaciones clínicas y el conjunto de la práctica médica basada en la evidencia, impactando de manera inevitable en la confianza de los profesionales en los resultados compartidos por la comunidad científica.

CONCLUSIONES

El proceso de toma de decisiones clínica es de por sí complejo y debería incorporar las preferencias del paciente, la experiencia clínica previa y la mejor evidencia disponible.

En este artículo hemos querido enfatizar que clínicos e investigadores en salud no debemos confiar de manera automática ni acrítica en conclusiones o recomendaciones derivadas de dicha investigación, sino considerarlas a la luz de problemas que dificultan la toma de decisiones basada en la evidencia, entre los que se incluyen los sesgos, los heurísticos y, de forma ocasional, el fraude, que pueden distorsionar la aplicación del conocimiento generado. Es fundamental comprender críticamente las suposiciones y retos subyacentes y preguntarnos desde la prudencia qué cambios supone el adoptar una nueva recomendación o tomar una decisión sin esta reflexión. Por último, queremos recordar la importancia de abogar por una ciencia basada en el rigor y la honestidad, y que desde la profesionalidad y la ética no permitamos que las actividades fraudulentas queden impunes o que sean rentables.

CONSIDERACIONES

Este artículo es una versión ampliada y revisada de una presentación en el XLV Congreso de la semFYC de Madrid 2025.

CONFLICTO DE INTERESES

Ambos autores afirman no presentar conflicto de intereses con lo expresado en el presente trabajo.

BIBLIOGRAFÍA

- Ioannidis JPA. Why Most Clinical Research Is Not Useful. *PLoS Med.* 2016;13(6):1-10.
- Smith R. Time to face up to research misconduct. *BMJ.* 1996 Mar 30;312(7034):789-90. doi: 10.1136/bmj.312.7034.789.
- Armond ACV, Gordijn B, Lewis J, Hosseini M, Bodnár JK, Holm S, et al. A scoping review of the literature featuring research ethics and research integrity cases. *BMC Med Ethics.* 2021 Apr 30;22(1):50. doi: 10.1186/s12910-021-00620-8.
- Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research—A review of common pitfalls. *Swiss Med Wkly.* 2007;137:44-9.
- Thiese MS, Walker S, Lindsey J. Truths, lies, and statistics. *J Thorac Dis.* 2017 Oct;9(10):4117-24. doi: 10.21037/jtd.2017.09.24.
- Ioannidis JPA. Why most published research findings are false. *PLoS Med.* 2005;2:e124.
- Clark AM, Sousa BJ, Ski CF, Thompson DR. Honest yet unacceptable research practices: when research becomes a health risk. *BMJ Open.* 2025 Jun 20;15(6):e097757. doi: 10.1136/bmjopen-2024-097757.
- Silva Aycaguer LC. Errores metodológicos frecuentes en la investigación clínica. *Med Intensiva.* 2018;42(9):541-6. doi: 10.1016/j.medint.2017.12.012.
- Carrasco-Labra A, Devji T, Qasim A, Phillips MR, Wang Y, Johnston BC, et al. Minimal important difference estimates for patient-reported outcomes: A systematic survey. *J Clin Epidemiol.* 2021 May;133:61-71. doi: 10.1016/j.jclinepi.2020.11.024.
- CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). CAPRIE Steering Committee. *Lancet.* 1996 Nov 16;348(9038):1329-39. doi: 10.1016/s0140-6736(96)09457-3.
- Martí-Villalta JL, Matías-Ghiu J. Antiagregación y anticoagulación en el ictus: incorporando la evidencia a la práctica clínica. *Rev Neurol.* 1999;29:767-88.
- Andrade C. Sample Size and its Importance in Research. *Indian J Psychol Med.* 2020 Jan 6;42(1):102-3. doi: 10.4103/IJPSYM.IJPSYM_504_19.
- Andrade C. Multiple testing and protection against a type 1 (false positive) error using the Bonferroni and Hochberg corrections. *Indian J Psychol Med.* 2019;41:99-100. doi: 10.4103/IJPSYM.IJPSYM_499_18.
- Sackett DL, Cook DJ. Can we learn anything from small trials? *Ann N Y Acad Sci.* 1993;703:25-31.
- Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol.* 2009 Feb;38(1):263-73. doi: 10.1093/ije/dyn147.
- Keller MB, Ryan ND, Strober M, Klein RG, Kutcher SP, Birmaher B, et al. Efficacy of paroxetine in the treatment of adolescent major depression: a randomized, controlled trial. *J Am Acad Child Adolesc Psychiatry* 2001;40:762-72. doi:10.1097/00004583-200107000-00010. PMID:11437014.
- Kondro W, Sibbald B. Drug company experts advised staff to withhold data about SSRI use in children. *CMAJ* 2004;170:783. doi:10.1503/cmaj.1040213.
- Dyer O. GlaxoSmithKline faces US lawsuit over concealment of trial results. *BMJ.* 2004 Jun 12;328(7453):1395. doi: 10.1136/bmj.328.7453.1395.
- Spurgeon D. GlaxoSmithKline staff told not to publicise ineffectiveness of its drug. *BMJ.* 2004 Feb 21;328(7437):422.
- Le Noury J, Nardo JM, Healy D, Jureidini J, Raven M, Tufanaru C, et al. Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. *BMJ.* 2015 Sep 16;351:h4320. doi: 10.1136/bmj.h4320.
- Peckham AM, Evoy KE, Ochs L, Covvey JR. Gabapentin for Off-Label Use: Evidence-Based or Cause for Concern? *Subst Abuse.* 2018 Sep 23;12:1178221818801311. doi: 10.1177/1178221818801311.
- Johansen ME. Gabapentinoid use in the United States 2002 through 2015. *JAMA Intern Med.* 2018;178(2):292-4.
- Goodman CW, Brett AS. A clinical overview of off-label use of gabapentinoid drugs. *JAMA Intern Med.* 2019;179(5):695-701.
- Evoy KE, Morrison MD, Saklad SR. Abuse and misuse of pregabalin and gabapentin. *Drugs.* 2017;77(4):403-26.
- Imani F, Rahimzadeh P. Gabapentinoids: gabapentin and pregabalin for postoperative pain management. *Anesth Pain Med.* 2012;2(2):52-53.
- Soriano JB, Ruano-Ravina A. The rising threat of predatory journals and paper mills in respiratory medicine and research. *Lancet Respir Med.* 2025 Jun;13(6):e30-e31. doi: 10.1016/S2213-2600(25)00117-1.
- Bareket R, Schonberg MA, Schonmann Y. Make quotations great again: a proposal to reduce false-knowledge. *BMJ Evid Based Med.* 2020 Feb;25(1):12-14. doi: 10.1136/bmjebm-2018-111109.
- Whelan DF, Conlon KC, Ridgway PF. Medicine and heuristics: cognitive biases and medical decision-making. *Ir J Med Sci.* 2020;189(4):1477-84.
- McDonagh TA, Metra M, Adamo M, Baumbach A, Böhm M, Burri H, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *European Heart Journal.* Oxford University Press. 2021;42:3599-726.
- McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2023 Focused Update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure Developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur Heart J.* 2023 Oct 1;44(37):3627-39.
- Solomon SD, McMurray JJV, Claggett B, De Boer RA, DeMets D, Hernandez AF, et al; DELIVER Trial Committees and Investigators. Dapagliflozin in Heart Failure with Mildly Reduced or Preserved Ejection Fraction. *N Engl J Med.* 2022 Sep 22;387(12):1089-98.
- Anker SD, Butler J, Filippatos G, Ferreira JP, Bocchi E, Böhm M, et al; EMPEROR-Preserved Trial Investigators. Empagliflozin in Heart Failure with a Preserved Ejection Fraction. *N Engl J Med.* 2021 Oct 14;385(16):1451-61.
- Vaduganathan M, Docherty KF, Claggett BL, Jhund PS, De Boer RA, Hernandez AF, et al. SGLT-2 inhibitors in patients with heart failure: a comprehensive meta-analysis of five randomised controlled trials. *Lancet.* 2022 Sep 3;400(10354):757-67. doi: 10.1016/S0140-6736(22)01429-5. Erratum in: *Lancet.* 2023 Jan 14;401(10371):104. doi: 10.1016/S0140-6736(23)00018-1.
- Pitt B, Pfeffer MA, Assmann SF, Boineau R, Anand IS, Claggett B, et al; TOPCAT Investigators. Spironolactone for heart failure with preserved ejection fraction. *N Engl J Med.* 2014 Apr 10;370(15):1383-92. doi: 10.1056/NEJMoa1313731.
- Vaduganathan M, Docherty KF, Claggett BL, Jhund PS, De Boer RA, Hernandez AF, et al. SGLT-2 Inhibitors in Patients with Heart Failure with Mildly Reduced or Preserved Ejection Fraction. PROSPERO 2024. [Internet]. Disponible en: <https://www.crd.york.ac.uk/PROSPERO/view/CRD42022327527>
- US Department of Health and Human Services. Office of Research Integrity, 2024 US Department of Health and Human Services. Office of Research Integrity. (2024). Part 93. Public Health Policies on Research Misconduct. [Internet]. Disponible en: <https://www.federalregister.gov/documents/2024/09/17/2024-20814/public-health-service-policies-on-research-misconduct>
- Rawle F, Parker R. MRC Policy and procedure for investigating allegations of research misconduct. 2014;1-25.
- Bauchner H, Steinbrook R, Redberg RF. Research Misconduct and Medical Journals. *J Law Med Ethics.* 2025 Mar 27;1-6. doi: 10.1017/jme.2025.35. Epub ahead of print.
- George SL, Buyse M. Data fraud in clinical trials. *Clin Investig (Lond).* 2015 Feb;5(2):161-73.
- Furuse Y. Characteristics of retracted research papers before and during the COVID-19 pandemic. *Front Med (Lausanne).* 2024 Jan 10;10:1288014. doi: 10.3389/fmed.2023.1288014.
- Retraction Watch. Retraction Watch Data. [Internet]. 2026 [cited 2026 Jan 4]. Disponible en: <https://gitlab.com/crossref/retraction-watch-data>
- Candal-Pedreira C, Ross JS, Rey-Brandariz J, Ruano-Ravina A. Retraction of publications in Spain: A retrospective analysis using the Retraction Watch database. *Med Clin (Barc).* 2024 Dec 27;163(12):626-33. English, Spanish. doi: 10.1016/j.medcli.2024.07.003
- Carlisle JB. The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia.* 2012 May;67(5):521-37. doi: 10.1111/j.1365-2044.2012.07128.x.
- Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia.* 2017 Aug;72(8):944-52. doi: 10.1111/anae.13938. Erratum in: *Anaesthesia.* 2018 Jul;73(7):915. doi: 10.1111/anae.14323. Erratum in: *Anaesthesia.* 2018 Sep;73(9):1176. doi: 10.1111/anae.14420.
- Carlisle JB. False individual patient data and zombie randomised controlled trials submitted to Anaesthesia. *Anaesthesia.* 2021 Apr;76(4):472-9. doi: 10.1111/anae.15263.
- Mascha EJ, Vetter TR, Pittet JF. An Appraisal of the Carlisle-Stouffer-Fisher Method for Assessing Study Data Integrity and Fraud. *Anesth Analg.* 2017;125(4):1381-5.
- Ioannidis JPA. Hundreds of thousands of zombie randomised trials circulate among us. *Anaesthesia.* 2021;76(4):444-7.
- Seife C. Research misconduct identified by the US Food and Drug Administration: Out of sight, out of mind, out of the peer-reviewed literature. *JAMA Intern Med.* 2015 Apr 1;175(4):567-77.
- Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, et al; VIGOR Study Group. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *N Engl J Med.* 2000 Nov 23;343(21):1520-8. 2 p following 1528. doi: 10.1056/NEJM200011233432103.
- McIntyre WF, Evans G. The Vioxx® legacy: Enduring lessons from the not so distant past. *Cadiol J.* 2014;21(2):203-5. doi: 10.5603/CJ.2014.0029.
- Jüni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet.* 2004 Dec 4-10;364(9450):2021-9. doi: 10.1016/S0140-6736(04)17514-4.

51. Turpie AG, Lassen MR, Davidson BL, Bauer KA, Gent M, Kwong LM, et al; RECORD4 Investigators. Rivaroxaban versus enoxaparin for thromboprophylaxis after total knee arthroplasty (RECORD4): a randomised trial. *Lancet*. 2009 May 16;373(9676):1673-80. doi: 10.1016/S0140-6736(09)60734-0. Erratum in: *Lancet*. 2022 Dec 10;400(10368):2048. doi: 10.1016/S0140-6736(22)02503-X.
52. Gómez-Outes A, Suárez-Gea ML, Blázquez-Pérez A, Pozo-Hernández C, Vargas-Castriellón E. Rivaroxaban versus enoxaparin after total knee arthroplasty. Vol. 374, *The Lancet*. *Lancet* 2009;374(9691):682; author reply 683. doi: 10.1016/S0140-6736(09)61552-X.
53. Turpie AGG. Revisiting RECORD4. *Lancet*. 2022 Dec 10;400(10368):2047-2048. doi: 10.1016/S0140-6736(22)02422-9. Erratum in: *Lancet*. 2022 Dec 10;400(10368):2048. doi: 10.1016/S0140-6736(22)02503-X.
54. Lancet Editors. Retraction--Valsartan in a Japanese population with hypertension and other cardiovascular disease (Jikei Heart Study): a randomised, open-label, blinded endpoint morbidity-mortality study. *Lancet*. 2013 Sep 7;382(9895):843. doi: 10.1016/S0140-6736(13)61847-4.
55. Sawada T, Yamada H, Dahlof B, Matsubara H. Effects of valsartan on morbidity and mortality in uncontrolled hypertensive patients with high cardiovascular risks: KYOTO HEART Study. *Eur Heart J*. 2009 Oct 2;30(20):2461-9.
56. Ozaki A, Harada K, Murayama A, Saito H, Sawano T, Tanimoto T, Shrestha S, et al. Japan's valsartan clinical trials shambles; time for fundamental changes. *Int J Health Plann Manage*. 2023 May;38(3):557-68. doi: 10.1002/hpm.3618.
57. Xu C, Fan S, Tian Y, Liu F, Furuya-Kanamori L, Clark J, et al; VITALITY Collaborative Research Network. Investigating the impact of trial retractions on the healthcare evidence ecosystem (VITALITY Study I): retrospective cohort study. *BMJ*. 2025 Apr 23;389:e082068. doi: 10.1136/bmj-2024-082068.