PAN AMERICAN JOURNAL OF PUBLIC HEALTH

Informe especial

Directrices para los protocolos de ensayos clínicos de intervenciones con inteligencia artificial: la extensión SPIRIT-AI*

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

Forma de citar Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, Grupo de Trabajo SPIRIT-AI y CONSORT-AI et al. Directrices para los protocolos de ensayos clínicos de intervenciones con inteligencia artificial: la extensión SPIRIT-AI. Rev Panam Salud Publica. 2024;48:e12. https://doi.org/10.26633/RPSP.2024.12

RESUMEN

La declaración SPIRIT 2013 tiene como objetivo mejorar la exhaustividad de los informes de los protocolos de los ensayos clínicos proporcionando recomendaciones basadas en la evidencia para el conjunto mínimo de elementos que deben abordarse. Esta guía ha sido fundamental para promover la evaluación transparente 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 los resultados médicos. La extensión SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence, por sus siglas en inglés) es una nueva directriz para el reporte de los protocolos de ensayos clínicos que evalúan intervenciones con un componente de IA. Esta directriz se desarrolló en paralelo con su declaración complementaria para los informes de ensayos clínicos: CONSORT-AI (Consolidated Standards of Reporting Trials-Artificial Intelligence). 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 26 ítems candidatos, que fueron consultados por un grupo internacional de múltiples partes interesadas en una encuesta Delphi de dos etapas (103 partes interesadas), acordados en una reunión de consenso (31 partes interesadas) y refinados a través de una lista de verificación piloto (34 participantes). La ampliación de SPIRIT-AI incluye 15 nuevos elementos que se consideraron suficientemente importantes para los protocolos de los ensayos clínicos con intervenciones de IA. Estos nuevos ítems deben ser reportados rutinariamente además de los ítems centrales de SPIRIT 2013. SPIRIT-AI recomienda que los investigadores proporcionen descripciones claras de la intervención de IA, incluyendo las instrucciones y las habilidades necesarias para su uso, el entorno en el que se integrará la intervención de IA, las consideraciones para el manejo de los datos de entrada y salida, la interacción entre el ser humano y la IA y el análisis de los casos de error. SPIRIT-AI ayudará a promover la transparencia y la exhaustividad de los protocolos de los ensayos clínicos de las intervenciones de IA. Su uso ayudará a los editores y revisores, así como a los lectores en general, a comprender, interpretar y valorar críticamente el diseño y el riesgo de sesgo de un futuro ensayo clínico.

[1] Centre for Patient Reported Outcomes Research, Institute of Applied Health Research, University of Birmingham, Birmingham, Reino Unido. [2]Institute of Applied Health Research, University of Birmingham, Birmingham, Reino Unido. [3]Birmingham Health Partners Centre for Regulatory Science and Innovation, University of Birmingham, Birmingham, Reino Unido. [4]Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, Reino Unido. [5]University Hospitals Birmingham NHS Foundation Trust, Birmingham, Reino Unido. [6]Health Data Research UK, Londres, Reino Unido. [7]Moorfields Eye Hospital NHS Foundation Trust, Londres, Reino Unido. [8]Department of Medicine, Women's College Research Institute, Women's College Hospital, University of Toronto, Ontario, Canadá. [9]National Institute of Health Research Biomedical Research Centre for Ophthalmology, Moorfields Hospital London NHS Foundation Trust and University College London, Institute of Ophthalmology, Londres, Reino Unido. [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. [a]Véase el listado de autores y sus afiliaciones 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;26:1351-1363. https://doi.org/10.1038/s41591-020-1037-7

El protocolo de un ensayo clínico es un documento esencial elaborado por los investigadores del estudio que detalla a priori la justificación, los métodos propuestos y los planes para la realización de un ensayo clínico[1, 2]. Este documento clave es utilizado por los revisores externos (agencias de financiación, organismos reguladores, comités de ética de la investigación, editores de revistas, revisores de pares, juntas de revisión institucional y, cada vez más, el público en general) para comprender e interpretar la justificación, el rigor metodológico y las consideraciones éticas del ensayo. Además, los protocolos de los ensayos clínicos proporcionan un punto de referencia compartido para apoyar al equipo de investigación en la realización de un estudio de alta calidad.

A pesar de su importancia, la calidad y la exhaustividad de los protocolos de ensayos clínicos publicados son variables[1, 2]. La declaración SPIRIT se publicó en 2013 para proporcionar orientación sobre el contenido mínimo de información de un protocolo de ensayo clínico y ha sido ampliamente respaldada como norma internacional[3-5]. La declaración SPIRIT publicada en 2013 proporciona una orientación mínima aplicable a todas las intervenciones de los ensayos clínicos, pero reconoce que ciertas intervenciones pueden requerir la extensión o aclaración de estos elementos[1, 2]. La inteligencia artificial (IA) es un área de enorme interés, con un fuerte impulso para acelerar las nuevas intervenciones desde su publicación e implementación hasta su comercialización[6]. 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 en salud. Los ejemplos de estas aplicaciones son muy variados e incluyen sistemas de IA para el cribado y el triaje[7, 8], el diagnóstico[9-12], el pronóstico[13, 14], el apoyo a la toma de decisiones[15] y la recomendación de tratamientos[16]. Sin embargo, en la mayoría de los casos recientes, la mayor parte de las pruebas publicadas han consistido en una validación *in silico* en fase inicial. Se ha reconocido que la mayoría de los estudios de IA recientes no se comunican de forma adecuada y que las directrices de información existentes no cubren totalmente las posibles fuentes de sesgo específicas de los sistemas de IA[17]. La bienvenida a la aparición de ensayos controlados aleatorizados que buscan evaluar la eficacia clínica de las nuevas intervenciones basadas en, o que incluyen, un componente de IA (denominadas aquí "intervenciones de IA")[15, 18-23] se ha encontrado igualmente con preocupaciones sobre su diseño y reporte[17,24-26]. 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.

SPIRIT-AI (como parte de la iniciativa SPIRIT-AI y CONSORT-AI) es una iniciativa internacional apoyada por SPIRIT y la red EQUATOR (Enhancing the Quality and Transparency of Health Research, por sus siglas en inglés) para ampliar o aclarar la declaración existente SPIRIT 2013 cuando sea necesario, con el fin de desarrollar una guía de protocolo específica para la IA basada en consenso[27, 28]. Es complementaria a la declaración CONSORT-AI, cuyo objetivo es promover la alta calidad del reporte en los ensayos de IA. Esta declaración de consenso describe los métodos utilizados para identificar y evaluar los ítems candidatos y obtener un consenso. Además, también proporciona la lista de verificación completa de SPIRIT-AI, incluyendo los nuevos ítems y las explicaciones que los acompañan.

MÉTODOS

Las extensiones de SPIRIT-AI y CONSORT-AI se desarrollaron simultáneamente para el reporte y los protocolos de ensayos clínicos. En octubre de 2019 se publicó el anuncio de la iniciativa SPIRIT-AI y CONSORT-AI (ref [27]), y las dos directrices se registraron en mayo de 2019 como directrices en desarrollo en la biblioteca de directrices de EQUATOR. Ambas directrices se desarrollaron de acuerdo con el marco metodológico de EQUATOR Network[29]. Se formó el Grupo Directivo de SPIRIT-AI y CONSORT-AI, compuesto por 15 expertos internacionales, 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 del 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 registrados, de los cuales 62 se habían completado y 7 habían publicado resultados[22, 30-35]. Dos estudios se informaron con referencia a la declaración CONSORT[22, 34], y un estudio proporcionó un protocolo de ensayo clínico no publicado[34]. El equipo de operaciones (X.L., S.C.R., M.J.C. y A.K.D.) identificó las consideraciones específicas de la IA relacionada con estos estudios y las reformuló como ítems de información candidatos. Los ítems candidatos también se basaron en los resultados de una revisión sistemática anterior que evaluó la precisión diagnóstica de los sistemas de aprendizaje profundo en las imágenes médicas[17]. Después de consultar con el Grupo Directivo y con 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 solo 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

RECUADRO 1. Glosario

Inteligencia artificial	La ciencia que desarrolla sistemas informáticos que pueden realizar tareas que normalmente requieren inteligencia humana.
Intervención de IA	Intervención sanitaria que se basa en un componente de IA/ML para cumplir su objetivo.
CONSORT	Normas consolidadas para el reporte de ensayos clínicos.
Elemento de extensión de CONSORT-AI	Un 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 especialmente 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 entre humanos e inteligencia artificial	El proceso de cómo los usuarios (humanos) interactúan con la intervención de IA, para que esta 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 la formación 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	Modificaciones o entrenamientos adicionales realizados en el modelo de intervención de la IA, con la intención de mejorar su desempeño.
Datos de entrada	Los datos que deben presentarse a la intervención de IA para que pueda cumplir su propósito.
Aprendizaje automático	Campo de la informática 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	El resultado predicho 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	Casos en los 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 desempeño, preferiblemente basado en evidencia previa. Puede ir desde pequeñas disminuciones de la precisión (en comparación con la precisión esperada) hasta predicciones erróneas o la incapacidad de producir una salida, en determinados casos.
SPIRIT	Standard Protocol Items (Elementos de protocolo estándar), por sus siglas en inglés. Recomendaciones para los ensayos 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 ítem existente de SPIRIT 2013 cuando se aplica a las intervenciones de IA.

sobre los ítems candidatos y sugerir elementos adicionales. Los expertos fueron identificados y contactados a través del Grupo Directivo y se les permitió una ronda de reclutamiento de "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[27]. 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 (en inglés). 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 las encuestas 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 un 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 (el 88% de los participantes de la primera ronda) para la segunda ronda.

Los resultados de los estudios Delphi sirvieron de base para la posterior reunión de consenso internacional. Los participantes en el estudio Delphi propusieron 12 nuevos ítems, que se añadieron para su discusión en la reunión de consenso. Los

CUADRO 1. Lista de control de SPIRIT-AI

Sección	Elemento	Ítem de SPIRIT 2013	Ítem de SPIRIT-AI	Abordado en el número de página[b]
Información administrativa				
Título	1	Título descriptivo que identifica el diseño del estudio, la población, las intervenciones y, si procede, el acrónimo del ensayo clínico	SPIRIT-AI 1 (i) Aclaración	Indique que la intervención implica inteligencia artificial/aprendizaje automático y especifique el tipo de modelo.
			SPIRIT-AI 1 (ii) Aclaración	Especifique el uso previsto de la intervención de IA.
Registro del ensayo clínico	2a	Identificador del ensayo y nombre del registro. Si aún no está registrado, nombre del registro previsto		
	2b	Todos los elementos del conjunto de datos de registro de ensayos de la Organización Mundial de la Salud		
Versión del protocolo	3	Fecha e identificador de la versión		
Financiación	4	Fuentes y tipos de apoyo financiero, material y de otro tipo		
Funciones y responsabilidades	5a	Nombres, afiliaciones y funciones de los colaboradores del protocolo		
	5b	Nombre e información de contacto del patrocinador del ensayo clínico		
	5c	Papel del patrocinador del estudio y de los financiadores, si los hay, en el diseño del estudio; en la recogida, gestión, análisis e interpretación de los datos; en la redacción del informe; y en la decisión de presentar el informe para su publicación, incluyendo si tendrán autoridad final sobre alguna de estas actividades		
	5d	Composición, funciones y responsabilidades del centro coordinador, el comité directivo, el comité de adjudicación de criterios de valoración, el equipo de gestión de datos y otras personas o grupos que supervisen el ensayo, si procede (véase el punto 21a para el comité de supervisión de datos)		
Introducción				
Antecedentes y justificación	6a	Descripción de la pregunta de investigación y justificación para realizar el ensayo, incluyendo un resumen de los estudios relevantes (publicados y no publicados) que examinan los beneficios y daños de cada intervención	SPIRIT-AI 6a (i) Extensión	Explique el uso previsto de la intervención de IA en el contexto de la vía clínica, incluyendo su propósito y sus usuarios previstos (p. ej., profesionales sanitarios, pacientes, público).
			SPIRIT-AI 6a (ii) Extensión	Describa cualquier evidencia preexistente para la intervención de IA.
	6b	Explicación de la elección de los comparadores		
Objetivos	7	Especifique objetivos o hipótesis		
Diseño del ensayo	8	Descripción del diseño del ensayo, incluido el tipo de ensayo (p. ej., grupo paralelo, cruzado, factorial, grupo único), la proporción de asignación y el marco (p. ej, superioridad, equivalencia, no inferioridad, exploratorio)		
Métodos: participantes, intervenciones y resultados				
Marco del estudio	9	Descripción del entorno del estudio (p. ej., clínica comunitaria, hospital académico) y lista de países donde se recogerán los datos. Referencia a dónde puede obtenerse la lista de lugares de estudio	SPIRIT-AI 9 Extensión	Describa los requisitos in situ y ex situ necesarios para integrar la intervención de IA en el entorno del ensayo clínico.

(Continuará)

CUADRO 1. (Cont.)

Sección	Elemento	Ítem de SPIRIT 2013	Ítem de SPIRIT-AI		Abordado en el número de página[b]
Criterios de elegibilidad	10	Criterios de inclusión y exclusión de los participantes. Si procede, criterios de elegibilidad para los centros de estudio y las personas que realizarán las intervenciones (p. ej., cirujanos, psicoterapeutas)	SPIRIT-AI 10 (i) Aclaración	Indique los criterios de inclusión y exclusión a nivel de los participantes.	
			SPIRIT-AI 10 (ii) Extensión	Indique los criterios de inclusión y exclusión a nivel de los datos de entrada.	
Intervenciones	11a	Intervenciones para cada grupo con suficiente detalle para permitir su replicación, incluyendo cómo y cuándo se administrarán	SPIRIT-AI 11a (i) Extensión	Indique qué versión del algoritmo de IA se utilizará.	
			SPIRIT-AI 11a (ii) Extensión	Especifique el procedimiento de adquisición y selección de los datos de entrada para la intervención de IA.	
			SPIRIT-AI 11a (iii) Extensión	Especifique el procedimiento para evaluar y tratar los datos de entrada de baja calidad o no disponibles.	
			SPIRIT-AI 11a (iv) Extensión	Especifique si hay interacción entre el ser humano y la IA en el manejo de los datos de entrada, y qué nivel de experiencia se requiere para los usuarios.	
			SPIRIT-AI 11a (v) Extensión	Especifique el resultado de la intervención de la IA.	
			SPIRIT-AI 11a (vi) Extensión	Explique el procedimiento de cómo el resultado de la intervención de IA contribuirá a la toma de decisiones u otros elementos de la práctica clínica.	
	11b	Criterios para interrumpir o modificar las intervenciones asignadas a un determinado participante en el ensayo clínico (p. ej., cambio de la dosis del fármaco en respuesta a los daños, a la petición del participante o a la mejora/empeoramiento de la enfermedad)			
	11c	Estrategias para mejorar la adherencia a los protocolos de intervención, y cualquier procedimiento para monitorear la adherencia (p. ej., devolución de tabletas de medicamentos, pruebas de laboratorio)			
	11d	Cuidados e intervenciones concomitantes relevantes que se permiten o prohíben durante el ensayo			
Resultados	12	Resultados primarios, secundarios y otros, incluyendo la variable de medición específica (p. ej., la presión arterial sistólica), la métrica de análisis (p. ej., el cambio desde la línea de base, el valor final, el tiempo hasta el evento), el método de agregación (p. ej., la mediana, la proporción) y el punto de tiempo para cada resultado. Se recomienda encarecidamente explicar la relevancia clínica de los resultados de eficacia y daño elegidos			

(Continuará)

CUADRO 1. (Cont.)

Sección	Elemento	Ítem de SPIRIT 2013	Ítem de SPIRIT-AI	Abordado en el número de página[b]
Cronología de los participantes	13	Cronograma de la inscripción, las intervenciones (incluidas las pruebas y los lavados), las evaluaciones y las visitas de los participantes. Se recomienda encarecidamente un diagrama esquemático (Fig. 1)		
Tamaño de la muestra	14	Número estimado de participantes necesarios para alcanzar los objetivos del estudio y cómo se determinó, incluyendo las hipótesis clínicas y estadísticas que respaldan cualquier cálculo del tamaño de la muestra		
Reclutamiento	15	Estrategias para lograr la inscripción adecuada de los participantes para alcanzar el tamaño de la muestra objetivo		
Métodos: asignación de intervenciones (para ensayos controlados)				
Generación de secuencias	16a	Método de generación de la secuencia de asignación (p. ej., números aleatorios generados por ordenador), y lista de cualquier factor de estratificación. Para reducir la previsibilidad de una secuencia aleatoria, los detalles de cualquier restricción planificada (p. ej., el bloqueo) deben proporcionarse en un documento separado que no esté disponible para quienes inscriben a los participantes o asignan las intervenciones		
Mecanismo de ocultación de la asignación	16b	Mecanismo de aplicación de la secuencia de asignación (p. ej., teléfono central; sobres numerados secuencialmente, opacos y sellados), describiendo cualquier paso para ocultar la secuencia hasta que se asignen las intervenciones		
Ejecución	16c	Quién generará la secuencia de asignación, quién inscribirá a los participantes y quién los asignará a las intervenciones		
Cegamiento (enmascaramiento)	17a	Quién estará cegado después de la asignación a las intervenciones (p. ej., los participantes en el ensayo, los proveedores de atención, los evaluadores de resultados, los analistas de datos), y cómo		
	17b	Si se ciega, las circunstancias en las que se permite el desenmascaramiento, y el procedimiento para revelar la intervención asignada a un participante durante el ensayo		
Métodos: Recolección, gestión y análisis de datos				
Métodos de recolección de datos	18a	Planes para la evaluación y recogida de datos de resultados, de referencia y de otros datos del ensayo, incluyendo cualquier proceso relacionado para promover la calidad de los datos (p. ej., mediciones duplicadas, formación de los evaluadores) y una descripción de los instrumentos del estudio (p. ej., cuestionarios, pruebas de laboratorio) junto con su fiabilidad y validez, si se conocen. Referencia a dónde pueden encontrarse los formularios de recogida de datos, si no están en el protocolo		
	18b	Planes para promover la retención de los participantes y el seguimiento completo, incluyendo una lista de los datos de resultados que se recogerán para los participantes que abandonen o se desvíen de los protocolos de intervención		

(Continuará)

CUADRO 1. (Cont.)

Sección	Elemento	Ítem de SPIRIT 2013	Ítem de SPIRIT-AI		Abordado en el número de página[b]
Gestión de los datos	19	Planes para la introducción de datos, codificación, seguridad y almacenamiento, incluyendo cualquier proceso relacionado para promover la calidad de los datos (p. ej., doble introducción de datos; comprobación de rangos para los valores de los datos). Referencia a dónde se pueden encontrar los detalles de los procedimientos de gestión de datos, si no están en el protocolo			
Métodos estadísticos	20a	Métodos estadísticos para analizar los resultados primarios y secundarios. Referencia al lugar donde se pueden encontrar otros detalles del plan de análisis estadístico, si no está en el protocolo			
	20b	Métodos para cualquier análisis adicional (p. ej., análisis de subgrupos y ajustados)			
	20c	Definición de la población de análisis relacionada con la no adherencia al protocolo (p. ej., como análisis aleatorio), y cualquier método estadístico para manejar los datos faltantes (p. ej., imputación múltiple)			
Métodos: Seguimiento					
Monitorización de datos	21a	Composición del comité de monitorización de datos (CMD); resumen de su función y estructura de información; declaración de si es independiente del patrocinador y de los intereses en competencia; y referencia a dónde se pueden encontrar más detalles sobre sus estatutos, si no están en el protocolo. Alternativamente, una explicación de por qué no es necesario un CMD.			
	21b	Descripción de los análisis interinos y las directrices de interrupción, incluyendo quién tendrá acceso a estos resultados interinos y tomará la decisión final de terminar el ensayo			
Daños	22	Planes para recoger, evaluar, notificar y gestionar los acontecimientos adversos solicitados y notificados espontáneamente y otros efectos no deseados de las intervenciones o la realización del ensayo	SPIRIT-AI 22 Extensión	Especifique cualquier plan para identificar y analizar los errores de ejecución. Si no hay planes para ello, justifique por qué no.	
Auditoría	23	Frecuencia y procedimientos para auditar la realización del ensayo, si los hay, y si el proceso será independiente de los investigadores y del patrocinador			
Ética y difusión					
Aprobación ética de la investigación	24	Planes para solicitar la aprobación del comité de ética de investigación/comité de revisión institucional (CEI/CRI)			
Modificaciones del protocolo	25	Planes para comunicar las modificaciones importantes del protocolo (p. ej., cambios en los criterios de elegibilidad, resultados, análisis) a las partes pertinentes (p. ej., investigadores, CEI/CRI, participantes en el ensayo, registros del ensayo, revistas, reguladores)			
Consentimiento o asentimiento	26a	Quién obtendrá el consentimiento informado o asentimiento de los posibles participantes en el ensayo o sustitutos autorizados, y cómo (véase el ítem 32)			
	26b	Disposiciones adicionales de consentimiento para la recogida y el uso de datos de los participantes y de muestras biológicas en estudios auxiliares, si procede			

(Continuará)

CUADRO 1. (Cont.)

Sección	Elemento	Ítem de SPIRIT 2013	Ítem de SPIRIT-AI	Abordado en el número de página[b]
Confidencialidad	27	Cómo se recogerá, compartirá y mantendrá la información personal de los participantes potenciales e inscritos para proteger la confidencialidad antes, durante y después del ensayo		
Declaración de intereses	28	Intereses económicos y otros intereses concurrentes de los investigadores principales del ensayo en general y de cada centro del estudio		
Acceso a los datos	29	Declaración de quién tendrá acceso al conjunto de datos final del ensayo, y revelación de los acuerdos contractuales que limitan dicho acceso para los investigadores	SPIRIT-AI 29 Extensión	Indicar si se puede acceder a la intervención de IA y/o a su código, y cómo, incluyendo cualquier restricción de acceso o reutilización.
Cuidados auxiliares y posteriores al ensayo	30	Disposiciones, si las hay, para la atención auxiliar y posterior al ensayo, y para la compensación a quienes sufran daños por la participación en el ensayo		
Política de difusión	31a	Planes para que los investigadores y el patrocinador comuniquen los resultados del ensayo a los participantes, a los profesionales de la salud, al público y a otros grupos relevantes (p. ej., a través de la publicación, la información en bases de datos de resultados u otros acuerdos para compartir datos), incluyendo cualquier restricción de publicación		
	31b	Directrices de elegibilidad para la autoría y cualquier uso previsto de escritores profesionales		
	31c	Planes, si los hay, para conceder acceso público al protocolo completo, al conjunto de datos de los participantes y al código estadístico		
Apéndices				
Materiales de consentimiento informado	32	Modelo de formulario de consentimiento y otra documentación relacionada entregada a los participantes y sustitutos autorizados		
Muestras biológicas	33	Planes para la recogida, evaluación en el laboratorio y almacenamiento de muestras biológicas para el análisis genético o molecular en el ensayo actual y para su uso futuro en estudios auxiliares, si procede		

[a]Se recomienda encarecidamente leer esta lista de comprobación junto con la Explicación y aclaración de SPIRIT 2013 para obtener aclaraciones importantes sobre los puntos. [b]Indica los números de página que deben completar los autores durante el desarrollo del protocolo.

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 temas y votar sobre su inclusión. Los participantes se seleccionaron para lograr una representación adecuada de todos los grupos interesados. Se discutieron 38 ítems, incluyendo los 26 generados en la revisión inicial de la literatura y la fase de generación de ítems (estos 26 ítems eran relevantes tanto para SPIRIT-AI como para CONSORT-AI; también se discutieron 3 ítems extra relevantes solo para CONSORT-AI) y los 12 nuevos ítems propuestos por los participantes durante las encuestas Delphi. Cada punto se presentó al grupo de consenso, junto con su puntuación en el ejercicio Delphi (mediana y rangos intercuartiles) y cualquier comentario realizado por los participantes en el Delphi relacionado con ese punto. Se invitó a los participantes en la reunión de consenso a comentar la importancia de cada elemento y si este 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 las pads de votación Turning Point (Turning Technologies, versión 8.7.2.14).

Lista de control piloto

Tras la reunión de consenso, los asistentes tuvieron la oportunidad de hacer comentarios finales sobre la redacción y acordar que los elementos actualizados de SPIRIT-AI y CONSORT-AI reflejaban 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 (Fig. suplementaria 1, en inglés). Se realizó una prueba piloto de las penúltimas listas de verificación con 34 participantes para garantizar 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 (Fig. suplementaria 2, en inglés).

RECOMENDACIONES

Elementos de la lista de comprobación de SPIRIT-AI y explicación. La extensión de SPIRIT-AI recomienda que, junto con los ítems existentes de SPIRIT 2013, se aborden 15 ítems (12 extensiones y 3 aclaraciones) para los protocolos de ensayo clínico de las intervenciones de IA. Estos ítems se consideraron lo suficientemente importantes para los protocolos de ensayos clínicos que incluyen intervenciones de IA como para ser informados de forma rutinaria además de los ítems de la lista de comprobación central de SPIRIT 2013.

El cuadro 1 enumera los ítems de SPIRIT-AI. Los 15 ítems incluidos en la Extensión SPIRIT-AI superaron el umbral del 80% para su inclusión en la reunión de consenso. SPIRIT-AI 6a (i), SPIRIT-AI 11a (v) y SPIRIT-AI 22 fueron el resultado de la fusión de dos ítems luego de la discusión. SPIRIT-AI 11a (iii) no cumplía los criterios de inclusión sobre la base de su redacción inicial (73% de votos a favor de la inclusión); sin embargo, tras un amplio debate y una nueva redacción, el grupo de consenso apoyó por unanimidad una nueva votación, momento en el que superó el umbral de inclusión (97% a favor de la inclusión).

Información administrativa

SPIRIT-AI 1 (i) Aclaración: Indique que la intervención implica inteligencia artificial/aprendizaje automático y especifique el tipo de modelo. *Explicación*. Se recomienda indicar en el título del protocolo y/o en el resumen 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 protocolo del ensayo clínico en bases de datos bibliográficas, registros y otros recursos en línea. 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.

SPIRIT-AI 1 (ii) Aclaración: Indique el uso previsto de la intervención de IA. *Explicación*. El uso previsto de la intervención de IA debe quedar claro en el título y/o el resumen del protocolo. Esto debe describir el propósito de la intervención de IA y el contexto de la enfermedad[19, 36]. 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

SPIRIT-AI 6a (i) Extensión: Explique el uso previsto de la intervención de IA en el contexto de la vía clínica, incluyendo su propósito y sus usuarios previstos (p. ej., profesionales sanitarios, pacientes, público). *Explicación*. Para aclarar cómo encaja la intervención de IA en un flujo clínico, debe incluirse en los antecedentes del protocolo una descripción detallada de su función. Las intervenciones de IA pueden estar diseñadas para interactuar con diferentes usuarios, incluidos los profesionales sanitarios, los pacientes y el público, y sus funciones pueden ser muy variadas (p. ej., la misma intervención de IA podría, en teoría, 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 evaluará la intervención de IA en el ensayo.

SPIRIT-AI 6a (ii) Extensión: Describa cualquier evidencia preexistente para la intervención de IA. *Explicación*. Los autores deben describir en el protocolo cualquier evidencia preexistente publicada (con referencias de apoyo) o evidencia no publicada relacionada con la validación de la intervención de IA o la falta de ella. Hay que tener en cuenta si las pruebas se refieren a un uso, un entorno y una población objetivo similares a los del ensayo previsto. Esto puede incluir el desarrollo previo del modelo de IA, las validaciones internas y externas y cualquier modificación realizada antes del ensayo.

Participantes, intervenciones y resultados

SPIRIT-AI 9 Extensión: Describa los requisitos in situ y externos necesarios para integrar la intervención de IA en el entorno del ensayo. *Explicación*. La generalización de los algoritmos de IA tiene limitaciones, una de las cuales es cuando se utilizan fuera de su entorno de desarrollo[37, 38] . Los sistemas de IA dependen de su entorno operativo, y el protocolo 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 requiere dispositivos específicos del proveedor, si es necesario un hardware informático especializado en cada centro, o si los centros deben soportar la integración en la nube, especialmente si esto es específico del proveedor. Si se requiere 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.

SPIRIT-AI 10 (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 protocolos de los ensayos de intervención no relacionados con la IA. Esto es distinto de los criterios de inclusión y exclusión realizados a nivel de los datos de entrada, que se abordan en el punto 10 (ii).

SPIRIT-AI 10 (ii) Extensión: 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 (p. ej., para un sistema de diagnóstico de cáncer de mama, los datos de entrada podrían ser la mamografía no procesada o procesada por un proveedor específico sobre la 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 protocolo 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 determinarían la elegibilidad previa a la aleatorización. Debe especificar cuándo, cómo y quién lo evaluará. Por ejemplo, si un participante cumple los criterios de elegibilidad para estar en decúbito dorsal para una TC según el punto 10 (i), pero la calidad de la exploración se ha visto comprometida (por cualquier motivo) hasta un nivel tal que ya no es apta para ser utilizada por el sistema de IA, esto debería considerarse como un criterio de exclusión a nivel de datos de entrada. Obsérvese que cuando los datos de entrada se adquieren después de la aleatorización (abordada por SPIRIT-20c), cualquier exclusión se considera del análisis, no de la inscripción (Fig. 1).

SPIRIT-AI 11a (i) Extensión: Indique qué versión del algoritmo de IA se utilizará. *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. El protocolo debe indicar qué versión del sistema de IA se utilizará en el ensayo clínico y si se trata de la misma versión que se utilizó en estudios anteriores que se han utilizado para justificar la justificación del estudio. Si procede, el protocolo debe describir qué ha cambiado entre las versiones pertinentes y la justificación de los cambios. Cuando esté disponible, el protocolo debe incluir una referencia de marcado reglamentario, como un identificador único de dispositivo, que requiere un nuevo identificador para las versiones actualizadas del dispositivo[39].

SPIRIT-AI 11a (ii) Extensión: Especificar el procedimiento de adquisición y selección de los datos de entrada para la intervención de IA. *Explicación*. El desempeño medido de cualquier sistema de IA puede depender críticamente de la naturaleza y la calidad de los datos de entrada[40]. Debe indicarse el procedimiento de tratamiento de los datos de entrada, incluida la adquisición, selección y preprocesamiento de los datos antes de su análisis por el sistema de IA. La integridad y la transparencia de este proceso son esenciales para la evaluación de la viabilidad y para la futura reproducción de la intervención más allá del ensayo clínico. También ayudará a identificar si los procedimientos de manejo de datos de entrada se estandarizarán en todos los centros del ensayo.

SPIRIT-AI 11a (iii) Extensión: Especifique el procedimiento para evaluar y manejar los datos de entrada de mala calidad o no disponibles. *Explicación*. Al igual que en SPIRIT-AI 10 (ii), los "datos de entrada" se refieren a los datos requeridos por la intervención de IA para cumplir su propósito. Como se indica en el punto 10 (ii), el desempeño de los sistemas de IA puede verse comprometido como resultado de la mala calidad o la falta de datos de entrada[41] (p. ej., un artefacto de movimiento excesivo en un electrocardiograma). El protocolo del estudio debe especificar si se identificarán y tratarán los datos de entrada de mala calidad o no disponibles, y cómo se hará. El protocolo también debe especificar un estándar mínimo requerido para los datos de entrada y el procedimiento para cuando no se cumpla el estándar mínimo (incluyendo el impacto en, o cualquier cambio en, la vía de atención al participante).

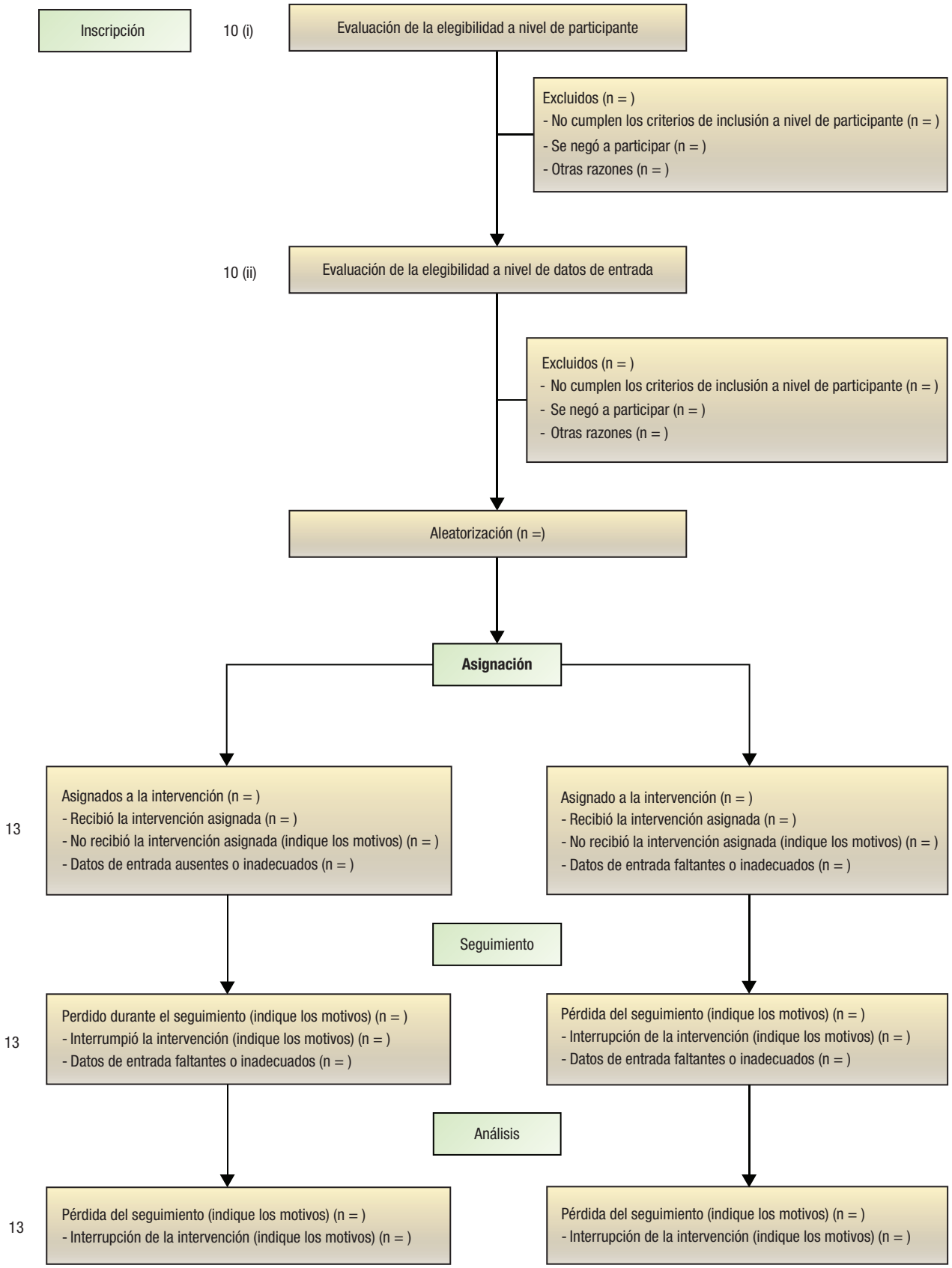
Los datos de mala calidad o no disponibles también pueden afectar a las intervenciones no relacionadas con la IA. Por ejemplo, una calidad inferior a la óptima de una exploración podría afectar a la capacidad de un radiólogo para interpretarla y realizar un diagnóstico. Por lo tanto, es importante que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

SPIRIT-AI 11a (iv) Extensión: Especifique si hay interacción entre el ser humano y la IA en el manejo de los datos de entrada, y qué nivel de experiencia se requiere para los usuarios. *Explicación*. Debe proporcionarse una descripción de la interfaz persona-inteligencia artificial y de los requisitos para una interacción satisfactoria cuando se manejan los datos de entrada. Los ejemplos incluyen la selección por parte del médico de las regiones de interés de un portaobjetos de histología que luego interpreta un sistema de diagnóstico de IA[42], o la selección por parte de un endoscopista de un vídeo de colonoscopia como datos de entrada para un algoritmo diseñado para detectar pólipos[21]. La descripción de cualquier formación prevista para el usuario y las instrucciones sobre cómo los usuarios manejarán los datos de entrada proporcionan transparencia y replicabilidad a los procedimientos del ensayo clínico. La falta de claridad en la interfaz persona-IA puede conducir a la falta de un enfoque estándar y puede tener implicaciones éticas, especialmente en caso de daño[43,44]. 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.

SPIRIT-AI 11a (v) Extensión: Especifique el resultado de la intervención de IA. *Explicación*. El resultado de la intervención de IA debe estar claramente definido en el protocolo. Por ejemplo, un sistema de IA puede dar como resultado una clasificación o probabilidad de diagnóstico, una recomendación de acción, una alarma que alerte de 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.

SPIRIT-AI 11a (vi) Extensión: Explicar el procedimiento de cómo los resultados de la intervención de IA contribuirán a la toma de decisiones u otros elementos de la práctica clínica. *Explicación*. Dado que los resultados de salud también pueden

FIGURA 1. Diagrama de flujo de CONSORT 2010 adaptado para los ensayos clínicos de inteligencia artificial.



SPIRIT-AI 10 (i): Indique los criterios de inclusión y exclusión a nivel de participantes. SPIRIT-AI 10 (ii): Establezca los criterios de inclusión y exclusión a nivel de los datos de entrada. SPIRIT 13 (elemento central de las directrices CONSORT): Cronograma de inscripción, intervenciones (incluyendo los recorridos y lavados), evaluaciones y visitas de los participantes. Se recomienda encarecidamente un diagrama esquemático.

depender de manera crítica de la forma en que los seres humanos interactúan con la intervención de IA, el protocolo del ensayo debe explicar cómo se utilizan 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 influir en los resultados. Como en el caso de SPIRIT-AI 11a (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 debe interpretarse este resultado y cómo debe actuar el usuario, especificando tanto las alternativas clínicas previstas (p. ej., la escisión de la lesión cutánea si el diagnóstico es positivo) como los umbrales para entrar en estas vías (p. ej., 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 manera similar, junto con una explicación de cómo se utilizó dicha información para llegar a las decisiones clínicas para el manejo del paciente, cuando sea relevante.

Monitoreo

SPIRIT-AI 22 Extensión: Especifique cualquier plan para identificar y analizar los errores de desempeño. Si no hay planes para ello, explique por qué no. *Explicación*. El reporte de errores de desempeño y el análisis de casos de fallo 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[45]. Por lo tanto, la identificación de casos de error y la definición de estrategias de mitigación de riesgos es importante para informar de cuándo puede aplicarse la intervención de forma segura, y para qué poblaciones. El protocolo debe especificar si hay planes para analizar los errores de funcionamiento. Si no hay planes para ello, debe incluirse una justificación en el protocolo.

Ética y difusión

SPIRIT-AI 29 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 protocolo debe dejar claro si se puede acceder a la intervención de IA y/o a su código o reutilizarlos, y cómo hacerlo. Esto debe incluir detalles sobre la licencia y cualquier restricción de acceso.

DISCUSIÓN

La extensión SPIRIT-AI proporciona una guía basada en el consenso internacional sobre la información específica de la IA que debe ser reportada en los protocolos de los ensayos clínicos, junto con SPIRIT 2013 y otras extensiones relevantes de SPIRIT[4,46]. Consta de 15 elementos: 3 aclaraciones de la guía SPIRIT 2013 existente en el contexto de los ensayos de IA, y 12 extensiones nuevas. La guía no pretende ser prescriptiva sobre el enfoque metodológico de los ensayos de IA; en cambio, pretende promover la transparencia en la comunicación del diseño y los métodos de un ensayo clínico para facilitar la comprensión, la interpretación y la revisión por pares.

Una serie de ítems de ampliación se refieren a la intervención (puntos 11 (i)-11 (vi)), su entorno (punto 9) y la función prevista (punto 6a (i)). Se formularon recomendaciones específicas relativas a los sistemas de IA relacionadas con la versión del algoritmo, los datos de entrada y salida, la integración en el entorno del ensayo, la experiencia de los usuarios y el protocolo para actuar según las recomendaciones del sistema de IA. Se acordó que estos detalles son fundamentales para la evaluación independiente del protocolo del estudio. Los editores de las revistas informaron que, a pesar de la importancia de estos elementos, en la actualidad suelen faltar en los protocolos e 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 la discusión del grupo de consenso fue la seguridad de los sistemas de IA. Esto se debe a que estos sistemas, a diferencia de otras intervenciones sanitarias, pueden producir errores imprevisibles 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[47,48]. 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. Se añadieron dos elementos de ampliación para abordar esta cuestión. El punto 6a (ii) de SPIRIT-AI requiere la especificación del nivel de evidencia previo para la validación de la intervención de IA. El punto 22 de SPIRIT-AI requiere la especificación de cualquier plan para analizar los errores de desempeño, para enfatizar la importancia de anticipar los errores sistemáticos cometidos por el algoritmo y sus consecuencias.

Un tema que se planteó en las respuestas de la encuesta Delphi y en la reunión de consenso y que no se incluye en las directrices finales es el de los sistemas de IA de "evolución continua" (también conocidos como sistemas de IA de "adaptación continua" o "aprendizaje continuo"). Se trata de sistemas de IA con capacidad para entrenarse continuamente con nuevos datos, lo que puede provocar cambios en el desempeño a lo largo del tiempo. El grupo observó que, aunque es interesante, este campo se encuentra en una fase relativamente temprana de desarrollo, sin ejemplos tangibles en aplicaciones sanitarias, y que no sería apropiado que SPIRIT-AI lo abordara en esta fase[49]. Este tema será objeto de seguimiento y revisión en futuras iteraciones de SPIRIT-AI. Cabe señalar que los cambios incrementales del software, ya sean continuos o iterativos, intencionados o no, podrían tener graves consecuencias en el desempeño de la seguridad luego 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, en el momento del desarrollo de SPIRIT-AI, solo había siete ensayos publicados y ningún protocolo de ensayo publicado en el ámbito de la IA en la salud. Por lo tanto, el debate y las decisiones tomadas durante el desarrollo de SPIRIT-AI no siempre están respaldados por ejemplos del mundo

real existentes. Esto surge de nuestro objetivo declarado de abordar los problemas de desarrollo de protocolos deficientes 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 estos estándares de reporte y garantizar su continua relevancia. En segundo lugar, la búsqueda bibliográfica de ensayos clínicos controlados aleatorizados de IA utilizó terminología como "inteligencia artificial", "aprendizaje automático" y "aprendizaje profundo", pero no términos como "sistemas de apoyo a la decisión clínica" y "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[50]. 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 ítems candidatos fue generada por un grupo relativamente pequeño de expertos formado por los miembros del Grupo Directivo 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 SPIRIT, la ampliación de SPIRIT-AI pretende ser una guía mínima para la presentación de informes, y hay consideraciones adicionales específicas de la IA para los protocolos de los ensayos que pueden justificar su consideración (Tabla Suplementaria 2, en inglés). Esta extensión está dirigida especialmente a los investigadores que planifican o realizan 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)[24] y a STARD-AI (Standards For Reporting Diagnostic Accuracy Studies-Artificial Intelligence)[51], ambas en fase de desarrollo. Otras directrices potencialmente relevantes, que son agnósticas al diseño del estudio, están registradas en la red EQUATOR[52]. Se espera que la extensión de SPIRIT-AI fomente la planificación temprana y cuidadosa de las intervenciones de IA para los ensayos clínicos y esto, junto con CONSORT-AI, debería ayudar a mejorar la calidad de los ensayos clínicos para las intervenciones de IA.

Existe un reconocimiento generalizado de que la IA es un campo que evoluciona rápidamente, por lo que será necesario actualizar SPIRIT-AI a medida que se desarrolle la tecnología y sus nuevas aplicaciones. En la actualidad, la mayoría de las aplicaciones de la IA/ML implican la detección, el diagnóstico y el triaje de enfermedades, y es probable que esto haya influido en la naturaleza y la priorización de los elementos dentro de SPIRIT-AI. A medida que surjan aplicaciones más amplias que utilicen la "IA como terapia", será importante reevaluar SPIRIT-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 el diseño de los estudios y en la presentación de informes para garantizar la transparencia, minimizar los posibles sesgos y asegurar que los hallazgos de un estudio de este tipo son fiables 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 datos. Las solicitudes de datos deben dirigirse al autor de correspondencia y su liberación estará sujeta a la consideración del Grupo Directivo de SPIRIT-AI y CONSORT-AI.

Contribución 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., A.W.C., M.J.C. y A.K.D.; obtención de financiación, A.K.D., M.J.C., C.Y. y C.H. 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. A los participantes en el estudio Delphi y el estudio piloto (Nota suplementaria); E. Marston (University of Birmingham, Reino Unido) por su apoyo estratégico; y C. Radovanovic (University Hospitals Birmingham NHS Foundation Trust, Reino Unido) y A. Walker (University of Birmingham, Reino Unido) por su apoyo administrativo; al Comité de Inteligencia Artificial de la Asociación Colombiana de Radiología (ACR) por la traducción; a Mauricio De Jesús Solano Diaz (Universidad de Antioquia) por su apoyo administrativo y editorial.

Financiación. Este trabajo fue financiado por un Fondo Institucional de Apoyo Estratégico del Wellcome Trust: Digital Health Pilot Grant Research England (parte de Investigación y Desarrollo del Reino Unido), Health Data Research UK y el Alan Turing Institute. El estudio fue patrocinado por la University of Birmingham, Reino Unido. Los financiadores y patrocinadores del estudio no participaron en el diseño y la realización del estudio; la recogida, la gestión, el análisis y la interpretación de los datos; la preparación, revisión o aprobación del manuscrito; o la decisión de presentar el manuscrito para su publicación.

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) fuera 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 ha participado 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.

M.J.C. es investigador senior del National Institute for Health Research (NIHR) y recibe fondos del National Institute for Health Research (NIHR) Birmingham Biomedical Research Centre; NIHR Surgical Reconstruction and Microbiology Research Centre y NIHR ARC West Midlands en la University of Birmingham y el University Hospitals Birmingham NHS Foundation Trust; Health Data Research UK; Innovate UK (parte de Investigación y Desarrollo del Reino Unido); Health Foundation; Macmillan Cancer Support; y UCB Pharma. A.D. y J.D. también son investigadores senior del NIHR. S.J.V. recibe fondos de Engineering and Physical Sciences Research Council, UK Research and Innovation (UKRI), Accenture, Warwick Impact Fund, Health Data Research UK y el European Regional Development Fund. S.R. es empleado del Medical Research Council (UKRI). D.M. recibe apoyo de University of Ottawa Research Chair. A.B. recibe apoyo de los National Institutes of Health (NIH) (asignación 7K01HL141771-02). M.K.E. recibe apoyo de la U.S. Food and Drug Administration (FDA), y D.P. recibe apoyo parcial de la Oficina del Director de la National Library of Medicine (NLM), US National Institutes of Health (NIH).

Declaración. Las opiniones expresadas en este manuscrito son únicamente responsabilidad de los autores, los participantes en el estudio Delphi y los participantes de las partes interesadas y no reflejan necesariamente las opiniones o políticas del NIH, la FDA, el NIHR, el Departamento de Salud y Asistencia Social, ni otras partes interesadas o instituciones huésped. Tampoco reflejan necesariamente los criterios ni la política de la *Revista Panamericana de Salud Pública* o de la Organización Panamericana de la Salud.

REFERENCIAS

1. Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. Ann Intern Med. 5 de febrero de 2013;158(3):200.
2. Chan AW, Tetzlaff JM, Gotzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ. 9 de enero de 2013;346(jan08 15):e7586-e7586.
3. Sarkis-Onofre R, Cenci MS, Demarco FF, Lynch CD, Fleming PS, Pereira-Cenci T, et al. Use of guidelines to improve the quality and transparency of reporting oral health research. J Dent. abril de 2015;43(4):397-404.
4. Calvert M, Kyte D, Mercieca-Bebber R, Slade A, Chan AW, King MT, et al. Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension. JAMA. 6 de febrero de 2018;319(5):483.
5. Dai L, Cheng CW, Tian R, Zhong LL, Li YP, Lyu AP, et al. Standard Protocol Items for Clinical Trials with Traditional Chinese Medicine 2018: Recommendations, Explanation and Elaboration (SPIRIT-TCM Extension 2018). Chin J Integr Med. enero de 2019;25(1):71-9.
6. 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.
7. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. enero de 2020;577(7788):89-94.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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. Lancet Digit Health. octubre de 2019;1(6):e271-97.
18. 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. diciembre de 2019;68(12):2161-9.
19. 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.
20. 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.

21. 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.
22. 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.
23. 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.
24. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 20 de abril de 2019;393(10181):1577-9.
25. 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.
26. 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.
27. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med. octubre de 2019;25(10):1467-8.
28. Liu X, Faes L, Calvert MJ, Denniston AK, CONSORT/SPIRIT-AI Extension Group. Extension of the CONSORT and SPIRIT statements. Lancet. 5 de octubre de 2019;394(10205):1225.
29. 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.
30. 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.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. 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.
37. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 29 de octubre de 2019;17(1):195.
38. 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 2024]. Disponible en: http://arxiv.org/abs/1909.01940
39. 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
40. Sabottke CF, Spieler BM. The Effect of Image Resolution on Deep Learning in Radiography. Radiol Artif Intell. enero de 2020;2(1):e190015.
41. Heaven D. Why deep-learning AIs are so easy to fool. Nature. octubre de 2019;574(7777):163-6.
42. 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.
43. 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.
44. 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.
45. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Proc ACM Conf Health Inference Learn (2020). abril de 2020;2020:151-9.
46. SPIRIT. Publications & downloads – GUIDANCE FOR CLINICAL TRIAL PROTOCOLS [Internet]. [citado 24 de marzo de 2020]. Disponible en: https://www.spirit-statement.org/publications-downloads/
47. 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.
48. 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.
49. Lee CS, Lee AY. Clinical applications of continual learning machine learning. Lancet Digit Health. junio de 2020;2(6):e279-81.
50. 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.
51. 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.
52. 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 DIRECTIVO SPIRIT-AI Y CONSORT-AI

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

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[8], Maria Beatrice Panico[26], Cecilia S. Lee[22], Charlotte Haug[27], Christophe J. Kelly[28], Christopher Yau[15,17], Cynthia Mulrow[29], Cyrus Espinoza[30], John Fletcher[31], David Moher[18,19], 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[2], Luke Oakden-Rayner[40], Melanie J. Calvert[1,2,3,6,10,11,12], Melissa McCradden[41], Pearse A. Keane1, Richard Savage[42], Robert Golub[43], Rupa Sarkar[44] y Samuel Rowley[45]

Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension

ABSTRACT

The SPIRIT 2013 statement aims to improve the completeness of clinical trial protocol reporting by providing evidence-based recommendations for the minimum set of items to be addressed. This guidance has been instrumental in promoting transparent 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 their impact on health outcomes. The SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence) extension is a new reporting guideline for clinical trial protocols evaluating interventions with an AI component. It was developed in parallel with its companion statement for trial reports: CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence). Both guidelines were developed through a staged consensus process involving literature review and expert consultation to generate 26 candidate items, which were consulted upon by an international multi-stakeholder group in a two-stage Delphi survey (103 stakeholders), agreed upon in a consensus meeting (31 stakeholders) and refined through a checklist pilot (34 participants). The SPIRIT-AI extension includes 15 new items that were considered sufficiently important for clinical trial protocols of AI interventions. These new items should be routinely reported in addition to the core SPIRIT 2013 items. SPIRIT-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 will be integrated, considerations for the handling of input and output data, the human–AI interaction and analysis of error cases. SPIRIT-AI will help promote transparency and completeness for clinical trial protocols for AI interventions. Its use will assist editors and peer reviewers, as well as the general readership, to understand, interpret and critically appraise the design and risk of bias for a planned clinical trial.

[13]Patient Safety Translational Research Centre, Imperial College London, Londres, Reino Unido. [14]Institute of Global Health Innovation, Imperial College London, Londres, Reino Unido. [15]Alan Turing Institute, Londres, Reino Unido. [16]Department of Statistics and Nuffield Department of Medicine, University of Oxford, Oxford, Reino Unido. [17]University of Manchester, Manchester, Reino Unido. [18]Centre for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canadá. [19]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Canadá. [20]Department of Ophthalmology, Cantonal Hospital Lucerne, Lucerna, Suiza. [21]University of Warwick, Coventry, Reino Unido. [22]Department of Ophthalmology, University of Washington, Seattle, Estados Unidos de América. [23]The National Institute for Health and Care Excellence, Londres, Reino Unido. [24]Salesforce Research, San Francisco, Estados Unidos de América. [25]Harvard T.H. Chan School of Public Health, Boston, Estados Unidos de América. [26]Medicines and Healthcare products Regulatory Agency, Londres, Reino Unido. [27]New England Journal of Medicine, Waltham, Estados Unidos de América. [28]Google Health, Londres, Reino Unido. [29]Annals of Internal Medicine, Filadelfia, Estados Unidos de América. [30]Patient Partner, Birmingham, Reino Unido. [31]British Medical Journal, Londres, Reino Unido. [32]National Institutes of Health, Bethesda, 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, Nueva York, Estados Unidos de América. [39]Food and Drug Administration, Silver Spring, Estados Unidos de América. [40]Australian Institute for Machine Learning, North Terrace, Adelaida, Australia. [41]The Hospital for Sick Children, Toronto, Canadá. [42]PinPoint Data Science, Leeds, Reino Unido. [43]Journal of the American Medical Association, Chicago, Estados Unidos de América. [44]The Lancet Group, Londres, Reino Unido. [45]Medical Research Council, Londres, Reino Unido.

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

RESUMO

A declaração SPIRIT 2013 tem como objetivo melhorar a integralidade dos relatórios dos protocolos de ensaios clínicos, fornecendo recomendações baseadas em evidências para o conjunto mínimo de itens que devem ser abordados. Essas orientações têm sido fundamentais para promover uma avaliação transparente 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 SPIRIT-AI (*Standard Protocol Items: Recommendations for Interventional Trials - Artificial Intelligence*) é uma nova diretriz de relatório para protocolos de ensaios clínicos que avaliam intervenções com um componente de IA. Essa diretriz foi desenvolvida em paralelo à sua declaração complementar para relatórios de ensaios clínicos, CONSORT-AI (*Consolidated Standards of Reporting 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 26 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 SPIRIT-AI inclui 15 itens novos que foram considerados suficientemente importantes para os protocolos de ensaios clínicos com intervenções que utilizam IA. Esses itens novos devem constar dos relatórios de rotina, juntamente com os itens básicos da SPIRIT 2013. A SPIRIT-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 será integrada, considerações sobre o manuseio dos dados de entrada e saída, a interação humano-IA e a análise de casos de erro. A SPIRIT-AI ajudará a promover a transparência e a integralidade nos protocolos 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 o delineamento e o risco de viés de um futuro estudo clínico.