deep learning AI model for detecting ARDS and explored the strengths, weaknesses, and blind spots of both physicians and AI systems to inform optimal system deployment. We then investigated several AI-physician collaboration strategies, including: 1) AI-aided physician: physicians interpret chest X-rays first and defer to the AI model if uncertain, 2) physician-aided AI: the AI model interprets chest X-rays first and defers to a physician if uncertain, and 3) AI model and physician interpreting chest X-rays separately and then averaging their interpretations. RESULTS/ANTICIPATED RESULTS: While the AI model (84.7% accuracy) had higher accuracy than physicians (80.8%), we found evidence that AI and physician expertise are complementary. When physicians lacked confidence in a chest X-ray's interpretation, the AI model had higher accuracy. Conversely, in cases of AI uncertainty, physicians were more accurate. The AI excelled with easier cases, while physicians were better with difficult cases, defined as those where at least two physicians disagreed with the majority label. Collaboration strategies tested include AI-aided physician (82.4%), physician-aided AI (86.9%), and averaging interpretations (86%). The physician-aided AI approach had the highest accuracy, could off-load the human expert workload on the reading of up to 79% chest X-rays, allowing physicians to focus on challenging cases. DISCUSSION/SIGNIFICANCE OF IMPACT: This study shows AI and physicians complement each other in ARDS diagnosis, improving accuracy when combined. A physician-aided AI strategy, where the AI defers to physicians when uncertain, proved most effective. Implementing AI-physician collaborations in clinical settings could enhance ARDS care, especially in low-resource environments.

22

Advancing clinical trial reporting and AI integration: Optimizing protocol data extraction using LLMs and regulatory best practices

Ramya Sri Baluguri and Nicholas Anderson
University of California, Davis

OBJECTIVES/GOALS: This study aimed to enhance clinical trial data management through large language model information retrieval and generation techniques within the clinical trial reporting workflow. We focused on improving compliance with reporting, reducing human labor, and promoting standardized reporting structure and data quality oversight. METHODS/STUDY POPULATION: We used approved study protocols from UC Davis IRB-approved investigator-initiated studies compared to the same studies reported to ClinicalTrials.gov. Our baseline data extraction system employs commercial large language models (LLMs) and retrieval augmented generation (RAG) to isolate data sources within the secure extraction environment. We stratified protocol documents into easy, complex, and random categories based on study focus, document complexity, the extent of amendments or modifications, and completion metrics from ClinicalTrials.gov. We developed a pilot web-based architecture to capture variations in categorization, labeling, and reporting style and compared generated extraction data. We primarily focused on qualitative evaluation through a review of expert staff. RESULTS/ANTICIPATED RESULTS: Our results revealed significant variations in reporting quality, with dependencies stemming from multiple authors and stages throughout the clinical trial protocol lifecycle. Based on these variations, we used prompt engineering to improve the pilot application's output compliance with the protocol registration and results system (PRS) structured data format for various study types. We piloted the assisted workflow with prospective studies by partnering with study investigators and the clinical trial office staff to assist in review and clinical trial reporting creation. Initial studies reported by our system were approved and released to the public by PRS staff. We are refining content generation and workflows to different components of studies and evaluating their use in quality and training areas. DISCUSSION/SIGNIFICANCE OF IMPACT: Our system fosters collaboration, efficient review, and compliance with clinical trial reporting standards. It supports the promise of AI-driven assistance in clinical trial management, design, and reporting. We focus on the multiple stakeholders, expertise, and data flows in the organizational management of clinical and translational science.

23

Generative artificial intelligence for automated unstructured MRI data extraction in prostate cancer care*†

William Pace, Andrew Liu, Marvin Carlisle, Robert Krumm, Janet Cowan, Peter Carroll, Matthew Cooperberg and Anobel Odisho
University of California, San Francisco

OBJECTIVES/GOALS: Magnetic resonance imaging (MRI) reports are stored as unstructured text in the electronic health record (EHR), rendering the data inaccessible. Large language models (LLM) are a new tool for analyzing and generating unstructured text. We aimed to evaluate how well an LLM extracts data from MRI reports compared to manually abstracted data. METHODS/STUDY POPULATION: The University of California, San Francisco has deployed a HIPAA-compliant internal LLM tool utilizing GPT-4 technology and approved for PHI use. We developed a detailed prompt instructing the LLM to extract data elements from prostate MRI reports and to output the results in a structured, computer-readable format. A data pipeline was built using the OpenAI Application Programming Interface (API) to automatically extract distinct data elements from the MRI report that are important in prostate cancer care. Each prompt was executed five times and data were compared with the modal responses to determine variability of responses. Accuracy was also assessed. RESULTS/ANTICIPATED RESULTS: Across 424 prostate MRI reports, GPT-4 response accuracy was consistently above 95% for most parameters. Individual field accuracies were 98.3% (96.3–99.3%) for PSA density, 97.4% (95.4–98.7%) for extracapsular extension, 98.1% (96.3–99.2%) for TNM Stage, had an overall median of 98.1% (96.3–99.2%), a mean of 97.2% (95.2–98.3%), and a range of 99.8% (98.7–100.0%) to 87.7% (84.2–90.7%). Response variability over five repeated runs ranged from 0.14% to 3.61%, differed based on the data element extracted (p DISCUSSION/SIGNIFICANCE OF IMPACT: GPT-4 was highly accurate in extracting data points from prostate cancer MRI reports with low upfront programming requirements. This represents an effective tool to expedite medical data extraction for clinical and research use cases.

https://doi.org/10.1017/cts.2024.714