



ПРЕДВАРИТЕЛЬНЫЙ
НАЦИОНАЛЬНЫЙ
СТАНДАРТ
РОССИЙСКОЙ
ФЕДЕРАЦИИ

ПНСТ
872—
2023

СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ С ПРИМЕНЕНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Методы клинических испытаний

Издание официальное

Москва
Российский институт стандартизации
2023

Предисловие

1 РАЗРАБОТАН 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 20 ноября 2023 г. № 64-пнст

Правила применения настоящего стандарта и проведения его мониторинга установлены в ГОСТ Р 1.16—2011 (разделы 5 и 6).

Федеральное агентство по техническому регулированию и метрологии собирает сведения о практическом применении настоящего стандарта. Данные сведения, а также замечания и предложения по содержанию стандарта можно направить не позднее чем за 4 мес до истечения срока его действия разработчику настоящего стандарта по адресу: 121205 Москва, Инновационный центр Сколково, улица Нобеля, д. 1, e-mail: info@tc194.ru и/или в Федеральное агентство по техническому регулированию и метрологии: 123112 Москва, Пресненская набережная, д. 10, стр. 2.

В случае отмены настоящего стандарта соответствующая информация будет опубликована в ежемесячном информационном указателе «Национальные стандарты» и также будет размещена на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

© Оформление. ФГБУ «Институт стандартизации», 2023

Настоящий стандарт не может быть полностью или частично воспроизведен, тиражирован и распространен в качестве официального издания без разрешения Федерального агентства по техническому регулированию и метрологии

II

Содержание

1 Область применения1

2 Нормативные ссылки.....1

3 Термины и определения1

4 Сокращения3

5 Методология клинических испытаний систем поддержки принятия врачебных решений.....3

Библиография10

Введение

Системы поддержки принятия врачебных решений (СППВР) являются существенным сегментом цифрового здравоохранения (см. [1], [2]). Внедрение СППВР совершенствует и ускоряет сбор и анализ данных, сокращает количество ошибок в медицинской документации, повышает качество медицинских данных, сокращает когнитивную нагрузку на медицинский персонал и, как следствие, способствует повышению качества принятия клинических решений и эффективности лечебно-диагностических процессов (см. [3]).

Вместе с тем, во время функционирования СППВР осуществляет интерпретацию имеющихся данных, что создает потенциальный риск причинения вреда здоровью пациентов, если в работе СППВР будут возникать ошибки [4]. Более того, применение технологий искусственного интеллекта (ИИ) и машинного обучения при создании и эксплуатации СППВР еще больше увеличивают данный риск в силу вероятностной природы алгоритмов, создаваемых с помощью ИИ [5], [6] (см. также [7]). С точки зрения законодательного регулирования СППВР относятся к программному обеспечению как медицинскому изделию, вследствие чего требуется их государственная регистрация в качестве медицинского изделия и последующий пострегистрационный мониторинг. Причем ключевым этапом в обеспечении контроля безопасности, качества и эффективности СППВР являются клинические испытания [8], [9].

В этой связи стандартизация процессов клинических испытаний СППВР является необходимой мерой по обеспечению контроля безопасности и эффективности СППВР, а также по формированию доверия к ним со стороны работников практического здравоохранения, инвесторов и регулирующих органов.

Настоящий стандарт входит в комплекс стандартов «Системы поддержки принятия врачебных решений» и устанавливает требования к методам клинических испытаний СППВР.

ПРЕДВАРИТЕЛЬНЫЙ НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ
С ПРИМЕНЕНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Методы клинических испытаний

Clinical decision support systems using artificial intelligence.
Clinical trials methodsСрок действия — с 2024—01—01
до 2027—01—01

1 Область применения

Настоящий стандарт устанавливает общие требования к разработке программ и методов клинических испытаний систем поддержки принятия врачебных решений (СППВР), включая:

- дизайн испытания;
- используемые методы испытания;
- расчет числа записей в наборе данных, предназначенном для клинического испытания СППВР.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ Р 59921.0 Системы искусственного интеллекта в клинической медицине. Основные положения

ГОСТ Р 59921.1 Системы искусственного интеллекта в клинической медицине. Часть 1. Клиническая оценка

ГОСТ Р 59921.5 Системы искусственного интеллекта в клинической медицине. Часть 5. Требования к структуре и порядку применения набора данных для обучения и тестирования алгоритмов

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 термины по ГОСТ Р 59921.0, а также следующие термины с соответствующими определениями:

3.1 система поддержки принятия клинических решений; СППВР (Clinical decision support system, CDSS): Программное обеспечение, позволяющее путем интерпретации собираемой информации поддерживать принятие врачом решения на всех этапах лечебно-диагностического процесса с целью снижения ошибок и повышения качества оказываемой медицинской помощи.

3.2 алгоритм системы поддержки принятия врачебных решений; алгоритм СППВР (algorithm in clinical decision support system): Последовательность и правила выполнения действий, применение логико-лингвистических моделей, проведение вычислительных операций в системе поддержки принятия врачебных решений.

3.3

программное обеспечение как медицинское изделие; ПОкМИ (software as medical device): Программное обеспечение, не являющееся составной частью медицинского изделия и предназначенное его изготовителем для применения с одной медицинской целью или более.
[ГОСТ Р 59765—2021, статья 2.1]

3.4

машинное обучение (machine learning): Процесс автоматического обучения и совершенствования поведения системы искусственного интеллекта на основе обработки массива обучающих данных без явного программирования.
[ГОСТ Р 59895—2021, статья 2.1.7]

3.5

данные (data): Предоставление информации в формальном виде, пригодном для передачи, интерпретации или обработки людьми или компьютерами.
[ГОСТ 33707—2016, статья 4.259]

3.6

программное обеспечение (программа, программное средство) (software): Упорядоченная последовательность инструкций (кодов) для вычислительного средства, находящаяся в памяти этого средства и представляющая собой описание алгоритма управления вычислительными средствами и действий с данными.
[ГОСТ Р 53622—2009, пункт 3.8]

3.7

искусственный интеллект (artificial intelligence): Комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение, поиск решений без заранее заданного алгоритма и достижение инсайта) и получать при выполнении конкретных практически значимых задач обработки данных результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека.
[ГОСТ Р 59277—2020, пункт 3.18]

3.8

набор данных (data set): Совокупность данных, прошедших предварительную подготовку (обработку) в соответствии с требованиями законодательства Российской Федерации об информации, информационных технологиях и о защите информации и необходимости для разработки программного обеспечения на основе искусственного интеллекта.
[ГОСТ Р 59921.5—2022, пункт 3.12]

3.9

метрика (в подходах к качеству) (metric): Материальная мера некоторых аспектов характеристик качества.
[ГОСТ Р 54837—2011, пункт 4.4]

3.10

жизненный цикл (life cycle): Развитие системы, продукции, услуги, проекта или другой создаваемой человеком сущности от замысла до списания.
[ГОСТ Р 57193—2016, пункт 4.1.19]

3.11 одномоментное (поперечное) исследование (cross-sectional): Однократное обследование участников или объектов исследования.

Примечание — Приведен в [10].

4 Сокращения

В настоящем стандарте использованы следующие сокращения:

ДИ	— доверительный интервал;
ИИ	— искусственный интеллект;
ОЧ	— относительная частота;
РКИ	— рандомизированное контролируемое испытание;
СППВР	— системы поддержки принятия врачебных решений;
AUC ROC	— площадь под характеристической кривой.

5 Методология клинических испытаний систем поддержки принятия врачебных решений

5.1 Общие требования к клиническим испытаниям СППВР

5.1.1 Основная задача клинических испытаний СППВР состоит в оценке того, будут ли достигнуты заявленные производителем метрики качества СППВР в случае их применения в реальной клинической практике в заявленных производителем условиях применения.

Выделяют несколько видов клинических испытаний СППВР (согласно жизненному циклу):

а) внутреннее тестирование, предназначенное для проверки соответствия продукта техническим условиям производителя (спецификациям). Подтвержденное прохождение внутреннего тестирования свидетельствует о готовности продукта к проведению клинических исследований или к следующему этапу жизненного цикла;

б) клиническая валидация (внешнее тестирование), предназначенная(ое) для независимой проверки достижения точности работы СППВР, в том числе в рамках процедур регистрации СППВР в качестве медицинского изделия. В этом испытании необходимо использовать верифицированный набор данных в медицинской организации, проводящей клинические испытания;

в) валидация по время эксплуатации (мониторинг), предназначенная(ый) для периодической проверки стабильности метрик работы СППВР во время применения в реальной клинической практике. В этом испытании допустимо использовать верифицированный набор данных организации, проводящей мониторинг, либо результаты проверки доступности работы СППВР, например количество положительных ответов на обращения к СППВР, количество сбоев и т. д.

Клинические испытания СППВР должен осуществлять подготовленный персонал, имеющий представление как о методологических основах проведения клинических испытаний, используемых методах и показателях качества работы СППВР, так и общее представление о предназначении и особенностях реализации испытуемой СППВР.

5.1.2 Для внутреннего тестирования достаточно применение тестового набора данных. Испытание необходимо проводить путем использования заранее подготовленного набора данных, содержащего входные признаки и сопоставленные им выходные признаки (эталонные значения). Записи из набора данных подают на вход алгоритма СППВР или всей СППВР в целом, затем их обрабатывают и хранят в применяемом наборе данных. Для каждой использованной строки данных фиксируют полученный результат СППВР. Полученные данные используют для расчета матрицы ошибок и соответствующих метрик качества.

5.1.3 Для внешней валидации алгоритма СППВР, например в рамках подготовки к последующей государственной регистрации в качестве медицинского изделия, в ряде случаев может потребоваться интеграция СППВР с медицинской информационной системой, используемой в работе врачей медицинской организации, участвующей в проведении клинического испытания. Это может быть сопряжено с рядом существенных сложностей. В связи с этим для проведения внешней валидации допускается использование заранее подготовленного верифицированного набора данных, соответствующего требованиям 5.3, при условии соблюдения следующих требований:

а) выборка записей для внешней валидации должна быть сформирована строго из целевой популяции;

- б) выборка должна быть случайной;
- в) должен быть использован надежный референсный тест (для задач диагностики) или должно быть фиксирование исходов (для задач прогнозирования);
- г) соотношение позитивных и негативных случаев должно соответствовать относительной частоте искомого состояния в целевой популяции;
- д) необходим расчет чувствительности, специфичности, а также прогностических ценностей положительного и отрицательного классов;
- е) для всех операционных характеристик должны быть рассчитаны 95 % ДИ;
- ж) должна быть оценена безопасность, прежде всего последствия ошибок гипо- и гипердиагностики;
- и) выборка должна быть сформирована на основании данных от тех медицинских организаций и условий оказания медицинской помощи, для которых предназначено применение данной СППВР.

При проведении внешней валидации целесообразно оценить влияние СППВР на принятие врачами клинических решений.

Испытания системы мониторинга проводят в форме наблюдательного исследования с оценкой получаемых в ходе работы СППВР метрик качества.

5.2 Требования к программе клинических испытаний СППВР

Клинические испытания СППВР проводят в соответствии с утвержденной программой клинических испытаний. Разработку программы клинических испытаний выполняют по ГОСТ Р 59921.1.

В случае внутреннего тестирования программа испытаний должна быть определена производителем СППВР.

В случае клинической валидации в рамках клинических испытаний программу клинических испытаний определяет производитель СППВР совместно с медицинской организацией, участвующей в клиническом испытании и/или проведении валидации.

Программа клинических испытаний должна подразумевать выполнение испытаний полного комплекта алгоритмов, включенного в состав СППВР и участвующего в интерпретации поданных на вход клинических данных, с фиксацией полученных результатов.

Принятие решения о завершенности клинического испытания СППВР должно быть основано на анализе достижения метрик качества всех алгоритмов, включенных в состав испытываемой версии СППВР. Если какой-то из алгоритмов не достиг какой-либо из заявленных метрик качества и безошибочности работы, клиническое испытание всей СППВР признают неудовлетворительным.

Повторные циклы испытаний дублируют для всей системы в целом, а не только для анализа исправленного алгоритма, т. к. архитектура и программный код СППВР могут иметь достаточно сложную природу с внутренними перекрестными зависимостями между алгоритмами, в результате чего изменение отдельного алгоритма может влиять на работу других компонентов СППВР.

Для мониторинга СППВР должен быть подготовлен протокол наблюдательного исследования.

5.3 Требования к наборам данных для проведения клинических испытаний

Подготовку набора данных выполняют по ГОСТ Р 59921.5.

В случае внутреннего тестирования набор данных подготавливает производитель СППВР. Записи набора данных для внутреннего тестирования не должны содержать записи, использованные в машинном обучении алгоритма (обучающий набор данных).

В случае клинической валидации набор данных подготавливает медицинская организация, участвующая в таком виде испытаний, используя при этом данные, которые не были доступны производителю СППВР на этапе создания и внутреннего тестирования.

При подготовке набора данных для проведения испытаний основным является определение числа записей набора данных, необходимого и достаточного для формирования надежных выводов о качестве работы СППВР. При подготовке набора данных должна быть обеспечена репрезентативность выборки, которая может быть достигнута вероятностным способом ее формирования (случайный отбор, систематический отбор, кластерный отбор и т. д.); данный подход в большинстве случаев сложно достигаем. Таким образом используют следующие вероятностные способы формирования набора данных: выборка удобства (произвольная); последовательная (сплошная) выборка; выборка добровольцев; квотный отбор и т. д. Необходимо обращать внимание на относительную частоту состояния в целевой популяции (ОЧ, долю, частоту) диагностируемого (или прогнозируемого) состояния в используемом

для проведения испытаний наборе данных, т. к. именно от него зависят наиболее значимые метрики качества СППВР.

Размер выборки для внутреннего тестирования определяет производитель. Размер выборки для клинической валидации и валидации во время эксплуатации определяет организация, проводящая данный вид испытаний.

При создании набора данных для испытаний следует использовать выборку из той же популяции пациентов и условий оказания медицинской помощи, в которых предполагают применять разработанную СППВР. Например, если СППВР предназначена для применения в условиях стационара, то и набор данных для испытаний СППВР должен быть создан на основе данных пациентов, получавших медицинскую помощь в таких же условиях. Если СППВР предназначена для применения в разных уровнях медицинских организаций (центральная районная больница, городская поликлиника, региональный клинический центр и т. д.), то в наборе данных для испытаний должны содержаться записи, собранные из всех заявленных производителем СППВР уровней.

5.4 Метрики качества работы алгоритма СППВР, оцениваемые при проведении клинических испытаний

При проведении клинических испытаний алгоритмов СППВР для каждого из них должны быть оценены следующие метрики качества:

- а) чувствительность — Чв (sensitivity, Se, recall);
- б) специфичность — Сп (specificity, Sp);
- в) прогностическая ценность положительного результата — ПЦПР (positive predictive value, PPV, precision);
- г) прогностическая ценность отрицательного результата — ПЦОР (negative predictive value, NPV);
- д) общая точность алгоритма (accuracy);
- е) AUC ROC.

Для метрик следует рассчитывать не только точечные оценки, но и интервальные — доверительные интервалы с доверительной вероятностью от 95 % до 99 %, т. е. те диапазоны, в которых фактически с той или иной вероятностью могут варьировать соответствующие метрики. Причем Чв и Сп, так же, как и ПЦПР и ПЦОР, изменяются реципрокно, и, оптимизируя один показатель, производитель СППВР, как правило, ухудшает другой.

Если в исследование отдельно набирались позитивные и негативные случаи, необходима поправка на ОЧ состояния в целевой популяции:

$$\text{ПЦПР} = \text{Чв} \cdot \text{ОЧ} / (\text{Чв} \cdot \text{ОЧ} + (1 - \text{Сп}) \cdot (1 - \text{ОЧ}));$$

$$\text{ПЦОР} = \text{Сп} \cdot (1 - \text{ОЧ}) / (\text{Сп} \cdot (1 - \text{ОЧ}) + (1 - \text{Чв}) \cdot \text{ОЧ}).$$

Прогностические ценности необходимо оценивать, т. к. именно с ними работает врач, оценивая результат диагностики или прогноза конкретного больного с учетом вероятности гипер- и гиподиагностики.

Расчет общей точности алгоритма СППВР основан на отношении суммы истинно-положительных и истинно-отрицательных результатов к общему числу наблюдений в выборке. Для точности также могут быть рассчитаны ДИ (99 %, 95 %). Точность зависима от относительной частоты состояния в целевой популяции, соответственно не может быть рассчитана для нерепрезентативной выборки без поправки на относительную частоту состояния в целевой популяции. Этот показатель является наиболее общим, недостаточно понятным врачам, таким образом принятие решения о качестве реализации алгоритма СППВР только на основе анализа общей точности не рекомендуется.

Площадь под характеристической кривой не является бинарным показателем. ROC-анализ может быть проведен как в координатах ((1 – Сп); Чв), так и в координатах ((1 – ПЦОР); ПЦПР). Последний анализ предпочтительнее, т. к. ориентирован на врача — лицо, принимающее решение в отношении конкретного пациента. ROC-анализ используют для предварительного сравнения точности изучаемых алгоритмов машинного обучения, особенно если их много. Однако наличия AUC ROC недостаточно для доказательства эффективности работы СППВР, в связи с чем должен быть определен порог активации бинарного классификатора (cut-off threshold) для классификации случаев, критериями оптимальности которой могут быть:

- минимум ошибки I рода (гипердиагностики) при приемлемой ошибке II рода (гиподиагностике);
- минимум ошибки II рода (гиподиагностики) при приемлемой ошибке I рода (гипердиагностике);
- баланс Чв и Сп;
- максимизация суммы Чв и Сп (критерий Юдена) и т. д.

После определения порога активации должен последовать расчет Чв, Сп, ПЦПР и ПЦОР. В задачах диагностики оптимизируют Чв и/или ПЦПР; в задачах скрининга — Сп и/или ПЦОР при приемлемых значениях Чв и ПЦПР.

Примечание — Возможно так называемое одностороннее использование алгоритма (бинарного классификатора), например использование только на подтверждение искомого состояния (т. е. для диагностики), если высока ПЦПР, а ПЦОР низкая, и при этом цена ошибок II рода (гиподиагностики) невелика. И наоборот — можно использовать для скрининга алгоритм, у которого высокая ПЦОР и низкая ПЦПР, если цена ошибок гипердиагностики невелика.

5.5 Методика испытаний диагностических алгоритмов СППВР

5.5.1 Испытание диагностического алгоритма СППВР осуществляется в дизайне одномоментного испытания. Основные черты испытаний: диагностические тесты (новый и референсный) могут быть применены к каждому пациенту одновременно (с минимальным интервалом времени), а полученные результаты тестов сравнивают. В схеме дизайна одномоментных исследований должны быть учтены рекомендации STARD [6].

Виды дизайнов испытаний диагностических алгоритмов:

- а) несравнительное исследование — сопоставление с референсным методом/тестом;
- б) сравнительное исследование, при проведении которого проверяют гипотезу превосходства точности/безопасности над существующим методом решения задачи при сопоставлении с референсным методом;
- в) сравнительное исследование, в ходе которого проверяют гипотезу не меньшей точности/безопасности метода по отношению к существующему методу решения задачи при сопоставлении с референсным методом.

5.5.2 Выбор дизайна испытаний диагностического алгоритма, включая выбор метода расчета необходимого набора данных, зависит от ответа на следующие вопросы:

- а) является ли исследование сравнительным?
- б) если исследование сравнительное, какую гипотезу проверяют?

5.5.3 Несравнительное исследование диагностического алгоритма

В данном дизайне референсным методом должен быть наиболее точный из имеющихся на текущий момент методов диагностики патологии; допускаемо, что референсный метод обеспечивает 100 %-ную точность диагностики по всем операционным показателям. Таким методом в медицине обычно считают гистологическое исследование, однако это исследование инвазивно и в большинстве случаев не может быть использовано. Обоснование выбора референсного метода следует приводить в отчетах и публикациях.

Показатели точности диагностического алгоритма (Чв, Сп, ПЦПР, ПЦОР) — пропорции (доли), для которых необходимо оценивать ДИ; как правило, используют доверительную вероятность 95 %. Нижняя граница ДИ должна быть поставлена как ориентир при расчете объема выборки, которая не должна быть ниже 85 %.

Расчет необходимого объема выборки для проведения испытания алгоритма СППВР в этом случае сводится к решению обратной задачи — расчету ДИ (обычно 95 %-ного) для ожидаемой величины показателя. Целевое значение показателя должно быть задано на основании клинической значимости, т. е. врачами, а не статистиками. Врачами должна быть задана минимально приемлемая величина показателя точности диагностического алгоритма при гипотетической 100 %-ной точности референсного метода. Должна быть задана и приемлемая величина альтернативного показателя (ПЦОР — альтернативный показатель для ПЦПР, ПЦПР — альтернативный показатель для ПЦОР). Полученные для двух альтернативных показателей объемы выборок следует суммировать.

Более высокие требования должны быть предъявлены к ПЦПР (с учетом ОЧ состояния в целевой популяции), если решают задачу диагностики, т. е. выявления состояния с высокой ОЧ состояния в целевой популяции. Если задача скрининга — выявление состояния с низкой ОЧ состояния в целевой популяции, следует, прежде всего, ориентироваться на ПЦОР. Чв и Сп менее значимы с практической точки зрения применения СППВР; при этом Чв ассоциирована с ПЦПР, а Сп — с ПЦОР.

Пример — *Врачами заданы приемлемые ПЦПР 90 % и ПЦОР 80 %. Это означает, что нижняя граница ДИ для ПЦПР должна быть не менее 90 %, для ПЦОР — не менее 80 %. Тогда приблизительно можно считать, что точечная оценка ПЦПР располагается в интервале между 90 % и 100 %, а именно 95 %. Для малых выборок такое допущение не обосновано. При условии, что будущая выборка будет репрезентативной, как минимум, с точки зрения ОЧ состояния в целевой популяции искомого состояния, не-*

обходимый объем выборки будет составлять 150 пациентов, т. к. 95 % ДИ для ПЦПР, рассчитанный по точному (exact) методу Клоппера—Пирсона, в этом случае равен (90,6 %; 98,1 %). Аналогично для ПЦОР: середину интервала между 80 % и 100 %, т. е. 90 %, можно принять за точечную оценку показателя. Объем выборки при расчете по ПЦОР составит 63 [для получения доли 90 % с 95 %-ным ДИ (80,5 %; 95,9 %)]. После суммирования $150 + 63$ получают 213 как итоговое значение. Если ОЧ состояния в целевой популяции 60 % (0,6), то в группу случаев должно быть включено $213 \cdot 0,6 = 128$ пациентов, в группу сравнения $213 - 128 = 85$ пациентов. Если ОЧ состояния в целевой популяции 10 % = 0,1, то распределение будет следующим: $213 \cdot 0,1 = 21$ пациент в группе случаев, $213 - 21 = 192$ пациента в группе контроля.

Примечание — Если алгоритм СППВР получен методами машинного обучения на так называемой сбалансированной обучающей выборке, т. е. эта выборка была нерепрезентативной в отношении ОЧ состояния в целевой популяции, то полученные при внутреннем тестировании оценки ПЦПР и ПЦОР смещены — тем больше, чем значительнее фактическая относительная частота состояния в целевой популяции будет отклонена от соотношения объемов групп в обучающей выборке. Вследствие этого получить такие же значения на правильно сформированной выборке в испытаниях СППВР будет сложно, если вообще возможно. При этом Чв и Сп не зависят от ОЧ состояния в целевой популяции, и поэтому их легче воспроизвести, однако Чв и Сп практического значения для врачей не имеют.

5.5.4 Сравнительное исследование, гипотеза превосходства диагностического алгоритма над рутинно используемым методом решения задачи при сопоставлении с референсным методом

В данном дизайне проверяют гипотезу превосходства диагностического алгоритма СППВР над рутинно используемым методом решения задачи при сопоставлении с референсным методом. Диагностика осуществляется тремя методами — референсным, новым и рутинно используемым (предлагаемым к замене) методом.

Расчет объема набора данных для проведения испытаний должен быть основан на клинически значимой величине превосходства нового метода над рутинно используемым. Основными параметрами расчета являются:

- ошибка I рода (альфа) — устанавливается не более 5 %;
- статистическая мощность — рекомендуется 90 %, минимально — 80 %;
- значение выбранного показателя оценки (например, ПЦПР) для рутинно используемого метода;
- значение выбранного показателя оценки для нового метода.

Расчет объема выборки возможен в различных статистических пакетах.

Пример — Разработан новый метод диагностики, который превышает рутинно используемый по точности на 10 %. Точность рутинно используемого метода (в интерфейсе «control group») — 80 %, нового — 90 %. Тогда необходимый объем выборки (при ошибке I рода 5 % и статистической мощности 90 %) — 263 пациента.

5.5.5 Сравнительное исследование, гипотеза неменьшей точности/безопасности диагностического алгоритма по отношению к существующему методу решения задачи при сопоставлении с референсным методом

В данном дизайне должна быть проверена гипотеза относительно того, что точность алгоритма СППВР не меньше точности рутинно используемого метода. При использовании данного дизайна необходимо учитывать, что для любых медицинских технологий значимы не только эффективность (в случае диагностики или скрининга — точность), но и безопасность. Повышение безопасности может быть доказано в технических и клинических испытаниях при проверке гипотезы превосходства в отношении критерия (или нескольких критериев) безопасности. Если рутинно используемый метод инвазивный или при его использовании необходимо облучение пациентов, то диагностика также должна быть осуществлена тремя методами — референсным, новым и рутинно используемым (предлагаемым к замене) методом.

Расчет объема набора данных должен быть основан на приемлемой величине порога неменьшей точности/безопасности нового метода по сравнению с рутинно используемым. Основными параметрами расчета являются:

- ошибка I рода (альфа) — устанавливается не более 5 %;
- статистическая мощность — рекомендуется 90 %, минимально — 80 %;
- значение выбранного показателя оценки (например, ПЦПР) для рутинно используемого метода;
- значение выбранного показателя оценки для нового метода;
- порог неменьшей точности/безопасности.

Последний параметр указывает разность между величинами оцениваемого показателя, которую врачи могут считать допустимой. Например, если новый метод должен быть строго таким же, как и рутинно используемый: точность рутинно используемого и нового метода — 80 %, порог равен нулю.

Расчет объема выборки возможен в различных статистических пакетах.

Пример — Точность нового и рутинно используемого метода установлена на уровне 80 %, порог 5 % при ошибке I рода 5 % и статистической мощности 90 %. В этом случае необходимо участие 1097 пациентов для доказательства данной гипотезы. Если порог установить на 7 %, требуемое количество пациентов в выборке почти вдвое меньше — 560.

Объем набора данных при проверке гипотезы неменьшей эффективности неизменно значительно больше, чем при проверке гипотезы превосходства.

5.6 Методика испытаний прогностических алгоритмов СППВР

Разработка правильной программы испытаний прогностического алгоритма СППВР сложнее, чем испытание диагностического алгоритма.

Возможно применение трех видов дизайна:

а) ретроспективные исследования:

- 1) исследование «случай—контроль»,
- 2) когортное исследование;

б) проспективное рандомизированное контролируемое испытание.

5.6.1 Ретроспективные исследования прогностических алгоритмов

Ретроспективные дизайны испытаний прогностических алгоритмов СППВР позволяют оценить только достижение алгоритмом заданных метрик, но не позволяют оценить эффективность и обоснованность внедрения СППВР в реальную клиническую практику.

В ретроспективном исследовании «случай—контроль» должен быть сформирован набор данных, содержащий положительный и отрицательный классы по наличию/отсутствию прогнозируемого исхода (события) и включающий данные пациентов за период, соответствующий сроку прогноза и фактические результаты прогнозируемого события. При обработке набора данных с помощью алгоритмов СППВР осуществляет сбор выданных алгоритмом прогнозов, на основании которых затем оценивают метрики полученной точности (Чв, Сп, ПЦПР, ПЦОР).

В ретроспективном когортном исследовании должны быть учтены следующие требования:

- а) целевая популяция и критерий синхронизации когорты;
- б) по данным в точке синхронизации должен быть сделан прогноз, позитивные случаи относят в основную группу, негативные — в группу контроля;
- в) частоты исхода в группах, возникшие за срок прогнозирования, и прогнозы. При этом учитывают выпадающие наблюдения и примененные медицинские вмешательства.

Метриками качества прогноза определены Чв, Сп, ПЦПР и ПЦОР. Расчет размера набора данных — несравнительное исследование диагностического алгоритма.

5.6.2 Рандомизированное контролируемое испытание прогностического алгоритма

В случае оценки эффективности и безопасности применения алгоритма в практическом здравоохранении следует применять проспективный дизайн РКИ. Для этого надо определить и оценить клинический сценарий применения прогностических алгоритмов СППВР, которые предназначены для выбора/изменения тактики ведения пациентов по сравнению с рутинно применяемой в настоящее время, в том числе для определения необходимости вторичной/третичной профилактики. Испытание прогностического алгоритма должно заключать в себя проверку комбинированной медицинской технологии «прогноз + прогноз-зависимое ведение пациента». При разработке программы испытаний алгоритмов прогнозирования следует сначала убедиться в том, что существуют эффективные способы предупреждения прогнозируемых неблагоприятных событий для пациента.

Существенным параметром прогностических алгоритмов является предельный срок прогнозирования, который зависит от содержания конкретной задачи. Существуют различия сроков прогнозирования — от нескольких дней (например, прогноз улучшения состояния пациента, находящегося в отделении интенсивной терапии) до 10 лет (например, оценка сердечно-сосудистого риска по шкале SCORE).

5.6.2.1 Основными требованиями к РКИ являются следующие:

- а) определение целевой популяции, синхронизированной по какому-либо событию (впервые установленный диагноз, достижение определенного возраста, выполнение хирургического вмешательства и т. д.);

б) рандомизирование объектов целевой популяции после подписания информированного согласия в основную и контрольную группы;

в) прогнозирование в основной группе для всех пациентов; в случае получения неблагоприятного прогноза к пациенту применяют модифицированную по сравнению с рутинной тактику ведения (например, более частые визиты к врачу для раннего обнаружения рецидивов заболевания после хирургической операции); при благоприятном прогнозе применяют рутинную тактику ведения пациента;

г) неприменение прогнозирования в контрольной группе, тактика ведения — рутинная;

д) установление периода наблюдений, в течение которого возникающие целевые события в каждой из групп должны быть зафиксированы. Длительность наблюдения должна быть такой, чтобы возникло достаточное число прогнозируемых событий в каждой из групп.

5.6.2.2 Метриками эффективности прогностических алгоритмов являются:

а) относительный риск;

б) снижение абсолютного риска.

При значительном выбывании из исследования (а это неизбежный спутник длительного наблюдения, необходимого при медленно накапливающимися событиями) требуется оценивать другой показатель — отношение угроз.

При испытании прогностического алгоритма СППВР могут проверять следующие гипотезы:

а) гипотеза превосходства ведения пациента с использованием прогнозирования над рутинным ведением пациента;

б) гипотеза меньшей эффективности ведения пациента с использованием прогнозирования по сравнению с рутинным ведением пациента.

Как правило, с помощью алгоритма прогнозируют неблагоприятное событие с целью снизить его частоту в основной группе по сравнению с контрольной за счет применения некоей медицинской технологии профилактики — вторичной (предотвращения заболевания) или третичной (предотвращения осложнений, рецидивов, обострений, инвалидизирующего течения заболевания).

В этом случае расчет набора данных должен быть проведен так же, как описано выше для гипотезы превосходства, однако при этом необходимы две выборки, каждая из которых будет состоять из вычисленного количества пациентов. Возможно формирование неравных по объему выборок (например, в отношении 3:1), однако статистическая мощность при этом падает, и, следовательно, потребуется больший объем выборки.

Пример — В контрольной группе заболевание возникает у 20 % пациентов. Для основной группы дизайном испытания установлено, что оно должно возникать не более чем у 10 % пациентов, при этом целевое значение выбрано в соответствии с ожидаемой клинической значимостью эффекта. Тогда при ошибке I рода 5 % и статистической мощности 90 %, доле успеха (отсутствия заболевания) в контрольной группе 80 % и доле успеха в основной группе 90 % необходимый объем каждой из групп — 263.

При использовании гипотезы меньшей эффективности в основу необходимо брать предположение о том, что при наличии прогноза можно упростить ведение пациента, не ухудшив при этом исходы (развитие заболевания, осложнения и т. д.).

Пример — Можно приглашать пациента на визиты не один раз в год после операции, а один раз в два года, не ухудшая исходы пациентов.

Расчет набора данных должен быть проведен так же, как описано выше для гипотезы меньшей эффективности, однако здесь необходимы две выборки, каждая из которых должна состоять из вычисленного количества пациентов.

Пример — В контрольной и опытных группах заболевание возникает у 20 % пациентов (т. е. доля «успеха» в обеих группах — 80 %), при этом порог (Δ) меньшей эффективности устанавливают на уровне 5 %. Тогда при ошибке I рода 5 % и статистической мощности 90 % необходимый объем каждой из групп — 1097.

Библиография

- [1] Sutton R.T., Pincock D., Baumgart D.C., Sadowski D.C., Fedorak R.N., Kroeker K.I. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020 Feb 6;3:17, <https://doi.org/10.1038/s41746-020-0221-y>
- [2] Muhiyaddin R., Abd-Alrazaq A.A., Househ M., Alam T., Shah Z. The Impact of Clinical Decision Support Systems (CDSS) on Physicians: A Scoping Review. Stud Health Technol Inform. 2020 Jun 26; 272:470—473, <https://doi.org/10.3233/shti200597>
- [3] Borges do Nascimento I., Abdulazeem H., Vasanthan L., Martinez E., Zucoloto M., Østen-gaard L., Azzopardi-Muscat N., Zapata T., Novillo-Ortiz D. The global effect of digital health technologies on health workers' competencies and health workplace: an umbrella review of systematic reviews and lexical-based and sentence-based meta-analysis. The Lancet Digital Health, Volume 5, Issue 8, 2023, Pages e534-e544, [https://doi.org/10.1016/S2589-7500\(23\)00092-4](https://doi.org/10.1016/S2589-7500(23)00092-4)
- [4] Кобринский Б. А. Системы поддержки принятия врачебных решений: история и современные решения // Методология и технология непрерывного профессионального образования. 2020. — № 4(4). — С. 22—38. — DOI 10.24075/MTCPE.2020.022
- [5] Тыров И.А., Васильев Ю.А., Арзамасов К.М. и др. Оценка зрелости технологий искусственного интеллекта для здравоохранения: методология и ее применение на материалах Московского Эксперимента по компьютерному зрению в лучевой диагностике // Врач и информационные технологии, 2022. № 4 — С. 76—92
- [6] Компьютерное зрение в лучевой диагностике: первый этап московского эксперимента. Владимирский А.В., Васильев Ю.А., Арзамасов К.М. и др. // Общество с ограниченной ответственностью «Издательские решения». — Москва, 2022 — 388 с.
- [7] Moazemi S., Vahdati S., Li J., Kalkhoff S., Castano L.J.V., Dewitz B., Bibo R., Sabouniaghdam P., Tootooni M.S., Bundschuh R.A., Lichtenberg A., Aubin H., Schmid F. Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review. Front Med (Lausanne). 2023 Mar 31;10:1109411, <https://doi.org/10.3389/fmed.2023.1109411>
- [8] Гусев А.В., Морозов С.П., Кутичев В.А., Новицкий Р.Э. Нормативно-правовое регулирование программного обеспечения для здравоохранения, созданного с применением технологий искусственного интеллекта, в Российской Федерации. Медицинские технологии. Оценка и выбор. 2021; (1): 36—45, <https://doi.org/10.17116/medtech20214301136>
- [9] Гусев А.В., Чижииков А.А., Столбов А.П., Кривоносова О.А. Государственная регистрация программного обеспечения как медицинского изделия по правилам Евразийского экономического союза. Медицинские технологии. Оценка и выбор. 2023;(2):32-49. <https://doi.org/10.17116/medtech20234502132>
- [10] Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. — М., МедиаСфера, 2002. 312 с.

УДК 615.841:006.354

ОКС 11.040.01

Ключевые слова: системы поддержки принятия врачебных решений, искусственный интеллект, методы испытаний

Редактор *Л.С. Зимилова*
Технический редактор *И.Е. Черепкова*
Корректор *М.И. Першина*
Компьютерная верстка *И.А. Налейкиной*

Сдано в набор 22.11.2023. Подписано в печать 05.12.2023. Формат 60×84%. Гарнитура Ариал.
Усл. печ. л. 1,86. Уч.-изд. л. 1,58.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru