Аудит справедливости в клинических системах ИИ с использованием симуляции на основе прослеживаемости: сравнительная и регуляторная перспектива

Источник: Frontiers in AI — Medicine

Оригинал: https://www.frontiersin.org/articles/10.3389/frai.2026.1756023

алгоритмическая предвзятость | аудит | клиническая поддержка принятия решений | регулирование | этика ИИ

Введение

Внедрение искусственного интеллекта (**AI**) в системы поддержки принятия клинических решений столкнулось с препятствиями, вызванными алгоритмической предвзятостью и отсутствием прозрачности. Чтобы решить эту проблему, мы разработали систему аудита, использующую подробное отслеживание происхождения данных (**data provenance**) для проверки справедливости алгоритмов.

Методы

Мы провели симуляцию аудита на синтетическом наборе данных пациентов (N = 1 000), сравнивая модели логистической регрессии и случайного леса (**random forest**) для обнаружения гендерной предвзятости с использованием 5-кратной перекрестной проверки (**5-fold cross-validation**) и перестановочного тестирования (**permutation testing**).

Результаты

Логистическая регрессия достигла точности 75,2 ± 1,0% (AUC = 0,806 ± 0,030), а модель случайного леса достигла точности 70,1 ± 1,4% (AUC = 0,745 ± 0,020). Журналы происхождения данных успешно выявили гендерную предвзятость в обеих моделях. Логистическая регрессия продемонстрировала статистически значимую предвзятость (EOD = +0,256, p = 0,0080), в то время как меньшее неравенство в модели случайного леса (EOD = +0,055, p = 0,5664) не было статистически значимым, что доказывает способность нашего аудита отличать систематическую дискриминацию от случайных отклонений. Анализ чувствительности подтвердил успешное обнаружение предвзятости при значениях величины от β = −0,10 до β = −0,80.

Обсуждение

Несмотря на более низкую точность, модель случайного леса показала на 57% меньше предвзятости, чем логистическая регрессия, что ставит под сомнение предположения о том, что интерпретируемость гарантирует справедливость. Мы представляем стандартизированную запись происхождения справедливости **AI** (**AI Fairness Provenance Record**), документирующую происхождение данных, выбор моделей и метрики предвзятости, что позволяет аудиторам отслеживать решения до их первоисточника. Данная структура соответствует руководствам **FDA** (Управление по санитарному надзору за качеством пищевых продуктов и медикаментов) по прозрачности и требованиям **ONC HTI-1** (Национальный координатор по вопросам информационных технологий в здравоохранении — требования к технологической целостности), демонстрируя, как аудит на основе отслеживания происхождения поддерживает соблюдение нормативных требований и прокладывает путь к более ответственному и справедливому использованию **AI** в клинических условиях.