PAN AMERICAN JOURNAL OF PUBLIC HEALTH

Informe especial

Directrices para presentación de informes de ensayos clínicos sobre intervenciones con inteligencia artificial: extensión CONSORT-AI*

Xiaoxuan Liu,[1,2,3,4,5] *Samantha Cruz Rivera,*[5,6,7] *David Moher,*[8,9] *Melanie J. Calvert,*[4,5,6,7,10,11,12,] *Alastair K. Denniston,*[2,3,4,5,6,13] *y Grupo de Trabajo SPIRIT-AI y CONSORT-AI* [a]

Forma de citar Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK y Grupo de Trabajo SPIRIT-AI y CONSORT-AI et al. Directrices para presentación de informes de ensayos clínicos sobre intervenciones con inteligencia artificial: extensión CONSORT-AI Rev Panam Salud Publica. 2024;48:e13. https://doi.org/10.26633/RPSP.2024.13

RESUMEN

La declaración CONSORT 2010 proporciona unas directrices mínimas para informar sobre los ensayos clínicos aleatorizados. Su uso generalizado ha sido fundamental para garantizar la transparencia en la evaluación de nuevas intervenciones. Más recientemente, se ha reconocido cada vez más que las intervenciones con inteligencia artificial (IA) deben someterse a una evaluación rigurosa y prospectiva para demostrar su impacto en la salud. La extensión CONSORT-AI (Consolidated Standards of Reporting Trials-Artificial Intelligence) es una nueva pauta de información para los ensayos clínicos que evalúan intervenciones con un componente de IA, esta se desarrolló en paralelo con su declaración complementaria para los protocolos de ensayos clínicos: SPIRIT-AI (Standard Protocol Items – Artificial Intelligence: Recomendaciones para ensayos clínicos de intervención - Inteligencia Artificial). Ambas directrices se desarrollaron a través de un proceso de consenso por etapas que incluía la revisión de la literatura y la consulta a expertos para generar 29 elementos candidatos, que fueron evaluados por un grupo internacional de múltiples partes interesadas en una encuesta Delphi de dos etapas (103 partes interesadas congregados en una reunión de consenso de dos días (31 partes interesadas) y refinados a través de una lista de verificación piloto (34 participantes). La ampliación del CONSORT-AI incluye 14 nuevos elementos que se consideraron lo suficientemente importantes para las intervenciones de IA como para que se informen de forma rutinaria, además de los elementos básicos del CONSORT 2010. CONSORT-AI recomienda que los investigadores proporcionen descripciones claras de la intervención de IA, incluyendo las instrucciones y las habilidades requeridas para su uso, el entorno en el que se integra la intervención de IA, el manejo de los datos de entrada y los datos de salida de la intervención de IA, la interacción entre el ser humano y la IA y la provisión de un análisis de los casos de error. CONSORT-AI ayudará a promover la transparencia y la exhaustividad en los informes de los ensayos clínicos de las intervenciones de AI, también ayudará a los editores y revisores, así como a los lectores en general, a entender, interpretar y valorar críticamente la calidad del diseño del ensayo clínico y el riesgo de sesgo en los resultados comunicados.

[1] Moorfields Eye Hospital NHS Foundation Trust, Londres, Reino Unido.
[2] Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, Reino Unido.
[3] University Hospitals Birmingham NHS Foundation Trust, Birmingham, Reino Unido.
[4] Health Data Research Reino Unido, Londres, Reino Unido.
[5] Birmingham Health Partners Centre for Regulatory Science and Innovation, University of Birmingham, Birmingham, Reino Unido.
[6] Centre for Patient Reported Outcomes Research, Institute of Applied Health Research, University of Birmingham, Birmingham.
[7] Institute of Applied Health Research, University of Birmingham, Birmingham,Reino Unido.
[8] Centre for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canadá.
[9] School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Canada.
[10] National Institute of Health Research Birmingham Biomedical Research Centre, University of Birmingham and University Hospitals Birmingham NHS Foundation Trust, Birmingham, Reino Unido.
[11] National Institute of Health Research Applied Research Collaborative West Midlands, Coventry, Reino Unido.
[12] National Institute of Health Research Surgical Reconstruction and Microbiology Centre, University of Birmingham and University Hospitals Birmingham NHS Foundation Trust, Birmingham, Reino Unido.
[13] NIHR Biomedical Research Center at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, Londres, Reino Unido.
[a] La lista de autores y sus afiliaciones aparecen al final del artículo. ✉ Alastair K. Denniston, a.denniston@bham.ac.uk

* Traducción al español efectuada por el comité de Inteligencia artificial de la Asociación Colombiana de Radiología (ACR), Medellín, Colombia y aprobada por los autores. En caso de discrepancia, prevalecerá la versión original en inglés publicada en Nat Med. 2020;26:1364-1374. https://doi.org/10.1038/s41591-020-1034-x

Los ensayos clínicos controlados aleatorizados (ECA) se consideran el diseño experimental de referencia para proporcionar pruebas de la seguridad y la eficacia de una intervención[1, 2]. Los resultados de los ensayos clínicos, si se comunican adecuadamente, tienen el potencial de informar sobre las decisiones reguladoras, las directrices clínicas y la política sanitaria. Por lo tanto, es crucial que los ECA se comuniquen con transparencia y exhaustividad para que los lectores puedan valorar críticamente los métodos y los resultados del ensayo clínico y evaluar la presencia de sesgos en los resultados[3-5].

La declaración CONSORT proporciona recomendaciones basadas en la evidencia para mejorar la exhaustividad de los informes de los ECA. La declaración se introdujo por primera vez en 1996 y desde entonces ha sido ampliamente respaldada por las revistas médicas a nivel internacional[5]. En las últimas dos décadas, ha sido objeto de dos actualizaciones y ha demostrado un impacto positivo sustancial en la calidad de los informes de los ECA[6, 7]. La declaración CONSORT 2010 más reciente proporciona una lista de control de 25 elementos del contenido mínimo de los informes aplicable a todos los ECA, pero reconoce que ciertas intervenciones pueden requerir la ampliación o aclaración de estos elementos. Existen varias extensiones de este tipo[8-13].

La IA es un área de enorme interés con fuertes impulsos para acelerar las nuevas intervenciones desde su publicación, aplicación hasta su comercialización[14]. Aunque los sistemas de IA se han investigado durante algún tiempo, los recientes avances en el aprendizaje profundo y las redes neuronales han ganado un considerable interés por su potencial en las aplicaciones sanitarias. Los ejemplos de estas aplicaciones son muy variados e incluyen sistemas de IA para el cribado y el triaje[15, 16], el diagnóstico[17-20], el pronóstico[21, 22], el apoyo a la toma de decisiones[23] y la recomendación de tratamientos[24]. Sin embargo, en los casos más recientes, las pruebas publicadas han consistido en una

RECUADRO 1 Glosario

Inteligencia artificial	Ciencia que desarrolla sistemas informáticos que pueden realizar tareas que normalmente requieren inteligencia humana.
Intervención de IA	Intervención en salud que se basa en un componente de IA/ML para cumplir su propósito.
CONSORT	Normas consolidadas de reporte de ensayos clínicos.
Elemento de ampliación de CONSORT-AI	Elemento adicional de la lista de verificación para abordar el contenido específico de la IA que no está adecuadamente cubierto por CONSORT 2010.
Mapa de activación de clases	Los mapas de activación de clases son particularmente relevantes para las intervenciones de IA de clasificación de imágenes. Los mapas de activación de clases son visualizaciones de los píxeles que tuvieron mayor influencia en la clase predicha, mostrando el gradiente del resultado predicho por el modelo con respecto a la entrada. También se denominan "mapas de saliencia" o "mapas de calor".
Resultado de salud	Variables medidas en el ensayo clínico que se utilizan para evaluar los efectos de una intervención.
Interacción humano-inteligencia	El proceso de cómo los usuarios (humanos) interactúan con la intervención de IA, para que ésta funcione como se pretende.
Resultado clínico	Variables medidas en el ensayo clínico que se utilizan para evaluar los efectos de una intervención.
Estudio Delphi	Método de investigación que obtiene las opiniones colectivas de un grupo mediante una consulta escalonada de encuestas, cuestionarios o entrevistas, con el objetivo de alcanzar un consenso al final.
Entorno de desarrollo	Entorno clínico y operativo en el que se generan los datos utilizados para el entrenamiento del modelo. Esto incluye todos los aspectos del entorno físico (como la ubicación geográfica, el entorno físico), el entorno operativo (como la integración con un sistema de registro electrónico, la instalación en un dispositivo físico) y el entorno clínico (como la atención primaria, secundaria y/o terciaria, el espectro de enfermedades del paciente).
Ajuste (Fine-Tuning)	Modificaciones o entrenamientos adicionales realizados en el modelo de intervención de la IA, con la intención de mejorar su rendimiento.
Datos de entrada	Datos que deben presentarse a la intervención de IA para que pueda cumplir su objetivo.
Aprendizaje automático	Campo de las ciencias de la computación que se ocupa del desarrollo de modelos/algoritmos que pueden resolver tareas específicas mediante el aprendizaje de patrones a partir de datos, en lugar de seguir reglas explícitas. Se considera un enfoque dentro del campo de la IA.
Entorno operativo	Entorno en el que se desplegará la intervención de IA, incluida la infraestructura necesaria para que la intervención de IA funcione.
Datos de salida	Resultado previsto por la intervención de IA basado en el modelado de los datos de entrada. Los datos de salida pueden presentarse de diferentes formas, incluida una clasificación (que incluye el diagnóstico, la gravedad o el estadio de la enfermedad, o una recomendación como la derivación), una probabilidad, un mapa de activación de clases, etc. Los datos de salida suelen proporcionar información clínica adicional y/o desencadenar una decisión clínica.
Error de desempeño	Instancias en las que la intervención de la IA no funciona como se esperaba. Este término puede describir diferentes tipos de fallos, y corresponde al investigador especificar lo que debe considerarse un error de rendimiento, preferiblemente basado en pruebas previas. Esto puede ir desde pequeñas disminuciones en la precisión (en comparación con la precisión esperada) hasta predicciones erróneas o la incapacidad de producir una salida, en ciertos casos.
Elementos del protocolo estándar SPIRIT:	Recomendaciones para los ensayos clínicos de intervención.
SPIRIT-AI	Un ítem adicional de la lista de verificación para abordar el contenido específico de la IA que no está adecuadamente cubierto por SPIRIT 2013.
Elemento de aclaración de SPIRIT-AI	Consideraciones adicionales a un elemento existente de SPIRIT 2013 cuando se aplica a las intervenciones de IA.

validación *in silico* en fase inicial. Se ha reconocido que la mayoría de los estudios recientes de IA están informados de forma inadecuada y las directrices de información existentes no cubren completamente las fuentes potenciales de sesgo específicas de los sistemas de IA[25]. La aparición de ECAs que buscan evaluar intervenciones más nuevas basadas en, o que incluyen, un componente de IA (denominadas aquí "intervenciones de IA")[23, 26-31] se ha encontrado igualmente con preocupaciones sobre el diseño y la presentación de informes[25, 32-34]. Esto ha puesto de manifiesto la necesidad de proporcionar una guía para la presentación de informes que sea "adecuada para el propósito" en este ámbito.

CONSORT-AI (como parte de la iniciativa SPIRIT-AI y CONSORT-AI) es una iniciativa internacional apoyada por CONSORT y la red EQUATOR (Enhancing the Quality and Transparency of Health Research. Mejorando la calidad y transparencia de la investigación de la salud, por sus siglas en inglés) que busca evaluar la declaración CONSORT 2010 existente y ampliar o aclarar puntos en esta guía cuando sea necesario, para apoyar la presentación de informes de ensayos clínicos para intervenciones de IA[35, 36]. Es complementaria a la declaración SPIRIT-AI, cuyo objetivo es promover reportes de protocolos de alta calidad para los ensayos clínicos de IA. Esta Declaración de Consenso describe los métodos utilizados para identificar y evaluar los elementos candidatos y obtener un consenso. Además, también proporciona la lista de verificación CONSORT-AI, que incluye los nuevos elementos de ampliación y las explicaciones que los acompañan.

MÉTODOS

Las extensiones de SPIRIT-AI y CONSORT-AI se desarrollaron simultáneamente para los protocolos e informes de ensayos clínicos. En octubre de 2019 se publicó un anuncio para la iniciativa SPIRIT-AI y CONSORT-AI (ref. 35), y las dos directrices se registraron como directrices de reporte en desarrollo en la biblioteca de directrices de reporte de EQUATOR en mayo de 2019. Ambas directrices se desarrollaron de acuerdo con el marco metodológico de EQUATOR Network37. El Grupo Directivo de SPIRIT-AI y CONSORT-AI, compuesto por 15 expertos internacionales, se formó para supervisar la realización y la metodología del estudio. Las definiciones de los términos clave se incluyen en el glosario (recuadro 1).

Aprobación ética

Este estudio fue aprobado por el comité de revisión ética de la Universidad de Birmingham, Reino Unido (ERN_19-1100). La información de los participantes en el Delphi se proporcionó por vía electrónica antes de completar la encuesta y antes de la reunión de consenso. Los participantes en el Delphi dieron su consentimiento informado por vía electrónica, y se obtuvo el consentimiento por escrito de los participantes en la reunión de consenso.

Revisión de la literatura y generación de ítems candidatos

Se generó una lista inicial de ítems candidatos para las listas de verificación SPIRIT-AI y CONSORT-AI mediante la revisión de la literatura publicada y la consulta con el Grupo Directivo y expertos internacionales conocidos. Se realizó una búsqueda el 13 de mayo de 2019 utilizando los términos "artificial intelligence", "machine learning" y "deep learning" para identificar los ensayos clínicos existentes para las intervenciones de IA que figuran dentro del registro de ensayos clínicos de la Biblioteca Nacional de Medicina de los Estados Unidos (ClinicalTrials.gov). Había 316 ensayos clínicos registrados, de los cuales 62 se habían completado y 7 habían publicado resultados[30, 38-43]. Dos estudios se informaron con referencia a la declaración CONSORT[30, 42], y un estudio proporcionó un protocolo de ensayo clínico no publicado[42]. El equipo de operaciones (X.L., S.C.R., M.J.C. y A.K.D.) identificó las consideraciones específicas de la IA de estos estudios y las reformuló como elementos de información candidatos. Los elementos candidatos también se basaron en los resultados de una revisión sistemática anterior que evaluó la precisión diagnóstica de los sistemas de aprendizaje profundo para la imagen médica[25]. Después de consultar con el Grupo Directivo y otros expertos internacionales (n = 19), se generaron 29 ítems candidatos, 26 de los cuales eran relevantes tanto para SPIRIT-AI como para CONSORT-AI y 3 de los cuales eran relevantes sólo para CONSORT-AI. El equipo de operaciones asignó estos ítems a los ítems correspondientes de SPIRIT y CONSORT, revisando la redacción y proporcionando el texto explicativo necesario para contextualizar los ítems. Estos ítems se incluyeron en las encuestas Delphi posteriores.

Proceso de consenso Delphi

En septiembre de 2019, se invitó a 169 expertos internacionales clave a participar en la encuesta Delphi en línea para votar sobre los elementos candidatos y sugerir elementos adicionales. Los expertos fueron identificados y contactados a través del Grupo Directivo y se les permitió una ronda de reclutamiento "bola de nieve" en la que los expertos contactados podían sugerir expertos adicionales. Además, se incluyó a las personas que se pusieron en contacto tras la publicación del anuncio[35]. El Grupo Directivo acordó que las personas con experiencia en ensayos clínicos y en IA y aprendizaje automático (ML), así como los usuarios clave de la tecnología, deberían estar bien representados en la consulta. Entre las partes interesadas se encontraban profesionales sanitarios, metodólogos, estadísticos, informáticos, representantes de la industria, editores de revistas, responsables políticos, "informáticos" sanitarios, expertos en derecho y ética, reguladores, pacientes y financiadores. Las características de los participantes se describen en la tabla suplementaria 1. Se realizaron dos encuestas Delphi en línea. Se utilizó el software DelphiManager (versión 4.0), desarrollado y mantenido por la iniciativa COMET (Core Outcome Measures in Effectiveness Trials), para realizar la encuesta e-Delphi. Los participantes recibieron información escrita sobre el estudio y se les pidió que indicaran su nivel de experiencia en los campos de (i) IA/ML, y (ii) ensayos clínicos. Se presentó cada punto para su consideración (26 para SPIRIT-AI y 29 para CONSORT-AI). Se pidió a los participantes que votaran sobre cada punto utilizando una escala de 9 puntos, de la siguiente manera 1-3, no importante; 4-6, importante pero no crítico; y 7-9, importante y crítico. Los encuestados proporcionaron calificaciones separadas para SPIRIT-AI y CONSORT-AI. Había una opción para no votar por cada ítem, y cada ítem incluía un espacio para comentarios de texto libre. Al final de la encuesta Delphi, los participantes tuvieron la oportunidad de sugerir nuevos elementos. Se recibieron 103 respuestas para la primera ronda Delphi, y 91 respuestas

CUADRO 1. Lista de control CONSORT-AI

Sección	Ítem de CONSORT 2010[a]		Item CONSORT-AI	Abordado en el número de página[b]
Título y resumen				
Título y resumen	1a	Identificación como ensayo clínico aleatorio en el título	CONSORT-AI 1a, b Aclaración	(i) Indicar que la intervención implica inteligencia artificial/aprendizaje automático en el título y/o el resumen y especificar el tipo de modelo.
	1b	Resumen estructurado del diseño, los métodos, los resultados y las conclusiones del ensayo clínico (para una orientación específica, véase el CONSORT para resúmenes)		(ii) Indique el uso previsto de la intervención de IA dentro del ensayo clínico en el título y/o el resumen.
Introducción				
Antecedentes y objetivos	2a	Antecedentes científicos y explicación de la justificación	CONSORT-AI 2a (i) Extensión	Explicar el uso previsto de la intervención de IA en el contexto de la vía clínica, incluyendo su propósito y sus usuarios previstos (por ejemplo, profesionales sanitarios, pacientes, público).
	2b	Objetivos o hipótesis específicos		
Métodos				
Diseño del ensayo clínico	3a	Descripción del diseño del ensayo clínico (por ejemplo, paralelo, factorial), incluida la proporción de asignación		
	3b	Cambios importantes en los métodos tras el inicio del ensayo clínico (como los criterios de elegibilidad), con las razones		
Participantes	4a	Criterios de elegibilidad de los participantes	CONSORT-AI 4a (i) Aclaración	Indique los criterios de inclusión y exclusión a nivel de los participantes.
			CONSORT-AI 4a (ii) Extensión	Indique los criterios de inclusión y exclusión a nivel de los datos de entrada.
	4b	Entornos y lugares donde se recogieron los datos	CONSORT-AI 4b Extensión	Describa cómo se integró la intervención de IA en el entorno del ensayo clínico, incluyendo cualquier requisito in situ o externo.
Intervenciones	5	Las intervenciones para cada grupo con detalles suficientes para permitir la replicación, incluyendo cómo y cuándo se administraron realmente	CONSORT-AI 5 (i) Extensión	Indique qué versión del algoritmo de IA se utilizó.
			CONSORT-AI 5 (ii) Extensión	Describa cómo se adquirieron y seleccionaron los datos de entrada para la intervención de IA.
			CONSORT-AI 5 (iii) Extensión	Describa cómo se evaluaron y manejaron los datos de entrada de baja calidad o no disponibles.
			CONSORT-AI 5 (iv) Extensión	Especifique si hubo interacción entre el ser humano y la IA en el manejo de los datos de entrada, y qué nivel de experiencia se requirió de los usuarios.
			CONSORT-AI 5 (v) Extensión	Especifique el resultado de la intervención de la IA
			CONSORT-AI 5 (vi) Extensión	Explique cómo los resultados de la intervención de IA contribuyeron a la toma de decisiones u otros elementos de la práctica clínica.
Resultados	6a	Medidas de resultado primarias y secundarias completamente definidas, incluyendo cómo y cuándo fueron evaluadas		
	6b	Cualquier cambio en los resultados del ensayo clínico después de su inicio, con las razones		

(Continuará)

CUADRO 1. (Cont.)

Sección	Ítem de CONSORT 2010[a]		Item CONSORT-AI	Abordado en el número de página[b]
Tamaño de la muestra	7a	Cómo se determinó el tamaño de la muestra		
	7b	Cuando proceda, explicación de los análisis intermedios y de las directrices de interrupción		
Aleatorización				
Generación de secuencias	8a	Método utilizado para generar la secuencia de asignación aleatoria		
	8b	Tipo de aleatorización; detalles de cualquier restricción (como el bloqueo y el tamaño del bloque)		
Mecanismo de ocultación de la asignación	9	Mecanismo utilizado para aplicar la secuencia de asignación aleatoria (como contenedores numerados secuencialmente), describiendo cualquier medida adoptada para ocultar la secuencia hasta la asignación de las intervenciones		
Implementación	10	Quién generó la secuencia de asignación aleatoria, quién inscribió a los participantes y quién los asignó a las intervenciones		
Cegamiento	11a	Si se hizo, quién fue cegado después de la asignación a las intervenciones (por ejemplo, los participantes, los proveedores de atención, los que evalúan los resultados) y cómo		
	11b	Si procede, descripción de la similitud de las intervenciones		
Métodos estadísticos	12a	Métodos estadísticos utilizados para comparar los grupos para los resultados primarios y secundarios		
	12b	Métodos para los análisis adicionales, como los análisis de subgrupos y los análisis ajustados		
Resultados				
Flujo de participantes (se recomienda encarecidamente un diagrama)	13a	Para cada grupo, el número de participantes que fueron asignados al azar, recibieron el tratamiento previsto y fueron analizados para el resultado primario		
	13b	Para cada grupo, pérdidas y exclusiones después de la aleatorización, junto con las razones		
Reclutamiento	14a	Fechas que definen los períodos de reclutamiento y seguimiento		
	14b	Razones por las que el ensayo clínico terminó o se detuvo		
Datos de referencia	15	Tabla con las características demográficas y clínicas de partida de cada grupo		
Números analizados	16	Para cada grupo, número de participantes (denominador) incluidos en cada análisis y si el análisis fue por grupos asignados originalmente		
Resultados y estimación	17a	Para cada resultado primario y secundario, resultados para cada grupo, y el tamaño del efecto estimado y su precisión (como el intervalo de confianza del 95%)		
	17b	Para los resultados binarios, se recomienda la presentación de los tamaños del efecto tanto absolutos como relativos		

(Continuará)

CUADRO 1. (Cont.)

Sección		Ítem de CONSORT 2010[a]	Item CONSORT-AI		Abordado en el número de página[b]
Análisis auxiliares	18	Resultados de cualquier otro análisis realizado, incluidos los análisis de subgrupos y los análisis ajustados, distinguiendo los preespecificados de los exploratorios			
Daños	19	Todos los daños importantes o efectos no deseados en cada grupo (para una orientación específica, véase CONSORT para los daños)	Extensión CONSORT-AI 19	Describa los resultados de cualquier análisis de los errores de ejecución y cómo se identificaron los errores, si procede. Si no se planificó o realizó dicho análisis, justifique por qué no.	
Discusión					
Limitaciones	20	Limitaciones del ensayo clínico, abordando las fuentes de sesgo potencial, imprecisión y, si es relevante, la multiplicidad de análisis			
Generalizabilidad	21	Generalización (validez externa, aplicabilidad) de los resultados del ensayo			
Interpretación	22	Interpretación coherente con los resultados, equilibrando los beneficios y los daños, y teniendo en cuenta otras pruebas pertinentes			
Otra información					
Registro	23	Número de registro y nombre del registro del ensayo clínico			
Protocolo	24	Dónde se puede acceder al protocolo completo del ensayo clínico, si está disponible			
Financiación	25	Fuentes de financiación y otras ayudas (como el suministro de medicamentos), papel de los financiadores	CONSORT-AI 25 Extensión	Indique si se puede acceder a la intervención de IA y/o a su código, y cómo, incluyendo cualquier restricción al acceso o a la reutilización.	

[a]Recomendamos encarecidamente la lectura de esta declaración junto con la Explicación y Aclaración del CONSORT 2010 para obtener aclaraciones importantes sobre todos los puntos. [b]Indica los números de página que deben completar los autores durante el desarrollo del protocolo.

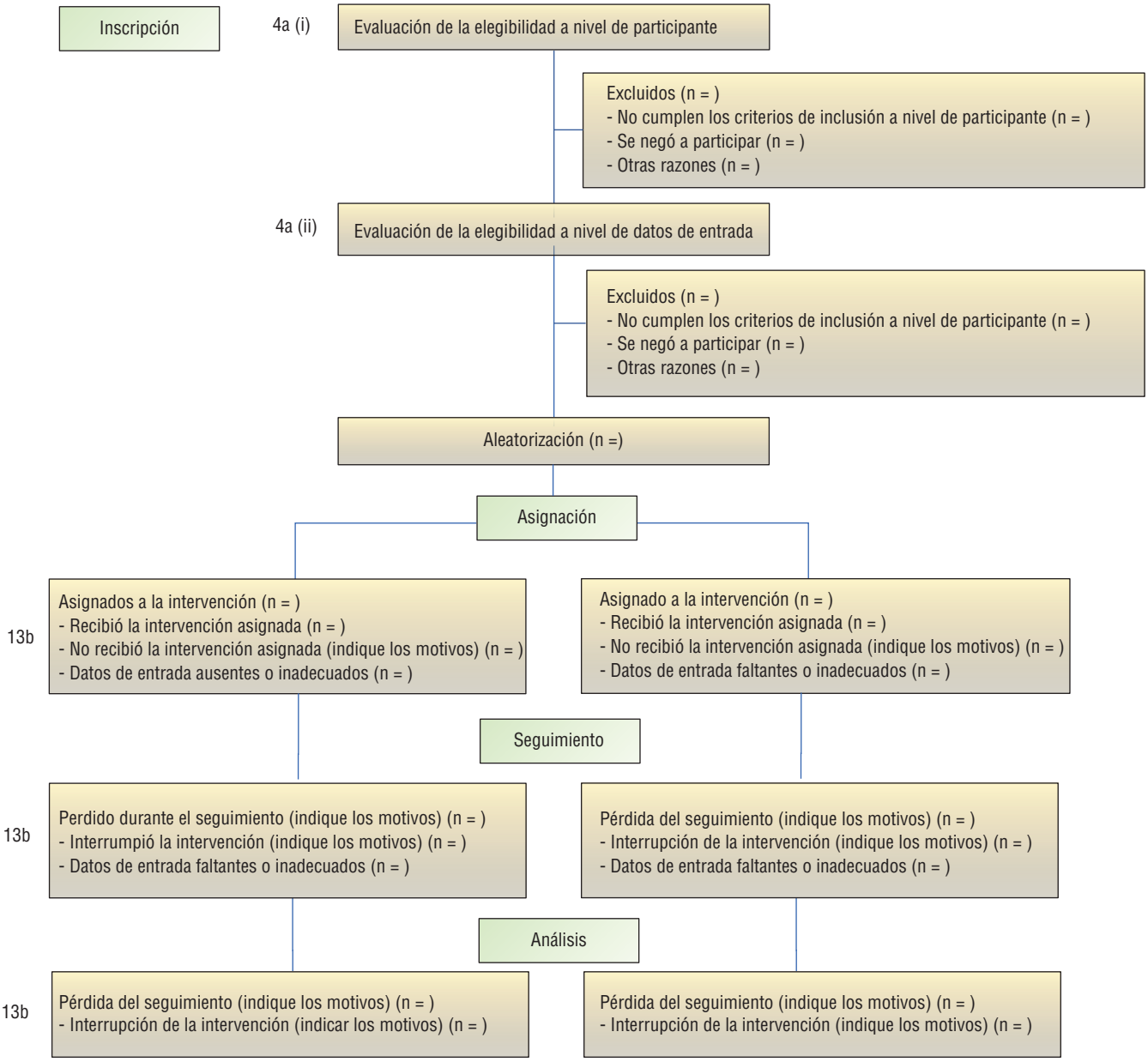
(el 88% de los participantes de la primera ronda) para la segunda ronda. Los resultados de la encuesta Delphi sirvieron de base para la posterior reunión de consenso internacional. Los participantes en el estudio Delphi propusieron 12 nuevos puntos que se añadieron para su discusión en la reunión de consenso. Los datos recogidos durante la encuesta Delphi se anonimizaron, y los resultados a nivel de ítems se presentaron en la reunión de consenso para su discusión y votación.

La reunión de consenso, de dos días de duración, tuvo lugar en enero de 2020 y fue organizada por la Universidad de Birmingham, Reino Unido, para buscar el consenso sobre el contenido de SPIRIT-AI y CONSORT-AI. Se invitó a 31 interesados internacionales de entre los participantes en la encuesta Delphi a debatir los puntos y votar sobre su inclusión. Los participantes se seleccionaron para lograr una representación adecuada de todos los grupos interesados. Se discutieron 41 ítems, que incluían los 29 ítems generados en la revisión inicial de la literatura y la fase de generación de ítems (26 ítems relevantes tanto para SPIRIT-AI como para CONSORT-AI; 3 ítems relevantes sólo para CONSORT-AI) y los 12 nuevos ítems propuestos por los participantes durante las encuestas Delphi. Cada ítem se presentó al grupo de consenso, junto con su puntuación en el ejercicio Delphi (mediana y rangos Inter cuartiles) y cualquier comentario realizado por los participantes en el Delphi relacionado con ese ítem. Se invitó a los participantes en la reunión de consenso a comentar la importancia de cada elemento y si éste debía incluirse en la ampliación de la IA. Además, se invitó a los participantes a comentar la redacción del texto explicativo que acompañaba a cada ítem y la posición de cada ítem en relación con las listas de verificación SPIRIT 2013 y CONSORT 2010. Tras la discusión abierta de cada ítem y la opción de ajustar la redacción, se realizó una votación electrónica, con la opción de incluir o excluir el ítem. Se preespecificó un umbral del 80% para la inclusión, que el Grupo Directivo consideró razonable para demostrar el consenso de la mayoría. Cada parte interesada votó de forma anónima utilizando los cuadernos de votación *Turning Point* (*Turning Technologies*, versión 8.7.2.14).

Lista de comprobación piloto

Tras la reunión de consenso, los asistentes tuvieron la oportunidad de hacer comentarios finales sobre la redacción y acordar que los ítems actualizados de SPIRIT-AI y CONSORT-AI reflejaran las discusiones de la reunión. El equipo de operaciones asignó cada elemento como extensión o aclaración sobre la base de un árbol de decisiones y elaboró un penúltimo borrador de las listas de comprobación de SPIRIT-AI y CONSORT-AI (figura 1 suplementaria). Se realizó una prueba piloto de las penúltimas listas de verificación con 34 participantes para garantizar

FIGURA 1. Diagrama de flujo de CONSORT 2010 - adaptado para los ensayos clínicos de IA



CONSORT-AI 4a (i): indique los criterios de inclusión y exclusión a nivel de participantes. CONSORT-AI 4a (ii): indicar los criterios de inclusión y exclusión a nivel de los datos de entrada. CONSORT 13b (elemento central de CONSORT): Para cada grupo, las pérdidas y exclusiones después de la aleatorización, junto con las razones

la claridad de la redacción. Los expertos que participaron en el piloto fueron los siguientes (a) participantes en el Delphi que no asistieron a la reunión de consenso, y (b) expertos externos que no habían participado en el proceso de desarrollo pero que se habían puesto en contacto con el Grupo Directivo después de que comenzara el estudio Delphi. El equipo de operaciones introdujo los últimos cambios en la redacción, únicamente para mejorar la claridad para los lectores (figura 2 suplementaria).

Recomendaciones

Elementos de la lista de verificación CONSORT-AI y explicación. La extensión del CONSORT AI recomienda que se añadan 14 nuevos ítems de la lista de verificación a la declaración CONSORT 2010 existente (11 extensiones y 3 aclaraciones). Estos ítems se consideraron lo suficientemente importantes para los informes de ensayos clínicos de intervenciones de IA como para que se informen de forma rutinaria, además de los ítems de la lista de comprobación básica del CONSORT 2010. El cuadro 1 enumera los ítems de CONSORT-AI.

Los 14 ítems siguientes superaron el umbral del 80% para su inclusión en la reunión de consenso. CONSORT-AI 2a, CONSORT-AI 5 (ii) y CONSORT-AI 19 fueron el resultado de la fusión de dos ítems tras la discusión con el grupo de consenso. CONSORT-AI 4a (i) y (ii) se dividió en dos puntos para mayor claridad y se votó por separado. CONSORT-AI 5(iii) no cumplía

los criterios de inclusión sobre la base de su redacción inicial (77% de votos a favor de su inclusión); sin embargo, tras un amplio debate y una nueva redacción, el grupo de consenso apoyó unánimemente una nueva votación, momento en el que superó el umbral de inclusión (97% a favor de su inclusión). Los resultados del Delphi y de la votación para cada elemento incluido y excluido se describen en el cuadro suplementario 2.

Título y resumen

CONSORT-AI 1a, b (i) Aclaración: indique que la intervención implica inteligencia artificial/aprendizaje automático en el título y/o el resumen y especifique el tipo de modelo. *Explicación*. Se recomienda indicar en el título y/o el resumen del informe del ensayo clínico que la intervención implica una forma de IA, ya que identifica inmediatamente la intervención como una intervención de IA/ML y también sirve para facilitar la indexación y la búsqueda del informe del ensayo clínico. El título debe ser comprensible para un público amplio; por lo tanto, se recomienda utilizar un término general más amplio, como "inteligencia artificial" o "aprendizaje automático". Los términos más precisos deben utilizarse en el resumen, en lugar del título, a menos que se reconozca ampliamente que son una forma de IA/ML. La terminología específica relacionada con el tipo de modelo y la arquitectura debe detallarse en el resumen.

CONSORT-AI 1a,b (ii) Aclaración: Indique el uso previsto de la intervención de IA dentro del ensayo clínico en el título y/o el resumen. Explicación. Describa el uso previsto de la intervención de IA en el título y/o el resumen del informe del ensayo clínico. Esto debe describir el propósito de la intervención de IA y el contexto de la enfermedad[26, 44]. Algunas intervenciones de IA pueden tener múltiples usos previstos, o el uso previsto puede evolucionar con el tiempo. Por lo tanto, documentar esto permite a los lectores comprender el uso previsto del algoritmo en el momento del ensayo.

INTRODUCCIÓN

CONSORT-AI 2a (i) Extensión: explique el uso previsto para la intervención de IA en el contexto de la vía clínica, incluyendo su propósito y sus usuarios previstos (por ejemplo, profesionales sanitarios, pacientes, público). *Explicación*. Para aclarar cómo se pretende que la intervención de IA encaje en un flujo de trabajo clínico, debe incluirse una descripción detallada de su función en los antecedentes del informe del ensayo clínico. Las intervenciones de IA pueden estar diseñadas para interactuar con diferentes usuarios, incluidos los profesionales de la salud, los pacientes y el público, y sus funciones pueden ser muy variadas (por ejemplo, la misma intervención de IA podría teóricamente sustituir, aumentar o adjudicar componentes de la toma de decisiones clínicas). Aclarar el uso previsto de la intervención de IA y su usuario previsto ayuda a los lectores a comprender el propósito para el que se evaluó la intervención de IA en el ensayo.

MÉTODOS

CONSORT-AI 4a (i) Aclaración: indique los criterios de inclusión y exclusión a nivel de participantes. *Explicación*. Los criterios de inclusión y exclusión deben definirse a nivel de los participantes, según la práctica habitual en los informes de ensayos de intervención no relacionados con la IA (Fig. 1). Esto es distinto de los criterios de inclusión y exclusión realizados a nivel de los datos de entrada, que se abordan en el punto 4a (ii).

CONSORT-AI 4ª Extensión (ii): indique los criterios de inclusión y exclusión a nivel de los datos de entrada. *Explicación*. Los "datos de entrada" se refieren a los datos requeridos por la intervención de IA para cumplir con su propósito (por ejemplo, para un sistema de diagnóstico de cáncer de mama, los datos de entrada podrían ser la mamografía no procesada o el postprocesamiento específico del proveedor sobre el que se realiza el diagnóstico; para un sistema de alerta temprana, los datos de entrada podrían ser las mediciones fisiológicas o los resultados de laboratorio de la historia clínica electrónica). El informe del ensayo debe especificar previamente si existen requisitos mínimos para los datos de entrada (como la resolución de la imagen, las métricas de calidad o el formato de los datos) que determinen la elegibilidad previa a la aleatorización. Debe especificar cuándo, cómo y quién lo evaluó. Por ejemplo, si un participante cumplía los criterios de elegibilidad para estar tumbado para una TC según el punto 4a (i), pero la calidad de la exploración estaba comprometida (por cualquier razón) a tal nivel que se consideraba inadecuada para su uso por el sistema de IA, esto debería ser informado como un criterio de exclusión a nivel de datos de entrada. Nótese que cuando los datos de entrada se adquieren después de la aleatorización, cualquier exclusión se considera del análisis, no de la inscripción (punto 13b de CONSORT y Fig. 1).

CONSORT-AI 4b Extensión: describa cómo se integró la intervención de IA en el entorno del ensayo, incluyendo cualquier requisito in situ o externo. *Explicación*. La generalización de los algoritmos de IA tiene limitaciones, una de las cuales es cuando se utilizan fuera de su entorno de desarrollo 45,46. Los sistemas de IA dependen de su entorno operativo, y el informe debe proporcionar detalles de los requisitos de hardware y software para permitir la integración técnica de la intervención de IA en cada centro de estudio. Por ejemplo, debe indicarse si la intervención de IA requirió dispositivos específicos del proveedor, si hubo hardware informático especializado en cada centro, o si el centro tuvo que soportar la integración en la nube, particularmente si esto fue específico del proveedor. Si fue necesario realizar algún cambio en el algoritmo en cada centro de estudio como parte del procedimiento de implementación (como el ajuste fino del algoritmo en los datos locales), este proceso también debe describirse claramente.

CONSORT-AI 5 (i) Extensión: indique qué versión del algoritmo de IA se utilizó. *Explicación*. Al igual que otras formas de software como dispositivo médico, es probable que los sistemas de IA sufran múltiples iteraciones y actualizaciones durante su vida útil. Por lo tanto, es importante especificar qué versión del sistema de IA se utilizó en el ensayo clínico, si es la misma que la versión evaluada en estudios anteriores que se han utilizado para justificar la justificación del estudio, y si la versión cambió durante la realización del ensayo. Si es el caso, el informe debe describir lo que ha cambiado entre las versiones pertinentes y la justificación de los cambios. Cuando esté disponible, el informe debe incluir una referencia de marcado reglamentario, como un identificador único de dispositivo, que requiere un nuevo identificador para las versiones actualizadas del dispositivo[47].

CONSORT-AI 5 (ii) Extensión: describa cómo se adquirieron y seleccionaron los datos de entrada para la intervención de IA. *Explicación*. El desempeño de cualquier sistema de IA puede depender críticamente de la naturaleza y la calidad de

los datos de entrada[48]. Debe proporcionarse una descripción del tratamiento de los datos de entrada, incluida la adquisición, la selección y el preprocesamiento antes del análisis por parte del sistema de IA. La exhaustividad y la transparencia de esta descripción son esenciales para la replicabilidad de la intervención más allá del ensayo clínico en el mundo real. También ayuda a los lectores a identificar si los procedimientos de manejo de datos de entrada se estandarizaron en todos los centros del ensayo clínico.

CONSORT-AI 5 (iii) Extensión: describa cómo se evaluaron y manejaron los datos de entrada de mala calidad o no disponibles. *Explicación*. Al igual que en CONSORT-AI 4a (ii), los "datos de entrada" se refieren a los datos requeridos por la intervención de IA para cumplir su propósito. Como se discute en el punto 4a (ii), el rendimiento de los sistemas de IA puede verse comprometido como resultado de la mala calidad o la falta de datos de entrada[49] (por ejemplo, un artefacto de movimiento excesivo en un electrocardiograma). El informe del ensayo clínico debe informar de la cantidad de datos que faltan, así como de la forma en que se identificaron y gestionaron. El informe también debe especificar si había un estándar mínimo requerido para los datos de entrada y, cuando no se alcanzó este estándar, cómo se manejó esto (incluyendo el impacto en, o cualquier cambio en, la vía de atención del participante).

Los datos de mala calidad o no disponibles también pueden afectar a las intervenciones no relacionadas con la IA. Por ejemplo, una calidad inferior a la óptima de una exploración podría afectar a la capacidad de un radiólogo para interpretarla y realizar un diagnóstico. Por lo tanto, es importante que esta información se comunique igualmente en la intervención de control, cuando sea pertinente. Si esta norma de calidad mínima fuera diferente de los criterios de inclusión de los datos de entrada utilizados para evaluar la elegibilidad antes de la aleatorización, debería indicarse.

CONSORT-AI 5 (iv) Extensión: especifique si hubo interacción entre humanos y la IA en el manejo de los datos de entrada, y qué nivel de experiencia se requirió de los usuarios. *Explicación*. Debe proporcionarse una descripción de la interfaz persona-IC y de los requisitos para una interacción satisfactoria cuando se manejan datos de entrada; por ejemplo, la selección por parte del clínico de regiones de interés de un portaobjetos de histología que luego interpreta un sistema de diagnóstico de IA[50], o la selección por parte de un endoscopista de un vídeo de colonoscopia como datos de entrada para un algoritmo diseñado para detectar pólipos[28]. La descripción de la formación del usuario y las instrucciones sobre cómo deben manejar los datos de entrada aportan transparencia y posibilidad de reproducción de los procedimientos del ensayo clínico. La falta de claridad en la interfaz persona-IC puede conducir a la falta de un enfoque estándar y puede tener implicaciones éticas, especialmente en caso de daño[51, 52]. Por ejemplo, puede no estar claro si un caso de error se produjo debido a una desviación humana del procedimiento instruido, o si fue un error cometido por el sistema de IA.

CONSORT-AI 5 (v) Extensión: especificar el resultado de la intervención de IA. *Explicación*. El resultado de la intervención de IA debe especificarse claramente en el informe del ensayo clínico. Por ejemplo, un sistema de IA puede producir una clasificación o probabilidad diagnóstica, una acción recomendada, una alarma que alerte sobre un evento, una acción instigada en un sistema de bucle cerrado (como la titulación de infusiones de medicamentos) u otra salida. La naturaleza del resultado de la intervención de la IA tiene implicaciones directas en su capacidad de uso y en cómo puede conducir a acciones y resultados posteriores.

CONSORT-AI 5 (vi) Extensión: expliqué cómo los resultados de la intervención de IA contribuyeron a la toma de decisiones u otros elementos de la práctica clínica. *Explicación*. Dado que los resultados de salud también pueden depender críticamente de la forma en que los seres humanos interactúan con la intervención de IA, el informe debe explicar cómo se utilizaron los resultados del sistema de IA para contribuir a la toma de decisiones u otros elementos de la práctica clínica. Esto debe incluir una descripción adecuada de las intervenciones posteriores que pueden afectar a los resultados. Al igual que en CONSORT-AI 5 (iv), cualquier efecto de la interacción entre el ser humano y la IA en los resultados debe describirse en detalle, incluyendo el nivel de experiencia requerido para entender los resultados y cualquier formación y/o instrucciones proporcionadas para este fin. Por ejemplo, un sistema de detección de cáncer de piel que produzca un porcentaje de probabilidad como resultado debe ir acompañado de una explicación de cómo se interpretó este resultado y cómo actuó el usuario, especificando tanto las vías previstas (por ejemplo, la escisión de la lesión cutánea si el diagnóstico es positivo) como los umbrales para entrar en estas vías (por ejemplo, la escisión de la lesión cutánea si el diagnóstico es positivo y la probabilidad es >80%). La información producida por las intervenciones de comparación debe describirse de forma similar, junto con una explicación de cómo se utilizó dicha información para llegar a las decisiones clínicas sobre el tratamiento del paciente, cuando sea pertinente. Debe informarse de cualquier discrepancia en la forma en que se tomó la decisión en comparación con la forma en que se pretendía que se tomara (es decir, según lo especificado en el protocolo del ensayo).

RESULTADOS

CONSORT-AI 19 Extensión: describa los resultados de cualquier análisis de los errores de ejecución y cómo se identificaron los errores, cuando corresponda. Si no se planificó o se hizo tal análisis, explique por qué no. *Explicación*. Informar sobre los errores de rendimiento y el análisis de los casos de fracaso es especialmente importante para las intervenciones de IA. Los sistemas de IA pueden cometer errores que pueden ser difíciles de prever pero que, si se permite su despliegue a escala, podrían tener consecuencias catastróficas[53]. Por lo tanto, la reporte de casos de error y la definición de estrategias de mitigación de riesgos son importantes para informar sobre cuándo, y para qué poblaciones, puede aplicarse la intervención con seguridad. Los resultados de cualquier análisis de errores de funcionamiento deben ser comunicados y las implicaciones de los resultados deben ser discutidas.

Otra información

CONSORT-AI 25 Extensión: indique si se puede acceder a la intervención de IA y/o a su código y cómo, incluyendo cualquier restricción de acceso o reutilización. *Explicación*. El informe del ensayo clínico debe aclarar si se puede acceder a la intervención de IA y/o a su código o reutilizarlos, y cómo

hacerlo. Debe incluir detalles sobre la licencia y cualquier restricción de acceso.

DISCUSIÓN

CONSORT-AI es una nueva extensión de las directrices de reporte desarrollada mediante un consenso internacional de múltiples partes interesadas. Su objetivo es promover el reporte transparente de los ensayos de intervenciones de IA y pretende facilitar la evaluación crítica y la síntesis de la evidencia. Los ítems de la extensión añadidos en CONSORT-AI abordan una serie de cuestiones específicas de la implementación y evaluación de las intervenciones de IA, que deben considerarse junto con la lista de verificación básica de CONSORT 2010 y otras extensiones de CONSORT[54]. Es importante tener en cuenta que se trata de requisitos mínimos y que puede ser útil incluir elementos adicionales no incluidos en las listas de verificación en el informe o en los materiales suplementarios (Cuadro suplementario 2).

Tanto en CONSORT-AI como en su proyecto complementario, SPIRIT-AI, se hizo hincapié en la adición de varios ítems nuevos relacionados con la propia intervención y su aplicación en el contexto clínico. Los puntos 5 (i)-5 (vi) se añadieron para abordar las consideraciones específicas de la IA en las descripciones de la intervención. Se hicieron recomendaciones específicas relativas a los sistemas de IA en relación con la versión del algoritmo, los datos de entrada y salida, la integración en los entornos de los ensayos clínicos, la experiencia de los usuarios y el protocolo para actuar según las recomendaciones del sistema de IA. Se acordó que estos detalles son críticos para la evaluación independiente o la replicación del ensayo clínico. Los editores de las revistas informaron de que, a pesar de la importancia de estos elementos, en la actualidad suelen faltar en los informes de los ensayos en el momento de su presentación para la publicación, lo que da más peso a su inclusión como elementos de extensión específicamente enumerados.

Un tema recurrente en los comentarios del Delphi y en el debate del grupo de consenso fue la seguridad de los sistemas de IA. Esto se debe al reconocimiento de que los sistemas de IA, a diferencia de otras intervenciones sanitarias, pueden producir errores impredecibles que no son fácilmente detectables o explicables por el juicio humano. Por ejemplo, los cambios en las imágenes médicas que son invisibles, o parecen aleatorios, para el ojo humano pueden cambiar por completo la probabilidad del resultado del diagnóstico[55, 56]. La preocupación es que, dada la facilidad teórica con la que los sistemas de IA podrían desplegarse a escala, cualquier consecuencia perjudicial no intencionada podría ser catastrófica. El punto 19 de CONSORT-AI, que requiere la especificación de cualquier plan para analizar los errores de rendimiento, se añadió para enfatizar la importancia de anticipar los errores sistemáticos cometidos por el algoritmo y sus consecuencias. Más allá de esto, también se debe alentar a los investigadores a explorar las diferencias en las tasas de rendimiento y error entre los subgrupos de población. Se ha demostrado que los sistemas de IA pueden estar sistemáticamente sesgados hacia diferentes resultados, lo que puede conducir a un tratamiento diferente o incluso injusto, sobre la base de las características existentes[53, 57-59].

El tema de los sistemas de IA de "evolución continua" (también conocidos como sistemas de IA de "adaptación continua" o "aprendizaje continuo") se debatió ampliamente durante la reunión de consenso, pero se acordó excluirlo del CONSORT-AI. Se trata de sistemas de IA con capacidad para entrenarse continuamente con nuevos datos, lo que puede provocar cambios en el rendimiento a lo largo del tiempo. El grupo señaló que, si bien es interesante, este campo se encuentra en una fase relativamente temprana de su desarrollo, sin ejemplos tangibles en aplicaciones sanitarias, y que no sería apropiado que se incluyera en CONSORT- IA en esta fase[60]. Este tema será objeto de seguimiento y se revisará en futuras iteraciones de CONSORT-AI. Cabe señalar que los cambios incrementales del software ya sean continuos o iterativos, intencionados o no, podrían tener graves consecuencias en el rendimiento de la seguridad después de su despliegue. Por lo tanto, es de vital importancia que dichos cambios se documenten e identifiquen por versión de software y que se establezca un sólido plan de vigilancia posterior al despliegue.

Este estudio se enmarca en el contexto actual de la IA en la salud, por lo que hay que señalar varias limitaciones. En primer lugar, hay relativamente pocos ensayos de intervención publicados en el campo de la IA para la salud; por lo tanto, los debates y las decisiones tomadas durante este estudio no siempre se apoyaron en los ejemplos existentes de ensayos clínicos completados. Esto se debe a nuestro objetivo declarado de abordar los problemas de información deficiente en este campo tan pronto como sea posible, reconociendo los fuertes impulsores en el campo y los desafíos específicos del diseño de estudios y la presentación de informes para la IA. A medida que la ciencia y el estudio de la IA evolucionan, acogemos con agrado la colaboración con los investigadores para coevolucionar estas normas de reporte y garantizar su continua relevancia. En segundo lugar, la búsqueda bibliográfica de ECA de IA utilizó terminología como "inteligencia artificial", "aprendizaje automático" y "aprendizaje profundo", pero no términos como "sistemas de apoyo a la decisión clínica" o "sistemas expertos", que se utilizaron más comúnmente en la década de 1990 para las tecnologías respaldadas por sistemas de IA y comparten riesgos similares a los de los ejemplos recientes[61]. Es probable que tales sistemas, si se publicaran hoy en día, se indexarían bajo "inteligencia artificial" o "aprendizaje automático"; sin embargo, los sistemas de apoyo a la decisión clínica no se discutieron activamente durante este proceso de consenso. En tercer lugar, la lista inicial de elementos candidatos fue generada por un grupo relativamente pequeño de expertos formado por los miembros del grupo de dirección y otros expertos internacionales; sin embargo, los elementos adicionales del grupo Delphi más amplio se sometieron a la consideración del grupo de consenso, y no se sugirieron nuevos elementos durante la reunión de consenso o la evaluación posterior a la reunión. Al igual que la declaración CONSORT, la extensión CONSORT-AI pretende ser una guía mínima para la presentación de informes, y hay consideraciones adicionales específicas de la IA para los informes de los ensayos que pueden justificar su consideración (Cuadro suplementario 2). Esta extensión está dirigida particularmente a los investigadores y lectores que informan o evalúan ensayos clínicos; sin embargo, también puede servir como guía útil para los desarrolladores de intervenciones de IA en las primeras etapas de validación de un sistema de IA. Los investigadores que deseen informar sobre estudios que desarrollen y validen las propiedades diagnósticas y predictivas de los modelos de IA deben remitirse a TRIPOD-ML (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis

or Diagnosis-Machine Learning, Informes transparentes de un modelo de predicción multivariable para el pronóstico individual o el aprendizaje de la máquina de diagnóstico) y a STARD-AI (Standards for Reporting Diagnostic Accuracy Studies-Artificial Intelligence), ambas en fase de desarrollo[32, 62]. Otras directrices potencialmente relevantes, que son agnósticas al diseño del estudio, están registradas en la red EQUATOR[63]. Se espera que la ampliación del CONSORT-AI fomente la planeación temprana y cuidadosa de las intervenciones de IA para los ensayos clínicos y esto, junto con el SPIRIT-AI, debería ayudar a mejorar la calidad de los ensayos para las intervenciones de IA. El desarrollo de la guía CONSORT-AI no incluye elementos adicionales dentro de la sección de discusión de los informes de los ensayos. Se consideró que las orientaciones proporcionadas por el CONSORT 2010 sobre las limitaciones, la generalización y la interpretación de los ensayos eran trasladables a los ensayos de intervenciones de IA.

También se reconoce que la IA es un campo que evoluciona rápidamente y que será necesario actualizar el CONSORT-AI a medida que se desarrolle la tecnología y sus nuevas aplicaciones. En la actualidad, la mayoría de las aplicaciones de la IA implican la detección, el diagnóstico y el triaje de enfermedades, y es probable que esto haya influido en la naturaleza y el orden de prioridad de los elementos del CONSORT-AI. A medida que surjan aplicaciones más amplias que utilicen la "IA como terapia", será importante seguir evaluando el CONSORT-AI a la luz de dichos estudios. Además, los avances en las técnicas computacionales y la capacidad de integrarlas en los flujos de trabajo clínicos aportarán nuevas oportunidades de innovación que beneficien a los pacientes. Sin embargo, pueden ir acompañados de nuevos retos en torno al diseño de los estudios y la presentación de informes. Con el fin de garantizar la transparencia, minimizar los posibles sesgos y promover la fiabilidad de los resultados y el grado en que pueden ser generalizables, el Grupo Directivo de SPIRIT-AI y CONSORT-AI seguirá vigilando la necesidad de actualizaciones.

Disponibilidad de los datos. La solicitudes de datos deben dirigirse al autor correspondiente y su publicación estará sujeta a la consideración del Grupo Directivo de SPIRIT-AI y CONSORT-AI.

Contribuciones de los autores. Concepto y diseño, y adquisición, análisis e interpretación de los datos, todos los autores; redacción del manuscrito, X.L., S.C.R., D.M., M.J.C. y A.K.D.; obtención de financiación, M.J.C., C.Y., C.H. y A.K.D. El grupo de trabajo de SPIRIT-AI y CONSORT-AI está formado por dos grupos que han sido clave en el desarrollo de las directrices: el grupo directivo de SPIRIT-AI y CONSORT-AI, que se encargó de supervisar el proceso de consenso y la metodología de desarrollo de las directrices (Alastair K. Denniston, An-Wen Chan, Ara Darzi, Christopher Holmes, Christopher Yau, David Moher, Hutan Ashrafian, Jonathan J. Deeks, Lavinia Ferrante di Ruffano, Livia Faes, Melanie J. Calvert, Pearse A. Keane, Samantha Cruz Rivera, Sebastian J. Vollmer y Xiaoxuan Liu); y el Grupo de Consenso de SPIRIT-AI y CONSORT-AI, que se encargó de llegar a un consenso sobre el contenido y la redacción de los elementos de las listas de verificación (Aaron Y. Lee, Adrian Jonas, Andre Esteva, Andrew L. Beam, An-Wen Chan, Maria Beatrice Panico, Cecilia S. Lee, Charlotte Haug, Christopher J. Kelly, Christopher Yau, Cynthia Mulrow, Cyrus Espinoza, David Moher, Dina Paltoo, Elaine Manna, Gary Price, Gary S Collins, Hugh Harvey, James Matcham, Joao Monteiro, John Fletcher, M. Khair ElZarrad, Lavinia Ferrante Di Ruffano, Luke Oakden-Rayner, Melanie J. Calvert, Melissa McCradden, Pearse A. Keane, Richard Savage, Robert Golub, Rupa Sarkar y Samuel Rowley.

Agradecimientos y financiación. Damos las gracias a los participantes en el estudio Delphi y el estudio piloto (Nota complementaria); a E. Marston (Universidad de Birmingham, Reino Unido) por su apoyo estratégico; a C. Radovanovic (University Hospitals Birmingham NHS Foundation Trust, Reino Unido) y A. Walker (Universidad de Birmingham, Reino Unido) por el apoyo administrativo. Las opiniones expresadas en esta publicación son las de los autores, participantes en el Delphi y de las partes interesadas y pueden no representar las opiniones de la institución anfitriona. Este trabajo ha sido financiado por Wellcome Trust Institutional

Strategic Support Fund: Digital Health Pilot Grant, Research England (parte de investigación e innovación del Reino Unido), Health Data Research UK y a Alan Turing Institute.

El estudio fue patrocinado por la Universidad de Birmingham, Reino Unido. Los financiadores y patrocinadores del estudio no intervinieron en el diseño ni en la realización del estudio, ni en la recopilación, gestión, análisis e interpretación de los datos, ni en la preparación de los informes; preparación, revisión o aprobación del manuscrito; ni en la decisión de presentar el manuscrito para su publicación. M.J.C. es investigador principal del National Institute for Health Research (NIHR) y recibe financiación del NIHR, de Birmingham; Birmingham Biomedical Research Centre; el

NIHR Surgical Reconstruction and Microbiology Research Centre y NIHR ARC West Midlands, University of Birmingham y University Hospitals Birmingham NHS Foundation Trust; Health Data Research UK; Innovate UK (parte de investigación e innovación del Reino Unido) Research and Innovation); the Health Foundation; Macmillan Cancer Support; and UCB Pharma. A.D.y J.D. son también investigadores principales del NIHR. Las opiniones expresadas en este artículo son de los autores y no necesariamente las del NIHR o el Departamento de Salud y Asistencia Social. S.J.V. recibe financiación del Engineering and Physical Sciences Investigación e Innovación del Reino Unido (UKRI), Accenture, Warwick Impact Fund, Health Data Research UK y el Fondo Europeo de Desarrollo Regional. S.R. Es empleado del Consejo de Investigación Médica (UKRI). D.M. Cuenta con el apoyo de una cátedra de investigación de la Universidad de Ottawa. M.K.E. Cuenta con el apoyo de la Administración de Alimentos y Medicamentos de EE.UU. (FDA), y D.P. Cuenta en parte con el apoyo de la Oficina del director de la Biblioteca Nacional de Medicina de EE.UU. (NLM) de los Institutos Nacionales de Salud (NIH). A.B. Cuenta con el apoyo de los 7K01HL141771-02 de los Institutos Nacionales de Salud (NIH). Este artículo puede no coincidir con las opiniones o políticas del NIH y de la FDA. Refleja únicamente los puntos de vista y opiniones de los autores. Agradecemos también al Comité de Inteligencia artificial de la Asociación Colombiana de Radiología (ACR) por su traducción, Mauricio De Jesús Solano Diaz (Universidad de Antioquia) por su apoyo administrativo y editorial.

Conflictos de intereses. M.J.C. ha recibido honorarios personales de Astellas, Takeda, Merck, Daiichi Sankyo, Glaukos,

GlaxoSmithKline y el Patient-Centered Outcomes Research Institute (PCORI) al margen del trabajo presentado. P.A.K. es consultor de DeepMind Technologies, Roche, Novartis y Apellis, y ha recibido honorarios de conferenciante o apoyo para viajes de Bayer, Allergan, Topcon y Heidelberg Engineering. C.J.K. es empleado de Google y posee acciones de Alphabet. A.E. es empleado de Salesforce CRM. R.S. es empleado de Pinpoint Science. J. Matcham era empleado de AstraZeneca en el momento de realizar este estudio. J. Monteiro es editor jefe de la revista Nature Medicine; se ha recusado de cualquier aspecto de la toma de decisiones sobre este manuscrito y no participó en la asignación de este manuscrito a los editores internos o a los revisores, y también fue separado y cegado del proceso editorial desde el inicio de la presentación hasta la decisión.

Declaración. Las opiniones expresadas en este manuscrito son responsabilidad del autor y no reflejan necesariamente los criterios ni la política de la *RPSP/PAJPH* o de la OPS.

REFERENCIAS

1. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? BMJ. 17 de enero de 1998;316(7126):201.
2. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol. enero de 1995;48(1):23-40.
3. Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ. 7 de julio de 2001;323(7303):42-6.
4. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1 de febrero de 1995;273(5):408-12.
5. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 24 de marzo de 2010;340:c869.
6. Moher D, Jones A, Lepage L, CONSORT Group (Consolidated Standards for Reporting of Trials). Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA. 18 de abril de 2001;285(15):1992-5.
7. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. Lancet. 18 de enero de 2014;383(9913):267-76.
8. Boutron I, Altman DG, Moher D, Schulz KF, Ravaud P, CONSORT NPT Group. CONSORT Statement for Randomized Trials of Nonpharmacologic Treatments: A 2017 Update and a CONSORT Extension for Nonpharmacologic Trial Abstracts. Ann Intern Med. 4 de julio de 2017;167(1):40-7.
9. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al. CONSORT for reporting randomised trials in journal and conference abstracts. Lancet. 26 de enero de 2008;371(9609):281-3.
10. MacPherson H, Altman DG, Hammerschlag R, Youping L, Taixiang W, White A, et al. Revised STandards for Reporting Interventions in Clinical Trials of Acupuncture (STRICTA): extending the CONSORT statement. PLoS Med. 8 de junio de 2010;7(6):e1000261.
11. Gagnier JJ, Boon H, Rochon P, Moher D, Barnes J, Bombardier C, et al. Reporting randomized, controlled trials of herbal interventions: an elaborated CONSORT statement. Ann Intern Med. 7 de marzo de 2006;144(5):364-7.
12. Cheng CW, Wu TX, Shang HC, Li YP, Altman DG, Moher D, et al. CONSORT Extension for Chinese Herbal Medicine Formulas 2017: Recommendations, Explanation, and Elaboration. Ann Intern Med. 18 de julio de 2017;167(2):112-21.
13. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA. 27 de febrero de 2013;309(8):814-22.
14. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. enero de 2019;25(1):30-6.
15. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. Addendum: International evaluation of an AI system for breast cancer screening. Nature. octubre de 2020;586(7829):E19.
16. Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. Invest Ophthalmol Vis Sci. 1 de octubre de 2016;57(13):5200-6.
17. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. septiembre de 2018;24(9):1342-50.
18. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2 de febrero de 2017;542(7639):115-8.
19. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. noviembre de 2018;15(11):e1002686.
20. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. marzo de 2020;46(3):383-400.
21. Yim J, Chopra R, Spitz T, Winkens J, Obika A, Kelly C, et al. Predicting conversion to wet age-related macular degeneration using deep learning. Nat Med. junio de 2020;26(6):892-9.
22. Kim H, Goo JM, Lee KH, Kim YT, Park CM. Preoperative CT-based Deep Learning Model for Predicting Disease-Free Survival in Patients with Lung Adenocarcinomas. Radiology. julio de 2020;296(1):216-24.
23. Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. octubre de 2019;68(10):1813-9.
24. Tyler NS, Mosquera-Lopez CM, Wilson LM, Dodier RH, Branigan DL, Gabo VB, et al. An artificial intelligence decision support system for the management of type 1 diabetes. Nat Metab. julio de 2020;2(7):612-9.
25. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health. 1 de octubre de 2019;1(6):e271-97.
26. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a Machine Learning-Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. JAMA. 17 de marzo de 2020;323(11):1052-60.
27. Gong D, Wu L, Zhang J, Mu G, Shen L, Liu J, et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. Lancet Gastroenterol Hepatol. abril de 2020;5(4):352-61.
28. Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol. abril de 2020;5(4):343-51.
29. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for

monitoring blind spots during esophagogastroduodenoscopy. Gut. 1 de diciembre de 2019;68(12):2161-9.
30. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. EClinicalMedicine. marzo de 2019;9:52-9.
31. Su JR, Li Z, Shao XJ, Ji CR, Ji R, Zhou RC, et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc. febrero de 2020;91(2):415-424.e4.
32. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 20 de abril de 2019;393(10181):1577-9.
33. Gregory J, Welliver S, Chong J. Top 10 Reviewer Critiques of Radiology Artificial Intelligence (AI) Articles: Qualitative Thematic Analysis of Reviewer Critiques of Machine Learning/Deep Learning Manuscripts Submitted to JMRI. J Magn Reson Imaging. julio de 2020;52(1):248-54.
34. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 25 de marzo de 2020;368:m689.
35. Liu X, Rivera SC, Faes L, Ferrante di Ruffano L, Yau C, Keane PA, et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med. octubre de 2019;25(10):1467-8.
36. Liu X, Faes L, Calvert MJ, Denniston AK. Extension of the CONSORT and SPIRIT statements. The Lancet. 5 de octubre de 2019;394(10205):1225.
37. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med. 16 de febrero de 2010;7(2):e1000217.
38. Caballero-Ruiz E, García-Sáez G, Rigla M, Villaplana M, Pons B, Hernando ME. A web-based clinical decision support system for gestational diabetes: Automatic diet prescription and detection of insulin needs. Int J Med Inform. junio de 2017;102:35-49.
39. Kim TWB, Gay N, Khemka A, Garino J. Internet-Based Exercise Therapy Using Algorithms for Conservative Treatment of Anterior Knee Pain: A Pragmatic Randomized Controlled Trial. JMIR Rehabil Assist Technol. 14 de diciembre de 2016;3(2):e12.
40. Labovitz DL, Shafner L, Reyes Gil M, Virmani D, Hanina A. Using Artificial Intelligence to Reduce the Risk of Nonadherence in Patients on Anticoagulation Therapy. Stroke. mayo de 2017;48(5):1416-9.
41. Nicolae A, Morton G, Chung H, Loblaw A, Jain S, Mitchell D, et al. Evaluation of a Machine-Learning Algorithm for Treatment Planning in Prostate Low-Dose-Rate Brachytherapy. Int J Radiat Oncol Biol Phys. 15 de marzo de 2017;97(4):822-9.
42. Voss C, Schwartz J, Daniels J, Kline A, Haber N, Washington P, et al. Effect of Wearable Digital Intervention for Improving Socialization in Children With Autism Spectrum Disorder: A Randomized Clinical Trial. JAMA Pediatr. 1 de mayo de 2019;173(5):446-54.
43. Mendes-Soares H, Raveh-Sadka T, Azulay S, Edens K, Ben-Shlomo Y, Cohen Y, et al. Assessment of a Personalized Approach to Predicting Postprandial Glycemic Responses to Food Among Individuals Without Diabetes. JAMA Netw Open. 1 de febrero de 2019;2(2):e188102.
44. Choi KJ, Jang JK, Lee SS, Sung YS, Shim WH, Kim HS, et al. Development and Validation of a Deep Learning System for Staging Liver Fibrosis by Using Contrast Agent-enhanced CT Images in the Liver. Radiology. diciembre de 2018;289(3):688-97.
45. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine. 29 de octubre de 2019;17(1):195.
46. Pooch EHP, Ballester PL, Barros RC. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification [Internet]. arXiv; 2020 [citado 24 de marzo de 2020]. Disponible en: http://arxiv.org/abs/1909.01940
47. International Medical Device Regulators Forum [Internet]. 2019 [citado 24 de marzo de 2020]. Unique Device Identification system (UDI system) Application Guide. Disponible en: https://www.imdrf.org/documents/unique-device-identification-system-udi-system-application-guide
48. Sabottke CF, Spieler BM. The Effect of Image Resolution on Deep Learning in Radiography. Radiol Artif Intell. enero de 2020;2(1):e190015.
49. Heaven D. Why deep-learning AIs are so easy to fool. Nature. octubre de 2019;574(7777):163-6.
50. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med. 2020;3:23.
51. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. septiembre de 2019;25(9):1337-40.
52. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ. 1 de abril de 2020;98(4):251-6.
53. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging [Internet]. arXiv; 2019 [citado 24 de marzo de 2020]. Disponible en: http://arxiv.org/abs/1909.12475
54. CONSORT. Extensions of the CONSORT Statement. [Internet]. [citado 24 de marzo de 2020]. Disponible en: http://www.consort-statement.org/extensions
55. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. PLoS Med. 6 de noviembre de 2018;15(11):e1002683.
56. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science. 22 de marzo de 2019;363(6433):1287-9.
57. Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol. 1 de noviembre de 2018;154(11):1247-8.
58. Zou J, Schiebinger L. AI can be sexist and racist - it's time to make it fair. Nature. julio de 2018;559(7714):324-6.
59. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nat Med. enero de 2020;26(1):16-7.
60. Lee CS, Lee AY. Clinical applications of continual learning machine learning. Lancet Digit Health. junio de 2020;2(6):e279-81.
61. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.
62. Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nat Med. junio de 2020;26(6):807-8.
63. Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykänen P, Rigby M. STARE-HI--Statement on reporting of evaluation studies in Health Informatics. Int J Med Inform. enero de 2009;78(1):1-9.

Manuscrito (original en inglés) recibido el 24 de abril de 2020. Aceptado el 23 de julio de 2020. Publicado en línea el 9 de septiembre de 2020.

GRUPO DE DIRECCIÓN SPIRIT-AI Y CONSORT-AI

Alastair K. Denniston[2,3,4,5,6,13], An-Wen Chan[14], Ara Darzi[15,16], Christopher Holmes[17,18], Christopher Yau[17,19], David Moher[8,9], Hutan Ashrafian[15,16], Jonathan J. Deeks[7,10], Lavinia Ferrante di Ruffano[7], Livia Faes[20], Melanie J. Calvert[4,5,6,7,10,11,12], Pearse A. Keane[13], Samantha Cruz Rivera[5,6,7], Sebastian J. Vollmer[17,21] y Xiaoxuan Liu[1,2,3,4,5]

GRUPO DE CONSENSO SPIRIT-AI Y CONSORT-AI

Aaron Y. Lee[22], Adrian Jonas[23], Andre Esteva[24], Andrew L. Beam[25], An-Wen Chan[14], Maria Beatrice Panico[26], Cecilia S. Lee[22], Charlotte Haug[27], Christopher J. Kelly[28], Christopher Yau[17,19], Cynthia Mulrow[29], Cyrus Espinoza[30], John Fletcher[31], David Moher[8,9], Dina Paltoo[32], Elaine Manna[33], Gary Price[34], Gary S. Collins[35], Hugh Harvey[36], James Matcham[37], Joao Monteiro[38], M. Khair ElZarrad[39], Lavinia Ferrante di Ruffano[7] , Luke Oakden-Rayner[40], Melanie J. Calvert[4,5,6,7,10,11,12], Melissa McCradden[41], Pearse A. Keane[13], Richard Savage[42], Robert Golub[43], Rupa Sarkar[44] y Samuel Rowley[45]

Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension

ABSTRACT

The CONSORT 2010 statement provides minimum guidelines for reporting randomized trials. Its widespread use has been instrumental in ensuring transparency in the evaluation of new interventions. More recently, there has been a growing recognition that interventions involving artificial intelligence (AI) need to undergo rigorous, prospective evaluation to demonstrate impact on health outcomes. The CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) extension is a new reporting guideline for clinical trials evaluating interventions with an AI component. It was developed in parallel with its companion statement for clinical trial protocols: SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence). Both guidelines were developed through a staged consensus process involving literature review and expert consultation to generate 29 candidate items, which were assessed by an international multi-stakeholder group in a two-stage Delphi survey (103 stakeholders), agreed upon in a two-day consensus meeting (31 stakeholders) and refined through a checklist pilot (34 participants). The CONSORT-AI extension includes 14 new items that were considered sufficiently important for AI interventions that they should be routinely reported in addition to the core CONSORT 2010 items. CONSORT-AI recommends that investigators provide clear descriptions of the AI intervention, including instructions and skills required for use, the setting in which the AI intervention is integrated, the handling of inputs and outputs of the AI intervention, the human–AI interaction and provision of an analysis of error cases. CONSORT-AI will help promote transparency and completeness in reporting clinical trials for AI interventions. It will assist editors and peer reviewers, as well as the general readership, to understand, interpret and critically appraise the quality of clinical trial design and risk of bias in the reported outcomes.

[14]Department of Medicine, Women's College Research Institute, Women's College Hospital, University of Toronto, Toronto, Ontario, Canada. [15]Patient Safety Translational Research Centre, Imperial College London, Londres, Reino Unido. [16]Institute of Global Health Innovation, Imperial College London, London, UK. [17]Alan Turing Institute, Reino Unido. [18]Department of Statistics and Nuffield Department of Medicine, University of Oxford, Oxford, Reino Unido. [19]University of Manchester, Manchester, Reino Unido. [20]Department of Ophthalmology, Cantonal Hospital Lucerne, Lucerne, Switzerland. [21]University of Warwick, Coventry, Reino Unido. [22]Department of Ophthalmology, University of Washington, Seattle, WA, Estados Unidos de América. [23]The National Institute for Health and Care Excellence, Londres, Reino Unido. [24]Salesforce Research, San Francisco, CA, Estados Unidos de América. [25]Harvard T.H. Chan School of Public Health, Boston, MA, USA. [26]Medicines and Healthcare products Regulatory Agency, London, UK. [27]New England Journal of Medicine, Waltham, MA, USA. [28]Google Health, London, UK. [29]Annals of Internal Medicine, Filadelfia, PA, Estados Unidos de América. [30]Patient Partner, Birmingham, Reino Unido. [31]British Medical Journal, Londres, Reino Unido. [32]National Institutes of Health, Bethesda, MD, Estados Unidos de América. [33]Patient Partner, Londres, Reino Unido. [34]Patient Partner, Centre for Patient Reported Outcome Research, Institute of Applied Health Research, University of Birmingham, Birmingham, Reino Unido. [35]Centre for Statistics in Medicine, University of Oxford, Oxford, Reino Unido. [36]Hardian Health, Londres, Reino Unido. [37]AstraZeneca, Cambridge, Reino Unido. [38]Nature Research, New York, NY, Estados Unidos de América. [39]Food and Drug Administration, Silver Spring, MD, Estados Unidos de América. [40]Australian Institute for Machine Learning, North Terrace, Adelaide, Australia. [41]The Hospital for Sick Children, Toronto, Canadá. [42]PinPoint Data Science, Leeds, Reino Unido. [43]Journal of the American Medical Association, Chicago, IL, Estados Unidos de América. [44]The Lancet Group, Londres Reino Unido. [45]Medical Research Council, Londres, Reino Unido.

Diretrizes para relatórios de ensaios clínicos com intervenções que utilizam inteligência artificial: a extensão CONSORT-AI

RESUMO

A declaração CONSORT 2010 apresenta diretrizes mínimas para relatórios de ensaios clínicos randomizados. Seu uso generalizado tem sido fundamental para garantir a transparência na avaliação de novas intervenções. Recentemente, tem-se reconhecido cada vez mais que intervenções que incluem inteligência artificial (IA) precisam ser submetidas a uma avaliação rigorosa e prospectiva para demonstrar seus impactos sobre os resultados de saúde. A extensão CONSORT-AI (*Consolidated Standards of Reporting Trials – Artificial Intelligence*) é uma nova diretriz para relatórios de ensaios clínicos que avaliam intervenções com um componente de IA. Ela foi desenvolvida em paralelo à sua declaração complementar para protocolos de ensaios clínicos, a SPIRIT-AI (*Standard Protocol Items: Recommendations for Interventional Trials – Artificial Intelligence*). Ambas as diretrizes foram desenvolvidas por meio de um processo de consenso em etapas que incluiu revisão da literatura e consultas a especialistas para gerar 29 itens candidatos. Foram feitas consultas sobre esses itens a um grupo internacional composto por 103 interessados diretos, que participaram de uma pesquisa Delphi em duas etapas. Chegou-se a um acordo sobre os itens em uma reunião de consenso que incluiu 31 interessados diretos, e os itens foram refinados por meio de uma lista de verificação piloto que envolveu 34 participantes. A extensão CONSORT-AI inclui 14 itens novos que, devido à sua importância para as intervenções de IA, devem ser informados rotineiramente juntamente com os itens básicos da CONSORT 2010. A CONSORT-AI preconiza que os pesquisadores descrevam claramente a intervenção de IA, incluindo instruções e as habilidades necessárias para seu uso, o contexto no qual a intervenção de IA está inserida, considerações sobre o manuseio dos dados de entrada e saída da intervenção de IA, a interação humano-IA e uma análise dos casos de erro. A CONSORT-AI ajudará a promover a transparência e a integralidade nos relatórios de ensaios clínicos com intervenções que utilizam IA. Seu uso ajudará editores e revisores, bem como leitores em geral, a entender, interpretar e avaliar criticamente a qualidade do desenho do ensaio clínico e o risco de viés nos resultados relatados.