Mitigating AI Bias & Privacy Risks in Digital Health Monitoring

Gaurangi Sinha, Saptarshi Das, Shreyas Kumar, Ruihong Huang, Srividhya Ragavan

Keywords: Artificial Intelligence, Cyber Security, Digital Privacy, healthcare innovations, Privacy

Category: Medical Science

Introduction: Artificial Intelligence (AI) powered public health surveillance systems are critical in tracking infectious diseases like COVID-19 and Ebola, enabling rapid response and containment. These technologies leverage vast amounts of data from various sources, including but not limited to electronic health records, social media, and wearable devices, to detect outbreaks and predict disease spread. However, despite their potential to enhance public health efforts, these systems present challenges concerning security, privacy, and ethics. Bias in AI algorithms could present themselves as data from external sources (a.k.a 'grounding') that are imbalanced or non-representative, which may perpetuate disparities in health monitoring, disproportionately affecting marginalized communities and leading to unequal healthcare responses. This integration of current industry AI and health monitoring raises concerns regarding inadequate privacy safeguards against current personal digital privacy, data security, and risks of unauthorized access or cyber attacks against one's sensitive personal health information. By addressing these concerns, we aim to enhance trust, equity, and the effectiveness of AI-driven disease surveillance in future public health initiatives.

Methods: The research will address biased evaluation and private information exposure risks from AI. Regarding bias evaluation & mitigation, 1) Identifying and refinding Data Sources by collecting publicly available health datasets in electronic health records (EHR). 2) Apply different AI models (GPT-4o, Claude 3.5, and LLaMA) against the different EHRs across different subgroups (e.g., by race, ethnicity, socioeconomic status, geographic location), and manipulate the data to simulate how the AI system's predictions would change if specific protected characteristics were different or altered. 3) Finally applying different statistical methods to analyze and identify significant differences in algorithm results and performance across subgroups against real world outcomes.

Results: As of Feb 26th, this research is in an early stage, with only a poster being created on the topic.

Conclusion: As of Feb 26th, this idea is in a very early stage, and conclusion will be provided after results