ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ	**ГОСТ Р (Проект, окончательная редакция)**

СИСТЕМЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В КЛИНИЧЕСКОЙ МЕДИЦИНЕ

Алгоритмы анализа медицинских изображений

Методы испытаний

Общие требования

Настоящий проект стандарта не подлежит применению до его утверждения

Москва

Российский институт стандартизации

20

Предисловие

1 РАЗРАБОТАН Федеральным государственным бюджетным учреждением «Российский институт стандартизации» (ФГБУ «РСТ»), Государственным бюджетным учреждением здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы» (ГБУЗ «НПКЦ ДиТ ДЗМ»)

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ приказом Федерального агентства по техническому регулированию и метрологии от №

4 ВВЕДЕН ВПЕРВЫЕ

*Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (*www.rst.gov.ru*)*

Содержание

1 Область применения 1
2 Нормативные ссылки 1
3 Термины и определения 2
4 Общие положения 4
5 Оценка соответствия системы искусственного интеллекта заявленным требованиям 5
6 Оценка клинической эффективности и безопасности 7
7 Оценка эффективности и безопасности СИИ на протяжении жизненного цикла 7
8 Функциональное тестирование 8
9 Уровень производительности 14
12 Тестирование на способность и устойчивость работы с разнородными данными 20
Приложение А (обязательное) Дизайн испытаний 25
Приложение Б (справочное) Пример заполнения протокола по результатам оценки клинической эффективности 31
Приложение В (справочное) Пример заполнения протокола по результатам функционального тестирования системы искусственного интеллекта 39
Библиография 40

Введение

Все медицинские изделия проходят стадию технических и клинических испытаний для целей регистрации в качестве медицинского изделия. Программное обеспечение является медицинским изделием в случае, если оно предназначено изготовителем для использования в медицинских целях, не является составной частью и (или) принадлежностью другого медицинского изделия, а также результат действия которого заключается в интерпретации данных в автоматическом режиме, в том числе с использованием технологий искусственного интеллекта, или по заданным медицинским работником параметрам, влияющим на принятие клинических решений, набора данных.

Системы искусственного интеллекта в клинической медицине представляют собой программное обеспечение, в котором используются технологии искусственного интеллекта. Системы искусственного интеллекта, функциональным назначением которых является анализ медицинских изображений (алгоритмы анализа медицинских изображений) относятся к программному обеспечению, являющемуся медицинским изделием. В связи с этим, должны быть проведены испытания алгоритмов анализа медицинских изображений для подтверждения воспроизводимости, надежности, безопасности использования и его эффективности в соответствии с предназначенным применением.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

СИСТЕМЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В КЛИНИЧЕСКОЙ МЕДИЦИНЕ

Алгоритмы анализа медицинских изображений

Методы испытаний

Общие требования

Artificial Intelligence Systems in Clinical Medicine. Algorithm of Medical Image Analysis. Test methods. General requirements

Дата введения — 20 – –

1 Область применения

Настоящий стандарт устанавливает общие требования к методам испытаний систем искусственного интеллекта, функциональным назначением которых является анализ медицинских изображений – алгоритмам анализа медицинских изображений.

Настоящий стандарт не приводит детальное описание методов испытаний стандартных параметров системы искусственного интеллекта как программного обеспечения, т.к. они представлены в соответствующих стандартах.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие документы:

ГОСТ ISO/IEC 17025−2019 Общие требования к компетентности испытательных и калибровочных лабораторий

ГОСТ Р 53114−2008 Защита информации. Обеспечение информационной безопасности в организации. Основные термины и определения

ГОСТ Р 56429−2021 Изделия медицинские. Клиническая оценка.

ГОСТ Р 59276−2020 Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения

ГОСТ Р 59898−2021 Оценка качества систем искусственного интеллекта. Общие положения

ГОСТ Р 59921.2−2021 Системы искусственного интеллекта в клинической медицине. Часть 2. Программа и методика технических испытаний

Издание официальное

ГОСТ Р 59921.3−2021 Системы искусственного интеллекта в клинической медицине. Часть 3. Управление изменениями в системах искусственного интеллекта с непрерывным обучением

ГОСТ Р 59921.4−2021 Системы искусственного интеллекта в клинической медицине. Часть 4. Оценка и контроль эксплуатационных параметров

ГОСТ Р ИСО 14155−2014 Клинические исследования. Надлежащая клиническая практика

ГОСТ Р ИСО/МЭК 9126−93 Информационная технология (ИТ). Оценка программной продукции. Характеристики качества и руководства по их применению

ГОСТ Р ИСО/МЭК 12119−2000 Информационная технология (ИТ). Пакеты программ. Требования к качеству и тестирование

ГОСТ Р ИСО/МЭК 25040−2014 Информационные технологии (ИТ). Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Процесс оценки

Примечание — При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающее положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

3 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями:

3.1 **автоматизированная система**: Система, состоящая из персонала и комплекса средств автоматизации его деятельности, реализующая информационную технологию выполнения установленных функций.

3.2

главный исследователь (principal investigator): Квалифицированное лицо, ответственное за проведение клинического исследования в исследовательском центре.

Примечания

1 Если клинические исследования проводятся группой лиц в исследовательском центре, главный исследователь ответственен за руководство данной группой.

2 Является ли ответственным конкретное лицо или институт, может зависеть от особенностей национальной системы регулирования медицинских изделий.

[ГОСТ Р ИСО 14155−2014, пункт 3.33]

3.3 **компаратор** (comparator): Медицинское изделие, лечебно-диагностическая процедура (например, активный контроль), плацебо или отсутствие лечения, используемые в контрольной группе при клиническом исследовании.

3.4

конечная точка (главная) (endpoint(s): Важнейший показатель, используемый для проверки основной гипотезы клинического исследования.

[ГОСТ Р ИСО 14155−2014, пункт 3.16]

3.5

отказ: Событие, заключающееся в нарушении работоспособного состояния объекта.

Примечания

1 Отказ может быть полным или частичным.

2 Полный отказ характеризуется переходом объекта в неработоспособное состояние.

Частичный отказ характеризуется переходом объекта в частично неработоспособное состояние.

[ГОСТ 27.002−2015]

3.6 **состязательный пример**: Входные данные для модели искусственного интеллекта, созданные путем применения небольших возмущений к рабочему примеру, что приводит к тому, что модель выводит неверный результат с высокой степенью достоверности.

Примечания

1 Обычно применяется к моделям искусственного интеллекта в форме нейронной сети.

2 См. [1]

4 Общие положения

4.1 Система искусственного интеллекта (СИИ), относящаяся к программному обеспечению, являющемуся медицинским изделием, должна проходить серию испытаний на стадии регистрации, включающую технические и клинические испытания. Кроме того, в течение жизненного цикла СИИ необходимо проводить контроль качества и безопасности СИИ.

4.2 В связи с тем, что СИИ относится к программному обеспечению, в рамках технических испытаний должны быть оценены ее метрики качества как программного продукта. Данные метрики и методы испытаний основаны на ГОСТ Р ИСО/МЭК 9126, ГОСТ Р ИСО/МЭК 25040, ГОСТ 28195, ГОСТ Р ИСО/МЭК 12119. Общие требования к номенклатуре метрик качества СИИ приведены в ГОСТ Р 59898 и ГОСТ Р 59276. Общие требования к данным испытаниям СИИ, включающую алгоритмы анализа медицинских изображений приведена в 5 разделе настоящего стандарта.

4.3 Для целей оценки клинической эффективности и безопасности СИИ, относящейся к программному обеспечению, являющемуся медицинским изделием, выполняют клинические испытания (раздел 6 настоящего стандарта).

4.4 Общая методология и описание испытаний приведены в таблице А.1 приложения А.

4.5 Для целей выполнения испытаний производитель СИИ должен предоставлять пользователям прозрачную информацию о предназначении разработки, включая следующие данные:

- клиническое предназначение СИИ;

- целевую популяцию и условия, для которых предназначена СИИ (учреждения первичного и (или) других звеньев здравоохранения, носимая электроника и др.);

- предполагаемые результаты использования СИИ для пациентов и системы здравоохранения.

4.6 Предоставленной производителем информации должно быть достаточно для того, чтобы пользователи могли оценить применимость СИИ для целевого клинического предназначения. В описании должны содержаться следующие данные:

- заявленные значения метрики, используемой для различения субъектов с наличием и отсутствием целевого признака. Метрика должна быть выбрана в согласии с медицинскими экспертами в целевой области и соответствовать медицинскому предназначению модели.

Пример − Если СИИ предназначена для оценки вероятности рака груди, она должна обладать высокой чувствительностью

- характеристики данных, используемых для обучения и тестирования модели;

- тип и формат входных данных;

- известные ограничения модели;

- целевой метод медицинской визуализации (рентгенография, магнитно-резонансная томография, ультразвук, компьютерная томография и т.д.). Элементы пользовательского интерфейса должны соответствовать требованиям стандартов ГОСТ Р 55544, [2] и [3].

5 Оценка соответствия системы искусственного интеллекта заявленным требованиям

5.1 Для оценки соответствия характеристик СИИ требованиям технической и эксплуатационной документации изготовителя, заявленным стандартам, применимым регулирующим требованиям и т.д. выполняют технические испытания. Результаты тестирования должны быть воспроизводимы на различном совместимом оборудовании и представлены с использованием стандартных метрик оценки соответствующих характеристик.

5.2 С целью выполнения технических испытаний разрабатывают программу испытаний (см. ГОСТ Р 59921.2), включающую методику проведения испытаний существенных характеристик СИИ, определенных в ГОСТ Р 59276 и ГОСТ Р 59898.

5.3 СИИ, применяемая в клинической медицине для анализа медицинских изображений, имеет отличия от общих СИИ. Например, качество данных медицинской визуализации может существенно варьировать в рамках системы здравоохранения: иметь разное соотношение сигнал/шум, дефекты различной природы, изменяться со временем из-за износа и (или) обновления оборудования. В связи с этим были введены дополнительных характеристики. Соответствие существенных характеристик СИИ и методов испытаний, приведенных в настоящем стандарте, перечислены в таблице 1.

Испытания некоторых характеристик СИИ будут аналогичными испытаниям общих программных продуктов, поэтому для данных пунктов приведены нормативные ссылки.

Таблица 1 – Соответствие существенных характеристик системы искусственного интеллекта и методов испытаний настоящего стандарта

Группа характеристик	Характеристика по ГОСТ Р 59276	Существенная характеристика	Пункт настоящего стандарта/нормативная ссылка
Функциональность	Функциональные возможности	Функциональные возможности (functionality) Способность к взаимодействию (compatibility)	пункт 8
	Эффективность	Уровень производительности (performance efficiency)	пункт 9
	Мобильность	Мобильность (portability)	ГОСТ Р ИСО/МЭК 12119, ГОСТ Р ИСО/МЭК 9126
	Практичность	Практичность (usability)	ГОСТ Р ИСО/МЭК 12119, ГОСТ Р ИСО/МЭК 9126
	Сопровождаемость	Сопровождаемость (maintainability)	ГОСТ Р ИСО/МЭК 12119, ГОСТ Р ИСО/МЭК 9126
Надежность	Надежность	Надежность (reliability)	пункты 10-12
Безопасность	—	Защищенность (security)	пункт 13

5.4 Перед выполнением тестирования необходимо осуществить подготовительные работы (см. ГОСТ Р 59898, пункт 7.2.1). Подготовительные работы включают следующие этапы: установление целей и задач тестирования; определение критериев к набору данных для тестирования; выбор и обоснование существенных характеристик и метрик их оценки; определение допустимого диапазона изменений метрик, а также наиболее существенных факторов, оказывающих влияние на работу СИИ; утверждение состава экспертной группы, составление методики тестирования и подготовка программы тестирования.

6 Оценка клинической эффективности и безопасности

6.1 Оценка клинической эффективности и безопасности СИИ является целью клинических испытаний. Эффективность и характеристики СИИ должны быть протестированы в клинически значимых условиях. Целью испытаний является сбор доказательств клинической эффективности СИИ при минимизации рисков для пациентов и потенциального негативного влияния СИИ на функционирование организации здравоохранения, в рамках которой проходят испытания.

6.2 Уровень вмешательства СИИ в клинических испытаниях может варьировать от работы в фоновом режиме, когда СИИ функционирует одновременно и параллельно с существующими методами оказания медицинских услуг, до полноценных испытаний СИИ в соответствии с целевым предназначением (см. ГОСТ Р 56429).

6.3 Метрики клинической эффективности и безопасности СИИ должны включать в себя конечные точки (см. ГОСТ Р ИСО 14155), имеющие значение для пациентов (например, детекцию и классификацию патологии, оценку клинического исхода и т.п.).

6.4 Программа проведения клинических испытаний должна соответствовать критериям прозрачности, а также рекомендовано включать описание пунктов, приведенных в приложении Б.

6.5 Представление результатов клинических испытаний должно соответствовать международным рекомендациям, таким как CONSORT-AI для клинических контролируемых испытаний вмешательств с использованием СИИ [4], эквивалентных инструментов для исследований диагностической точности СИИ (STARD-AI) [5], и испытаний предсказательных моделей (TRIPOD-AI) [6].

7 Оценка эффективности и безопасности СИИ на протяжении жизненного цикла

7.1 Качество данных, методы сбора данных (в том числе модели и версии медицинского оборудования), характеристики популяции, принятые клинические практики могут изменяться со временем, что способно оказать влияние на достоверность результатов и клиническую применимость СИИ.

7.2 Внесение изменений в СИИ и ее (само)обучение на этапе эксплуатации также способны повлиять на их рабочие характеристики. В зависимости от вида изменений и модификаций СИИ, а также при ее (само)обучении может возникнуть необходимость проведения новых испытаний (см. ГОСТ Р 59921.3).

7.3 Выполнение контроля клинической эффективности СИИ на протяжении жизненного цикла должен предоставить доказательства отсутствия снижения технических и клинических характеристик модели. Рекомендуется проводить оценку пользы от применения СИИ для пациентов и системы здравоохранения, а также идентифицировать и устранять потенциальные и реальные источники непредумышленного вреда и систематических ошибок.

7.4 Контроль безопасности и эффективности СИИ необходимо оценивать путем проведения мониторинга в соответствии с действующими нормативными правовыми актами.

8 Функциональное тестирование

8.1 Назначение

Функциональное тестирование выполняют с целью оценки соответствия функциональных возможностей СИИ требованиям, указанным в техническом задании на создание это продукта. Согласно ГОСТ Р 59898 в рамках функционального тестирования рекомендовано оценивать следующие метрики: функциональная пригодность (functional appropriateness), функциональная корректность (functional correctness), согласованность (compliance), функциональная полнота (functional completeness) и способность к самообучению (ability to learn).

Согласно ГОСТ Р ИСО 25040 функциональное тестирование СИИ проводят методом «черного ящика» с контролем документации изготовителя.

Данный вид тестирования соответствует понятию аналитической валидации СИИ, как программного обеспечения, являющегося медицинским изделием, в рамках клинической оценки СИИ (см. [7]).

8.2 Требования к квалификации персонала

Для выбора номенклатуры характеристик, подготовки набора данных, а также для проведения тестирования и оценки результатов создают экспертную группу, которая должна удовлетворять требованиям ГОСТ Р 59898 (пункт 7.1.2).

Непосредственно выполнение тестирования в рамках функционального тестирования проводят силами квалифицированных технических специалистов, имеющих опыт тестирования СИИ, обработки изображений и анализа данных результатов. Также возможно привлечение к работе СИИ для анализа функциональных характеристик и подготовки набора данных медицинских специалистов, которые имеют опыт работы и квалификацию по областях, соответствующих решаемым задачам СИИ.

Тестирование проводит независимый от изготовителя СИИ коллектив исследователей; в противном случае в протоколе указывают источники конфликта интересов.

8.3 Описание метода тестирования

8.3.1 Тестирование выполняют на небольшом наборе данных путем расчета выбранного набора метрик. Возможно использование автоматических методов тестирования (см. [1] пункт 7.9).

8.3.2 Метрики функциональной пригодности и способности к самообучению определяют согласно общим требованиям ГОСТ Р 59898 (пункт 8.2).

8.3.3 Оценку метрик функциональной полноты выполняют регистрационным методом (см. ГОСТ 28195, пункт 1.5), путем расчета отношения количества недостающих или неправильно реализованных функций к общему количеству функций СИИ, указанных в технической и эксплуатационной документации (см. ГОСТ Р 59898, пункт 8.2.2).

Пример − Может быть произведена анализ следующих параметров:

- визуальная оценка выходных данных СИИ;

- тип принимаемого решения;

- набор заключений СИИ;

- формат входных данных;

- тип анализируемого объекта;

- и тд.

8.3.4 Метрику согласованности оценивают путем анализа документации на соответствие стандартам или нормативным правовым актам, или других рекомендаций (см. ГОСТ Р ИСО/МЭК 9126, пункт А.2.1.4). Техническая и эксплуатационная документация на СИИ должна включать описание эксплуатационных параметров, приведенных в ГОСТ Р 59921.4.

8.3.5 Метрики функциональной корректности (правильности) используют для подтверждения того, что СИИ генерирует выходные данные с надлежащим уровнем точности (accuracy), а также повторяемости (repeatability) и воспроизводимости (reproducibility) (см. ГОСТ Р 59921.2). Оценку данной метрики выполняют расчетным методом (см. ГОСТ 28195, пункт 1.5), на небольшом наборе данных с последующем определением соответствующих метрик (см. ГОСТ Р 59898, пункт 8.2.3). Набор метрик определяют на основании вида решаемой задачи СИИ и устанавливают решением экспертной группы.

8.4 Требования к наборам данных

Функциональное тестирование проводят на небольшом наборе данных, включающим изображения с отсутствием целевой патологии, а также изображения с целевой патологией. Набор данных должен быть верифицирован, а также должен включать данные из разных медицинских организаций и разных моделей/производителей оборудования, обработку данных с которых изготовитель СИИ включает в функциональное назначение.

8.5 Требования к оборудованию для проведения тестирования

Функциональное тестирование необходимо проводить с применением испытательного стенда (может быть как отдельным автоматизированным рабочим местом, так и виртуальной тестовой средой, которая удовлетворяет техническим требованиям СИИ (см. [1] пункт 10).

Возможно использование виртуальной тестовой среды, позволяющей выполнять автоматические тестирования (benchmarking) (см. [1] пункт 7.9).

Перед выполнением тестирования необходимо удостовериться в отсутствии существенных различий между средой проведения тестирования и средой эксплуатации СИИ (см. ГОСТ Р 59898, пункт 7.2).

8.6 Порядок выполнения тестирования

8.6.1 Тестирование проводят в соответствии с программой испытаний после выполнения подготовительных работ (см. 5.4).

8.6.2 Набор данных загружают на испытательный стенд, в котором инсталлирована СИИ.

8.6.3 Обрабатывают элементы набора данных с использованием СИИ и фиксируют выходные данные СИИ.

8.6.4 При проведении тестирования документируют все программные и прочие ошибки, возникающие при эксплуатации СИИ.

8.6.5 При тестировании функциональной полноты необходимо предварительно определить номенклатуру функций СИИ, описанных в технической и эксплуатационной документации. Рекомендовано включить в перечень параметры, указанные в ГОСТ Р 59921.4:

- визуальная оценка выходных данных СИИ;

- тип принимаемого решения (тип (бинарное, вероятностное, локализация, классификация) и формат возвращаемого СИИ заключения);

- набор заключений СИИ (СИИ может формировать одно или больше заключений (многофакторное заключение));

- формат входных данных и тип анализируемого объекта (входные данные для систем анализа медицинских изображений могут быть представлены статичными и динамическими изображениями анализируемой области; сериями двумерных изображений, соответствующих сечениям анализируемой области; содержать метаданные).

При выполнении тестирования каждой функции, результаты оформляют в табличном виде, пример представлен в таблице 2.

Таблица 2 – Анализ и оценка результатов тестирования на функциональную полноту

Наименование параметра	Оценочные данные	Результат
Визуальная оценка выходных данных СИИ	Соответствие заявленным в технической документации	Соответствует/не соответствует
Тип принимаемого решения	Соответствие заявленным в технической документации	Соответствует/не соответствует
Формат возвращаемого СИИ заключения	Соответствие заявленным в технической документации	Соответствует/не соответствует
Набор заключений СИИ	Соответствие заявленным в технической документации	Соответствует/не соответствует
Формат входных данных	Соответствие заявленным в технической документации	Соответствует/не соответствует
Тип анализируемого объекта	Соответствие заявленным в технической документации	Соответствует/не соответствует

8.6.6 При тестировании функциональной корректности выполняют тестирование СИИ на небольшом наборе данных, достаточного объема для получения статистически значимого результата.

Порядок тестирования определяют исходя из установленного набора метрик:

- для оценки метрик точности выполняют регистрацию выходных данных СИИ при обработке набора данных;

- для оценки повторяемости выполняют повторную обработку набора данных в одинаковых условиях тестирования и регистрацию выходных данных СИИ;

- для оценки воспроизводимости выполняют повторную обработку одного набора данных в разных условиях испытаний (разные испытательные стенды, операторы, распределение во времени и т.д.) и регистрацию выходных данных СИИ.

Пример − Есть задача сегментации ткани на изображении магнитно-резонансной томографии, которая требует от пользователя ввода начальных точек или выбора срезов для анализа СИИ. Для данной задачи определение повторяемости будет заключаться в повторном анализе одного изображения 7 раз одним и тем же оператором (для получения степени свободы не менее 6). Определение воспроизводимости может быть выполнено путем анализа данного изображения 2 операторами или одним оператором на разных испытательных стендах.

8.7 Анализ и оценка результатов тестирования

8.7.1 Оценка функциональной полноты возможна путем расчета метрики полноты реализации функций. Допустимо использовать другие показатели.

Полнота реализации функций определяется согласно (см. ГОСТ Р 59898, пункт 8.2.2):

$$M_1 = 1 - \frac{A}{B}, \qquad (1)$$

где A – количество недостающих или неправильно реализованных функций, обнаруженных при тестировании;

B – общее количество функций, описанных в технической и эксплуатационной документации СИИ.

8.7.2 Оценка функциональной корректности возможна методом расчета метрик точности, воспроизводимости и повторяемости.

Метрики точности в зависимости от решаемой задачи СИИ и формулы расчета приведены в ГОСТ Р 59898 (пункт 8.2.3):

1) общая метрика – результативность;

2) для задач регрессии – средняя квадратичная ошибка, средняя абсолютная ошибка;

3) в задачах ранжирования – приведенная суммарная эффективность.

4) для задач классификации и обнаружения – доля правильных исходов, точность, чувствительность, специфичность, F-мера, площадь под кривой ROC, площадь под кривой PRC. Возможно расширение набора метрик [8], [9], на основании решения экспертной группы.

Методы расчета воспроизводимости и повторяемости приведены в ГОСТ Р ИСО 5725-2.

8.8 Критерии оценки и представление результатов

Все метрики нормируют, чтобы их значения были в интервале от 0 до 1. Весовые коэффициенты метрик и допустимый интервал определяют члены экспертной группы.

Критерии оценки определяют в соответствии с предназначением СИИ и заявленными характеристиками.

По завершении тестирования составляют отчет, в котором должны быть указаны в том числе перечень метрик, методы испытаний, характеристики наборов данных, допустимые интервалы, а также определенные значения параметров (см. таблицу 3).

Результаты функционального тестирования оформляют в виде протокола испытаний в соответствии с ГОСТ ISO/IEC 17025.

Таблица 3 – Анализ и оценка результатов функционального тестирования

Наименование характеристики	Нормативное значение	Определенное значение	Результат
Функциональная полнота	от … до …		Соответствует/не соответствует
Функциональная корректность			
- точность	от … до …		Соответствует/не соответствует
- воспроизводимость	от … до …		Соответствует/не соответствует
- повторяемость	от … до …		Соответствует/не соответствует
…			
Функциональная пригодность	от … до …		Соответствует/не соответствует
Способность к самообучению	от … до …		Соответствует/не соответствует

Оценка согласованности выполняется описательным методом, в связи с этим заполняется таблица по примеру 4.

Таблица 4 – Анализ и оценка результатов функционального тестирования - согласованность

Наименование параметра	Оценочные данные	Результат
Согласованность	Соответствие нормативным требованиям	Соответствует/не соответствует

Пример заполнения протокола по результатам функционального тестирования СИИ приведен в приложении В.

9 Уровень производительности

9.1 Назначение

В рамках оценки уровня производительности (performance efficiency) рекомендуют оценивать следующие характеристики: характер изменения во времени, характер изменения ресурсов и производительные возможности (см. ГОСТ Р 59898, пункт 8.3,).

9.2 Общие требования к проведению испытания

Для выбора номенклатуры характеристик, подготовки набора данных, а также для проведения тестирования и оценки результатов создают экспертную группу, которая должна удовлетворять требованиям ГОСТ Р 59898 (пункт 7.1.2).

Непосредственно выполнение тестирования в рамках функционального тестирования проводят силами квалифицированных технических специалистов, имеющих опыт тестирования СИИ.

Набор данных для тестирования уровня производительности должен включать такой объем, чтобы выполнить тестирование на максимальном уровне загрузке СИИ. Наборы данных для данного тестирования не должны включать примеры медицинских изображений, резко отличающиеся от нормы и (или) патологии, применительно как к качеству изображения, так и к его содержанию.

Тестирование выполняют на испытательном стенде (может быть как отдельным автоматизированным рабочим местом, так и виртуальной тестовой средой, которая удовлетворяет техническим требованиям СИИ (см. [1] пункт 10).

Возможно использование виртуальной тестовой среды, позволяющей выполнять автоматические тестирования (benchmarking) (см. [1] пункт 7.9).

Перед выполнением тестирования необходимо удостовериться в отсутствии существенных различий между средой проведения тестирования и средой эксплуатации СИИ (см. ГОСТ Р 59898, пункт 7.2).

9.3 Метрики

Метрики характера изменения во времени (time behaviour) характеризуют соответствие требованиям временных ресурсов на выполнение операции, а также производительность – количество однотипных задач за определенное время. Методика расчета приведена в ГОСТ Р 59898 (пункт 8.3.4).

Метрики характера изменения ресурсов (resource utilization) относят к объему используемых ресурсов и продолжительности такого использования при выполнении функции (см. ГОСТ Р 59898, пункт 8.3.5).

Метрики производительных возможностей (capacity) характеризуют степень соответствия требованиям таких параметров СИИ как количество параллельно обрабатываемых наборов данных, количество параллельно работающих пользователей, емкость канала, пропускная способность по транзакциям.

9.4 Обработка результатов и оформление протокола

В протокол испытания включают параметры, влияющие на уровень производительности: параметры центрального процессора, объем памяти, объем хранения, сетевой трафик и прочее.

Результаты тестирования оформляют в виде протокола испытаний в соответствии с ГОСТ ISO/IEC 17025 (см. таблицу 5).

Таблица 5 – Анализ и оценка результатов тестирования уровня производительности

Наименование характеристики	Нормативное значение	Определенное значение	Результат
Характер изменения во времени	от … до …		Соответствует/не соответствует
Характер изменения ресурсов	от … до …		Соответствует/не соответствует
Производительная возможность	от … до …		Соответствует/не соответствует

10 Надежность

10.1 Назначение

Надежность СИИ определяют как способности СИИ сохранять свой уровень качества функционирования при установленных условиях за установленный период времени (см. ГОСТ Р ИСО/МЭК 9126).

К метрикам надежности СИИ относят как общие характеристики, такие как стабильность (maturity), устойчивость к ошибке (fault tolerance), восстанавливаемость (recoverability), робастность (robustness). Также к метрикам надежности СИИ относят устойчивость к состязательным атакам (см. раздел 11) и показатели способности и устойчивости работы с разнородными данными (см. раздел 12).

10.2 Общие требования

Для выбора номенклатуры характеристик, подготовки набора данных, а также для проведения тестирования и оценки результатов создают экспертную группу, которая должна удовлетворять требованиям ГОСТ Р 59898 (пункт 7.1.2).

Непосредственно выполнение тестирования в рамках тестирования надежности проводят силами квалифицированных технических специалистов, имеющих опыт тестирования СИИ.

Наборы данных для данного тестирования не должны включать примеры медицинских изображений, резко отличающиеся от нормы и (или) патологии, применительно как к качеству изображения, так и к его содержанию.

Тестирование выполняют на испытательном стенде (может быть как отдельным автоматизированным рабочим местом, так и виртуальной тестовой средой, которая удовлетворяет техническим требованиям СИИ (см. [1] пункт 10),

Возможно использование виртуальной тестовой среды, позволяющей выполнять автоматические тестирования (benchmarking) (см. [1] пункт 7.9).

Перед выполнением тестирования необходимо удостовериться в отсутствии существенных различий между средой проведения тестирования и средой эксплуатации СИИ (см. ГОСТ Р 59898, пункт 7.2).

Для тестирования надежности многократно повторяют обработку тестовых наборов, объем которых определяется экспертной группой, но должен быть не менее 1000 циклов.

В процессе тестирования регистрируют выходные данные СИИ с указанием тип ошибки, триггер ошибки, влияние ошибки, а также определяют время работы СИИ до момента возникновения ошибки.

Проводят оценку ущерба, связанного с каждой возникающей ошибкой. Важным фактором, влияющим на оценку надежности СИИ, являются ошибки, приводящие к снижению безопасности системы.

10.3 Метрики

10.3.1 Метрики стабильность и робастность приведены в ГОСТ Р 59898 (пункт 8.8). Метрика робастности включает тестирования на устойчивость к состязательным атакам (см. раздел 11) и тестированию на способность и устойчивость работы с разнородными данными (см. раздел 12).

10.3.2 Среднее время наработки на отказ определяют как отношение суммы интервалов безотказной работы вне зависимости от типа ошибки к количеству отказов в системе.

10.3.3 Устойчивость к ошибкам и отказам, показывающая способность СИИ выполнять функции при аномальных условиях (сбой аппаратуры, некорректные действия оператора и т.д.). Определяется как отношение количества разных типов отказов, для которых предусмотрены средства восстановления, к общему типу отказов в испытаниях.

10.3.3 Восстанавливаемость, показывающая способность СИИ возобновить работу после возникновения ошибки. Вычисляется как сумма временных интервалов восстановления работоспособности системы после ошибки к количеству ошибок и отказов, зафиксированному в испытаниях.

10.4 Обработка результатов и оформление протокола

Результаты тестирования надежности оформляют в виде протокола испытаний в соответствии с ГОСТ ISO/IEC 17025 (см. таблицу 6). Данные по тестированию на устойчивость к состязательным атакам и к работе с разнородными данными приводят отдельно.

Таблица 6 – Анализ и оценка результатов тестирования надежности

Наименование характеристики	Нормативное значение	Определенное значение	Результат
Метрики стабильности	от … до …		Соответствует/не соответствует
Метрики устойчивости к ошибке	от … до …		Соответствует/не соответствует
Метрики восстанавливаемости	от … до …		Соответствует/не соответствует
Метрики робастности	от … до …		Соответствует/не соответствует
Среднее время наработки на отказ	от … до …		Соответствует/не соответствует

11 Тестирование на устойчивость к состязательным атакам

11.1 Назначение

Состязательной атакой называют такой состязательный пример, который приводит к некорректной работе СИИ. Состязательный пример возникает, в случае крайне небольшого изменения, внесенного во входные данные алгоритма ИИ, который приводит к неожиданному (и неправильному) большому изменению выходных данных, т. е. к совершенно другому результату, чем при неизменных входных данных. Состязательные примеры могут возникать, например, при изменении несколько пикселей на медицинском изображении, искажение которых незаметно человеческому глазу.

Тестирование на устойчивость к состязательным атакам основано на формировании и обработки состязательных примеров для выявления дефектов в СИИ. Выполняя данное тестирование можно оценить устойчивость и надежность СИИ к наличию на входе системы данных, приводящих к возникновению состязательного примера [10], [1].

11.2 Требования к квалификации персонала

Тестирование на устойчивость к состязательным атакам проводят силами квалифицированных технических специалистов, имеющих опыт тестирования СИИ, обработки изображений и анализа результатов.

11.3 Описание метода тестирования

Оценка точности предсказаний алгоритма ИИ и (или) других метрик в условиях состязательных атак и в нормальных условиях.

11.4 Требования к оборудованию

Тестирование на устойчивости СИИ к состязательным атакам должно проводиться с применением стенда, в виртуальной тестовой среде, которая удовлетворяет техническим требованиям СИИ (см. [1]).

11.5 Требования к набору данных

Набор данных для выполнения данного тестирования должен состоять из изображений, которые соответствуют функциональному назначению СИИ. При этом половину набора данных составляют данные в нормальных условиях, а вторую – данные с внесенными искажениями для формирования состязательных примеров.

Методы формирования состязательных примеров могут быть добавление универсального случайного возмущения (universal random perturbation), шума (imperceptible noise), и другими методами (FGSM, inverse FGSM, JSMA) [11].

Специалисты, выполняющие данное тестирование, устанавливают виды состязательных примеров, влияние которых будет оценено, в соответствии с указанными в технической документации производителя.

11.6 Порядок выполнения тестирования

Формируют набор данных для выполнения тестирования на устойчивость к состязательным атакам согласно установленных требований.

Определяют выходные данные для каждого элемента набора данных, поданного на вход СИИ.

11.7 Анализ и оценка результатов тестирования

Сопоставляют выходные данные СИИ с заданными значениями набора данных. Вычисляют метрики качества, которые определяются типом принимаемого решения СИИ и ее функциональным назначением [12].

Составляют ранжированный список состязательных атак, упорядоченный по степени влияния на алгоритм.

11.8 Критерии оценки (метрики) и оформление результатов

Критерии оценки устойчивости СИИ к состязательным атакам определяют исходя из требований, приведенных в технической документации СИИ.

Пример − СИИ анализирует оптические изображения глаза для сегментации области глаукомы. На исходном изображении метрика точности сегментации (индекс Jaccard) составляла 71,1%. После внесения изменений в исходное изображение, чтобы получить состязательный пример, значение метрики снизилось до 16,4% [13].

Результаты оценки устойчивости СИИ к состязательным атакам оформляют в виде протокола испытаний в соответствии с ГОСТ ISO/IEC 17025 (см. таблицу 7).

Таблица 7 – Анализ и оценка тестирования на устойчивость СИИ к состязательным атакам

Наименование характеристики	Нормативное значение	Определенное значение	Результат
Метрика качества на исходных изображениях	от … до …	…	Соответствует/ не соответствует

Окончание таблицы 7

Наименование характеристики	Нормативное значение	Определенное значение	Результат
Метрика качества на изображениях – состязательных примерах (1 вид)	от … до …	…	Соответствует/ не соответствует
Метрика качества на изображениях – состязательных примерах (2 вид)	от … до …	…	Соответствует/ не соответствует
…			

12 Тестирование на способность и устойчивость работы с разнородными данными

12.1 Назначение

Тестирование проводят для проверки соответствия требованиям СИИ к способности и устойчивости работы с разнородными входными данными. С целью выполнения данного тестирования формируют соответствующий набор данных.

12.2 Требования к квалификации персонала

Тестирование на способность и устойчивость работы с разнородными данными проводят силами квалифицированных технических специалистов, имеющих опыт тестирования СИИ, обработки изображений и анализа результатов.

12.3 Описание метода тестирования

Моделирование типовых задач СИИ с различными наборами данных и диагностическими устройствами.

Данное тестирование оценивает способность СИИ успешно обрабатывать входные данные, содержащие разного рода искажения путем расчета метрик качества.

12.4 Требования к наборам данных

12.4.1 Наборы данных должны содержать входные данные в формате, описанном в технической документации СИИ. Допускается включение в набор примеров данных в ином формате для исследования способности СИИ распознавать и анализировать альтернативно организованную информацию.

12.4.2 Наборы данных должны включать данные, полученные на разной аппаратуре, исследования с которой СИИ потенциально способна обработать (в том числе не указанные в технической документации).

12.4.3 Наборы данных должны включать весь возможный диапазон размеров и разрешений изображений, который можно получить на медицинском оборудовании, актуальном на момент проведения тестирования.

12.4.4 Допускается включение в набор данных зашумленных или искаженных данных (изображения, подверженные разного рода трансформациям, изображения, содержащие артефакты).

12.4.5 Наборы данных должны содержать изображения, включающие объекты, которые не соответствуют функциональному назначению СИИ, но могут быть в результате ошибки маршрутизации поданы на вход СИИ (например, изображения с другими анатомическими областями, белый шум, инородные предметы, и др.).

12.4.6 Наборы данных должны включать также сложные для интерпретации экспертами изображения.

12.5 Требования к оборудованию

Тестирование должно быть проведено с применением стенда, в виртуальной тестовой среде, которая удовлетворяет техническим требованиям СИИ [1].

12.6 Порядок выполнения тестирования

Для каждого элемента сформированного набора данных получают выходные данные СИИ с использованием стенда.

Регистрируют выходные данные и сформированные ошибки и сообщения СИИ.

Перечисляют формат и тип входных данных, наименования медицинского оборудования, размеры и разрешения изображений, типы зашумленных и искаженных изображений, которые были поданы на вход системы (см. таблицу 8).

Таблица 8 – Характеристики входных данных

Наименование	Характеристика
Формат входных данных	
Тип входных данных	
Наименование медицинского оборудования	
Наименование диагностического устройства	
Размер изображений	
Разрешение изображений	

Окончание таблицы 8

Наименование	Характеристика
Характеристики изображений (зашумленных, искаженных)	- шум - искажение изображения (описание) - другая анатомическая область (описание) - инородные предметы (описание) - др.

12.7 Анализ и оценка результатов тестирования

Рассчитывают показатели качества СИИ для соответствующего набора данных.

Для каждой выявленной ошибки СИИ определяют параметры: тип ошибки и триггер ошибки.

12.8 Критерии оценки (метрики)

Анализируют изменяются ли метрики качества при подаче на вход СИИ разнородного набора данных.

Возможно использовать также метрики точности определения некорректных изображений, которые не соответствуют функциональному назначению СИИ.

Результаты тестирования на способность и устойчивость работы с разнородными данными оформляют в виде протокола испытаний в соответствии с ГОСТ ISO/IEC 17025 (см. таблицы 9, 10).

Таблица 9 – Анализ и оценка результатов тестирования на способность и устойчивость работы с разнородными входными данными с оценочными результатами

Наименование параметра	Оценочные данные	Результат
Визуальная оценка выходных данных СИИ	Соответствие заявленным в технической документации	Соответствует/не соответствует
Уведомление пользователя о поступлении на вход изображения инородного предмета	Имеется	Соответствует/не соответствует

Окончание таблицы 9

Наименование параметра	Оценочные данные	Результат
Уведомление пользователя о поступлении на вход изображения не соответствующей назначению анатомической области	Имеется	Соответствует/не соответствует
….		

Таблица 10 − Анализ и оценка результатов тестирования на способность и устойчивость работы с разнородными входными данными с количественными результатами

Наименование	Нормативное значение	Определенное значение	Результат
Метрика качества на исходных изображениях	от … до …	…	Соответствует/ не соответствует

13 Тестирование на защищенность и тестирование на безопасность

Тестирование на защищенность и тестирование на безопасность выполняют согласно утвержденным нормативным документам путем анализа программного кода и технической и эксплуатационной документации СИИ.

Номенклатура утвержденных нормативных документов:

- ГОСТ Р 51904;
- ГОСТ 19781;
- ГОСТ Р ИСО 53114;
- ГОСТ Р 56939;
- [14].

В рамках данных испытаний в том числе проверяют:

- разграничение пользовательского доступа, организацию пользовательской авторизации;

- записывает ли СИИ персональные данные пользователей (пол, возраст, место работы и должность, клинические данные, семейное положение и др.) для постоянного или временного хранения;

- требуется ли СИИ потенциально вредоносное дополнительное (стороннее) программное обеспечение для корректной работы;

- требуется ли СИИ постоянное подключение к сети интернет для корректной работы.

Приложение А

(обязательное)

Дизайн испытаний

Таблица А.1 − Дизайн испытаний в зависимости от фазы испытаний

Этап	Дизайн испытания	Исследование	Метод	Задачи	Комментарии
0	Доказательство концепции (proof of concept)	Контроль качества данных	Описательный анализ	Убедиться, что качество данных отвечает необходимым стандартам (требованиям) и что состав (диапазон) данных соответствует целевой популяции	Качество и диапазон данных могут изменяться в зависимости от цели исследования (анализируемых параметров)
0		Тестирование алгоритма	Статистический анализ	Оценка точности предсказаний алгоритма ИИ и (или) других метрик	Использование раздельных обучающего и тестирующего наборов данных; В качестве эталона выступает среднестатистический показатель профильных медицинских сотрудников при выполнении данной задачи; Приемлемые значения измеряемых метрик зависят от клинических последствий возможных ошибок.
0, 1		Популяционное исследование	Наблюдения и опросы; Анализ последовательности операций (workflow analysis). .	Понимание причин, влияющих на принятие клинических решений; Определение и детальное изучение процессов, требующих автоматизации (требований пользователя); Определение полезных функциональных возможностей и вариантов их реализации.	Проводят оценку необходимых ресурсов на реализацию и обеспечение функционирования вариантов автоматизированной системы; оценку преимуществ и недостатков каждого варианта; сопоставление требований пользователя и характеристик предлагаемой системы и выбор оптимального варианта; определение порядка оценки качества и условий приемки системы; оценку эффектов, получаемых от автоматизированной системы

Продолжение таблицы А.1

Этап	Дизайн испытания	Исследование	Метод	Задачи	Комментарии
1, 2	Технические испытания (аналитическая валидация)	Функциональное тестирование	Описательный анализ и расчетные методы	Функциональная полнота	В СИИ присутствует и соответствует документации набор функций
				Функциональная корректность	СИИ генерирует выходные данные с надлежащим уровнем точности (accuracy), а также повторяемости (repeatability) и воспроизводимости (reproducibility)
				Функциональная пригодность	Оценка степени функционального упрощения выполнения определенных задач
				Способность к самообучению	Оценка способности СИИ извлекать знания из накопленного опыта и применять их для улучшения качества поставленных задач
				Согласованность	СИИ соответствует стандартам или соглашениям, или положениям законов, или подобных рекомендаций
		Уровень производительности	Регистрационный метод (моделирование типовых задач СИИ)	Характер изменения во времени	Соответствие требованиям временных ресурсов на выполнение операции, а также производительность – количество однотипных задач за определенное время: время, затрачиваемое на каждую операцию в последовательности (workflow) на рекомендуемой конфигурации аппаратного оборудования
				Характер изменения ресурсов	Объем используемых ресурсов и продолжительности такого использования при выполнении функции: потребление ресурсов центрального и (или) графического процессора; возможность многопоточной работы; потребление оперативной памяти;

Продолжение таблицы А.1

Этап	Дизайн испытания	Исследование	Метод	Задачи	Комментарии
1, 2	Технические испытания (аналитическая валидация)			Производительные возможности	Степень соответствия требованиям таких параметров СИИ как количество параллельно обрабатываемых наборов данных, количество параллельно работающих пользователей, емкость канала, пропускная способность по транзакциям
		Надежность	Регистрационный метод (моделирование типовых задач), способность СИИ сохранять свой уровень качества функционирования при установленных условиях за установленный период времени	Стабильность	Вероятность ошибочного срабатывания СИИ при N циклов непрерывной обработки тестового набора медицинских изображений
				Устойчивость к ошибке	Способность СИИ поддерживать определенный уровень качества функционирования в случаях программных ошибок или нарушения определенного интерфейса
				Восстанавливаемость	Способность СИИ возобновить работу после возникновения ошибки
				Робастность	Способность СИИ демонстрировать требуемую точность выходных данных при наличии разного рода выбросов (соответствует Тестирования на устойчивость к состязательным атакам и Тестированию на способность и устойчивость работы с разнородными данными)
		Тестирования на устойчивость к состязательным атакам (adversarial attacks) (относится к метрики надежности)	Сравнительный статистический анализ	Оценка точности предсказаний алгоритма ИИ и (или) других метрик для каждой модальности в отдельности и для всех модальностей вместе в условиях состязательных атак и в нормальных условиях	Составляют ранжированный список состязательных атак, упорядоченный по степени влияния на алгоритм

Продолжение таблицы А.1

Этап	Дизайн испытания	Исследование	Метод	Задачи	Комментарии
1, 2	Технические испытания (аналитическая валидация)	Тестирование на защищенность	Анализ программного кода	Проверка защиты от взлома, несанкционированного доступа и прочих внешних воздействий, а также приватности данных.	Проверяют разграничение пользовательского доступа, организацию пользовательской авторизации; Проверяют, записывает ли СИИ персональные данные пользователей (пол, возраст, место работы и должность, клинические данные, семейное положение и др.) для постоянного или временного хранения; Проверяют, требуется ли СИИ потенциально вредоносное дополнительное (стороннее) программное обеспечение для корректной работы; Проверяют, требуется ли СИИ постоянное подключение к сети интернет для корректной работы
		Тестирование на способность и устойчивость работы с разнородными данными (относится к метрики надежности)	Моделирование типовых задач СИИ с различными наборами данных и диагностическими устройствами	Проверка соответствия технической документации, если требования к качеству входных данных определены	Для проверки способности обработки некорректных данных производится последовательно два индекс-теста на выборке с наличием искаженных данных и на выборке без искаженных данных
				Тестирование на совместимость с диагностическим устройством	Проверка обработки данных с устройств, указанных в эксплуатационной документации производителя; Тестирование на корректную работу с примерами, сложными к интерпретации экспертами; Адаптируемость к новым типам данных при вводе СИИ в эксплуатацию.
3,4*	Клинические испытания (клиническая валидация)	Неконтролируемые испытания (uncontrolled trials)	Анализ медицинских изображений, содержащих целевую патологию (анатомическую структуру)	Количественная оценка способности СИИ распознавать целевую анатомическую структуру	Используют для алгоритмов, предназначенных для высокоточной сегментации целевых анатомических структур; Не рекомендуется для оценки диагностической точности алгоритма, поскольку данный тип исследований ассоциирован с существенной переоценкой этой метрики.

Продолжение таблицы А.1

Этап	Дизайн испытания	Исследование	Метод	Задачи	Комментарии
3,4*	Клинические испытания (кли-ническая вали-дация)	Рандомизированные контролируемые испытания с параллельными группами (parallel group design)	Группы сравнения: Интерпретация медицинских изображений с использованием СИИ; Интерпретация медицинских изображений без использования СИИ	Количественная оценка эффекта СИИ на принятие медицинских решений; Количественная оценка диагностической точности СИИ и (или) других метрик	Участники исследования (медицинские сотрудники) остаются в группе сравнения на протяжении всего исследования; Распределение участников между группами сравнения проводят случайным образом; Необъяснимую вариабельность результатов ассоциируют с разницей между участниками исследования; Групп сравнения может быть больше двух.
		Рандомизированные контролируемые испытания с перекрестными группами (crossover design)	Испытание разбивают на два этапа. Группы сравнения: Интерпретация медицинских изображений с использованием СИИ (первый этап), интерпретация медицинских изображений без использования ИИ (второй этап); Обратный порядок.	Количественная оценка эффекта СИИ на принятие медицинских решений; Количественная оценка диагностической точности СИИ и (или) других метрик	Требуют меньшего объема выборки, чем испытания с параллельными группами; Каждый участник выступает собственным контролем; Меньшая вариабельность результатов по сравнению с испытаниями с параллельными группами; Возможно включение периода паузы между этапами для снижения психологического эффекта использования СИИ; Требуют более длительного времени проведения, чем испытания с параллельными группами.

Окончание таблицы А.1

Этап	Дизайн испытания	Исследование	Метод	Задачи	Комментарии
		Рандомизированные контролируемые факториальные испытания (factorial design)	Группы сравнения: Интерпретация медицинских изображений с использованием СИИ 1; Интерпретация медицинских изображений с использованием СИИ 2; Интерпретация медицинских изображений с использованием СИИ 1 и СИИ 2; Интерпретация медицинских изображений без использования СИИ	Сравнение двух алгоритмов искусственного интеллекта, предназначенных для решения одной и той же задачи	Позволяют оценить комбинированный эффект альтернативных СИИ; Позволяют получить ответ сразу на несколько исследовательских вопросов; Требуют меньшего объема выборки, чем испытания с параллельными группами;
		Пре- и пост-имплементационные сравнения (pre-post comparison)	Статистический анализ	Мультицентровая скорректированная с учетом времени (time-adjusted) оценка эффекта внедрения СИИ на частоту детекции целевой патологии и (или) другие метрики	В качестве контроля используют ретроспективные данные для той же когорты пациентов до внедрения СИИ в медицинской организации, на базе которой проходят клинические испытания; Потенциальные источники систематической ошибки: факторы со временной зависимостью (сезонные заболевания и др.), неочевидные тенденции развития и распространения заболевания.
*Аналогичные типы исследований допускается проводить на Фазах испытаний 1 и 2. В этом случае они носят название пилотных проектов, требуют значительно меньших размеров выборки, и направлены на доказательство потенциальной работоспособности алгоритма в клинической медицине. Успешное завершение пилотного проекта (статистически значимый эффект от внедрения) не является доказательством клинической эффективности алгоритма и служит только обоснованием для проведения полноценных испытаний.					

Приложение Б

(справочное)

Пример заполнения протокола по результатам оценки клинической эффективности

При проведении испытаний алгоритмов анализа медицинских изображений составляют протокол, в котором должны быть отражены следующие данные:

а) Название, в котором определяют дизайн испытания, целевую популяцию, тип алгоритма искусственного интеллекта, предназначенное применение СИИ;

б) Дата испытания и версия программного обеспечения;

в) Детализированное описание предназначенного применения СИИ: тип/типы медицинских изображений, предполагаемая роль СИИ в клинической практике, конкретная решаемая задача;

г) Описание цели испытания, обоснование необходимости его проведения, включая анализ опубликованной релевантной литературы (научные статьи, патенты, грантовые заявки, в том числе зарубежные), указание потенциальной пользы и вреда использования СИИ;

д) Обоснование выбора компаратора (компараторов);

е) Описание дизайна испытаний.

1) Используют разные методологии в зависимости от фазы испытаний (Таблица А.1, приложение А). СИИ в качестве медицинских устройств отличаются от лекарственных препаратов и медицинских приборов тем, что предназначены для воздействия на принятие решений медицинским работником. По этой причине оценка эффектов СИИ в клинической медицине не может проводиться независимо от предполагаемых пользователей (медицинских сотрудников), которые являются такими же участниками исследования, как пациенты и анализируемая технология искусственного интеллекта (ИИ). Таким образом, в каждом испытании должно быть минимум две переменных, определяющих характеристики группы сравнения:

- распределение пациентов между группами сравнения на имеющих целевое заболевание и (или) детектируемый средствами медицинской визуализации признак (категория «случай») и с отсутствием заболевания и (или) детектируемого средствами медицинской визуализации признака (категория «контроль»). В зависимости от фазы испытаний вместо данных специально включенных в испытания пациентов допускается использовать медицинские изображения из специализированных наборов данных, в том числе находящихся в публичном доступе. Соотношение распределения пациентов (исследований) указывают

в виде случай/контроль, например, 1/1, 1/2 и т.д.; конкретное соотношение выбирают в зависимости от цели испытаний и распространенности заболевания (признака), приводя обоснование (пункт г);

- распределение медицинских сотрудников между группами сравнения на проводящих интерпретацию медицинских изображений с использованием СИИ и без использования СИИ. Указывают соотношение численности между этими группами и уровень квалификации медицинских сотрудников (трудовой стаж в годах, если применимо -- пройденное обучение по целевому медицинскому вмешательству, медицинской технологии, медицинскую специализацию), которые выбирают в зависимости от цели испытаний, приводя обоснование (пункт г).

2) Также указывают гипотезу испытаний: превосходство, неуступающая эффективность, эквивалентность.

Гипотеза превосходства предполагает, что эффективность и (или) другая метрика исследуемой СИИ по выбранному критерию выше, чем у компаратора (компараторов).

Гипотеза неуступающей эффективности предполагает, что эффективность и (или) другая метрика исследуемой СИИ по выбранному критерию не хуже, чем у компаратора (компараторов).

Гипотеза эквивалентности предполагает, что эффективность и (или) другая метрика исследуемой СИИ по выбранному критерию является эквивалентной с компаратором (компараторами).

ж) Описание условий проведения испытаний (например, амбулаторное учреждение, стационар). Описание требований к аппаратному и программному обеспечению, необходимости дообучения алгоритма на локальных данных для успешной интеграции СИИ в цифровой контур организации, проводящей испытания. Существуют ограничения обобщаемости алгоритмов ИИ при их использовании вне среды разработки и обучения. В протоколе указывают:

- требует ли внедрение СИИ каких-либо устройств от конкретной фирмы-производителя;

- есть ли потребность в локальном вычислительном аппаратном обеспечении;

- существует ли потребность в обеспечении интеграции облачных решений, при необходимости с уточнением конкретной фирмы-поставщика услуг;

- при необходимости внесения любых изменений в алгоритм в рамках процедуры внедрения, этот процесс должен быть описан.

и) Описание критериев включения и исключения на уровне (а) участников исследования (пациентов), (б) участников исследования (медицинских сотрудников) и (в) входных данных.

1) Критерии включения для участников исследования (пациентов) определяют целевую демографию СИИ: тип и (или) тяжесть заболевания и (или) признака, наличие сочетанных заболеваний, диагностические процедуры и др. показатели.

2) Критерии включения для участников исследования (медицинских сотрудников) определяют предполагаемых пользователей СИИ, задавая уровень квалификации, имеющей отношение к изучаемому вмешательству.

3) Под входными данными понимают данные, необходимые СИИ для выполнения предполагаемой функции. В протоколе указывают, есть ли минимальные требования к входным данным (разрешение изображения, формат данных и др.).

4) Указывают, как будет оцениваться соответствие критериям включения.

Примеры

1 Если пациент соответствует критериям включения на уровне участников исследования, но качество полученного исследования компьютерной томографии по каким-либо причинам было неудовлетворительным для использования СИИ, это необходимо расценивать как критерий исключения на уровне входных данных.

2 «Участники должны иметь компетенции в интерпретации исследований компьютерной томографии грудной клетки согласно системе LUNG-Rads, версия 1.1. Проверку компетенции проводят тестированием на предварительно размеченном наборе данных компьютерной томографии, содержащем примеры рака легкого в разных стадиях развития, а также исследования без патологий».

к) Описание вмешательства для каждой группы сравнения в деталях, необходимых для воспроизведения результатов исследования.

1) Указывают, какая версия алгоритма ИИ будет использована. СИИ, как правило, подвергаются неоднократному изменению и обновлению программного кода течение своего жизненного цикла. В протоколе указывают, какую версию используют в испытаниях, и является ли она той же версией, которую использовали в предыдущих исследованиях, на основании которых обосновывали необходимость проведения испытаний. Когда это применимо, указывают, какие изменения были внесены в текущую версию и обоснования внесения этих изменений.

2) Указывают процедуру получения, отбора и предварительной обработки входных данных для СИИ.

3) Указывают процедуру оценки качества данных и действия, предпринимаемые с низкокачественными (не соответствующими минимальным требованиям к качеству) или недостающими данными. Низкокачественные данные могут осложнить интерпретацию также медицинским сотрудникам, не использующим СИИ. По этой причине необходимо, в случае применимости, дополнительно указывать такую информацию для контрольного вмешательства.

4) Указывают, существует ли необходимость участия медицинского сотрудника в обработке (подготовке) входных данных СИИ, и требуемую квалификацию медицинского сотрудника, включая обучение и инструктаж по работе с СИИ.

Пример – Медицинский сотрудник должен отметить исследуемую область на медицинском изображении, которую затем будет анализировать алгоритм ИИ.

5) Описание выводимых данных СИИ-вмешательства. СИИ может предоставлять данные о диагностической классификации, вероятности существования либо развития патологии, рекомендуемые мероприятия, либо другую информацию. Тип выводимых данных имеет прямую связь с эксплуатационными свойствами СИИ, а также медицинскими решениями, на которые она может повлиять.

л) Описывают ожидаемые первичные, вторичные и другие исходы СИИ-вмешательства, включая конкретный анализируемый признак, метрику анализа, метод объединения данных.

Пример – Анализируемый признак: легочный узел. Метрика анализа: линейные размеры. Метод объединения данных: среднее значение со стандартным отклонением.

м) Диаграмма дизайна исследования, на которой показаны этапы включения участников в исследование, вмешательства (в том числе пауза между этапами в случае использования перекрестных групп), и анализа данных (см. рисунок 1).

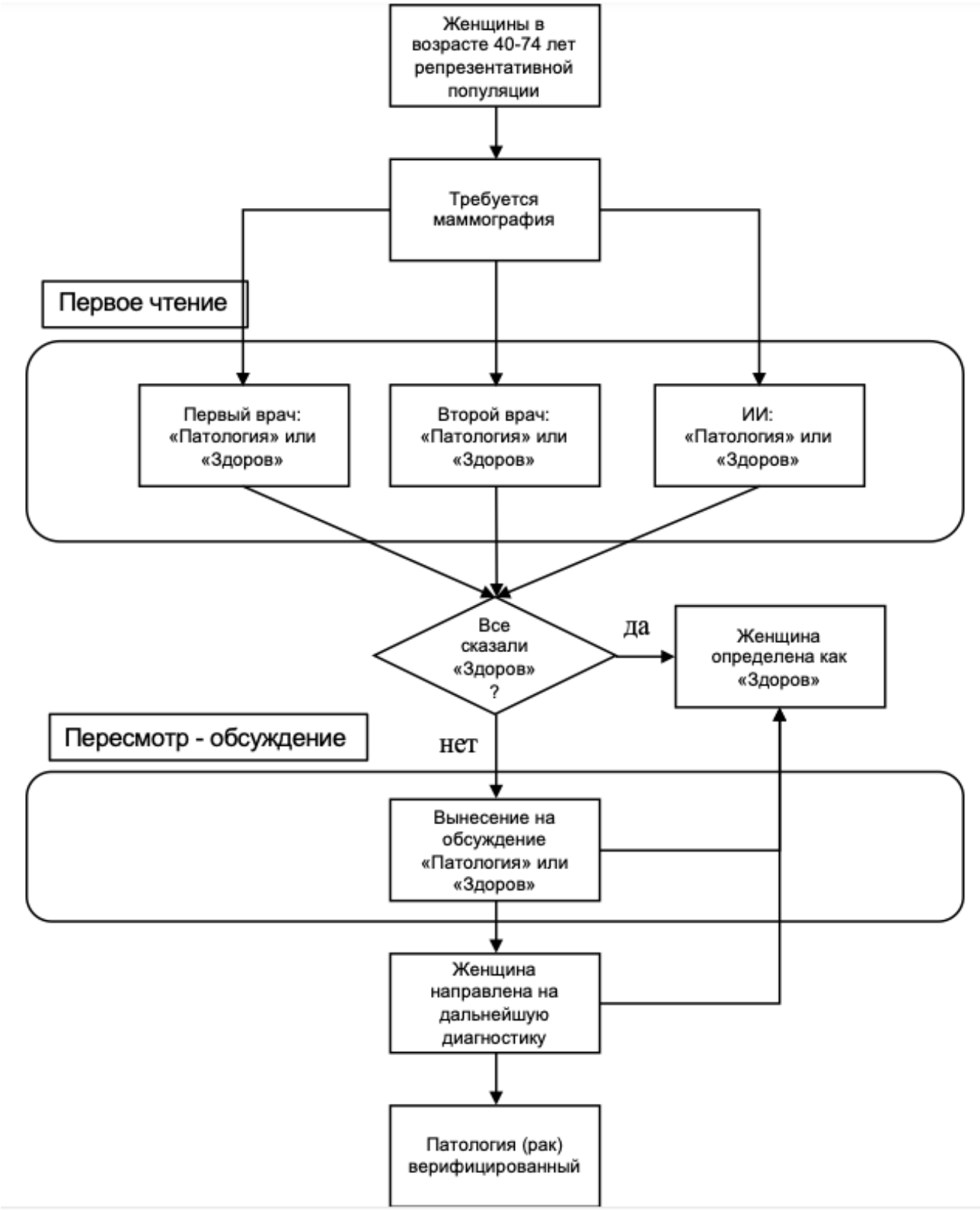

Рисунок 1 − Пример диаграммы дизайна исследования (клинические испытания WHO ID STGKS001)

н) Оценка объема выборки пациентов (размера набора данных) согласно проверяемой гипотезе исследования. Расчет объема выборки проводят по первичному исходу, исходя из предположения, что недостающих данных не будет. Явным образом указывают клинические и статистические предположения, на которые опирались при вычислениях: (если применимо) доля популяции с целевым заболеванием (признаком) либо среднее значение со стандартным отклонением, статистический критерий, величина ошибки первого и второго рода.

п) Оценка числа медицинских сотрудников, принимающих участие в исследовании, с описанием ролей и указанием квалификации и специализации. Описывают предусмотренные меры в случае отказа или невозможности дальнейшего участия медицинского сотрудника в исследовании.

р) Описание подхода к распределению анализируемых данных между медицинскими сотрудниками в группах сравнения. Как правило, необходимые размеры выборок являются слишком большими, чтобы их мог полностью обработать один медицинский сотрудник. Допускается частичное распределение набора данных между участниками исследования, с соблюдением обязательных условий:

- каждое исследование должно быть просмотрено хотя бы дважды разными участниками из группы, проводящей интерпретацию с использованием СИИ, и дважды разными участниками из группы, не использующей СИИ при интерпретации медицинских изображений. Это позволит оценить согласие между экспертами и влияние СИИ на этот показатель;
- распределение исследований между участниками (медицинскими сотрудниками) должно проводиться случайным образом.

Пример – Сгенерированное компьютером случайное распределение идентификационных номеров исследований в наборе данных.

Если дизайн испытаний подразумевает группировку данных (по медицинской организации, полу и (или) возрасту пациентов, стадии заболевания и др.), указывают список факторов, по которым проводили группировку (стратификацию) данных, с обоснованием этих факторов.

с) Если предполагается использование референсного стандарта, указывают, проводится ли ослепление участников испытания к этим результатам с конкретным указанием, для кого и как проводят ослепление: для пациентов, СИИ, медицинских сотрудников, специалистов по обработке данных. Возможно также проводить ослепление участников испытания к проверяемой гипотезе.

Пример – При проверке гипотезы эквивалентности допускается информировать участников, что проводится проверка гипотезы превосходства одного из вмешательств.

т) Методика сбора данных. Указывают, каким образом будет проводиться оценка целевых признаков на медицинских изображениях для первичного, вторичного и других исходов исследования, с указанием специализированных программных инструментов в случае их использования.

Пример – Медицинские сотрудники ищут на КТ-снимках легочные очаги размерами от 4 до 30 мм, сохраняя такую информацию о находках, как локализация легочного очага (положение центра находки по двум измерениям на изображении и номеру среза); диаметр находки; тип легочного очага (солидный, полусолидный или очаг по типу матового стекла) с помощью программного обеспечения FAnTom. Рекомендовано не отмечать кальцинированные и перифиссуральные очаги в легких, а также не отмечать более пяти крупнейших легочных очагов на одном КТ-снимке.

у) Управление данными. Описывают планы по хранению данных исследования, обеспечению их безопасности и защищенности, а также любые мероприятия, направленные на обеспечение качества данных.

Пример – Данные будут храниться в цифровом контуре медицинской организации, на базе которой будут проходить испытания. При необходимости для промежуточного или итогового анализа данные будут извлечены для исследовательских целей под ответственность главного исследователя. Перед проведением любого анализа, в том числе статистического, будет проведена анонимизация (псевдонимизация) данных. Каждый параметр данных будет проходить проверку на достоверность по типу и диапазону.

ф) Статистические методы. В протоколе должны быть указаны и описаны планируемые методы статистического анализа в случае, если их возможно предусмотреть [15]. Должны быть отражены все предполагаемые методы анализа при сравнении групп исследования. Результаты испытаний могут подвергаться существенному влиянию со стороны методов статистического анализа. Если для конкретного исхода, особенно первичного, предполагают использовать более одной стратегии анализа, возникает потенциальная

возможность недопустимого выборочного представления наиболее ярких и интересных результатов. В протоколе указывают основной метод статистического анализа первичного исхода.

Указывают планы по дополнительному статистическому анализу, предназначенному для подгрупп пациентов. Анализ подгрупп позволяет определить статистически значимые различия испытываемой технологии ИИ для разных категорий пациентов с целевой патологией и (или) признаком, в конечном итоге обеспечивая персонализированный подход в медицине. Однако, некорректно проведенный анализ подгрупп, а также выборочное представление результатов, особенно в случае, если обработку данных проводили посредством апостериорного анализа, сопряжены с риском сомнительных (ложных) выводов.

х) Мероприятия, направленные на согласование исследования в этическом комитете. В протоколе необходимо указать, было ли получено согласование этического комитета с указанием даты и номера согласования и названия этического комитета, либо наметить планы по получению согласования.

В случае, если необходимо получение информированного добровольного согласия на сбор и использование данных от пациентов, соответствующую форму прикладывают к протоколу.

ц) Декларация интересов. Указывают (при наличии) конфликт финансовых и иных интересов для коллектива, проводящего испытания, и для всех медицинских организаций, в рамках которых оно будет проходить.

ч) Протокол заверяют в установленном порядке в организации коллектива, проводящего испытания, с указанием даты заверения. Любые изменения в протоколе испытаний отражают в поправках к протоколу, которые также должны быть согласованы с этическим комитетом.

Испытания алгоритмов анализа медицинских изображений проводят в соответствии с протоколом испытаний.

Приложение В

(справочное)

Пример заполнения протокола по результатам функционального тестирования системы искусственного интеллекта

Наименование параметра	Данные, указанные в технической документации	Результат в соответствии с порядковым номером исследования			Комментарий
		1	2	…	
Возможность приоритизации (триаж)	Имеется	Да	Да		
Наличие графического обозначения	Имеется	Да	Да		
Наличие дополнительной серии от СИИ	Имеется	Да	Да		
Название дополнительной серии	Имеется	Да	Да		
Возможность синхронизации серий	Не применимо	–	–		
Отображение информации о СИИ	Имеется	Да	Да		
Отображение вероятности(-ей) находки(-ок)	Отсутствует	–	–		
Указание категории находок	Не применимо	–	–		
Возможность отключения маркировки	Отсутствует	–	–		
Наличие DICOM SR 1)	Имеется	Да	Да		
Создание шаблона протокола	Имеется				
Возможность сравнения исследований в динамике	Не применимо	–	–		
Прочее (указать)	–	N	#		

Библиография

[1] ISO/IEC TR 29119-11:2020(E) — Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems.

[2] ISO 9241-161:2016 Ergonomics of human-system interaction -- Part 161: Guidance on visual user-interface elements

[3] ISO 9241-171:2008 Ergonomics of human-system interaction -- Part 171: Guidance on software accessibility

[4] Liu, X., Cruz Rivera, S., Moher, D. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 26, 1364–1374 (2020).

[5] Sounderajah V, Ashrafian H, Golub RM On behalf of the STARD-AI Steering Committee, et alDeveloping a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocolBMJ Open 2021;11:e047709.

[6] Collins GS, Dhiman P, Andaur Navarro CL, et alProtocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligenceBMJ Open 2021;11:e048008.

[7] IMDRF/SaMD WG/N41 — Software as a Medical Device (SaMD): Clinical Evaluation, 2017.

[8] Fenster A, Chiu B. Evaluation of segmentation algorithms for medical imaging. In: Conf Proc IEEE Eng Med Biol Soc. Shanghai; 2005. p. 7186–189.

[9] Taha A.A., Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging (2015) 15:29

[10] Hirano, H., Minagi, A. & Takemoto, K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med Imaging* 21, 9 (2021).

[11] Apostolidis, K.D.; Papakostas, G.A. A Survey on Adversarial Deep Learning Robustness in Medical Image Analysis. Electronics 2021, 10, 2132.

[12] Клинические испытания программного обеспечения на основе интеллектуальных технологий (лучевая диагностика) / С.П. Морозов [и др.] / Серия «Лучшие практики лучевой и инструментальной диагностики». — М., 2019. — Вып. 57. — 51 с.

[13] Defending Deep Learning-Based Biomedical Image Segmentation from Adversarial Attacks: A Low-Cost Frequency Refinement Approach| SpringerLink. Available online: https://link.springer.com/chapter/10.1007/978-3-030-59719-1_34 (accessed on 4 June 2021).]

[14] Методические рекомендации по порядку проведения экспертизы качества, эффективности и безопасности медицинских изделий (в части программного обеспечения) для государственной регистрации в рамках национальной системы» (утв. Росздравнадзором 24.08.2018 г.).

[15] Методические рекомендации по оценке качества статистического анализа в клинических исследованиях.

УДК 615.841:006.354 ОКС 11.040.01

Ключевые слова: система искусственного интеллекта, искусственный интеллект, клиническая медицина, алгоритмы обработки медицинских изображений, методы испытаний

Руководитель организации-разработчика
ГБУЗ «НПКЦ ДиТ ДЗМ»

Руководитель разработки — Директор _____ С.П. Морозов