Ретроспективное сравнительное исследование ChatGPT-5.4 и DeepSeek-VL2 для стратификации рисков узлов щитовидной железы на основе C-TIRADS по ультразвуковым изображениям

Источник: Frontiers in Digital Health

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1877420

диагностика | мультимодальные модели | онкология | ультразвуковое исследование | щитовидная железа

Введение/цели

Мультимодальные большие языковые модели всё чаще исследуются для интерпретации медицинских изображений, однако их эффективность при ультразвуковой оценке узлов щитовидной железы остается неопределенной. В данном исследовании проводилось сравнение **ChatGPT-5.4** и **DeepSeek-VL2** для стратификации риска узлов щитовидной железы на основе системы **C-TIRADS** (Korean Thyroid Imaging Reporting and Data System) с использованием статических ультразвуковых изображений.

Методы

Данное одноцентровое ретроспективное исследование диагностической точности включало 206 патологически подтвержденных узлов щитовидной железы, из которых 117 были доброкачественными и 89 — злокачественными. Для каждого узла пары поперечных и продольных ультразвуковых изображений анализировались обеими моделями с использованием одного и того же стандартизированного англоязычного промпта. Модели генерировали структурированные выходные данные по

ультразвуковым признакам, связанным с **C-TIRADS**, и оценивали категорию **C-TIRADS**. Патоморфологическое исследование служило эталонным стандартом для определения злокачественности, а категория **C-TIRADS**, присвоенная экспертами, служила эталонным стандартом визуализации. Категории **C-TIRADS**, присвоенные моделями, использовались в качестве порядковых показателей риска. В основном анализе в качестве бинарного положительного порога для прогнозирования патологически подтвержденной злокачественности использовалось значение **C-TIRADS ≥4B**, а в анализе чувствительности использовалось значение **C-TIRADS ≥4A**.

Результаты

При пороге **≥4B** модель **ChatGPT-5.4** достигла более высоких показателей точности, чувствительности и специфичности, чем **DeepSeek-VL2**: 70,9%, 61,8% и 77,8% против 57,8%, 50,6% и 63,2% соответственно. Её показатель недиагностических случаев также был ниже (6,3% против 15,5%). В ROC-анализе (Receiver Operating Characteristic — рабочая характеристика приемника) валидных выходных данных значения **AUC** (Area Under the Curve — площадь под кривой) составили 0,903 для экспертной оценки, 0,790 для **ChatGPT-5.4** и 0,757 для **DeepSeek-VL2**; ROC-анализы на основе моделей исключали неопределенные результаты и должны интерпретироваться как оценки согласно протоколу (per-protocol). На уровне категорий **C-TIRADS** модель **ChatGPT-5.4** показала более высокую степень точного совпадения с экспертной оценкой, чем **DeepSeek-VL2** (45,6% против 36,4%), и более высокую степень совпадения в пределах одной категории (87,6% против 82,2%). Анализ на уровне признаков показал, что обе модели лучше всего справились с оценкой эхогенности, но продемонстрировали плохие результаты при оценке контуров и эхогенных фокусов. При пороге **≥4A** чувствительность увеличилась, но специфичность существенно снизилась для обеих моделей.

Выводы

ChatGPT-5.4 продемонстрировала более благоприятный профиль эффективности, чем **DeepSeek-VL2**, однако различия в чувствительности и **AUC** валидных выходных данных не были статистически значимыми. Обе модели остаются исследовательскими вспомогательными инструментами и не должны использоваться для принятия клинических решений без контроля специалиста или в качестве самостоятельного средства сортировки пациентов.