РАСЧЕТ ОБЪЕМА ВЫБОРКИ ДЛЯ КЛИНИЧЕСКИХ ИСПЫТАНИЙ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ С БИНАРНЫМ ОТКЛИКОМ

DOI: 10.17691/stm2022.14.3.01
УДК 614.2:004.891.3
Поступила 20.02.2022 г.

© **О.Ю. Реброва,** д.м.н., профессор кафедры медицинской кибернетики и информатики медико-биологического факультета[1]; профессор кафедры эндокринологии Института высшего и дополнительного профессионального образования[2]; ведущий научный сотрудник сектора динамических нейросетей отдела нейроинформатики Центра оптико-нейронных технологий[3]; главный научный сотрудник лаборатории доказательной медицины и биостатистики[4];
А.В. Гусев, к.т.н., директор по развитию[5]; старший научный сотрудник отдела научных основ организации здравоохранения[6]; эксперт сектора клинических и технических испытаний[7]

[1]Российский национальный исследовательский медицинский университет им. Н.И. Пирогова, ул. Островитянова, 1, Москва, 117997;
[2]Национальный медицинский исследовательский центр эндокринологии Минздрава России, ул. Дмитрия Ульянова, 11, Москва, 117292;
[3]Федеральный научный центр Научно-исследовательский институт системных исследований РАН, Нахимовский проспект, 36, кор. 1, Москва, 117218;
[4]Научный центр психического здоровья, Каширское шоссе, 34, Москва, 115522;
[5]ООО «К-Скай», набережная Варкауса, 17, Петрозаводск, 185031;
[6]Центральный научно-исследовательский институт организации и информатизации здравоохранения Минздрава России, ул. Добролюбова, 11, Москва, 127254;
[7]Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы, ул. Петровка, 24, стр. 1, Москва, 127051

В настоящее время идет активная разработка программных продуктов для применения в медицине. Среди них доминирующую долю занимают системы поддержки принятия врачебных решений (СППВР), которые могут быть интеллектуальными (основанными на математических моделях, полученных методами машинного обучения, или на других технологиях искусственного интеллекта) или неинтеллектуальными. Государственная регистрация СППВР как программных медицинских продуктов предусматривает проведение клинических испытаний, протокол которых разрабатывается совместно разработчиком и уполномоченной медицинской организацией. Одним из обязательных компонентов протокола является расчет объема выборки.

В данной статье рассмотрен расчет объема выборки для наиболее распространенного случая — бинарного отклика в диагностических/скрининговых и прогностических системах. Для диагностических/скрининговых моделей рассмотрены случаи несравнительного исследования, сравнительного исследования с проверкой гипотезы превосходства, сравнительного исследования с проверкой гипотезы не меньшей точности в исследованиях одномоментного дизайна. Для прогностических моделей рассмотрены случаи рандомизированных контролируемых испытаний комплексного вмешательства «прогноз + прогноз-зависимое ведение пациента» с проверкой гипотезы превосходства и не меньшей точности.

Подчеркивается, что не менее важным, чем объем выборки, аспектом клинических испытаний является также репрезентативность выборки и другие компоненты дизайна. Они даже более важны, так как систематические ошибки в клинических испытаниях первичны, и самый изощренный статистический анализ не может возместить дефекты дизайна. Редукция клинических испытаний до внешней валидизации моделей (оценки метрик точности на внешних данных) представляется совершенно необоснованной.

Для контактов: Гусев Александр Владимирович, e-mail: agusev@webiomed.ai

Рекомендуется проводить клинические испытания с адекватным задачам дизайном, с тем чтобы далее был возможен клинико-экономический анализ и комплексная оценка медицинских технологий.

Описанные в статье методы расчетов объема выборки потенциально могут быть применены и к более широкому спектру медицинских изделий.

Ключевые слова: системы поддержки принятия врачебных решений; диагностические модели; прогностические модели; объем выборки; бинарный исход; клинические испытания; внешняя валидация.

Как цитировать: Rebrova O.Yu., Gusev A.V. Sample size calculation for clinical trials of medical decision support systems with binary outcome. Sovremennye tehnologii v medicine 2022; 14(3): 6, https://doi.org/10.17691/stm2022.14.3.01

English

Sample Size Calculation for Clinical Trials of Medical Decision Support Systems with Binary Outcome

O.Yu. Rebrova, MD, DSc, Professor, Department of Medical Cybernetics and Informatics[1]; Professor, Department of Endocrinology, Institute for Higher Education and Additional Professional Training[2]; Leading Researcher, Sector of Dynamic Neural Networks, Department of Neuroinformatics, Center for Optical-Neuron Technologies[3]; Chief Researcher, Laboratory of Evidence-Based Medicine and Biostatistics[4];
A.V. Gusev, PhD, Head of Business Development[5]; Senior Researcher, Department of Scientific Fundamentals of Health Organization[6]; Expert of the Sector of Clinical and Technical Trials[7]

[1]Pirogov Russian National Research Medical University, 1 Ostrovityanova St., Moscow, 117997, Russia;
[2]Endocrinology Research Centre, 11 Dmitriya Ulyanova St., Moscow, 117292, Russia;
[3]Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences", 36/1 Nakhimovsky Prospect, Moscow, 117218, Russia;
[4]Mental Health Research Center, 34 Kashirskoye Shosse, Moscow, 115522, Russia;
[5]K-SkAI LLC, 17 Naberezhnaya Varkausa, Petrozavodsk, The Republic of Karelia, 185031, Russia;
[6]Russian Research Institute of Health, 11 Dobrolyubova St., Moscow, 127254, Russia;
[7]Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, 24/1 Petrovka St., Moscow, 127051, Russia

Currently, software products for use in medicine are actively developed. Among them, the dominant share belongs to clinical decision support systems (CDSS), which can be intelligent (based on mathematical models obtained by machine learning methods or other artificial intelligence technologies) or non-intelligent. For the state registration of CDSSs as software medical products, clinical trials are required, and the protocol of trial is developed jointly by the developer and an authorized medical organization. One of the mandatory components of the protocol is the calculation of the sample size.

This article discusses the calculation of the sample size for the most common case, the binary outcome in diagnostic/screening and predictive systems. For diagnostic/screening models, cases of a non-comparative study, comparative study with testing of the superiority hypothesis, comparative study with testing of a hypothesis of non-inferiority in cross-sectional studies are considered. For predictive models, cases of randomized controlled trials of the complex intervention "prediction + prediction-dependent patient management" with testing of the hypothesis of superiority and non-inferiority are considered.

It is emphasized that representativeness of the sample and other design components are no less important in clinical trials than sample size. They are even more important since systematic biases in clinical trials are primary, and even the most sophisticated statistical analysis cannot compensate for design defects. The reduction of clinical trials to external validation of models (i.e. evaluation of accuracy metrics on external data) seems completely unreasonable. It is recommended to perform clinical trials with the design adequate to the tasks, so that further clinical and economic analysis and comprehensive assessment of medical technologies are possible.

The sample size calculation methods described in the article can potentially be applied to a wider range of medical devices.

Key words: clinical decision support systems; diagnostic models; predictive models; sample size; binary outcome; clinical trials; external validation.

Введение

В настоящее время идет активная разработка программных продуктов для применения в медицине. Среди них доминирующую долю занимают системы поддержки принятия врачебных решений (СППВР), которые могут быть интеллектуальными (основанными на математических моделях, полученных методами машинного обучения, или на других технологиях искусственного интеллекта) или неинтеллектуальными.

Согласно действующему в России законодательству, такое программное обеспечение подлежит государственной регистрации в качестве медицинского изделия [1, 2], для чего в свою очередь необходимо проведение клинических испытаний. Целью этих испытаний является оценка эффективности и безопасности медицинского изделия в части программного обеспечения [2], причем они проводятся в двух формах:

1) исследования (анализ и оценка клинических данных);

2) испытания.

Содержание термина «исследование» в приказе Министерства здравоохранения РФ от 30 августа 2021 г. №885 [2] не определено. Согласно рекомендациям International Medical Device Regulators Forum [3], под ним понимается «клиническая оценка» как совокупность оценки достоверности клинической связи, аналитической валидации и клинической валидации.

Как правило, программа клинических испытаний создается разработчиком СППВР совместно с внешней медицинской организацией, имеющей в соответствии с действующим нормативным регулированием право на проведение таких испытаний.

Несмотря на то, что целью клинических испытаний является оценка эффективности и безопасности СППВР, фактически вместо нее в настоящее время проводится внешняя валидация СППВР, которая позволяет оценить, будут ли достигнуты заявленные производителем метрики качества работы модели на данных, которые не были использованы при создании или тестировании такой модели.

Из литературы [4–6] и нашего опыта известно, что при использовании моделей в условиях реальной клинической практики возможна деградация метрик их точности (в частности, чувствительности — Ч, специфичности — С) в силу того, что модели начинают работать на неизвестных для них клинических случаях, которые отсутствовали в обучающем и/или тестовом наборах данных. Это связано в том числе и с тем, что нередко модели разрабатываются на данных одного медицинского учреждения, при этом репрезентативность (типичность) использованной выборки зачастую весьма сомнительна, но разработчики об этом не задумываются. В результате обобщаемость этих моделей (надежность их работы в других медицинских учреждениях) бывает довольно низкой. Подчеркнем, что внешняя валидация очень важна, но совершенно недостаточна для оценки эффективности и безопасности СППВР.

При создании СППВР используются в основном модели, созданные для решения диагностических и прогностических задач (например, для выявления патологических изменений на МРТ различных органов, прогнозирования неблагоприятных событий). На этапе планирования клинических испытаний и внешней валидации одним из сложных практических вопросов является определение размера выборки пациентов, необходимой и достаточной для формирования надежных выводов о качестве работы диагностической или прогнозной модели. В конечном счете это необходимо для формирования доверия к результатам клинических испытаний программного продукта, содержащего данную модель, в процессе регистрации его в качестве медицинского изделия. Таким образом, корректность определения размера выборки может стать важнейшим фактором успешности проведения клинических испытаний и последующего получения разработчиком регистрационного удостоверения Росздравнадзора.

Важно отметить, что объем является не единственной принципиально важной характеристикой выборки, важна также ее репрезентативность [7]. Репрезентативность выборки может быть успешно достигнута при вероятностном способе ее формирования (случайный отбор, систематический отбор, кластерный отбор и т.д.), однако на практике такой подход в большинстве случаев невозможен. Обычно используются следующие не вероятностные способы формирования выборок: выборка удобства (произвольная), последовательная (сплошная) выборка, выборка добровольцев, квотный отбор и т.д. В этой ситуации важно обращать внимание на преваленс (П, долю, частоту) диагностируемого (или прогнозируемого) состояния в обучающей и валидационной выборках, так как именно от него зависят наиболее важные операционные характеристики (метрики точности) модели.

Следует формировать выборку из той же популяции пациентов, в которой предполагается применять разработанную СППВР, — целевой популяции. Так, если модель, предназначенная для встраивания в программный продукт, разработана на данных пациентов из историй болезни стационара, то и валидировать, и применять данный продукт следует для таких же пациентов, а не для пациентов, например, поликлиники. В идеале и обучающая выборка должна быть репрезентативной хотя бы в отношении преваленса диагностируемого (прогнозируемого) состояния. Однако зачастую разработчики специально добиваются сбалансированности обучающей выборки (равенства объемов распознаваемых классов), так как в этом случае они получают модели с более высокими оценками метрик точности. При этом часто разработчики не осознают, что неудача настигнет их позже, когда модель почти гарантированно окажется неработоспособной в реальной практике, где такая же сбалансированность классов не будет встречаться.

Описание в публикациях и отчетах разработки и валидации моделей диагностики и прогноза в целом должно соответствовать современным рекомендациям STARD [8] и TRIPOD [9]. В обоих документах отмечается, что должен быть описан расчет объема выборки. В литературе обсуждаются различные способы определения достаточного размера выборки для целей внешней валидации прогнозных и диагностических моделей искусственного интеллекта в зависимости от изучаемого исхода (синонимы: отклика,

функции, зависимой переменной, выходного показателя) [10–14]. Такими откликами могут быть бинарный признак, категориальный признак (когда распознаются три или более класса/события), количественный признак, время до события.

В данной работе мы рассматриваем подходы к расчету объема выборки в клинических испытаниях с целью оценки эффективности и безопасности диагностических и прогностических моделей с бинарным откликом.

Диагностические модели

Эффективность и безопасность диагностических моделей следует изучать в одномоментном исследовании. Его основные черты следующие: диагностические методы (как минимум новый и референсный) применяются у пациента одновременно (с минимальным интервалом времени), а их результаты взаимно ослепляются. Схема дизайна одномоментных исследований может быть легко понята из шаблона для их описания в публикациях STARD [8].

Выбор метода расчета необходимого объема выборки для бинарного отклика в задаче диагностики/скрининга зависит от ответа на следующие вопросы:

1. Является исследование сравнительным или нет?
2. Если исследование сравнительное, какая гипотеза проверяется?

Соответственно в зависимости от ответов на эти вопросы возникают следующие ситуации:

1. Бинарный признак, несравнительное исследование — сопоставление с референсным методом.
2. Бинарный признак, сравнительное исследование:

а) проверяется гипотеза превосходства точности/безопасности над существующим методом решения задачи при сопоставлении с референсным методом;

б) проверяется гипотеза не меньшей точности/безопасности метода по отношению к существующему методу решения задачи при сопоставлении с референсным методом.

В принципе в сравнительном исследовании возможна также проверка гипотезы эквивалентности точности нового метода и точности существующего метода при сопоставлении с референсным методом. Однако такая ситуация является крайне редкой, поэтому здесь мы ее рассматривать не будем.

Прежде чем перейти к рассмотрению методов расчета, перечислим основные метрики оценки качества диагностических моделей с бинарным откликом в одномоментных исследованиях:

1. Оценки чувствительности (в машинном обучении обычно называется откликом, *англ*. recall) и специфичности — устойчивых, не зависящих от преваленса (частоты) идентифицируемого состояния в целевой популяции операционных характеристик модели и их доверительных интервалов (confidence interval, CI) — 95%-го, а лучше — 99%-го. Отметим, что чувствительность и специфичность изменяются реципрокно, и потому, оптимизируя один показатель, мы ухудшаем другой.

2. Оценки прогностической ценности положительного и отрицательного результатов (ПЦПР и ПЦОР соответственно) — зависящих от преваленса операционных характеристик модели и их CI (95%, 99%).

В машинном обучении ПЦПР обычно называется английским термином precision. В случае, если выборка репрезентативна по отношению к целевой популяции в отношении преваленса (так происходит обычно при использовании сплошного или случайного методов формирования выборки), расчет прогностических ценностей прост. Однако, если в исследование отдельно набирались позитивные и негативные случаи, необходима поправка на преваленс:

$$ПЦПР=Ч\cdot П/[Ч\cdot П+(1-С)\cdot(1-П)];$$
$$ПЦОР=С\cdot(1-П)/[С\cdot(1-П)+(1-Ч)\cdot П].$$

Прогностические ценности крайне важны, так как именно с ними работает врач, оценивая результат диагностики или прогноза конкретного больного с учетом вероятности гипер- и гиподиагностики.

В ряде случаев оценивают также общую точность модели — отношение суммы истинно-положительных и истинно-отрицательных результатов к общему числу наблюдений в выборке. В машинном обучении эта метрика обычно называется accuracy. Иногда под точностью также понимают среднее между значениями Ч и С. Кроме того, для точности могут быть рассчитаны CI (95%, 99%). Точность тоже преваленс-зависима, соответственно не может быть рассчитана для нерепрезентативной выборки без поправки на преваленс. Этот показатель является слишком общим, недостаточно понятным врачам, его использовать не рекомендуется.

Общим характером обладает и такой популярный показатель, как площадь под характеристической кривой — AUROC. Подчеркнем, что этот показатель не является бинарным и, соответственно, для него неприменимы те расчеты объема выборки, о которых далее пойдет речь. ROC-анализ может проводиться как в координатах [(1−С); Ч], так и в координатах [(1−ПЦОР); ПЦПР]. Последний анализ предпочтительнее, так как ориентирован на врача — лицо, принимающее решение в отношении конкретного пациента. Часто ROC-анализ используют для предварительного сравнения точности изучаемых моделей, особенно если их много (что часто бывает при построении моделей машинного обучения). Однако ROC-анализа совершенно недостаточно для доказательства эффективности модели.

Далее должен быть выполнен поиск отрезной точки, если отклик модели имеет непрерывную область значений. Критериями оптимальности отрезной точки могут быть:

1) минимум ошибки I рода (гипердиагностики) при приемлемой ошибке II рода (гиподиагностике);

2) минимум ошибки II рода (гиподиагностики) при приемлемой ошибке I рода (гипердиагностике);

3) оптимизация их соотношения;

4) максимизация их суммы (критерий Юдена) и т.д.

После определения отрезной точки должен последовать расчет показателей Ч, С, ПЦПР и ПЦОР для этой выбранной точки. Обычно в задачах диагностики оптимизируют Ч и/или ПЦПР и жертвуют (до приемлемого значения) С и ПЦОР. В задачах скрининга наоборот: оптимизируют С и/или ПЦОР при приемлемых значениях Ч и ПЦПР. Заметим, что в принципе возможно так называемое одностороннее использование модели (бинарного классификатора), например использование модели только на подтверждение искомого состояния (т.е. для диагностики), если ПЦПР высокая, а ПЦОР — низкая, и при этом цена ошибок II рода (гиподиагностики) невелика. И наоборот, можно использовать для скрининга модель, у которой высокая ПЦОР и низкая ПЦПР, если цена ошибок гипердиагостики невелика.

Бинарный признак, несравнительное исследование способа диагностики/скрининга

В исследовании нового (индексного) способа диагностики референсным методом должен являться лучший из имеющихся на данный момент методов диагностики состояния. При этом принимается допущение, что референсный метод обеспечивает 100% точность диагностики по всем операционным показателям. Таким методом в медицине обычно считается гистологическое исследование, однако оно инвазивно и в большинстве случаев не может быть использовано. Тем не менее обоснование выбора референсного метода следует приводить в отчетах и публикациях.

Показатели точности диагностического теста — Ч, С, ПЦПР, ПЦОР — являются пропорциями (долями), для которых необходимо оценивать CI (обычно используется доверительная вероятность 95%). Именно нижняя граница CI должна быть поставлена как ориентир при расчете объема выборки. Обычно следует стремиться к тому, чтобы эта граница не была ниже 85%. То есть модель хороша, если CI для любого из показателей лежит в диапазоне 85–100%. При этом очевидно, что если CI включает 50% или даже приближается к этой величине, то модель неработоспособна, вместо ее использования проще положиться на случай, бросив монетку.

Таким образом, расчет необходимого объема выборки в этом случае сводится к решению обратной задачи — расчету CI (обычно 95%-го) для ожидаемого значения показателя. При этом целевое значение показателя должно быть задано на основании клинической значимости, т.е. врачами, а не статистиками. Это означает, что именно врачами должно быть задано минимально приемлемое значение показателя точности диагностики при гипотетической 100% точности референсного метода. Затем должно быть задано и приемлемое значение 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 предъявляться к ПЦПР (с учетом преваленса), если решается задача диагностики, т.е. выявления состояния с высоким преваленсом в целевой популяции. Если же решается задача скрининга, т.е. выявления состояния с низким преваленсом в целевой популяции, следует прежде всего ориентироваться на ПЦОР. Показатели Ч и С менее важны с практической точки зрения применения СППВР, при этом показатель Ч ассоциирован с ПЦПР, а С — с ПЦОР.

Расчет CI «вручную» сложен, поэтому формулу мы здесь не приводим. В любом профессиональном пакете программ, конечно, есть удобные опции для таких расчетов. Однако можно воспользоваться не очень удобным, но надежным онлайн-калькулятором https://www.graphpad.com/quickcalcs/confInterval1/ (хотя есть много и других подобных калькуляторов), используя процедуру «математического подгона» значений числителя и знаменателя для заданной доли.

Пример 1. Врачами заданы приемлемые значения показателей: ПЦПР — 90% и ПЦОР — 80%. Это означает, что нижняя граница CI для ПЦПР должна быть не меньше 90%, для ПЦОР — не меньше 80%. Тогда приближенно можно считать, что точечная оценка ПЦПР располагается в середине интервала между 90 и 100%, т.е. равна 95%. Заметим, что для малых выборок такое допущение не обосновано. При условии, что будущая выборка будет являться репрезентативной как минимум с точки зрения преваленса искомого состояния, необходимый объем выборки, полученный с помощью приведенного калькулятора, будет составлять 150 пациентов, так как 95% CI 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, рассчитанный с помощью того же калькулятора, составит 63 — для получения доли 90% с 95% CI (80,5%; 95,9%). После суммирования 150+63 получаем число 213 как итоговое значение.

Далее полученный объем выборки должен быть распределен между положительными и отрицательными случаями (определенными референсным методом) в соответствии с преваленсом состояния в целевой популяции.

Пример 2 (в продолжение Примера 1). Если преваленс искомого состояния в целевой популяции 60% (0,6), то в группу случаев должно быть включено 213·0,6=128 пациентов, в группу сравнения — 213−128=85 пациентов. Если преваленс состояния в целевой популяции 10% (0,1), то распределение будет другим: 213·0,1=21 пациент — в группе случаев, 213−21=192 пациента — в группе контроля.

Следует подчеркнуть, что если модель разрабатывалась на так называемой сбалансированной

обучающей выборке, т.е. эта выборка была нерепрезентативной в отношении преваленса, то полученные при внутреннем тестировании оценки ПЦПР и ПЦОР смещены — и тем более, чем сильнее фактический преваленс отклоняется от соотношения объемов групп в обучающей выборке. Вследствие этого получить такие же значения на правильно сформированной выборке в клинических испытаниях будет сложно, если не невозможно. При этом показатели Ч и С не зависят от преваленса и поэтому их легче воспроизвести, однако они практического значения для врачей не имеют.

Внешняя валидизация модели диагностики/скрининга, которая зачастую, к сожалению, подменяет клинические испытания этой модели, фактически соответствует вышеописанному дизайну несравнительного исследования: формируется выборка позитивных и негативных случаев и рассчитываются Ч и С. Это в принципе можно считать приемлемым, если соблюдены следующие принципы:

1) выборка должна формироваться строго из целевой популяции;

2) должен быть использован надежный референсный тест;

3) соотношение позитивных и негативных случаев должно соответствовать преваленсу искомого состояния в целевой популяции;

4) необходим расчет не только показателей Ч и С, но и ПЦПР и ПЦОР;

5) для всех операционных характеристик должны быть рассчитаны 95% CI;

6) должна быть оценена безопасность, прежде всего — последствия ошибок гипо- и гипердиагностики.

Бинарный признак, сравнительное исследование, гипотеза превосходства точности/безопасности способа диагностики/скрининга

Проверяется гипотеза превосходства модели над существующим методом решения задачи при сопоставлении с референсным методом. Таким образом, диагностика осуществляется тремя методами — референсным, новым и старым (рутинно используемым, предлагаемым к замене).

В этом случае расчет объема выборки основывается на клинически значимой величине превосходства нового метода над старым. Основными параметрами расчета являются:

1) ошибка I рода (альфа) — обычно устанавливается 5%;

2) статистическая мощность — рекомендуется 90%, минимально — 80%;

3) значение выбранного показателя оценки (например, ПЦПР) для старого метода;

4) значение выбранного показателя оценки для нового метода.

Заметим, что точность рутинно используемого метода далеко не всегда известна. В этом случае необходимо провести предварительное исследование с оценкой его точности.

Если превышение точности нового метода над старым ожидается небольшим (например, 5%), то такой метод вряд ли будет внедрен в медицинскую практику. Зачастую новый метод дороже рутинно используемого, соответственно в этой ситуации за проведением клинического испытания должен последовать клинико-экономический анализ, в ходе которого оценен инкрементальный показатель «стоимость/эффективность». Другими словами, должно быть определено, является ли приращение стоимости обоснованным с точки зрения приращения точности диагностики. Кроме того, медицинская практика в принципе весьма консервативна, и небольшое повышение точности может не стать серьезным аргументом в пользу внедрения нового метода. Соответственно для расчета выборки необходимо установление врачами того минимального значения точности (по сравнению с рутинным методом), которое убедит их использовать новый, потенциально более точный метод.

Расчет объема выборки возможен в различных статистических пакетах, однако можно пользоваться и надежными онлайн-калькуляторами, например https://sealedenvelope.com/power/binary-superiority/ (конечно, со ссылкой на калькулятор и ту литературу, которая лежит в основе расчетов и указана на веб-странице калькулятора).

Пример 3. Разработан новый метод диагностики, который превышает старый по точности на 10%. Точность старого метода (в интерфейсе control group) — 80%, нового — 90%. Тогда необходимый объем выборки (при ошибке I рода 5% и статистической мощности 90%) — 263 пациента.

Бинарный признак, сравнительное исследование, гипотеза не меньшей точности/безопасности способа диагностики/скрининга

Проверяется гипотеза, что точность нового метода не меньше точности старого метода. Конечно, может возникнуть вопрос, зачем тогда вообще нужен новый метод. Однако для любых медицинских технологий важна не только эффективность (в случае диагностики или скрининга — точность), но и безопасность. Тогда повышение безопасности может быть также доказано в клинических испытаниях — при проверке гипотезы превосходства в отношении критерия (или нескольких критериев) безопасности. Особенно это важно, если старый метод — инвазивный или при его использовании необходимо облучение пациента. Кроме того, имеет значение и экономический аспект. Так, новый метод может быть дешевле при прежней точности, что создаст аргументацию в пользу внедрения этого нового метода.

В данном случае диагностика также осуществляется тремя методами — референсным, новым и старым (рутинно используемым, предлагаемым к замене). Расчет объема выборки тоже основывается на клини-

чески значимой величине превосходства нового метода над старым. Основными параметрами расчета являются:

1) ошибка I рода (альфа) — обычно устанавливается 5%;

2) статистическая мощность: рекомендуется 90%, минимально — 80%;

3) значение выбранного показателя оценки (например, ПЦПР) для старого метода;

4) значение выбранного показателя оценки для нового метода;

5) порог не меньшей точности/безопасности.

Последний параметр показывает разность между значениями оцениваемого показателя, которая может считаться врачами допустимой. Например, если новый метод должен быть строго таким же, как и старый (точности старого и нового методов — 80%), то порог равен нулю. Для доказательства этого потребуется бесконечное число наблюдений. Увеличивая порог, мы допускаем, что все же новый метод может быть несколько хуже старого. Чем больше эта разница, тем легче доказать не меньшую эффективность, так как необходимый объем выборки будет снижаться.

Расчет объема выборки возможен в онлайн-калькуляторе на том же онлайн-сервисе (https://sealedenvelope.com/power/binary-noninferior/).

Пример 4. Точность нового и старого методов установлена на уровне 80%, порог — 5% при ошибке I рода 5% и статистической мощности 90%. Тогда потребуется 1097 пациентов для доказательства данной гипотезы. Если порог установить на 7%, необходимое число пациентов в выборке будет почти вдвое меньше — 560.

Возможно, что новый метод чуть лучше старого на клинически значимую величину (например, на 2%). Тогда необходимый объем выборки будет меньше.

Пример 5. Точность нового метода — 82%, старого — 80%, порог — 5%, ошибка I рода — 5%, статистическая мощность — 90%. Тогда потребуется 538 пациентов.

Отметим, что объем выборки при проверке гипотезы не меньшей эффективности всегда значительно больше, чем при проверке гипотезы превосходства.

Прогностические модели

Такие модели оценивать значительно сложнее. Прежде всего надо определить, как будет использоваться прогноз. Обычно он применяется для изменения тактики ведения пациентов по сравнению с рутинно используемой в настоящее время и в целях вторичной/третичной профилактики. Таким образом, тестирование такой модели фактически заключается в тестировании комбинированной медицинской технологии «прогноз + прогноз-зависимое ведение пациента». Причем если прогноз осуществляется точно, но способов воздействия на пациента (например, для предотвращения неблагоприятных событий) нет или они малоэффективны, то смысла в таком прогнозе нет. Более того, негативный прогноз будет вреден для психики пациента, если последнего информируют о нем.

Подчеркнем еще раз, что важно наличие эффективного способа воздействия именно на той стадии заболевания/жизни, на которой осуществляется прогноз. Известно, что методы лечения, эффективные на поздних стадиях болезни, могут быть совершенно бесполезными на ранних стадиях заболевания. Таким образом, начиная решать задачу прогнозирования, нужно сначала убедиться в том, что существуют эффективные способы предупреждения прогнозируемых неблагоприятных событий.

Еще одним важным параметром прогностических моделей является предельный срок прогнозирования, который зависит от содержания конкретной задачи. Конечно, речь идет о сроке прогноза не ровно на 1 год, 5 лет и т.д., а на период до 1 года, до 5 лет и т.д. Чем меньше срок прогноза, тем легче его построить — это связано с полнотой данных, отсутствием исторического смещения и т.д. Например, прогнозировать исход госпитализации в связи с острым заболеванием гораздо легче, чем предсказывать инфаркт миокарда на срок до 5 лет.

Прогностические модели должны оцениваться с помощью другого дизайна исследований — в рандомизированном контролируемом испытании (а не одномоментном, как методы диагностики/скрининга).

Основные черты таких исследований:

1) определяется целевая популяция, синхронизированная по какому-либо событию (впервые установленный диагноз, достижение определенного возраста, выполнение хирургического вмешательства и т.д.);

2) объекты целевой популяции после подписания информированного согласия рандомизируются в основную и контрольную группы;

3) в основной группе для всех пациентов выполняется прогнозирование, и в случае получения неблагоприятного прогноза к пациенту применяется модифицированная по сравнению с рутинной тактика ведения (например, более частые визиты к врачу для раннего обнаружения рецидивов заболевания после хирургической операции); при благоприятном прогнозе применяется рутинная или упрощенная тактика ведения пациента;

4) в контрольной группе прогнозирование не осуществляется, тактика ведения — рутинная;

5) устанавливается период наблюдений, в течение которого фиксируются возникающие неблагоприятные события в каждой из групп. Длительность наблюдения должна быть такой, чтобы возникло достаточное число прогнозируемых событий в контрольной группе.

Метриками эффективности модели в таких испытаниях являются два показателя — относительный риск и снижение абсолютного риска. В случае если было большое выбывание из исследования (а это неизбежный спутник длительного наблюдения, необходимого при медленно накапливающихся событиях), то требу-

ется оценивать другой показатель — отношение угроз (мы не останавливаемся на этом случае в настоящей статье). Фактически здесь проверяется обычно одна из двух упомянутых гипотез — превосходства или не меньшей эффективности.

Рассмотрим перечисленные выше ситуации последовательно.

Бинарный признак, испытание способа прогноза, гипотеза превосходства

При разработке прогностических моделей обычно предполагают, что при наличии прогноза удастся улучшить исходы пациента. Обычно прогнозируют неблагоприятное событие с целью снизить его частоту в основной группе по сравнению с контрольной за счет применения некой медицинской технологии профилактики — вторичной (предотвращения заболевания) или третичной (предотвращения осложнений, рецидивов, обострений, инвалидизирующего течения заболевания).

В этом случае расчет объема выборки проводится так же, как описано выше для гипотезы превосходства, однако здесь уже необходимы две выборки, каждая из которых будет состоять из вычисленного количества пациентов. Вычисления можно делать в калькуляторе https://sealedenvelope.com/power/binary-superiority/. В принципе возможно формирование и неравных по объему выборок (например, в отношении 3:1), однако статистическая мощность при этом падает и, следовательно, потребуется больший объем выборки.

Пример 6. В контрольной группе заболевание возникает у 20% пациентов, в основной группе хотелось бы, чтобы оно возникало не более чем у 10% пациентов (последнее значение выбирается в соответствии с ожиданиями врачей, т.е. 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 составит 263 пациента.

Бинарный признак, испытание способа прогноза, гипотеза не меньшей эффективности

В таких испытаниях обычно предполагается, что при наличии прогноза удастся упростить ведение пациента, не ухудшив исходы (развитие заболевания, осложнения и т.д.). Например, можно приглашать пациента на визиты не 1 раз в год после операции, а 1 раз в 2 года, не ухудшая исхода.

Расчет объема выборки проводится так же, как описано выше для гипотезы не меньшей эффективности, однако здесь необходимы две выборки, каждая из которых должна состоять из вычисленного количества пациентов. Вычисления можно делать в калькуляторе https://sealedenvelope.com/power/binary-noninferior/.

Пример 7. В контрольной и опытной группах заболевание возникает у 20% пациентов (т.е. доля «успеха» в обеих группах — 80%), при этом порог не меньшей эффективности устанавливается на уровне 5%. Тогда при ошибке I рода 5% и статистической мощности 90% необходимый объем каждой из групп — 1097 пациентов.

Внешняя валидизация прогностической модели, к которой (необоснованно) сводятся в настоящее время клинические испытания прогностических моделей, проводится в дизайне ретроспективного исследования случай–контроль: формируются основная и контрольная выборки по наличию/отсутствию прогнозируемого исхода (события), извлекаются данные пациентов за период, соответствующий сроку прогноза, и оценивается точность прогноза (в терминах AUROC, Ч, С). Такой подход чреват серьезными систематическими ошибками, не позволяющими правильно оценить эффективность и безопасность модели, в частности:

1) когорта не синхронизирована;

2) в анализе не участвуют пациенты с пропусками (как в данных, используемых для прогноза, так и в исходах);

3) срок прогноза фиксирован, в то время как события у пациентов возникают в разные сроки;

4) игнорируются все медицинские вмешательства, которые применялись к пациентам за этот срок.

Таким образом, дизайн ретроспективного исследования случай–контроль совершенно не годится для оценки прогнозной модели. Паллиативной мерой могло бы быть проведение ретроспективного когортного исследования, основные черты дизайна которого следующие:

1) определяются целевая популяция и критерий синхронизации когорты;

2) по данным в точке синхронизации строится прогноз, позитивные случаи относятся в основную группу, негативные — в группу контроля;

3) сравниваются частоты исхода в группах, возникшие за срок прогнозирования, и прогнозы; при этом учитываются выпадающие наблюдения и примененные медицинские вмешательства.

Метриками качества прогноза в этой ситуации являются показатели Ч, С, ПЦПР и ПЦОР, а расчет размера выборки сводится к случаю несравнительного исследования способа диагностики/скрининга. При прогнозировании часто предпочтение отдается гипердиагностике как консервативной тактике.

Конечно, ретроспективное когортное исследование является лишь меньшим из зол, так как и оно не позволяет получить несмещенные оценки эффективности и безопасности прогнозной модели.

Заключение

Мы рассмотрели расчет размера выборки для наиболее распространенного вида моделей — с бинарным исходом. Однако при любом расчете все же же-

лательно увеличивать полученное число пациентов на 5–10% для надежности, особенно если возможно их выбывание или статистическая мощность установлена на уровне 80%.

Расчет объема выборки — лишь один из компонентов протокола клинических испытаний, которые планируются совместно разработчиком и уполномоченной медицинской организацией. Другие аспекты дизайна клинических испытаний важны не менее и даже более, так как систематические ошибки в клинических испытаниях первичны и самый изощренный статистический анализ не может возместить дефекты дизайна. Редукция клинических испытаний до внешней валидизации моделей представляется совершенно необоснованной. Рекомендуется проводить клинические испытания с адекватным задачам дизайном, с тем чтобы далее были возможны клинико-экономический анализ и комплексная оценка медицинских технологий.

Финансирование исследования. Авторы заявляют об отсутствии финансирования для выполнения работы.

Конфликт интересов. Авторы декларируют отсутствие конфликтов интересов.

Литература/References

1. 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.

Gusev A.V., Morozov S.P., Kutichev V.A., Novitsky R.E. Legal regulation of artificial intelligence software in healthcare in the Russian Federation. *Medicinskie tehnologii. Ocenka i vybor* 2021; 1: 36–45, https://doi.org/10.17116/medtech20214301136.

2. *Приказ Министерства здравоохранения РФ от 30 августа 2021 г. №885 «Об утверждении Порядка проведения оценки соответствия медицинских изделий в форме технических испытаний, токсикологических исследований, клинических испытаний в целях государственной регистрации медицинских изделий».* URL: https://docs.cntd.ru/document/608935477.

Prikaz Ministerstva zdravookhraneniya RF ot 30 avgusta 2021 g. No.885 "Ob utverzhdenii Poryadka otsenki sootvetstviya meditsinskikh izdeliy v forme tekhnicheskikh ispytaniy, toksikologicheskikh issledovaniy, klinicheskikh ispytaniy v tselyakh gosudarstvennoy registratsii meditsinskikh izdeliy" [Order of the Ministry of Health of the Russian Federation of August 30, 2021 Np.885 "On approval of the Procedure for assessing the conformity of medical devices in the form of technical tests, toxicological studies, clinical trials for the purpose of state registration of medical devices"]. URL: https://docs.cntd.ru/document/608935477.

3. *MDRF/SaMD WG/N41FINAL:2017. Software as a Medical Device (SaMD): Clinical Evaluation.* URL: http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf.

4. Wallert J., Tomasoni M., Madison G., Held C. Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data. *BMC Med Inform Decis Mak* 2017; 17(1): 99, https://doi.org/10.1186/s12911-017-0500-y.

5. Ye C., Fu T., Hao S., Zhang Y., Wang O., Jin B., Xia M., Liu M., Zhou X., Wu Q., Guo Y., Zhu C., Li Y.M., Culver D.S., Alfreds S.T., Stearns F., Sylvester K.G., Widen E., McElhinney D., Ling X. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. *J Med Internet Res* 2018; 20(1): e22, https://doi.org/10.2196/jmir.9268.

6. Park J., Kim J.W., Ryu B., Heo E., Jung S.Y., Yoo S. Patient-level prediction of cardio-cerebrovascular events in hypertension using nationwide claims data. *J Med Internet Res* 2019; 21(2): e11757, https://doi.org/10.2196/11757.

7. Реброва О.Ю. Жизненный цикл систем поддержки принятия врачебных решений как медицинских технологий. *Врач и информационные технологии* 2020; 1: 27–37, https://doi.org/10.37690/1811-0193-2020-1-27-37.

Rebrova O.Yu. Life cycle of decision support systems as medical technologies. *Vrac i informacionnye tehnologii* 2020; 1: 27–37, https://doi.org/10.37690/1811-0193-2020-1-27-37.

8. Bossuyt P.M., Reitsma J.B., Bruns D.E., Gatsonis C.A., Glasziou P.P., Irwig L., Lijmer J.G., Moher D., Rennie D., de Vet H.C.W., Kressel H.Y., Rifai N., Golub R.M., Altman D.G., Hooft L., Korevaar D.A., Cohen J.F.; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015; 351: h5527, https://doi.org/10.1136/bmj.h5527.

9. Collins G.S., Reitsma J.B., Altman D.G., Moons K.G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; 350: g7594, https://doi.org/10.1136/bmj.g7594.

10. Snell K.I.E., Archer L., Ensor J., Bonnet L., Debray T.P.A., Philips B., Collins G.S., Riley R.D. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *J Clin Epidemiol* 2021; 135: 79–89, https://doi.org/10.1016/j.jclinepi.2021.02.011.

11. Riley R.D., Debray T.P.A., Collins G.S., Archer L., Ensor J., van Smeden M., Snell K.I.E. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med* 2021; 40(19): 4230–4251, https://doi.org/10.1002/sim.9025.

12. Archer L., Snell K.I.E., Ensor J., Hudda M.T., Collins G.S., Riley R.D. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Stat Med* 2021; 40(1): 133–146, https://doi.org/10.1002/sim.8766.

13. Riley R.D., Collins G.S., Ensor J., Archer L., Booth S., Mozumder S.I., Rutherford M.J., van Smeden M., Lambert P.C., Snell K.I.E. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat Med* 2022; 41(7): 1280–1295, https://doi.org/10.1002/sim.9275.

14. Feng D., Cortese G., Baumgartner R. A comparison of confidence/credible interval methods for the area under the ROC curve for continuous diagnostic tests with small sample size. *Stat Methods Med Res* 2017; 26(6): 2603–2621, https://doi.org/10.1177/0962280215602040.